ADVERSARIAL ROBUSTNESS IN CLASSIFICATION VIA THE LENS OF OPTIMAL TRANSPORT: THEORY AND NUMERICS

by

Jakwang Kim

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2023

Date of final oral examination: 05/01/2023

The dissertation is approved by the following members of the Final Oral Committee:
Nicolás García Trillos, Assistant Professor, Statistics
Robert D. Nowak, Professor, Electrical and Computer Engineering
Xiaojin Jerry Zhu, Professor, Computer Science
Hanbaek Yu, Assistant Professor, Mathematics
Yiqiao Zhong, Assistant Professor, Statistics

To my family.

ACKNOWLEDGMENTS

First and foremost, I would like to express my most sincere acknowledgement to my advisor Nicolas García Trillos. It is undoubtedly fortunate to get advised by him. At the beginning of studying this topic, I had failed and made mistakes so many times. But, Nicolas has always been so nice and supportive and give an insightful guidance on our projects. I had a very happy and fruitful experience at Madison and I am deeply grateful to him.

I would also like to express my gratitude to Rob Nowak, Jerry Zhu, Hanbaek Lyu and Yiqiao Zhong for being on my dissertation defense committee members. In particular, Hanbaek Lyu gave a lot of supports in more ways than one: as an expert for both probability and optimization, his comments have been helpful to my Ph.d period.

Also, I also thank Matt Jacobs who is a friend of Nicolas and our collaborator. His insight and knowledge are really essential for our projects. I would also thank Matt Werenski, another collaborator, especially his efforts and talents about coding and computing for our projects.

Finally, special thanks to my parents and family for their support.

CONTENTS

LIST OF FIGURES

ABSTRACT

Through this thesis, I purse rigorous understanding of adversarial training in multiclass classification problem which is now one of the most important tasks in modern machine learning community. Especially, after the great success of deep learning based algorithms, there have been a numerous demands to understand the robustness of machine learning models after their training, especially since the discovery that they exhibit a critical vulnerability to adversarial perturbations that are imperceptible to humans was known. Since 2010 there have been numerous papers regarding adversarial training in order to defend against such adversarial attacks, and hence to obtain more robust machine learning models.

However, in spite of huge efforts devoted to in this field, until very recently no rigorous mathematical understanding was achieved, to the author's best knowledge. Because of the lack of rigorous understanding regarding this problem, many properties, or even the existence of such robust classifiers, is unveiled.

Since this problem has both interesting and important for both theoretical and practical reasons, the ultimate goal of this thesis is not only to provide the mathematically rigorous foundation for the adversarial training in multiclass classification but also to propose practical algorithms. Part of new algorithms we purse in this thesis is guided by the new mathematical framework for adversarial training perspective that we develop.

The contributions of this thesis can be summarized in the following three themes, which we will develop in succeeding three chapters:

1. In chapter 3, *Adversarial learning, Generalized barycenter problem and their connection by multimarginal optimal transport*, the adversarial training problem is connected to multimarginal optimal transport problem. To obtain this beautiful connection, the generalized (Wasserstein) barycenter problem which is indeed the generalization of classical barycenter problem is introduced. Based on that, various equivalent formulas are derived including a multimarginal optimal transport formulation. Through these equivalent formulations we are

able to prove that the adversarial training problem, especially, *distributional-perturbing adversarial training model*, is equivalent to the generalized barycenter problem and the associating multimarginal optimal transport problem. One of advantages of these equivalences is that it allows to use many computational optimal transport tool to calculate the adversarial risk. We will leverage such computational advantages in chapter 5.

2. In chapter 4, *On the existence of solutions to adversarial training in multiclass classification*, the mathematical understanding of three variant adversarial training models is pursued. In particular, we show that the well-posedness of adversarial training models. The existence of Borel measurable optimal robust classifiers is proved in the distributional-perturbing adversarial training model. Furthermore, other two models also impose Borel measurable optimal robust classifiers through the previous result. Lastly, a unifying perspective of all three models is provided, through which we will see that the three models are almost the same.

3. In chapter 5, *Two approaches for computing adversarial training problem based on optimal transport frameworks*, based on the theoretical understanding of the previous two chapters, we propose two numerical implementations for adversarial training. One employs the geometric structure of the generalized barycenter problem in chapter 3 which suggests a way to count all possible interactions efficiently. The other relies on a multimarginal optimal transport formulation of the adversarial training problem, also developed in chapter 3, which implicitly hints the idea of truncation. Numerical results on real data sets obtained by these algorithms are also provided.

1 INTRODUCTION

This thesis is the attempt to provide a rigorous bridge between adversarial learning community and optimal transport community. The aim is to build up the understanding of adversarial training model in classification via the lens of modern optimal transport theory, especially so-called multimarginal optimal transport.

Modern machine learning models, in particular those generated with deep learning, perform remarkably well, in many cases much better than humans, at classifying data in a variety of challenging application fields like image recognition, medical image reconstruction, and natural language processing. However, the robustness of these learning models to data perturbations is a completely different story. For example, in image recognition, it has been widely documented (e.g., Goodfellow et al. (2015) and Zhang et al. (2019); see Figure 1) that certain structured but human-imperceptible modifications of images at the pixel level can fool an otherwise well-performing image classification model. These small data perturbations, known as *adversarial attacks*, when deployed at scale can make a model's prediction accuracy drop substantially and in many cases collapse altogether. As such, they are a significant obstacle to the deployment of machine learning systems in security-critical applications, e.g. Biggio and Roli (2018). To defend against these attacks, many researchers have investigated the problem of adversarial training, i.e., training methods that produce models that are robust to attacks. In adversarial training, one typically pits the adversary against the learner during the training step, forcing the learner to select a model that is robust against attacks. Nonetheless, despite the attention that has been devoted to understanding these problems, theoretically and algorithmically, there are still several important mathematical questions surrounding them that have not been well understood.

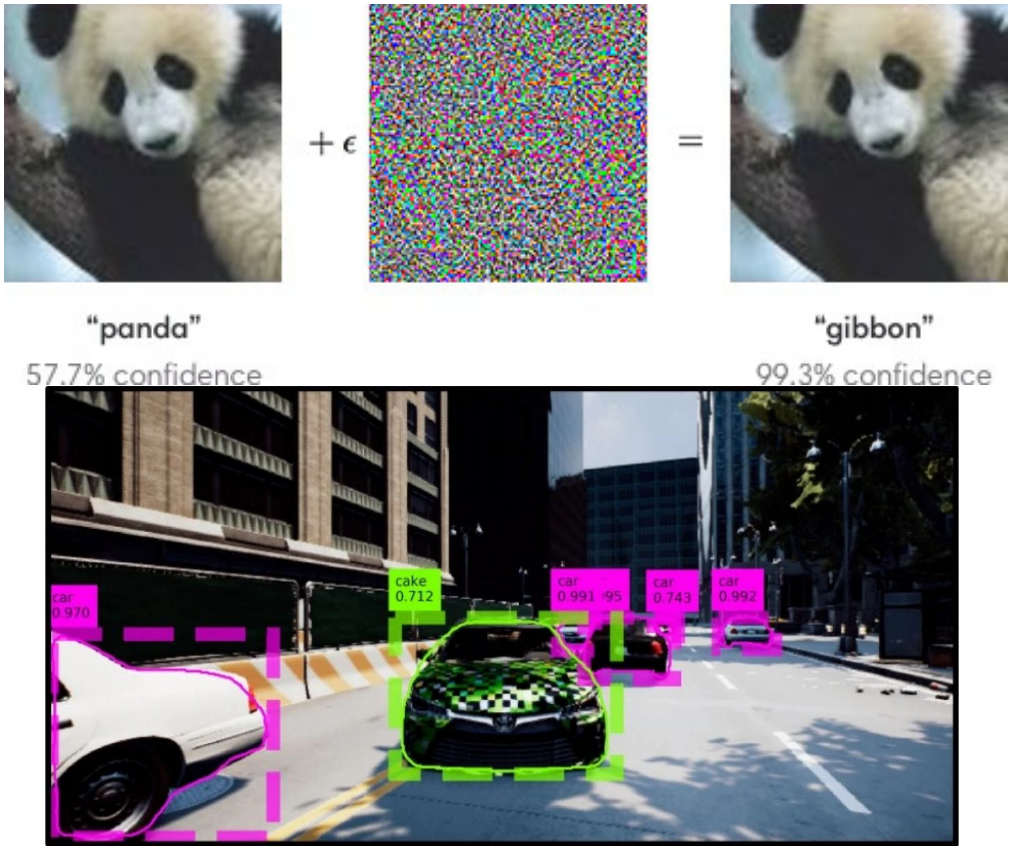

Figure 1.1: Top : Adversarial examples generated for GoogLeNet. (Left) is a correctly predicted sample, (center) difference between correct image and (right) adversarial example from Goodfellow et al. (2015).
Bottom : The car with a camouflage pattern is misdetected as a "cake" from Zhang et al. (2019).

1.1 Generalized barycenter problem and multimarginal optimal transport formulation

The first attempts Bhagoji et al. (2019); Pydi and Jog (2021a,b) to understand the adversarial training problem are based on an somewhat accidental but intriguing observation that the optimal adversarial risk in the binary classification with the agnostic learning setting has the same form of an optimal transport problem which

is also known as "Strassen's theorem", equivalently known as the total-variation distance on the space of probability measures. This is the first breakthrough to reach some meaningful mathematical result on this field in a very general sense. Also, through the equivalence of optimal transport problem they can explicitly compute the adversarial risk. However, their achievements are not fully satisfactory. Although those papers open a new door to explore this field, it is still mysterious that what is really going on the coincidence of formula. In fact, no explanation was provided about this coincidence. Because of this reason, no one truly understands the meaning of the adversarial learning in some sense beyond this miracle coincidence.

In chapter 3, base on Garcıa Trillos et al. (2023), we generalize the adversarial training problem to multiclass classification and prove the equivalence between this problem and several fomulae written as multimarginal optimal transport problems. In addition to these, we can provide the mathematical explanation regarding the adversarial training problem, especially the geometric understanding regarding the adversarial attack. The key object for all works is the *generalized* (*Wasserstein*) *barycenter problem* which is the following:

$$\inf_{\lambda,\widetilde{\mu}_1,\dots,\widetilde{\mu}_K}\left\{\lambda(\mathcal{X})+\sum_{i\in\mathcal{Y}}C(\mu_i,\widetilde{\mu}_i):\ \lambda\geqslant\widetilde{\mu}_i \text{ for all } i\in\mathcal{Y}\right\}.$$

where $\mu_1,\dots,\mu_K$ are given positive measures and

$$C(\mu_i,\widetilde{\mu}_i)=\inf_{\pi_i\in\Pi(\mu_i,\widetilde{\mu}_i)}\int_{\mathcal{X}}c(x,x')d\pi_i(x,x')$$

which is classical optimal transport cost from μ_i to $\widetilde{\mu}_i$. The first main theorem is that

Theorem 1.1. *Under some assumptions on* $c(x,x')$*, we have*

$$\textit{adversarial risk}=1-\inf_{\lambda,\widetilde{\mu}_1,\dots,\widetilde{\mu}_K}\left\{\lambda(\mathcal{X})+\sum_{i\in\mathcal{Y}}C(\mu_i,\widetilde{\mu}_i):\ \lambda\geqslant\widetilde{\mu}_i \textit{ for all } i\in\mathcal{Y}\right\}.$$

This equivalence shows the nice interpretation of adversarial training problem, especially from the adversary's perspective because variables $\widetilde{\mu}_i$'s are taken by the adversary. Geometrically and intuitively, the adversary wants to make more overlaps among $\widetilde{\mu}_i$'s provided that the total transporting cost is not too big.

Once obtaining it, following the philosophy of classical barycenter problems, one would attain an equivalent multimarginal optimal transport formulation of the adversarial training model. Like classical barycenter problems, the generalized barycenter problem has an equivalent multimarginal optimal transport (MOT) formulation. To be precise, we use a *stratified* multimarginal optimal transport problem to obtain an equivalent reformulation.

Theorem 1.2. *Under some assumptions on* $c(x, x')$, *let* $S_K := \{A \subseteq \mathcal{Y} : A \neq \emptyset\}$. *Given* $A \in S_K$, *define* $c_A : \mathcal{X}^K \to [0, \infty]$ *as* $c_A(x_1, \dots, x_K) := \inf_{x' \in \mathcal{X}} \sum_{i \in A} c(x_i, x')$.

Consider the problem:

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

where $\mathcal{P}_i$ *is the projection map* $\mathcal{P}_i : (x_1, \dots, x_K) \mapsto x_i$, *and* $S_K(i) := \{A \in S_K \, : \, i \in A\}$. *Then*

$$\inf \lambda(\mathcal{X}) + \sum_{i \in \mathcal{Y}} C(\mu_i, \widetilde{\mu}_i) = \inf \sum_{A \in S_K} \int_{\mathcal{X}^K} \big(c_A(x_1, \dots, x_K) + 1\big) \, d\pi_A(x_1, \dots, x_K).$$

In particular, if you solve this stratified multimargianl optimal transport problem, then you can obtain the optimal adversarial attiacks $\widetilde{\mu}_i$'s from π_A's. In this sense, this formulation is the problem for the adversary.

Lastly, one would get a nice single MOT formulation. But, it is still not like usual MOT problems because its cost function $\mathbf{c}$, we call it a MOT cost function to distinguish a given cost c, has a very special structure. Heuristically, a MOT cost function is itself an optimization problem. So, we have a nested optimization

problem; first solve a local problem given K-many points and solve a global problem by taking an optimal coupling which organizes the mass from μ_i's in a most efficient way.

Theorem 1.3. *Under some assumptions on* $c(x, x')$*, there is some MOT cost function* $\mathbf{c}$ *such that*

$$\inf \lambda(\mathcal{X}) + \sum_{i \in \mathcal{Y}} C(\mu_i, \widetilde{\mu}_i) = \inf_{\pi \in \Pi(\widehat{\mu}_1, \dots, \widehat{\mu}_K)} \int_{\mathcal{X}_*^K} \mathbf{c}(x_1, \dots, x_K) d\pi(x_1, \dots, x_K).$$

One can regard the above formulae for the adversary's problem. In words, if the adversary solves one of the above three problems, then the optimal adversarial attacks follow. A natural question is about the learner, or classifier. How can we obtain a robust classifier? A very natural guess is to study the dual of the adversary's minimzation problems. Regarding the adversarial training as a two-players minimax game, it is an intuitive consequence that its dual is a problem for the learner. Without the previous observations, however, this guess is not trivial at all. The following results are about the learner's problems.

Theorem 1.4. *Let* $\mathcal{C}_b(\mathcal{X})$ *be the set of bounded real-valued continuous functions over* $\mathcal{X}$*. Under some assumptions on* $c(x, x')$*, The dual of the generalized barycenter problem is*

$$\begin{aligned} &\sup_{f_1, \dots, f_K \in \mathcal{C}_b(\mathcal{X})} \sum_{i \in \mathcal{Y}} \int_{\mathcal{X}} f_i^c(x_i) d\mu_i(x_i) \\ &\text{s.t. } f_i(x) \geqslant 0, \ \sum_{i \in \mathcal{Y}} f_i(x) \leqslant 1, \ \text{for all } x \in \mathcal{X}, \ i \in \{1, \dots, K\}, \end{aligned}$$

where $f_i^c(x) := \inf_{x'} \{f_i(x) + c(x, x')\}$*. Also, there is no duality gap between primal and dual problems.*

The dual of the stratified MOT problem is

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

and there is no duality gap between primal and dual problems.

If f *is a solution for the dual of the generalized barycenter problem, then it is a* (*optimal*) *robust classifier.*

Every detail will be described and formulated in chapter 3. We also need to mention that to solve the first dual problem is not trivial because it lacks any structure. Instead, one can detour through the second dual problem which has more structures thanks to a specific constraints to obtain a robust classifier.

Theorem 1.5. *Suppose that* $(\widetilde{\mu}^*, g^*)$ *is a solution pair for the generalized barycenter problem and the dual of its MOT formulation. Let* f^* *be defined as*

$$f_i^*(\widetilde{x}) := \sup_{x \in \mathrm{spt}(\mu_i)} \{g_i^*(x) - c(x, \widetilde{x})\},$$

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

If f^* *is Borel-measurable, then* $(f^*, \widetilde{\mu}^*)$ *is a saddle solution for the adversarial training problem. In particular,* f^* *is a robust classifier.*

To summarize, using the generalized barycenter problem, the corresponding MOT formulation and its duality, we can completely characterize a pair of solutions, optimal adversarial attack and robust classifier, for the adversarial training model. Furthermore, we can prove the existence of an optimal adversarial attack and explain its geometry of the generalized barycenter problem.

1.2 The existence of robust classifier and unifying perspective of adversarial models

The last theorem describes how to obtain a robust classifier. A problem is that, although we have a very explicit transformation from an optimal dual potential g^* to f^*, it is not trivial that such g^* exists. If a cost function c is bounded and Lipschitz, it is well known that there is an optimal g^* which is also bounded and Lipschitz,

hence, so as f^*. But, most cost functions of interest are not either continuous nor bounded. Thus, it is not clear whether an optimal dual potential g^* exists.

Furthermore, even if such g^* exists, it is even more unclear that f^* obtained by the above transformation is measurable in general. It is well-known that the supremum of an uncountably large family of measurable functions is not measurable in general. Hence, the statement imposes the precondition that "f^* is measurable". In fact, this issue is common, because the transformation, known as c-transform in optimal transport community, does not preserve the measurability.

In this sense, the existence of (Borel measurable) robust classifier is not straightforward, and this is the reason why, through this thesis, we choose *agnostic setting*, in other words, the solution space is the set of all Borel measurable functions whose values are between 0 and 1. For the adversarial attack, thanks to the semi-continuity and uniform boundedness of the values of problem, we are able to use compactness argument to guarantee the minimum and a minimizer easily.

A fundamental difficulty for obtaining a roubst classifier in adversarial training, in contrast to standard training of learning models, is the fact that the adversary has the power to alter the underlying data distribution. In particular, model training becomes an implicit optimization problem over a space of measures, as a result, one may be forced to leave the prototypical setting of equivalence classes of functions defined over a single fixed measure space. In general, measurability issues become more delicate for adversarial training problems at the moment of providing a rigorous mathematical formulation for the problem. Due to these difficulties, there are several subtle variations of the adversarial training model in the literature and it has not been clear whether these models are fully equivalent. More worryingly, for some models, even the *existence* of optimal robust classifiers is unknown, essentially due to convexity and compactness issues.

Let us emphasize that these issues arise even in what can be regarded as the simplest possible setting of the agnostic learner, i.e. where the space of classifiers is taken to be the set of all possible Borel measurable weak (probabilistic) classifiers. While this setting is trivial in the absence of an adversary (there the optimal choice for the learner is always the Bayes classifier), the structure of the problem is much

more subtle in the adversarial setting (in other words the analog of the Bayes classifier is not fully understood). With an adversary, the training process can be viewed as a two-player min-max game (learner versus adversary) Bose et al. (2020); Meunier et al. (2021); Pydi and Jog (2021b); Balcan et al. (2023) and as a result, the optimal strategies for the two players are far from obvious. By relaxing the problem to the agnostic setting, one at least is working over a convex space, but again measurability issues pose a problem for certain formulations of adversarial training.

The existence of measurable "robust" solutions to optimization problems has been a topic of interest not only in the context of adversarial training Pydi and Jog (2021b); Frank and Niles-Weed (2022); Frank (2022); Awasthi et al. (2021a,b) but also in the general *distributionally robust optimization* literature, e.g., Blanchet and Murthy (2019). Previous studies of robust classifiers use *the universal* σ*-algebra* not only to formulate optimization problems rigorously, but also as a feasible search space for robust classifiers. The proofs of these existence results rely on the pointwise topology of a sequence of universally measurable sets, the weak topology on the space of probability measures, and lower semi-continuity properties of $\overline{R}_\varepsilon(\cdot)$. The (universal) measurability of a minimizer is then guaranteed immediately by the definition of the universal σ-algebra. We want to emphasize that all the works Pydi and Jog (2021b); Frank and Niles-Weed (2022); Frank (2022); Awasthi et al. (2021a,b) prove their results in the binary ($K = 2$) classification setting where $\mathcal{X}$ is a subset of Euclidean space.

In light of the above considerations, the purpose of chapter 4, based on García Trillos et al. (2023a), of my thesis is twofold. On one hand, we provide rigorous justification for the existence of *Borel*-measurable robust classifiers in the multiclass classification setting for three different models of adversarial training. Notably, our analysis includes a widely used model for which the existence of Borel classifiers was not previously known and existence of solutions had only been guaranteed when enlarging the original Borel σ-algebra of the data space. On the other hand, we develop a series of connections between the three mathematical models of adversarial training discussed throughout the paper exploiting ideas from optimal

transportation and total variation minimization. By developing these connections, we hope to present a unified formulation of adversarial training and highlight the prospective advantages of using tools in computational optimal transport for solving these problems in the agnostic-classifier setting (and perhaps beyond the agnostic setting too). We also highlight, in concrete terms, the connection between adversarial training and the direct regularization of learning models. To achieve all the aforementioned goals, we expand and take advantage of chapter 3 as well as of the work Bungert et al. (2023) exploring the connection between adversarial training and perimeter minimization in the binary classification setting.

Two main results discussed in chapter 4 are the following:

Theorem 1.6. *Under some assumptions on* $c(x, x')$*, there exists a (Borel) robust classifier* f^**. Furthermore, there exists* $\tilde{\mu}^* \in \mathcal{P}(\mathcal{Z})$ *such that* $(f^*, \tilde{\mu}^*)$ *is a saddle point for the adversarial training problem.*

Theorem 1.7. *For almost all* $\varepsilon \geqslant 0$*, there exists a Borel robust classifiers for three different adversarial training models.*

To summarize, based on optimal transport duality theorem, we can prove the existence of Borel robust classifiers for three variants of adversarial training problem. We want to emphasize that all the proofs does rely on the ingredients developed in chapter 3. Based on that, an elementary but critical observation leads to the unifying perspective for different adversarial training models.

1.3 More tractable numerics

The adversarial training problem is not just a toy example to understand something abstract. Rather, practitioners in reality require its understanding and do something based on it everyday. Therefore, a natural next question is *how to compute it*.

Although there are a lot of demands for computing several objects in the adversarial training problem, for example an optimal adversarial attack, an optimal

robust classifier and the optimal adversarial risk, until very recently there has been no clear way, at least theoretically, for this purpose.

The purpose of chapter 5, based on García Trillos et al. (2023b), is to suggest more tractable algorithms to achieve such purpose: to compute everything we need in this problem. We propose two different algorithms, *exact solving* and *truncated Sinkhorn iteration*. Interestingly, thanks to various formulations of the adversarial training problems developed in chapter 3, each of algorithms is based on different theoretical backgrounds: exact solving relies on the generalized barycenter problem and truncated Sinkhorn iteration does on the stratified MOT problem.

The common idea shared by two approaches is that in real data sets *there might not be many higher-order interactions*. Regarding a data point of each class as a particle, one is able to imagine that the adversary has more chance to fool the learner if particles are close to each other. Then, the terminology *interaction* indicates how these particles to interact with each other, or, how they are close to each other. If we assume that two particles cannot interact with each other when the distance between them is too large, larger than the given adversarial budget $\varepsilon \geqslant 0$, then the adversary is not able to use them together to deceive the learner.

In this sense, one can reach the hypothesis that if each of classes are not so close to each other, then there may not be many interactions so that computing the optimal adversarial risk is not too hard. In particular, although higher-order interactions are critical for the learner, it might be unlikely to confront them due to weak interactions.

In other words, there is a combinatorial problem at the core of the adversarial training problem: Which classes of points should be moved onto a single point and where should that single point be placed? Naively, if there are K classes, each of which has n points this leads to a total of $(n+1)^K - 1$ possible classes of 1 to K points to consider, and this quickly becomes infeasible for even moderately chosen n and K. However, if the previous guess is indeed true, we avoid this overwhelming number and somehow compute what we want.

Here, we state two algorithms: details will be discussed in chapter 5.

Algorithm 1 Exact solving

Input: X : data set, $\mu = (\mu_1, \ldots, \mu_K)$: empirical distribution, ε : adversarial budget.
Construct $C(\varepsilon)$.
Construct the ε-incidence matrix $I_\varepsilon \in \{0,1\}^{X \times C(\varepsilon)}$.
Solve LP.
Output: $\lambda^* = \sum_{C \in C(\varepsilon)} w^*(C)\delta_{F(C)}$, $\widetilde{\mu}_i = \sum_{A \in S_K(i)} \sum_{C \in C_A(\varepsilon)} w^*(C)\delta_{F(C)}$ and value $= \sum_{C \in C(\varepsilon)} w^*(C)$.

Algorithm 2 Truncated entropic regularization(Sinkhorn)

Input: X : data set, $\eta > 0$: entropic parameter, L : trucation level, $\mu = (\mu_1, \ldots, \mu_K)$: empirical distribution, ε : adversarial budget.
Initialization : $\lambda_i = \eta \log \mu_i$ for all $i \in \mathcal{Y}$.
while not converge **do**

$$\lambda_1(\cdot) \leftarrow \eta \log \mu_1(\cdot) - \eta \log \left(\sum_{A \in S_k^L(1)} G(\lambda_{A \setminus \{1\}})(\cdot) \right),$$

$$\vdots$$

$$\lambda_K(\cdot) \leftarrow \eta \log \mu_K(\cdot) - \eta \log \left(\sum_{A \in S_k^L(K)} G(\lambda_{A \setminus \{K\}})(\cdot) \right).$$

end while
Compute $\pi_A^*(x_A) = \exp\left(\frac{1}{\eta}\left(\sum_{i \in A} \lambda_i^*(x_{l_i}^i)\right)\right) \exp\left(-\frac{1}{\eta}\left(1 + c_A(x_A)\right)\right)$ for all $A \in S_K^L$.
Output: $\{\pi_A^*\}_{A \in S_k^L}$ and value $= \sum_{A \in S_k^L} \sum_{x^A} (1 + c_A(x_A))\pi_A^*(x_A)$.

1.4 Organization

The remaining of my thesis is organized in this way.

In chapter 2, we will elaborate some preliminaries to state our problem rigorously.

In chapter 3, we reformulate the adversarial training problem in terms of equivalent formulae which are based on optimal transport theory. They include the generalized (Wasserstein) barycenter problem, the stratified MOT problem, another MOT problem and their dual problems.

In chapter 4, we will prove the existence of Borel measurable robust classifiers for the adversarial training problem and provide the unifying perspective for its variant models.

In chapter 5, we will propose more tractable numerical implementations for the adversarial training problem based on the theoretical results developed in the previous two chapters.

Lastly, in chapter 6, we will finish this long for long, short for short journey with the summary of my thesis and advance some of possible future works in this field.

2 PRELIMINARIES

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

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

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

Typically, a (multiclass) classification rule in the above setting is simply a Borel measurable map $f : \mathcal{X} \to \mathcal{Y}$. In this paper, however, it will be convenient to expand this notion slightly and interpret general classification rules as Borel measurable maps from $\mathcal{X}$ into $\Delta_{\mathcal{Y}} := \left\{(u_i)_{i\in\mathcal{Y}} : 0 \leqslant u_i \leqslant 1, \sum_{i\in\mathcal{Y}} u_i \leqslant 1\right\}$, the set of (up to normalization constants) probability distributions over $\mathcal{Y}$ (see remark 2.1); oftentimes these functions are known as *soft-classifiers*. For future reference, we denote by $\mathcal{F}$ the set

$$\mathcal{F} := \{f : \mathcal{X} \to \Delta_{\mathcal{Y}} : f \text{ is Borel measurable}\}. \tag{2.1}$$

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

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

$$\inf_{f \in \mathcal{F}} R(f, \mu), \tag{2.2}$$

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

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

The loss function $\ell : \Delta_{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$ appearing in the definition of the risk can be chosen in multiple reasonable ways, but here we restrict to the choice

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

which, in lieu of the fact that $\ell(e_j, i)$ is equal to 1 if $i \neq j$ and 0 if $i = j$ (e_j is the extremal point of $\Delta_{\mathcal{Y}}$ with entry one in its j-th coordinate), will be referred to as the 0-1 loss. Note that under the 0-1 loss function the risk $R(f, \mu)$ can be rewritten as

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

Remark 2.1. *Given the structure of the 0-1 loss function considered here, we may replace the set $\mathcal{F}$ with the set of those $f \in \mathcal{F}$ for which $\sum_i f_i = 1$. Indeed, given $f \in \mathcal{F}$ we can always consider $\tilde{f} \in \mathcal{F}$ defined according to $\tilde{f}_{i_0} := f_{i_0} + (1 - \sum_{i \in \mathcal{Y}} f_i)$ and $\tilde{f}_i = f_i$ for $i \neq i_0$ to obtain a value of risk that is no greater than the one of the original f.*

Moreover, one can observe that solutions to the risk minimization problem (2.2) are the standard multiclass Bayes classifiers from statistical learning theory (e.g., see Bousquet et al. (2004); von Luxburg and Schölkopf (2011)). These classifiers are characterized by the condition $f^*_{Bayes,i}(x) = 0$ if $\mathbf{P}_{(X,Y)\sim\mu}(Y = i|X = x) \neq \max_{j \in \mathcal{Y}} \mathbf{P}_{(X,Y)\sim\mu}(Y = j|X = x)$ for all i, and it is always possible to select a Bayes

classifier of the form $f^*_{Bayes}(x) = (\mathbb{1}_{A_1^*}(x), \ldots, \mathbb{1}_{A_K^*}(x))$, where $A_1^*, \ldots, A_K^*$ form a measurable partition of $\mathcal{X}$. In other words, there always exist hard classifiers that solve the risk minimization problem (2.2).

By definition, a solution to (2.2) classifies *clean* data optimally; by clean data here we mean data distributed according to the original distribution μ. However, one should not expect the standard Bayes classifier to perform equally well when inputs have been adversarially contaminated, and the goal in adversarial training is precisely to create classifiers that are less susceptible to data corruption. One possible way to enforce this type of robustness is to replace the objective function in (2.2) with one that incorporates the actions of a well-defined adversary, and then search for the classifier that minimizes the new notion of (adversarial) risk. This adversarial risk can be defined in multiple ways, but two general ways stand out in the literature and will be the emphasis of our discussion; we will refer to these two alternatives as *data-perturbing adversarial model* and *distribution-perturbing adversarial model*. As it turns out, there exist connections between the two (see Pydi and Jog (2021b) for more details) and we will develop further connections shortly.

For the *data-perturbing adversarial model* we will consider the following two versions:

$$R^o_\varepsilon := \inf_{f\in\mathcal{F}} R^o_\varepsilon(f) := \inf_{f\in\mathcal{F}} \left\{ \sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} \sup_{\tilde{x}\in B_\varepsilon(x)} \{1 - f_i(\tilde{x})\} d\mu_i(x) \right\}, \tag{2.3}$$

$$\inf_{f\in\mathcal{F}} \left\{ \sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} \sup_{\tilde{x}\in\overline{B}_\varepsilon(x)} \{1 - f_i(\tilde{x})\} d\mu_i(x) \right\}. \tag{2.4}$$

Here $B_\varepsilon(x)$ ($\overline{B}_\varepsilon(x)$, respectively) denotes an open (closed) ball with radius ε centered at x. In both versions, the adversary can substitute any given input x with a $\tilde{x}$ that belongs to a small ball of radius ε around the original x. In this setting, the learner's goal is to minimize the worst-loss that the adversary may induce by carrying out one of their feasible actions. Although at the heuristic level the difference between the two models is subtle (in the first model the adversary optimizes over open balls and in the second over closed balls), at the mathematical level these two

models can be quite different. For starters, the problem (2.4) is not well-formulated, as it follows from a classical result in Luzin and Sierpiński (1919), which discusses that, in general, the function $x \mapsto \sup_{\tilde{x} \in \overline{B}_\varepsilon(x)} \{1 - f_i(\tilde{x})\}$ may not be Borel-measurable when only the Borel-measurability of the function f_i has been assumed. For this reason, the integral with respect to μ_i in (2.4) (which is a Borel positive measure, i.e., it is only defined over the Borel σ-algebra) may not be defined for all $f \in \mathcal{F}$. In (4.2.2) we provide a rigorous formulation of (2.4) (which at this stage should only be interpreted informally). This reformulation will require the use of an extension of the Borel σ-algebra, known as the universal σ-algebra, as well as an extension of the measures μ_i to this enlarged σ-algebra. Problem (2.3), on the other hand, is already well formulated, as no measurability issues arise when taking the sup over open balls. At a high level, this is a consequence of the fact that arbitrary unions of open balls are open sets and thus Borel-measurable; see, for example, Remark 2.3 in Bungert et al. (2023). Regardless of which of the two models one adopts, and putting aside for a moment the measurability issues mentioned above, it is unclear whether it is possible to find minimizers for any of the problems (2.3) and (2.4) within the family $\mathcal{F}$.

The *distributional-perturbing adversarial model* is defined as a minimax problem that can be described as follows: after the learner has chosen a classifier $f \in \mathcal{F}$, an adversary selects a new data distribution $\widetilde{\mu} \in \mathcal{P}(\mathcal{Z})$, and, by paying some cost $C(\mu, \widetilde{\mu})$, attempts to make the risk $R(f, \widetilde{\mu})$ be as large as possible. Precisely, we consider the problem

$$R^*_{DRO} := \inf_{f \in \mathcal{F}} \sup_{\widetilde{\mu} \in \mathcal{P}(\mathcal{Z})} \{R(f, \widetilde{\mu}) - C(\mu, \widetilde{\mu})\}, \tag{2.5}$$

where $C : \mathcal{P}(\mathcal{Z}) \times \mathcal{P}(\mathcal{Z}) \to [0, \infty]$ has the form:

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

for some Borel measurable cost function $c_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to [0, \infty]$. Here and in the remainder of the paper, we use $\Pi(\cdot, \cdot)$ to represent the set of couplings between

two positive measures over the same space; for example, $\Pi(\mu, \widetilde{\mu})$ denotes the set of positive measures over $\mathcal{Z} \times \mathcal{Z}$ whose first and second marginals are μ and $\widetilde{\mu}$, respectively. Note that problem (2.5) is an instance of a *distributionally robust optimization* (DRO) problem. Problem (2.5) is well-defined given that all its terms are written as integrals of Borel measurable integrands against Borel measures.

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

$$c_{\mathcal{Z}}(z, \tilde{z}) := \begin{cases} c(x, \tilde{x}) & \text{if } y = \tilde{y} \\ \infty & \text{otherwise ,} \end{cases} \tag{2.6}$$

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

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

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

$$C(\mu_i, \widetilde{\mu}_i) = \min_{\pi \in \Pi(\mu_i, \widetilde{\mu}_i)} \int c(x, \tilde{x}) d\pi(x, \tilde{x}). \tag{2.7}$$

Remark 2.2. *Throughout my thesis, we use the convention that* $C(\mu_i, \widetilde{\mu}_i) = \infty$ *whenever the set of couplings* $\Pi(\mu_i, \widetilde{\mu}_i)$ *is empty. This is the case when* μ_i *and* $\widetilde{\mu}_i$ *have different total masses.*

A typical example of a cost c that we will discuss in detail throughout this paper is the cost function:

$$c(x, \tilde{x}) = c_\varepsilon(x, \tilde{x}) := \begin{cases} \infty & \text{if } d(x, \tilde{x}) > \varepsilon \\ 0 & \text{if } d(x, \tilde{x}) \leqslant \varepsilon, \end{cases} \tag{2.8}$$

where in the above ε is a positive parameter that can be interpreted as *adversarial budget*.

Example 2.3. *Notice that in this case, the* c*-transform* f^c *of a given function* f *takes the form:*

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

In this setting, (2.5) *can be written as*

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

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

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

Example 2.4. *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* (2.5) *can be written as*

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

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*** *2.3. 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*.*

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

$$f^c(x) := \inf_{x' \in \mathcal{X}} \{f(x') + c(x', x)\}, \tag{2.9}$$

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

$$g^{\bar{c}}(x') := \sup_{x \in \mathcal{X}} \{g(x) - c(x', x)\}. \tag{2.10}$$

They have an important role in optimal transport problems. More details will be explained in section 4.6.

Through my thesis, we always impose some assumptions on the cost $c : \mathcal{X} \times \mathcal{X} \to [0, \infty]$ stated below.

