Journal of Machine Learning Research 25 (2024) 1-45 Submitted 2/24; Revised 11/24; Published 12/24

An Optimal Transport Approach for Computing Adversarial Training Lower Bounds in Multiclass Classification

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

Matt Jacobs MAJACO@UCSB.EDU
Department of Mathematics
UC Santa Barbara
552 University Rd, Isla Vista, CA 93117, USA

Jakwang Kim JAKWANG.KIM@MATH.UBC.CA
Department of Mathematics
University of British Columbia
1984 Mathematics Road, Vancouver, British Columbia, V6T 1Z2, Canada

Matthew Werenski MATTHEW.WERENSKI@TUFTS.EDU
Department of Computer Science
Tufts University
420 Joyce Cummings Center, 177 College Avenue, Medford, MA 02155, USA

Editor: Zaid Harchaoui

Abstract

Despite the success of deep learning-based algorithms, it is widely known that neural networks may fail to be robust. A popular paradigm to enforce robustness is *adversarial training* (AT), however, this introduces many computational and theoretical difficulties. Recent works have developed a connection between AT in the multiclass classification setting and multimarginal optimal transport (MOT), unlocking a new set of tools to study this problem. In this paper, we leverage the MOT connection to propose computationally tractable numerical algorithms for computing universal lower bounds on the optimal adversarial risk and identifying optimal classifiers. We propose two main algorithms based on linear programming (LP) and entropic regularization (Sinkhorn). Our key insight is that one can harmlessly *truncate the higher order interactions* between classes, preventing the combinatorial run times typically encountered in MOT problems. We validate these results with experiments on MNIST and CIFAR-10, which demonstrate the tractability of our approach.

Keywords: Adversarial learning, Optimization, Linear programming, Sinkhorn algorithm, Multiclass classification, Optimal transport, Multimarginal optimal transport, Wasserstein barycenter, Generalized barycenter problem

1. Introduction

While neural networks have achieved state-of-the-art accuracy in classification problems, it is by now well known that networks trained with standard *error risk minimization* (ERM)

can be exceedingly brittle (Goodfellow et al., 2014; Chen et al., 2017; Qin et al., 2019; Cai et al., 2021). As a result, various works have suggested replacing ERM with alternative training procedures that enforce robustness. In this paper, we focus on *adversarial training* (AT), which converts standard risk minimization into a min-max problem where the learner is pitted against an adversary with the power to perturb the training data (see, e.g., Madry, Makelov, Schmidt, Tsipras, and Vladu (2017); Tramèr, Kurakin, Papernot, Goodfellow, Boneh, and McDaniel (2018); Sinha, Namkoong, and Duchi (2018)). This provides a powerful defense against adversarial attacks at the cost of increased computation time and worse performance on clean data (Tsipras et al., 2018). Hence, balancing robustness against efficiency of computation and clean accuracy is central to the effective deployment of adversarial training.

A key hyperparameter in AT is the *adversarial budget* ε, which controls how far the adversary is allowed to move each individual data point. As ε increases, an adversarially trained network will become more robust but will lose accuracy, as the learner is forced to be robust against stronger and stronger attacks. Due to the computational cost of adversarial training (one must solve a min-max problem rather than a pure min problem), hyperparameter tuning of the adversarial budget is very expensive. As a result, there is a great practical need for theory and algorithms that can predict good choices of ε without requiring massive computation.

A recent body of work has attempted to address this issue by providing bounds on the gap between the optimal adversarial risk (with a given budget ε) and the optimal standard risk. Indeed, provided that these bounds are reasonably tight and efficiently computable, they may provide a more efficient route for a practitioner to choose and tune ε.

Thus far, the literature has largely focused on the theoretical side of characterizing such bounds. A number of authors have provided bounds for specific types of classifiers and neural network architectures by Weng, Zhang, Chen, Yi, Su, Gao, Hsieh, and Daniel (2018); Yin, Kannan, and Bartlett (2019); Khim and Loh (2019), though it is unclear whether any of these results continue to hold if the underlying model changes. On the other hand, a more recent body of work has established lower bounds that are classifier agnostic (see Bhagoji, Cullina, and Mittal (2019); Pydi and Jog (2021); García Trillos and Murray (2022); Bungert and Stinson (2022) for the binary classification case and Garcıa Trillos, Jacobs, and Kim (2023a); Dai, Ding, Bhagoji, Cullina, Zhao, Zheng, and Mittal (2023) for the multiclass case). The classifier agnostic lower bounds are obtained by relaxing the training problem to allow the learner to select any measurable probabilistic classifier (note that in the modern era of neural networks with billions of parameters, this relaxation may be relatively tight). As a result, the lower bounds are in fact universal and have no dependence on the learning model. Nonetheless, despite this fertile body of work, the actual computation of these bounds along with implementable algorithms has been left largely unexplored.

The main focus of this paper is to provide efficient algorithms for computing classifier agnostic lower bounds on the optimal adversarial risk. To do so, we utilize an equivalence discovered by Garcıa Trillos, Jacobs, and Kim (2023a) between the relaxed adversarial training problem and a *multimarginal optimal transport* (MOT) problem related to finding barycenters using the ∞-Wasserstein distance. Thanks to this connection, we can leverage tools from computational optimal transport to develop efficient algorithms for solving the equivalent MOT barycenter problem.

In general, MOT problems (including the Wasserstein barycenter problem) are NP-Hard (Altschuler and Boix-Adserà, 2022). However, we will show that when ε is not too large (the relevant regime for AT) it is possible to efficiently solve the particular MOT problem related to AT. The key insight is that for small values of ε, the search space for the problem can be significantly truncated, allowing for very efficient computations. Notably, this truncation can only decrease the value of the problem, preserving the guarantee that computed values are truly lower bounds on the optimal adversarial risk. Equally important is the fact that this truncation is compatible with both linear programming and the Sinkhorn algorithm, two of the most popular methods for solving MOT problems. Leveraging these two approaches, we propose two very efficient algorithms for solving this problem. We then validate our algorithms with experiments on MNIST and CIFAR-10 to demonstrate the tractability of our approach.

The closest paper to this work is Dai, Ding, Bhagoji, Cullina, Zhao, Zheng, and Mittal (2023), where independently to Garcıa Trillos, Jacobs, and Kim (2023a), the authors relate the multiclass relaxed adversarial learning problem to an equivalent combinatorial optimization problem using the notion of a conflict hypergraph. Although the work Dai, Ding, Bhagoji, Cullina, Zhao, Zheng, and Mittal (2023) also explores the idea of truncation, the authors do not provide any tailored algorithms to solve the problem. In contrast, we present two concrete algorithms and provide their computational complexity.

1.1 Our contributions

Our main contributions in this paper are the following.

- We introduce and analyze a new algorithm for approximating adversarial attacks based on a stratified and multi-marginal form of Sinkhorn's algorithm. In addition we provide a publicly available implementation of our algorithm.[1]
- We give rigorous bounds on the computational complexities of our algorithms based on the user chosen truncation rate. We show that a certain class of MOT problems arising in adversarial training models can be solved very efficiently, despite the fact that MOT problems are in general NP-Hard.
- We implement, discuss, and compare against an exact solver based on Linear Programming as was done in Dai, Ding, Bhagoji, Cullina, Zhao, Zheng, and Mittal (2023). We also implement fast constructions of multiclass variants of the Čech and Rips complexes which may be of broader use, particularly for topological data analysis.

1.2 Outline

The rest of the paper is organized as follows. In section 2, we will formally introduce mathematical backgrounds for the adversarial training problem and notation used throughout the rest of the paper. In section 3, we will present our two main algorithms along with informal explanations for their efficiency and run time. In section 3.4 we present our main theoretical results on the analysis of our algorithm based on Sinkhorn iterations. Rigorous proofs of our main theoretical results will be delayed until the appendix. In section 4, we provide and

1. Code can be found at `https://github.com/MattWerenski/Adversarial-OT`

discuss empirical results obtained from running our proposed algorithms. We will finish off the paper in section 5, with some conclusions and discussions of future directions. Finally, all technical details and proofs will be discussed in appendix A.

2. Preliminaries

2.1 Basic concepts and notation

Let $(\mathcal{X}, d) = (\mathbb{R}^p, ||\cdot||)$ denote the feature space, and let $\mathcal{Y} := \{1, \dots, K\}$ be the set of K classes for a given classification problem of interest. Let $S_K := \{A \subseteq \mathcal{Y} : A \neq \emptyset\}$ and $S_K(i) := \{A \in S_K : i \in A\}$. Let $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ denote the set of feature-class pairs. Let $\mu \in \mathcal{P}(\mathcal{Z})$ be a Borel probability measure that represents the ground-truth data distribution. For convenience, we will often describe the measure μ in terms of its class weights $\mu = (\mu_1, \dots, \mu_K)$, where each μ_i is a positive Borel measure (not necessarily a probability measure) over $\mathcal{X}$ defined according to:

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

for all Borel measurable $E \subseteq \mathcal{X}$. Notice that the measures μ_i's are, up to normalization factors, the conditional distributions of features given the specific labels, and $\sum_{i \in \mathcal{Y}} \mu_i(\mathcal{X}) = ||\mu|| = 1$.

The typical goal of (deterministic) multiclass classification is to find a Borel measurable map $f : \mathcal{X} \to \mathcal{Y}$ within a certain class (a.k.a. hypothesis class) which minimizes $\mathbb{E}[\ell(f(\mathcal{X}), Y)]$, where $(\mathcal{X}, Y) \sim \mu$, and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a loss function. Due to the non-convexity of the space of multiclass deterministic classifiers, often one needs to relax the space and instead consider *probabilistic* multiclass classifiers $f : \mathcal{X} \to \Delta_{\mathcal{Y}}$ where

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

is the probability simplex over the label space $\mathcal{Y}$. In what follows, we denote the set of *all* such Borel maps by $\mathcal{F} := \{f : \mathcal{X} \to \Delta_{\mathcal{Y}} : f \text{ is Borel measurable}\}$. Given $f \in \mathcal{F}$ and $x \in \mathcal{X}$ we use $f_i(x)$ to denote it's ith component at the point x. The value of $f_i(x)$ should be interpreted as the estimated probability that x is in the ith class under the classification rule f. This extension to probabilistic classifiers is typically unavoidable, especially for adversarial problems in multiclass classification Garcıa Trillos, Jacobs, and Kim (2023a, 2024). The objective in this setting is given by

$$\inf_{f \in \mathcal{F}} R(f, \mu) := \mathbb{E}_{(X,Y) \sim \mu}[\ell(f(X), Y))]. \tag{1}$$

(1) represents the standard (agnostic) multiclass Bayes learning problem. Through this paper, the loss function $\ell : \Delta_{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$ is set to be $\ell(u, i) := 1 - u_i$ for $(u, i) \in \Delta_{\mathcal{Y}} \times \mathcal{Y}$, which is usually referred to as the 0-1 loss. Under the 0-1 loss function the risk $R(f, \mu)$ can be rewritten as

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

and is well known that its solution is the so called Bayes classifier, which admits an explicit form in terms of the distribution μ.

2.2 Adversarial training

One popular approach used in adversarial training is *distributionally robust optimization* (DRO). Here the training model is based on a *distribution-perturbing adversary* which can be formulated through the min-max optimization problem

$$R^*_{DRO} := \inf_{f\in\mathcal{F}} \sup_{\widetilde{\mu}\in\mathcal{P}(\mathcal{Z})} \{R(f,\widetilde{\mu}) - C(\mu,\widetilde{\mu})\}\,. \tag{2}$$

Here, the adversary has the power to select a new data distribution $\widetilde{\mu}$, but they must pay a cost to transform μ into $\widetilde{\mu}$ given by $C : \mathcal{P}(\mathcal{Z})\times\mathcal{P}(\mathcal{Z}) \to [0,\infty]$, which measures how different $\widetilde{\mu}$ is from μ. This forces the learner to select a classifier f that is robust to perturbations of the ground truth data μ within a certain distance determined by the properties of the chosen cost C.

In this work, we consider the family of costs $C_\varepsilon, \varepsilon > 0$ which are transportation costs from μ to $\widetilde{\mu}$ given by

$$C_\varepsilon(\mu,\widetilde{\mu}) := \sum_{i\in\mathcal{Y}} \inf_{\pi_i\in\Pi(\mu_i,\widetilde{\mu}_i)} \int c_\varepsilon(x,\tilde{x}) d\pi_i(x,\tilde{x}),$$

where $\widetilde{\mu}_i$ is defined analogously to μ_i, $\Pi(\mu_i,\widetilde{\mu}_i)$ is the set of probability measures over $\mathcal{X}\times\mathcal{X}$ whose first and second marginals are μ_i and $\widetilde{\mu}_i$ respectively, and c_ε is given by

$$c_\varepsilon(x,\tilde{x}) = \begin{cases} 0 & d(x,\tilde{x}) \le \varepsilon \\ +\infty & \text{otherwise} \end{cases} \tag{3}$$

for some distance d on $\mathcal{X}$ and some adversarial budget ε. We will slightly abuse notation and write $C_\varepsilon(\mu_i,\widetilde{\mu}_i)$ to mean the transport cost between μ_i and $\widetilde{\mu}_i$ with cost c_ε. If $\Pi(\mu_i,\widetilde{\mu}_i)$ is empty, which is the case when $\|\mu_i\| \neq \|\widetilde{\mu}_i\|$, then we take $C_\varepsilon(\mu_i,\widetilde{\mu}_i) = \infty$.

An equivalent perspective is that $\widetilde{\mu}$ is a feasible attack for the adversary only if for all $i \in \mathcal{Y}$ it holds that $W_\infty(\mu_i,\widetilde{\mu}_i) \le \varepsilon$, where W_∞ denotes the ∞-OT distance between measures. In other words, the adversary is only allowed to move each individual data point in the distribution by a distance ε in the feature space $\mathcal{X}$. This shows how the choice of budget ε is related to the strength of the adversary. As ε increases, the adversary can make stronger and stronger attacks.

Since the min in problem 2 is over all possible measurable probabilistic classifiers, the value of problem (2) provides a universal lower bound for learning problems over *any* family of (Borel measurable) classifiers (e.g., neural networks, kernel machines, etc) with the same type of robustness enforcing mechanism (i.e., same adversarial cost). For this reason, computing (2) is of relevance, and in particular an estimated value for (2) can be used as benchmark when training structured classifiers in practical settings, as has been discussed in the introduction.

2.3 An equivalent MOT problem

In Garcıa Trillos, Jacobs, and Kim (2023a), the authors showed that the DRO training problem (2) is equivalent to an MOT problem related to solving a generalized version of the Wasserstein barycenter problem. In what follows we provide some discussion on this

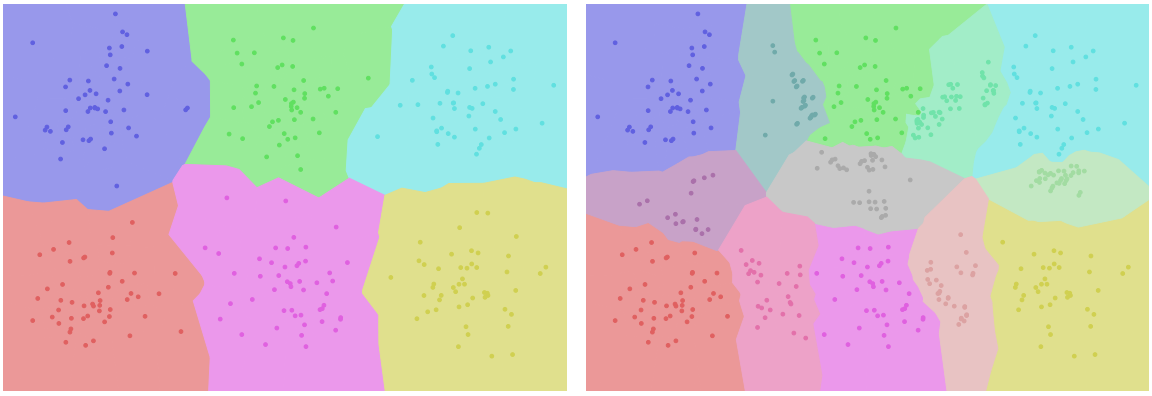

Figure 1: **(Left)** A simple six class dataset with 50 points in each class. Filled regions are colored according to the class of the nearest point. **(Right)** The optimal adversarial attack applied to the dataset on the left. The shared colors from the left figure represent the singleton classes $\widetilde{\mu}_{\{1\}}, ..., \widetilde{\mu}_{\{6\}}$ and the blended colors represent the various $\widetilde{\mu}_A$ for $|A| \geq 2$. Points are colored according to the combination of classes they are associated with. Filled regions are colored according to the combination of the nearest point.

equivalence. Given $\mu = (\mu_1, \dots, \mu_K)$, the *generalized barycenter problem* associated to the cost c_ε in (3) is

$$\min_{\lambda, \widetilde{\mu}_1, \dots, \widetilde{\mu}_K} \left\{ \lambda(\mathcal{X}) + \sum_{i \in \mathcal{Y}} C_\varepsilon(\mu_i, \widetilde{\mu}_i) : \lambda \succeq \widetilde{\mu}_i \text{ for all } i \in \mathcal{Y} \right\}, \tag{4}$$

where by $\lambda \succeq \widetilde{\mu}_i$ we mean that the positive measure λ dominates the measure μ_i. That is, for any non-negative measurable function g, it holds that $\int_{\mathcal{X}} g(x) d\lambda(x) \geq \int_{\mathcal{X}} g(x) d\widetilde{\mu}_i(x)$.

