IEEE CSS

Received XX October 2025; revised ; accepted XX October 2025; Date of publication XX Octobe 2025; date of current version XX October 2025.

Digital Object Identifier 10.1109/OJCSYS.2025.Doi Number

Machine Learning Adversarial Attacks using Partial Sinkhorn Optimization

André Bertolace [1] (Student Member, IEEE), Konstantinos Gatsis [2] (Member, IEEE), Kostas Margellos [1] (Senior Member, IEEE)

[1] Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, U.K.
[2] Electrical and Computer Engineering, Villanova University, Villanova, PA, USA.
CORRESPONDING AUTHOR: André Bertolace (email: andre.bertolace@eng.ox.ac.uk)

ABSTRACT The vulnerability of machine learning models to adversarial perturbations has motivated the development of robust optimization frameworks that ensure reliability under distributional uncertainty. In this work, we frame adversarial attacks as a Distributionally Robust Optimization problem, modeling adversarial shifts in the data distribution measured by the Wasserstein metric rather than isolated perturbations of individual samples. This formulation provides a link between adversarial robustness and optimal transport theory, enabling a more general and structured characterization of adversarial effects.

We show that this formulation naturally captures a distributionally robust approach to modeling attacks but leads to a non-convex optimization problem with linear constraints. To address the resulting computational challenges, we propose an entropic relaxation to obtain a Difference-of-Convex structure. Building on this reformulation, we develop the Partial Sinkhorn algorithm, a novel iterative method inspired by Sinkhorn-type updates that approximates local optima while guaranteeing convergence to stationary points.

Numerical experiments on synthetic and benchmark datasets demonstrate that our method yields more effective and computationally efficient adversarial attacks. Beyond adversarial learning, the proposed framework establishes a theoretical and algorithmic bridge between distributional robustness, optimal transport, and control-oriented optimization, contributing to the design of systems resilient to structured distributional shifts.

INDEX TERMS Optimization, Robust Optimization, Non-convex optimization, Machine Learning, Adversarial Learning, Optimal Transport

I. Introduction

Since the reporting of intriguing limitations in neural networks regarding their susceptibility to adversarial manipulation [1]–[3], research has increasingly focused on studying adversarial attacks [1], [3]–[10] and in addressing vulnerabilities to enhance model robustness and reliability via robust [4], [11] and adversarial training [10], [12]–[17]. Yet, even before these findings, researchers had already identified and questioned the security limitations of these models [18], [19].

In most cases, adversarial attacks have focused on minimal modifications for misclassification, approached through either penalty-based optimization or distance-specific methods, often using p-norms [20]. An alternative to using p-norms is to produce adversarial perturbations through a more flexible and general, yet computationally more demanding, approach: Distributionally Robust Optimization (DRO) procedures with Wasserstein metric constraints [21]–[27]. In fact, Adversarial Training (AT) is shown to be a special case of DRO [21].

DRO [28]–[32] provides a structured approach to integrate data with decision-making, immunizing against uncertainty in the probability distribution. Two established paradigms for handling uncertainty are Stochastic Optimization (SO) and Robust Optimization (RO), which differ fundamentally in their modeling approaches. RO represents uncertainty as deterministic variability in the parameters of a problem or its solution, aiming for performance that is reliable even in worst-case scenarios. In contrast, SO models uncertainty probabilistically, using random variables to capture randomness in objectives or constraints and often optimizing expected outcomes.

Building on these two paradigms DRO serves as a unifying framework that balances the probabilistic rigor of SO with the conservatism of RO. Unlike stochastic optimization (SO), which requires complete distributional knowledge, DRO relies on partial information. If the uncertainty set includes only the true distribution, DRO reduces to SO whereas if it includes all distributions over the support, it reduces to Robust Optimization (RO). Thus, an appropriate choice of the uncertainty set positions DRO between SO and RO, making it a less conservative alternative to RO and a unifying framework for both [31].

Particularly to adversarial attacks, adversarial training aims at achieving robustness by countering individual perturbations to each example. In contrast, DRO provides a framework for robustness by considering adversarial shifts to the entire training set [21].

A. Related work

Research on adversarial robustness in machine learning has advanced through approaches that connect Adversarial Training (AT) with Distributionally Robust Optimization (DRO). AT can be interpreted as a specific instance of DRO, as shown by [21], which also proposes an efficient DRO algorithm to improve neural network resilience against adversarial perturbations. Subsequent work applies DRO with a Wasserstein penalty to train neural networks robust to adversarial examples by accounting for worst-case data perturbations, achieving provable robustness at minimal cost and outperforming heuristic defenses for subtle attacks [22]. In addition, a related study introduces a threat model based on the Wasserstein distance that captures natural image transformations such as scaling and rotation, generating Wasserstein adversarial examples that significantly degrade model accuracy, while adversarial training partially restores robustness [23].