Assumption 2.5. *We assume that the cost* $c : \mathcal{X} \times \mathcal{X} \to [0, \infty]$ *is a lower semi-continuous and symmetric function satisfying* $c(x, x) = 0$ *for all* $x \in \mathcal{X}$*. We also assume the the following compactness property holds: 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 satisfying* $\sup_{n \in \mathbb{N}} c(x_n, x'_n) < \infty$*, then* $\{(x_n, x'_n)\}_{n \in \mathbb{N}}$ *is precompact in* $\mathcal{X} \times \mathcal{X}$ *(endowed with the product topology).*

Remark 2.6. *Notice that Assumption 2.5 implicitly requires bounded subsets of* $\mathcal{X}$ *to be precompact.*

3 ADVERSARIAL LEARNING, GENERALIZED BARYCENTER PROBLEM AND THEIR CONNECTION BY MULTIMARGINAL OPTIMAL TRANSPORT

This chapter is based on Garcıa Trillos et al. (2023) which is a joint work with Nicolas García Trillos and Matt Jacobs.

The main goal of this chapter is regarding (2.5). For a large family of functionals C in (2.5) we show that the adversarial problem (2.5) is equivalent to a *multimarginal optimal transport problem* (*MOT*) of the form:

$$\inf_{\pi \in \Pi_K(\mu)} \int \mathbf{c}(z_1, \dots, z_K) d\pi(z_1, \dots, z_K), \tag{3.0.1}$$

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 3.1. As part of this equivalence, we explicitly describe how to construct solutions to the original problem (2.5) from solutions to the problem (3.0.1) and its dual, offering in this way new computational strategies for solving problem (2.5). 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 (2.5) by running one such OT algorithm. The equivalence between (2.5) and (3.0.1) 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 et al. (2019); Nakkiran (2019); Pydi and Jog (2021a,b); García Trillos and Murray (2022).

In order to establish the equivalence between (2.5) and (3.0.1), we develop another interesting equivalent reformulation of (2.5) that reveals a rich geometric structure of the original adversarial problem. This reformulation takes the form of a *generalized barycenter problem*

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

which is a novel variant of the Wasserstein barycenter problems introduced in Agueh and Carlier (2011a); Carlier and Ekeland (2010). In the classical Wasserstein barycenter problem, given K *probability* measures $\rho_1, \ldots, \rho_K$ 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 $\widetilde{\mu}_i \leqslant \lambda$ and transport μ_i to $\widetilde{\mu}_i$ in the classical optimal transport sense. This problem will be studied in detail in section 3.2. 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 (2011a); 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 (3.0.1): for the adversary to obtain their optimal strategy, they can actually *localize* their problem to sets of K or fewer data points —see section 3.1. Other theoretical, methodological, and computational implications of these reformulations will be pursued in future work. See section 3.5 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 et al. (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 et al. (2007); Buttazzo et al. (2012); Cotar et al. (2013); Mendl and Lin (2013); Colombo et al. (2015), and in economics Ekeland (2005); Chiappori et al. (2010); Carlier and Ekeland (2010). In the machine learning community, researchers have recently explored many interesting applications, including generative adversarial networks(GANs) Choi et al. (2018); Cao et al. (2019) and Wasserstein Barycenters Agueh and Carlier (2011a); Cuturi and Doucet (2014); Benamou et al. (2015); Carlier et al. (2015); Srivastava et al. (2018); Delon and Desolneux (2020), where MOTs are used. Recent works like Di Marino and Gerolin (2020); Haasler et al. (2021) develop a connection between the Schrödinger bridge problem and MOT. MOT problems have been extended to the unbalanced setting —see Beier et al. (2021).

Outline of this chapter

In section 3.1, we introduce the generalized Wasserstein barycenter problem, which can be interpreted as dual of (2.5), and define in detail the MOT problem (3.0.1). In section 3.2, 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 3.3 we discuss the equivalence between (2.5) and (3.0.1) through the duality results in earlier sections. In section 3.4, 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 chapter in section 3.5 by presenting some summary.

3.1 The generalized barycenter problem and the MOT problem

Recall the solution space $\mathcal{F}$ and the space of probability measures $\mathcal{P}(\mathcal{Z})$. For a given pair $(f, \widetilde{\mu}) \in \mathcal{F} \times \mathcal{P}(\mathcal{Z})$, we define the risk:

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

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

$$B(f, \widetilde{\mu}) := \sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} f_i(\widetilde{x})\, d\widetilde{\mu}_i(\widetilde{x}). \tag{3.1.1}$$

With these new definitions, problem (2.5) is immediately seen to be equivalent to

$$\sup_{f\in\mathcal{F}} \inf_{\widetilde{\mu}\in\mathcal{P}(\mathcal{Z})} \{B(f, \widetilde{\mu}) + C(\mu, \widetilde{\mu})\}. \tag{3.1.2}$$

Moreover, if we denote by $\widetilde{B}^*_\mu$ the optimal value of (3.1.2), and by R^*_μ the optimal value of (2.5), we have the identity:

$$R^*_{DRO} = 1 - \widetilde{B}^*_\mu.$$

We write 1 explicitly, although for the most part 1 can be thought of as being equal to one.

The dual of (3.1.2) is obtained by swapping the sup and the inf:

$$\inf_{\widetilde{\mu}\in\mathcal{P}(\mathcal{Z})} \sup_{f\in\mathcal{F}} \{B(f, \widetilde{\mu}) + C(\mu, \widetilde{\mu})\}. \tag{3.1.3}$$

Notice that the value of (3.1.3) is always greater than or equal to the value of

(3.1.2). 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 2.5, there is no duality gap in this problem. In what follows we focus on the dual problem (3.1.3) and only return to problem (3.1.2), which is equivalent to the original adversarial problem (2.5), in section 3.3. Notice, however, that the statement of Theorem 3.3 mentions the adversarial problem explicitly.

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

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

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

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

the constraint in λ can be simply written as $\lambda\geqslant\widetilde{\mu}_i$ for all $i\in\mathcal{Y}$. Combining the above with the structure of the cost $C(\mu,\widetilde{\mu})$, we conclude that problem (3.1.3) is equivalent to the generalized barycenter problem mentioned in the introduction:

$$B^*_\mu := \inf_{\lambda,\widetilde{\mu}_1,\dots,\widetilde{\mu}_K}\left\{\lambda(\mathcal{X})+\sum_{i\in\mathcal{Y}}C(\mu_i,\widetilde{\mu}_i):\ \lambda\geqslant\widetilde{\mu}_i \text{ for all } i\in\mathcal{Y}\right\}, \tag{3.1.4}$$

where we use the notation B^*_μ for future reference; see Figure 3.1 for a pictorial explanation.

Remark 3.1. *It is straightforward to see from* (3.1.3) *that* B^*_μ *is* 1*-homogeneous in* μ*. That is, if* $a>0$*, then* $B^*_{a\mu}=aB^*_\mu$.

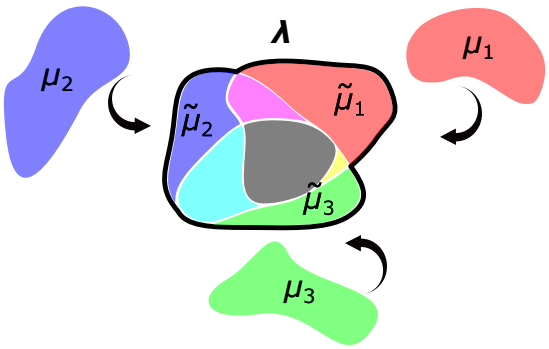

Figure 3.1: Picture for (3.1.4). μ_i's are first moved to $\widetilde{\mu}_i$'s and λ is chosen to cover all $\widetilde{\mu}_i$'s: it is the smallest positive measure which is larger than all $\widetilde{\mu}_i$'s.

The MOT problem

General MOT problems

Before providing the details of our MOT problem (3.0.1), 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 $1 \leqslant k \leqslant K$, 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:

$$\sup_{\phi\in\Phi}\left\{\sum_{j=1}^{K}\int_{\mathcal{S}_j}\phi_j(\xi_j)d\nu_j(\xi_j)\right\}, \tag{3.1.5}$$

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

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

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 et al. (2010); Agueh and Carlier (2011a). 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

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

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

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

Indeed, let π^* be a minimizer of the above MOT problem. Defining $\nu^*=T_{\#}\pi$, where $T(x_1,\dots,x_K):=\operatorname{argmin}_{x'}\sum_{i\in\mathcal{Y}}c(x',x_i)$, 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 (2011a).

From adversarial robustness to MOT

Now we are ready to state problem (3.0.1) 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 👻. The marginals of the couplings in the desired MOT problem will be probability measures over the set $\mathcal{Z}_* := \mathcal{Z} \cup \{👻\}$. More precisely, letting P_i represent the projection onto the i-th coordinate, we consider the set:

$$\Pi_K(\mu) := \left\{ \pi \in \mathcal{P}(\mathcal{Z}_*^K) \ : \ P_{i\sharp}\pi = \frac{1}{2}\mu(\cdot \cap \mathcal{Z}) + \frac{1}{2}\delta_{👻} \text{ for all } i \in \mathcal{Y} \right\}. \tag{3.1.6}$$

Notice that in this set all K marginals are the same. Dividing by the factor $\frac{1}{2}$, 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

$$\mathbf{c}(z_1, \dots, z_K) := B^*_{\widehat{\mu}_{\vec{z}}}, \tag{3.1.7}$$

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

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

Recall that $B^*_{\widehat{\mu}_{\vec{z}}}$ is equal to (3.1.4) (alternatively, equal to (3.1.3)) when μ is equal to $\widehat{\mu}_{\vec{z}}$. In this sense, $\mathbf{c}(z_1, \dots, z_K)$ of (3.1.7) is the value of the generalized barycenter problem given $\widehat{\mu}_{\vec{z}}$ as the data distribution, or *local* generalized barycenter problem.

Remark 3.2. *Notice that $\widehat{\mu}_{\vec{z}}$ is a probability measure if and only if no element in the tuple $\vec{z}$ is 👻.*

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

$$\sup_{\phi\in\Phi}\left\{\sum_{j=1}^{K}\int_{\mathcal{X}\times\mathcal{Y}}\phi_j(z_j)\frac{1}{2}d\mu(z_j)+\frac{1}{2}\sum_{j=1}^{K}\phi_j(\text{🙁})\right\}, \tag{3.1.8}$$

where

$$\Phi:=\left\{\phi=(\phi_1,\dots,\phi_K)\in\prod_{j=1}^{K}L^1\big(\frac{1}{2}\mu+\frac{1}{2}\delta_{\text{🙁}}\big):\sum_{j=1}^{K}\phi_j(z_j)\leqslant B^*_{\widehat{\mu}_{\vec{z}}},\quad\forall\vec{z}\in\mathcal{Z}_*^K\right\}. \tag{3.1.9}$$

We will later show that under Assumption 2.5 there is no duality gap between the MOT problem and its dual (3.1.8) —see Corollary 3.28.

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

Theorem 3.3. *Suppose that Assumption 2.5 holds. Let μ be a finite positive measure over $\mathcal{Z}$. Then* (3.1.3) *is equivalent to the MOT problem* (3.0.1) *with set of couplings $\Pi_K(\mu)$ defined as in* (3.1.6)*, and cost function* $\mathbf{c}$ *defined as in* (3.1.7)*. Specifically,*

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

Furthermore, (3.1.2) $=$ (3.1.3)*. 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* (3.1.3) *and its dual, i.e. problem* (3.1.2)*. The pair $(f^*,\widetilde{\mu}^*)$ is also a saddle point for the original adversarial problem* (2.5)*.*

One immediate consequence of Theorem 3.3 is that with the identity

$$R^*_{DRO}=1-\widetilde{B}^*_\mu,$$

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 et al. (2015, 2019); Lin et al. (2022); Tupitsa et al. (2020); Haasler et al. (2021); Altschuler and Boix-Adsera (2021); Carlier (2022). Our numerical results for a subsample of MNIST and CIFAR 10, shown in Figure 3.6, are obtained using the algorithm discussed in Lin et al. (2022), also known as MOT Sinkhorn algorithm; see subsection 3.4 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 3.3 is presented throughout section 3.3; the expression for $(f^*, \widetilde{\mu}^*)$ in terms of (ϕ^*, π^*) is presented in Corollary 3.30. Given the definition of the cost function $\mathbf{c}$, Theorem 3.3 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, 👻, 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 (3.0.1) is a problem solved by the adversary: ideally, the adversary wants to group together points $(z_1, \dots, z_K)$ for which there is a low classification power $B^*_{\widehat{\mu}_{\vec{z}}}$ (or alternatively large robust risk). On the other hand, the dual of (3.0.1) can be interpreted as a maximization problem solved by the learner. We formalize this novel connection in subsection 3.3: see Corollary 3.30.

In order to prove Theorem 3.3, we will first obtain a series of equivalent reformulations of problem (3.1.4) 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.

3.2 The generalized barycenter problem

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

Proposition 3.4. *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* $c(x, x') \leqslant c(x, x'')$*. Given finite positive measures* $\mu_1, \dots, \mu_K$ *and* c *as above, there exists at least one solution to problem* (3.1.4)*.*

Remark 3.5. *If* c *is a cost that satisfies Assumption 2.5, then* c *satisfies the hypothesis of Proposition 3.4.*

Remark 3.6. *Nearly identical arguments can be used to prove that the various reformulations of* (3.1.4) *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 (3.1.4), we can rewrite the problem in the following form

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

Note that a feasible solution to this problem exists since we may choose $\lambda, \pi_1, \dots, \pi_K$ such that $\lambda := \sum_{i \in \mathcal{Y}} \mu_i$ 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 $\sum_{i \in \mathcal{Y}} \mu_i(\mathcal{X})$.

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

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

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

$$\sum_{i \in \mathcal{Y}} \int_{\mathcal{X} \times E} d\pi_i^n(x, x') \geqslant \lambda^n(E),$$

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 $\sum_{i \in \mathcal{Y}} \mu_i(\mathcal{X} \setminus E_\epsilon) \leqslant \epsilon$. 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 $c(x, x') \leqslant c(x, x'')$. 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 $\lambda^n(\mathcal{X} \setminus F_\epsilon) \leqslant \epsilon$. Therefore, the λ^n are a tight family.

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

A first MOT reformulation of (3.2.1) and geometric consequences

In the rest of what follows, we shall let S_K denote the power set of $\mathcal{Y}$ except for the empty set and for every $i \in \mathcal{Y}$ we let $S_K(i) = \{A \in S_K : i \in A\}$. We can reduce (3.1.4) 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 3.7. *Let* $u_1, \ldots, u_K \in [0, 1]$ *be such that* $\max_{i=1,\ldots,K} u_i = 1$. *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,

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

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 3.8. *Problem* (3.1.4) *is equivalent to*

$$\begin{aligned} &\inf_{\{\lambda_A, \mu_{i,A} : i \in \mathcal{Y}, A \in S_K\}} \sum_{A \in S_K} \left\{ \lambda_A(\mathcal{X}) + \sum_{i \in A} C(\lambda_A, \mu_{i,A}) \right\} \\ &\text{s.t.} \sum_{A \in S_K(i)} \mu_{i,A} = \mu_i \text{ for all } i \in \mathcal{Y}. \end{aligned} \tag{3.2.1}$$

Proof. We split the proof into two parts.

Step 1: Suppose that $\lambda, \tilde{\mu}_1, \dots, \tilde{\mu}_K$ is feasible for problem (3.1.4). In particular, $\lambda \geqslant \tilde{\mu}_i$ 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 $\frac{d\tilde{\mu}_i}{d\lambda} \leqslant 1$ 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 (3.1.4) while maintaining the constraints.

For each $x \in \mathrm{spt}(\lambda)$ we apply Lemma 3.7 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 $i = 1, \ldots, k$.

Notice that the functions $r_A(\cdot)$ can be constructed in a measurable way as it follows from the proof of Lemma 3.7. 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 3.2 (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 $\pi_i \in \Pi(\mu_i, \tilde{\mu}_i)$ be a coupling realizing the cost $C(\mu_i, \tilde{\mu}_i)$, i.e., a minimizer of (2.7), 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 (3.2.1). Moreover,

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

Step 2: Conversely, suppose that $\{\lambda_A\}_A, \{\mu_{i,A}\}_A$ is feasible for (3.2.1). 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 $\lambda \geqslant \tilde{\mu}_i$ 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 3.2 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

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

□

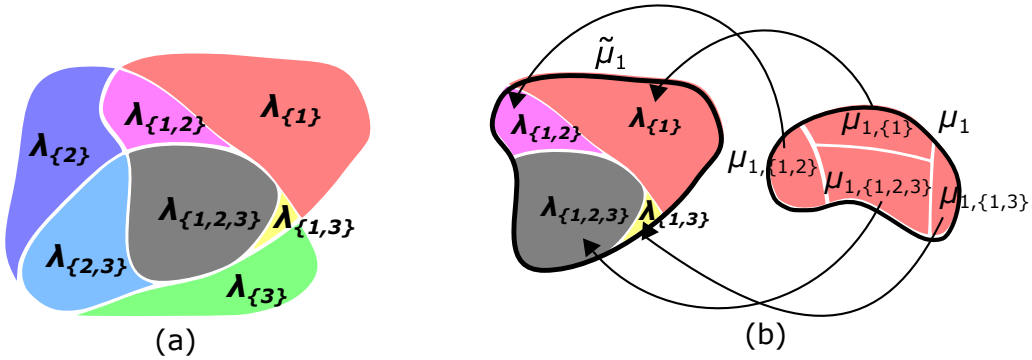

Figure 3.2: (a) : Illustration of a partition of λ. (b) : Illustration of the transport from $\mu_{1,A}$'s to λ_A's.

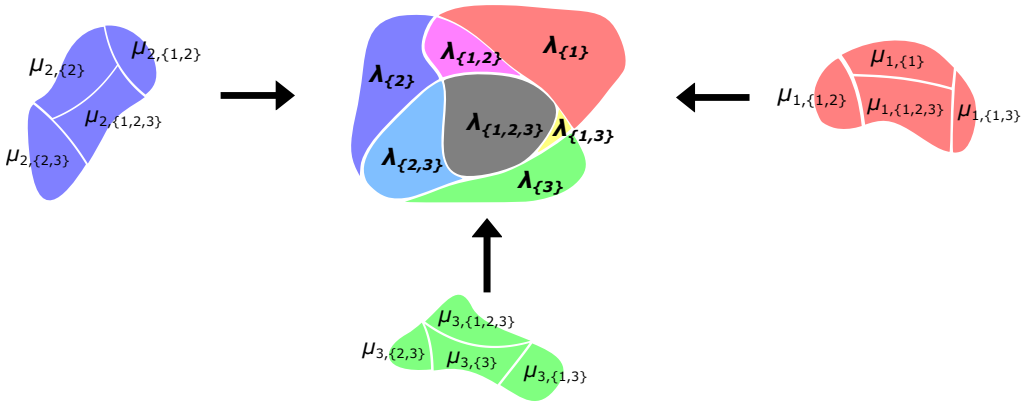

Figure 3.3: Picture for (3.2.1). Each of $\mu_{i,A}$'s is transported to λ_A for all $i \in A$.

Remark 3.9. *Figure 3.3 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 $i \in \mathcal{Y}$ and $A \subseteq S_K$, however, note that if $i \notin A$, then $\mu_{i,A}$ plays no role in the optimization* (3.2.1).

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

$$\operatorname*{argmin}_{\lambda_A} \sum_{i \in A} C(\lambda_A, \mu_{i,A}). \tag{3.2.2}$$

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 (3.2.2) can be reformulated as a multimarginal optimal transport problem Agueh and Carlier (2011a); see also our subsection 3.1. To that end, given $A \subseteq \mathcal{Y}$, define $c_A : \mathcal{X}^K \to \mathbb{R}$

$$c_A(x_1, \dots, x_K) := \inf_{x' \in \mathcal{X}} \sum_{i \in A} c(x', x_i), \tag{3.2.3}$$

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

$$T_A(x_1, \dots, x_K) := \operatorname*{argmin}_{x' \in \mathcal{X}} \sum_{i \in A} c(x', x_i). \tag{3.2.4}$$

Remark 3.10. *If* $\operatorname{argmin}_{x' \in \mathcal{X}} \sum_{i \in A} c(x', x_i)$ *is not unique, we can consider using an additional selection procedure. For example, when* $\mathcal{X} = \mathbb{R}^d$ *we can still recover a unique mapping by choosing* T_A *to be the element of* $\operatorname{argmin}_{x' \in \mathcal{X}} \sum_{i \in A} c(x', x_i)$ *that is closest (in the Euclidean distance) to the Euclidean barycenter* $\frac{1}{|A|} \sum_{i \in A} x_i$.

With the definition of c_A, we can rewrite (3.2.2) as the multimarginal optimal transport problem

$$\inf_{\pi_A} \int_{\mathcal{X}^K} c_A(x_1, \dots, x_K) d\pi_A(x_1, \dots, x_K) \quad \text{s.t. } \mathcal{P}_{i\#}\pi_A = \mu_{i,A} \text{ for all } i \in A, \tag{3.2.5}$$

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 (3.2.5) we can now eliminate λ_A and all of the $\mu_{i,A}$'s from problem (3.2.1) and reformulate the optimization as the multimarginal problem

$$\begin{aligned} &\inf_{\{\pi_A : A \in S_K\}} \sum_{A \in S_K} \int_{\mathcal{X}^K} \big(c_A(x_1, \dots, x_K) + 1\big) d\pi_A(x_1, \dots, x_K) \\ &\text{s.t.} \sum_{A \in S_K(i)} \mathcal{P}_{i\#}\pi_A = \mu_i \text{ for all } i \in \mathcal{Y}. \end{aligned} \tag{3.2.6}$$

The next two propositions formally prove the equivalence between (3.2.1) and (3.2.6). They will also allow us to establish some important geometric properties of optimal generalized barycenters.

Proposition 3.11. *Let* c *be a cost satisfying Assumption 2.5. Given measures* $\mu_1, \dots, \mu_K$, *let* $\{\pi_A\}_{A \in S_K}$ *be a feasible solution to* (3.2.6). *For each* $(x_1, \dots, x_K) \in \mathcal{X}^K$ *and* $A \in S_K$, *let* $f_A(x_1, \dots, x_K)$ *be a choice of element in* $T_A(x_1, \dots x_K)$, *where we recall the definition of* $T_A(x_1, \dots, x_K)$ *from* (3.2.4).

If for each $A \in S_K$ *and* $i \in A$ *we set* $\tilde{\lambda}_A = f_{A\#}\pi_A$ *and* $\tilde{\mu}_{i,A} = \mathcal{P}_{i\#}\pi_A$*, then* $\{\tilde{\lambda}_A, \tilde{\mu}_{i,A} : A \in S_K, i \in A\}$ *is a feasible solution to* (3.2.1) *and*

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

Proof. Since $\sum_{A\in S_K(i)} \mathcal{P}_{i\#}\pi_A = \mu_i$, it is automatic that $\sum_{A\in S_K(i)} \tilde{\mu}_{i,A} = \mu_i$. Since pushforwards do not affect the total mass of a measure, so we also have $\tilde{\mu}_{i,A}(\mathcal{X}) = \tilde{\lambda}_A(\mathcal{X})$ for all $A \in S_K$ and $i \in A$. Hence, $\{\tilde{\lambda}_A, \tilde{\mu}_{i,A}\}_{A\in S_K, i\in A}$ is a feasible solution to (3.2.1).

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

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

We can then compute

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

Thus,

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

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

Hence,

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

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

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

which is the desired result we want. □

In the next proposition we will show that any feasible solution of problem (3.2.1) induces a feasible solution of (3.2.6) with lesser or equal value. This will prove the equivalence between problems (3.2.1) and (3.2.6) and will provide a powerful geometric characterization of optimal generalized barycenters.

Proposition 3.12. *Let* c *be a cost satisfying Assumption 2.5. Given measures* $\mu_1,\dots,\mu_K$*, let* $\mu_{i,A},\lambda_A$ *be feasible solutions to problem* (3.2.1)*. Let* $\gamma_{i,A}\in\mathcal{M}(\mathcal{X}\times\mathcal{X})$ *be an optimal plan for the transport of* $\mu_{i,A}$ *to* 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* (3.2.6) *and*

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

Therefore, (3.2.1) = (3.2.6).

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

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

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 (3.2.1) and (3.2.6), Proposition 3.11 and Proposition 3.12 have the following very important geometric consequences.

Corollary 3.13. *Let c be a cost satisfying Assumption 2.5. 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 (3.2.1). Recalling (3.2.4), 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 $T = \cup_{A \subseteq y} T_A$, then $\lambda_A(\mathcal{X}) = \lambda_A(T_A)$, $\lambda(\mathcal{X}) = \lambda(T)$ and the optimal measures $\widetilde{\mu}_i$ in (3.1.4) 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 3.14. *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 3.12, 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

$$\sum_{A \in S_K} \int_{\mathcal{X}^K} \big(c_A(x_1, \dots, x_K) + 1\big) \, d\tilde{\pi}_A(x_1, \dots, x_K) \leqslant \sum_{A \in S_K} \lambda_A(\mathcal{X}) + \sum_{i \in A} C(\lambda_A, \mu_{i,A}). \tag{3.2.7}$$

From Proposition 3.11, 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

$$\sum_{A \in S_K} \tilde{\lambda}_A(\mathcal{X}) + \sum_{i \in A} C(\tilde{\lambda}_A, \tilde{\mu}_{i,A}) \leqslant \sum_{A \in S_K} \int_{\mathcal{X}^K} \big(c_A(x_1, \dots, x_K) + 1\big) \, d\tilde{\pi}_A(x_1, \dots, x_K). \tag{3.2.8}$$

Examining the proof of Proposition 3.12, it follows that the inequality in (3.2.7) is strict if $\lambda_A(\mathcal{X}) > \lambda_A(T_A)$. In that case, combining (3.2.7) and (3.2.8) 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 3.11. □

When $\mu_1, \dots, \mu_K$ are supported on a finite set of points, Corollary 3.13 has the following consequence.

Corollary 3.15. *If $\mu_1, \dots, \mu_K$ are measures that are supported on a finite set of points and c is a cost satisfying Assumption 2.5, then there exists a solution λ to the optimal generalized barycenter problem (3.1.4) 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 $\sum_{A \in S_K} \prod_{i \in A} n_i \leqslant 2^K \prod_{i=1}^K n_i$ points.

Remark 3.16. *Notice that the bound mentioned at the end of Corollary 3.15 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 3.4.*

Proof. For each $i \in \mathcal{Y}$ 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 (3.2.4). We can now construct the set

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

which is necessarily finite. Indeed, if we set $n_i = |X_i|$, then F has at most $\sum_{A \in S_K} \prod_{i \in A} n_i$ elements. By Corollary 3.13, there exists an optimal barycenter supported on F only. □

A second MOT reformulation of (3.2.1)

Note that in problem (3.2.6) 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 (3.2.6) 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 $\mu_1, \ldots, \mu_K$, then we cannot recover the full generality of (3.2.6).

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

$$\begin{aligned} &\widetilde{c}((x_1, r_1), \ldots, (x_K, r_K)) \\ &:= \inf_{m : S_K \to \mathbb{R}} \sum_{A \in S_K} m_A \big(c_A(x_1, \ldots, x_K) + 1\big) \quad \text{s.t.} \sum_{A \in S_K(i)} m_A = r_i. \end{aligned} \tag{3.2.9}$$

For each $i \in \mathcal{Y}$, let $\tilde{\mathcal{P}}_i$ be the projection $((x_1, r_1), \ldots, (x_K, r_K)) \mapsto x_i$. In what follows, we use $(\vec{x}, \vec{r})$ to denote the tuple $((x_1, r_1), \ldots, (x_K, r_K))$. We then claim that problem (3.2.6) is equivalent to

$$\inf_{\gamma} \int_{(\mathcal{X} \times [0,1])^K} \widetilde{c}(\vec{x}, \vec{r}) \, d\gamma(\vec{x}, \vec{r}) \quad \text{s.t. } \tilde{\mathcal{P}}_{i\#}(r_i \gamma) = \mu_i \text{ for all } i \in \mathcal{Y}. \tag{3.2.10}$$

Proposition 3.17. *Problems* (3.2.6) *and* (3.2.10) *are equivalent, and thus* (3.2.10) *is also equivalent to* (3.1.3), (3.1.4) *and* (3.2.1).

Proof. Given a feasible solution $\pi_{\{1\}}, \ldots, \pi_{\mathcal{Y}}$ to problem (3.2.6), define γ such that for every continuous and bounded function $f : (\mathcal{X} \times [0,1])^K \to \mathbb{R}$ we have

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

where $\chi_A(i) = 1$ if $i \in A$ and zero otherwise. We can then check that γ is feasible for (3.2.10), since for any function $g : \mathcal{X} \to \mathbb{R}$

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

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

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

Therefore,

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

Conversely, suppose that γ is a feasible solution to (3.2.10). Given a tuple $(\vec{x}, \vec{r})$, let

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

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 (3.2.6). Evaluating the objective in

(3.2.6) we see that

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

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. □

Localization

In this section we show that the cost function $\widetilde{c}$ in problem (3.2.10) is equal to $B^*_{\widehat{\mu}}$ for a measure $\widehat{\mu}$ that depends on the arguments of $\widetilde{c}$. This result can be interpreted as a localization property for problem (3.1.4) (and hence for problem (3.1.3) as well). Compare with the discussion after Theorem 3.3.

Lemma 3.18. *Let* $\tilde{x}_1,\dots,\tilde{x}_k\in\mathcal{X}$, *and let* $0\leqslant\tilde{r}_1,\dots,\tilde{r}_k\leqslant 1$. *Then* $\widetilde{c}((\tilde{x}_1,\tilde{r}_1),\dots,(\tilde{x}_K,\tilde{r}_K))$ *defined in* (3.2.9) *is equal to* $B^*_{\widehat{\mu}}$, *where*

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

Proof. To prove this claim we first notice that by Proposition 3.17 $B^*_{\widehat{\mu}}$ 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 (3.2.10). For a feasible γ, notice that γ must concentrate on the set $\{(\vec{x},\vec{r}):x_i=\tilde{x}_i,i\in\mathcal{Y}\}$. Applying the disintegration

theorem to γ, we can rewrite the objective function evaluated at γ as

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

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

$$\int_{[0,1]} r_i d\gamma_r(r_1, \ldots, r_K) = \tilde{r}_i, \quad \forall i = 1, \ldots, K. \tag{3.2.11}$$

It is clear that the map associating a feasible γ to a γ_r satisfying (3.2.11) is onto, and thus, we can rewrite $B^*_{\widehat{\mu}}$ as

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

In the above, the set $G(r_1, \ldots, r_K)$ is the set of $\{m_A\}_{A \in S_K}$ satisfying the constraints in (3.2.9) for the specific tuple $\big((\tilde{x}_1, r_1), \ldots, (\tilde{x}_K, r_K)\big)$, while G is the set of $\{m_A\}_A$ where each m_A is a functions with inputs $r_1, \ldots, r_K$ satisfying $\{m_A(r_1, \ldots, r_K)\}_A \in G(r_1, \ldots, r_K)$.

We can now write the term

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

where we define

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

Notice that

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

Conversely, notice that given a collection of functions $\tilde{m}_A$ satisfying the constraint in (3.2.9) for the tuple $(\tilde{x}_1,\tilde{r}_1),\dots,(\tilde{x}_K,\tilde{r}_K)$, it is straightforward to find γ_r such that $\tilde{m}_A = m_{A,\gamma_r}$ for all A. It now follows that

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

as we wanted to prove. □

Dual Problems

In this section we discuss the dual problems of the different formulations of the generalized barycenter problem studied in section 3.2.1.

Proposition 3.19. *The dual problems to* (3.1.4), (3.2.6), *and* (3.2.10) *can be written as*

$$\begin{aligned}&\sup_{f_1,\dots,f_K\in C_b(\mathcal{X})} \sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} f_i^c(x_i)\,d\mu_i(x_i)\\ &\text{s.t. } f_i(x) \geqslant 0,\ \sum_{i\in\mathcal{Y}} f_i(x) \leqslant 1,\ \text{for all } x\in\mathcal{X}, i\in\mathcal{Y},\end{aligned} \tag{3.2.12}$$

$$\begin{aligned} &\sup_{g_1,\ldots,g_K \in C_b(\mathcal{X})} \sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} g_i(x_i) d\mu_i(x_i) \\ &\text{s.t. } \sum_{i\in A} g_i(x_i) \leqslant 1 + c_A(x_1,\ldots,x_K) \text{ for all } (x_1,\ldots,x_K) \in \mathcal{X}^K, A \in S_K, \end{aligned} \tag{3.2.13}$$

and

$$\begin{aligned} &\sup_{h_1,\ldots,h_K \in C_b(\mathcal{X})} \sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} h_i(x_i) d\mu_i(x_i) \\ &\text{s.t. } \sum_{i\in\mathcal{Y}} r_i h_i(x_i) \leqslant \widetilde{c}(\vec{x},\vec{r}) \text{ for all } (\vec{x},\vec{r}) \in (\mathcal{X}\times[0,1])^K, \end{aligned} \tag{3.2.14}$$

respectively.

Let $f_1,\ldots,f_K; g_1,\ldots,g_K; h_1,\ldots,h_K$ *be feasible solutions to problems* (3.2.12), (3.2.13), *and* (3.2.14) *respectively. Problems* (3.2.13) *and* (3.2.14) *have the same feasible set and hence are identical. Furthermore,* $g_i' := f_i^c$ *is a feasible solution to* (3.2.13) *and* $f_i' = \max\{g_i, 0\}^{\bar{c}}$ *is a feasible solution to* (3.2.12), *hence the optimization of* (3.2.13) *can be restricted to nonnegative* g_i *that satisfy* $g_i = g_i^{\bar{c}c}$. *In particular,* (3.2.12), (3.2.13), *and* (3.2.14) *all have the same optimal value.*

Proof. The derivation of the dual problems is standard.

To see the equivalence between problems (3.2.13) and (3.2.14), fix some $h_1,\ldots,h_K$ that are feasible for (3.2.14) and choose some $B \in S_K$ and $(x_1,\ldots,x_K) \in \mathcal{X}^K$ such that $c_B(x_1,\ldots,x_K) < \infty$. Choose

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

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

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

Thus, we see that the h_i are feasible for (3.2.13) since B and $(x_1, \ldots, x_K)$ were arbitrary.

Conversely, fix some $g_1, \ldots, g_K$ that are feasible for (3.2.13) and some $(\vec{x}, \vec{r}) \in (\mathcal{X} \times [0,1])^K$. Choose

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

and observe that

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

where we used the feasibility of the g_i. Thus, the g_i are feasible for (3.2.14). Since both problems are optimizing the same functional over the same constraint set, we see that (3.2.13) and (3.2.14) are identical.

Now suppose that $f_1, \ldots, f_K$ and $g_1, \ldots, g_K$ are feasible solutions to problems (3.2.12) and (3.2.13) respectively and define $g'_i = f_i^c$ and $f'_i = \max\{g_i, 0\}^{\bar{c}}$. Given

$A \in S_K$, $x_1, \ldots, x_K \in \mathcal{X}^K$, and $r > 0$ we can choose x_r such that

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

Then we see that

$$\sum_{i \in A} g_i'(x_i) \leqslant \sum_{i \in A} f(x_r) + c(x_r, x_i) \leqslant r + 1 + c_A(x_1, \ldots, x_K).$$

Letting $r \to 0$, we see that the g_i' are feasible for (3.2.13). Hence, the optimal value of (3.2.13) cannot lie strictly below the optimal value of (3.2.12).