For any feasible collection $\lambda, \widetilde{\mu}_1, ..., \widetilde{\mu}_K$ it is possible to perform the following decomposition

$$\lambda = \sum_{A \in S_k} \lambda_A, \quad \lambda_A = \widetilde{\mu}_{i,A} \ \forall \ i \in A, \quad \widetilde{\mu}_i = \sum_{A \in S_K(i)} \widetilde{\mu}_{i,A} \ \forall \ i = 1, ..., K, \tag{5}$$

where λ_A is a measure which accounts for the jointly overlapping mass of the $\widetilde{\mu}_i$ with $i \in A$: see Figure 1 for the pictorial explanation. The indices A's of the decomposition represent the interactions between different classes. To decrease $\lambda(\mathcal{X})$, which can be interpreted as the classification power by the learner if the adversary were to choose the distributions $\tilde{\mu}_i$, the adversary would attempt to make the overlap between different classes be as large as possible, i.e., aim to make $\widetilde{\mu}_i \approx \widetilde{\mu}_j$. Under the additional transportation cost constraints, the optimal strategy for the adversary is to decompose the $\widetilde{\mu}_i$'s according to $\widetilde{\mu}_{i,A} = \widetilde{\mu}_{j,A} = \lambda_A$ for all $i, j \in A$. From the decomposition of $\widetilde{\mu}_i$ one can also decompose μ_i into $\mu_i = \sum_{A \in S_K(i)} \mu_{i,A}$ in such a way that $C_\varepsilon(\mu_i, \widetilde{\mu}_i) = \sum_{A \in S_K(i)} C_\varepsilon(\mu_{i,A}, \widetilde{\mu}_{i,A})$. Using the decomposition (5) and

the identity for the cost one can convert (4) into the equivalent problem

$$\min_{(\lambda_A,\mu_{i,A})\in F}\left\{\sum_{A\in S_K}\left(\lambda_A(\mathcal{X})+\sum_{i\in A}C_\varepsilon(\lambda_A,\mu_{i,A})\right)\right\}, \tag{6}$$

where F is the feasible set

$$F:=\left\{(\lambda_A,\mu_{i,A})\ :\ \sum_{A\in S_K(i)}\mu_{i,A}=\mu_i\ \forall\ i\in\mathcal{Y}\right\}.$$

One can rigorously prove that an optimal λ_A is a barycenter for the set of measures, $\{\mu_{i,A}: i\in A\}$. As in classical (Wasserstein) barycenter problems (Ekeland, 2005; Agueh and Carlier, 2011b,a), (6) has an equivalent *stratified MOT* formulation: letting $\mathcal{X}^A:=\prod_{i\in A}\mathcal{X}$ and $x_A:=(x_i: i\in A)\in\mathcal{X}^A$,

$$\begin{aligned}&\min_{\{\pi_A\}_{A\in S_K}}\sum_{A\in S_K}\int_{\mathcal{X}^A}(1+c_{\varepsilon,A}(x_A))\,d\pi_A(x_A)\\ &\text{s.t.}\sum_{A\in S_K(i)}\mathcal{P}_{i\#}\pi_A=\mu_i\quad\forall\ i\in\mathcal{Y},\end{aligned} \tag{7}$$

where $\mathcal{P}_i$ is the projection map $(x_A)\mapsto x_i$ onto the i-th component for $i\in A$, and for each $A\in S_K$, c_A is defined as

$$c_{\varepsilon,A}(x_A):=\inf_{x'\in\mathcal{X}}\sum_{i\in A}c_\varepsilon(x',x_i) \tag{8}$$

with the convention that $c_{\{i\}}=0$ for all $i\in\mathcal{Y}$. It is proved in Garcıa Trillos, Jacobs, and Kim (2023a) that

$$(2)=1-(4)=1-(6)=1-(7).$$

This paper is devoted to developing and understanding algorithms to solve (2), or equivalently to solve (6) and (7), respectively.

The dual formulation of (7), studied in Garcıa Trillos, Jacobs, and Kim (2023a), takes the form

$$\begin{aligned}&\sup_{g_1,\dots,g_K\in C_b(\mathcal{X})}\sum_{i\in[K]}\int_{\mathcal{X}}g_i(x_i)d\mu_i(x_i)\\ &\text{s.t.}\sum_{i\in A}g_i(x_i)\le 1+c_{\varepsilon,A}(x_A)\text{ for all }x_A\in\mathcal{X}^A, A\in S_K,\end{aligned} \tag{9}$$

where $C_b(\mathcal{X})$ is the set of bounded continuous functions. The primal problem can be thought of as a problem solved by the adversary, whereas the dual can be associated to the learner. While this can be seen more directly for the adversary, some discussion is needed to explain this interpretation for the learner.

Suppose $(g_1^*,\dots,g_K^*)$ is a solution of (9). Defining $f^*=(f_1^*,\dots,f_K^*)$ as

$$f_i^*(x):=\max\left\{\sup_{x'\in\mathrm{spt}(\mu_i)}\left\{g_i^*(x')-c_\varepsilon(x,x')\right\},0\right\}, \tag{10}$$

it is possible to show that f^* is indeed an optimal robust classifier of (2) provided that f^* is Borel measurable: see Garcıa Trillos et al. (2023a, Corollary 33) and Garcıa Trillos et al. (2023b, Corollary 4.7 and Remark 4.9). Borel measurability is guaranteed, for example, if we assume the measures μ_i to be supported on finitely many points (i.e., the setting of interest in this work). For more general measures μ_i, however, existence of (optimal) Borel measurable robust classifiers is a delicate issue that is carefully studied in Garcıa Trillos et al. (2024, Theorem 2.5 and Proposition 4.1).

3. Algorithms

The main contribution of this paper is to suggest two numerical schemes for solving (2) when the measure μ is an empirical measure associated to a finite data set. From now on, we assume that each μ_i is a measure supported over a finite set

$$\mathcal{X}_i := \mathrm{spt}(\mu_i) \subset \mathcal{X}, \quad |\mathcal{X}_i| = n_i = O(n) \text{ for all } i \in \mathcal{Y}.$$

Also, we use $\mathcal{X}^A := \prod_{i\in A} \mathcal{X}_i$ and $x_A := (x_i : i \in A) \in \mathcal{X}^A$ as before. We say that a tuple of points x_A is a feasible interaction if $c_{\varepsilon,A}(x_A) < \infty$, which corresponds to there existing a x' as in (8) with finite cost. The points $(x_i : i \in A)$ are then capable of interacting in an adversarial attack by assigning $\pi_A[\{x_A\}]$ positive mass, which is equivalent to assigning λ_A positive mass at x', the location of the interaction. If x_A is not a feasible interaction, then any adversarial attack which assigns $\pi_A[\{x_A\}]$ positive mass will immediately have infinite cost. The key ideas that make these methods feasible even when K is big is to *truncate interactions* and leverage the fact that the set of *feasible interactions* is small when the adversarial budget ε is small, that is to say that when $|A|$ is large and ε is small, there are typically very few feasible interactions x_A. The truncation of (6) to interactions of level $L \leq K$ is given by

$$\min_{(\lambda_A,\mu_{i,A})\in F_L} \left\{ \sum_{A\in S_K} \left(\lambda_A(\mathcal{X}) + \sum_{i\in A} C_\varepsilon(\lambda_A, \mu_{i,A}) \right) \right\}. \tag{11}$$

where

$$F_L := \{(\lambda_A, \mu_{i,A}) \in F : |A| \leq L\}.$$

In words, we set $\mu_{i,A} = 0$ (hence, $\lambda_A = 0$) for all A such that $|A| > L$. Similarly, the truncation of (7) to interactions of level L is given by

$$\begin{aligned} &\min_{\{\pi_A\}_{A\in S_K^L}} \sum_{A\in S_K^L} \int_{\mathcal{X}^A} (1 + c_{\varepsilon,A}(x_A))\, d\pi_A(x_A) \\ &\text{s.t.} \sum_{A\in S_K^L(i)} \mathcal{P}_{i\#}\pi_A = \mu_i \text{ for all } i \in \mathcal{Y}, \end{aligned} \tag{12}$$

where $S_K^L := \{A \in S_K : |A| \leq L\}$ and $S_K^L(i) := \{A \in S_K(i) : |A| \leq L\}$. In words, we set $\pi_A = 0$ for all A such that $|A| > L$.

The truncations of these problems satisfy the following approximation guarantees.

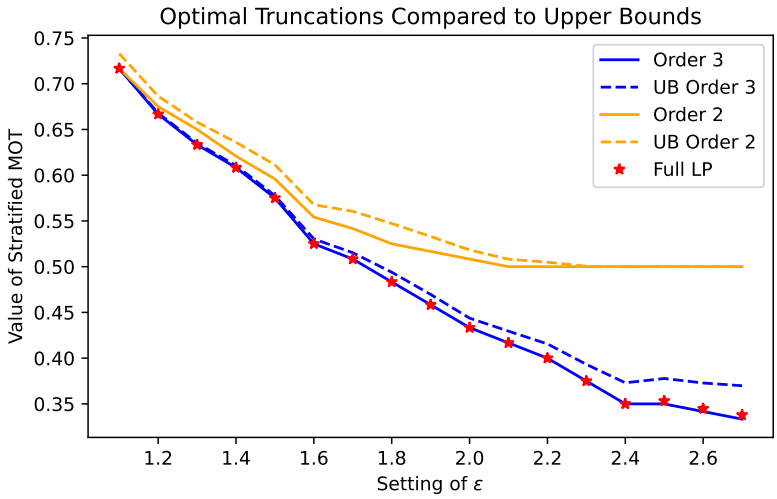

Figure 2: Plots of the value of (12) and the upper bound provided by Proposition 1 for a range of settings of ε and $K = 2, 3$ as well as the untruncated values. These are derived from synthetic data using 20 samples from six classes.

Proposition 1 *Let* $\{\lambda_A^*\}$ *be the optimal measures in* (6). *For* $1 \le L \le K$ *we have*

$$(11) - (6) \le \sum_{k>L}^{K} \left(\frac{k}{L} - 1\right) \sum_{|A|=k} \|\lambda_A^*\|.$$

Let $\{\pi_A^*\}$ *be the optimal measures in* (7). *For* $1 \le L \le K$ *we have*

$$(12) - (7) \le \sum_{k>L}^{K} \frac{k}{L} \sum_{|A|=k} \|\pi_A^*\|.$$

The proofs of both facts, presented in Appendix A.1, are given by constructing measures $\{\lambda_A^L : |A| \le L\}$ and $\{\pi_A^L : |A| \le L\}$ from the optimal measures $\{\lambda_A^*\}$ and $\{\pi_A^*\}$, respectively, which are feasible for (11) and (12) and obtain the bounds above.

An important takeaway from Proposition 1 is that if the optimal measures do not heavily utilize interactions beyond the truncation level L, i.e., the measures λ_A^* ad π_A^* have small mass when $|A| > L$, then one can faithfully recreate the attack without leveraging these interactions at all. Empirically, we observe that this is often the case: see experiments in section 4. We also illustrate these bounds on synthetic data in Figure 2 where the upper bounds are (7) plus the right hand side in Proposition 1. We observe that these bounds in this example are close to the actual values of (12). Importantly, notice that our truncation procedure does not reduce the total number of classes K, indeed, all classes $\{1, \ldots, K\}$ continue to influence adversarial attacks.

3.1 Feasible labeled interactions

In addition to truncation, the other key step in our method is to reduce the computational complexity by restricting the search space to feasible interactions only. Note that this does not change the problem whatsoever, as the adversary cannot combine points that do not lie in a common ε ball (or more generally points whose joint transport cost is infinite) whenever ε is small. We make this concept rigorous in the following definition.

Definition 2 *A set of points $\{(x_i, y_i)\}_{i=1}^k \subset \mathrm{spt}(\mu)$ is a valid interaction with $1 \leq k \leq K$ if the labels y_i are all distinct and*

$$\inf_{x' \in \mathcal{X}} \sum_{i=1}^{k} c_\varepsilon(x', x_i) < \infty. \tag{13}$$

The set of all valid interactions will be denoted by $\mathcal{J}_K$ and the set of all valid interactions of size at most L will be denoted by $\mathcal{J}_L$.

Note that this definition allows valid interaction sets to be of size 1 to K and in fact every singleton set is a valid interaction by choosing $x' = x_i$.

To leverage the computational gains of our approach, we must compute the feasible ordered interactions. Here we present an iterative method for computing these interactions, starting from singleton interactions (namely, all points of each class) and then building up to interactions of order $L \leq K$ where L is the user-chosen truncation level. To facilitate the description of our iterative method it will be useful to introduce for each $A \in S_K$ the feasible sets

$$F_A := \{\{(x_i, y_i)\}_{i=1}^{|A|} : \inf_{x' \in \mathcal{X}} \sum_{i=1}^{|A|} c_\varepsilon(x', x_i) < \infty, \bigcup_{i=1}^{|A|} \{y_i\} = A\}.$$

Note that F_A is the set of all feasible interactions with labels corresponding to the set A. Now we are ready to present Algorithm 1 for computing feasible interactions.

Algorithm 1 Construct $\mathcal{J}_L$

Input: X : data set, ε : adversarial budget, L: truncation level
  For each $i \in \mathcal{Y}$, set $F_{\{i\}} = \{(i, 1), ..., (i, n_i)\}$.
  for $k = 2, \ldots, L$ **do**
    for $A, A' \in S_K^L$ with $|A| = |A'| = k-1, |A \cap A'| = k-2$ and $F_A, F_{A'} \neq \varnothing$ **do**
      for Each $C \in F_A, C' \in F_{A'}$ 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 $F_{A \cup A'}$.
      end for
    end for
  end for
Output: $\mathcal{J}_L := \bigcup_{A \in S_K^L} F_A$.

The main difficulty in implementing Algorithm 1 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 L spheres in $\mathbb{R}^n$ of radius ε have a mutual intersection. One geometry where this calculation is particularly simple is when using the ℓ_∞ norm, where the problem is reduced to finding the intersection of axis-aligned rectangles.

In general the speed of Algorithm 1 is $O(L|\mathcal{J}_L|m(L))$ where $m(L)$ is the computational complexity required to check the existence of a point x for groups of size at most L (this is typically polynomial in L and the dimension of the space d). The overall complexity is essentially at worst the same as trying every possible group of L or fewer points, which is what may be required if ε is large enough that a majority of all the groups of size L are feasible. However, in practice, there are often far fewer higher-order interactions, which leads to a much faster algorithm.

3.2 Linear Programming Approach

The first method for solving problem (7) or its truncated version (12) that we discuss in this paper is based on *linear programming* (LP). The key object in the LP approach is the *interaction matrix* denoted by $J \in \{0,1\}^{\mathrm{spt}(\mu) \times \mathcal{J}_K}$. The rows of this matrix are indexed by points $z = (x, y) \in \mathrm{spt}(\mu)$ while the columns are indexed by the valid interactions $\iota \in \mathcal{J}_K$. For $z \in \mathrm{spt}(\mu)$ and $\iota \in \mathcal{J}_K$ the corresponding entry in J is given by

$$J[z, \iota] = \begin{cases} 1 & z \in \iota \\ 0 & \text{otherwise.} \end{cases}$$

We will also need to define a *marginal vector* $m \in [0,1]^{\mathrm{spt}(\mu)}$ with $m[z] = \mu[\{z\}]$. With these definitions, we can formulate a linear program which solves (6):

$$\begin{aligned} \min_{w \in [0,1]^{\mathcal{J}_K}} \quad & \sum_{\iota \in \mathcal{J}_K} w[\iota] \\ \text{s.t.} \quad & Jw = m. \end{aligned} \tag{14}$$

In an analogous way to the above, one can truncate the problem with the truncation level L, and obtain the truncated LP

$$\begin{aligned} \min_{w \in [0,1]^{\mathcal{J}_L}} \quad & \sum_{\iota \in \mathcal{J}_L} w[\iota] \\ \text{s.t.} \quad & Jw = m. \end{aligned} \tag{15}$$

Proposition 3 *The optimization problems* (7) *and* (14) *are equivalent. The optimization problems* (12) *and* (15) *are equivalent. As a result the truncated LP obtains the same approximation as in Proposition 1.*

Proof We will show how to convert between the two problems. We will only do this for the untruncated versions as the truncated versions are done in an identical fashion.

First let $\{\pi_A\}$ be feasible and have finite cost for (7). To each point $(x_1, ..., x_K) \in \mathrm{spt}(\pi_A)$ we can assign a set $\iota = \{(x_i, i) \mid i \in A\}$. Clearly the labels in ι are unique. In addition, since we assume that the $\{\pi_A\}$ achieve a finite cost we must have $c_A(x_1, ..., x_K) = \inf_{x' \in \mathcal{X}} \sum_{i \in A} c(x', x_i) < \infty$ which shows that ι is a valid interaction. Set

$$w[\iota] = \pi_A \left[\left\{ (x'_1, .., x'_K) \mid x'_i = x_i \ \forall \ i \in A \right\} \right]. \tag{16}$$

The projection sum constraint ensures that $Jw = m$.

Now consider a feasible w. For a valid interaction $\iota = \{(x_i, y_i)\}$ let $A = \{y_i\}$. For $i \notin A$ let x_i be an arbitrary point in the support of μ_i. Now set

$$\pi_A \left[\{(x_1, ..., x_K)\} \right] = w[\iota]. \tag{17}$$

The summation constraint in the LP ensures that the π_A are feasible. ■

The gain in (14) is that the optimization occurs over a space of dimension $|\mathcal{J}_K|$ (which we expect to be small when ϵ is small), while the dimension of the problem in (7) is $(2^K - 1)n^K$ when $\mathrm{spt}(\mu)_i = O(n)$ for every i (since there are $2^K - 1$ sets $A \in S_K$ and each π_A is of size n^K). In the worst case setting, it may be that $|\mathcal{J}_K| = (2^K - 1)n^K$, which happens for example when $\bigcup_{i=1}^K \mathrm{spt}(\mu_i) \subset \overline{B}(0, \varepsilon)$, a closed ball with radius ε. However, in typical settings, one should expect $|\mathcal{J}_K|$ to be much smaller.