It remains to verify the feasibility of the f_i'. We begin by showing that if $g_1, \ldots, g_K$ are feasible for (3.2.13) then $\max\{g_1, 0\}, \ldots, \max\{g_K, 0\}$ are also feasible. Fix $A \in S_K$ and $(x_1, \ldots, x_K) \in \mathcal{X}^K$. Let $A' = \{i \in A : g_i(x_i) > 0\}$. We then see that

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

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 $f_i'(x) \geqslant 0$ since $c(x, x) = 0$ for all $x \in \mathcal{X}$. Given $x \in \mathcal{X}$, fix $r > 0$ and for each $i \in \mathcal{Y}$, choose $x_{i,r} \in X$ such that

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

We then have

$$\begin{aligned}\sum_{i \in \mathcal{Y}} \max\{g_i, 0\}^{\bar{c}}(x) &\leqslant \sum_{i \in \mathcal{Y}} \max\{g_i(x_{i,r}), 0\} - c(x_{i,r}, x) + r \\ &\leqslant 1 + r + c_{\mathcal{Y}}(x_{1,r}, \ldots, x_{k,r}) - \sum_{i \in \mathcal{Y}} c(x_{i,r}, x),\end{aligned}$$

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

Notice that the above arguments prove that whenever $g_1, \ldots, g_K$ are feasible

for (3.2.13), then $\max\{g_1, 0\}^{\bar{c}c}, \ldots, \max\{g_K, 0\}^{\bar{c}c}$ are also feasible for (3.2.13). Since $u \leqslant u^{\bar{c}c}$ for any function $u : \mathcal{X} \to \mathbb{R}$, it follows that

$$\sum_{i \in \mathcal{Y}} \int_{\mathcal{X}} g_i(x) d\mu_i(x) \leqslant \sum_{i \in \mathcal{Y}} \int_{\mathcal{X}} \max\{g_i, 0\}^{\bar{c}c}(x) d\mu_i(x).$$

Since we showed that $\max\{g_i, 0\}^{\bar{c}}$ was feasible for (3.2.12), it follows that (3.2.13) cannot attain a larger value than (3.2.12). Hence, we have shown that (3.2.13) and (3.2.12) have the same optimal value. □

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

Lemma 3.20. *Suppose that* c *is a bounded Lipschitz cost that satisfies the hypotheses of Proposition 3.4. If* $\mathcal{B} \subset \mathcal{M}(\mathcal{X})$ *is a weakly compact and convex set, then given measures* $\mu_1, \ldots, \mu_K, \in \mathcal{M}(\mathcal{X})$*, let we have the following minimax formula*

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

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

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

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 $\Phi_c = \{(\varphi_i, \psi_i) \in C_b(\mathcal{X}) \times C_b(\mathcal{X}) : \varphi_i(x) - \psi_i(x') \leqslant c(x, x') \text{ for all } x, x' \in \mathcal{X}\}$. For each $\varphi_i, \psi_i \in C_b(\mathcal{X})$ fixed, the mapping $(\rho, \nu_i) \mapsto J_i(\nu_i, \varphi_i, \psi_i)$ is linear and lower semicontinuous with respect to the weak convergence of measures. For any

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 $\nu_i \leqslant \rho$ and Φ_c are convex, we are in a situation where Sion's minimax theorem applies. Therefore,

$$\min_{\rho,\nu_i\in\mathcal{B},\nu_i\leqslant\rho} \sup_{\varphi_i,\psi_i\in\Phi_c} \sum_{i\in\mathcal{Y}} J_i(\nu_i,\varphi_i,\psi_i) = \sup_{\varphi_i,\psi_i\in\Phi_c} \min_{\rho,\nu_i\in\mathcal{B},\nu_i\leqslant\rho} \sum_{i\in\mathcal{Y}} J_i(\nu_i,\varphi_i,\psi_i)$$

Since

$$\min_{\nu_i\leqslant\rho} \sum_{i\in\mathcal{Y}} J_i(\nu_i,\varphi_i,\psi_i) = \sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} \varphi_i(x)\mathrm{d}\mu_i(x) - \max(\psi_i(x'),0)\mathrm{d}\rho(x'),$$

we have

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

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 $\psi_i \geqslant 0$. Thus, the supremum is actually attained by some pair $(\varphi_i^*, \psi_i^*) \in \Phi_c$ with $\psi_i^* \geqslant 0$, $\varphi_i^* = (\psi_i^*)^c$ and $(\psi_i^*)^{c\bar{c}} = \psi_i^*$. □

Using Lemma 3.20 we can prove that there is no duality gap for bounded and Lipschitz costs. We will then show that there is no duality gap for general costs by approximation.

Proposition 3.21. *Given measures $\mu_1, \ldots, \mu_K$ and a bounded Lipschitz cost c satisfying the assumptions in Proposition 3.4, suppose that $\lambda, \tilde{\mu}_1, \ldots, \tilde{\mu}_K$ are optimal solutions to* (3.1.4)*. 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 3.20), then $\varphi_1^*, \ldots, \varphi_K^*$ are optimal solutions to problem* (3.2.13)*, $\psi_1^*, \ldots, \psi_K^*$ are optimal solutions to* (3.2.12)*, and the values of* (3.2.12)-(3.2.14) *are equal to* (3.1.4)*. 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 3.20 and the optimality of λ that there exists φ_i^*, ψ_i^* such that

$$\lambda(\mathcal{X}) + \sum_{i \in \mathcal{Y}} C(\mu_i, \tilde{\mu}_i) = \min_{\rho \in \mathcal{B}} \rho(\mathcal{X}) + \sum_{i \in \mathcal{Y}} \int_{\mathcal{X}} \varphi_i^*(x) d\mu_i(x) - \psi_i^*(x') d\rho(x'), \quad (3.2.15)$$

$\psi_i^*(x') \geqslant 0$, and $(\varphi_i^*)^{\bar{c}}(x') = \psi_i^*(x')$, $(\psi_i^*)^c(x) = \varphi_i^*(x)$ for all $1 \leqslant i \leqslant K$ and $x, x' \in \mathcal{X}$. If there exists $x' \in \mathcal{X}$ such that $\sum_{i \in \mathcal{Y}} \psi_i^*(x') > 1$, then we can make the right hand side of (3.2.15) smaller than the left hand side by choosing $\rho = M\delta_{x'}$ for some sufficiently large value of M. Hence, it follows that $\sum_{i \in \mathcal{Y}} \psi_i^*(x) \leqslant 1$ everywhere. Thus, the ψ_i^* are feasible solutions to problem (3.2.12) and, by Proposition 3.19 $(\psi_i^*)^c = \varphi_i^*$ are feasible solutions to (3.2.13). Finally, if we choose $\rho = 0$, it follows that

$$(3.1.4) = \lambda(\mathcal{X}) + \sum_{i \in \mathcal{Y}} C(\mu_i, \tilde{\mu}_i) \leqslant \sum_{i \in \mathcal{Y}} \varphi_i^*(x) d\mu_i(x) \leqslant (3.2.13) = (3.2.12) \leqslant (3.1.4)$$

where the second last equality follows from Proposition 3.19 and the last inequality holds trivially by duality. Therefore, we can infer that there is no duality gap. □

Proposition 3.22. *Given measures $\mu_1, \ldots, \mu_K$, if c is a cost that satisfies Assumption 2.5, then problems* (3.2.12)-(3.2.14) *all have the same value as* (3.1.4)*.*

Remark 3.23. *Note that we do not claim that the supremums in* (3.2.12)-(3.2.14) *are attained.*

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

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

and $c_j(x, x') := j\eta(\frac{\tilde{c}_j(x,x')}{j})$. It then follows that c_j is a bounded Lipschitz cost that satisfies the assumptions of Proposition 3.4. 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 (3.1.4) and (3.2.13) respectively with cost c_j. From Proposition 3.21 we know that $\alpha_j = \beta_j$. Let α, β denote the optimal values of Problems (3.1.4) and (3.2.13) respectively with the original cost c. Since we already know that $\beta \leqslant \alpha$, our goal is to show that $\alpha \leqslant \beta$.

Exploiting the fact that c_j is increasing with respect to j, if $g_1^{j_0}, \ldots, g_K^{j_0}$ is a feasible solution to (3.2.13) for the cost c_{j_0}, then it is also a feasible solution to (3.2.13) for c. Therefore, $\lim_{j\to\infty} \beta_j \leqslant \beta$.

On the other hand, let λ^j and $\tilde{\mu}_1^j, \ldots, \tilde{\mu}_K^j$ be optimal solutions to (3.1.4) 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 3.4, 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 $j \geqslant j_0$

$$\alpha_j = \lambda^j(\mathcal{X}) + \sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} c_j(x,x')\mathrm{d}\pi_i^j(x,x') \geqslant \lambda^j(\mathcal{X}) + \sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} c_{j_0}(x,x')\mathrm{d}\pi_i^j(x,x').$$

Therefore,

$$\liminf_{j\to\infty} \alpha_j \geqslant \lambda(\mathcal{X}) + \sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} c_{j_0}(x,x')\mathrm{d}\pi_i(x,x').$$

Taking a supremum over j_0, it follows that

$$\liminf_{j\to\infty} \alpha_j \geqslant \lambda(\mathcal{X}) + \sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} c(x,x')\mathrm{d}\pi_i(x,x') \geqslant \alpha.$$

Thus, $\alpha \leqslant \liminf_{j\to\infty} \alpha_j = \liminf_{j\to\infty} \beta_j = \beta$. Thanks to Proposition 3.19, it follows that (3.1.4) and (3.2.12)-(3.2.14), all have the same optimal value. □

3.3 Proof of Theorem 3.3

In this section, we prove Theorem 3.3 and return to the adversarial problem (2.5).

Theorem 3.3: upper bound

First we show that

$$\frac{1}{21}B_\mu^* \leqslant \inf_{\pi\in\Pi_K(\mu)} \int_{\mathcal{Z}_*^K} \mathbf{c}(z_1,\dots,z_K)\,d\pi(z_1,\dots,z_K).$$

To see this, recall that B_μ^* is, according to Proposition 3.17, equal to

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

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

Let $\pi \in \Pi_K(\mu)$, and for given $\vec{z} = (z_1,\dots,z_K) \in \mathcal{Z}_*^K$, let $\gamma_{\vec{z}} \in \Upsilon_{\widehat{\mu}_{\vec{z}}}$ be a solution for problem (3.2.10) (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 $\gamma \in \Upsilon_{\frac{1}{21}\mu}$. Indeed, for any test function $g : \mathcal{X} \to \mathbb{R}$ we have:

$$\begin{aligned}
\int_{(\mathcal{X}\times[0,1])^K} r_i g(x_i)\,d\gamma(\vec{x},\vec{r}) &= \int_{\mathcal{Z}_*^K} \left(\int_{(\mathcal{X}\times[0,1])^K} r_i g(x_i)\,d\gamma_{\vec{z}}(\vec{x},\vec{r}) \right) d\pi(z_1,\dots,z_K) \\
&= \frac{1}{K} \int_{\mathcal{Z}_*^K} \left(\sum_{j:z_j\neq \text{🕱}} g(x_j)\mathbb{1}_{i_j=i} \right) d\pi(z_1,\dots,z_K) \\
&= \frac{1}{21} \int_{\mathcal{X}} g(x)\,d\mu_i(x).
\end{aligned}$$

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* 3.1, we conclude that

$$\frac{1}{21}B^*_\mu = B^*_{\frac{1}{21}\mu} = \inf_{\gamma\in\Upsilon_{\frac{1}{21}\mu}}\int_{(\mathcal{X}\times[0,1])^K}\widetilde{c}(\vec{x},\vec{r})\,d\gamma(\vec{x},\vec{r}) \leqslant \inf_{\pi\in\Pi_K(\mu)}\int_{\mathcal{Z}_*^K}\mathbf{c}(\vec{z})\,d\pi(\vec{z}).$$

Theorem 3.3: lower bound

Now, it is sufficient to show

$$\inf_{\pi\in\Pi_K(\mu)}\int \mathbf{c}(z_1,\dots,z_K)\,d\pi(z_1,\dots,z_K) \leqslant \frac{1}{21}B^*_\mu.$$

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

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

For each $i \in \mathcal{Y}$, define

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

It is thus clear from the above computation and definition that

$$\sum_{j=1}^{K}\int_{\mathcal{X}\times\mathcal{Y}}\phi_j(z_j)\frac{1}{2}d\mu(z_j)+\frac{1}{2}\sum_{j=1}^{K}\phi_j(\text{👻})=\sum_{i\in\mathcal{Y}}\int_{\mathcal{X}}\psi_i(x_i)\frac{1}{2}d\mu_i(x_i). \tag{3.3.1}$$

Our goal is now to show that $\{\psi_i : i \in \mathcal{Y}\}$ is feasible for problem (3.2.13) (working with the normalized measure $\frac{1}{2|}\mu$). We start with a preliminary lemma and an example illustrating the strategy behind the proof of this fact. The precise statement appears in Proposition 3.25 below.

Lemma 3.24. *Given* $(z_1,\dots,z_K)\in\mathcal{Z}_*^K$, *let* $A=\{j\in\mathcal{Y} : z_j\neq\text{👻}\}$. *Suppose that for each* $j\in A$ $z_j=(x_j,j)$. *Then, for each* $\phi\in\Phi$,

$$\sum_{j=1}^{K}\phi_j(z_j)\leqslant\frac{1}{K}+\frac{1}{K}c_A. \tag{3.3.2}$$

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

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

where

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

For simplicity, assume that $A=\{1,\dots,S\}$. By Lemma 3.18,

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

where we can pick $x_{S+1},\dots,x_K$ arbitrarily. Let $m_A=\frac{1}{K}$ and $m_{A'}=0$ for $A'\neq A$. It is easy to check that such m is feasible for (3.2.9) since $r_s=\frac{1}{K}$ for $1\leqslant s\leqslant S$ and $r_j=0$ for $j\notin A$. So, (3.2.9) implies

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

The conclusion follows. □

We now present specific examples which illustrate why $\{\psi_i : i \in \mathcal{Y}\}$ is feasible for (3.2.13), that is, we need to show that for any $(x_1, \ldots, x_K) \in \mathcal{X}^K$ and for any $A \in S_K$ we have

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

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

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

or, after a rearrangement of the summands:

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

We can bound the first line above using (3.3.2):

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

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

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

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 $\phi_j(\text{👻})$'s available is in this case $K|A| = 18$ while the required number of $\phi_j(\text{👻})$'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 $\phi_j(\text{👻})$'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 $1 \leqslant j \leqslant 9$ is exactly 1. Filling $\emptyset_j$'s with $\phi_j(\text{👻})$'s, and using the

fact that the number of $\phi_j(\text{👻})$'s for each $1 \leqslant j \leqslant 9$ is 2, it follows that

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

Observe that for $(z_1, \dots, z_K) = ((x_1, 1), (x_2, 2), \dots, (x_1, 1), (x_2, 2), \text{👻})$, $\widehat{\mu}_{\vec{z}} = \frac{4}{9}\delta_{(x_1,1)} + \frac{4}{9}\delta_{(x_2,2)}$. Factoring out the 4 (see *Remark* 3.1) and applying (3.3.2), what we obtain is

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

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 $\sum_{j=1}^{9} \phi_j(\text{👻}) \leqslant 0$, it follows that

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

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 3.24 in the most effective way possible. In the following proposition we state precisely our aim and prove it by such strategy.

Proposition 3.25. *Let* $(\phi_1, \dots, \phi_K) \in \Phi$ *be a feasible dual potential. For each* $i \in \mathcal{Y}$*, define*

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

Then $\{\psi_i : i \in \mathcal{Y}\}$ *is feasible for* (3.2.13).

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

$$\sum_{i \in A} \psi_i(x_i) \leqslant 1 + c_A. \tag{3.3.3}$$

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

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

Note that for each $1 \leqslant s \leqslant S$, $|\{u(s,j) : 1 \leqslant j \leqslant K\}| = \frac{K}{S}$, 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 (3.3.2),

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

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

$$\begin{aligned}\sum_{i\in A} \psi_i(x_i) &= \sum_{s=1}^{S}\sum_{j=1}^{K} \phi_j(x_{u(s,j)}, u(s,j)) + S\sum_{j=1}^{K} \phi_j(\text{👻}) \\ &\leqslant \sum_{s=1}^{S} \left(\frac{1}{S} + \frac{1}{S} c_A\right) \\ &= 1 + c_A,\end{aligned}$$

proving the desired inequality in the first case.

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

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

Construct a $K \times K$ arrangement in the following way: for each $1 \leqslant s \leqslant S$ we set the $\nu(s,k)$-th element to be $\phi_{\nu(s,k)}(x_s, s)$, 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 $\phi_{\nu(1,k)}(x_1, 1), \ldots, \phi_{\nu(S,k)}(x_S, S)$ as non-empty elements, which are placed from the $\nu(1,k)$-th coordinate to the $\nu(S,k)$-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 $K(K-S) \leqslant KS$. For each $1 \leqslant k \leqslant K$, collect all the $\phi_j(\text{👻})$'s such that $j \notin A_k := \{\nu(1,k), \ldots, \nu(S,k)\}$. 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 $\phi_j(\text{👻})$. From the above and an application of

(3.3.2), we deduce that

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

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 $1 \leqslant s \leqslant S$, 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 $\phi_{\nu(s,(m-1)S+1)}(x_1,1),\dots,\phi_{\nu(s,mS)}(x_S,S)$ 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 $1 \leqslant j \leqslant K$, we would be done since the number of $\phi_j(\text{👻})$ is S for each j, whence it would be possible to replace the $\emptyset_j$'s with $\phi_j(\text{👻})$'s. We show that this is indeed the case.

For each merged row, its non-empty elements are

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

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

$\nu(s, K)$. The definition of $\nu(s, k)$ yields that

$$\nu(s, \lfloor \frac{K}{S} \rfloor S + 1) = \lfloor \frac{K}{S} \rfloor S + s \text{ if } 1 \leqslant s \leqslant K - \lfloor \frac{K}{S} \rfloor S, \tag{3.3.4}$$

$$\nu(s, \lfloor \frac{K}{S} \rfloor S + 1) = \lfloor \frac{K}{S} \rfloor S + s - K \text{ if } K - \lfloor \frac{K}{S} \rfloor S + 1 \leqslant s \leqslant S \tag{3.3.5}$$

and

$$\nu(s, K) = K \text{ if } s = 1, \tag{3.3.6}$$

$$\nu(s, K) = s - 1 \text{ if } 2 \leqslant s \leqslant S. \tag{3.3.7}$$

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

(i) $\lfloor \frac{K}{S} \rfloor S + 1 \leqslant j \leqslant K$: By (3.3.4), if $1 \leqslant s \leqslant K - \lfloor \frac{K}{S} \rfloor S$, then the s-th row has $\emptyset_j$ for $\lfloor \frac{K}{S} \rfloor S + s \leqslant j \leqslant K$. Also, by (3.3.5) and (3.3.7), if $K - \lfloor \frac{K}{S} \rfloor S + 1 \leqslant s \leqslant S$, then no merged row has such $\emptyset_j$. Hence, the number of $\emptyset_j$ is $j - \lfloor \frac{K}{S} \rfloor S$.

(ii) $S \leqslant j \leqslant \lfloor \frac{K}{S} \rfloor S$: It follows from (3.3.4) and (3.3.5) that either $\nu(s, \lfloor \frac{K}{S} \rfloor S + 1) > \lfloor \frac{K}{S} \rfloor S$ or $\nu(s, \lfloor \frac{K}{S} \rfloor S + 1) < S$. Similarly, it follows from (3.3.6) and (3.3.7) that either $\nu(s, K) > \lfloor \frac{K}{S} \rfloor S$ or $\nu(s, K) < S$. Since the index j of $\emptyset_j$ of the s-th merged row varies from $\nu(s, \lfloor \frac{K}{S} \rfloor S + 1)$ to $\nu(s, K)$, the number of $\emptyset_j$ is 0.

(iii) $S - (K - \lfloor \frac{K}{S} \rfloor S) + 1 \leqslant j \leqslant S - 1$: By (3.3.5) and (3.3.7), if $S - (K - \lfloor \frac{K}{S} \rfloor S) + 1 \leqslant j \leqslant S - 1$, 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) $1 \leqslant j \leqslant S - (K - \lfloor \frac{K}{S} \rfloor S)$: Similar to (iii), if $1 \leqslant j \leqslant S - (K - \lfloor \frac{K}{S} \rfloor S)$, 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

$$\text{the number of } \emptyset_j = \begin{cases} j - \lfloor \frac{K}{S} \rfloor S & \text{for } \lfloor \frac{K}{S} \rfloor S + 1 \leqslant j \leqslant K, \\ 0 & \text{for } S \leqslant j \leqslant \lfloor \frac{K}{S} \rfloor S, \\ S - j & \text{for } S - (K - \lfloor \frac{K}{S} \rfloor S) + 1 \leqslant j \leqslant S - 1, \\ K - \lfloor \frac{K}{S} \rfloor S & \text{for } 1 \leqslant j \leqslant S - (K - \lfloor \frac{K}{S} \rfloor S). \end{cases} \tag{3.3.8}$$

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

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

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

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

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

To summarize, in the remaining original rows

$$\text{the number of } \emptyset_j = \begin{cases} K - j & \text{for } \lfloor \frac{K}{S} \rfloor S + 1 \leqslant j \leqslant K, \\ K - \lfloor \frac{K}{S} \rfloor S & \text{for } S \leqslant j \leqslant \lfloor \frac{K}{S} \rfloor S, \\ K - \lfloor \frac{K}{S} \rfloor S - S + j & \text{for } S - (K - \lfloor \frac{K}{S} \rfloor S) + 1 \leqslant j \leqslant S - 1, \\ 0 & \text{for } 1 \leqslant j \leqslant S - (K - \lfloor \frac{K}{S} \rfloor S). \end{cases} \tag{3.3.9}$$

Combining (3.3.8) with (3.3.9), the total number of $\emptyset_j$ is always exactly equal to $K - \lfloor \frac{K}{S} \rfloor S$ which is always less than S. This allows us to replace every $\emptyset_j$ with a $\phi_j(\text{👻})$. Accordingly, using $\sum \phi_j(\text{👻}) \leqslant 0$, we deduce that

$$\sum_{i \in A} \psi_i(x_i) \leqslant \sum_{\text{merged rows}} \left(\sum_{\nu(s,j)} \phi_{\nu(s,j)}(x_s, s) + \sum_{l \neq \nu(s,j)} \phi_l(\text{👻}) \right) + \sum_{\text{remaining rows}} \left(\sum_{\nu(s,j)} \phi_{\nu(s,j)}(x_s, s) + \sum_{l \neq \nu(s,j)} \phi_l(\text{👻}) \right).$$

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 (3.3.2), we obtain

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

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

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

(3.3.2) immediately implies

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

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

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

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 (3.3.1)) is feasible for (3.2.13). Consequently, this leads to

$$(3.1.8) \leqslant \frac{1}{21} (3.2.13) \tag{3.3.10}$$

In turn, by the equivalence between (3.2.13) and (3.1.4) by Proposition 3.22, this automatically implies that

$$(3.1.8) \leqslant \frac{1}{21} B_\mu^*.$$

Finally, combining with Corollary 3.28 below (which establishes that under Assumption 2.5 there is no duality gap for the MOT problem (3.0.1)) we obtain the desired inequality relating the minimum value for the MOT problem and B_μ^*.

Returning to the adversarial problem (2.5)

We begin by establishing that, under Assumption 2.5, the cost $\mathbf{c}$ is lower semi-continuous with respect to a suitable notion of convergence.

Proposition 3.26. *Let $\mathcal{Z}_* = \mathcal{Z} \cup \{\text{👻}\}$ on which 👻 is considered as an isolated point. Let $\widehat{d}$ be defined according to:*

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

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

$$\widehat{d}_K((z_1, \dots, z_K), (z'_1, \dots, z'_K)) := \max_{i \in \mathcal{Y}} \widehat{d}(z_i, z'_i).$$

Recall

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

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

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

Under Assumption 2.5, $\mathbf{c}$ is lower semi-continuous on $(\mathcal{Z}_*^K, \widehat{d}_K)$.

Remark 3.27. *Note that $(\mathcal{Z}_*^K, \widehat{d}_K)$ is still a Polish space.*

Proof. Suppose $\vec{z}^n = (z_1^n, \dots, z_K^n)$ converges to $\vec{z} = (z_1, \dots, z_K)$ in $(\mathcal{Z}_*^K, \widehat{d}_k)$. Without loss of generality, assume that $z_1, \dots, z_L = \text{👻}$ for all $1 \leqslant L \leqslant K$. If $L = K$, the claim would be trivial and so we can focus on the case $L < K$. By the definition of $\widehat{d}_K$, without loss of generality we can further assume that $z_1^n, \dots, z_L^n = \text{👻}$ for all n, and likewise, for each $L + 1 \leqslant j \leqslant K$, 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 3.18 we have

$$\mathbf{c}(z_1^n, , \dots, z_K^n) = B^*_{\widehat{\mu}_{\vec{z}^n}} = \inf_{m : S_K \to \mathbb{R}} \sum_{A \subseteq \{L+1, \dots, K\}} m_A \big(c_A(x_{L+1}^n, \dots, x_K^n) + 1 \big), \quad (3.3.11)$$

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

$\frac{1}{K}, \quad \forall i = L+1, \dots, K.$

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

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

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

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

By the compactness property in Assumption 2.5 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

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

as we wanted to show.

Returning to (3.3.11), we can find for each $n \in \mathbb{N}$ a collection of feasible $\{m^n_A\}_{A\subseteq\{L+1,\dots,K\}}$ such that

$$\liminf_{n\to\infty} \sum_{A\subseteq\{L+1,\dots,K\}} m^n_A \big(c_A(x^n_{L+1}, \dots, x^n_K) + 1\big) = \liminf_{n\to\infty} \mathbf{c}(z^n_1, \dots, z^n_K).$$

Using the Heine-Borel theorem in Euclidean space, we can assume without the loss of generality that for every A, m^n_A 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:

$$\mathbf{c}(z_1, \dots, z_K) \leqslant \sum_{A\subseteq\{L+1,\dots,K\}} m_A \big(c_A(x^n_{L+1}, \dots, x^n_K) + 1\big) \leqslant \liminf_{n\to\infty} \mathbf{c}(z^n_1, \dots, z^n_K).$$

□

Corollary 3.28. (*Duality of MOT*) *Under Assumption 2.5,*

$$\inf_{\pi\in\Pi_K(\mu)}\int_{\mathcal{Z}_*{}^K}\mathbf{c}(z_1,\dots,z_K)\mathrm{d}\pi(z_1,\dots,z_K)$$
$$=\sup_{\phi\in\Phi}\left\{\sum_{j=1}^{K}\int_{\mathcal{X}\times\mathcal{Y}}\phi_j(z_j)\frac{1}{2}\mathrm{d}\mu(z_j)+\frac{1}{2}\sum_{j=1}^{K}\phi_j(\text{👻})\right\}.$$

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

Proof. From Proposition 3.26 it follows that the cost function $\mathbf{c}(z_1,\dots,z_K)$ 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 $\mathbf{c}(z_1,\dots,z_K)$ and the compactness of $\Pi_K(\mu)$. □

Corollary 3.29. *Under Assumption 2.5,* (3.1.2)=(3.1.3).

Proof. By the upper bound from section 3.3 we have

$$\frac{1}{21}B_\mu^*\leqslant\min_{\pi\in\Pi_K(\mu)}\int\mathbf{c}(z_1,\dots,z_K)\mathrm{d}\pi(z_1,\dots,z_K).$$

On the other hand, from (3.3.10) and Corollary 3.28 we have

$$\min_{\pi\in\Pi_K(\mu)}\int\mathbf{c}(z_1,\dots,z_K)\mathrm{d}\pi(z_1,\dots,z_K)=(3.1.8)\leqslant\frac{1}{21}(3.2.13)\leqslant\frac{1}{21}B_\mu^*.$$

Combining these two inequalities we conclude that all the above terms must be equal. In particular, $(3.2.13)=B_\mu^*$. Finally, by Proposition 3.19 we know that $(3.2.13)=(3.2.12)=(3.1.2)$. In particular, $(3.1.3)=B_\mu^*=(3.1.2)$. □

Corollary 3.30. *Suppose that Assumption 2.5 holds and that (π^*,ϕ^*) is a solution pair*

for the MOT problem and its dual. Define f^* *and* $\widetilde{\mu}^*$ *according to:*

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

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* (2.5)*.*

Remark 3.31. *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 (3.1.3). More explicitly, we show that for any $f \in \mathcal{F}$ and for any $\widetilde{\mu}$,

$$B(f, \widetilde{\mu}^*) + C(\mu, \widetilde{\mu}^*) \leqslant B(f^*, \widetilde{\mu}^*) + C(\mu, \widetilde{\mu}^*) \leqslant B(f^*, \widetilde{\mu}) + C(\mu, \widetilde{\mu}). \tag{3.3.12}$$

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 (2.10),

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

where the third inequality follows from (3.3.3). Hence,

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

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

$$f_i^*(\widetilde{x}_i) \geqslant \sum_{j=1}^{K} \phi_j^*(x_i, i) + \sum_{j=1}^{K} \phi_j^*(\text{👻}) - c(\widetilde{x}_i, x_i). \tag{3.3.13}$$

Using $\sum_{A\in S_K(i)} \mu_{i,A} = \mu_i$ and (3.3.13), the optimality of ϕ^* implies that

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

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 (3.3.12). The first inequality of (3.3.12) is straightforward, since the definition of B_μ^* in (3.1.4) and the optimality of $\widetilde{\mu}^*$ imply that

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

For the second inequality of (3.3.12), let arbitrary $\widetilde{\mu}$ be fixed and $\pi_i \in \Gamma(\widetilde{\mu}_i, \mu_i)$ be an optimal coupling for each $i \in \mathcal{Y}$. Then,

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

Applying (3.3.13) yields that

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

Therefore, $(f^*, \widetilde{\mu}^*)$ is a saddle point for (3.1.3), hence for (3.1.2) and (2.5) also. □

Remark 3.32. *Many recent papers have tried to analyze adversarial learning from a game-theoretic perspective Bose et al. (2020); Meunier et al. (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 3.33. *Let π^* be a solution of the MOT problem (3.0.1) 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 $\sigma : \mathcal{Y} \to \mathcal{Y}$ 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 **c** is invariant under permutations and the fact that all marginals of π^* are the same. □

3.4 Examples and Numerical experiments

Through this section, the cost c is as in Example 2.3. 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.

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 👻, then $\mathbf{c}(z_1, z_2) = 0$. If only one of them is 👻, then the cost is $\frac{1}{2}$. Finally, consider the case where $z_1, z_2 \neq$ 👻. 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 $d(x_1, x_2) \leqslant 2\varepsilon$. Thus,

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

To summarize,

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

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

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

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

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

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 (3.0.1) (such a solution exists thanks to Proposition 3.28). Thanks to Corollary 3.33, 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:

$$\int_{\mathcal{Z}\times\mathcal{Z}} \tilde{\varphi}(z_1,z_2)\,d\tilde{\pi}(z_1,z_2) := 2\int_{\mathcal{Z}\times\mathcal{Z}} \tilde{\varphi}(z_1,z_2)\,d\pi(z_1,z_2) + \int_{\mathcal{Z}\times\{\text{🕵}\}} \tilde{\varphi}(z_1,z_1)\,d\pi(z_1,z_2) + \int_{\{\text{🕵}\}\times\mathcal{Z}} \tilde{\varphi}(z_2,z_2)\,d\pi(z_1,z_2),$$

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

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

The above computations show that our results indeed recover those from García Trillos and Murray (2022) for the binary case.

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 $\omega_1 \geqslant \omega_2 \geqslant \omega_3 > 0$ and $\sum \omega_i = 1$. Let $\varepsilon > 0$ be given and consider the cost from Example 2.3 with d as the Euclidean distance (for simplicity). We will explicitly construct an optimal robust classifier and an optimal adversarial attack for this problem. Even in this simple scenario, one can observe non-trivial situations.

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

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

where $\overline{B}(x,r) = \{x' : d(x,x') \leqslant r\}$, 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 (2.5) can not disambiguate the values of f outside of this set). We consider 4 non-trivial configurations and one trivial one. Figure 3.4 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 $1 \leqslant i \neq j \leqslant 3$. This is a trivial case. We claim that for any $\widetilde{\mu}_i$ such that $W_\infty(\omega_i \delta_{x_i}, \widetilde{\mu}_i) \leqslant \varepsilon$, $((\mathbb{1}_{\overline{B}(x_1,\varepsilon)}, \mathbb{1}_{\overline{B}(x_2,\varepsilon)}, \mathbb{1}_{\overline{B}(x_3,\varepsilon)}), (\widetilde{\mu}_1, \widetilde{\mu}_2, \widetilde{\mu}_3))$ is a saddle point for (2.5). 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 $d(\overline{x}, x_i) \leqslant \varepsilon$ for all $1 \leqslant i \leqslant 3$. 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 $1 \leqslant i \leqslant 3$, 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 $\omega_1 \geqslant \omega_2 \geqslant \omega_3$. 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 $d(x_1, x_2) \leqslant 2\varepsilon$, $d(x_1, x_3) > 2\varepsilon$ 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 $W_\infty(\widetilde{\mu}_3, \omega_3 \delta_{x_3}) \leqslant \varepsilon$. We claim that $(\widehat{f}, \widehat{\mu})$ is a saddle point. For any $(f_1, f_2, f_3) \in \mathcal{F}$ we have

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

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

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

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. $d(x_i, x_j) \leqslant 2\varepsilon$ 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 $K = 2$, $d(x_1, x_2) \leqslant 2\varepsilon$ if and only if $\overline{B}(x_1, \varepsilon) \cap \overline{B}(x_2, \varepsilon) \neq \emptyset$. However, when $K \geqslant 3$, 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 $1 \leqslant i \leqslant 3$ 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, $\alpha_i \geqslant 0$ since $\omega_1 \leqslant \omega_2 + \omega_3$. 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 $d(x_i, x_j) \leqslant 2\varepsilon$ 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 $W_\infty(\widetilde{\mu}_i, \omega_i \delta_{x_i}) \leqslant \varepsilon$, it must be the case that $A_{ij} \cap \mathrm{spt}(\widetilde{\mu}_k) = \emptyset$ if $k \neq i, j$. For each $1 \leqslant i \leqslant 3$, 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 $(f_1, f_2, f_3) \in \mathcal{F}$,

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

where the second to last inequality follows from the fact that $\sum f_i(x) \leqslant 1$ 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 $(\widetilde{\mu}_1, \widetilde{\mu}_2, \widetilde{\mu}_3)$

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

Note that since $W_\infty(\widetilde{\mu}_i, \omega_i \delta_{x_i}) \leqslant \varepsilon$, $\mathrm{spt}(\widetilde{\mu}_i) \cap A_j = \emptyset$ 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

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

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 $d(x_1, x_2) \leqslant 2\varepsilon$ and $d(x_1, x_3) \leqslant 2\varepsilon$, 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

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

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) $\omega_1 \geqslant \omega_2 + \omega_3$. 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 $\kappa = \omega_1 - (\omega_2 + \omega_3) \geqslant 0$ 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 $(f_1, f_2, f_3) \in \mathcal{F}$,

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

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, $\omega_1 \geqslant \frac{1}{2}$, since $\omega_1 \geqslant \omega_2 + \omega_3$ 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.

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 (3.0.1) 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 et al. (2015, 2019). Works that study the computational complexity of generic MOT problems include: Lin et al. (2022); Tupitsa et al. (2020); Haasler et al. (2021); Carlier (2022). In particular, Lin et al. (2022) and Tupitsa et al. (2020) prove the complexity of MOT Sinkhorn algorithm to be $\widetilde{O}(K^3 n^K \epsilon^{-2})$ and $\widetilde{O}(K^3 n^{K+1} \epsilon^{-1})$, 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

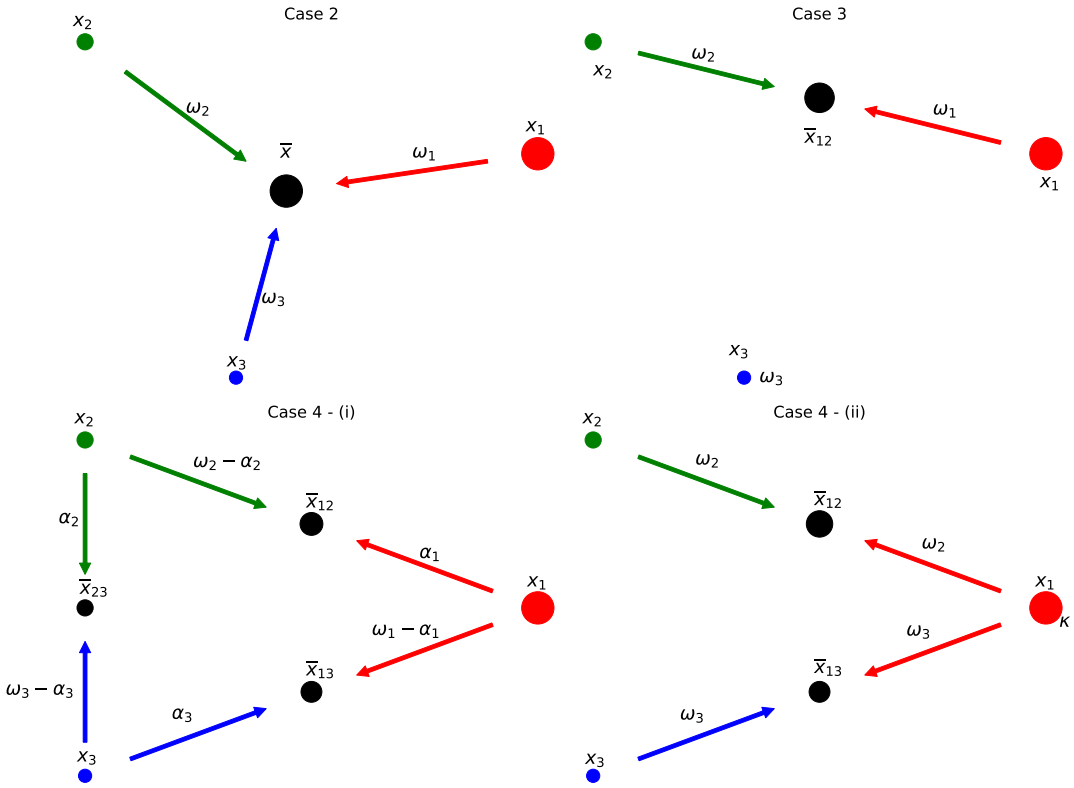

Figure 3.4: Illustrations of the adversarial attacks in all cases from section 3.4. Weights on arrows indicate the amount of mass the adversary moves to a perturbed point. $\bar{x}$'s are the support of λ in (3.1.4). One observes that the support of λ depends on both the geometry of data distributions and their magnitudes.