Truncating (14) down to (15) allows the problem to be solved even more quickly since we eliminate interactions of order larger than L. Note that in the worst case $|\mathcal{J}_L|$ has size at most $(2^L - 1)n^L$, but again we expect this to be much smaller when ε is not too large.

Once one has obtained an optimizer w^* of (14) one can easily compute an optimal adversarial attack $\widetilde{\mu} = (\widetilde{\mu}_1, \ldots, \widetilde{\mu}_K)$ and a corresponding optimal generalized barycenter λ. Let w^* be an optimizer for (14). An optimal generalized barycenter can be recovered via

$$\lambda = \sum_{C \in \mathcal{J}_K} w^*(C) \delta_{b(C)}$$

where $b(C)$ returns any point x such that $\{x^i_{l_i} : (i, l_i) \in C\} \subset \overline{B}(x, \varepsilon)$, furthermore, an optimal adversarial attack $\{\widetilde{\mu}_1, \ldots, \widetilde{\mu}_K\}$ can be recovered via

$$\widetilde{\mu}_i = \sum_{A \in S_K(i)} \sum_{C \in F_A} w^*(C) \delta_{b(C)}.$$

The total mass of the measures is correctly preserved because of the constraint $Jw = [\mu_1, \ldots, \mu_K]^T$. From the preceding equations, it is clear that λ dominates $\widetilde{\mu}_i$ for each $i \in \mathcal{Y}$. In addition, it is also easy to recover the transformation $\mu_i \mapsto \widetilde{\mu}_i$. The above analysis implies that (4) = (14), hence, the optimal adversarial risk is obtained by

$$1 - (14) = 1 - \sum_{C \in \mathcal{J}_K} w^*(C).$$

Given an optimizer w^*_L to the level L truncated problem, one can analogously compute all of the corresponding truncated quantities (barycenter, adversarial attacks, and risk) by replacing $w^*, \mathcal{J}_K, S_K(i)$ by $w^*_L, \mathcal{J}_L, S_L(i)$ respectively in the above formulas.

3.2.1 Complexity Considerations of LP approach

In general, the optimization problem involves a vector w whose length is determined by $|\mathcal{J}_L|$ as well as a sparse matrix I with at most $L|\mathcal{J}_L|$ non-zero entries (although this is quite pessimistic) for some choice of truncation level $L \leq K$. It is therefore essential to control $|\mathcal{J}_L|$. A straightforward calculation can show that

$$\mathbb{E}|\mathcal{J}_L| = \sum_{A \in S_K, |A| \leq L} \left[\prod_{i \in A} n_k \right] \mathbb{P}\left\{ \{X_i\}_{i \in A} \subset \overline{B}(x, \varepsilon) \text{ for some } x \right\}$$

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 worthwhile to analyze cases where one can cleanly bound the probability on the right hand side. For example, if $X_i \sim N(m_i, \Sigma_i)$, then one may reasonably expect to bound the probability by a function of the values of $\{(m_i, \Sigma_i)\}$.

3.3 Entropic Regularization Approach

The second approach that we consider in this paper is based on Sinkhorn iterations, which here are adapted to be able to solve the entropy-regularized truncated version of problem, (12). Sinkhorn iterations were originally proposed in Sinkhorn (1964); Sinkhorn and Knopp (1967), and in the past decade have been extensively studied in Cuturi (2013); Cuturi and Doucet (2014); Benamou, Carlier, Cuturi, Nenna, and Peyré (2015); Altschuler, Niles-Weed, and Rigollet (2017); Lin, Ho, Cuturi, and Jordan (2022) in a variety of settings under the optimal transport contexts. These extensive algorithmic and theoretical developments have been key factors in the increased use of optimal transport in modern machine learning by practitioners.

Let $L \leq K$ be the fixed level of truncation in problem (12) and recall the notation $S_K^L = \{A \in S_K : |A| \leq L\}$ and $S_K^L(i) = \{A \in S_K^L : i \in A\}$. Throughout the discussion in this section, we will consider a general family of cost tensors $\{c_A\}_A$ that are non-negative (and that may possibly take the value ∞) and satisfy $c_{\{i\}} \equiv 0$ for all $i \in \mathcal{Y}$. The main example to keep in mind is the collection of cost tensors defined as in (8), since these are the cost tensors that are connected to the adversarial training problem.

The L-level truncated entropic regularization problem associated to (12) (adapted in the obvious way to arbitrary cost tensors) is defined as

$$\begin{aligned} &\min_{\{\pi_A\}_{A \in S_K^L}} \sum_{A \in S_K^L} \sum_{\mathcal{X}^A} (1 + c_A(x_A))\, \pi_A(x_A) - \eta H(\pi_A) \\ &\text{s.t.} \sum_{A \in S_K^L(i)} \mathcal{P}_{i\#}\pi_A = \mu_i \qquad \text{for all } i \in \mathcal{Y}, \end{aligned} \tag{18}$$

where $H(\pi_A) := -\sum_{x_A} (\log \pi_A(x_A) - 1)\, \pi_A(x_A)$. In general, a Sinkhorn-based algorithm for finding solutions to a regularized transport problem over couplings aims at solving the corresponding dual problem by a coordinatewise greedy update. In this case, as discussed in Appendix A.2, the dual problem associated to (18) (here written as a minimization problem

for convenience) takes the form:

$$\begin{aligned}&\min_{\{g_i\}_{i\in\mathcal{Y}}} \mathcal{G}^L(\{g_i\})\\ &:= \sum_{A\in S_K^L}\sum_{\mathcal{X}^A}\exp\left(\frac{1}{\eta}\left(\sum_{i\in A} g_i(x_i) - (1+c_A(x_A))\right)\right) - \sum_{i\in\mathcal{Y}}\sum_{\mathcal{X}_i}\frac{g_i(x_i)}{\eta}\mu_i(x_i).\end{aligned} \tag{19}$$

For a given value of the dual variables $\{g_i\}_{i\in\mathcal{Y}}$ we define an induced family of couplings $\{\pi_A(g)\}_{A\in S_K^L}$ according to

$$\pi_A(g)(x_A) := \exp\left(\frac{1}{\eta}\left(\sum_{i\in A} g_i(x_i) - (1+c_A(x_A))\right)\right), \quad \forall x_A\in\mathcal{X}^A, \quad \forall A\in S_K^L. \tag{20}$$

In general, these couplings do not satisfy the constraints in (18), but if $g=g^*$ is a solution to (19), then $\{\pi_A^*\}_{A\in S_K^L}$ is optimal for problem (18). Moreover, it can be shown that a collection of couplings of the form (20) that in addition satisfy the constraints in (18) is in fact the global solution of (18).

Remark 4 *We want to highlight two new challenges in our setting in relation to other settings that have been studied in the literature of optimal transport, including standard MOTs. The first obstacle is that we must consider the set of coupling tensors of different orders simultaneously, rather than a single coupling tensor. The second obstacle is to account for the imbalance of marginal distributions. Unlike typical MOT problems, each marginal μ_i has a distinct mass. We address these issues using and expanding on the methods from previous works such as Altschuler, Niles-Weed, and Rigollet (2017); Lin, Ho, Cuturi, and Jordan (2022). Further details regarding the convergence analysis will be provided in section 3.4 and in appendices A.3, A.4 and A.5.*

Remark 5 *It is well known that the dual variables to standard MOT problems (i.e. Kantorovich potentials) satisfy a useful invariance. For standard problems, given dual variables $(g_1,\dots,g_K)$, the value of the MOT dual problem at $(g_1,\dots,g_K)$ will remain unchanged if one adds a constant vector with mean zero $(h_1,\dots,h_K)$ (i.e. $\sum_{i=1}^K h_i = 0$) to $(g_1,\dots,g_K)$ (see Vialard (2019); Di Marino and Gerolin (2020); Carlier (2022)). However, our dual problem does not have this property. This creates an additional layer of difficulty that we will need to overcome, as this invariance is often leveraged in Sinkhorn-type algorithms.*

Remark 6 *If $c_A(x_A) = +\infty$, then it is clear that the term x_A in the sum over $\mathcal{X}^A$ does not contribute to the value of (19). Therefore, the sum over $\mathcal{X}^A$ can be reduced to the sum over F_A, which is computed by Algorithm 1. As in the LP approach, this reduction can significantly reduce the computational complexity of the problem when the adversarial budget is small.*

To solve the entropic regularization problem (19) and as a consequence also solve (18), we introduce Algorithm 2 below. Note in light of the last remark, when computing $\mathcal{P}_{i\#}\pi_A(g^t)$, one needs only to sum over $x_A\in F_A$.

Algorithm 2 Truncated Multi-Sinkhorn (without rounding)

Input: X : data set, $\{c_A\}_A$: cost tensors, $\mu = (\mu_1, \ldots, \mu_K)$: empirical distribution, ε : adversarial budget, η : entropic parameter, L : truncation level, δ': parameter for stopping criterion.

Initialization. $t = 0$ and $g_i = \mathbf{0} \in \mathbb{R}^{n_i}$ for each $i \in \mathcal{Y}$.

while $E_t > \delta'$ **do**

Step 1. Choose the greedy coordinate I (with arbitrary tie breaking) by

$$I := \underset{1 \le i \le K}{\operatorname{argmax}} D_{\mathrm{KL}} \left(\mu_i || \sum_{A \in S_K^L(i)} \mathcal{P}_{i\#} \pi_A(g^t) \right).$$

Step 2. Compute $g^{t+1} = (g_1^{t+1}, \ldots, g_K^{t+1})$ by

$$g_i^{t+1} = \begin{cases} g_i^t + \eta \log \mu_i - \eta \log \left(\sum_{A \in S_K^L(i)} \mathcal{P}_{i\#} \pi_A(g^t) \right), & \text{if } i = I \\ g_i^t, & \text{otherwise .} \end{cases}$$

Step 3. Set new t to be $t + 1$.

end while

Output: $\{\pi_A(g^t)\}_{A \in S_K^L}$.

The stopping criterion we use for Algorithm 2 is $E_t \le \delta'$ for some prespecified $\delta' > 0$, where E_t is

$$E_t := \sum_{i \in \mathcal{Y}} ||\mu_i - \sum_{A \in S_K^L(i)} \mathcal{P}_{i\#} \pi_A(g^t)||_1, \tag{21}$$

i.e., E_t is the addition of the discrepancies in marginal constraints at iteration t.

Following the analysis presented in section A.3, one can deduce that the sequence of iterates produced by Algorithm 2 induces a collection of couplings $\{\pi_A(g^t)\}_{A \in S_K^L}$ (via (20)) that converge toward the unique solution of (18) as $t \to \infty$. However, when Algorithm 2 is stopped at a finite iteration T, it is not guaranteed that the current $\{\pi_A(g^T)\}_{A \in S_K^L}$ is feasible for the original problem (19). In order to obtain feasibility at every iteration, it is important to introduce a rounding scheme. The scheme we use here is an adaptation of one proposed by Altschuler, Niles-Weed, and Rigollet (2017) and later extended by Lin, Ho, Cuturi, and Jordan (2022) in multimarginal setting. However, our rounding scheme needs to take into account the fact that multiple couplings appear in each individual marginal constraint.

The truncated entropic regularization algorithm (with rounding) is the following.

Algorithm 3 Rounding

Input: $\{\pi_A\}_{A\in S_K^L}$: couplings, $\mu = (\mu_1, \dots, \mu_K)$: empirical distribution.

Initialization. $\pi_A^{(0)} = \pi_A$ for all $A \in S_K^L$.

for $i = 1, \dots, K$ **do**

Compute $z_i := \min\left\{\mathbb{1}_{n_i}, \mu_i / \sum_{A\in S_K^L(i)} \mathcal{P}_{i\#}\pi_A^{(i-1)}\right\} \in \mathbb{R}_+^{n_i}$.

for $x_i \in \mathcal{X}_i$ **do**

Set for all A and all x_A containing x_i in its coordinate i:

$$\pi_A^{(i)}(x_A) = \begin{cases} z_i(x_i)\pi_A^{(i-1)}(x_A), & \text{if } i \in A \\ \pi_A^{(i-1)}(x_A), & \text{otherwise.} \end{cases}$$

end for

end for

Compute $\mathrm{err}_i := \mu_i - \sum_{A\in S_K^L(i)} \mathcal{P}_{i\#}\pi_A^{(K)}$ for each $i \in \mathcal{Y}$

Compute

$$\widehat{\pi}_A = \begin{cases} \pi_{\{i\}}^{(K)} + \mathrm{err}_i, & \text{if } A = \{i\} \\ \pi_A^{(K)}, & \text{otherwise.} \end{cases}$$

Output: $\{\widehat{\pi}_A\}_{A\in S_K^L}$.

Algorithm 4 Entropic regularization with rounding,

Input: Fix $\delta > 0$ and $L \le K$. Set $\eta = \frac{\delta/2}{2L\log(C^* K n)}$ and $\delta' = \frac{\delta/2}{2L \max_{A\in S_K^L} |1 + c_A \mathbb{1}_{c_A<\infty}|}$.

Step 1. For each $i \in \mathcal{Y}$, set

$$\mu_i' := \left(1 - \frac{\delta'}{4K}\right)\mu_i + \frac{\delta' ||\mu_i||}{4K n_i}\mathbf{1}_{n_i}.$$

Let $\mu' := (\mu_1', \dots, \mu_K')$.

Step 2. Compute $\{\widetilde{\pi}_A : A \in S_K^L\}$ by Algorithm 2 with $\{c_A\}_A$, η, μ' and $\delta'/2$.

Step 3. Obtain $\{\widehat{\pi}_A : A \in S_K^L\}$ by Algorithm 3 with $\{\widetilde{\pi}_A : A \in S_K^L\}$ and μ'.

Output: $\{\widehat{\pi}_A\}$.

As we will discuss in Section 3.4, Algorithm 4 returns a δ-approximate solution to the unregularized truncated problem (12).

Remark 7 *Here, $C^* = \max_{i\in\mathcal{Y}} \frac{n_i}{n}$ appearing in the choice of entropic parameter η is a constant independent of all other parameters, n, K, L and δ.*

***Step 1** of Algorithm 4 is necessary unless each μ_i is dense on $\mathcal{X}_i$. Algorithm 2 suffers when updating potentials if there is some μ_j with a very small mass. This weakness is not specific to this setup but a commonly observed phenomenon in variants of Sinkhorn-type algorithms.*

The choice of η is adapted for the theoretical analysis which considers the worst case. In practice, however, a larger choices of η works well. Sinkhorn folklore suggests that 0.05

is a good choice for η in most applications. See Section 4.3 for further discussion of how to choose η along with empirical results.

*If one is interested in the lower bound of the adversarial risk only, it is fine to skip **Step 3**, or Algorithm 3. Skipping **Step 3** has almost no effect on computing the risk. However, the main cost of the algorithm is **Step 2**, Algorithm 2, and the computational cost of **Step 3** is minor comparing to that of **Step 2**.*

Remark 8 *At the end of subsection 2.3 we discussed how to obtain a Borel measurable optimal robust classifier from optimal dual potentials of the dual formulation of the stratified MOT problem* (9)*. On the other hand, the objective function of the truncated Multi-Sinkhorn,* (19)*, is the entropic version of* (9)*. By solving* (19) *we can thus produce approximations of the optimal dual potentials for* (9) *that are parametrized by the entropic parameter η. In this sense, Algorithm 2 produces approximations not only for adversarial attacks but also for robust classifiers after plugging its outputs in* (20) *and* (10)*.*

A classifier obtained in this manner, however, is limited in the sense that it is only defined at most up to the closure of the ε-expansion of the supports of the μ_i's. In practice, these measures are typically empirical measures built from finite data sampled from some population distribution, and it is thus a priori unclear whether these classifiers can provide reliable outputs for all inputs in the ε-vicinity of the support of the population distribution. It would thus be important to study the sample complexity analysis of robust classifiers built in this fashion and to carefully investigate the dependence of the error on the parameter η. We leave this for future work.

3.4 Theoretical Results of Entropic Regularization Approach

In this Section, we state our main theoretical results, where we summarize our analysis of the approach for solving (12) that is based on the entropic regularization presented in section 3.3. Our first main result describes the number of iterations required to achieve the stoppin criteria for Algorithm 2.

Theorem 9 *Let $\{g^t\}_{t\in\mathbb{N}}$ be generated by Algorithm 2. For a sufficiently small fixed δ', the number of iterations T to achieve the stopping criterion $E_T \leq \delta'$ satisfies*

$$T \leq 2 + \left\lceil \frac{\mathcal{G}^L(g^0)}{\min_{i\in\mathcal{Y}}\|\mu_i\|_1} \right\rceil + \frac{14K^2\overline{R}}{\eta\delta'}, \tag{22}$$

where

$$\overline{R} := L - \eta \log \min_{j\in\mathcal{Y}, y\in\mathcal{X}_j} \mu_j(y) + \eta L \log(KC^*n). \tag{23}$$

Remark 10 *Recall that $g^0 = \mathbf{0}$. The initial value of the objective function is bounded above by*

$$\mathcal{G}^L(g^0) \leq \sum_{A\in S_K^L} \sum_{\mathcal{X}^A} \exp\left(-\frac{1}{\eta}\right) \leq C^* \exp\left(L\log(Kn) - \frac{1}{\eta}\right).$$

Taking $\eta = O\left(\frac{1}{L\log(Kn)}\right)$ *as we will do in Algorithm 4, we see that* $\mathcal{G}^L(g^0) = O(1)$. *Hence,* $\frac{\mathcal{G}^L(g^0)}{\min_{j\in\mathcal{Y}}||\mu_j||_1} = O(1)$. *Moreover, if we assume that* $\mu_i(x_i) \sim \frac{1}{n}$ *for all* x_i *and all* i, *the last term in* (22) *is* $O(\log^2(n))$. *As a consequence, Algorithm 2 exhibits almost linear convergence.*

Remark 11 *Notice that the number of iterations in Theorem 9 does not depend on the specific cost tensors* c_A. *Only the non-negativity of the cost tensors and the assumption* $c_{\{i\}} \equiv 0$ *for all* $i \in \mathcal{Y}$ *play a role in the estimated number of iterations.*

To prove Theorem 9, we adapt to our setting the analysis for standard MOT problems presented in Altschuler, Niles-Weed, and Rigollet (2017); Lin, Ho, Cuturi, and Jordan (2022). In our setting, the marginal distributions need not have the same total mass and each marginal constraint depends on multiple couplings of different orders simultaneously. In addition, the dual potentials in our setting lack an invariance property that is present in the standard setting, which facilitates the analysis in that case (see Remark 5). As a result, we require a more careful analysis for the decrement of energy at each step of the algorithm. The proof of Theorem 9 is presented at the end of appendix A.3, after proving a series of preliminary estimates.

In order to analyze Algorithm 4, we need the following estimates on the output of the rounding scheme.

Theorem 12 *Let* $\{\pi_A : A \in S_K^L\}$ *be a set of couplings and* $\mu = (\mu_1, \dots, \mu_K)$ *be a sequence of finite positive vectors. Then Algorithm 3 returns a set of couplings* $\{\widehat{\pi}_A : A \in S_K^L\}$ *which satisfies: for all* $i \in \mathcal{Y}$

$$\sum_{A\in S_K^L(i)} \mathcal{P}_{i\#}\widehat{\pi}_A = \mu_i,$$

as well as the error bound

$$\sum_{A\in S_K^L} ||\widehat{\pi}_A - \pi_A||_1 \leq L\sum_{i\in\mathcal{Y}} ||\sum_{A\in S_K^L(i)} \mathcal{P}_{i\#}\pi_A - \mu_i||_1.$$

Finally, we can combine Theorems 9 and 12 to prove that Algorithm 4 outputs a δ-approximate solution for (12), the truncated version of (7).

Theorem 13 *Algorithm 4 returns a* δ*-approximate optimal solution for* (12). *Moreover, if* $\min_{j\in\mathcal{Y},y\in\mathcal{X}_j} \mu_j(y) = \Omega(n^{-1})$ *and* $C^* := \max_{i\in\mathcal{Y}} \frac{n_i}{n} = O(1)$, *Algorithm 4 requires*

$$O\left(\frac{L^2K^2 \max_{A\in S_K^L}(1 + c_A\mathbb{1}_{c_A<\infty})|\mathcal{J}_L|\log(C^*Kn)}{\delta^2}\right)$$

operations to produce its output.

Remark 14 *Theorem 13 precisely quantifies the benefit of truncation in Algorithm 2: while all* K *classes still play a role in the formation of an adversarial attack, by truncating the number of interactions between classes, in the worst case* $|\mathcal{J}_L|$ *scales like* $\widetilde{O}(n^L)$ *as opposed to the worst case of scaling of* $|\mathcal{J}_K|$ *which scales like* $\widetilde{O}(n^K)$.

4. Empirical results

4.1 Numerical experiments for MNIST and CIFAR-10

In this section, we present experimental results obtained from applying our algorithms to datasets drawn from MNIST and CIFAR-10. For both data sets, there are 10 classes, and each class contains 100 points. Two different underlying ground metrics, ℓ^2 and ℓ^∞, are used. Due to dimensional scaling effects, the adversary requires a larger budget when the ground metric is ℓ^2. Note that MNIST images are 28×28 pixel grayscale images, while CIFAR-10 images are 32×32 pixels with 3 color channels.

Figure 3 shows the adversarial risks computed by the LP and Sinkhorn approaches using truncations of orders 2 and 3, along with their associated time complexities. Even though we restrict the interactions to order 2 or 3, the plots show that the adversarial risk does not change much when going from order 2 to order 3 interactions when the budget is not too large. This indicates that the truncated problem indeed provides meaningful lower bounds for the true problem when the adversarial budget is reasonable. Indeed, the curves for truncations of orders 2 and 3 are nearly identical for adversarial risk values below .3. Let us emphasize that the adversarial risks obtained by the LP and Sinkhorn approaches do not coincide, and in fact, the former is always larger than the latter as Sinkhorn gives a lower bound for the true optimal adversarial risk. Here, we do not scale down the entropic parameter η in terms of the number of points. One can reduce this gap by decreasing η, but this may cause numerical issues, which is a common phenomenon in computational optimal transport methods based on entropic regularization.

We should emphasize that the size of the adversarial budget has an enormous impact on the computational complexities of both algorithms. Indeed, both algorithms need the interaction matrix $\mathcal{J}_L$ constructed in Algorithm 1 to be relatively sparse to run efficiently. For small values of ε we expect $\mathcal{J}_L$ to be very sparse; however, as ε increases, it will become more dense slowing down both algorithms.

For the datasets that we consider, the worst-case complexities of the LP and Sinkhorn without truncation are $O(100^{30})$ and $\widetilde{O}(100^{10})$, respectively. With truncation up to order 3, the worst-case complexities become $O(100^9)$ and $\widetilde{O}(100^3)$, respectively. Note that these numbers are not usually achieved with adversarial budgets of small or moderate size (i.e. budgets that are relevant for adversarial training). Note, however, that the worst-case complexity of the LP approach can still be problematic even with truncation. In practice, this does happen in our experiments once ϵ becomes sufficiently large, in this case, we terminate the computation once it exceeds a certain wall-clock time. For Sinkhorn, since its worst-case complexity is almost linear with respect to n, the computation remains feasible even for budgets where the LP approach is infeasible (at least when η is not too small). This is one significant advantage of the Sinkhorn approach. However, one should still keep in mind that Sinkhorn will only return the exact value of the adversarial risk when the entropic regularization parameter is sent to zero (i.e. the regime where the algorithm gets slower and slower). Nonetheless, in our experiments, with an appropriate choice of the entropic parameter η, the Sinkhorn solution is quite close to the exact solution provided by the LP, while offering a much shorter computation time.

In Figure 4, we run experiments on a smaller subset of the data where there are 4 classes with 50 points (for both MNIST and CIFAR-10). This allows us to compare our computed

value of the adversarial risk from the order 2 and 3 truncated problems to the computed value for the untruncated problem. Readers can observe from those plots that truncations of orders 2 and 3 barely underestimate or even match the full order 4 risk, especially when the adversarial budget isn't too large. This is thanks to the fact that there are almost no valid interactions of higher order for reasonable values of ε. Indeed in almost all of the plots, the order 2 truncated value matches the untruncated value when the adversarial risk is in the range of 0% to 30%. Once ε is large enough that the risk grows beyond .3 we do start to see some discrepancy between the values of the truncated problem and the untruncated problem (especially when the ground metric is ℓ^∞). However, this regime is not so relevant as an adversarial risk of 20% is already extremely large and suggests that one should be training with a smaller budget.

From the experiments, we see that the truncation method works well in terms of both accurately approximating the adversarial risk and reducing the computational complexity. This is surprisingly nice because computing or even approximating the adversarial risk with many classes is hard in general: one must deal with a tensor of large order, requiring immense computational power. As long as classes are separated well, the truncation method will significantly reduce the order of tensors appearing in optimization (enhancing the efficiency of computing), while barely changing the optimal value of the problem, i.e. one should expect a very tight relaxation.

4.2 Fluctuation of interactions: Gaussian mixture, Iris and Glass data sets

In this section we further investigate in three settings the number of available higher-order interactions as well as how much mass the optimal multicoupling in (7) uses for each order. The three settings we consider are the following.

Synthetic We make a simple two-dimensional synthetic dataset which consists of six classes. For each class 30 samples are generated from 2-d Gaussian distribution with a mean $c_i \in \{(-2,2),(2,2),(6,2),(-2,-2),(2,-2),(6,-2)\}$, and the identity covariance matrix. This gives a total of 180 samples.

Iris This is the Iris dataset by Fisher (1988) which is four dimensional (measurements of sepal length, sepal width, petal length, and petal width) and has three classifications (setosa, versicolor, and virginica). One must classify the type of iris from the four given measurements. There are 50 samples for each type of iris and a total of 150 samples.

Glass This is the Glass dataset by German (1987) which is a ten dimensional (refractive index and percent composition of 9 atoms) and has six classifications (types of glass). There are 214 total samples non-uniformly distributed across the six classifications.

In Figure 5 we solve (7) and plot the amount of mass used in the optimal multicoupling, weighted by the number of interactions (we omit orders with negligible contributions). The order L line corresponds to $L \cdot \sum_{|A|=L} \|\pi_A\|$. This places the curves for higher-order interactions on the same scale and represents how much total mass of the marginals is accounted for by each order of interaction.

Interestingly the use of lower-order interactions (high-order interactions respectively) does not monotonically decrease (increase). A simple example where this happens can be furnished using six points and three classes and is illustrated in Figure 6. In this example

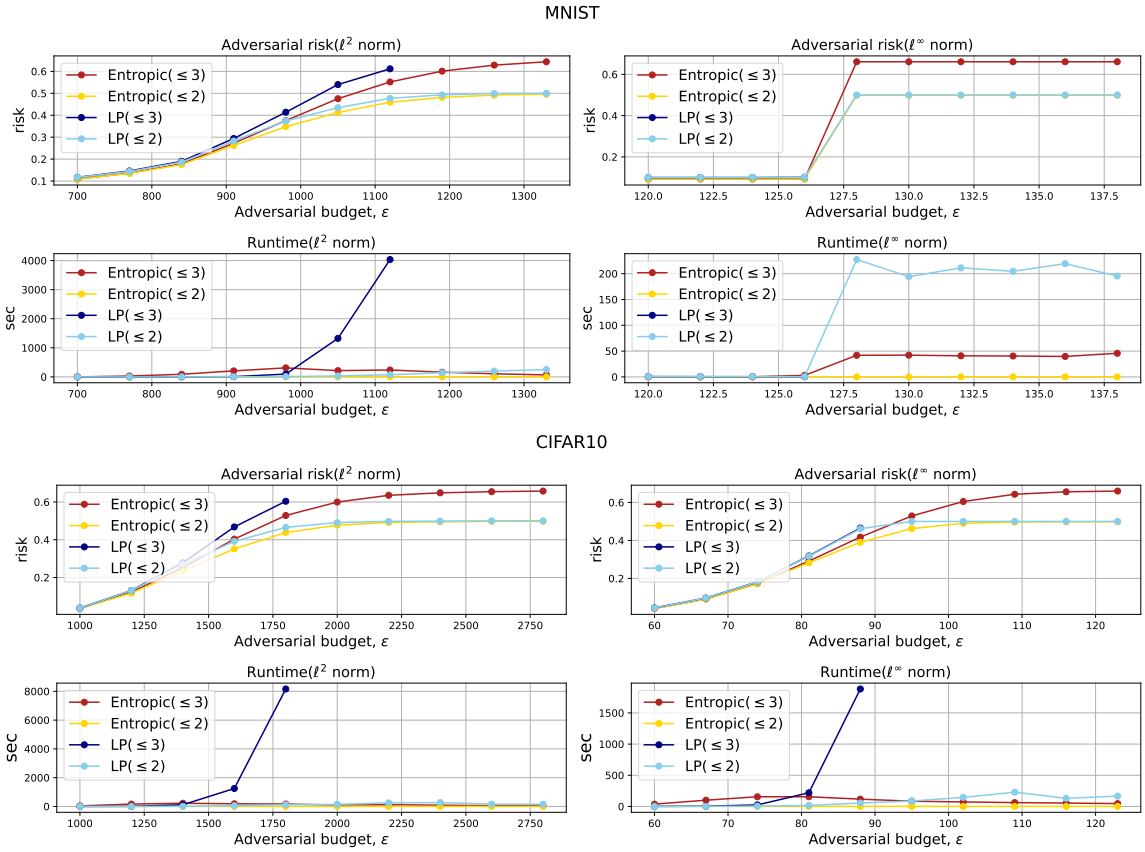

Figure 3: Lower bound of adversarial risk of and runtimes of the entropic regularization and LP for MNIST and CIFAR-10. The left plots and the right ones are equipped with ℓ^2 norm and ℓ^∞ norm, respectively. For LP with truncation up to 3, due to the huge complexity we stop the computing earlier.

if each point has unit weight the optimal perturbation for small budget has a mass of 4 (1 central point and 3 exterior points) and for a higher budge the optimal perturbation has a mass of 3 (3 midpoints) while using fewer order 3 interactions.

In Figure 7 we also show the number of feasible interactions that must be considered of each order as the budget varies. For the Glass dataset, though there are six classes, given this range of budgets, the highest order of interactions is 4; of course, one will see higher order interactions as the adversarial budget increases. For small budgets ε there are typically only interactions of lower order. However there are sharp thresholds where the number of higher order interactions rapidly increases. After these thresholds the computational complexity of the optimization problems rapidly increases due to an explosion in the number of optimization variables.

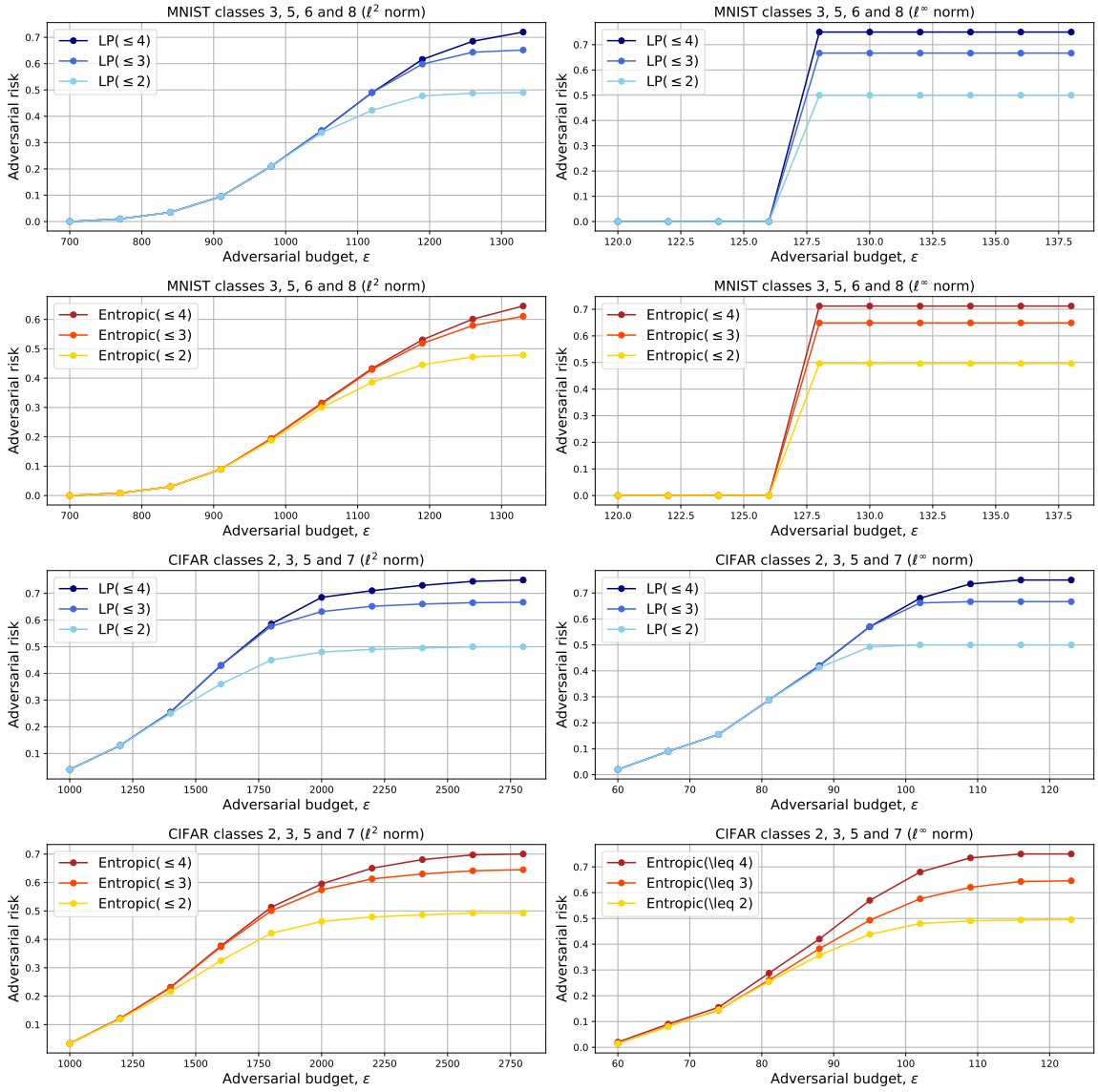

Figure 4: The optimal adversarial risks for MNIST and CIFAR-10 with 4 classes.