$c = c_\varepsilon$ from Example 2.3 with d the Euclidean distance in $\mathbb{R}^2$ and different values of ε. In Figure 3.5 we show the labels assigned to the data by the (approximate) robust classifier, which we computed using Corollary 3.30 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 (2.5)) 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

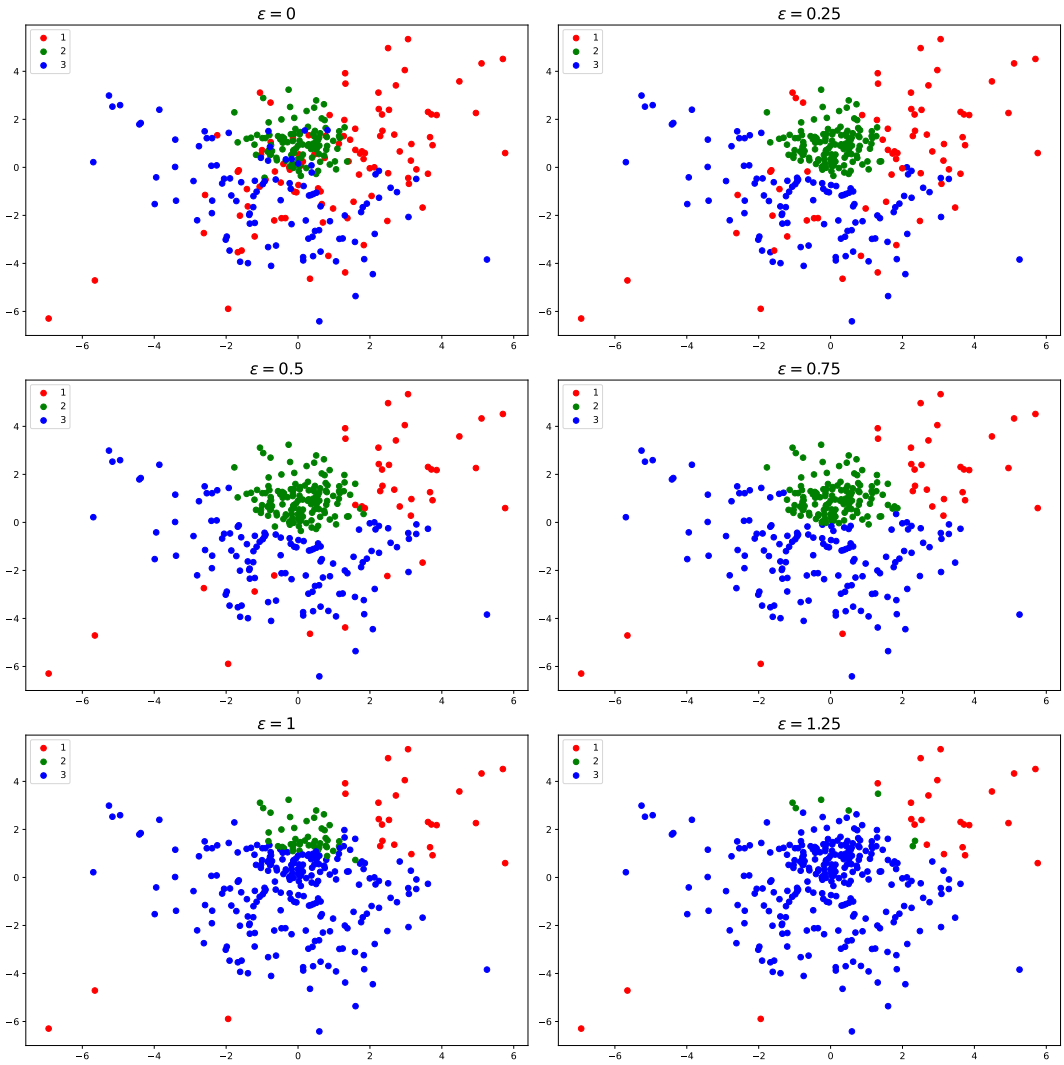

Figure 3.5: Three Gaussians in $\mathbb{R}^2$. One can observe that as ε grows the robust classifying rule becomes simpler, as expected.

Example 2.3 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 3.6. 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

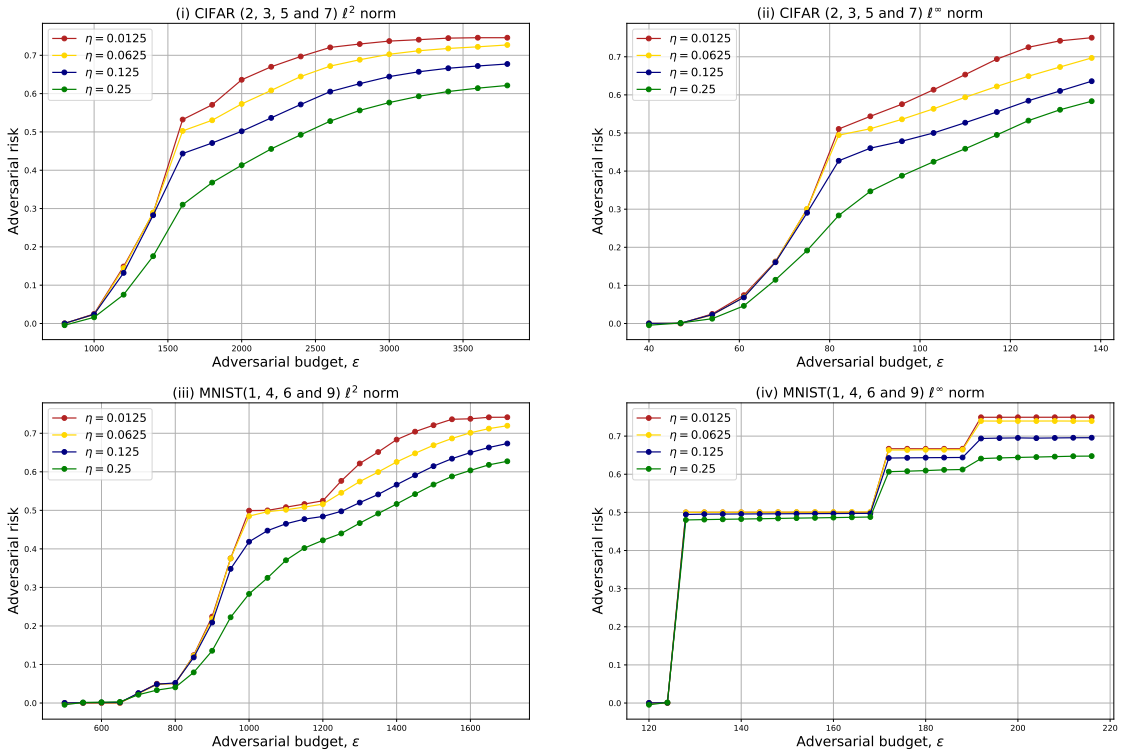

Figure 3.6: Adversarial risks (2.5) computed using the multimarginal Sinkhorn algorithm. η is the entropic regularization parameter of the Sinkhorn algorithm. The maximum adversarial risk in all cases is 0.75 because we consider 4 classes and an equal number of points in each class. Due to the entropic 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^*.

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 3.6 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 will always produce an upper bound for B_μ^* since the regularization term effectively restricts the solution space of (3.0.1). 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 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.

3.5 Summary

In this chapter 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.

4 ON THE EXISTENCE OF SOLUTIONS TO ADVERSARIAL TRAINING IN MULTICLASS CLASSIFICATION

This chapter is based on García Trillos et al. (2023a) which is a joint work with Nicolas García Trillos and Matt Jacobs.

4.1 Outline of this chapter

This chapter is organized as follows. In section 4.2, we introduce the main results of this chapter. In section 4.3 we lay down the main mathematical tools for analyzing the DRO model. Part of these tools come directly from chapter 3, while others are developed here. In section 4.4 we prove our main results: first, we prove the existence of solutions for the DRO model; then we prove that solutions to the DRO model are solutions to the closed-ball model; finally, we relate the closed-ball model with the open-ball model. Lastly, in section 4.5 we wrap up the chapter by summarizing it.

4.2 Main results

Our first main theorem discusses the existence of (Borel) solutions for problem (2.5) under Assumption 2.5 on the cost c.

Theorem 4.1. *Suppose that* $c : \mathcal{X} \times \mathcal{X} \to [0, \infty]$ *satisfies Assumption 2.5. Then there exists a (Borel) solution* f^* *of* (2.5). *Furthermore, there exists* $\tilde{\mu}^* \in \mathcal{P}(\mathcal{Z})$ *such that* $(f^*, \tilde{\mu}^*)$ *is a saddle point for* (2.5). *In other words, the following holds: for any* $\widetilde{\mu} \in \mathcal{P}(\mathcal{Z})$ *and any* $f \in \mathcal{F}$ *we have*

$$R(f^*, \widetilde{\mu}) - C(\mu, \widetilde{\mu}) \leqslant R(f^*, \widetilde{\mu}^*) - C(\mu, \widetilde{\mu}^*) \leqslant R(f, \widetilde{\mu}^*) - C(\mu, \widetilde{\mu}^*). \tag{4.2.1}$$

When the cost function c is regular enough or when μ is an empirical measure, we can reduce the problem of finding a solution f^* of (2.5) to the problem of solving

the dual of a *generalized barycenter problem* or the dual of a *multimarginal optimal transport* problem. These connections were first put forward in our earlier work García Trillos et al. (2023) and will be discussed again in section 4.3, concretely in Proposition 4.19. Unfortunately, when the cost is only lower semi-continuous (e.g., for $c = c_\varepsilon$ as in (2.8)) and when μ is an arbitrary Borel probability measure, we can not directly use the content of Proposition 4.19 to guarantee the existence of (Borel) solutions f^*. One possible way around this issue is to consider an approximation argument where, instead of working directly with the cost c, we work with smoother approximations c_n to c for which we can use some of our previous theory. At a high level, we can thus reduce finding solutions for the DRO problem (2.5) to that for an MOT or a generalized barycenter (or sequences thereof).

Remark 4.2. *When the cost c has the form c_ε in (2.8), Assumption 2.5 reduces to the requirement that bounded subsets in $\mathcal{X}$ are precompact, which we are anyway assuming in Assumption 2.5, according to remark 2.6. This is the case for Euclidean space or for a smooth manifold of finite dimension endowed with its geodesic distance.*

In order to discuss the existence of solutions to the problem (2.4) we actually first need to modify the problem and define it properly. To do this, we first introduce the *universal* σ-algebra of the space $\mathcal{X}$.

Definition 4.3 (Definition 2.2 in Nishiura (2008))**.** *Let $\mathcal{B}(\mathcal{X})$ be the Borel σ-algebra over $\mathcal{X}$ and let $\mathcal{M}(\mathcal{X})$ be the set of all signed σ-finite Borel measures over $\mathcal{X}$. For each $\nu \in \mathcal{M}(\mathcal{X})$, let $\mathcal{L}_\nu(\mathcal{X})$ be the completion of $\mathcal{B}(\mathcal{X})$ with respect to ν. The universal σ-algebra of $\mathcal{X}$ is defined as*

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

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

$$\overline{\mu}(A \times \{i\}) := \overline{\mu}_i(A), \quad \mathcal{F}orall A \in \mathcal{U}(\mathcal{X}),$$

where $\overline{\mu}_i$ is the extension of μ_i to $\mathcal{U}(\mathcal{X})$. Finally, we will use $\overline{\mathcal{U}}(\mathcal{Z})$ to denote the set of all $f = (f_1, \ldots, f_K)$ for which each f_i is universally measurable.

Remark 4.4. *If $(\mathcal{X}, d) = (\mathbb{R}^n, \|\cdot\|)$, then $\mathcal{U}(\mathcal{X})$ is the set of all Lebesgue measurable sets; see Theorem 4.2 in Nishiura (2008). So, any Lebesgue-measurable function is universally measurable and vice-versa.*

Having introduced the above notions, we can reformulate problem (2.4) as:

$$\overline{R}_\varepsilon := \inf_{f \in \mathcal{F}} \overline{R}_\varepsilon(f) := \inf_{f \in \mathcal{F}} \left\{ \sum_{i \in \mathcal{Y}} \int_{\mathcal{X}} \sup_{\tilde{x} \in \overline{B}_\varepsilon(x)} \{1 - f_i(\tilde{x})\} d\overline{\mu}_i(x) \right\}. \tag{4.2.2}$$

Although the difference with (2.4) is subtle (in (2.4) we use μ_i whereas in (4.2.2) we use $\overline{\mu_i}$), problem (4.2.2) is actually well-defined. Indeed, combining Lemma 4.2 in Pydi and Jog (2021b) with Corollary 7.42.1 in Bertsekas and Shreve (1996), originally from Luzin and Sierpiński (1919), it follows that for any Borel measurable f_i the function $x \mapsto \sup_{x \in \overline{B}_\varepsilon(x)} \{1 - f_i(\tilde{x})\}$ is universally measurable and thus the integrals on the right hand side of (4.2.2) are well defined.

Our second main result relates solutions of (2.5) with solutions of (4.2.2).

Theorem 4.5. *There exists a Borel solution of (2.5) for the cost function $c = c_\varepsilon$ from (2.8) that is also a solution of (4.2.2). In particular, there exists a (Borel) solution for (4.2.2).*

Finally, we connect problem (4.2.2) with problem (2.3).

Theorem 4.6. *For all but at most countably many $\varepsilon \geqslant 0$, we have $R^o_\varepsilon = \overline{R}_\varepsilon$. Moreover, for those $\varepsilon \geqslant 0$ for which this equality holds, every solution f^* of (4.2.2) is also a solution of (2.3).*

Remark 4.7. *In general, we can not expect the optimal adversarial risks of open-ball and closed-ball models to agree for all values of ε. To illustrate this, consider the simple setting of a two class problem (i.e., $K = 2$) where $\mu_1 = \frac{1}{2}\delta_{x_1}$ and $\mu_2 = \frac{1}{2}\delta_{x_2}$. Let $\epsilon_0 = \frac{1}{2}d(x_1, x_2)$. It is straightforward to check that $R^o_{\epsilon_0} = 0$ whereas $\overline{R}_{\epsilon_0} = 1/2$. Naturally, if we had selected any other value for $\varepsilon > 0$ different from ϵ_0 we would have obtained $R^o_\varepsilon = \overline{R}_\varepsilon$.*

From Theorem 4.1, Proposition 4.5, and Theorem 4.6 we may conclude that it is essentially sufficient to solve problem (2.5) to find a solution for all other formulations of the adversarial training problem discussed in this paper. Our results thus unify all notions of adversarial robustness into the single DRO problem (2.5). The advantage of (2.5) over the other formulations of the adversarial training problem is that it can be closely related to a generalized barycenter problem or an MOT problem, as has been discussed in detail in our previous work Garcıa Trillos et al. (2023) (see also section 4.3 below). In turn, either of those problems can be solved using computational optimal transport tools. From a practical perspective, it is thus easier to work with the DRO formulation than with the other formulations of adversarial training.

Discussion and literature review

The existence of measurable "robust" solutions to optimization problems has been a topic of interest not only in the context of adversarial training Pydi and Jog (2021b); Frank and Niles-Weed (2022); Frank (2022); Awasthi et al. (2021a,b) but also in the general *distributionally robust optimization* literature, e.g., Blanchet and Murthy (2019). Previous studies of robust classifiers use *the universal* σ*-algebra* not only to formulate optimization problems rigorously, but also as a feasible search space for robust classifiers. The proofs of these existence results rely on the pointwise topology of a sequence of universally measurable sets, the weak topology on the space of probability measures, and lower semi-continuity properties of $\overline{R}_\varepsilon(\cdot)$. The (universal) measurability of a minimizer is then guaranteed immediately by the definition of the universal σ-algebra. We want to emphasize that all the works Pydi and Jog (2021b); Frank and Niles-Weed (2022); Frank (2022); Awasthi et al. (2021a,b) prove their results in the binary ($K = 2$) classification setting with $\mathcal{X}$ the Euclidean space.

In contrast to the closed-ball model formulation, the objective in (2.5) is well-defined for all Borel probability measures $\tilde{\mu}$ and all $f \in \mathcal{F}$, as has been discussed in previous sections. The papers Pydi and Jog (2021b); Frank and Niles-Weed (2022);

Frank (2022); Awasthi et al. (2021a,b) can only relate, in the binary case, problems (2.5) and (4.2.2) when problem (2.5) is appropriately extended to the universal σ-algebra, yet it is not clear that such extension is necessary. For concreteness, we summarize some of the results in those works in the following theorem.

Theorem 4.8 (Pydi and Jog (2021b); Awasthi et al. (2021a,b); Frank (2022))**.** *Suppose* $K = 2$ *and* $\overline{\mu} \in \overline{\mathcal{P}}(\mathcal{Z})$*. Then, for any* $f \in \mathcal{F}$*, we have* $\sup_{\widetilde{x} \in \overline{B}_\varepsilon(x)} \{1 - f_i(\widetilde{x})\} \in \mathcal{U}(\mathcal{Z})$ *and*

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

where C *is defined in terms of the cost* c_ε *from* (2.8)*.*

Assume further that $(\mathcal{X}, d) = (\mathbb{R}^n, \|\cdot\|)$*. Then, for any* $f \in \mathcal{U}(\mathcal{Z})$*, it holds that* $\sup_{\widetilde{x} \in \overline{B}_\varepsilon(\cdot)} \{1 - f_i(\widetilde{x})\}$ *is universally measurable for each* i*. In addition, there exists a minimizer of the objective in* (4.2.2) *in the class of soft-classifiers that are universally measurable. Finally,* (4.2.2) *and* (2.5) *are equivalent, provided that the latter is interpreted as an optimization problem over the space of universally measurable soft-classifiers.*

In this paper, we use the universal σ-algebra to rigorously define the objective function in (4.2.2), but we will only consider elements in $\mathcal{F}$ (thus, Borel measurable soft-classifiers) as feasible classifiers. Indeed, based on some of our previous results in Garcıa Trillos et al. (2023), we prove the existence of Borel measurable robust classifiers of (2.5) for general lower semi-continuous c satisfying Assumption 2.5 only. Then, back to the closed-ball model, we prove the existence of Borel robust classifiers of (4.2.2). When we specialize our results to the binary classification setting (i.e., $K = 2$), we obtain the following improvement upon the results from Bhagoji et al. (2019); Pydi and Jog (2021a); Frank (2022).

Corollary 4.9. *Let* $K = 2$ *and let* $f^* \in \mathcal{F}$ *be any solution to the problem* (4.2.2)*. Then, for Lebesgue a.e.* $t \in [0, 1]$*, the pair* $(\mathbb{1}_{\{f_1^* \geqslant t\}}, \mathbb{1}_{\{f_1^* \geqslant t\}^c})$ *is also a solution to* (4.2.2)*.*

In particular, there exist solutions to the problem

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

Notice that Corollary 4.9 implies, for the binary case, the existence of robust hard-classifiers for the adversarial training problem, a property shared with the nominal risk minimization problem (2.2) that we discussed at chapter 2. Analogous results on the equivalence of the hard-classification and soft-classification problems in adversarial training under the binary setting have been obtained in Pydi and Jog (2021a,b); Bungert et al. (2023); García Trillos and Murray (2022). Unfortunately, when the number of classes is such that $K > 2$, the hard-classification and soft-classification problems in adversarial training may not be equivalent, as has been discussed in Section 5.2 of our work Garcıa Trillos et al. (2023).

In light of Theorem 4.6, one can conclude from Corollary 4.9 that for all but countably many $\varepsilon > 0$ the problem

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

admits solutions; notice that the above is the open-ball version of the optimization problem in Corollary 4.9. However, notice that the results in Bungert et al. (2023) guarantee existence of solutions for *all* values of ε. It is interesting to note that the technique used in Bungert et al. (2023) can not be easily adapted to the multiclass case $K > 2$. Specifically, it does not seem to be straightforward to generalize Lemma C.1 in Bungert et al. (2023) to the multiclass case. For example, if one used the aforementioned lemma to modify the coordinate functions f_i of a multiclass classifier f, one could end up producing functions for which their sum may be greater than one for some points in $\mathcal{X}$, thus violating one of the conditions for belonging to $\mathcal{F}$.

We observe, on the other hand, that the total variation regularization interpretation for the open ball model in the binary case discussed in Bungert et al. (2023)

continues to hold in the multiclass case. To make this connection precise, let us introduce the non-local TV functionals:

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

It is then straightforward to show that problem (2.3) is equivalent to

$$\inf_{f \in \mathcal{F}} \sum_{i=1}^{K} \int_{\mathcal{X}} (1 - f_i(x)) d\mu_i(x) + \varepsilon \sum_{i=1}^{K} \widetilde{\mathrm{TV}}_\varepsilon(f_i, \mu_i), \tag{4.2.3}$$

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

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

when f_i is a smooth enough function. The functional $\widetilde{\mathrm{TV}}_\varepsilon(f_i, \mu_i)$ is thus connected to more standard notions of (weighted) total variation in Euclidean space. This heuristic can be formalized further via variational tools, as has been done recently in Bungert and Stinson (2022).

Total variation regularization with general TV functionals is an important methodology in imaging, and also in unsupervised and supervised learning on graphs, where it has been used for community detection, clustering, and graph trend-filtering; e.g., see Hu et al. (2013); Merkurjev et al. (2013); van Gennip et al. (2014); Bertozzi and Flenner (2016); Luo and Bertozzi (2017); Caroccia et al. (2020); Boyd et al. (2020); Cristofari et al. (2020); García Trillos et al. (2022); García Trillos and Murray (2017) and references therein.

4.3 Distributional-perturbing model and its generalized barycenter formulation

In this section we reintroduce some tools from chapter 3 and develop a collection of technical results that we will use in section 4.4 to prove Theorem 4.1.

Generalized barycenter and MOT problems

Given $\mu \in \mathcal{P}(\mathcal{Z})$, recall the *generalized barycenter problem*, (3.1.4), in chapter 3:

$$\inf_{\lambda, \tilde{\mu}_1, \dots, \tilde{\mu}_K} \left\{ \lambda(\mathcal{X}) + \sum_{i \in \mathcal{Y}} C(\mu_i, \tilde{\mu}_i) : \lambda \geqslant \tilde{\mu}_i \text{ for all } i \in \mathcal{Y} \right\}.$$

In the above, the infimum is taken over positive (Borel) measures $\tilde{\mu}_1, \dots, \tilde{\mu}_K$ and λ satisfying the constraints $\lambda \geqslant \tilde{\mu}_i$ for all $i \in \mathcal{Y}$. This constraint must be interpreted as: $\lambda(A) \geqslant \tilde{\mu}_i(A)$ for all $A \in \mathfrak{B}(\mathcal{X})$. Problem (3.1.4) can be understood as a generalization of the standard (Wasserstein) barycenter problem studied in Agueh and Carlier (2011b). Indeed, if all measures $\mu_1, \dots, \mu_K$ had the same total mass and the term $\lambda(\mathcal{X})$ in (3.1.4) was rescaled by a constant $\alpha \in (0, \infty)$, then, as $\alpha \to \infty$, the resulting problem would recover the classical barycenter problem with pairwise cost function c. As stated, one can regard (3.1.4) as a partial optimal transport barycenter problem: we transport each μ_i to a part of λ while requiring the transported masses to overlap as much as possible (this is enforced by asking for the term $\lambda(\mathcal{X})$ to be small).

One of the essential results of chapter 3 is that the generalized barycenter problem (3.1.4) is the dual of (2.5).

Theorem 4.10 (Proposition 3.4 and Corollary 3.29 in chapter 3)**.** *Suppose that* c *satisfies Assumption 2.5. Then*

$$(2.5) = 1 - (3.1.4).$$

Furthermore, the infimum of (3.1.4) *is attained. In other words, there exists* $(\lambda^*, \widetilde{\mu}^*)$ *which minimizes* (3.1.4).

Like classical barycenter problems, (3.1.4) has an equivalent multimarginal optimal transport (MOT) formulation which is already shown in subsection 3.19. To be precise, we recall a *stratified* multimarginal optimal transport problem to obtain an equivalent reformulation of (3.1.4).

Theorem 4.11 (Proposition 3.11 and 3.12)**.** *Suppose that* c *satisfies Assumption 2.5. Let* $S_K := \{A \subseteq \mathcal{Y} : A \neq \emptyset\}$. *Given* $A \in S_K$, *define* $c_A : \mathcal{X}^K \to [0, \infty]$ *as* $c_A(x_1, \dots, x_K) := \inf_{x' \in \mathcal{X}} \sum_{i \in A} c(x', x_i)$.

Let's recall (3.2.6)*:*

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

where $\mathcal{P}_i$ *is the projection map* $\mathcal{P}_i : (x_1, \dots, x_K) \mapsto x_i$, *and* $S_K(i) := \{A \in S_K \, : \, i \in A\}$. *Then* (3.1.4) = (3.2.6). *Also, the infimum in* (3.2.6) *is attained.*

Remark 4.12. *Even though* c_A *and* π_A *above are defined over* $\mathcal{X}^K$*, only the coordinates* i *where* $i \in A$ *actually play a role in the optimization problem. Also, notice that* (3.2.6) *is not a standard MOT problem since in* (3.2.6) *we optimize over several couplings* π_A (*each with its own cost function* c_A) *that are connected to each other via the marginal constraints. We refer to this type of problem as a stratified MOT problem.*

The following theorem is the recast regarding the duals of the generalized barycenter problem and its MOT formulation proven in subsection 3.2. Recall the notions of c-transform and $\bar{c}$-transform, (2.9) and (2.10), introduced in chapter 2, play an important role in these results: revisit in section 4.6 for more details.

Theorem 4.13 (Proposition 3.19 and Proposition 3.21)**.** *Suppose that* c *satisfies Assumption 2.5. Let* $\mathcal{C}_b(\mathcal{X})$ *be the set of bounded real-valued continuous functions over* $\mathcal{X}$.

The dual of (3.1.4) *is* (3.2.12)*:*

$$\sup_{f_1,\dots,f_K\in\mathcal{C}_b(\mathcal{X})} \sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} f_i^c(x_i)\,d\mu_i(x_i)$$
$$\text{s.t. } f_i(x) \geqslant 0,\ \sum_{i\in\mathcal{Y}} f_i(x) \leqslant 1,\ \text{for all } x\in\mathcal{X},\ i\in\{1,\dots,K\},$$

and there is no duality gap between primal and dual problems. In other words, (3.1.4) = (3.2.12)*. In the above,* f_i^c *denotes the* c*-transform of* f_i*,* (2.9)*, as introduced in chapter 2.*

The dual of (3.2.6) *is* (3.2.13)*:*

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

and there is no duality gap between primal and dual problems. In other words, (3.2.6) = (3.2.13)*.*

If in addition the cost function c *is bounded and Lipschitz, then* (3.2.13) *is achieved by* $g\in\mathcal{C}_b(\mathcal{X})^K$*. Also, for* f *feasible for* (3.2.12)*,* $g' := f^c$ *is feasible for* (3.2.13)*. Similarly, for* g *feasible for* (3.2.13)*,* $f' = \max\{g,0\}^{\overline{c}}$ *is feasible for* (3.2.12)*. Therefore, the optimization of* (3.2.13) *can be restricted to non-negative* g *satisfying* $g_i = g_i^{\overline{c}c}$*, or* $0\leqslant g_i\leqslant 1$ *for all* $i\in\mathcal{Y}$*.*

Remark 4.14. *Combining 4.10, 4.11, and 4.13 we conclude that* $1-$(3.2.13) $=$ (2.5)*.*

Remark 4.15. *A standard argument in optimal transport theory shows that problem* (3.2.13) *is equivalent to*

$$\sup_{g_1,\dots,g_K} \sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} g_i(x_i)\,d\mu_i(x_i), \tag{4.3.1}$$

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

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

for $\otimes_i \mu_i$*-almost every tuple* $(x_1, \dots, x_K)$*. Indeed, notice that since* (3.2.13) *has already been shown to be equal to* (3.2.6)*, the claim follows from the observation that any feasible* $g_1, \dots, g_K$ *for* (4.3.1) *satisfies the condition*

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

for every $\{\pi_A\}_{A \in S_K}$ *satisfying the constraints in* (3.2.6)*.*

Existence of optimal dual potentials g for general lower-semicontinuous costs

We already know from the last part in Theorem 4.13 that if c is bounded and Lipschitz, then there is a feasible $g \in \mathcal{C}_b(\mathcal{X})^K$ that is optimal for (3.2.13). In this subsection we prove an analogous existence result in the case of a general lower semi-continuous cost function c satisfying Assumption 2.5. More precisely, we prove existence of maximizers for (4.3.1). We start with an approximation result.

Lemma 4.16. *Let* c *be a cost function satisfying Assumption 2.5. For each* $n \in \mathbb{N}$ *let*

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

where

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

Then the following properties hold:

1. c_n *is bounded and Lipschitz.*
2. $c_n \leqslant c_{n+1} \leqslant c$ *and* $\tilde{c}_n \leqslant \tilde{c}_{n+1}$ *for all* $n \in \mathbb{N}$*.*

3. $\lim_{n\to\infty} c_n(x, x') = c(x, x')$ *for all* $(x, x') \in \mathcal{X} \times \mathcal{X}$.

Proof. Items 1. and 2. are straightforward to prove. To prove item 3., notice that due to the monotonicity of the cost functions we know that $\lim_{n\to\infty} c_n(x, x')$ exists in $[0, \infty]$ and $\lim_{n\to\infty} c_n(x, x') \leqslant c(x, x')$. If $\lim_{n\to\infty} c_n(x, x') = \infty$, then we would be done. Hence, we may assume that $\lim_{n\to\infty} c_n(x, x') < \infty$. From the definition of c_n it then holds that $\lim_{n\to\infty} \tilde{c}_n(x, x') = \lim_{n\to\infty} c_n(x, x') < \infty$. Let $(x_n, x'_n) \in \mathcal{X} \times \mathcal{X}$ be such that

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

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

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

from where the desired claim follows. □

Lemma 4.17. *Let* c *be a cost function satisfying Assumption 2.5, and let* c_n *be the cost function defined in Lemma 4.16. For each* $A \in S_K$*, let*

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

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

Proof. Fix $A \in S_K$ and $x_A := (x_i)_{i \in A} \in \mathcal{X}^{|A|}$. From Lemma 4.16 it follows $c_{A,n} \leqslant c_{A,n+1} \leqslant c_A$. Therefore, for a given x_A, $\lim_{n\to\infty} c_{A,n}(x_A)$ exists in $[0, \infty]$ and is less than or equal to $c_A(x_A)$. If the limit is ∞, we are done. We can then assume without the loss of generality that $\lim_{n\to\infty} c_{A,n}(x_A) < \infty$. We can then find sequences

$\{x_{n,i}\}_{n\in\mathbb{N}}$, $\{x'_{n,i}\}_{n\in\mathbb{N}}$, and $\{x'_n\}_{n\in\mathbb{N}}$ such that for all large enough $n \in \mathbb{N}$

$$\sum_{i\in A} c(x'_{n,i}, x_{n,i}) + n(\sum_{i\in A}(d(x_{n,i}, x_i) + d(x'_{n,i}, x'_n))) \leqslant c_{A,n}(x_A) + \frac{1}{n}.$$

From the above we derive that $\lim_{n\to\infty} d(x'_{n,i}, x'_n) = 0$ and $\lim_{n\to\infty} d(x_{n,i}, x_i) = 0$. Hence, it follows that $\limsup_{n\to\infty} c(x'_{n,i}, x_{n,i}) < \infty$. Combining the previous facts with Assumption 2.5, we conclude that $\{x'_n\}_{n\in\mathbb{N}}$ is precompact, and thus, up to subsequence (that we do not relabel), we have $\lim_{n\to\infty} d(x'_n, \hat{x}) = 0$ for some $\hat{x} \in \mathcal{X}$. Combining with $\lim_{n\to\infty} d(x'_{n,i}, x'_n) = 0$ we conclude that $\lim_{n\to\infty} d(x'_{n,i}, \hat{x}) = 0$ for all $i \in A$. Using the lower semi-continuity of c we conclude that

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

□

Proposition 4.18. *Let c be a cost function satisfying Assumption 2.5. Then there exists a solution for* (4.3.1).

Proof. Let $\{c_n\}_{n\in\mathbb{N}}$ be the sequence of cost functions introduced in Lemma 4.16. Notice that for each $n \in \mathbb{N}$ there is a solution $g^n = (g_1^n, \dots, g_K^n) \in \mathcal{C}_b(\mathcal{X})^K$ for the problem (3.2.13) (with cost c_n) that can be assumed to satisfy $0 \leqslant g_i^n \leqslant 1$ for each $i \in \mathcal{Y}$. Therefore, for each $i \in \mathcal{Y}$ the sequence $\{g_n^i\}_{n\in\mathbb{N}}$ is weakly* precompact in $L^\infty(\mathcal{X}; \mu_i)$. This implies that there exists a subsequence of $\{g^n\}_{n\in\mathbb{N}}$ (not relabeled) for which g^n weakly* converges toward some $g^* \in \prod_{i\in\mathcal{Y}} L^\infty(\mathcal{X}; \mathbb{R}, \mu_i)$, which would necessarily satisfy $0 \leqslant g_i^* \leqslant 1$ for all $i \in \mathcal{Y}$; see section 4.6 for the definition of weak* topologies. We claim that this g^* is feasible for (4.3.1). Indeed, by Lemma 4.17 we know that $c_{A,n} \leqslant c_A$ for all $A \in S_K$. In particular, since $c_{A,n} \leqslant c_A$, and $\sum_{i\in A} g_i^n(x_i) \leqslant 1 + c_{A,n} \leqslant 1 + c_A$ for all $A \in S_K$ and all $n \in \mathbb{N}$, it follows that $\sum_{i\in A} g_i^*(x_i) \leqslant 1 + c_A$, $\otimes_i \mu_i$-almost everywhere, due to the weak* convergence of g_i^n toward g_i^*. This verifies that g^* is indeed feasible for (4.3.1).

Let α_n and β_n be the optimal values of (3.1.4) and (3.2.13), respectively, for the cost c_n. Likewise, let α and β be the optimal values of (3.1.4) and (3.2.13),

respectively, for the cost c. Recall that, thanks to Theorem 4.11 and Theorem 4.13, we have $\alpha_n = \beta_n$ for all $n \in \mathbb{N}$ and $\alpha = \beta$. Suppose for a moment that we have already proved that $\lim_{n\to\infty} \alpha_n = \alpha$. Then we would have

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

which would imply that g^* is optimal for (4.3.1).

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

Let λ^n and $\tilde{\mu}_1^n, \dots, \tilde{\mu}_K^n$ be an optimal solution of (3.1.4) with the cost c_n and let π_i^n be a coupling realizing $C(\mu_i, \tilde{\mu}_i^n)$. We first claim that $\{\tilde{\mu}_i^n\}_{n\in\mathbb{N}}$ is weakly precompact for each $i \in \mathcal{Y}$. To see this, notice that for every n we have $\tilde{\mu}_i^n(\mathcal{X}) = \mu_i(\mathcal{X}) \leqslant 1$, for otherwise $C(\mu_i, \tilde{\mu}_i^n) = \infty$. Thus, by Prokhorov theorem it is enough to show that for every $\eta > 0$ there exists a compact set $\mathcal{K} \subseteq \mathcal{X}$ such that $\tilde{\mu}_i^n(\mathcal{X} \setminus \mathcal{K}) \leqslant C\eta$ for all $n \in \mathbb{N}$ and some C independent of n, η or $\mathcal{K}$. To see that this is true, let us start by considering a compact set G such that $\mu_i(G^c) \leqslant \eta$. Let $n_0 \in \mathbb{N}$ be such that $n_0 - 1 > \frac{1}{\eta}$. For $n \geqslant n_0$ we have

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

Consider the set

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

using the definition of c_{n_0} and Assumption 2.5 it is straightforward to show that $\tilde{\mathcal{K}}$ is a compact subset of $\mathcal{X}$. We see that $\alpha_\infty \geqslant \frac{1}{\eta}(\tilde{\mu}_i^n(\tilde{\mathcal{K}}^c) - \mu_i(G^c))$, from where we can conclude that $\tilde{\mu}_i^n(\tilde{\mathcal{K}}^c) \leqslant (\alpha_\infty + 1)\eta$ for all $n \geqslant n_0$. We now consider a compact set $\hat{\mathcal{K}}$ for which $\tilde{\mu}_i^n(\hat{\mathcal{K}}^c) \leqslant \eta$ for all $n = 1, \dots, n_0$, and set $\mathcal{K} := \tilde{\mathcal{K}} \cup \hat{\mathcal{K}}$, which is compact. Then for all $n \in \mathbb{N}$ we have $\tilde{\mu}_i^n(\mathcal{K}^c) \leqslant (\alpha_\infty + 1)\eta$. This proves the desired claim.