4.3 Setting η and δ' In Practice

In Theorem 13 we obtain bounds which achieve approximation ratios based on a parameter δ. As one can see in Algorithm 4 however, this requires setting η and δ' as functions of δ. This can be problematic in practice because taking η too small (for example below 0.01) leads to numerical instability, which is indeed a common issue for entropic regularization based methods (Feydy et al., 2019; Pooladian et al., 2022; Kassraie et al., 2024), and setting δ' too small may cause prohibitively many iterations in Algorithm 2. Instead, we often set η to be sufficiently small, typically close to $\eta = 0.01$ and $\delta' = 0.001$. We observe empirically

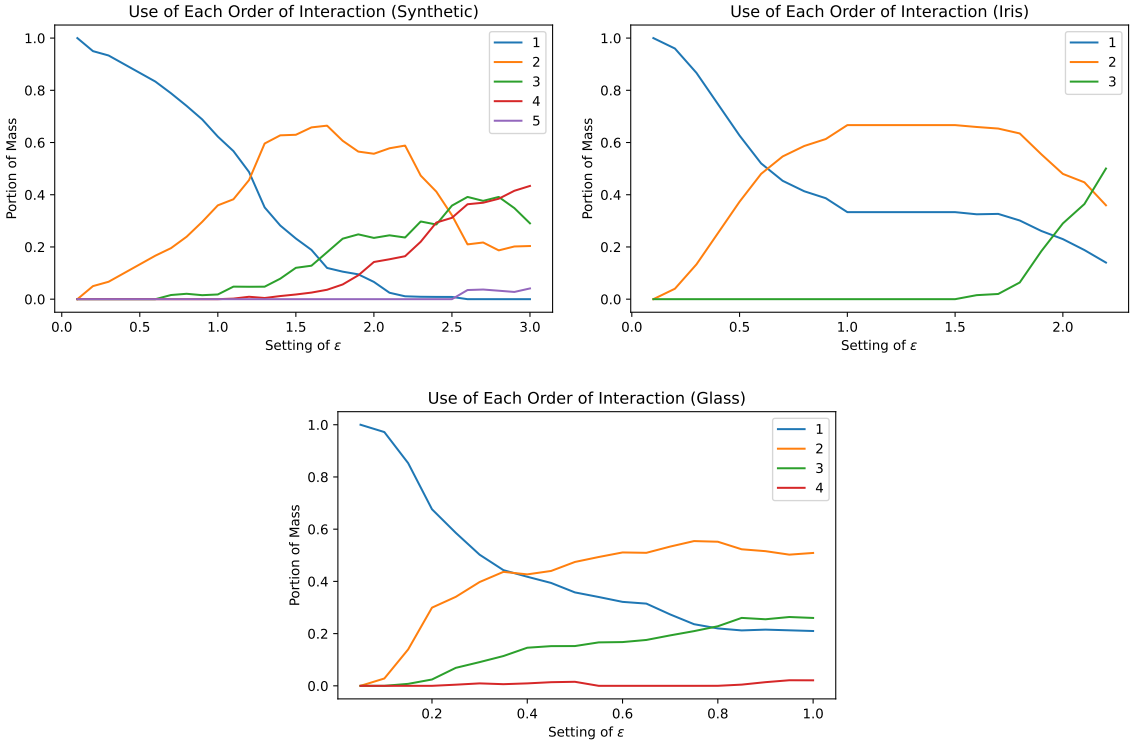

Figure 5: Contribution by interactions of each order to the optimal multicoupling as the budget ϵ varies in three different settings.

that these lead to high-quality solutions. This is illustrated on a synthetic example in Figure 8.

To create this figure we used six classes in two dimensions with data drawn according to $N(\mu_i, I/2)$ where $\mu_i \in \{(-2,2),(2,2),(6,2),(-2,-2),(2,-2),(6,-2)\}$ and 100 samples from each class. The maximum allowed interaction size was three. The parameter δ' in Algorithm 2 was fixed at 0.001. At least in this setting, a choice of moderate $\eta = 0.1$ achieves close to the exact adversarial risk (when only considering interactions up to size three).

5. Conclusion and future works

In this paper, we propose two algorithms that demonstrate impressive performance in synthetic data sets by utilizing MOT formulations of the adversarial training problem. The key idea powering our algorithms is that we can reduce the problem to a very sparse search space provided that the adversarial budget is not too large and the data is well-separated. Furthermore, by disallowing interactions of a certain size, we can cut down the search space even further, while barely affecting the optimal value of the problem. We validate these results through several experiments on popular machine-learning datasets.

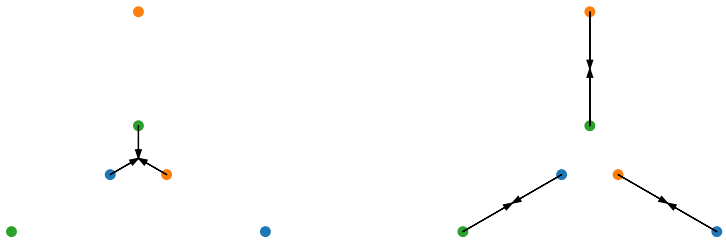

Figure 6: Configuration of six points with colors representing class. the central triangle has edge length 1 and the distance between corresponding points in the triangles is 2.0207. **(Left)** When $\varepsilon \in (0.5774, 1.0104)$ the optimal merging is achieved by placing the three interior points together. **(Right)** When $\varepsilon \in (1.0104, 1.0729)$ the optimal merging is achieved by pairing the interior triangle with the exterior triangle.

While our empirical results demonstrate that the training problem is indeed sparse for reasonably sized ϵ and that one can safely truncate higher-order interactions without affecting accuracy, it would be highly desirable to back up these results theoretically. A natural question is how to quantify the separation and impact of truncation on error. In particular, what is the ideal truncation level to reduce computational expenses while maintaining acceptable error levels? It would be highly beneficial to identify a sufficient criterion based on fundamental statistical measurements like mean, variance, and covariance to advise on the optimum truncation level. A Gaussian mixture model would be a promising starting point to address this inquiry.

Let us also emphasize that there is still significant room for improving the performances of both the linear programming entropic regularization approaches. Neither of our algorithms exploit specialized data structures nor parallelization. We anticipate that a more sophisticated handling of the sparsity structure of the problem could reduce the computation time by orders of magnitude. While the linear program should be fairly straightforward to parallelize, it may be quite nontrivial to parallelize the Sinkhorn version. To the best of our knowledge, it is unclear how to run Sinkhorn-type algorithms for MOT in parallel once there are three or more marginals (see Peyré et al. (2019) however for parallelization in the binary setting). Developing an appropriately parallelized version of our algorithms would be a very interesting line of future inquiry.

Appendix A. Appendix

A.1 Proof of Proposition 1

Proof Let $(\lambda_A^*, \mu_{i,A}^*)$ be the optimal measures in (6). Let $\lambda_A^L = \lambda_A^*$, $\mu_{i,A}^L = \mu_{i,A}^*$ for every A with $|A| < L$ and let $\lambda_A^L = 0, \mu_{i,A}^L = 0$ for every A with $|A| > L$. The approach (made precise below) is for each A with $|A| > L$ to distribute the mass in the sets λ_A^* and $\mu_{i,A}^*$ uniformly over the subsets of size L of A.

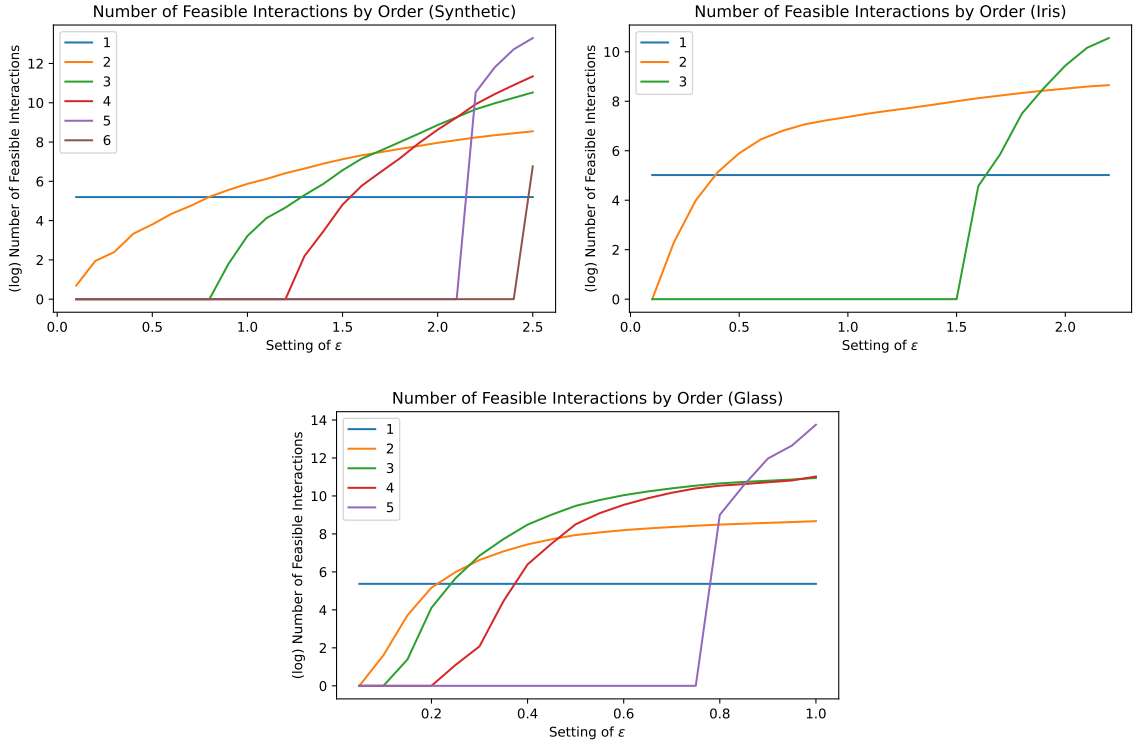

Figure 7: Number of interactions (plus one, in log scale) of each order which are feasible as the budget varies in the three different settings.

Now carrying out the details, for every A with $|A| = L$ let

$$\lambda_A^L = \sum_{\substack{B \in S_K \\ A \subset B}} \frac{|B|}{L} \binom{|B|}{L}^{-1} \lambda_B^* = \sum_{\substack{B \in S_K \\ A \subset B}} \binom{|B|-1}{L-1}^{-1} \lambda_B^*$$

and for every $i \in A$ set

$$\mu_{i,A}^L = \sum_{\substack{B \in S_K \\ A \subset B}} \binom{|B|-1}{L-1}^{-1} \mu_{i,B}^*.$$

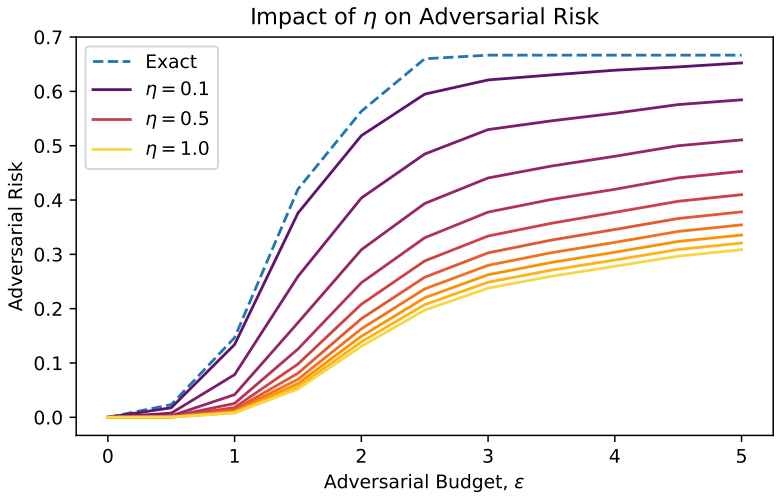

Figure 8: Impact on the adversarial risk as η varies from 0.1 to 1.0 in steps of 0.1 on synthetic data with $\delta' = 0.001$.

Clearly the $(\lambda_A^L, \mu_{i,A}^L)$ do not place any mass on sets A of size exceeding L. In addition

$$\begin{aligned}
\sum_{A\in S_K(i)} \mu_{i,A}^L &= \sum_{\substack{A\in S_K(i)\\ |A|<L}} \mu_{i,A}^* + \sum_{\substack{A\in S_K(i)\\ |A|=L}} \mu_{i,A}^L \\
&= \sum_{\substack{A\in S_K(i)\\ |A|<L}} \mu_{i,A}^* + \sum_{\substack{A\in S_K(i)\\ |A|=L}} \sum_{\substack{B\in S_K\\ A\subset B}} \binom{|B|-1}{L-1}^{-1} \mu_{i,B}^* \\
&= \sum_{\substack{A\in S_K(i)\\ |A|<L}} \mu_{i,A}^* + \sum_{\substack{B\in S_K(i)\\ |B|\geq L}} \sum_{\substack{A\subset B\\ |A|=L\\ A\in S_K(i)}} \binom{|B|-1}{L-1}^{-1} \mu_{i,B}^* \\
&= \sum_{\substack{A\in S_K(i)\\ |A|<L}} \mu_{i,A}^* + \sum_{\substack{B\in S_K(i)\\ |B|\geq L}} \mu_{i,B}^* = \mu_i.
\end{aligned}$$

where we have used in order the definition of $\mu_{i,A}^L$, a change in the order of summation, that the inner summand is constant with A and is counted precisely $\binom{|B|-1}{L-1}$ times, and that $\mu_{i,A}^*$ is feasible so it must sum to μ_i. This shows that the $\mu_{i,A}^L$ terms also satisfy the summation constraint and are therefore feasible for (11).

Next we will check that $C_\varepsilon(\lambda_A, \mu_{i,A}^L) = 0$ for every A. Let $\pi_{i,A}^*$ be the couplings of $\lambda_A^*, \mu_{i,A}^*$ which achieve

$$0 = C_\varepsilon(\lambda_A^*, \mu_{i,A}^*) = \int c_\varepsilon(x, x') d\pi_{i,A}^*.$$

For every A with $|A| < L$ we can use the coupling $\pi^*_{i,A}$ to couple $\mu^L_{i,A}$ and λ^L_A since these equal $\mu^*_{i,A}$ and λ^*_A respectively. For $|A| > L$ we can use $\pi_{i,A} = 0$ since $\mu^L_{i,A} = \lambda^L_A = 0$. All that remains is to handle $|A| = L$.

In this case note that $\pi^*_{i,A}$ has first marginal $\mu^*_{i,A}$ and second marginal λ^*_A. From this it follows that if we define

$$\pi^L_{i,A} = \sum_{\substack{B \in S_K \\ A \subset B}} \binom{|B|-1}{L-1}^{-1} \pi^*_{i,B}$$

then $\pi^L_{i,A}$ will have marginals $\mu^L_{i,A}$ and λ^L_A. We can also check that the cost is zero via

$$\int c_\varepsilon(x,x')d\pi^L_{i,A}(x,x') = \sum_{\substack{B \in S_K \\ A \subset B}} \binom{|B|-1}{L-1}^{-1} \int c_\varepsilon(x,x')d\pi^*_{i,B}(x,x') = \sum_{\substack{B \in S_K \\ A \subset B}} \binom{|B|-1}{L-1}^{-1} 0.$$

This shows that $\pi^L_{i,A}$ is a coupling of $\mu_{i,A}$ and λ_A with zero cost. Finally we can compare the objective costs of $(\lambda^*_A, \pi^*_{i,A})$ with $(\lambda^L_A, \pi^L_{i,A})$ as follows

$$\begin{aligned}
&\sum_{A \in S_K} \left(\lambda^L_A(\mathcal{X}) + \sum_{i \in A} C_\varepsilon(\lambda^L_A, \mu^L_{i,A}) \right) - \left(\lambda^*_A(\mathcal{X}) + \sum_{i \in A} C_\varepsilon(\lambda^*_A, \mu^*_{i,A}) \right) \\
&= \sum_{\substack{A \in S_K \\ |A|=L}} \lambda^L_A(\mathcal{X}) - \sum_{\substack{A \in S_K \\ |A| \geq L}} \lambda^*_A(\mathcal{X}) \\
&= \sum_{\substack{A \in S_K \\ |A|=L}} \left(\sum_{\substack{B \in S_K \\ A \subset B}} \frac{|B|}{L} \binom{|B|}{L}^{-1} \lambda^*_B(\mathcal{X}) \right) - \sum_{\substack{A \in S_K \\ |A| \geq L}} \lambda^*_A(\mathcal{X}) \\
&= \sum_{\substack{A \in S_K \\ |A| \geq L}} \left(\frac{|A|}{L} - 1 \right) \lambda^*_A(\mathcal{X}) = \sum_{k > L} \left(\frac{k}{L} - 1 \right) \sum_{|A|=k} \|\lambda^*_A\|.
\end{aligned}$$

In the jump to the second line we have used that the C_ε terms are all zero, that $\lambda^L_A = \lambda^*_A$ for $|A| < L$ and that $\lambda^L_A = 0$ for $|A| > L$. The third line uses the definition of λ^L_A. The final line is a term counting argument. This completes the first part of the proof.

The second part of the proof uses an analogous treatment which we only sketch. Let (π^*_A) be the optimal multicouplings in (7) and define (π^L_A) by taking $\pi^L_A = \pi^*_A$ for $|A| < L$, $\pi^L_A = 0$ for $|A| > L$ and

$$\pi^L_A = \sum_{\substack{B \in S_K \\ A \subset B}} \frac{|B|}{L} \binom{|B|}{L}^{-1} \lambda^*_B = \sum_{\substack{B \in S_K \\ A \subset B}} \binom{|B|-1}{L-1}^{-1} \mathcal{P}_A \pi^*_B$$

where $\mathcal{P}_A \pi^*_B$ is the projection of π^*_B onto its marginals corresponding to the set A. The remainder of the proof follows essentially the same structure once we observes $c_{\varepsilon,A} \leq c_{\varepsilon,B}$ for all $A \subset B$ which is helpful for showing that π^L_A has finite cost. ■

A.2 Dual of (18)

The Lagrangian for problem (18) is

$$\mathcal{L}(\{\pi_A\}_{A\in S_K^L}, \{g_i\}_{i\in\mathcal{Y}})$$

$$:= \sum_{A\in S_K^L}\sum_{\mathcal{X}^A}(1+c_A(x_A))\,\pi_A(x_A) - \eta H(\pi_A) - \sum_{i\in\mathcal{Y}}\sum_{\mathcal{X}_i} g_i(x_i)\left(\sum_{A\in S_K^L(i)} \mathcal{P}_{i\#}\pi_A(x_i) - \mu_i(x_i)\right),$$

where $\{g_i \in \mathbb{R}^{n_i}\}_{i\in\mathcal{Y}}$ is the collection of dual variables (one for each marginal constraint). The corresponding dual objective is defined as:

$$\mathcal{G}(\{g_i\}_{i\in\mathcal{Y}}) := \min_{\{\pi_A\}_{A\in S_K^L}} \mathcal{L}(\{\pi_A\}, \{g_i\}). \tag{24}$$

Notice that for every fixed $\{g_i\}$ (24) is a strictly convex optimization problem and thus its first order optimality conditions are sufficient to guarantee optimality. In turn, a straightforward computation shows that these first order optimality conditions read:

$$0 = \partial_{\pi_A(x_A)}\mathcal{L}(\{\pi_A\}, \{g_i\}) = 1 + c_A(x_A) + \eta\log\pi_A(x_A) - \sum_{i\in A} g_i(x_i), \quad \forall x_A \in \mathcal{X}^A, \quad \forall A \in S_K^L.$$

As a result, rearranging the above, we conclude that the unique solution of (24) is given by $\{\pi_A(g)\}_{A\in S_K^L}$, the set of couplings of the form (20).

Since

$$\begin{aligned}\sum_{i\in\mathcal{Y}}\sum_{\mathcal{X}_i} g_i(x_i) \sum_{A\in S_K^L(i)} \mathcal{P}_{i\#}\pi_A(x_i) &= \sum_{i\in\mathcal{Y}}\sum_{A\in S_K^L(i)}\sum_{\mathcal{X}^A} g_i(x_i)\pi_A(x_A)\\ &= \sum_{A\in S_K^L}\sum_{\mathcal{X}^A}\left(\sum_{i\in A} g_i(x_i)\right)\pi_A(x_A),\end{aligned}$$

it follows that the dual of (18) is the maximization problem

$$\max_{\{g_i\}_{i\in\mathcal{Y}}} \quad \sum_{i\in\mathcal{Y}}\sum_{\mathcal{X}_i} g_i(x_i)\mu_i(x_i) - \eta\sum_{A\in S_K^L}\sum_{\mathcal{X}^A}\exp\left(\frac{1}{\eta}\left(\sum_{i\in A} g_i(x_i) - (1+c_A(x_A))\right)\right).$$

The above is equivalent to the minimization problem (19).

A.3 Analysis of Algorithm 2

Our goal in this section is to prove Theorem 9. In preparation for its proof, we state and prove a series of auxiliary results. We start recalling the definition of the Kullback–Leibler divergence between measures with possibly different total masses.

Definition 15 *Given two finite positive measures μ and ν (not necessarily with the same total mass) sharing a common finite support $\mathcal{Z}$, their KL divergence is defined as*

$$D_{KL}(\mu||\nu) := \sum_{z\in\mathcal{Z}}(\nu(z) - \mu(z)) + \sum_{z\in\mathcal{Z}}\mu(z)\log\frac{\mu(z)}{\nu(z)}.$$

Notice that when μ and ν are probability measures, the above definition coincides with the usual one. Also, like for the usual KL-divergence, $D_{KL}(\mu||\nu)$ is non-negative and is equal to 0 if and only if $\mu = \nu$.

The next lemma is a variant of Altschuler, Niles-Weed, and Rigollet (2017, Lemma 6) adapted to the setting where μ and ν are allowed to have different total masses.

Lemma 16 *Let μ and ν be finite positive measures over a finite set $\mathcal{Z}$ such that $\mu \leq \nu$. If $D_{KL}(\mu||\nu) \leq ||\mu||_1$, then*

$$D_{KL}(\mu||\nu) \geq \frac{1}{7||\mu||_1}||\mu - \nu||_1^2. \tag{25}$$

Proof Let $\overline{\mu}, \overline{\nu}$ be normalized probability vectors obtained from μ and ν, respectively. One can write

$$\begin{aligned} D_{\mathrm{KL}}(\mu||\nu) &= ||\nu||_1 - ||\mu||_1 + \sum_{z\in\mathcal{Z}} \mu(z) \log \frac{\mu(z)}{\nu(z)} \\ &= ||\mu||_1 \log ||\mu||_1 + ||\nu||_1 - ||\mu||_1 - ||\mu||_1 \log ||\nu||_1 + ||\mu||_1 D_{\mathrm{KL}}(\overline{\mu}||\overline{\nu}) \\ &= ||\mu||_1 \left(\frac{||\nu||_1}{||\mu||_1} - 1 - \log \frac{||\nu||_1}{||\mu||_1} + D_{\mathrm{KL}}(\overline{\mu}||\overline{\nu}) \right). \end{aligned}$$

Note that $\frac{||\nu||_1}{||\mu||_1} - 1 - \log \frac{||\nu||_1}{||\mu||_1}$ and $D_{\mathrm{KL}}(\overline{\mu}||\overline{\nu})$ are both non-negative. In particular, if $D_{\mathrm{KL}}(\mu||\nu) \leq ||\mu||_1$, then

$$\frac{||\nu||_1}{||\mu||_1} - 1 - \log \frac{||\nu||_1}{||\mu||_1} \leq 1.$$

With the aid of some basic calculus and algebra one can show that for those $s \in (0, \infty)$ satisfying $s - 1 - \log(s) \leq 1$ one has the lower bound $s - 1 - \log(s) \geq (s-1)^2/5$. Therefore,

$$\frac{||\nu||_1}{||\mu||_1} - 1 - \log \frac{||\nu||_1}{||\mu||_1} \geq \frac{(\frac{||\nu||_1}{||\mu||_1} - 1)^2}{5}.$$

An application of Pinsker's inequality for probability measures yields

$$D_{\mathrm{KL}}(\mu||\nu) \geq ||\mu||_1 \left(\frac{(\frac{||\nu||_1}{||\mu||_1} - 1)^2}{5} + \frac{||\overline{\mu} - \overline{\nu}||_1^2}{2} \right).$$

Finally, by the triangle inequality and Young's inequality,

$$\begin{aligned} ||\mu - \nu||_1^2 &= ||\mu||_1^2 ||\overline{\mu} - \frac{\nu}{||\mu||_1}||_1^2 \\ &\leq ||\mu||_1^2 \left(||\overline{\nu} - \frac{\nu}{||\mu||_1}||_1 + ||\overline{\mu} - \overline{\nu}||_1 \right)^2 \\ &= ||\mu||_1^2 \left(\left| \frac{||\nu||_1}{||\mu||_1} - 1 \right| + ||\overline{\mu} - \overline{\nu}||_1 \right)^2 \\ &\leq \frac{7||\mu||_1^2}{5} \left(\frac{||\nu||_1}{||\mu||_1} - 1 \right)^2 + \frac{7||\mu||_1^2}{2} ||\overline{\mu} - \overline{\nu}||_1^2. \end{aligned}$$

This completes the proof. ■

In what follows we use $\langle \cdot, \cdot \rangle$ to denote the inner product of any two vectors in the same Euclidean space. In particular, if $\mathcal{Z}$ is a finite set, $h : \mathcal{Z} \to \mathbb{R}$, and ν is a measure over $\mathcal{Z}$, then $\langle h, \nu \rangle := \sum_{z \in \mathcal{Z}} h(z)\nu(z)$. We provide a lower bound on the decrement of the energy $\mathcal{G}^L$ along the iterates in Algorithm 2.

Proposition 17 *Let $\{g^t\}_{t \in \mathbb{N}}$ be generated by Algorithm 2 and let $\mathcal{T}$ be the collection of iterates t for which the following holds:*

$$D_{KL}(\mu_i || \sum_{A \in S_K^L(i)} \mathcal{P}_{i\#}\pi_A(g^t)) \leq \|\mu_i\|_1 \quad \forall i \in \mathcal{Y}. \tag{26}$$

Then the following holds for any iterate t:

$$\mathcal{G}^L(g^t) - \mathcal{G}^L(g^{t+1}) \geq \frac{1}{7}\left(\frac{E_t}{K}\right)^2, \quad \text{if } t \in \mathcal{T}, \tag{27}$$

and

$$\mathcal{G}^L(g^t) - \mathcal{G}^L(g^{t+1}) \geq \min_{j \in \mathcal{Y}} \|\mu_j\|_1, \quad \text{if } t \notin \mathcal{T}. \tag{28}$$

Proof Consider an iterate t and let I be the greedy coordinate at $t+1$ in **Step 1** of Algorithm 2. **Step 2** of Algorithm 2 produces

$$g_I^{t+1} = g_I^t + \eta \log \mu_I - \eta \log \sum_{A \in S_K^L(I)} \mathcal{P}_{I\#}\pi_A(g^t). \tag{29}$$

It is straightforward to see that

$$\begin{aligned}
&\mathcal{G}^L(g^t) - \mathcal{G}^L(g^{t+1}) \\
&= \sum_{x_I \in \mathcal{X}_I} \left(\sum_{A \in S_K^L(I)} \mathcal{P}_{I\#}\pi_A(g^t)(x_I) - \sum_{A \in S_K^L(I)} \mathcal{P}_{I\#}\pi_A(g^{t+1})(x_I) \right) - \langle \frac{g_I^t}{\eta} - \frac{g_I^{t+1}}{\eta}, \mu_I \rangle \\
&= \sum_{x_I \in \mathcal{X}_I} \left(\sum_{A \in S_K^L(I)} \mathcal{P}_{I\#}\pi_A(g^t)(x_I) - \mu_I(x_I) \right) + \langle \log \mu_I - \log \sum_{A \in S_K^L(I)} \mathcal{P}_{I\#}\pi_A(g^t), \mu_I \rangle \\
&= D_{\mathrm{KL}}(\mu_I || \sum_{A \in S_K^L(I)} \mathcal{P}_{I\#}\pi_A(g^t)) \\
&\geq D_{\mathrm{KL}}(\mu_i || \sum_{A \in S_K^L(i)} \mathcal{P}_{i\#}\pi_A(g^t)), \quad \forall i \in \mathcal{Y}.
\end{aligned}$$

At this stage we split the analysis into two cases. First, if we assume that $t \notin \mathcal{T}$, then there is an $i \in \mathcal{Y}$ for which $D_{\mathrm{KL}}(\mu_i || \sum_{A \in S_K^L(i)} \mathcal{P}_{i\#}\pi_A(g^t)) > \|\mu_i\|_1$. In particular, this implies (28). On the other hand, if $t \in \mathcal{T}$, we can use the above chain of inequalities to obtain

$$\mathcal{G}^L(g^t) - \mathcal{G}^L(g^{t+1}) \geq \frac{1}{K} \sum_{i \in \mathcal{Y}} D_{\mathrm{KL}}(\mu_i || \sum_{A \in S_K^L(i)} \mathcal{P}_{i\#}\pi_A(g^t))$$

and apply (25), (26), and the fact that $\|\mu_i\|_1 \leq 1$ for all $i \in \mathcal{Y}$ to deduce

$$\mathcal{G}^L(g^t) - \mathcal{G}^L(g^{t+1}) \geq \frac{1}{7} \sum_{i \in \mathcal{Y}} \frac{1}{K} \|\mu_i - \sum_{A \in S_K^L(i)} \mathcal{P}_{i\#} \pi_A(g^t))\|_1^2.$$

Applying Jensen's inequality, we deduce

$$\mathcal{G}^L(g^t) - \mathcal{G}^L(g^{t+1}) \geq \frac{1}{7} \left(\sum_{i \in \mathcal{Y}} \frac{1}{K} \|\mu_i - \sum_{A \in S_K^L(i)} \mathcal{P}_{i\#} \pi_A(g^t))\|_1 \right)^2 = \frac{1}{7} \left(\frac{E_t}{K} \right)^2.$$

■

Next, we find an upper bound for the energy gap between g^t as generated by Algorithm 2 and an optimal g^*.

Proposition 18 *Let $\{g^t\}_{t \in \mathbb{N}}$ be generated by Algorithm 2 and let g^* be a minimizer for (19). Then, for all $t \geq 1$,*

$$\mathcal{G}^L(g^t) - \mathcal{G}^L(g^*) \leq \frac{\overline{R}}{\eta} E_t,$$

where we recall E_t was defined in (21) and where $\overline{R}$ was defined in (23).

In order to prove Proposition 18 we first prove some auxiliary estimates.

Lemma 19 *Let g^* be a minimizer for (19). Then, for all $i \in \mathcal{Y}$,*

$$-(L-1) + \eta \log \min_{j \in \mathcal{Y}, y \in \mathcal{X}_j} \mu_j(y) - \eta L \log(KC^* n) \leq \min_{x_i \in \mathcal{X}_i} g_i^*(x_i) \leq \max_{x_i \in \mathcal{X}_i} g_i^*(x_i) \leq 1. \quad (30)$$

Proof The first order optimality conditions for g^* imply that, for each $i \in \mathcal{Y}$ and $x_i \in \mathcal{X}_i$,

$$\sum_{A \in S_K^L(i)} \mathcal{P}_{i\#} \pi_A(g^*)(x_i) = \mu_i(x_i) \geq \min_{j \in \mathcal{Y}, y \in \mathcal{X}_j} \mu_j(y). \quad (31)$$

Expanding the left-hand side of the above inequality we get

$$\begin{aligned}
&\sum_{A \in S_K^L(i)} \mathcal{P}_{i\#} \pi_A(g^*)(x_i) \\
&= \exp\left(\frac{1}{\eta} \left(g_i^*(x_i) - (1 + c_{\{i\}}(x_i)) \right) \right) \\
&\quad + \exp\left(\frac{1}{\eta} g_i^*(x_i) \right) \sum_{A \neq \{i\} \in S_K^L(i)} \sum_{x_{A \setminus \{i\}} \in \mathcal{X}^{A \setminus \{i\}}} \exp\left(\frac{1}{\eta} \sum_{j \in A} g_j^*(x_j) - (1 + c_A(x_A)) \right) \\
&\geq \exp\left(\frac{1}{\eta} \left(g_i^*(x_i) - 1 \right) \right)
\end{aligned} \quad (32)$$

since the double summation in the second line is always non-negative and $c_{\{i\}}(x_i) = 0$. Hence,

$$\exp\left(\frac{1}{\eta}\left(g_i^*(x_i) - 1\right)\right) \leq \sum_{A \in S_K^L(i)} \mathcal{P}_{i\#}\pi_A(g^*)(x_i) = \mu_i(x_i).$$

Taking logarithms, it follows

$$\frac{1}{\eta} g_i^*(x_i) - \frac{1}{\eta} \leq \log \mu_i(x_i) \leq 0.$$

Thus,

$$\max_{x_i \in \mathcal{X}_i} g_i^*(x_i) \leq 1.$$

To get a lower bound for g_i^*, we can take logarithms in (31) and use the fact that $c_A \geq 0$ to deduce

$$\frac{g_i^*(x_i)}{\eta} \geq \log \min_{j \in \mathcal{Y}, y \in \mathcal{X}_j} \mu_j(y) - \log \sum_{A \in S_K^L(i)} \sum_{x_{A\setminus\{i\}}} \exp\left(\frac{1}{\eta} \sum_{j \in A\setminus\{i\}} g_j^*(x_j)\right) + \frac{1}{\eta}.$$

In turn, since we already know that $\max_{x_j} g_j^*(x_j) \leq 1$ for all $j \in \mathcal{Y}$, we can further obtain

$$\frac{g_i^*(x_i)}{\eta} \geq \log \min_{j \in \mathcal{Y}, y \in \mathcal{X}_j} \mu_j(y) - \log(K^L (C^* n)^L \exp(L/\eta)) + \frac{1}{\eta} = \log \min_{j \in \mathcal{Y}, y \in \mathcal{X}_j} \mu_j(y) - L \log(KC^* n) - \frac{(L-1)}{\eta}.$$

The desired lower bound now follows.

■

Lemma 20 *Let $\{g^t\}_{t \in \mathbb{N}}$ be generated by Algorithm 2. Then, for all $t \geq 1$ and all $i \in \mathcal{Y}$,*

$$-(L-1) + \eta \log \min_{j \in \mathcal{Y}, y \in \mathcal{X}_j} \mu_j(y) - \eta L \log(KC^* n) \leq \min_{x_i \in \mathcal{X}_i} g_i^t(x_i) \leq \max_{x_i \in \mathcal{X}_i} g_i^t(x_i) \leq 1. \quad (33)$$

Proof Since $g_i^0 = \mathbf{0}$ for all $i \in \mathcal{Y}$, (33) holds trivially in the case $t = 0$. Assume that (33) holds for all $s \leq t-1$. Let I be the greedy coordinate chosen in **Step 1** of Algorithm 2 at $t-1$. For all $i \neq I$, the induction hypothesis implies (33). On the other hand, thanks to (29), which guarantees that the marginal constraint is satisfied by class I, it follows

$$\sum_{A \in S_K^L(I)} \mathcal{P}_{I\#}\pi_A(g^t)(x_I) = \mu_I(x_I) \geq \min_{j \in \mathcal{Y}, y \in \mathcal{X}_j} \mu_j(y).$$

Taking logarithms and using the fact that $c_A \geq 0$ we obtain

$$\frac{g_I^t(x_I)}{\eta} \geq \log \min_{j \in \mathcal{Y}, y \in \mathcal{X}_j} \mu_j(y) - \log \sum_{A \in S_K^L(I)} \sum_{x_{A\setminus\{I\}}} \exp\left(\frac{1}{\eta} \sum_{j \in A\setminus\{I\}} g_j^{t-1}(x_j)\right) + \frac{1}{\eta}.$$

We can then use the induction hypothesis $\max_{x_j} g_j^{t-1}(x_j) \leq 1$ for all j to get the desired lower bound.