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

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

where $\overline{\mu}^n(x) = \sum_{i=1}^K \tilde{\mu}_i^n$. Indeed, notice that the above is the smallest positive measure greater than $\tilde{\mu}_1^n, \dots, \tilde{\mu}_K^n$. Given the form of λ^n and the weak precompactness of each of the sequences $\{\tilde{\mu}_i^n\}_{n\in\mathbb{N}}$, we can conclude that $\{\lambda^n\}_{n\in\mathbb{N}}$ is weakly precompact and so are the sequences $\{\pi_i^n\}_{n\in\mathbb{N}}$. We can thus assume that, up to subsequence, $\tilde{\mu}_i^n$ converges weakly toward some $\tilde{\mu}_i$; π_i^n converges weakly toward some $\pi_i \in \Gamma(\mu_i, \tilde{\mu}_i)$; and λ^n converges weakly toward some λ satisfying $\lambda \geqslant \tilde{\mu}_i$ for each $i \in \mathcal{Y}$. In particular, $\lambda, \tilde{\mu}_1, \dots, \tilde{\mu}_K$ is feasible for (3.1.4).

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

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

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

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

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

□

From dual potentials to robust classifiers for continuous cost functions

Having discussed the existence of solutions g^* for (4.3.1), we move on to discussing the connection between g^* and solutions f^* of problem (2.5).

Proposition 4.19 (Corollary 3.30 in chapter 3)**.** *Let $c : \mathcal{X} \times \mathcal{X} \to [0, \infty]$ be a lower-semicontinuous function and suppose that $(\widetilde{\mu}^*, g^*)$ is a solution pair for the generalized barycenter problem (3.1.4) and the dual of its MOT formulation (4.3.1). Let f^* be defined as*

$$f_i^*(\widetilde{x}) := \sup_{x \in \mathrm{spt}(\mu_i)} \{g_i^*(x) - c(x, \widetilde{x})\}, \tag{4.3.2}$$

for $i = 1, \dots, n$.

If f^* is Borel-measurable, then $(f^*, \widetilde{\mu}^*)$ is a saddle solution for the problem (2.5). In particular, f^* is a minimizer of (2.5).

The reason why we can not directly use Proposition 4.19 to prove existence of solutions to (2.5) for arbitrary c and μ is because it is a priori not guaranteed that f_i^*, as defined in (4.3.2), is Borel measurable; notice that the statement in Proposition 4.19 is conditional. If $\mathrm{spt}(\mu_i)$ was finite for all i, then the Borel measurability of f_i^* would follow immediately from the fact that the maximum of finitely many lower-semicontinuous functions is Borel; this is of course the case when working with empirical measures. Likewise, the Borel measurability of f_i^* is guaranteed when μ is arbitrary and c is a bounded Lipschitz function (in fact, it is sufficient for the cost to be continuous), as is discussed in Definitions 5.2 and 5.7 and Theorem 5.10 in Villani (2009). However, nothing can be said about the Borel measurability of f_i^* without further information on g_i^* (which in general is unavailable) when c is only assumed to be lower-semicontinuous (as is the case for the cost c_ε from (2.8)) and $\mathrm{spt}(\mu_i)$ is an uncountable set.

Our strategy to prove Theorem 4.1 in section 4.4 will be to approximate an arbitrary cost function c from below with a suitable sequence of bounded and Lipschitz cost functions c_n (the costs defined in Lemma 4.16), and, in turn, consider

a limit of the robust classifiers f_n^* associated to each of the c_n. This limit (lim sup, to be precise) will be our candidate solution for (2.5).

4.4 Proofs of our main results

In this section we prove the existence of a Borel measurable robust classifier for problem (2.5) when c is an arbitrary lower semi-continuous cost function satisfying Assumption 2.5. We also establish the existence of minimizers of (4.2.2) and establish Theorem 4.6 and Corollary 4.9.

Well-posedness of the DRO model

Proof of Theorem 4.1. Let $\{c_n\}_{n\in\mathbb{N}}$ be the sequence of cost functions converging to c from below defined in Lemma 4.16. For each $n \in \mathbb{N}$, we use Theorem 4.13 and let $g^n = (g_1^n, \dots, g_K^n) \in \mathcal{C}_b(\mathcal{X})^K$ be a solution of (3.2.13) with cost c_n; recall that we can assume that $0 \leqslant g_i^n \leqslant 1$. In turn, we use g^n and the cost c_n to define $f^n := (f_1^n, \dots, f_K^n)$ following (4.3.2). Since the g_i^n and c_n are continuous, and given that the pointwise supremum of a family of continuous functions is lower semi-continuous, we can conclude that f_i^n is lower semi-continuous and thus also Borel measurable for each $n \in \mathbb{N}$. Thanks to Proposition 4.19, f^n is optimal for (3.2.12) with cost function c_n.

From the proof of Proposition 4.18, we know that there exists a subsequence (that we do not relabel) such that the g_i^n converge in the weak* topology, as $n \to \infty$, toward limits g_i^* that form a solution for (4.3.1) with cost c. Using this subsequence, we define $f^* \in \mathcal{F}$ according to

$$f_i^*(\tilde{x}) := \limsup_{n\to\infty} f_i^n(\tilde{x}), \quad \tilde{x} \in \mathcal{X}. \tag{4.4.1}$$

Notice that each f_i^* is indeed Borel measurable since it is the lim sup of Borel measurable functions. In addition, notice that $0 \leqslant f_i^* \leqslant 1$, due to the fact that $0 \leqslant f_i^n \leqslant 1$ for all $i \in \mathcal{Y}$ and all $n \in \mathbb{N}$. We'll prove that f^* is a solution for (2.5).

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

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

Choosing $x' = x$ in the sup term (notice that indeed x can be assumed to belong to $\mathrm{spt}(\mu_i)$ since π_i has first marginal equal to μ_i), and applying reverse Fatou's lemma, we find that

$$\begin{aligned}
R(f^*, \widetilde{\mu}) - C(\mu, \widetilde{\mu}) &\leqslant 1 - \sum_{i \in \mathcal{Y}} \int_{\mathcal{X} \times \mathcal{X}} \limsup_{n \to \infty} \{ g_i^n(x) - c_n(x, \widetilde{x}) + c_n(x, \widetilde{x}) \} \, d\pi_i(x, \widetilde{x}) \\
&= 1 - \sum_{i \in \mathcal{Y}} \int_{\mathcal{X} \times \mathcal{X}} \limsup_{n \to \infty} \{ g_i^n(x) \} \, d\pi_i(x, \widetilde{x}) \\
&= 1 - \sum_{i \in \mathcal{Y}} \int_{\mathcal{X}} \limsup_{n \to \infty} \{ g_i^n(x) \} \, d\mu_i(x) \\
&\leqslant 1 - \limsup_{n \to \infty} \sum_{i \in \mathcal{Y}} \int_{\mathcal{X}} g_i^n(x) \, d\mu_i(x) \\
&= 1 - \sum_{i \in \mathcal{Y}} \int_{\mathcal{X} \times \mathcal{X}} g_i^*(x) \, d\mu_i(x) \\
&= 1 - (3.2.13) \\
&= R^*_{\mathrm{DRO}},
\end{aligned}$$

where the third equality follows from the weak* convergence of g_i^n toward g_i^*, the second to last equality follows from the fact that g^* is a solution for (4.3.1) (combined with 4.15), and the last equality follows from remark 4.14. Taking the

sup over $\tilde{\mu} \in \mathcal{P}(\mathcal{Z})$, we conclude that

$$\sup_{\tilde{\mu}\in\mathcal{P}(\mathcal{Z})} \{R(f^*, \tilde{\mu}) - C(\mu, \tilde{\mu})\} \leqslant R^*_{DRO},$$

and thus f^* is indeed a minimizer of (2.5).

Let now $\tilde{\mu}^*$ be a solution of (3.1.4) (which exists due to Theorem 4.10). The fact that $(\tilde{\mu}^*, f^*)$ is a saddle for (2.5) follows from the above computations and the fact that by Theorem 4.10 and Corollary 3.29 we have

$$R^*_{DRO} = \sup_{\tilde{\mu}\in\mathcal{P}(\mathcal{Z})} \inf_{f\in\mathcal{F}} \{R(f, \tilde{\mu}) - C(\mu, \tilde{\mu})\} = \inf_{f\in\mathcal{F}} \{R(f, \tilde{\mu}^*) - C(\mu, \tilde{\mu}^*)\}.$$

□

The next proposition states that the function g_i^* constructed in the proof of Proposition 4.18 is a Borel measurable version of the c-transform of f_i^*, where f_i^* was defined in (4.4.1).

Proposition 4.20. *Let $\{g^n\}_{n\in\mathbb{N}}$ and $\{f_n\}_{n\in\mathbb{N}}$ be as in the proof of 4.18, let g^* be the weak* limit of the g_n, and let f^* be as defined in* (4.4.1)*. Then, for every $i \in \mathcal{Y}$,*

$$g_i^*(x) = \inf_{\tilde{x}\in\mathcal{X}} \{f_i^*(\tilde{x}) + c(x, \tilde{x})\} \tag{4.4.2}$$

for μ_i-a.e. $x \in \mathcal{X}$. This statement must be interpreted as: the set in which (4.4.2) *is violated is contained in a Borel measurable set with zero μ_i measure.*

Proof. From the proof of Theorem 4.1 it holds that for each $i \in \mathcal{Y}$

$$\int_{\mathcal{X}} f_i^*(\tilde{x}) d\tilde{\mu}_i^*(\tilde{x}) + \int_{\mathcal{X}\times\mathcal{X}} c(x, \tilde{x}) d\pi_i^*(x, \tilde{x}) = \int_{\mathcal{X}} g_i^*(x) d\mu_i(x). \tag{4.4.3}$$

On the other hand, from the definition of f_i^n it follows that

$$g_i^n(x) \leqslant f_i^n(\tilde{x}) + c_n(x, \tilde{x}), \quad \mathcal{F}or all \tilde{x} \in \mathcal{X}, \text{ and } \mu_i\text{-a.e. } x \in \mathcal{X}.$$

We can then combine the above with Lemma 4.23 to conclude that for μ_i-a.e. $x \in \mathcal{X}$ and every $\tilde{x} \in \mathcal{X}$ we have

$$g_i^*(x) \leqslant \limsup_{n\to\infty} g_i^n(x) \leqslant \limsup_{n\to\infty} f_i^n(\tilde{x}) + c_n(x,\tilde{x}) = f_i^*(\tilde{x}) + c(x,\tilde{x}).$$

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

$$g_i^*(x) \leqslant \inf_{\tilde{x}\in\mathcal{X}} \{f_i^*(\tilde{x}) + c(x,\tilde{x})\}. \tag{4.4.4}$$

From this and (4.4.3) we see that $g_i^* \in L^1(\mu_i)$ and $-f_i^* \in L^1(\tilde{\mu}_i)$ are optimal dual potentials for the optimal transport problem $C(\mu_i, \tilde{\mu}_i^*)$. If (4.4.4) did not hold with equality for μ_i-a.e. $x \in \mathcal{X}$, then we would be able to construct a Borel-measurable version h_i of the right hand side of (4.4.4) (see Lemma 4.27) which would be strictly greater than g_i^* in a set of positive μ_i-measure. In addition, we would have that $(h_i, -f_i^*)$ is a feasible dual pair for the OT problem $C(\mu_i, \tilde{\mu}_i)$. However, the above would contradict the optimality of the dual potentials $(g_i^*, -f_i^*)$. We thus conclude that (4.4.4) holds with equality except on a set contained in a set of μ_i measure zero. □

Well-posedness of the closed-ball model (4.2.2)

Proof of Proposition 4.5. We actually prove that for arbitrary cost c satisfying Assumption 2.5, the solution f^* to (2.5) constructed in the proof of Theorem 4.1 is also a solution for the problem:

$$\inf_{f\in\mathcal{F}} \left\{ \sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} \sup_{\widetilde{x}\in\mathcal{X}} \{1 - f_i(\widetilde{x}) - c(x,\widetilde{x})\} \, d\overline{\mu}_i(x) \right\}. \tag{4.4.5}$$

4.5 will then be an immediate consequence of this more general result when applied to $c = c_\varepsilon$.

Let f^* be the Borel solution of (2.5) constructed in the proof of Theorem 4.1. It

suffices to show that for any $f \in \mathcal{F}$

$$\sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} \inf_{\widetilde{x}\in\mathcal{X}} \{f_i^*(\widetilde{x}) + c(x,\widetilde{x})\}\, d\overline{\mu}_i(x) \geqslant \sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} \inf_{\widetilde{x}\in\mathcal{X}} \{f_i(\widetilde{x}) + c(x,\widetilde{x})\}\, d\overline{\mu}_i(x).$$

Suppose not. Then there exists some $\widehat{f} \in \mathcal{F}$ which provides a strict inequality in the opposite direction. Now, on one hand, (4.4.2) of Proposition 4.20 implies

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

On the other hand, by Lemma 4.27, for each $i \in \mathcal{Y}$ there exists a Borel measurable function $\widehat{g}_i$ equal to $\inf_{\widetilde{x}\in\mathcal{X}}\{\widehat{f}_i(\widetilde{x}) + c(x,\widetilde{x})\}$ $\overline{\mu}_i$-almost everywhere. Let $\widehat{g} := (\widehat{g}_1, \ldots, \widehat{g}_K)$. Combining the existence of such $\widehat{g}$ with the above equation, and using (4.6.2), it follows that $\widehat{g}$ satisfies

$$\sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} g_i^*(x) d\mu_i(x) < \sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} \widehat{g}_i(x) d\mu_i(x). \tag{4.4.6}$$

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

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

From the above we conclude that $\widehat{g}$ is feasible for (4.3.1). However, this and (4.4.6) combined contradict the fact that g^* is optimal for (4.3.1), as had been shown in Proposition 4.18. □

Proof of Corollary 4.9. It is straightforward to verify (e.g., see Bungert et al. (2023)) that for $(f_1, 1-f_1) \in \mathcal{F}$ we can write

$$\overline{R}_\varepsilon((f_1, 1-f_1)) = \int_0^1 \overline{R}_\varepsilon((\mathbb{1}_{\{f_1\geqslant t\}}, \mathbb{1}_{\{f_1\geqslant t\}^c}))dt. \tag{4.4.7}$$

It is also straightforward to see that

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

Let $(f_1, 1 - f_1)$ be a solution to (4.2.2) (which by remark 2.1 can indeed be taken of this form). It follows from (4.4.7) that for almost every $t \in [0, 1]$ the pair $(\mathbb{1}_{\{f_1 \geqslant t\}}, \mathbb{1}_{\{f_1 \geqslant t\}^c})$ is also a solution for that same problem and thus also for the problem restricted to hard-classifiers. This proves the desired result. □

Connection between closed-ball model and open-ball model

Proof of Theorem 4.6. One can easily observe that for any fixed $\varepsilon > 0$ and $\delta > 0$ we have

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

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

Now, let f^* be solution of (4.2.2) and assume we have $R^o_\varepsilon = \overline{R}_\varepsilon$. Then

$$R^o_\varepsilon(f^*) \leqslant \overline{R}_\varepsilon(f^*) = \overline{R}_\varepsilon = R^o_\varepsilon,$$

which means f^* is a solution of 2.3. □

4.5 Conclusion

Through this chapter, we prove the existence of optimal Borel measurable robust classifier for three models of the problem of adversarial training in multiclass classification and provide not only a unifying framework of three models but also

a connection from optimal transport to total variation regularization. This is the first result of the existence of optimal Borel measurable robust classifiers.

4.6 Technical results

In this section, we state some technical results used in chapter 4.

Weak* topology

Definition 4.21 (Weak* topology)**.** *Let* $\mu = (\mu_1, \dots, \mu_K) \in \prod_{i=1}^{K} \mathcal{M}_+(\mathcal{X})$*. We say that a sequence* $\{h^n\}_{n\in\mathbb{N}} \subseteq \prod_{i\in\mathcal{Y}} L^\infty(\mathcal{X};\mu_i)$ *weak*-converges to* $h \in \prod_{i\in\mathcal{Y}} L^\infty(\mathcal{X};\mu_i)$ *if for any* $q \in \prod_{i\in\mathcal{Y}} L^1(\mathcal{X};\mu_i)$*, it holds that*

$$\lim_{n\to\infty} \int_{\mathcal{X}} h_i^n(x) q_i(x) d\mu_i(x) = \int_{\mathcal{X}} h_i(x) q_i(x) d\mu_i(x) \tag{4.6.1}$$

for all $i \in \mathcal{Y}$*.*

Remark 4.22. *Note that for a Borel positive measure* ρ *which is either finite or* σ*-finite over a Polish space, the dual of* $L^1(\rho)$ *is* $L^\infty(\rho)$*, which justifies the definition* (4.6.1)*.*

Lemma 4.23. *Suppose* $\{g_i^n\}_{n\in\mathbb{N}}$ *is a sequence of measurable real-valued functions over* $\mathcal{X}$ *satisfying* $0 \leqslant g_i^n \leqslant 1$ *for every* $n \in \mathbb{N}$*. Suppose that* g_i^n *converges in the weak* topology of* $L^\infty(\mathcal{X};\mu_i)$ *toward* g_i*, where* μ_i *is a finite positive measure. Then, for* μ_i*-a.e.* $x \in X$*, we have*

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

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

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

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

c-transform

c-transform has an important role in optimal transport theory. One can characterize an optimizer of a dual problem by iterating c-transform: see Villani (2003, 2009) for more details.

Definition 4.24 (c-transform in Villani (2009))**.** *Let $\mathcal{X}, \mathcal{X}'$ be measurable spaces, and let $c : \mathcal{X} \times \mathcal{X}' \to (-\infty, \infty]$. Given a measurable function $h : \mathcal{X} \to \mathbb{R} \cup \{\infty, -\infty\}$, its c-transform is defined as*

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

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

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

Proposition 4.25. *For any mearurable functions h over $\mathcal{X}$ and g over $\mathcal{X}'$, and cost function $c : \mathcal{X} \times \mathcal{X}' \to (-\infty, \infty]$, it holds that for every $(x, x') \in \mathcal{X} \times \mathcal{X}'$,*

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

Theorem 4.26 (Theorem 5.10 in Villani (2009))**.** *Let $\mathcal{X}$ be a Polish space and $c(\cdot, \cdot)$ be a cost function bounded from below and lower semi-continuous. Then, for $\nu, \widetilde{\nu} \in \mathcal{P}(\mathcal{X})$,*

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

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

Decomposition of universally measurable functions

Lemma 4.27. *Let $\mathcal{X}$ be a Polish space, μ and let $\overline{\mu}$ be a Borel probability measure and its extension to the universal σ-algebra, respectively. Let f be a universally measurable function for which $\int_{\mathcal{X}} |f(x)| d\overline{\mu}(x) < \infty$. Then there exists a Borel measurable function g such that $f = g$ $\overline{\mu}$-almost everywhere. Also,*

$$\int_{\mathcal{X}} f(x) d\overline{\mu}(x) = \int_{\mathcal{X}} g(x) d\mu(x). \tag{4.6.2}$$

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

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

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

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

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

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

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

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

$$\int_X f(x)d\overline{\mu}(x) = \int_X g(x)d\overline{\mu}(x) = \int_X g(x)d\mu(x),$$

from which (4.6.2) follows. □

5 TWO APPROACHES FOR COMPUTING ADVERSARIAL TRAINING PROBLEM BASED ON OPTIMAL TRANSPORT FRAMEWORKS

This chapter is based on García Trillos et al. (2023b) which is a joint work with Nicolas García Trillos, Matt Jacobs and Matt Werenski.

Outline

The rest of this chapter is organized as follows. In section 5.1, we will present two main algorithms and their informal explanations. In section 5.2, some empirical results based on proposed algorithms and their interpretation will be offered. We will finish off this chapter in section 5.3, where the summary of this chapter will be concerned. All technical details and delayed proofs will be discussed in section 5.4.

5.1 Main results

The main contribution of this chapter is to suggest two new algorithms which solves (2.5), the adversarial training problem in multiclass classification, using two equivalent problems developed in chapter 3. The first algorithm is called *exact solving* algorithm inspired by (3.1.4). The geometric intuition of (3.1.4) is to find $\widetilde{\mu}_i$'s which can be stacked over each other as much as possible, equivalently collecting higher-order interactions, as many as points from different classes, in the cheapest way. After collecting all possible such interactions, solving (3.1.4) reduces to a certain linear programming. Pseudocode appears in Algorithm 3.

Algorithm 3 Exact solving

Input: X : data set, $\mu = (\mu_1, \ldots, \mu_K)$: empirical distribution, ε : adversarial budget.
Use Algorithm 5 to construct $C(\varepsilon)$.
Construct the ε-incidence matrix $I_\varepsilon \in \{0,1\}^{X \times C(\varepsilon)}$.
Solve (5.4.1).

Output: $g^* = \sum_{C \in C(\varepsilon)} w^*(C)\delta_{F(C)}$, $\widetilde{\mu}_i = \sum_{A \in S_K(i)} \sum_{C \in C_A(\varepsilon)} w^*(C)\delta_{F(C)}$ and value $= \sum_{C \in C(\varepsilon)} w^*(C)$.

Algorithm 3 contains two main steps: use Algorithm 5 to construct $C(\varepsilon)$, the set of all possible interactions, and solve an appropriate linear programming with a matrix I_ε. Figure 5.1 below describes Algorithm 5 pictorially. It shows the procedure demonstrating the optimistic (and probably realistic) setting that there are several lower-order interactions and the number of higher-order interactions quickly decays. However, it may be possible in some settings that most, or all of the lower-order interactions can be merged together so that the number of higher-order interactions is comparable to that of lower-order ones. The details about this algorithm will be discussed in section 5.4.

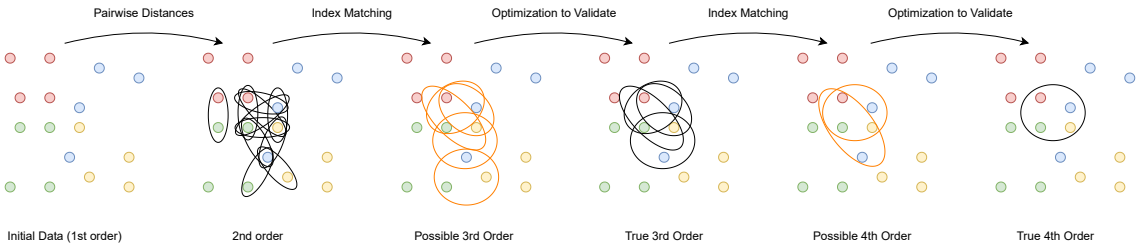

Figure 5.1: Steps to compute the labeled interactions.

The first main theorem is about the worst complexity of Algorithm 3. We want to emphasize that this is *worst case* scenario: one could expect much smaller complexity of Algorithm 3 in many real data sets. Since the size of I_ε notably depends on the size of $C(\varepsilon)$, therefore, the complexity of Algorithm 3 is affected by the size of $C(\varepsilon)$. The size of $C(\varepsilon)$ is determined by the possibility of interactions,

or, the geometry(separability) of classes. One can expect that if classes are far from each other and given adversarial budget is not too big, it is very likely that higher-order interactions vanish, in other words, I_ε *is sparse*. We expect that in most reasonable situations, I_ε is super spares. Under this prediction, we expect that Algorithm 5 stops earlier whence Algorithm 3 terminates faster than its worst case complexity.

Theorem 5.1 (Complexity of Algorithm 3)**.** *Assume that* $n_i = O(n)$ *for all* $i \in \mathcal{Y}$*. Algorithm 3 gives a* δ*-approximation of* (2.5) *in* $O(K^2 n^K \left(\log \frac{1}{\delta}\right)^2)$ *time.*

Remark 5.2. *Algorithm 3 has also a truncated version: namely, when constructing* $C(\varepsilon)$*, one can ignore all the higher-order interactions and restrict* $C(\varepsilon)$ *to at most* L*-order interactions. Then, the resulted* ε*-incidence matrix becomes much smaller, hence the complexity decreases. We will also discuss truncation in subsection 5.4.*

The second approach is the *truncated entropic regularization*: it is also known as *truncated Sinkhorn iteration* in computational optimal transport community. Sinkhorn iteration, firstly proposed by Cuturi (2013), has been extensively used and studied in various fields, especially in machine learning community, due to its faster implementation and empirical successes. This approach is based on the formula (3.2.6).

Observe that c_A's of (3.2.6) is indeed the cost for the adversary to pay for the consideration of $|A|$-order interaction associated with A classes: x_A is a local plan of forming an interaction, and π_A's are a global (transporting) plan of $|A|$-order interactions consisting of subclass A. Thanks to this special structure, we can explicitly truncate this problem by only focusing on π_A's whose $|A|$ is at most L. It means that we explicitly ignore all higher-order interactions. Of course, this truncated Sinkhorn not only reduces computational costs but also still have a chance to approximate o a solution well provided that the well-separability of data sets and the reasonable adversarial budget. This is the same intuition that we have in the previous algorithm.

Let $S_K^L := \{A \in S_K : |A| \leqslant L\}$ and $S_K^L(i) := \{A \in S_K^L : i \in A\}$. Pseudocode appears in Algorithm 4.

Algorithm 4 Truncated entropic regularization(Sinkhorn)

Input: X : data set, $\eta > 0$: entropic parameter, L : trucation level, $\mu = (\mu_1, \dots, \mu_K)$: empirical distribution, ε : adversarial budget.

Initialization : $g_i = \eta \log \mu_i$ for all $i \in \mathcal{Y}$.

while not converge **do**

$$g_1(\cdot) \leftarrow \eta \log \mu_1(\cdot) - \eta \log \left(\sum_{A \in S_K^L(1)} G(g_{A \setminus \{1\}})(\cdot) \right),$$

$$\vdots$$

$$g_K(\cdot) \leftarrow \eta \log \mu_K(\cdot) - \eta \log \left(\sum_{A \in S_K^L(K)} G(g_{A \setminus \{K\}})(\cdot) \right).$$

end while

Compute $\pi_A^*(x_A) = \exp\left(\frac{1}{\eta}\left(\sum_{i \in A} g_i^*(x^i_{l_i})\right)\right) \exp\left(-\frac{1}{\eta}(1 + c_A(x_A))\right)$ for all $A \in S_K^L$.

Output: $\{\pi_A^*\}_{A \in S_K^L}$ and value $= \sum_{A \in S_K^L} \sum_{\mathcal{X}^A} (1 + c_A(x_A)) \pi_A^*(x_A)$.

Algorithm 4 is basically a block coordinate ascent method which is the nature of Sinkhorn-type algorithms. Since (5.4.8) is strictly convex, a solution is unique.

We want to emphasize one strong advantage of Algorithm 4 compared to Algorithm 3: it proposes not only optimal couplings but also robust classifiers. If everything is well conditioned, for example well-separability of classes, it would be reasonable to expect that Algorithm 3 works faster than Algorithm 4 and its error is also smaller(or even vanishes). However, since Algorithm 3 solves (3.1.4), it does not give any information about (3.2.12), the problem for the learner to obtain robust classifiers. In other words, it does not care about anything for the learner's perspective. On the other hand, however, since Algorithm 4 naturally relies on the duality (with the entropic term), (3.2.13), it automatically solves both (3.1.4) and (3.2.13) simultaneously, hence solves (3.1.4) and (3.2.12) simultaneously. Therefore, Algorithm 4 implements an approximation of saddle point of (2.5). In this sense, it (approximately) solves the problem in two sides perfectly, while Algorithm 3 does only in one side. We state this as the following theorem.

Theorem 5.3. *Let* g^* *be a fixed point of Algorithm 4. Let* $f_i^* := \max\{g_i^*, 0\}^{\bar{c}}$ *for each* $i \in \mathcal{Y}$. *Then,* f^* *is an approximation for* (3.2.12). *Therefore,* f^* *is an approximation of an optimal robust classifier.*

Remark 5.4. *Notice that Algorithm 4 in general does not provide feasible* $\{\pi^*_A\}_{A \in S^L_K}$ *for* (5.4.8) *in a finite time. However, this violation becomes moderate as iterating Algorithm 4 more and converges to* 0 *in the limit.*

The authors of Altschuler et al. (2017) *provide the rounding scheme which enables an output within a finite time to be feasible for* (5.4.8)*. This rounding scheme is recently used in MOT Sinkhorn setting Lin et al.* (2022) *to achieve the complexity of general MOT Sinkhorn. Here, we do not adopt their rounding scheme but only apply the vanilla Sinkhorn iteration. We leave it to the future work.*

Remark 5.5. *In practice, one may face one numerical issue of Algorithm 4 which is that* c_A *can be* $+\infty$*. This fact hampers at least numerically the success of this algorithm. If one thinks that it is a really issue, Algorithm 6 is an alternative one. This algorithm has pros and cons: it can avoid the infinite cost function but should require a slightly large computational cost to recover the original coupling* π^*_A*. In fact, Algorithm 4 and Algorithm 6 solve two different regularized problems. However, for small* η*, both problems are close enough and give two approximated solutions for* (5.4.8)*: see subsection 5.4 for more details.*

Our second main theorem is about the convergence rate of Algorithm 4. With the bounded cost function, it is known Carlier (2022) that the MOT Sinkhorn iteration has the linaer convergence rate. A similar result holds in our setting because, although c_A is not bounded, it behaves like a hard constraint so that cost function on $c_A = 0$ is bounded nicely. More technical details will be discussed in section 5.4.

Theorem 5.6 (Informal)**.** *Algorithm 4 has the linear convergence rate.*

5.2 Empirical results

In this section, we present empirical results: see figure 5.2 and figure 5.3. We apply our algorithms, Exact solving and Truncated entropic regularization, to famous MNIST and CIFAR10 data sets. Also, we put the result computed by the MOT Sinkhorn proposed by Lin et al. (2022) based on another equivalent

formulation obtained in chapter 3 which plays a role of state-of-the-art algorithm for approximating MOT problems.

In both MNIST and CIFAR-10 data sets, we consider 4 classes and its possible truncations, 2 and 3-levels. As we expect, unless the adversarial budget ε is large, truncation is a good approximation for the full consideration. Both figure 5.2 and figure 5.3 show our heuristic: if the adversarial budget is not too large and if classes are well-separated, the truncated problem is a good approximation for the original problem. Here, we see the upper bound of the adversarial budget at which each level of truncation works nicely.

In particular, looking at the plots of Algorithm 3(bottom two plots for both MNIST and CIFAR-10), one can completely characterize when 3-rd order and 4-th order interactions become effective: for example, regarding MNIST plots, with ℓ^2 distance, 3-rd order and4-th order interactions occur after $\varepsilon \geqslant 1200$ and $\varepsilon \geqslant 1400$, respectively. In fact the number 1200 is roughly the half of the maximum distance between two different classes and it makes sense high-order interactions arise after this number.

In this sense, one can reversely estimate the geometry of MNIST and CIFAR-10 data sets from those plots. As everyone knows that the dimension of both data sets is quite huge, 28×28 for MNIST and 32×32 for CIFAR 10. For this high-dimensionality, it is hard to understand their geometry directly and most of their aspects are quite counter intuitive. But, interestingly, one can get geometric information from these adversarial risks because the adversarial risk highly depends on the geometry of data sets(and the adversarial budget). For example, in terms of ℓ^∞ metric, MNIST data sets are super separeted in a way that each class almost is concentrated on each vertex of ℓ^∞ hypercube. It is reflected in the plots: until less than 120 adversarial budget, no iteraction happens and then until 160, no 3-rd order interaction happens an so on and forth. We have a similar observation in CIFAR-10 data, but it is not separated as much as MNIST: the adversarial risk gradually increases from 80 to 110 and after 110, 3-rd order interaction seems to exist. Therefore, one may conclude that *MNIST is more separated than CIFAR*-10. We expect such estimate can give more rigorous and pedagogical information about inaccessible geometry of many

high-dimensional data sets.

Notice that, unless LP and its truncated versions, there are some gaps between the full MOT Sinkhorn and its truncated versions. This is because of the entropic parameter η. The more decreasing η, the less such gaps while one should pay more computing costs. This gap is not huge, and interestingly, in ℓ^∞ metric, the gap is even smaller than in ℓ^2 metric. We don't know why such asymmetry happens and probably, there are hidden geometric reasons to induce such difference.

Lastly, we give a comment about the complexity of the MOT Sinkhorn. In our case, its complexity is

$$O\left(\frac{K^5 n^K \log n}{\delta^2}\right).$$

The proof will be presented in subsection 5.4. To achieve the correct complexity, we will give a slightly simpler single MOT formulation of (2.5), a small improvement of Theorem 3.3 in chapter 3: see Proposition 5.4.32 and Corollary 5.23.

5.3 Conclusion

In this chapter, we propose new algorithms to solve the adversarial training problem, exact solving(LP), its trucated version and the truncated entropic regularization(Sinkhorn). Each of them is built on a different equivalent optimal transport type problem which is developed in chapter 3: the generalized barycenter problem, (3.1.4), and the stratified MOT problem, (3.2.13).

The novelty of these algorithms is that we do not need to perform the full n^K-order computation when data is well-separated and the adversarial budget is not too big. If these two conditions are satisfied, it is very unlikely that higher-order interactions have meaningful contributions of the adversarial risk so that, one can safely ignore them, as a result, the computational complexity drops significantly.

Also, these two approaches have their own pros and cons. For exact solving, thanks to the sparsity of I_ε, in spite of its huge size, it performs quite reasonably. Furthermore, it works faster than its worst case scenario because there are not many high-order interactions in real data sets with a reasonable adversarial budget.

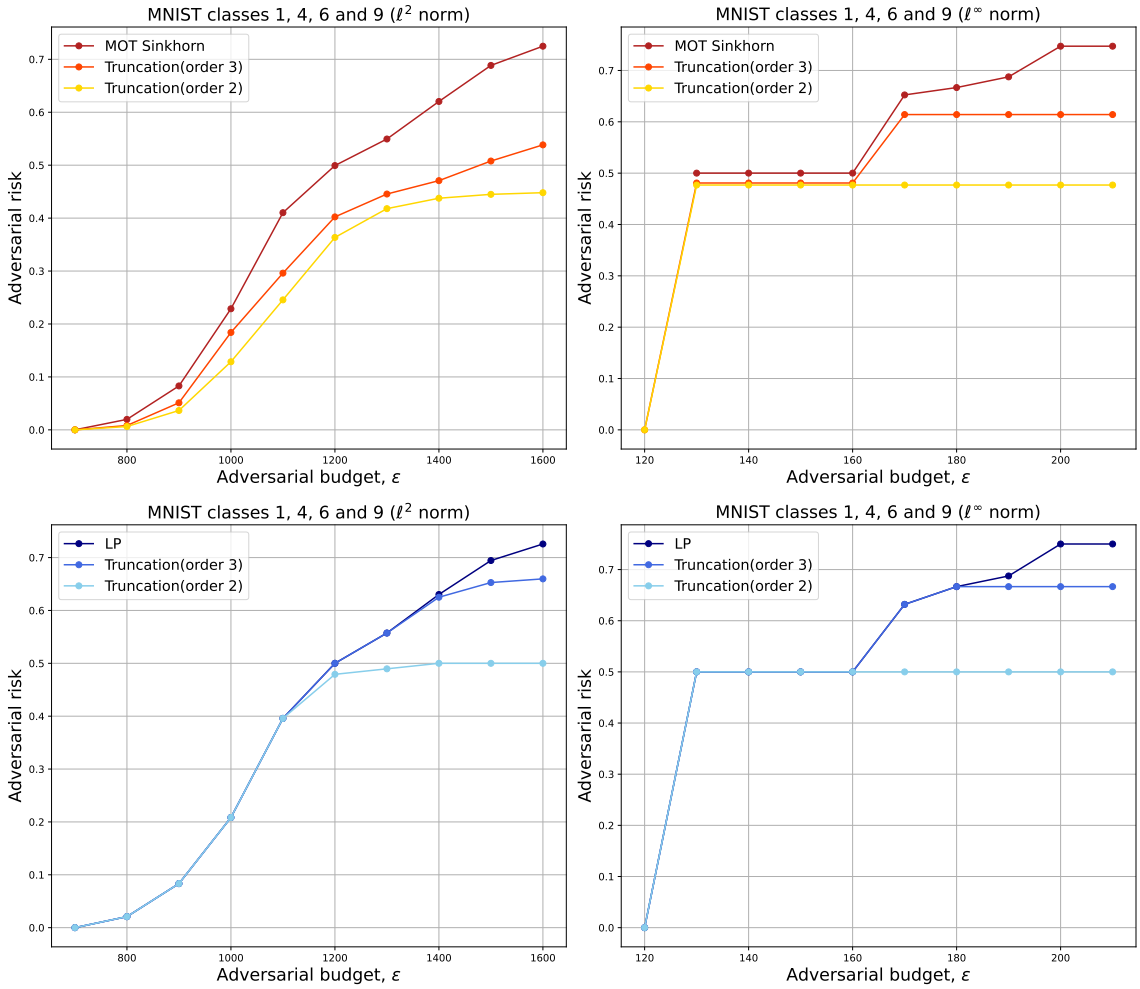

Figure 5.2: MNIST comparison.