For the upper bound, we notice that

$$\begin{aligned}\mu_I(x_I) &\geq \sum_{A\in S_K^L(I)} \mathcal{P}_{I\#}\pi_A(g^t)(x_I)\\ &= \exp\left(\frac{1}{\eta}\left(g_I^t(x_I) - (1 + c_{\{I\}}(x_I))\right)\right)\\ &\quad + \sum_{A\neq\{i\}}\sum_{x_A}\exp\left(\frac{1}{\eta}\sum_{j\in A} g_j^t(x_j) - (1 + c_A(x_A))\right)\\ &\geq \exp\left(\frac{1}{\eta}\left(g_I^t(x_I) - 1\right)\right),\end{aligned}$$

from where we can deduce that $g_I(x_I) \leq 1$ for all x_I. ■

We are ready to prove Proposition 18.

Proof [Proof of Proposition 18] Recall that $\mathcal{G}^L$ is convex and differentiable. We thus have

$$\mathcal{G}^L(g^t) - \mathcal{G}^L(g^*) \leq \langle g^t - g^*, \nabla_g \mathcal{G}^L(g^t)\rangle = \sum_{i=1}^K \langle g_i^t - g_i^*, \partial_{g_i}\mathcal{G}^L(g^t)\rangle.$$

For notational convenience, in the remainder of this proof we use

$$P_i^t := \sum_{A\in S_K^L(i)} \mathcal{P}_{i\#}\pi_A(g^t).$$

Since $\partial_{g_i}\mathcal{G}^L(g^t) = \frac{1}{\eta}\left(P_i^t - \mu_i\right)$ (as can be seen from a direct computation), we obtain

$$\begin{aligned}\mathcal{G}^L(g^t) - \mathcal{G}^L(g^*) &\leq \sum_{i=1}^K \frac{1}{\eta}\langle g_i^t - g_i^*, P_i^t - \mu_i\rangle\\ &\leq \frac{\overline{R}}{\eta} E_t,\end{aligned}$$

where we have used (30) and (33) to bound $\|g_i^* - g_i^t\|_\infty$ by $\overline{R}$.

■

Lemma 21 *Suppose that $\{a_t\}_t$ is a decreasing sequence of positive numbers (finite or infinite) satisfying:*

$$a_t - a_{t+1} \geq \max\{B\delta'^2, Aa_t^2\}, \quad \forall t,$$

for positive constants A, B, δ'. Then for all T we have

$$T \leq \min_{h \ s.t. \ h\geq a_T}\left(2 + \frac{1}{Ah} + \frac{h}{B\delta'^2}\right). \tag{34}$$

Proof Let $t' \leq T$. From the fact that $a_t - a_{t+1} \geq B\delta'^2$ for all t we have

$$a_{t'} \geq a_{t'} - a_T = \sum_{t=t'}^{T-1}(a_t - a_{t+1}) \geq \sum_{t=t'}^{T-1} B\delta'^2 \geq B\delta'^2(T-t').$$

On the other hand, from $a_t - a_{t+1} \geq Aa_t^2$ we get

$$\frac{1}{a_{t'}} \geq \frac{1}{a_{t'}} - \frac{1}{a_1} = \sum_{t=1}^{t'-1}\left(\frac{1}{a_{t+1}} - \frac{1}{a_t}\right) = \sum_{t=1}^{t'-1}\frac{a_t - a_{t+1}}{a_t a_{t+1}} \geq A\sum_{t=1}^{t'-1}\frac{a_t}{a_{t+1}} \geq A(t'-1),$$

using the fact that, by the assumptions, a_t is a decreasing sequence.

Combining the above inequalities we get

$$T - 1 = (T-t') + (t'-1) \leq \frac{1}{Aa_{t'}} + \frac{a_{t'}}{B\delta'^2}.$$

In particular,

$$T - 1 \leq \min_{t'=1,\dots,T}\left(\frac{1}{Aa_{t'}} + \frac{a_{t'}}{B\delta'^2}\right).$$

To be able to obtain (34) we need to modify the above argument slightly. We will show that for any $h \geq a_T$ we have

$$T - 2 \leq \frac{1}{Ah} + \frac{h}{B\delta'^2}.$$

First, consider $h \in [a_{t'+1}, a_{t'}]$ for some $t' < T$. We modify the sequence $\{a_t\}$ by adding the extra value h in the sequence. Precisely, let

$$\tilde{a}_t := \begin{cases} a_t & \text{if } t \leq t' \\ h & \text{if } t = t'+1 \\ a_{t-1} & \text{if } t > t'+1. \end{cases}$$

Notice that

$$h = \tilde{a}_{t'+1} \geq \tilde{a}_{t'+1} - a_T = \tilde{a}_{t'+1} - \tilde{a}_{t+2} + \tilde{a}_{t'+2} - a_T \geq \tilde{a}_{t'+2} - a_T = a_{t'+1} - a_T \geq B\delta'^2(T-(t'+1))$$

where the second inequality follows from the fact that, by construction, $\tilde{a}_{t'+1} - \tilde{a}_{t'+2} \geq 0$. Likewise, we have

$$\frac{1}{h} = \frac{1}{\tilde{a}_{t+1}} \geq \frac{1}{a_{t'}} - \frac{1}{a_1} \geq A(t'-1).$$

From the above it follows that

$$T - 2 = (T - (t'+1)) + t' - 1 \leq \frac{1}{Ah} + \frac{h}{B\delta'^2}.$$

It remains to consider the case $h \geq a_1$. In this case

$$h \geq a_1 \geq a_1 - a_T \geq B\delta'^2(T-1)$$

from where it follows that

$$T-2 \leq T-1 \leq \frac{h}{B\delta'^2} \leq \frac{1}{Ah} + \frac{h}{B\delta'^2}$$

in this case as well.

■

With Propositions 17 and 18 and the above lemma in hand, we are ready to prove Theorem 9.

Proof [Proof of Theorem 9]

Let $\Delta^t := \mathcal{G}^L(g^t) - \mathcal{G}^L(g^*)$. Notice that, thanks to Proposition 17, the sequence Δ_t is decreasing in t. Let us denote by T the iteration at which the stopping criterion for Algorithm 2 is met. Notice that T is indeed finite, as can be easily verified from Proposition 17.

Let $t_1, t_2, t_3, \ldots$ be the iterations in $\mathcal{T}$, where we recall $\mathcal{T}$ was defined in Proposition 17, and let t_s be the largest element in $\mathcal{T}$ that is strictly smaller than T. If such element does not exist, it follows that all iterations before stopping are not in $\mathcal{T}$, but in that case we would have

$$T-1 \leq \lceil \frac{\mathcal{G}^L(g^0)}{\min_{i\in\mathcal{Y}}\|\mu_i\|_1} \rceil,$$

since the decrement of energy at each of these iterations is at least $\min_{i\in\mathcal{Y}}\|\mu_i\|_1$. If t_s does exist, by a similar reasoning as before there must also be a first next iteration t_{s+1} in $\mathcal{T}$ (although larger than or equal to T). Now, for any $r \leq s$ we have

$$\Delta^{t_r} - \Delta^{t_{r+1}} \geq \Delta^{t_r} - \Delta^{t_r+1} \geq \frac{1}{7}\left\{ \left(\frac{\delta'}{K}\right)^2 \vee \left(\frac{\eta\Delta^{t_r}}{K\overline{R}}\right)^2 \right\}. \tag{35}$$

Indeed, the first inequality follows from the fact that Δ_t is decreasing in t, and the second inequality follows from the fact that

$$\Delta^{t_r} - \Delta^{t_r+1} \geq \frac{1}{7}\left(\frac{E_{t_r}}{K}\right)^2 \geq \frac{1}{7}\left(\frac{\eta\Delta^{t_r}}{K\overline{R}}\right)^2,$$

thanks to Propositions 17 and 18. We can thus apply Lemma 21 to the sequence $\Delta^{t_1}, \ldots, \Delta^{t_{s+1}}$ and deduce that

$$s+1 \leq \min_{h \text{ s.t. } h\geq\Delta^{t_{s+1}}} \left\{ 2 + 7\frac{K^2\overline{R}^2}{\eta^2 h} + 7h\left(\frac{K}{\delta'}\right)^2 \right\}.$$

From the definition of t_s, we deduce that $\Delta_{t_{s+1}} \leq \Delta_T \leq \frac{\overline{R}}{\eta}E_T \leq \frac{\overline{R}}{\eta}\delta'$. Therefore, taking $h := \frac{\overline{R}}{\eta}\delta'$ we obtain

$$s+1 \leq 2 + \frac{14K^2\overline{R}}{\eta\delta'}.$$

Finally, the number of iterations not in $\mathcal{T}$ before the stopping criterion is met satisfies

$$T-(s+1) \leq \lceil \frac{\mathcal{G}^L(g^0)}{\min_{i\in\mathcal{Y}}\|\mu_i\|_1} \rceil,$$

since, again, the decrement of energy at each of these iterations is at least $\min_{i\in\mathcal{Y}}\|\mu_i\|_1$. The desired estimate on T now follows from the previous two inequalities. ■

A.4 Analysis of Round Scheme (Algorithm 3)

Proof [Proof of Theorem 12] Notice that the $\pi_A^{(i)}$'s are non-negative for all i. In addition, from the definitions of the z_i's (in particular also the fact that they are less than or equal to one) and the $\pi_A^{(i)}$'s we get

$$\mathrm{err}_i := \mu_i - \sum_{A\in S_K^L(i)} \mathcal{P}_{i\#}\pi_A^{(K)} \geq 0.$$

Hence, the $\widehat{\pi}_A$'s are non-negative. Furthermore, the collection $\{\widehat{\pi}_A : A \in S_K^L\}$ satisfies the marginal constraints, since for each $i \in \mathcal{Y}$ we have

$$\sum_{A\in S_K^L(i)} \mathcal{P}_{i\#}\widehat{\pi}_A = \sum_{A\in S_K^L(i)} \mathcal{P}_{i\#}\pi_A^{(K)} + \mathrm{err}_i = \mu_i.$$

In what follows, we obtain an upper bound on the ℓ^1 distance between the π_A's and the $\widehat{\pi}_A$'s. Letting $\pi_A^{(0)} = \pi_A$, the difference between the mass of π_A's and $\pi_A^{(K)}$'s can be written using a telescoping sum as follows:

$$\sum_{A\in S_K^L} \left(||\pi_A||_1 - ||\pi_A^{(K)}||_1 \right) = \sum_{i=1}^{K} \sum_{A\in S_K^L} \left(||\pi_A^{(i-1)}||_1 - ||\pi_A^{(i)}||_1 \right).$$

Since $\pi_A^{(1)} = \pi_A$ when $1 \notin A$, a direct computation yields

$$\begin{aligned}
&\sum_{A\in S_K^L} (||\pi_A||_1 - ||\pi_A^{(1)}||_1) \\
&= \sum_{A\in S_K^L(1)} \sum_{x_A} \pi_A(x_A) - \sum_{x_1} \left(1 \wedge \frac{\mu_1(x_1)}{\sum_{A\in S_K^L(1)} \mathcal{P}_{1\#}\pi_A(x_1)} \right) \sum_{x_{A\setminus\{1\}}} \pi_A(x_A) \\
&= \sum_{A\in S_K^L(1)} \sum_{x_1} \frac{1}{\sum_{A\in S_K^L(1)} \mathcal{P}_{1\#}\pi_A(x_1)} \left(\left\{ \sum_{A\in S_K^L(1)} \mathcal{P}_{1\#}\pi_A(x_1) - \mu_1(x_1) \right\} \vee 0 \right) \sum_{x_{A\setminus\{1\}}} \pi_A(x_A) \\
&= \sum_{x_1} \frac{1}{\sum_{A\in S_K^L(1)} \mathcal{P}_{1\#}\pi_A(x_1)} \left(\left\{ \sum_{A\in S_K^L(1)} \mathcal{P}_{1\#}\pi_A(x_1) - \mu_1(x_1) \right\} \vee 0 \right) \sum_{A\in S_K^L(1)} \sum_{x_{A\setminus\{1\}}} \pi_A(x_A) \\
&= \sum_{x_1} \left(\left\{ \sum_{A\in S_K^L(1)} \mathcal{P}_{1\#}\pi_A(x_1) - \mu_1(x_1) \right\} \vee 0 \right) \\
&= \frac{1}{2} \left(|| \sum_{A\in S_K^L(1)} \mathcal{P}_{1\#}\pi_A - \mu_1||_1 + || \sum_{A\in S_K^L(1)} \mathcal{P}_{1\#}\pi_A||_1 - ||\mu_1||_1 \right).
\end{aligned}$$

In addition, since $z_i(x_i) \leq 1$ for all $i \in \mathcal{Y}$ and all x_i, it follows that for all $i \in \mathcal{Y}$,

$$\mathcal{P}_{i\#}\pi_A^{(K)} \leq \cdots \leq \mathcal{P}_{i\#}\pi_A^{(0)} = \mathcal{P}_{i\#}\pi_A.$$

Hence, similarly as above,

$$\begin{aligned}\sum_{A\in S_K^L}\left(||\pi_A^{(i-1)}||_1 - ||\pi_A^{(i)}||_1\right) &= \sum_{x_i}\left(\left\{\sum_{A\in S_K^L(i)}\mathcal{P}_{i\#}\pi_A^{(i-1)}(x_i) - \mu_i(x_i)\right\}\vee 0\right)\\ &\leq \sum_{x_i}\left(\left\{\sum_{A\in S_K^L(i)}\mathcal{P}_{i\#}\pi_A^{(0)}(x_i) - \mu_i(x_i)\right\}\vee 0\right)\\ &= \frac{1}{2}\left(||\sum_{A\in S_K^L(i)}\mathcal{P}_{i\#}\pi_A - \mu_i||_1 + ||\sum_{A\in S_K^L(i)}\mathcal{P}_{i\#}\pi_A||_1 - ||\mu_i||_1\right).\end{aligned} \tag{36}$$

As a result,

$$\begin{aligned}&\sum_{A\in S_K^L}\left(||\pi_A||_1 - ||\pi_A^{(K)}||_1\right)\\ &\leq \frac{1}{2}\sum_{i\in\mathcal{Y}}\left(||\sum_{A\in S_K^L(i)}\mathcal{P}_{i\#}\pi_A - \mu_i||_1 + ||\sum_{A\in S_K^L(i)}\mathcal{P}_{i\#}\pi_A||_1 - ||\mu_i||_1\right).\end{aligned} \tag{37}$$

On the other hand, from (36) we also get

$$\sum_{A\in S_K^L}\left(||\pi_A^{(i-1)}||_1 - ||\pi_A^{(i)}||_1\right) \leq ||\sum_{A\in S_K^L(i)}\mathcal{P}_{i\#}\pi_A - \mu_i||_1. \tag{38}$$

Recalling the definition of $\widehat{\pi}_A$'s and using the facts that $\mu_i \geq \sum_{A \in S_K^L(i)} \mathcal{P}_{i\#}\pi_A^{(K)}$ for all $i \in \mathcal{Y}$ and $\pi_A \geq \pi_A^{(K)}$ for all $A \in S_K^L$, it follows that

$$
\begin{aligned}
&\sum_{A \in S_K^L} ||\widehat{\pi}_A - \pi_A||_1 \\
&\leq \sum_{A \in S_K^L} ||\widehat{\pi}_A - \pi_A^{(K)}||_1 + \sum_{A \in S_K^L} ||\pi_A^{(K)} - \pi_A||_1 \\
&= \sum_{i \in \mathcal{Y}} ||\widehat{\pi}_{\{i\}} - \pi_{\{i\}}^{(K)}||_1 + \sum_{A \in S_K^L} ||\pi_A^{(K)} - \pi_A||_1 \\
&= \sum_{i \in \mathcal{Y}} ||\mathrm{err}_i||_1 + \sum_{A \in S_K^L} ||\pi_A^{(K)} - \pi_A||_1 \\
&= \sum_{i \in \mathcal{Y}} (||\mu_i||_1 - ||\sum_{A \in S_K^L(i)} \mathcal{P}_{i\#}\pi_A^{(K)}||_1) + \sum_{A \in S_K^L} ||\pi_A||_1 - ||\pi_A^{(K)}||_1 \\
&= \sum_{i \in \mathcal{Y}} (||\mu_i||_1 - ||\sum_{A \in S_K^L(i)} \mathcal{P}_{i\#}\pi_A||_1) + \underbrace{\sum_{i \in \mathcal{Y}} (||\sum_{A \in S_K^L(i)} \mathcal{P}_{i\#}\pi_A||_1 - ||\sum_{A \in S_K^L(i)} \mathcal{P}_{i\#}\pi_A^{(K)}||_1)}_{\mathrm{I}} \\
&\quad + \underbrace{\sum_{A \in S_K^L} (||\pi_A||_1 - ||\pi_A^{(K)}||_1)}_{\mathrm{II}}.
\end{aligned}
$$

Let's consider the I term first. Note that for each $A \in S_K^L$, π_A and $\pi_A^{(K)}$ appear at most $|A| \leq L$ times in the sum and we also have $\pi_A \geq \pi_A^{(K)}$. Thus, by (37) and (38) we obtain

$$
\begin{aligned}
\mathrm{I} &= \sum_{i \in \mathcal{Y}} (||\sum_{A \in S_K^L(i)} \mathcal{P}_{i\#}\pi_A||_1 - ||\sum_{A \in S_K^L(i)} \mathcal{P}_{i\#}\pi_A^{(K)}||_1) \\
&= \sum_{i \in \mathcal{Y}} \sum_{A \in S_K^L(i)} \sum_{x_A \in \mathcal{X}^A} \pi_A(x_A) - \sum_{i \in \mathcal{Y}} \sum_{A \in S_K^L(i)} \sum_{x_A \in \mathcal{X}^A} \pi_A^{(K)}(x_A) \\
&= \sum_{A \in S_K^L} \sum_{i \in A} \sum_{x_A \in \mathcal{X}^A} \pi_A(x_A) - \sum_{A \in S_K^L} \sum_{i \in A} \sum_{x_A \in \mathcal{X}^A} \pi_A^{(K)}(x_A) \\
&\leq L \sum_{A \in S_K^L} (||\pi_A||_1 - ||\pi_A^{(K)}||_1) \\
&= (L-2) \sum_{A \in S_K^L} \left(||\pi_A||_1 - ||\pi_A^{(K)}||_1 \right) + 2 \sum_{A \in S_K^L} \left(||\pi_A||_1 - ||\pi_A^{(K)}||_1 \right) \\
&\leq (L-1) \sum_{i \in \mathcal{Y}} ||\sum_{A \in S_K^L(i)} \mathcal{P}_{i\#}\pi_A - \mu_i||_1 + \sum_{i \in \mathcal{Y}} (||\sum_{A \in S_K^L(i)} \mathcal{P}_{i\#}\pi_A||_1 - ||\mu_i||_1).
\end{aligned}
$$

Applying (38) to II similarly, we deduce

$$\begin{aligned}\sum_{A\in S_K^L}||\widehat{\pi}_A-\pi_A||_1 &\leq \sum_{i\in\mathcal{Y}}||\mu_i||_1-||\sum_{A\in S_K^L(i)}\mathcal{P}_{i\#}\pi_A||_1\\ &\quad+(L-1)\sum_{i\in\mathcal{Y}}||\sum_{A\in S_K^L(i)}\mathcal{P}_{i\#}\pi_A-\mu_i||_1+\sum_{i\in\mathcal{Y}}||(\sum_{A\in S_K^L(i)}\mathcal{P}_{i\#}\pi_A||_1-||\mu_i||_1)\\ &\quad+\sum_{i\in\mathcal{Y}}||\sum_{A\in S_K^L(i)}\mathcal{P}_{i\#}\pi_A-\mu_i||_1\\ &=L\sum_{i\in\mathcal{Y}}||\sum_{A\in S_K^L(i)}\mathcal{P}_{i\#}\pi_A-\mu_i||_1.\end{aligned}$$

This completes the proof. ■

A.5 Analysis of Algorithm 4

Proof [Proof of Theorem 13] Let $\{\widetilde{\pi}_A : A\in S_K^L\}$ be an output of **Step 2** of Algorithm 4. Given that $c_{\{i\}}\equiv 0$ for each i, we deduce that for every $A\in S_K^L$ we have $\mathrm{spt}(\widetilde{\pi}_A)\subseteq\{c_A<\infty\}$ and in turn also $\mathrm{spt}(\widehat{\pi}_A)\subseteq\{c_A<\infty\}$. By Theorem 12,

$$\begin{aligned}&\sum_{A\in S_K^L}\sum_{x_A\in\mathcal{X}^A}(1+c_A(x_A))\left(\widehat{\pi}_A(x_A)-\widetilde{\pi}_A(x_A)\right)\\ &\leq L\left(1+\max_{A\in S_K^L}|c_A\mathbb{1}_{c_A<\infty}|\right)\sum_{i=1}^K||\sum_{A\in S_K^L(i)}\mathcal{P}_{i\#}\widetilde{\pi}_A-\mu_i||_1.\end{aligned}\tag{39}$$

Let $\{\pi_A^*:A\in S_K^L\}$ be a set of optimal couplings for

$$\min_{\{\pi_A:A\in S_K^L\}}\sum_{A\in S_K^L}\sum_{\mathcal{X}^A}(1+c_A(x_A))\,\pi_A(x_A)\text{ s.t. }\sum_{A\in S_K^L(i)}\mathcal{P}_{i\#}\pi_A=\mu_i\quad\text{ for all } i\in\mathcal{Y}$$

and $\{\pi'_A : A\in S_K^L\}$ be an output of Algorithm 3 with input $\{\pi_A^*:A\in S_K^L\}$ and $\nu:=(\nu_1,\ldots,\nu_K)$, where

$$\nu_i:=\sum_{A\in S_K^L(i)}\mathcal{P}_{i\#}\widetilde{\pi}_A.$$

Then, since $\sum_{A\in S_K^L(i)}\mathcal{P}_{i\#}\pi_A^*=\mu_i$ for all $i\in\mathcal{Y}$,

$$\sum_{A\in S_K^L}||\pi'_A-\pi_A^*||_1\leq L\sum_{i=1}^K||\nu_i-\mu_i||_1.\tag{40}$$

Note that $\{\widetilde{\pi}_A:A\in S_K^L\}$ is a solution for (18) when μ is replaced by ν. Since $\{\pi'_A:A\in S_K^L\}$ is also feasible for this problem,

$$\sum_{A\in S_K^L}\sum_{x_A\in\mathcal{X}^A}(1+c_A(x_A))\left(\widetilde{\pi}_A(x_A)-\pi'_A(x_A)\right)\leq\sum_{A\in S_K^L}\eta H(\widetilde{\pi}_A)-\eta H(\pi'_A).$$

Recall the log sum inequality: for $(a_1, \dots, a_n), (b_1, \dots, b_n) \in \mathbb{R}^n_+$,

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^n a_i\right) \log \frac{(\sum_{i=1}^n a_i)}{(\sum_{i=1}^n b_i)}. \tag{41}$$

Using the fact that $H(\pi'_A) \geq 0$ (since $\pi'_A \leq 1$ for all A) and applying (41) we obtain

$$\begin{aligned}
&\sum_{A \in S_K^L} H(\widetilde{\pi}_A) - H(\pi'_A) \\
&\leq \sum_{A \in S_K^L} H(\widetilde{\pi}_A) \\
&= \sum_{A \in S_K^L} \sum_{x_A \in \mathcal{X}^A} (1 - \log \widetilde{\pi}_A(x_A))\, \widetilde{\pi}_A(x_A) \\
&\leq \left(\sum_{A \in S_K^L} \sum_{x_A \in \mathcal{X}^A} \widetilde{\pi}_A(x_A)\right) - \left(\sum_{A \in S_K^L} \sum_{x_A \in \mathcal{X}^A} \widetilde{\pi}_A(x_A)\right) \log \frac{\left(\sum_{A \in S_K^L} \sum_{x_A \in \mathcal{X}^A} \widetilde{\pi}_A(x_A)\right)}{\left(\sum_{A \in S_K^L} \sum_{x_A \in \mathcal{X}^A} 1\right)} \\
&= \left(\sum_{A \in S_K^L} \sum_{x_A \in \mathcal{X}^A} \widetilde{\pi}_A(x_A)\right) - \left(\sum_{A \in S_K^L} \sum_{x_A \in \mathcal{X}^A} \widetilde{\pi}_A(x_A)\right) \log \left(\sum_{A \in S_K^L} \sum_{x_A \in \mathcal{X}^A} \widetilde{\pi}_A(x_A)\right) \\
&\quad + \left(\sum_{A \in S_K^L} \sum_{x_A \in \mathcal{X}^A} \widetilde{\pi}_A(x_A)\right) \log \left(\sum_{A \in S_K^L} \sum_{x_A \in \mathcal{X}^A} 1\right) \\
&\leq 1 + \left(\sum_{A \in S_K^L} \sum_{x_A \in \mathcal{X}^A} \widetilde{\pi}_A(x_A)\right) \log \left(\sum_{A \in S_K^L} \sum_{x_A \in \mathcal{X}^A} 1\right) \\
&\leq 1 + \frac{3L}{2} \log(C^* K n),
\end{aligned}$$

where the second to last inequality follows from the fact that $x - x \log x$ is bounded from above by 1 and the last inequality is a consequence of

$$\sum_{A \in S_K^L} ||\widetilde{\pi}_A||_1 \leq \sum_{i \in \mathcal{Y}} ||\nu_i||_1 \leq \sum_{i \in \mathcal{Y}} ||\nu_i - \mu_i||_1 + ||\mu_i||_1 \leq \frac{3}{2},$$

due to the stopping criterion $E_t := \sum_{i \in \mathcal{Y}} ||\nu_i - \mu_i||_1 < \delta' \leq \frac{1}{2}$. Hence,

$$\sum_{A \in S_K^L} \sum_{x_A \in \mathcal{X}^A} (1 + c_A(x_A)) \left(\widetilde{\pi}_A(x_A) - \pi'_A(x_A)\right) \leq \eta 2L \log(C^* K n). \tag{42}$$

Combining (40) and (42) leads to

$$\begin{aligned}
&\sum_{A\in S_K^L}\sum_{\mathcal{X}^A}(1+c_A(x_A))\left(\widetilde{\pi}_A(x_A)-\pi_A^*(x_A)\right)\\
&=\sum_{A\in S_K^L}\sum_{\mathcal{X}^A}(1+c_A(x_A))\left(\widetilde{\pi}_A(x_A)-\pi'_A(x_A)\right)+\sum_{A\in S_K^L}\sum_{\mathcal{X}^A}(1+c_A(x_A))\left(\pi'_A(x_A)-\pi_A^*(x_A)\right)\\
&\leq \eta 2L\log(C^*Kn)+L\left(1+\max_{A\in S_K^L}|c_A\mathbb{1}_{c_A<\infty}|\right)\sum_{i=1}^K||\nu_i-\mu_i||_1. \qquad (43)
\end{aligned}$$

Lastly, combining (39) and (43) leads to

$$\begin{aligned}
&\sum_{A\in S_K^L}\sum_{\mathcal{X}^A}(1+c_A(x_A))\left(\widehat{\pi}_A(x_A)-\pi_A^*(x_A)\right)\\
&\leq \eta 2L\log(C^*Kn)+2L\left(1+\max_{A\in S_K^L}|c_A\mathbb{1}_{c_A<\infty}|\right)\sum_{i=1}^K||\sum_{A\in S_K^L(i)}\mathcal{P}_{i\#}\widetilde{\pi}_A-\mu_i||_1. \qquad (44)
\end{aligned}$$

Recall the choices of η and δ':

$$\eta=\frac{\delta/2}{2L\log(C^*Kn)},\quad \delta'=\frac{\delta/2}{2L\max_{A\in S_K^L}|1+c_A\mathbb{1}_{c_A<\infty}|}.$$

Using the definition of $\mu'=(\mu'_1,\ldots,\mu'_K)$ and the fact that

$$\sum_{i\in\mathcal{Y}}||\mu'_i-\sum_{A\in S_K^L(i)}\mathcal{P}_{i\#}\widetilde{\pi}_A||_1\leq\frac{\delta'}{2},$$

we obtain

$$\sum_{i\in\mathcal{Y}}||\mu_i-\sum_{A\in S_K^L(i)}\mathcal{P}_{i\#}\widetilde{\pi}_A||_1\leq\sum_{i\in\mathcal{Y}}||\mu_i-\mu'_i||_1+||\mu'_i-\sum_{A\in S_K^L(i)}\mathcal{P}_{i\#}\widetilde{\pi}_A||_1\leq\delta'.$$

Applying the above inequality, and using our choices of η and δ' in (44), we obtain

$$\sum_{A\in S_K^L}\sum_{\mathcal{X}^A}(1+c_A(x_A))\left(\widehat{\pi}_A(x_A)-\pi_A^*(x_A)\right)\leq\delta.$$

Now, it remains to bound the computational complexity of Algorithm 4. Using (22) and the definition of $\overline{R}$ we see that Algorithm 2 in Step 2 requires at most T iterations to stop, where

$$\begin{aligned}
T&\leq O(1)+\frac{14K^2\overline{R}}{\eta\delta'}\\
&\leq O(1)+O\left(\frac{L^2K^2\max_{A\in S_K^L}(1+c_A\mathbb{1}_{c_A<\infty})\log(C^*Kn)}{\delta^2}\right).
\end{aligned}$$

Since each iteration of Algorithm 2 requires at most $O(\mathcal{J}_L)$ operations, the total computational complexity of **Step 2** of Algorithm 4 is $O\left(\frac{L^2K^2 \max_{A\in S_K^L}(1+c_A\mathbb{1}_{c_A<\infty})|\mathcal{J}_L|\log(C^*Kn)}{\delta^2}\right)$. **Step 1** and **Step 3** of Algorithm 4 require $O(Kn)$ and $O(\mathcal{J}_L)$ operations, respectively. Therefore, the conclusion follows. ■

Acknowledgments

Authors' names are listed in alphabetic order by family name. This material is based upon work supported by the National Science Foundation under Grant Number DMS 1641020 and was started during the summer of 2022 as part of the AMS-MRC program *Data Science at the Crossroads of Analysis, Geometry, and Topology*. NGT was supported by the NSF grant DMS-2236447. MJ is supported by NSF-grant DMS-2400641. JK thanks to PIMS Kantorovich Initiative supported through a PIMS PRN and NSF-DMS 2133244.

References

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

Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011b. doi: 10.1137/100805741. URL `https://doi.org/10.1137/100805741`.

Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *Advances in neural information processing systems*, 30, 2017.

Jason M. Altschuler and Enric Boix-Adserà. Wasserstein barycenters are NP-hard to compute. *SIAM Journal on Mathematics of Data Science*, 4(1):179–203, February 2022. doi: 10.1137/21m1390062. URL `https://doi.org/10.1137/21m1390062`.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

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

Leon Bungert and Kerrek Stinson. Gamma-convergence of a nonlocal perimeter arising in adversarial machine learning. *arXiv preprint arXiv:2211.15223*, 2022.

HanQin Cai, Yuchen Lou, Daniel McKenzie, and Wotao Yin. A zeroth-order block coordinate descent algorithm for huge-scale black-box optimization. In *International Conference on Machine Learning*, pages 1193–1203. PMLR, 2021.

Guillaume Carlier. On the linear convergence of the multimarginal sinkhorn algorithm. *SIAM Journal on Optimization*, 32(2):786–794, 2022.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017.

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

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

Sihui Dai, Wenxin Ding, Arjun Nitin Bhagoji, Daniel Cullina, Ben Y. Zhao, Haitao Zheng, and Prateek Mittal. Characterizing the optimal 0-1 loss for multi-class classification with a test-time attacker. *CoRR*, abs/2302.10722, 2023. URL `https://doi.org/10.48550/arXiv.2302.10722`.

Simone Di Marino and Augusto Gerolin. An optimal transport approach for the schrödinger bridge problem and convergence of sinkhorn algorithm. *Journal of Scientific Computing*, 85(2):1–28, 2020.

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

Jean Feydy, Thibault Séjourné, Francois-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR, 2019.

R. A. Fisher. Iris. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C56C76.

Nicolás García Trillos and Ryan Murray. Adversarial classification: Necessary conditions and geometric flows. *Journal of Machine Learning Research*, 23(187):1–38, 2022.

Nicolás Garcıa Trillos, Matt Jacobs, and Jakwang Kim. The multimarginal optimal transport formulation of adversarial multiclass classification. *Journal of Machine Learning Research*, 24(45):1–56, 2023a.

Nicolás Garcıa Trillos, Matt Jacobs, and Jakwang Kim. The multimarginal optimal transport formulation of adversarial multiclass classification, 2023b. URL `https://arxiv.org/abs/2204.12676`.

Nicolás Garcıa Trillos, Matt Jacobs, and Jakwang Kim. On the existence of solutions to adversarial training in multiclass classification. *European Journal of Applied Mathematics*, page 1–21, 2024. doi: 10.1017/S0956792524000822.

B. German. Glass Identification. UCI Machine Learning Repository, 1987. DOI: https://doi.org/10.24432/C5WW2P.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Parnian Kassraie, Aram-Alexandre Pooladian, Michal Klein, James Thornton, Jonathan Niles-Weed, and Marco Cuturi. Progressive entropic optimal transport solvers, 2024. URL `https://arxiv.org/abs/2406.05061`.

Justin Khim and Po-Ling Loh. Adversarial risk bounds via function transformation, 2019.

Tianyi Lin, Nhat Ho, Marco Cuturi, and Michael I Jordan. On the complexity of approximating multimarginal optimal transport. *Journal of Machine Learning Research*, 23(65): 1–43, 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Aram-Alexandre Pooladian, Marco Cuturi, and Jonathan Niles-Weed. Debiaser beware: Pitfalls of centering regularized transport maps. In *International Conference on Machine Learning*, pages 17830–17847. PMLR, 2022.

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

Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *International conference on machine learning*, pages 5231–5240. PMLR, 2019.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=Hk6kPgZA-`.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rkZvSe-RZ`.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Francois-Xavier Vialard. An elementary introduction to entropic regularization and proximal methods for numerical optimal transport. 2019.

Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=BkUHlMZ0b`.

Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *International conference on machine learning*, pages 7085–7094. PMLR, 2019.