For the truncated entropic regularization, there is an intrinsic error due to the entropic parameter η. Also, its speed is not faster than exact solving because after fixing the truncation level L, it requires computing n^{L-1} tensor in any case. If data is well-separated so that the number of high-order interaction decays fast enough, Exact solving will not compute such large(but moderate than n^K) tensor.

However, there is an advantage of the truncated entropic regularization. Due

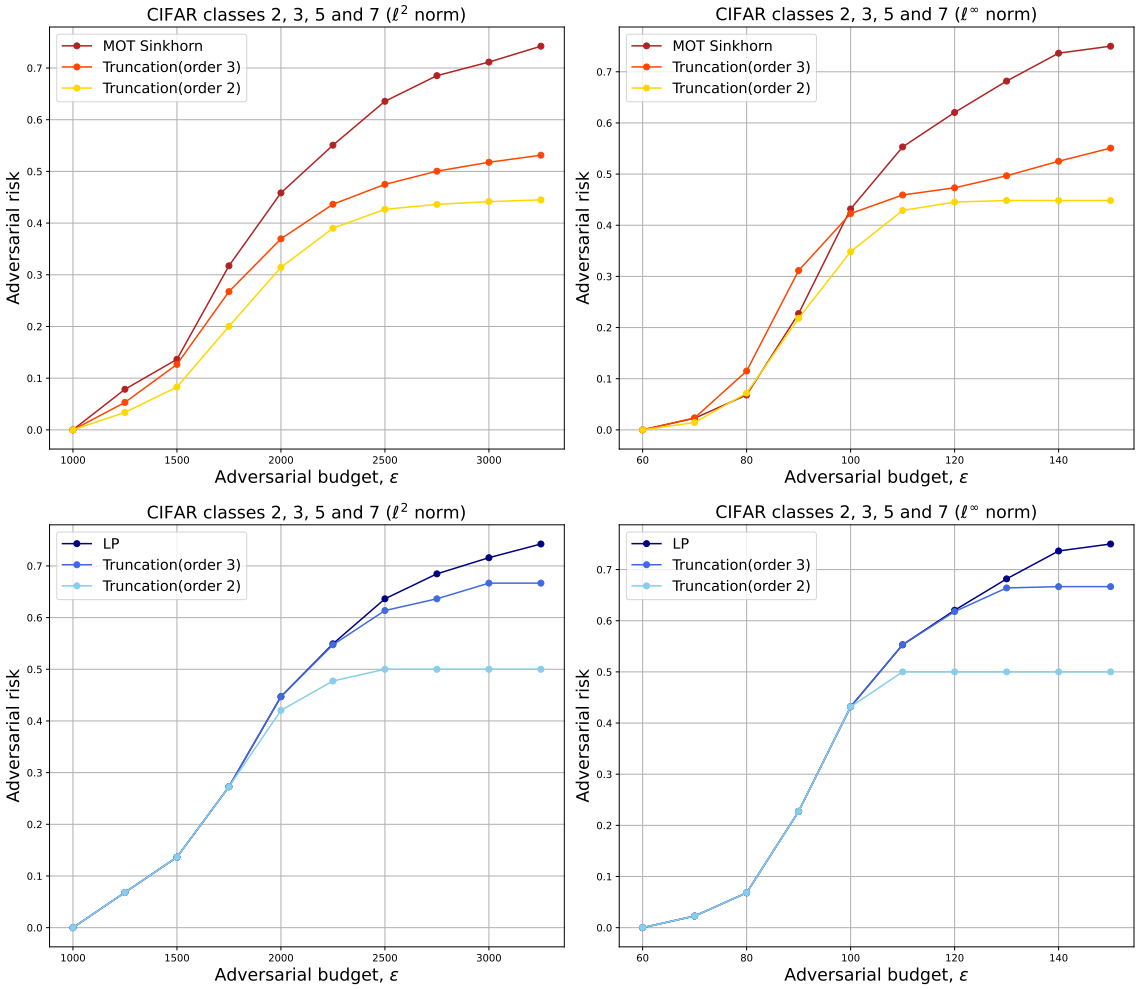

Figure 5.3: CIFAR10 comparison.

to its nature, computing the dual problem, it not only approximates (3.2.6) but also its dual (3.2.13). Hence, thanks to Proposition 3.19, we can obtain not only an optimal adversarial attack but also an optimal robust classifier at the same time.

5.4 Analysis of Algorithms 3 and 4

In this section, we will provide all the details of our proposed algorithms. Before doing that, we introduce some notations that we will use through this section.

- Let $\mathcal{X}_i = \{x^i_{l_i} \in \mathbb{R}^N : l_i = 1, \dots, n_i\}$ be the set of points for the label i where n_i is the total number of points of that label and $\mathcal{X} = \bigcup_{i=1}^K \mathcal{X}_i$ be the set of all points (with replacement if a point occurs for multiple labels). For each $A \in S_K$, we use

$$\mathcal{X}^A := \prod_{i \in A} \mathcal{X}_i, \quad x_A := (x^i_{l_i} : i \in A) \in \mathcal{X}^A.$$

- Let $\mu_i = \sum_{l_i=1}^{n_i} \mu_i(x^i_{l_i}) \delta_{x^i_{l_i}}$ be the (positive) empirical measure for the label i.

- For $k = 2, \dots, K$, a kth order labeled complex will be tracked by k (label, index)-pairs $\{(i_1, l_{i_1}), \dots, (i_k, l_{i_k})\}$ where each $i_p \neq i_q$ for $p \neq q$ and the index l_{i_p} corresponds to the $x^{i_p}_{l_{i_p}}$, the l_{i_p}-th point for the label i_p.

- Fix $A \in S_K$ and $\varepsilon > 0$. We use C_A and $C_A(\varepsilon)$ to denote the set of all labeled interactions and (A, ε)-feasible interactions, respectively,

$$\begin{aligned} C_A &:= \{\{(i, l_i)\}_{i \in A}\}, \\ C_A(\varepsilon) &:= \left\{\{(i, l_i)\}_{i \in A} \in C_A : \{(x^i_{l_i})\}_{i \in A} \subset \overline{B}(x, \varepsilon) \text{ for some } x\right\}. \end{aligned}$$

- The set of all ε-feasible interactions will be denoted by

$$C(\varepsilon) := \bigcup_{A \in S_K} C_A(\varepsilon).$$

  Given $C(\varepsilon)$, the ε-incidence matrix $I_\varepsilon \in \{0,1\}^{\mathcal{X} \times C(\varepsilon)}$ is defined as

$$I_\varepsilon((i, x^i_{l_i}), C) := \begin{cases} 1 & \text{if } (i, l_i) \in C, \\ 0 & \text{otherwise.} \end{cases}$$

Assuming that you can keep the data well-organized, the main bottleneck will be in one of two places

1. Simply enumerating all of the pairs of interactions which have $(k-2)$ matching labels.

2. Performing the check to see if the interactions can be fused.

Exact solving: Linear programming

Executing Algorithm 3 requires three steps. First, constructing $C(\varepsilon)$, the set of all ways of merging points given a budget ε. Second, convert $C(\varepsilon)$ into a sparse matrix I_ε. Third, use the existing libraries to solve a linear program with constraint matrix I_ε. We detail each of the steps in the following subsections.

Remark 5.7. *At the same time the authors of Dai et al. (2023) also proposed the same idea using the notion conflict hypergraph independently. In this paper, we elaborate why this algorithm succedes in practice based on the generalized barycenter problem which was introduced in Garcıa Trillos et al. (2023).*

Feasible labeled interactions

There is an iterative method for computing the feasible interactions, starting from order 1-st order interaction(namely, all points of each class) and then building up to order K. This is done by the following algorithm:

Algorithm 5 Construct $C(\varepsilon)$

Input: X : data set, ε : adversarial budget.
  For each $i \in \mathcal{Y}$, set $C_{\{i\}}(\varepsilon) = \{(i,1),...,(i,n_i)\}$.
  for $k = 2,\dots,K$ **do**
    for A, A' with $|A| = |A'| = k-1, |A \cap A'| = k-2$ and $C_A(\varepsilon), C_{A'}(\varepsilon) \neq \varnothing$ **do**
      for Each $C \in C_A(\varepsilon), C' \in C_{A'}(\varepsilon)$ with $|C \cap C'| = k-2$ **do**
        Check if there exists a point x within ε of every point in $C \cup C'$.
        If so, add $C \cup C'$ to the set $C_{A\cup A'}(\varepsilon)$.
      end for
    end for
  end for
Output: $C(\varepsilon) = \bigcup_{A \in S_K} C_A(\varepsilon)$.

The main difficulty in implementing Algorithm 5 is to ensure that the checks for $|A \cap A'| = k-2$ and $|C \cap C'| = k-2$ are efficient and are not done by enumerating all possibilities. With a proper implementation, the most time consuming step is checking when a point x is within ε of every point in $C \cup C'$. This is often a non-trivial geometric problem. For example in $\mathbb{R}^n$ with the Euclidean distance it requires checking if as many as K spheres in $\mathbb{R}^n$ of radius ε have a mutual intersection. One geometry where this calculation is particularly simple is when using $d(x,y) = \|x-y\|_\infty$, where the problem is reduced to finding the intersection of axis-aligned rectangles.

In general the speed of Algorithm 5 can be estimated by $O(K|C(\varepsilon)|m(K))$ where $m(K)$ is the computational complexity required to check the existence of a point x for groups of size at most K. This complexity can at the worst be essentially the same as trying every possible group of K or fewer points, which is what would be required if for example ε is so large that every selection of points are close enough. However, in practice there are often far fewer higher-order interactions which leads to a much faster algorithm. This output-dependent complexity is often a substantial gain.

Optimization

With the notation above, we describe the optimization problem to get the exact solution for (3.2.6). The linear programming(LP) for the exact solution is

$$\begin{aligned} &\min \sum_{C \in C(\varepsilon)} w(C) \\ &\text{s.t. } w(C) \geqslant 0 \text{ for all } C \in C(\varepsilon), \quad I_\varepsilon w = [\mu_1, \ldots, \mu_K]^T \end{aligned} \tag{5.4.1}$$

Let w^* be the minimizer of this problem. Then g, the generalized barycenter of the optimal adversarial attacks, can be recovered as

$$\lambda = \sum_{C \in C(\varepsilon)} w^*(C) \delta_{F(C)}$$

where $F(C)$ returns any point x such that $\{x^i_{l_i} : (i, l_i) \in C\} \subset \overline{B}(x, \varepsilon)$, and such a point must exist by the condition $C \in C(\varepsilon)$. Furthermore the optimal adversarial attacks $\{\widetilde{\mu}_1, \ldots, \widetilde{\mu}_K\}$ can be recovered as

$$\widetilde{\mu}_i = \sum_{A \in S_K(i)} \sum_{C \in C_A(\varepsilon)} w^*(C) \delta_{F(C)}$$

The mass works correctly in this problem because of the constraint $I_\varepsilon w = [\mu_1, \ldots, \mu_K]^T$. From the preceding equations it is clear that g dominates $\widetilde{\mu}_i$ for each $i \in \mathcal{Y}$. In addition, it is also easy to recover the transformation $\mu_k \mapsto \widetilde{\mu}_i$.

This leads to (3.1.4) = (5.4.1), hence, the optimal adversarial risk is obtained by

$$(2.5) = 1 - (5.4.1) = 1 - \sum_{C \in C(\varepsilon)} w^*(C).$$

Complexity Considerations of Algorithm 3

In general the optimization problem involves a vector w whose length is determined by $|C(\varepsilon)|$ as well as a sparse matrix I_ε with at most $K|C(\varepsilon)|$ non-zero entries (although this is quite pessimistic). It is therefore essential to control $|C(\varepsilon)|$. The expected

size of this set is given by

$$\begin{aligned}
\mathbb{E}|C(\varepsilon)| &= \mathbb{E}\left[\sum_{A\in S_K} |C_A(\varepsilon)|\right] \\
&= \sum_{A\in S_K} \mathbb{E}\left[\sum_{C\in C_A} \mathbf{1}[C_A \in C_A(\varepsilon)]\right] \\
&= \sum_{A\in S_K} \sum_{C\in C_A} \mathbb{E}\left[\mathbf{1}[C_A \in C_A(\varepsilon)]\right] \\
&= \sum_{A\in S_K} \left[\prod_{i\in A} n_k\right] \mathbb{P}\left\{\{X_i\}_{i\in A} \subset \overline{B}(x,\varepsilon) \text{ for some } x\right\}
\end{aligned}$$

where $X_i \sim \mu_i$ are independent random variables. It is therefore crucial that the classes are in some sense well-separated as this will control the probability of the formation of an ε-interaction. It may be interesting in its own right to try and come up with interesting cases where we can cleanly bound the probability on the right hand side. For example, if $X_i \sim N(m_i, \Sigma_i)$, then we can reasonably expect to bound the probability by a function of the values of $\{(m_i, \Sigma_i)\}$.

Theorem 5.8 (Complexity of Algorithm 3)**.** *Assume that* $n_i = O(n)$ *for all* $i \in \mathcal{Y}$*. Then, Algorithm 3 gives* $\{g', \widetilde{\mu}_1', \ldots, \widetilde{\mu}_K\}$ *which satisfies*

$$\sum_{x\in\mathcal{X}} \lambda'(x) + \sum_{i\in\mathcal{Y}} C(\mu_i, \widetilde{\mu}_i') \leqslant (3.1.4) + \delta$$

in $O(K^2 n^K \left(\log \frac{1}{\delta}\right)^2)$ *time. Therefore, it gives a* δ*-approximation for* (2.5) *in* $O(K^2 n^K \left(\log \frac{1}{\delta}\right)^2)$ *time.*

Proof. First, the computational complexity of constructing $C(\varepsilon)$ is $O(K|C(\varepsilon)|m(K))$ where $m(K)$ is the computational complexity required to check the existence of a point x for groups of size at most K. Since $m(K) = O(K)$, its complexity is $O(K^2|C(\varepsilon)|)$.

In the worst case, the optimization part of Algorithm 3 requires a complexity of $O(K|C(\varepsilon)|\left(\log \frac{1}{\delta}\right)^2)$ where δ is an error tolerance if one solves the associated linear

program using the method of Yen et al. (2015) which leverages the sparsity in the constraint matrix. As noted above, $|C(\varepsilon)|$ can in the worst case be on the order of

$$|C(\varepsilon)| = \sum_{k=1}^{K} \binom{K}{k} O(n^k) \leqslant (O(n)+1)^K = O(n^K)$$

when each class i has $n_i = O(n)$ points. Therefore, in total, a computational complexity would be $O(K^2 n^K \left(\log \frac{1}{\delta}\right)^2)$. □

Truncation: Deterministic Approximation Ratio

The main reason that the complexity can explode in the previous section is that the number of ways of forming an interaction of order k is on the order of $O(n^k)$, and there are $\binom{K}{k}$ types of these interactions.

In practice we can avoid this blow up by simply truncating the maximum size of the interaction being considered. For example, suppose there are K classes and we choose a maximum interaction size of L. There are two reasons why this may be perfectly fine. First, is because it provides a provably a $\frac{K}{L}$-approximation of the optimal risk.

Proposition 5.9. *Let* $L < K$ *be a truncation level and* w^* *be the optimal value of* (5.4.1). *Consider the* L*-truncated version of* (5.4.1) *with the restriction to*

$$C^L(\varepsilon) := \bigcup_{A \in S_K^L} C_A(\varepsilon).$$

Let w_L^* *be the optimal value of the* L*-truncated version of* (5.4.1). *Then,*

$$\frac{w_L^*}{w^*} \leqslant \frac{K}{L}$$

Proof. Suppose that there is an interaction of order K that is assigned mass m. Since every sub-interaction of an admissible interaction is also admissible, we can re-allocate tha mass amongst the $\binom{K}{L}$-many sub-interactions of order L. Doing so

uniformly requires placing $\frac{mK}{L}\binom{K}{L}^{-1}$ on each of the $\binom{K}{L}$-many sub-interactions of order L .

This requires a total mass of

$$\frac{mK}{L}\binom{K}{L}^{-1} \cdot \binom{K}{L} = \frac{mK}{L}.$$

This satisfies the mass allocation constraints since each point in the original K interaction has m mass to allocate and belongs to precisely

$$\binom{K-1}{L-1} = \binom{K}{L} \cdot \frac{L}{K}$$

total interactions.

Let $w(C)$ be the optimal weight allocation vector for the full order K attack. Repeating the reallocation for every interaction C of order $L+1$ or greater we have an attack which requires mass

$$\begin{aligned}
&\sum_{C\in\mathcal{C}(\varepsilon):|C|\leqslant L} w(C) + \sum_{C\in\mathcal{C}(\varepsilon):|C|=L}\ \sum_{C'\in\mathcal{C}(\varepsilon):C\subsetneq C'} \frac{|C'|w(C')}{|C|}\binom{|C'|}{|C|}^{-1} \\
&\leqslant \sum_{C\in\mathcal{C}(\varepsilon):|C|\leqslant L} w(C) + \sum_{C\in\mathcal{C}(\varepsilon):|C|=L}\ \sum_{C'\in\mathcal{C}(\varepsilon):C\subsetneq C'} \frac{Kw(C')}{L}\binom{|C'|}{L}^{-1} \\
&= \sum_{C\in\mathcal{C}(\varepsilon):|C|\leqslant L} w(C) + \sum_{C'\in\mathcal{C}(\varepsilon):|C'|>L} \frac{Kw(C')}{L}
\end{aligned}$$

where $C \subsetneq C'$ means C is a strict subset of C'. The inequality comes from $|C'| \leqslant K$.

To complete the proof divide this cost by the cost of the full attack. Note that for $a \geqslant b \geqslant 0$ and any $c \geqslant 0$ such that $a + c, b + c > 0$ we it holds $\frac{c+a}{c+b} \leqslant \frac{a}{b}$. Applying this we have

$$\frac{\sum_{C\in\mathcal{C}(\varepsilon):|C|\leqslant L} w(C) + \sum_{C'\in\mathcal{C}(\varepsilon):|C'|>L} \frac{Kw(C')}{L}}{\sum_{C\in\mathcal{C}(\varepsilon):|C|\leqslant L} w(C) + \sum_{C'\in\mathcal{C}(\varepsilon):|C'|>L} w(C')} \leqslant \frac{\sum_{C'\in\mathcal{C}(\varepsilon):|C'|>L} \frac{Kw(C')}{L}}{\sum_{C'\in\mathcal{C}(\varepsilon):|C'|>L} w(C')} = \frac{K}{L}$$

□

The proof above always works. However it may be pessimistic for the second reason that truncating is most likely not a bad thing to do: There is likely to be much more mass in small interactions than in large interactions. A better book keeping can be used to show that the restriction to order L interactions gives an approximation error of

$$1P_{\leqslant L} + \sum_{L=L+1}^{K} \frac{L}{L} P_L$$

where

$$P_{\leqslant L} = \frac{\sum_{C \in \mathcal{C}(\varepsilon):|C| \leqslant L} w(C)}{\sum_{C \in \mathcal{C}(\varepsilon)} w(C)}, \quad P_L = \frac{\sum_{C \in \mathcal{C}(\varepsilon):|C|=L} w(C)}{\sum_{C \in \mathcal{C}(\varepsilon)} w(C)}.$$

If P_L decays rapidly with L, then the bound will be much tighter than the pessimistic $\frac{K}{L}$.

In addition, the truncation leads to a much smaller set of interactions, denoted by $\mathcal{C}(\varepsilon; L)$ and improves the speed of solving the linear program to $O(L|\mathcal{C}(\varepsilon; L)| \left(\log \frac{1}{\delta}\right)^2)$ where the factor of L is because each interaction leads to a row in the constraint matrix with at most L non-zero entries instead of K. Again in the worst case we have

$$|\mathcal{C}(\varepsilon; L)| \leqslant \sum_{k=1}^{L} \binom{K}{k} n^k = O(K^L n^L) \tag{5.4.2}$$

which leads to a total complexity of $O(L^2 K^L n^L \left(\log \frac{1}{\delta}\right)^2)$ again using the method of Yen et al. (2015) to solve the linear program.

Truncated Entropic Regularization

Derivation of Truncated Entropic Regularization

Recalling (2.5) = 1 − (3.2.6), now the goal is to solve (3.2.6). One popular way to solve optimal transport problems numerically is the entropic regularization, also known as Sinkhorn interation, introduced in Cuturi (2013)(originally proposed in Sinkhorn (1964); Sinkhorn and Knopp (1967). The idea is that to add a certain

regularization term which enforces a problem to be strictly convex: the most popular choice of regularization term is entropy.

The entropic regularization of (3.2.6) is defined as

$$\begin{aligned}&\min_{\{\pi_A\}_{A\in S_K}} \sum_{A\in S_K} \int_{\mathcal{X}^K} (1 + c_A(x_1, ..., x_K))\, d\pi_A(x_1, ..., x_K) - \eta\,(\mathrm{Ent}\,(\pi_A) + \|\pi_A\|)\\ &\text{s.t.} \sum_{A\in S_K(i)} \mathcal{P}_{i\#}\pi_A = \mu_i \quad \text{for all } i \in \mathcal{Y}\end{aligned} \tag{5.4.3}$$

where $\mathrm{Ent}\,(\pi_A)$ and $\|\pi_A\|$ are the entropy and the total mass of π_A, respectively. Here, when computing $\mathrm{Ent}\,(\pi_A)$ we regard $0 \cdot \log 0 = 0$ according to the convention. The term $\eta\|\pi_A\|$ is introduced for the computational convenience.

From now on, we consider (5.4.3) with the finite support case to derive Sinkhorn-type update scheme. Introducing the Lagrangain dual variables $g_1 \in \mathbb{R}^{n_1}, \ldots, g_K \in \mathbb{R}^{n_K}$, we need to solve

$$\begin{aligned}&\min_{\{\pi_A\},\{g_i\}} \mathcal{L}(\{\pi_A\},\{g_i\})\\ &:= \sum_{A\in S_K} \sum_{x_A\in\mathcal{X}^A} (1 + c_A(x_A) + \eta\,(\log \pi_A - 1))\,\pi_A(x_A)\\ &\quad - \sum_{i\in\mathcal{Y}} \sum_{x_i\in\mathcal{X}_i} g_i(x_i) \left(\sum_{A\in S_K(i)} \sum_{x_{A\setminus\{i\}}\in\mathcal{X}^{A\setminus\{i\}}} \pi_A(x_i, x_{A\setminus\{i\}}) - \mu_i(x_i) \right).\end{aligned} \tag{5.4.4}$$

Since it is strictly convex, the first order condition is sufficient for characterizing a solution. Differentiating with respect to $\pi_A(x_A)$ for each $A \in S_K$ and each $x_A \in \mathcal{X}^A$ yields

$$0 = \partial_{\pi_A(x_A)}\mathcal{L}(\{\pi_A\},\{g_i\}) = 1 + c_A(x_A) + \eta \log \pi_A(x_A) - \sum_{i\in A} g_i(x_i).$$

From the above, it is deduced that

$$\log \pi_A(x_A) = \frac{1}{\eta}\left(\sum_{i\in A} g_i(x_i) - 1 - c_A(x_A)\right)$$

Plugging such $\log \pi_A$'s into (5.4.4), we have

$$\begin{aligned}\mathcal{L}(\{g_i\}) = &\sum_{A\in S_K}\sum_{x_A\in\mathcal{X}^A}\left(1 + c_A(x_A) + \sum_{i\in A} g_i(x_i) - 1 - c_A(x_A) - \eta\right)\pi_A(x_A)\\ &-\sum_{i\in\mathcal{Y}}\sum_{x_i\in\mathcal{X}_i} g_i(x_i)\left(\sum_{A\in S_K(i)}\sum_{x_{A\setminus\{i\}}\in\mathcal{X}^{A\setminus\{i\}}}\pi_A(x_i, x_{A\setminus\{i\}}) - \mu_i(x_i)\right).\end{aligned}$$

Since

$$\begin{aligned}&\sum_{i\in\mathcal{Y}}\sum_{x_i\in\mathcal{X}_i} g_i(x_i)\sum_{A\in S_K(i)}\sum_{x_{A\setminus\{i\}}\in\mathcal{X}^{A\setminus\{i\}}}\pi_A(x_i, x_{A\setminus\{i\}})\\ &= \sum_{i\in\mathcal{Y}}\sum_{A\in S_K(i)}\sum_{x_A\in\mathcal{X}^A} g_i(x_i)\pi_A(x_A)\\ &= \sum_{A\in S_K}\sum_{x_A\in\mathcal{X}^A}\left(\sum_{i\in A} g_i(x_i)\right)\pi_A(x_A),\end{aligned}$$

it follows that the dual of (5.4.3) is

$$\begin{aligned}\max_{\{g_i : i\in\mathcal{Y}\}}\mathcal{L}(\{g_i\}) = &\sum_{i\in\mathcal{Y}}\sum_{x_i\in\mathcal{X}_i} g_i(x_i)\mu_i(x_i)\\ &-\eta\sum_{A\in S_K}\sum_{x_A\in\mathcal{X}^A}\exp\left(\frac{1}{\eta}\left(\sum_{i\in A} g_i(x_i) - 1 - c_A(x_A)\right)\right).\end{aligned} \tag{5.4.5}$$

Using $g_A := (g_i)_{i\in A}$, let us introduce the tensor product associated with $A \in S_K$

by

$$\begin{aligned}&G(g_{A\setminus\{i\}})(x_i)\\&:=\sum_{x_{A\setminus\{i\}}\in\mathcal{X}^{A\setminus\{i\}}}\exp\left(\frac{1}{\eta}\left(\sum_{j\in A\setminus\{i\}}g_j(x^j_{l_j})\right)\right)\exp\left(-\frac{1}{\eta}\left(1+c_A(x_i,x_{A\setminus\{i\}})\right)\right).\end{aligned}$$

For the convenience, we define $G(g_\emptyset) := 1$.

For each $i \in \mathcal{Y}$ and each $x_i \in \mathcal{X}_i$, the first order condition is

$$0 = \partial_{g_i(x_i)}\mathcal{L}(\{g_i\}) = \mu_i(x_i) - \exp\left(\frac{1}{\eta}g_i(x_i)\right)\sum_{A\in S_K(i)}G(g_{A\setminus\{i\}})(x_i).$$

Then, the above first order condition yields

$$g_i(x_i) = \eta\log\mu_i(x_i) - \eta\log\left(\sum_{A\in S_K(i)}G(g_{A\setminus\{i\}})(x_i)\right). \tag{5.4.6}$$

Let $\{g_i^*\}$ be a maximizer of (5.4.5). Then, the minimizer of (5.4.3), $\{\pi_A^*\}$, is obtained by

$$\pi_A^*(x_A) = \exp\left(\frac{1}{\eta}\left(\sum_{i\in A}g_i^*(x_i)\right)\right)\exp\left(-\frac{1}{\eta}\left(1+c_A(x_A)\right)\right)$$

and the corresponding approximated minimum value is

$$\sum_{A\in S_K}\sum_{x_A\in\mathcal{X}^A}\left(1+c_A(x_A)\right)\pi_A^*(x_A).$$

All the arguments above are similar to the derivation of Sinkhorn iteration for classical optimal transport: but the denominator is not just the matrix multiplication but the tensor product. For each $i \in \mathcal{Y}$ we update g_i according to (5.4.6) while other g_j's are fixed. Since we fix x_i, the computational complexity of $\sum_{A\in S_K(i)}G(g_{A\setminus\{i\}})$ is $O(n^{K-1})$ caused by the largest contribution $A = \mathcal{Y}$. Of course, whenever K is

large, its complexity is huge so that computing this problem is quite intractable in practice.

Remark 5.10. (5.4.6) *can be formulated in a different way. Let* $\phi_i := \exp\left(\frac{1}{\eta} g_i\right)$. *Then,* (5.4.6) *is equivalent to*

$$\phi_i(x_i) = \frac{\mu_i(x_i)}{\sum_{A \in S_K(i)} G(g_{A\setminus\{i\}})(x_i)}. \tag{5.4.7}$$

This different point of view is also widely adopted in computational optimal transport community: see (*Peyré et al., 2019, section 4*) *for more details.*

One can truncate (5.4.3) by considering only $A \in S_K$ with $|A| \leqslant L \ll K$. This means that we only care about the lower-order interactions and ignore the higher-order ones. This truncation or ignorance of the higher-order interactions can be justified by empirical observations: in most real data sets, with a reasonable adversarial budget, an adversarial attack is concentrated on at most 2nd order or 3rd order interactions. See section 5.2.

Fix the truncation level $L < K$. Recall $S_K^L = \{A \in S_K : |A| \leqslant L\}$ and $S_K^L(i) = \{A \in S_K^L : i \in A\}$. The L-level truncated entropic regularization problem is

$$\begin{aligned} &\min_{\pi_A} \sum_{A \in S_K^L} \int_{\mathcal{X}^K} (1 + c_A(x_1, ..., x_K))\, d\pi_A(x_1, ..., x_K) - \eta\left(\mathrm{Ent}\,(\pi_A) + \|\pi_A\|\right) \\ &\text{s.t.} \sum_{A \in S_K^L(i)} \mathcal{P}_{i\#}\pi_A = \mu_i \qquad \text{for all } i \in \mathcal{Y}. \end{aligned} \tag{5.4.8}$$

(5.4.8) would be a good approximation of (5.4.3) provided that $\pi_A \approx 0$ for all $|A| > L$. Applying the same argument in subsection 5.4, the dual formulation suggests the following update scheme:

$$g_i(x_i) := \eta \log \mu_i(x_i) - \eta \log \left(\sum_{A \in S_K^L(i)} G(g_{A\setminus\{i\}})(x_i) \right). \tag{5.4.9}$$

Complexity Considerations of Algorithm 4

In this subsection, we obtain an upper bound of the computational complexity of Algorithm 4. The proof is based on Carlier (2022) which is built on Beck and Tetruashvili (2013) which is regarding the block coordinate descent method.

First of all, we claim that an upper and lower bound of g^* for (5.4.5).

Remark 5.11. *Notice that $\mathcal{L}(\{g_i\})$ might have many solutions unless the restriction of $\{g_i\}$. For example, assume that $\mu_i = \frac{1}{K}\delta_{x_i}$ for all $i \in \mathcal{Y}$ and there is some x' such that $d(x_i, x') \leqslant \varepsilon$ for all $i \in \mathcal{Y}$. Then, the optimal π^* is $\{\pi^*_{\mathcal{Y}} = \frac{1}{K}\delta_{x'}\}$. In this case, any $g = (g_1, \ldots, g_K)$ such that $\sum_{i\in\mathcal{Y}} g_i = 1$ can be a solution for $\mathcal{L}(\{g_i\})$. This is a common issue in entropic optimal transport problems(equivalently, in Sinkhorn iteration). So, people have used Hilbert–Birkhoff projective metric or centring method with bounded cost setting to manage this issue: see Franklin and Lorenz (1989); Chen et al. (2016); Di Marino and Gerolin (2020); Carlier (2022).*

Lemma 5.12. *Let $\gamma := \min_{i\in\mathcal{Y}, x_i\in\mathcal{X}_i} \log \mu_i(x_i) > -\infty$. If g^* is a maximizer of (5.4.5), then*

$$-(L-2) + \eta\left(\gamma - (L-1)\log(K-1) - L\log n\right) \leqslant g_i(x_i) \leqslant 1$$

for all $i \in \mathcal{Y}$ and all $x_i \in \mathcal{X}_i$.

Proof. Recalling the first order condition, one has

$$g_i^*(x_i) = \eta \log \mu_i(x_i) - \eta \log\left(\sum_{A\in S_K^L(i)} G(g^*_{A\setminus\{i\}})(x_i)\right) \leqslant -\eta \log\left(\sum_{A\in S_K^L(i)} G(g^*_{A\setminus\{i\}})(x_i)\right)$$

since $\mu_i \leqslant 1$. Notice that the right-hand side above is the usual log-sum-exp operator. Let

$$\overline{G}(g, i)(x) := \max_{x_{A\setminus\{i\}}\in\mathcal{X}^{A\setminus\{i\}}, A\in S_K^L(i)} \left(\sum_{j\in A\setminus\{i\}} g_j^*(x_j) - \left(1 + c_A(x, x_{A\setminus\{i\}})\right)\right). \quad (5.4.10)$$

Then,

$$\overline{G}(g,i)(x_i) \leqslant \eta \log\left(\sum_{A\in S_k^L(i)} G(g^*_{A\setminus\{i\}})(x_i)\right) \leqslant \overline{G}(g,i)(x) + \eta\log(\text{number of terms}).$$

By taking $A = \{i\}$, it is easy to check that $\overline{G}(g,i)(x) \geqslant -1$. Hence,

$$g_i^*(x_i) \leqslant -\eta\log\left(\sum_{A\in S_k^L(i)} G(g^*_{A\setminus\{i\}})(x_i)\right) \leqslant 1.$$

Observe that the number of terms of $\sum_{A\in S_k^L(i)} G(g^*_{A\setminus\{i\}})(x_i)$ is bounded by $(K-1)^{L-1}n^L$ where $n := \max\{n_1,\dots,n_K\}$. Since $g_i^* \leqslant 1$,it follows that

$$\begin{aligned} g_i^*(x_i) &\geqslant \eta\gamma - (L-2) - \eta\log\left((K-1)^{L-1}n^L\right) \\ &= -(L-2) + \eta\left(\gamma - \log(K-1) - L\log n\right). \end{aligned}$$

This completes the proof. □

Notice that g_i's in the dual problem (5.4.5) correspond to g_i's of (3.2.13) of Proposition 3.19. It is also proved there that it suffices to consider $0 \leqslant g_i \leqslant 1$. With the entropic regularization term, this is not trivial. But, if $\eta \to 0$, (5.12) recovers the usual bounds of g_i's which can be also obtained by c-transform: see Villani (2009).

Recall Algorithm 4. We elaborate the algorithm in detail: for each $t = 1, 2, \dots$ let's consider the subroutine $k = 1,\dots,K$ such that given $g^t = (g_1^t,\dots,g_K^t)$, update

$$g_1^{t+1}(\cdot) \leftarrow \eta\log\mu_1(\cdot) - \eta\log\left(\sum_{A\in S_k^L(1)} G(g^t_{A\setminus\{1\}})\right)(\cdot)$$

and set $g^{t+1,1} = (g_1^{t+1}, g_2^t,\dots,g_K^t)$. In general, given $g^{t+1,k} = (g_1^{t+1},\dots,g_k^{t+1}, g_{k+1}^t,\dots g_K^t)$,

update

$$g_{k+1}^{t+1}(\cdot) \leftarrow \eta \log \mu_1(\cdot) - \eta \log \left(\sum_{A \in S_K^L(1)} G(g_{A\setminus\{1\}}^{t+1,k}) \right)(\cdot).$$

Finally, $g^{t+1} = g^{t+1,K} = (g_1^{t+1}, g_2^{t+1}, \ldots, g_K^{t+1})$ and move to $t+2$ and consider the subroutine $k = 1, \ldots, K$ again. For the consistency of notation, let $g^t = g^{t+1,0} = (g_1^t, g_2^t, \ldots, g_K^t)$

The following lemma gives an upper and a lower bounds of g^t.

Lemma 5.13. *Let* $g^0 = (0, \ldots, 0)$ *and* $\gamma := \min_{i \in \mathcal{Y}, x_i \in \mathcal{X}_i} \log \mu_i(x_i) > -\infty$. *For each* $t = 1, 2, \ldots$

$$-(L-2) + \eta\left(\gamma - (L-1)\log(K-1) - L\log n\right) \leqslant g_i^t(x_i) \leqslant 1. \tag{5.4.11}$$

for each $i \in \mathcal{Y}$ *and each* $x_i \in \mathcal{X}_i$.

Proof. Recall (5.4.10). Since $g_j^0 = 0$ for all $j \in \mathcal{Y}$, $\overline{G}(g^0, 1)(x_1) = -1$. So, it follows from the above that for each $x_1 \in \mathcal{X}_1$,

$$g_1^1(x_1) \leqslant 1. \tag{5.4.12}$$

We use an induction argument on the subroutine k. Let's Assume that (5.4.12) holds until $k-1$. Again, it holds that

$$\overline{G}(g^{1,k-1}, k)(x_k) \leqslant \eta \log \left(\sum_{A \in S_K^L(k)} G(g_{A\setminus\{k\}}^{1,k-1})(x_k) \right).$$

Notice that

$$-1 \leqslant \overline{G}(g^{1,k-1}, k)(x)$$

by taking $A = \{k\}$. Thus,

$$g_k^1(x_k) \leqslant 1.$$

This verifies the induction hypothesis for the subroutine, $k = 1, \ldots, K$.

Let's use induction on $t = 1, 2, \ldots$. Assume that (5.4.12) holds until $t-1$. It is easy to check that the setting of the subroutine remains the same as before, so one can directly apply the above argument. Therefore, the induction hypothesis is satisfied for $t = 1, 2, \ldots$. This verifies an upper bound.

With an upper bound, a lower bound can be obtained analogously as in Lemma 5.12 by using induction argument again. □

Recall the objective function of the dual maximization problem:

$$\begin{aligned}\mathcal{L}^L(\{g_i\}) := &\sum_{i\in\mathcal{Y}}\sum_{x_i\in\mathcal{X}_i} g_i(x_i)\mu_i(x_i) \\ &-\eta\sum_{A\in S_K^L}\sum_{x_A\in\mathcal{X}^A}\exp\left(\frac{1}{\eta}\left(\sum_{i\in A} g_i(x_i) - 1 - c_A(x_A)\right)\right).\end{aligned} \tag{5.4.13}$$

It is easy to see that it is concave. For simplicity, let

$$\begin{aligned}\Gamma &:= \left|\min\{-(L-2)+\eta\left(\gamma-(L-1)\log(K-1)-L\log n\right),-1\}\right|, \\ \mathcal{K} &:= \prod_{i\in\mathcal{Y}}[-\Gamma,1]^{n_i}.\end{aligned} \tag{5.4.14}$$

Then, $\mathcal{L}^L(\{g_i\})$ is strongly concave on $\mathcal{K}$. Also, $g^* \in \mathcal{K}$

The following lemma is analogous to Carlier (2022)[Lemma 3.2].

Lemma 5.14. *For each* $t = 0, 1, \ldots$

$$\mathcal{L}^L(g^{t+1}) - \mathcal{L}^L(g^t) \geqslant \frac{e^{-(\Gamma+\frac{1}{\eta})}}{2\eta}\sum_{k=1}^{K}\sum_{x_k\in\mathcal{X}_k}\left(g_k^{t+1}(x_k) - g_k^t(x_k)\right)^2\mu_k(x_k).$$

Proof. Writing $\mathcal{L}^L(g^{t+1}) - \mathcal{L}^L(g^t)$ in a telescopic fashion,

$$\mathcal{L}^L(g^{t+1}) - \mathcal{L}^L(g^t) = \sum_{k=1}^{K}\mathcal{L}^L(g^{t+1,k}) - \mathcal{L}^L(g^{t+1,k-1})$$

where $g^{t+1,k-1} = g^t$. Let us focus on $\mathcal{L}^L(g^{t+1,1}) - \mathcal{L}^L(g^{t+1,0})$. The other cases are analogous.

A straightforward calculation leads to

$$\begin{aligned}
&\mathcal{L}^L(g^{t+1,1}) - \mathcal{L}^L(g^{t+1,0}) \\
&= \sum_{x_1 \in \mathcal{X}_1} \left(g_1^{t+1}(x_1) - g_1^t(x_1)\right) \mu_1(x_1) \\
&\quad + \eta \sum_{A \in S_k^L(1)} \sum_{x_A \in \mathcal{X}^A} \left(\exp\left(\frac{1}{\eta} g^t(x_1)\right) - \exp\left(\frac{1}{\eta} g^{t+1}(x_1)\right)\right) \\
&\qquad\qquad \times \exp\left(\frac{1}{\eta}\left(\sum_{i \in A \setminus \{1\}} g_i^t(x_i) - (1 + c_A(x_A))\right)\right).
\end{aligned} \tag{5.4.15}$$

Notice that if $a, b \in [-m, m]$,

$$e^b - e^a \geqslant e^a(b-a) + \frac{e^{-m}}{2}(b-a)^2. \tag{5.4.16}$$

Combined with (5.4.11), (5.4.16) yields that for each $x_1 \in \mathcal{X}_1$,

$$\begin{aligned}
&\exp\left(\frac{1}{\eta} g^t(x_1)\right) - \exp\left(\frac{1}{\eta} g^{t+1}(x_1)\right) \\
&\geqslant \frac{1}{\eta}\left(g^t(x_1) - g^{t+1}(x_1)\right) \exp\left(\frac{1}{\eta} g_1^{t+1}(x_1)\right) + \frac{e^{-\Gamma}}{2\eta^2}\left(g^t(x_1) - g^{t+1}(x_1)\right)^2.
\end{aligned} \tag{5.4.17}$$

Let $\pi_A^{t+1,k}$ denote $\pi_A(g^{t+1,k})$ defined as

$$\begin{aligned}
&\pi_A(g^{t+1,k})(x_A) \\
&:= \exp\left(\frac{1}{\eta}\left(\sum_{i \in A : i \leqslant k} g_i^{t+1}(x_i) + \sum_{i \in A : i > k} g_i^t(x_i)\right)\right) \exp\left(-\frac{1}{\eta}\left(1 + c_A(x_A)\right)\right)
\end{aligned}$$

and

$$\sum_{A \in S_K(i)} (\mathcal{P}_{i\#} \pi_A^{t+1,k})(x_i) := \sum_{A \in S_k^L(i)} \sum_{x_{A \setminus \{i\}} \in \mathcal{X}^{A \setminus \{i\}}} \pi_A(g^{t+1,k})(x_i, x_{A \setminus \{i\}}).$$

An equality associated with the first term of the right-hand side of (5.4.17) is expressed as

$$\sum_{A\in S_k^L(1)}\sum_{x_A\in\mathcal{X}^A}\left(g^t(x_1)-g^{t+1}(x_1)\right)\exp\left(\frac{1}{\eta}\left(g_1^{t+1}(x_1)+\sum_{i\neq 1\in A}g_i^t(x_i)-(1+c_A(x_A))\right)\right)$$
$$=\sum_{x_1\in\mathcal{X}_1}\left(g^t(x_1)-g^{t+1}(x_1)\right)\sum_{A\in S_K(1)}(\mathcal{P}_{1\#}\pi_A^{t+1,1})(x_1).$$

The other inequality associated with the second term of the right-hand side of (5.4.17) is written similarly as

$$\sum_{A\in S_k^L(1)}\sum_{x_A\in\mathcal{X}^A}\frac{e^{-\frac{1}{\eta}}}{2\eta}\left(g^t(x_1)-g^{t+1}(x_1)\right)^2\exp\left(\frac{1}{\eta}\left(\sum_{i\neq 1\in A}g_i^t(x_i)-(1+c_A(x_A)\right)\right)$$
$$=\frac{e^{-\Gamma}}{2\eta}\sum_{x_1\in\mathcal{X}_1}\left(g^t(x_1)-g^{t+1}(x_1)\right)^2\exp\left(-\frac{1}{\eta}g_1^{t+1}(x_1)\right)\sum_{A\in S_K(1)}(\mathcal{P}_{1\#}\pi_A^{t+1,1})(x_1).$$

Returning to (5.4.15), we have

$$\mathcal{L}^L(g^{t+1,1})-\mathcal{L}^L(g^{t+1,0})$$
$$\geqslant\sum_{x_1\in\mathcal{X}_1}\left(g_1^{t+1}(x_1)-g_1^t(x_1)\right)\mu_1(x_1)+\sum_{x_1\in\mathcal{X}_1}\left(g^t(x_1)-g^{t+1}(x_1)\right)\sum_{A\in S_K(1)}(\mathcal{P}_{1\#}\pi_A^{t+1,1})(x_1)$$
$$+\frac{e^{-\Gamma}}{2\eta}\sum_{x_1\in\mathcal{X}_1}\left(g^t(x_1)-g^{t+1}(x_1)\right)^2\exp\left(-\frac{1}{\eta}g_1^{t+1}(x_1)\right)\sum_{A\in S_K(1)}(\mathcal{P}_{1\#}\pi_A^{t+1,1})(x_1).$$

Observe that, by the construction of $g^{t+1,i}$,

$$\sum_{A\in S_K(i)}(\mathcal{P}_{i\#}\pi_A^{t+1,i})(x_i)=\mu_i(x_i). \tag{5.4.18}$$

Hence,

$$\begin{aligned}
&\mathcal{L}^{L}(g^{t+1,1}) - \mathcal{L}^{L}(g^{t+1,0}) \\
&\geqslant \sum_{x_1 \in \mathcal{X}_1} \left(g_1^{t+1}(x_1) - g_1^{t}(x_1)\right) \mu_1(x_1) + \sum_{x_1 \in \mathcal{X}_1} \left(g^{t}(x_1) - g^{t+1}(x_1)\right) \mu_1(x_1) \\
&\quad + \frac{e^{-\Gamma}}{2\eta} \sum_{x_1 \in \mathcal{X}_1} \left(g^{t}(x_1) - g^{t+1}(x_1)\right)^2 \exp\left(-\frac{1}{\eta} g_1^{t+1}(x_1)\right) \sum_{A \in S_K(1)} (\mathcal{P}_{1\#} \pi_A^{t+1,1})(x_1) \\
&\geqslant \frac{e^{-\Gamma}}{2\eta} \sum_{x_1 \in \mathcal{X}_1} \left(g^{t}(x_1) - g^{t+1}(x_1)\right)^2 \exp\left(-\frac{1}{\eta} g_1^{t+1}(x_1)\right) \sum_{A \in S_K(1)} (\mathcal{P}_{1\#} \pi_A^{t+1,1})(x_1).
\end{aligned}$$

Since $\exp\left(-\frac{1}{\eta} g_1^{t+1}(x_1)\right) \geqslant \exp\left(-\frac{1}{\eta}\right)$ due to the fact that $g_i^t \leqslant 1$ for all $i \in \mathcal{Y}$, it follows that

$$\mathcal{L}^{L}(g^{t+1,1}) - \mathcal{L}^{L}(g^{t+1,0}) \geqslant \frac{e^{-(\Gamma + \frac{1}{\eta})}}{2\eta} \sum_{x_1 \in \mathcal{X}_1} \left(g_1^{t}(x_1) - g_1^{t+1}(x_1)\right)^2 \mu_i(x_i).$$

The conclusion, therefore, follows. □

Theorem 5.15 (Linear convergence of Algorithm 4)**.** *For* $t = 0, 1, \ldots$

$$\mathcal{L}^{L}(g^*) - \mathcal{L}^{L}(g^t) \leqslant \left(1 - \frac{e^{-2\Gamma + (\Gamma - 1)\frac{1}{\eta}}}{4K}\right)^t \left(\mathcal{L}^{L}(g^*) - \mathcal{L}^{L}(g^0)\right). \tag{5.4.19}$$

Remark 5.16. *Recall that* $\Gamma \approx L + \eta(\gamma - L \log n)$*. Hence, with* $\gamma = \min_{i, x_i} \log \mu_i(x_i) < 0$,

$$(\Gamma - 1)\frac{1}{\eta} \approx \gamma - L \log n < 0.$$

So, on the reasonable regime of $\eta \ll 1$*, the multiplicative factor of* (5.4.19) *is less than* 1.

Proof. First of all, the convergence follows as usual: since $\mathcal{K}$ is compact, one can find a convergent subsequence (after relabeling properly), $\{g^t\}$. Then the limit of this subsequence, g^*, readily satisfies (5.4.9) since it is a fixed point, hence due to the strong concavity of $\mathcal{L}^L$ on $\mathcal{K}$, it is the unique maximizer. Thus, the whole

sequence $\{g^t\}$ converges to g^*.

Let $g^* \in \mathcal{K}$ be the maximizer of $\mathcal{L}^L$. Since $-\mathcal{L}^L$ is $\frac{e^{-(\Gamma+\frac{1}{\eta})}}{\eta}$-strongly convex on $\mathcal{K}$,

$$-\mathcal{L}^L(g^*) - (-\mathcal{L}^L(g^t)) \geqslant \sum_{i=1}^{K} \langle \partial_{g_i}(-\mathcal{L}^L(g^t)), g_i^* - g_i^t \rangle + \frac{e^{-(\Gamma+\frac{1}{\eta})}}{2\eta} \|g^* - g^t\|^2.$$

By (5.4.9), $\partial_{g_i}\mathcal{L}^L(g^{t+1,i}) = 0$. So,

$$\begin{aligned} &-\mathcal{L}^L(g^*) - (-\mathcal{L}^L(g^t)) \\ &\geqslant \sum_{i=1}^{K} \langle \partial_{g_i}(-\mathcal{L}^L(g^t)) - \partial_{g_i}(-\mathcal{L}^L(g^{t+1,i})), g_i^* - g_i^t \rangle + \frac{e^{-(\Gamma+\frac{1}{\eta})}}{2\eta} \|g^* - g^t\|^2. \end{aligned}$$

Recall Young's inequality which is

$$ab \leqslant \frac{a^2}{2q} + \frac{qb^2}{2}.$$

Applying this to the above inequality, we have

$$\mathcal{L}^L(g^*) - \mathcal{L}^L(g^t) \leqslant \frac{2\eta}{e^{-(\Gamma+\frac{1}{\eta})}} \|\partial_{g_i}\mathcal{L}^L(g^t)\|^2.$$

Again, writing $\|\partial_{g_i}\mathcal{L}^L(g^t)\|^2$ in a telescopic fashion,

$$\|\partial_{g_i}\mathcal{L}^L(g^t)\|^2 \leqslant \sum_{j=1}^{K} \|\partial_{g_i}\mathcal{L}^L(g^{t+1,j}) - \partial_{g_i}\mathcal{L}^L(g^{t+1,j-1})\|^2$$

where $g^{t+1,j} = (g_1^{t+1}, \ldots, g_j^{t+1}, g_{j+1}^t, \ldots g_K^t)$. A straightforward computation yields

that

$$\begin{aligned}
&\partial_{g_i}\mathcal{L}^L(g^{t+1,j})(x_i) - \partial_{g_i}\mathcal{L}^L(g^{t+1,j-1})(x_i) \\
&= -\exp\left(\frac{1}{\eta}g_i^t(x_i)\right) \\
&\quad \times \left\{ \sum_{A\in S_k^L(i,j)} \sum_{x_j\in\mathcal{X}_j} \left(\exp\left(\frac{1}{\eta}g_j^{t+1}(x_j)\right) - \exp\left(\frac{1}{\eta}g_j^t(x_j)\right)\right) G(g_{A\setminus\{i,j\}}^{t+1,j-1})(x_i,x_j) \right\}
\end{aligned}$$

if $j < i$. Here $S_k^L(i,j) = \{A \in S_k^L : i,j \in A\}$ and

$$\begin{aligned}
&G(g_{A\setminus\{i,j\}})(x_i,x_j) \\
&:= \sum_{x_{A\setminus\{i,j\}}\in\mathcal{X}^{A\setminus\{i,j\}}} \exp\left(\frac{1}{\eta}\left(\sum_{k\in A\setminus\{i,j\}} g_k(x_k)\right)\right) \exp\left(-\frac{1}{\eta}\left(1 + c_A(x_i,x_j,x_{A\setminus\{i,j\}})\right)\right).
\end{aligned}$$

For $j > i$,

$$\begin{aligned}
&\partial_{g_i}\mathcal{L}^L(g^{t+1,j})(x_i) - \partial_{g_i}\mathcal{L}^L(g^{t+1,j-1})(x_i) \\
&= -\exp\left(\frac{1}{\eta}g_i^{t+1}(x_i)\right) \\
&\quad \times \left\{ \sum_{A\in S_k^L(i,j)} \sum_{x_j\in\mathcal{X}_j} \left(\exp\left(\frac{1}{\eta}g_j^{t+1}(x_j)\right) - \exp\left(\frac{1}{\eta}g_j^t(x_j)\right)\right) G(g_{A\setminus\{i,j\}}^{t+1,j-1})(x_i,x_j) \right\}
\end{aligned}$$

and for $j = i$,

$$\begin{aligned}
&\partial_{g_i}\mathcal{L}^L(g^{t+1,i})(x_i) - \partial_{g_i}\mathcal{L}^L(g^{t+1,i-1})(x_i) \\
&= \sum_{A\in S_k^L(i)} \sum_{x_i\in\mathcal{X}_i} \left(\exp\left(\frac{1}{\eta}g_i^{t+1}(x_i)\right) - \exp\left(\frac{1}{\eta}g_i^t(x_i)\right)\right) G(g_{A\setminus\{i\}}^{t+1,i-1})(x_i).
\end{aligned}$$

Let's consider $j < i$ case first. Notice that if $a, b \in [-m, m]$,

$$|e^b - e^a| \leqslant e^m|b-a|. \tag{5.4.20}$$

Using (5.4.20) and Jensen's inequality,

$$
\begin{aligned}
&\left(\partial_{g_i}\mathcal{L}^L(g^{t+1,j})(x_i) - \partial_{g_i}\mathcal{L}^L(g^{t+1,j-1})(x_i)\right)^2 \\
&= \exp\left(\frac{1}{\eta}g_i^t(x_i)\right)^2 \\
&\quad \times \left\{\sum_{A\in S_k^L(i,j)}\sum_{x_j\in\mathcal{X}_j}\left(\exp\left(\frac{1}{\eta}g_j^{t+1}(x_j)\right) - \exp\left(\frac{1}{\eta}g_j^t(x_j)\right)\right)G(g_{A\setminus\{i,j\}}^{t+1,j-1})(x_i,x_j)\right\}^2 \\
&\leqslant \exp\left(\frac{1}{\eta}g_i^t(x_i)\right)^2 \\
&\quad \times \left\{\sum_{A\in S_k^L(i,j)}\sum_{x_j\in\mathcal{X}_j}\left(\exp\left(\frac{1}{\eta}g_j^{t+1}(x_j)\right) - \exp\left(\frac{1}{\eta}g_j^t(x_j)\right)\right)^2 G(g_{A\setminus\{i,j\}}^{t+1,j-1})(x_i,x_j)\right\} \\
&\quad \times \left(\sum_{A\in S_k^L(i,j)}\sum_{x_j\in\mathcal{X}_j} G(g_{A\setminus\{i,j\}}^{t+1,j-1})(x_i,x_j)\right) \\
&\leqslant \frac{e^\Gamma}{\eta^2}\exp\left(\frac{1}{\eta}g_i^t(x_i)\right)^2\left\{\sum_{A\in S_k^L(i,j)}\sum_{x_j\in\mathcal{X}_j}\left(g_j^{t+1}(x_j) - g_j^t(x_j)\right)^2 G(g_{A\setminus\{i,j\}}^{t+1,j-1})(x_i,x_j)\right\} \\
&\quad \times \left(\sum_{A\in S_k^L(i,j)}\sum_{x_j\in\mathcal{X}_j} G(g_{A\setminus\{i,j\}}^{t+1,j-1})(x_i,x_j)\right).
\end{aligned}
$$

Along the above lines, it follows that

$$
\begin{aligned}
&\|\partial_{g_i}\mathcal{L}^L(g^{t+1,j}) - \partial_{g_i}\mathcal{L}^L(g^{t+1,j-1})\|^2 \\
&\leqslant \frac{e^\Gamma}{\eta^2}\sum_{x_i\in\mathcal{X}_i}\exp\left(\frac{1}{\eta}g_i^t(x_i)\right)^2\left\{\sum_{A\in S_K^L(i,j)}\sum_{x_j\in\mathcal{X}_j}\left(g_j^{t+1}(x_j)-g_j^t(x_j)\right)^2 G(g_{A\setminus\{i,j\}}^{t+1,j-1})(x_i,x_j)\right\} \\
&\quad\times\left(\sum_{A\in S_K^L(i,j)}\sum_{x_j\in\mathcal{X}_j} G(g_{A\setminus\{i,j\}}^{t+1,j-1})(x_i,x_j)\right) \\
&= \frac{e^\Gamma}{\eta^2}\sum_{x_i\in\mathcal{X}_i}\sum_{A\in S_K^L(i,j)}\sum_{x_j\in\mathcal{X}_j}\left(g_j^{t+1}(x_j)-g_j^t(x_j)\right)^2\exp\left(\frac{1}{\eta}g_i^t(x_i)\right) G(g_{A\setminus\{i,j\}}^{t+1,j-1})(x_i,x_j) \\
&\quad\times\exp\left(\frac{1}{\eta}g_i^t(x_i)\right)\left(\sum_{A\in S_K^L(i,j)}\sum_{x_j\in\mathcal{X}_j} G(g_{A\setminus\{i,j\}}^{t+1,j-1})(x_i,x_j)\right).
\end{aligned}
$$

Observe that

$$
\begin{aligned}
&\sum_{x_i\in\mathcal{X}_i}\sum_{A\in S_K^L(i,j)}\exp\left(\frac{1}{\eta}g_i^t(x_i)\right) G(g_{A\setminus\{i,j\}}^{t+1,j-1})(x_i,x_j) \\
&= \sum_{A\in S_K^L(j)}\exp\left(-\frac{1}{\eta}g_j^{t+1}(x_j)\right)(\mathcal{P}_{i\#}\pi_A^{t+1,j})(x_j).
\end{aligned}
$$

Also, since $\exp\left(\frac{1}{\eta}g_i^t(x_i)\right)\left(\sum_{A\in S_K^L(i,j)}\sum_{x_j\in\mathcal{X}_j} G(g_{A\setminus\{i,j\}}^{t+1,j-1})(x_i,x_j)\right) > 0$,

$$
\begin{aligned}
&\exp\left(\frac{1}{\eta}g_i^t(x_i)\right)\left(\sum_{A\in S_K^L(i,j)}\sum_{x_j\in\mathcal{X}_j} G(g_{A\setminus\{i,j\}}^{t+1,j-1})(x_i,x_j)\right) \\
&\leqslant \sum_{x_i\in\mathcal{X}_i}\exp\left(\frac{1}{\eta}g_i^t(x_i)\right)\left(\sum_{A\in S_K^L(i,j)}\sum_{x_j\in\mathcal{X}_j} G(g_{A\setminus\{i,j\}}^{t+1,j-1})(x_i,x_j)\right) \\
&= \sum_{x_j\in\mathcal{X}_j}\sum_{A\in S_K^L(j)}\exp\left(-\frac{1}{\eta}g_j^{t+1}(x_j)\right)(\mathcal{P}_{i\#}\pi_A^{t+1,j})(x_j).
\end{aligned}
$$

Combining (5.4.11) and (5.4.18) with the above two computations, we have

$$\begin{aligned}
&\|\partial_{g_i}\mathcal{L}^L(g^{t+1,j}) - \partial_{g_i}\mathcal{L}^L(g^{t+1,j-1})\|^2 \\
&= \frac{e^{\Gamma}}{\eta^2}\sum_{x_j\in\mathcal{X}_j}\sum_{A\in S_K^L(j)}\left(g_j^{t+1}(x_j) - g_j^t(x_j)\right)^2 \exp\left(-\frac{1}{\eta}g_j^{t+1}(x_j)\right)(\mathcal{P}_{i\#}\pi_A^{t+1,j})(x_j) \\
&\quad\times\left(\sum_{x_j\in\mathcal{X}_j}\sum_{A\in S_K^L(j)}\exp\left(-\frac{1}{\eta}g_j^{t+1}(x_j)\right)(\mathcal{P}_{i\#}\pi_A^{t+1,j})(x_j)\right) \\
&= \frac{e^{\Gamma(1-\frac{2}{\eta})}}{\eta^2}\sum_{x_j\in\mathcal{X}_j}\left(g_j^{t+1}(x_j) - g_j^t(x_j)\right)^2\mu_j(x_j). \qquad (5.4.21)
\end{aligned}$$

One can check that (5.4.21) holds for $j \geqslant i$ cases. As a result,

$$\begin{aligned}
\|\partial_{g_i}\mathcal{L}^L(g^t)\|^2 &\leqslant \sum_{j=1}^{K}\|\partial_{g_i}\mathcal{L}^L(g^{t+1,j}) - \partial_{g_i}\mathcal{L}^L(g^{t+1,j-1})\|^2 \\
&\leqslant \frac{e^{\Gamma(1-\frac{2}{\eta})}}{\eta^2}\sum_{j=1}^{K}\sum_{x_j\in\mathcal{X}_j}\left(g_j^{t+1}(x_j) - g_j^t(x_j)\right)^2\mu_j(x_j).
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathcal{L}^L(g^*) - \mathcal{L}^L(g^t) &\leqslant \frac{2\eta}{e^{-(\Gamma+\frac{1}{\eta})}}\|\partial_{g_i}\mathcal{L}^L(g^t)\|^2 \\
&\leqslant \frac{2Ke^{\Gamma(1-\frac{2}{\eta})}}{\eta}\sum_{j=1}^{K}\sum_{x_j\in\mathcal{X}_j}\left(g_j^{t+1}(x_j) - g_j^t(x_j)\right)^2\mu_j(x_j).
\end{aligned}$$

Applying Lemma 5.14, finally one obtains

$$\mathcal{L}^L(g^*) - \mathcal{L}^L(g^t) \leqslant \frac{2Ke^{2\Gamma(1-\frac{1}{\eta})+\frac{1}{\eta}}}{\eta}\left(\mathcal{L}^L(g^{t+1}) - \mathcal{L}^L(g^t)\right).$$

(5.4.19), therefore, immediately follows from the above. □

Alternative Truncated Entropic Regularization

Algorithm 4 may have some issue which is that c_A is not bounded. Since c_A is either 0 or ∞, it impedes not only theoretical understanding of this algorithm but also implementation in practice. At the same time, we already know that the value of (3.2.6) is bounded by 1 so that c_A never attains ∞ at the optimum. With this hint, we can further reformulate (5.4.9) replacing $1 + c_A$ by a bounded cost function.

Recall the generalized barycenter problem (3.1.4). With the choice of a cost function $c = c_\varepsilon$ as in (2.8), it has a solution. More generally, if c satisfies Assumption 2.5, (3.1.4) always has a solution: see Proposition 3.4 in chapter 3.

Define a new cost function

$$B^*_A(x_A) := B^*_{\widehat{\mu}_{x_A}} \tag{5.4.22}$$

where $\widehat{\mu}_{x_A}$ is a non-negative measure defined as

$$\widehat{\mu}_{x_A} := \sum_{i \in A} \delta_{x_i}. \tag{5.4.23}$$

for given given $x_A \in \mathcal{X}^A$. In words, $B^*_A(x_A)$ is the value of generalized barycenter problem regarding the input distribution given by $\widehat{\mu}_{x_A}$. Trivially, $\widehat{\mu}_{x_A} = 0$ when $A = \emptyset$.

Proposition 5.17 justifies the replacement of $(1 + c_A)$ by $B^*_A(x_A)$.

Proposition 5.17. *Recall* $B^*_A(x_A)$ *defined as* (5.4.22). *Define*

$$\begin{aligned} &\min_{\{\pi_A\}_{A \in S_K}} \sum_{A \in S_K} \int_{\mathcal{X}^K} B^*_A(x_A) \, d\pi_A(x_1, ..., x_K) \\ &\text{s.t.} \sum_{A \in S_K(i)} \mathcal{P}_{i\#}\pi_A = \mu_i \quad \textit{for all } i \in \mathcal{Y}. \end{aligned} \tag{5.4.24}$$

Then,

$$(3.2.6) = (5.4.24).$$

Furthermore, let $\{\widehat{\pi}_A\}_{A\in S_K}$ *be a solution for* (5.4.24)*. Then, we can obtain a solution* $\{\pi^*_A\}_{A\in S_K}$ *for* (3.2.6) *from* $\{\widehat{\pi}_A\}_{A\in S_K}$*.*

Proof. First, it follows from (5.4.28) and (5.4.29) that

$$B^*_A(x_A) \leqslant 1 + c_A(x_A).$$

Hence, (3.2.6) $\geqslant$ (5.4.24) is direct.

For proving the other direction, let $\{\widehat{\pi}_A\}_{A\in S_K}$ be a solution for (5.4.24). Observe that for each x_A, according to (5.4.29) we can find a collection of $\{u^*_B\}_{B\neq\emptyset\subseteq A}$ which satisfies

$$B^*_A(x_A) = \sum_{B\neq\emptyset\subseteq A} (1 + c_B(x_B))\, u^*_B(x_A). \tag{5.4.25}$$

Fixing arbitrary feasible $\{\pi_A\}_{A\in S_K}$,

$$\begin{aligned}&\sum_{A\in S_K}\int_{\mathcal{X}^K} B^*_A(x_A)\mathrm{d}\pi_A(x_1,...,x_K)\\ &= \sum_{A\in S_K}\int_{\mathcal{X}^K}\left(\sum_{B\neq\emptyset\subseteq A}(1 + c_B(x_B))\, u^*_B(x_A)\right)\mathrm{d}\pi_A(x_1,...,x_K).\end{aligned}$$

Note that for $\mathrm{d}\pi_A(x_1,...,x_K)$, since x_j with $j\notin A$ is a dummy variable, $\mathrm{d}\pi_A(x_1,...,x_K) = \mathrm{d}\pi_A(x_A)$. For each $B\in S_K$, we can define

$$\mathrm{d}\pi^*_B := \sum_{A:B\subseteq A}\int_{\mathcal{X}^{A\setminus B}} u^*_A(x_{A\setminus B}, x_B)\mathrm{d}\pi_A(x_{A\setminus B}, x_B). \tag{5.4.26}$$

Here, $\int_{\mathcal{X}^{A\setminus B}} u^*_A(x_{A\setminus B}, x_B)\mathrm{d}\pi_A(x_{A\setminus B}, x_B)$ is understood as the integration with respect to the coordinates $x_{A\setminus B}$ given x_B and $x_\emptyset = 1$. Note that this integration is well-defined since u^* is Borel measurable. Changing the role of A and B, it follows

that

$$\begin{aligned}
&\sum_{A\in S_K}\int_{\mathcal{X}^K} B^*_A(x_A)\,d\pi_A(x_1,...,x_K)\\
&\geqslant \sum_{A\in S_K}\int_{\mathcal{X}^K}\left(\sum_{B\neq\emptyset\subseteq A}(1+c_B(x_B))\,u^*_B(x_A)\right)d\pi_A(x_1,...,x_K)\\
&=\sum_{B\in S_K}\int_{\mathcal{X}^B}(1+c_B(x_B))\left(\sum_{A:B\subseteq A}\int_{\mathcal{X}^{A\setminus B}}u^*_A(x_{A\setminus B},x_B)\,d\pi_A(x_{A\setminus B},x_B)\right)\\
&=\sum_{B\in S_K}\int_{\mathcal{X}^K}(1+c_B(x_B))\,d\pi^*_B(x_B).
\end{aligned}$$

Since the inequality holds for any feasible $\{\pi_A\}_{A\in S_K}$, this yields (3.2.6) $\leqslant$ (5.4.24).

Lastly, it remains to show that $\{\pi^*_A\}_{A\in S_K}$ is feasible for (3.2.6). In other words, for each $i\in\mathcal{Y}$,

$$\sum_{B\in S_K(i)}\int_{\mathcal{X}^{B\setminus\{i\}}}\sum_{A:B\subseteq A}\int_{\mathcal{X}^{A\setminus B}}u^*_A(x_{A\setminus B},x_B)\,d\pi_A(x_{A\setminus B},x_B)=d\mu_i(x_i).$$

Notice that for any $A\in S_K$ and any $x_A\in\mathcal{X}^A$

$$\sum_{B\in A(i)}u^*_B(x_A)=1$$

due to the definition of (5.4.28). Also, $\{\pi_A\}_{A\in S_K}$ satisfy $\sum_{A\in S_K(i)}\mathcal{P}_{i\#}\pi_A=\mu_i$.

Hence, a similar calculation as above provides

$$\begin{aligned}
d\mu_i(x_i) &= \sum_{A\in S_K(i)} \int_{\mathcal{X}^{A\setminus\{i\}}} d\pi_A(x_{A\setminus\{i\}}|x_i) \\
&= \sum_{A\in S_K(i)} \int_{\mathcal{X}^{A\setminus\{i\}}} \left(\sum_{B\in\mathcal{A}(i)} u_B^*(x_A) \right) d\pi_A(x_{A\setminus\{i\}}, x_i) \\
&= \sum_{B\in S_K(i)} \int_{\mathcal{X}^{B\setminus\{i\}}} \sum_{A:B\subseteq\mathcal{A}} \int_{\mathcal{X}^{A\setminus B}} u_B^*(x_{A\setminus B}, x_{B\setminus\{i\}}, x_i) d\pi_A(x_{A\setminus B}, x_{B\setminus\{i\}}, x_i) \\
&= \sum_{B\in S_K(i)} \int_{\mathcal{X}^{B\setminus\{i\}}} \sum_{A:B\subseteq\mathcal{A}} \int_{\mathcal{X}^{A\setminus B}} u_B^*(x_{A\setminus B}, x_{B\setminus\{i\}}, x_i) d\pi_A(x_{A\setminus B}, x_{B\setminus\{i\}}, x_i).
\end{aligned}$$

It verifies that $\{\pi_A^*\}_{A\in S_K}$ is feasible for (3.2.6). □

Thanks to Proposition 5.17, we can propose a slight variation of truncated Sinkhorn iteration. Here, we follow the fashion stated in remark 5.10 for diversity: in other words, this algorithm iterates over the space of exponential of dual variables, ϕ_i's. Of course, taking the logarithm, you can boil down the algorithm to the variables g_i's as usual.

The final update rule of $\phi_i = \exp\left(\frac{1}{\eta} g_i\right)$ is

$$\phi_i(x_i) = \mu_i(x_i) \Bigg/ \sum_{A\in S_K^L(i)} \left\langle \exp\left(-\frac{1}{\eta} B_A^*(x_i, \cdot)\right), \bigotimes_{j\in A\setminus\{i\}} \phi_j(\cdot) \right\rangle$$

and the optimal value of this iteration is written as

$$\sum_{A\in S_K^L} \sum_{\mathcal{X}^A} (1 + c_A(x_A)) \pi_A^*(x_A) = \sum_{A\in S_K^L} \sum_{\mathcal{X}^A} B_A^*(x_A) \widehat{\pi}_A(x_A).$$

Notice that if one only needs to compute the optimal adversarial risk, the step of transformation, defined in (5.4.26), is redundant due to the above identity.

Pseudocode appears in Algorithm 6.

Algorithm 6 (Alternative) Truncated entropic regularization(Sinkhorn)

Input: X : data set, $\eta > 0$: entropic parameter, L : trucation level, $\mu = (\mu_1, \dots, \mu_K)$: empirical distribution, ε : adversarial budget.

Initialization : $\phi_i = \mu_i$. Compute $\{B^*_A : A \in S^L_K\}$.

while not converge **do**

$$\phi_1(x^1_{l_1}) \leftarrow \mu_1(x^1_{l_1}) \Big/ \sum_{A \in S^L_K(1)} \left\langle \exp\left(-\tfrac{1}{\eta} B^*_A(x^1_{l_1}, \cdot)\right), \bigotimes_{j \in A \setminus \{1\}} \phi_j(\cdot) \right\rangle,$$

$$\vdots$$

$$\phi_K(x^K_{l_K}) \leftarrow \mu_K(x^K_{l_K}) \Big/ \sum_{A \in S^L_K(K)} \left\langle \exp\left(-\tfrac{1}{\eta} B^*_A(x^K_{l_K}, \cdot)\right), \bigotimes_{j \in A \setminus \{K\}} \phi_j(\cdot) \right\rangle.$$

end while

Compute $\widehat{\pi}_A(x_A) = \prod_{i \in A} \phi_i(x^i_{l_i}) \exp\left(-\tfrac{1}{\eta} B^*_A(x_A)\right)$ for all $A \in S^L_K$.

Output: $\{\pi^*_A = T(\{\widehat{\pi}_A\}_{A \in S^L_K})\}_{A \in S^L_K}$ and value $= \sum_{A \in S^L_K} \sum_{\mathcal{X}^A} B^*_A(x_A) \widehat{\pi}_A(x_A)$.

Remark 5.18. *Of course, Algorithm 6 solves the following problem*

$$\begin{aligned} &\min_{\{\pi_A\}_{A \in S^L_K}} \sum_{A \in S^L_K} \int_{\mathcal{X}^K} B^*_A(x_1, \dots, x_K) d\pi_A(x_1, \dots, x_K) - \eta \left(Ent\left(\pi_A\right) + \|\pi_A\| \right) \\ &s.t. \sum_{A \in S^L_K(i)} \mathcal{P}_{i\#} \pi_A = \mu_i \quad \textit{for all } i \in \mathcal{Y}. \end{aligned} \tag{5.4.27}$$

Even though Proposition 5.17 holds, when adopting the entropic term, however, the replacement $(1 + c_A)$ *by* B^*_A *engenders the discordance between* (5.4.8) *and* (5.4.27).

Recall the transformation from π_A*'s to* π^*_A*:*

$$d\pi^*_B := \sum_{A : B \subseteq A} \int_{\mathcal{X}^{A \setminus B}} u^*_A(x_{A \setminus B}, x_B) d\pi_A(x_{A \setminus B}, x_B).$$

Hence, for each $A \in S^L_K$*, while the entropy terms of* (5.4.27) *are*

$$\sum_{\mathcal{X}^A} -\pi_A(x_A) \log \pi_A(x_A),$$

those of (5.4.8) *are*

$$\begin{aligned}
&\sum_{x^A} -\pi^*_A(x_A)\log\pi^*_A(x_A)\\
&= \sum_{x^A} -\left(\sum_{C:A\subseteq C}\sum_{x^{C\setminus A}} u^*_C(x_{C\setminus A}, x_B A)\pi_A C(x_{C\setminus A}, x_A\right)\\
&\qquad\qquad \times \log\left(\sum_{C:A\subseteq C}\sum_{x^{C\setminus A}} u^*_C(x_{C\setminus A}, x_B A)\pi_A C(x_{C\setminus A}, x_A\right).
\end{aligned}$$

Thus, (5.4.8) $\neq$ (5.4.27) *in general. But,* (5.4.27) *still gives a good approximation for the* L*-level truncated version of original problem due to Proposition 5.17. Also, one can prove that, as* $\eta \to 0$*, both* (5.4.8) *and* (5.4.27) *converge to the solution for* L*-truncated version of original problem which has the maximum entropy among all other solutions.*

Remark 5.19 (Some argument about the complexity of Algorithm 6)**.** *Although, we cannot fully characterize the complexity, we can give some argument regarding that of Algorithm 6, heuristically. First, for each* x_A *you use Algorithm 3 to compute* $B^*_A(x_A)$*. The computational complexity of* $B^*_A(x_A)$ *is at most* $O(L^3)$ *since the number of inputs is at most* L*. So, the total computational complexity to obtain* $\{B^*_A : A \in S^L_K\}$ *is* $O(K^L L^3 n^L)$*.*

At the last step of Algorithm6, we need to compute $\{\pi^*_A\}$ *through the transformation defined in* (5.4.26)*. There are two extreme cases which contribute the total computational complexity. One is the case that* $|A|$ *is very small, say a singleton. In this case, the computational complexity of computing the summation term is* $O(K^L n^{L-1})$ *since there are at most* $O(K^L)$ *many* C *including* A *and the inner sum requires* $O(n^{L-1})$ *arithmetic operations. Since there are* $O(n)$ *points related to* A*, the total complexity is* $O(K^L n^L)$*. The other is the case that* $|A|$ *is very large, say* $|A| = L-1$*. In this case, the computational complexity of computing the summation term is* $O(Kn)$ *and there are* $O(n^{L-1})$ *points so that the total complexity is* $O(Kn^L)$*. Hence, the computational complexity of computing* $\{\pi^*_A\}_{A\in S^L_K}$ *is* $O(K^{2L} n^L)$*.*

We want to emphasize that, for sure, Algorithm 6 has the same convergence rate of Algorithm 4 proved in Theorem 5.15.

Complexity of the full MOT Sinkhorn

In this section, we consider the complexity of the (full) MOT Sinkhorn used in section 5.2. To achieve its correct complexity, we will provide the simpler MOT formulation. In fact, the following argument can apply for a general cost function c provided that c satisfies Assumption 2.5.

Recall that $\mathcal{X}$ is a Polish space with a metric d and $\mathcal{Y}$ is the set of classes. We denote by 👻 the so-called ghost element introduced in chapter 3 and let $\mathcal{X}_* := \mathcal{X} \cup \{\text{👻}\}$. Let us define $\widehat{d}$ to $\mathcal{X}_*$ as

$$\widehat{d}(x, x') := \begin{cases} d(x, x') & \text{if } x \neq \text{👻} \text{ and } x' \neq \text{👻}, \\ 0 & \text{if } x = x' = \text{👻}, \\ \infty & \text{otherwise.} \end{cases}$$

As similar, one can prove that $(\mathcal{X}_*, \widehat{d})$ is again a Polish space: see Proposition 3.26.

Let us define a proper MOT cost function. Recall the generalized barycenter problem (3.1.4) and (5.4.23). One can regard $\widehat{\mu}_{x_A} = \sum_{i:x_i \neq \text{👻}} \delta_{x_i}$. Then, we are able to extend it to $\mathcal{X}_*^K$ in this way: for $\mathbf{x} = (x_1, \dots, x_K) \in \mathcal{X}_*^K$,

$$\widehat{\mu}_{\mathbf{x}} = \sum_{i:x_i \neq \text{👻}} \delta_{x_i}.$$

Now, we define the MOT cost function as

$$\mathbf{c}(x_1, \dots, x_K) := B^*_{\widehat{\mu}_{\mathbf{x}}}.$$

If $x_1 = \cdots = x_K = \text{👻}$, $B^*_{\widehat{\mu}_{\mathbf{x}}} = 0$ trivially. Suppose $A = A(x_1, \dots, x_K) \neq \emptyset$. Let us use $A(i) := \{B \subseteq A : i \in B\}$. Reformulating $B^*_{\widehat{\mu}_{\mathbf{x}}}$ by the following equivalent

multimarginal problem which is proved in Proposition 3.11 and 3.12,

$$B^*_{\widehat{\mu}_x} = \inf_{\pi_B : B \neq \emptyset \subseteq A} \left\{ \sum_{B \neq \emptyset \subseteq A} \int_{\mathcal{X}^K} (1 + c_B(\boldsymbol{x}))\, d\pi_B(\boldsymbol{x}) : \sum_{B \in A(i)} \mathcal{P}_{i\#}\pi_B = \delta_{x_i} \text{ for all } i \in A \right\}. \tag{5.4.28}$$

In addition, under Assumption 2.6, (5.4.28) always attains a solution $\{\pi^*_B\}_{B \neq \emptyset \subseteq A}$ which satisfies $\mathrm{spt}(\pi^*_B) \subseteq \prod_{i \in B}\{x_i\}$ for any non-empty $B \subseteq A$ due to its constraints. So, one can characterize $\mathbf{c}(x_1, \ldots, x_K)$ as

$$\mathbf{c}(x_1, \ldots, x_K) = \sum_{B \neq \emptyset \subseteq A} (1 + c_B(\boldsymbol{x}))\, u^*_B(\boldsymbol{x}) \tag{5.4.29}$$

such that $u^*_B \geqslant 0$ for all B and for each $i \in A$,

$$\sum_{B \in A(i)} u^*_B(\boldsymbol{x}) = \delta_{x_i}$$

In the following lemma, we prove that $\mathbf{c}(x_1, \ldots, x_K)$ is lower semi-continuous under an appropriate topology on $\mathcal{X}^K_*$.

Lemma 5.20. *Define $\widehat{d}_K$ on $\mathcal{X}^K_*$ by*

$$\widehat{d}_K((x_1, \ldots, x_K), (x'_1, \ldots, x'_K)) := \max_{1 \leqslant i \leqslant K} \widehat{d}(x_i, x'_i).$$

Then, under Assumption 2.6, $\mathbf{c}$ is lower semi-continuous on $(\mathcal{X}^K_*, \widehat{d}_K)$.

Remark 5.21. *Note that $(\mathcal{X}^K_*, \widehat{d}_K)$ is still a Polish space.*

Proof. Suppose $\boldsymbol{x}^n = (x^n_1, \ldots, x^n_K)$ converges to $\boldsymbol{x} = (x_1, \ldots, x_K)$ in $(\mathcal{X}^K_*, \widehat{d}_k)$. Without loss of generality, assume that $x_1, \ldots, x_L = \text{👻}$ for all $1 \leqslant L \leqslant K$. If $L = K$, $\mathbf{c} = 0$ and the claim would be trivial. So we focus on the case $L < K$. By the definition of $\widehat{d}_K$, without loss of generality we can further assume that $x^n_1, \ldots, x^n_L = \text{👻}$ for all n, and likewise, for each $L + 1 \leqslant j \leqslant K$, we can assume that $i^n_j = i_j$ for all n, for otherwise the convergence would not hold due to the definition of $\widehat{d}_K$.

Let $A = \{L+1, \dots, K\}$ and $A(i) = \{B \subseteq A : i \in B\}$. We now claim that for every $B \neq \emptyset \subseteq A$,

$$c_B(x_{L+1}, \dots, x_K) \leqslant \liminf_{n\to\infty} c_B(x^n_{L+1}, \dots, x^n_K).$$

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

$$\liminf_{n\to\infty} \sum_{i\in B} c(\widetilde{x}^n, x^n_i) = \liminf_{n\to\infty} c_B(x^n_{L+1}, \dots, x^n_K) < \infty.$$

By Assumption 2.6, up to subsequence (not relabeled) we can find a sequence $\{\widetilde{x}^n\}_{n\in\mathbb{N}}$ converging toward a point $\widetilde{x} \in \mathcal{X}$. Combining with the lower semi-continuity of c, we deduce that

$$c_B(x_{L+1}, \dots, x_K) \leqslant \sum_{i\in B} c(\widetilde{x}, \mathcal{X}_i) \leqslant \liminf_{n\to\infty} c_B(x^n_{L+1}, \dots, x^n_K).$$

Since (5.4.29) holds, we can find for each $n \in \mathbb{N}$ a collection of feasible $\{\pi^n_B\}_{B\neq\emptyset\subseteq A}$ such that

$$\liminf_{n\to\infty} \sum_{B\neq\emptyset\subseteq A} \left(1 + c_B(x^n_{L+1}, \dots, x^n_K)\right) \pi^n_B = \liminf_{n\to\infty} \mathbf{c}(x^n_1, \dots, x^n_K).$$

Using the Heine-Borel theorem in Euclidean space, we can assume without the loss of generality that for every $B \neq \emptyset \subseteq A$, π^n_B converges to some π_B as $n \to \infty$. The resulting collection of π_B is feasible for the problem defining $\mathbf{c}(x_1, \dots, x_K)$ and thus, using the lower semi-continuity of c_B established earlier, we deduce:

$$\mathbf{c}(x_1, \dots, x_K) \leqslant \sum_{B\neq\emptyset\subseteq A} (1 + c_B(x_{L+1}, \dots, x_K))\pi_B \leqslant \liminf_{n\to\infty} \mathbf{c}(z^n_1, \dots, z^n_K).$$

□

In order to describe of the primal MOT problem and its the dual formally, we

need some preliminary works. For each $i \in \mathcal{Y}$, let $\widehat{\mu}_i$ be the extension of μ_i to $\mathcal{X}_*$ as

$$d\widehat{\mu}_i := d\mu_i + (1 - \|\mu_i\|)\delta_{\text{Ω}}.$$

Here, $\|\mu_i\|$ is the total mass of μ_i. By the construction of $\widehat{\mu}_i$'s, all of them have the same mass, 1, allowing the application of the usual MOT theory directly.

Let us define the set of couplings by

$$\Pi(\widehat{\mu}_1, \dots, \widehat{\mu}_K) := \left\{\pi \in \mathcal{P}(\mathcal{X}_*^K) : \mathcal{P}_{i\#}\pi = \widehat{\mu}_i \text{ for all } i \in \mathcal{Y}\right\} \tag{5.4.30}$$

and the set of feasible potentials by

$$\Phi := \left\{ \phi = (\phi_1, \dots, \phi_K) \in \prod_{j=1}^{K} L^1(\widehat{\mu}_i) : \sum_{i \in \mathcal{Y}} \phi_i(x_i) \leqslant \mathbf{c}(x_1, \dots, x_K) \right\}. \tag{5.4.31}$$

Proposition 5.22. *Under Assumption 2.6, we have*

$$(3.2.6) = \inf_{\pi \in \Pi(\widehat{\mu}_1, \dots, \widehat{\mu}_K)} \left\{ \int_{\mathcal{X}_*^K} \mathbf{c}(x_1, \dots, x_K) d\pi(x_1, \dots, x_K) \right\}. \tag{5.4.32}$$

Furthermore, it satisfies the duality which is

$$\inf_{\pi \in \Pi(\widehat{\mu}_1, \dots, \widehat{\mu}_K)} \left\{ \int_{\mathcal{X}_*^K} \mathbf{c}(x_1, \dots, x_K) d\pi(x_1, \dots, x_K) \right\} = \sup_{\phi \in \Phi} \left\{ \sum_{i \in \mathcal{Y}} \int_{\mathcal{X}_*} \phi_i(x_i) d\widehat{\mu}_i(x_i) \right\} \tag{5.4.33}$$

and both of the primal and the dual have solutions.

Proof. (5.4.33) is straightforward thanks to the fact that $\mathbf{c}(x_1, \dots, x_K)$ is lower semi-continuous and bounded below by 0: see (Villani, 2009, Theorem 5.10). The existence of a minimizer for a primal problem is also classic and the existence of a maximizer for dual problem follows from (5.4.32) and Proposition 4.18.

Let $g = (g_1, \dots, g_K) \in \mathcal{C}_b(\mathcal{X})^K$ be a feasible solution for (3.2.13). Define $\widehat{g} :=$

$(\widehat{g}_1, \dots, \widehat{g}_K)$ as, for each $i \in \mathcal{Y}$

$$\widehat{g}_i(x) := \begin{cases} g_i(x) & \text{if } x \in \mathcal{X}, \\ 0 & \text{if } x = \text{👻}. \end{cases}$$

We want to show that for any $(x_1, \dots, x_K) \in \mathcal{X}_*^K$,

$$\sum_{i \in \mathcal{Y}} \widehat{g}_i(x_i) \leqslant \mathbf{c}(x_1, \dots, x_K).$$

First, if $x_1 = \cdots = x_K = \text{👻}$, it holds trivially. Let $A = \{i : x_i \neq \text{👻}\}$ and $A(i) = \{B \subseteq A : i \in B\}$. Combining (5.4.29) with the fact that g satisfies $\sum_{i \in A} g_i(x_i) \leqslant 1 + c_A(x_A)$ for any $A \subseteq S_K$, it follows that for any non-empty $B \subseteq A$,

$$\pi_B^* \sum_{i \in B} g_i(x_i) \leqslant (1 + c_B(x_B))\, \pi_B^* = \int_{\mathcal{X}^K} (1 + c_B(\boldsymbol{x}))\, d\pi_B^*(\boldsymbol{x}). \tag{5.4.34}$$

Using (5.4.29), (5.4.34), $\sum_{B \subseteq A(i)} \|\pi_B\| = 1$ for all $i \in A$ and $\sum_{i \in \mathcal{Y}} \widehat{g}(\mathcal{X}_i) = \sum_{i \in A} g_i(\mathcal{X}_i)$, one can deduce

$$\begin{aligned} \sum_{i \in \mathcal{Y}} \widehat{g}(\mathcal{X}_i) &= \sum_{B \neq \emptyset \subseteq A} \sum_{i \in B} \pi_B^* g_i(\mathcal{X}_i) \\ &\leqslant \sum_{B \neq \emptyset \subseteq A} \int_{\mathcal{X}^K} (1 + c_B(\boldsymbol{x}))\, d\pi_B^*(\boldsymbol{x}) \\ &= \mathbf{c}(x_1, \dots, x_K), \end{aligned}$$

which verifies that $\widehat{g} \in \Phi$. Applying Corollary 3.28 and 3.29, it follows that

$$(3.2.6) \leqslant \sup_{\phi \in \Phi} \left\{ \sum_{i \in \mathcal{Y}} \int_{\mathcal{X}_*} \phi_i(x_i) d\widehat{\mu}_i(x_i) \right\}.$$

On the other hand, given a dual potential $\phi \in \Phi$ which is feasible for the dual

problem, define $\psi := (\psi_1, \dots, \psi_K) \in \prod_{i \in \mathcal{Y}} L^1(\mathcal{X}; \mu_i)$ as, for each $i \in \mathcal{Y}$

$$\psi_i(x) := \phi_i(x) + \sum_{j \in \mathcal{Y}} \phi_j(\text{👻}) - \phi_i(\text{👻}).$$

It is straightforward that

$$\begin{aligned}
\sum_{i \in \mathcal{Y}} \int_{\mathcal{X}} \psi_i(\mathcal{X}_i) d\mu_i(\mathcal{X}_i) &= \sum_{i \in \mathcal{Y}} \int_{\mathcal{X}} \left(\phi_i(\mathcal{X}_i) + \sum_{j \in \mathcal{Y}} \phi_j(\text{👻}) - \phi_i(\text{👻}) \right) d\mu_i(\mathcal{X}_i) \\
&= \sum_{i \in \mathcal{Y}} \left\{ \int_{\mathcal{X}} \phi_i(\mathcal{X}_i) d\mu_i(\mathcal{X}_i) + \|\mu_i\| \left(\sum_{j \in \mathcal{Y}} \phi_j(\text{👻}) - \phi_i(\text{👻}) \right) \right\} \\
&= \sum_{i \in \mathcal{Y}} \int_{\mathcal{X}_*} \phi_i(\mathcal{X}_i) d\widehat{\mu}_i(\mathcal{X}_i).
\end{aligned}$$

Fix $A \in S_K$. Consider

$$\sum_{i \in A} \psi_i(\mathcal{X}_i) = \sum_{i \in A} \phi_i(\mathcal{X}_i) + |A| \sum_{j \in \mathcal{Y}} \phi_j(\text{👻}) - \sum_{i \in A} \phi_i(\text{👻})$$

Letting $\phi_j^k(\text{👻}) := \phi_j(\text{👻})$, k-th copy of $\phi_j(\text{👻})$, rewrite the second term as

$$|A| \sum_{j \in \mathcal{Y}} \phi_j(\text{👻}) = \sum_{i \in A} \sum_{j \in \mathcal{Y}} \phi_j^k(\text{👻}).$$

Fix $l \in A$. By (5.4.28) and the definition of ϕ,

$$\begin{aligned}
\sum_{i \in A} \phi_i(\mathcal{X}_i) + \sum_{j \in \mathcal{Y}} \phi_j^l(\text{👻}) - \sum_{i \in A} \phi_i(\text{👻}) &= \sum_{i \in A} \phi_i(\mathcal{X}_i) + \sum_{j \in A^c} \phi_j^l(\text{👻}) \\
&\leqslant \mathbf{c}(x_1, \dots, x_K) \\
&\leqslant 1 + c_A(x_A).
\end{aligned}$$

Hence, since $\sum_{j\in\mathcal{Y}} \phi_j(\text{👻}) \leqslant 0$,

$$\begin{aligned}\sum_{i\in\mathcal{A}} \psi_i(\mathcal{X}_i) &= \sum_{i\in\mathcal{A}} \phi_i(\mathcal{X}_i) + \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{Y}} \phi_j^k(\text{👻}) - \sum_{i\in\mathcal{A}} \phi_i(\text{👻}) \\ &= \sum_{i\in\mathcal{A}} \phi_i(\mathcal{X}_i) + \sum_{j\in\mathcal{Y}} \phi_j^l(\text{👻}) - \sum_{i\in\mathcal{A}} \phi_i(\text{👻}) + \sum_{k\neq l\in\mathcal{A}}\sum_{j\in\mathcal{Y}} \phi_j^k(\text{👻}) \\ &\leqslant 1 + c_A(x_A),\end{aligned}$$

which verifies that ψ is feasible for (3.2.13) in $\prod_{i\in\mathcal{Y}} L^1(\mu_i)$-sense. This yields

$$\sup_{\phi\in\Phi}\left\{\sum_{i\in\mathcal{Y}}\int_{\mathcal{X}_*} \phi_i(\mathcal{X}_i)\mathrm{d}\widehat{\mu}_i(\mathcal{X}_i)\right\} \leqslant (3.2.6).$$

The conclusion follows. □

This new MOT formula (5.4.32) gives a correct upper bound of the complexity of approximated solution for adversarial training.

Corollary 5.23. *Assume that* $n_i = O(n)$ *for all* $i \in \mathcal{Y}$*. The computational complexity of finding a* δ*-approximate solution for the adversarial training problem is at most*

$$O\left(\frac{K^4 n^{K+\frac{1}{2}}\log n}{\delta} \wedge \frac{K^5 n^K \log n}{\delta^2}\right)$$

Proof. Recall

$$(2.5) = 1 - (3.2.6).$$

With (5.4.32) of Proposition 5.22, the conclusion follows from Tupitsa et al. (2020); Lin et al. (2022). Note that $\|\mathbf{c}\|_\infty = K$. □

Remark 5.24. *Notice that each marginal* μ *of the previous MOT formula in García Trillos et al. (2023) has the size* $O(Kn)$*. Then, the computational complexity corresponding the*

previous MOT formula is

$$O\left(\frac{K^{K+\frac{7}{2}}n^{K+\frac{1}{2}}\log(Kn)}{\delta} \wedge \frac{K^{K+4}n^{K}\log(Kn)}{\delta^2}\right).$$

6 CONCLUSION AND FUTURE WORKS

My thesis provide both theoretical understanding and practical implementations regarding the adversarial training problem in multiclass classification. The hammer we have used to tackle this problem is optimal transport, especially multimarginal optimal transport which has been recently payed attention from pure mathematics to applied fields.

More formally, we have three adversarial training models, DRO model (2.5), closed-ball model (2.4) and open-ball model (2.3), which also has an exciting connection to "total-variation regularization". The main object is DRO model intensively studied in chapter 3. The main theme of chapter 3 is that DRO model has twofold: the adversary's part and the learner's part. The adversary's part is equivalent to the generalized barycenter problem (3.1.4), in other words, solving the generalized barycenter problem is sufficient from the perspective of the adversary. Based on this observation, we derive the stratified MOT problem (3.2.6) which in turn connects everything to optimal transport type formulations. From the dual of the stratified MOT problem, (3.2.13), we can construct a robust classifier, hence, it is indeed a problem for the learner.

Then, in chapter 4, first of all we analyze the learner's problem more rigorously. We prove that DRO model has a Borel measurable optimal robust classifier in a very general setting, as a result, we can prove the existence of a saddle point of DRO model. Based on that, we give a connection between DRO model and closed-ball model again by the argument of optimal transport theory. Finally, a simple but critical observation allows us to conclude that indeed closed-ball model and open-ball model are almost equivalent: hence the conclusion is that three models are equivalent in some sense.

These theories guide new algorithms to solve the adversarial problems numerically, which is the main theme of chapter 5. We propose two algorithms, exact solving(Algorithm 3) and truncated entropic regularization(Algorithm 4). More precisely, exact solving solves the generalized barycenter problem and truncated

entropic regularization solves both the stratified MOT problem and its dual. The idea of them is that under well separability of classes and the reasonable adversarial budget, problematic higher-order interactions, which is indeed a barrier of solving the problem, are really rare in real data set. We provide some numerical results obtained by them supporting the success of our algorithms. Also, they show how our heuristic belief can be justified in real data set.

There are really exciting and demanding remained questions in this field. First, it would be of interest to use (2.5) to help in the training of robust classifiers within specific families of models. Notice that (2.5) 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 (2.5) 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 (2.5) is still meaningful for the above model-specific problem: 1) the optimal $\widetilde{\mu}^*$ computed from the problem (2.5) 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 (2.5) can serve as a benchmark for robust training within *any* family of models.

Another question is more geometric one. For an optimal Borel robust classifier f^*, can we say some regularity of f^* as in Bungert et al. (2023)? Since we do not have a hard classifier in general, a question might be about the level set of f^*. For instance, letting $A_i = \{x \in \mathcal{X} : f_i^* = 1\}$, is this set regular in some sense? Or, is there an optimal Borel robust classifier whose level set is regular? Answering this question will give a more rigorous understanding of robustness or regularity of classifiers in adversarial training problem.

A next one is about the extension of our framework to other settings. In this paper, we assume that the loss function is 0-1 loss but practitioners prefer convex

loss functions, for example cross entropy, for faster optimization. Can we obtain a similar DRO-type formulation with non-linear loss functions? If so, it helps expand rigorous understanding of adversarial training and develop more tractable and accurate numerics for practice.

About algorithms, definitely one of the most important questions is the sample complexity. Note that even for W_2-Wasserstein metric and its entropic regularization, their sample complexity is not trivial to understand: see Manole and Niles-Weed (2021); Niles-Weed and Bach (2019); Harchaoui et al. (2022). One big potential issue of the adversarial training problem is the singularity of cost function. Unlike ℓ^2 norm cost function, a typical cost function in our setting is 0-∞ cost function which is really singular. Also, with this cost function, c_A appearing in the stratified MOT problem is not uniquely defined. Such properties are potential barriers against this question. Furthermore, it is well-known that many optimal transport problems exhibit the curse of dimensionality. We also expect that dimension really matters in the adversarial training.

Also, can we obtain a better complexity of Algorithm 3 and Algorithm 4? For example, using some elementary statistics of data sets, can we a priori choose an appropriate interaction-order level L to implement them faster and still obtain a good approximation? What is the correct notion of separation among different classes? Also, we reasonably guess that adopting greedy descent method and the rounding scheme suggested in Altschuler et al. (2017) helps obtain near-linear time$(O(K^L n^L))$ complexity.

Another practically important question is how to enforce a specific robust classifier or an adversarial attack by our algorithms. It is naturally expected that there are multiple Nash equilibria, in other words, multiple saddle points of robust classifiers and adversarial attacks. Suppose that we want some additional aspects of them: for example, among all possible robust classifiers, we want to pick a specific one which has the maximum entropy. Then, how can we implement such classifier? This is indeed a very practical issue but quite non-trivial task. Answering this questions would have an impact in reality for sure.

Finally, it is of interest to investigate the geometric content that profiles like the

ones presented in Figure 3.6, 5.2 and 5.3 carry about a specific data set. As illustrated, these plots are probably 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 and the geometry, separability or concentration etc., of given data set.

REFERENCES

Agueh, Martial, and Guillaume Carlier. 2011a. Barycenters in the Wasserstein space. *SIAM J. Math. Anal.* 43(2):904–924.

———. 2011b. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis* 43(2):904–924. https://doi.org/10.1137/100805741.

Altschuler, Jason, Jonathan Niles-Weed, and Philippe Rigollet. 2017. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *Advances in neural information processing systems* 30.

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

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

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

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

Balcan, Maria-Florina, Rattana Pukdee, Pradeep Ravikumar, and Hongyang Zhang. 2023. Nash equilibria and pitfalls of adversarial training in adversarial robustness games. In *Proceedings of the 26th international conference on artificial intelligence and statistics*, ed. Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, vol. 206 of *Proceedings of Machine Learning Research*, 9607–9636. PMLR.

Beck, Amir, and Luba Tetruashvili. 2013. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization* 23(4):2037–2060. https://doi.org/10.1137/120887679.

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

Benamou, Jean-David, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. 2015. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing* 37(2):A1111–A1138.

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

Bertozzi, Andrea L., and Arjuna Flenner. 2016. Diffuse interface models on graphs for classification of high dimensional data. *SIAM Review* 58(2):293–328.

Bertsekas, Dimitri, and Steven E Shreve. 1996. *Stochastic optimal control: the discrete-time case*, vol. 5. Athena Scientific.

Bhagoji, Arjun Nitin, Daniel Cullina, and Prateek Mittal. 2019. Lower bounds on adversarial robustness from optimal transport. In *Advances in neural information processing systems*, ed. H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, and R. Garnett, vol. 32. Curran Associates, Inc.

Biggio, Battista, and Fabio Roli. 2018. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition* 84:317–331.

Blanchet, Jose, and Karthyek Murthy. 2019. Quantifying distributional model risk via optimal transport. *Math. Oper. Res.* 44(2):565–600.

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

Bousquet, Olivier, Stéphane Boucheron, and Gábor Lugosi. 2004. *Introduction to statistical learning theory*, 169–207. Berlin, Heidelberg: Springer Berlin Heidelberg.

Boyd, Zachary M., Mason A. Porter, and Andrea L. Bertozzi. 2020. Stochastic block models are a discrete surface tension. *J. Nonlinear Sci.* 30(5):2429–2462.

Bungert, Leon, Nicolás García Trillos, and Ryan Murray. 2023. The geometry of adversarial training in binary classification. *Information and Inference: A Journal of the IMA* 12(2):921–968.

Bungert, Leon, and Kerrek Stinson. 2022. Gamma-convergence of a nonlocal perimeter arising in adversarial machine learning. *arXiv preprint arXiv:2211.15223*.

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

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

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

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

Carlier, Guillaume, Adam Oberman, and Edouard Oudet. 2015. Numerical methods for matching for teams and Wasserstein barycenters. *ESAIM Math. Model. Numer. Anal.* 49(6):1621–1642.

Caroccia, Marco, Antonin Chambolle, and Dejan Slepčev. 2020. Mumford–shah functionals on graphs and their asymptotics. *Nonlinearity* 33(8):3846.

Chen, Yongxin, Tryphon Georgiou, and Michele Pavon. 2016. Entropic and displacement interpolation: A computational approach using the hilbert metric. *SIAM Journal on Applied Mathematics* 76(6):2375–2396. `https://doi.org/10.1137/16M1061382`.

Chiappori, Pierre-André, Robert J. McCann, and Lars P. Nesheim. 2010. Hedonic price equilibria, stable matching, and optimal transport: equivalence, topology, and uniqueness. *Econom. Theory* 42(2):317–354.

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

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

Colombo, Maria, Luigi De Pascale, and Simone Di Marino. 2015. Multimarginal optimal transport maps for one-dimensional repulsive costs. *Canad. J. Math.* 67(2): 350–368.

Cotar, Codina, Gero Friesecke, and Claudia Klüppelberg. 2013. Density functional theory and optimal transportation with Coulomb cost. *Comm. Pure Appl. Math.* 66(4):548–599.

Cristofari, Andrea, Francesco Rinaldi, and Francesco Tudisco. 2020. Total variation based community detection using a nonlinear optimization approach. *SIAM J. Appl. Math.* 80(3):1392–1419.

Cuturi, Marco. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, ed. C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, vol. 26. Curran Associates, Inc.

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

Dai, Sihui, Wenxin Ding, Arjun Nitin Bhagoji, Daniel Cullina, Ben Y. Zhao, Haitao Zheng, and Prateek Mittal. 2023. Characterizing the optimal 0-1 loss for multi-class classification with a test-time attacker. 2302.10722.

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

Di Marino, Simone, and Augusto Gerolin. 2020. An optimal transport approach for the schrödinger bridge problem and convergence of sinkhorn algorithm. *Journal of Scientific Computing* 85(2):1–28.

Ekeland, Ivar. 2005. An optimal matching problem. *ESAIM Control Optim. Calc. Var.* 11(1):57–71.

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

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

Franklin, Joel, and Jens Lorenz. 1989. On the scaling of multidimensional matrices. *Linear Algebra Appl.* 114/115:717–735.

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

Garcıa Trillos, Nicolás, Matt Jacobs, and Jakwang Kim. 2023. The multimarginal optimal transport formulation of adversarial multiclass classification. *Journal of Machine Learning Research* 24(45):1–56.

García Trillos, Nicolas, Matt Jacobs, and Jakwang Kim. 2023a. On the existence of solutions to adversarial training in multiclass classification. 2305.00075.

García Trillos, Nicolas, Matt Jacobs, Jakwang Kim, and Matt Werenski. 2023b. Two approaches for computing adversarial training problem based on optimal transport frameworks. In preparation.

García Trillos, Nicolás, and Ryan Murray. 2017. A new analytical approach to consistency and overfitting in regularized empirical risk minimization. *European Journal of Applied Mathematics* 28(6):886–921.

García Trillos, Nicolás, Ryan Murray, and Matthew Thorpe. 2022. From graph cuts to isoperimetric inequalities: convergence rates of Cheeger cuts on data clouds. *Arch. Ration. Mech. Anal.* 244(3):541–598.

García Trillos, Nicolás, and Ryan W. Murray. 2022. Adversarial classification: Necessary conditions and geometric flows. *Journal of Machine Learning Research* 23(187):1–38.

van Gennip, Yves, Nestor Guillen, Braxton Osting, and Andrea L. Bertozzi. 2014. Mean curvature, threshold dynamics, and phase field theory on finite graphs. *Milan J. Math.* 82(1):3–65.

Goodfellow, Ian, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *International conference on learning representations*.

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

Harchaoui, Zaid, Lang Liu, and Soumik Pal. 2022. Asymptotics of discrete schrödinger bridges via chaos decomposition. `2011.08963`.

Hu, Huiyi, Thomas Laurent, Mason A. Porter, and Andrea L. Bertozzi. 2013. A method based on total variation for network modularity optimization using the MBO scheme. *SIAM J. Appl. Math.* 73(6):2224–2246.

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

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

Lin, Tianyi, Nhat Ho, Marco Cuturi, and Michael I Jordan. 2022. On the complexity of approximating multimarginal optimal transport. *Journal of Machine Learning Research* 23(65):1–43.

Luo, Xiyang, and Andrea L Bertozzi. 2017. Convergence of the graph allen–cahn scheme. *Journal of Statistical Physics* 167(3):934–958.

von Luxburg, Ulrike, and Bernhard Schölkopf. 2011. Statistical learning theory: Models, concepts, and results. In *Inductive logic*, ed. Dov M. Gabbay, Stephan Hartmann, and John Woods, vol. 10 of *Handbook of the History of Logic*, 651–706. North-Holland.

Luzin, Nikolaj Nikolaevič, and Wacław Sierpiński. 1919. Sur quelques propriétés des ensembles (a).

Manole, Tudor, and Jonathan Niles-Weed. 2021. Sharp convergence rates for empirical optimal transport with smooth costs. *arXiv preprint arXiv:2106.13181*.

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

Merkurjev, Ekaterina, Tijana Kostić, and Andrea L. Bertozzi. 2013. An mbo scheme on graphs for classification and image processing. *SIAM Journal on Imaging Sciences* 6(4):1903–1930.

Meunier, Laurent, Meyer Scetbon, Rafael B Pinot, Jamal Atif, and Yann Chevaleyre. 2021. Mixed nash equilibria in the adversarial examples game. In *Proceedings of the 38th international conference on machine learning*, ed. Marina Meila and Tong Zhang, vol. 139 of *Proceedings of Machine Learning Research*, 7677–7687. PMLR.

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

Niles-Weed, Jonathan, and Francis Bach. 2019. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli* 25(4A):2620–2648.

Nishiura, Togo. 2008. *Absolute measurable spaces*, vol. 120 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge.

Pass, Brendan. 2015. Multi-marginal optimal transport: theory and applications. *ESAIM Math. Model. Numer. Anal.* 49(6):1771–1790.

Peyré, Gabriel, Marco Cuturi, et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning* 11(5-6): 355–607.

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

———. 2021b. The many faces of adversarial risk. In *Advances in neural information processing systems*, ed. M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, vol. 34, 10000–10012. Curran Associates, Inc.

Seidl, Michael, Paola Gori-Giorgi, and Andreas Savin. 2007. Strictly correlated electrons in density-functional theory: A general formulation with applications to spherical densities. *Physical Review A* 75(4):042511.

Sinkhorn, Richard. 1964. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics* 35(2):876–879.

Sinkhorn, Richard, and Paul Knopp. 1967. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics* 21(2):343–348.

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

Tupitsa, Nazarii, Pavel Dvurechensky, Alexander Gasnikov, and César A Uribe. 2020. Multimarginal optimal transport by accelerated alternating minimization. In *2020 59th ieee conference on decision and control (cdc)*, 6132–6137. IEEE.

Villani, Cédric. 2003. *Topics in optimal transportation*, vol. 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI.

———. 2009. *Optimal transport*, vol. 338 of *Grundlehren der mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*]. Springer-Verlag, Berlin. Old and new.

Yen, Ian En-Hsu, Kai Zhong, Cho-Jui Hsieh, Pradeep K Ravikumar, and Inderjit S Dhillon. 2015. Sparse linear programming via primal and dual augmented coordinate descent. In *Advances in neural information processing systems*, ed. C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, vol. 28. Curran Associates, Inc.

Zhang, Yang, Hassan Foroosh, Philip David, and Boqing Gong. 2019. CAMOU: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In *International conference on learning representations*.