Adversarial Distributional Training (ADT) offers a broader immunization framework by learning adversarial distributions around natural examples through a minimax optimization approach, incorporating an entropic regularizer to prevent the inner maximization from collapsing into standard adversarial training [24]. This method expands robustness beyond AT or DRO alone. Further work reformulates DRO and regularization in deep neural networks as a calculus of variations problem, linking adversarial robustness with optimal control techniques and deriving regularized risk minimization approximations [25]. This approach models neural network layers within an infinite structure as a discretized optimal control problem, showing that regularized minimization can be viewed as a DRO problem.

A distributionally robust classification model with Wasserstein ambiguity which minimizes the conditional value-at-risk of the misclassification distance and connects with maximum-margin classifiers and previous adversarial models has also been developed [26]. Wasserstein DRO has also been applied to adversarial attacks using a novel threat model that permits non-uniform perturbations across inputs. First-order attack algorithms extend methods like Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), and an asymptotic estimate of adversarial accuracy provides out-of-sample performance guarantees [27], validated on large-scale datasets.

Furthermore, stochastic programs with limited data have been approached by constructing a Wasserstein ball around the empirical distribution and optimizing for the worst-case distribution within it, enabling tractable convex or linear reformulations under convexity assumptions and yielding strong finite-sample performance guarantees [30]. Building on these developments, [33] fruther advances the theory of DRO by introducing the Sinkhorn DRO framework. The work establishes strong duality results, characterizes continuous worst-case distributions under entropic regularization of the Wasserstein distance and proposes a stochastic mirror-descent algorithm with convergence guarantees for convex formulations.

These studies collectively highlight the close relationship between adversarial robustness and distributional robustness, motivating the development of more general frameworks that can handle non-convex objectives and complex perturbation structures.

B. Contribution

Building on these foundations [21]–[23], our approach extends distributionally robust optimization techniques to non-convex settings, thereby accommodating a wider class of objective functions. Unlike previous work that focuses primarily on convex formulations and relies on approximate projection methods [23], [30], we reformulate the problem as a Difference-of-Convex (DC) program with linear constraints and introduce an iterative algorithm to solve it efficiently.

In parallel with recent theoretical progress on entropically regularized distributionally robust optimization, particularly the Sinkhorn DRO framework [33], our work investigates similar robustness principles within the adversarial machine-learning setting. While [33] provides theoretical results on Sinkhorn-based robustness and duality, our formulation focuses on the data-driven, often non-convex nature of adversarial optimization problems. We introduce a computationally tractable framework that retains the regularization benefits of the Sinkhorn distance while operating directly on empirical data. In this context, we propose the Partial Sinkhorn algorithm, which generalizes Sinkhorn-type iterations to handle non-convex objectives and guarantees convergence to stationary points within a practical, sample-based optimization framework.

More precisely, we propose an adversarial attack based on Optimal Transport, framed as a DRO problem, which addresses the non-convex nature of the adversarial process through a tailored optimization procedure. Our main contributions are:

- *Adversarial attack modeling:* We formulate adversarial attacks in a more general and flexible manner, namely, as a DRO problem, using Optimal Transport, and provide a data-driven robustness framework that extends existing adversarial methods. This development constitutes a novel adversarial methodology, thus complementing the related literature [21].

- *Reformulation as DC-problem:* We show that this formulation leads to a non-convex optimization problem with linear constraints, a problem which is computationally hard to solve. To address this challenge, we introduce an entropic penalization term to the original problem which allows casting it as a DC-programming problem. Beyond enabling convergence, the entropic regularization mitigates the tendency of adversarial methods to further manipulate already misclassified samples solely to increase loss, encouraging instead the generation of new adversarial examples.
- *Convergent algorithm:* Exploiting the DC structure, we introduce the *Partial Sinkhorn* algorithm, which efficiently approximates local optima and whose convergence we provably guarantee. This algorithm resembles the Sinkhorn method but differs crucially in the rescaling step. As will be explained in the sequel, instead of projecting onto both fixed adversarial marginals, it projects only partially on the original marginal while iteratively adjusting toward an extremum on the adversarial one.
- *Numerical validation:* We provide an empirical analysis on both synthetic and real-world datasets, demonstrating that our approach achieves more efficient attacks compared to conventional adversarial training techniques.

Overall, our work contributes a new class of DRO-driven adversarial attacks that are both theoretically grounded and practically efficient, offering provable convergence, computational tractability, and enhanced adversarial power, as validated by empirical results.

The remainder of this paper is organized as follows. We formulate the problem in Section II formalizing both learner's and adversary's problems as well as presenting the adversary model under the lens of optimal transport, followed by a section highlighting our main results, Section III, the data-driven DRO and the partial Sinkhorn algorithm and a discussion on numerical results on Section IV.

II. Problem Formulation

Let $(\Omega, \mathcal{F}, \mu)$ be a probability space and consider the following mappings,

$$\begin{aligned} X &:(\Omega, \mathcal{F}) \to (\Xi, \mathcal{X}), \\ Y &:(\Omega, \mathcal{F}) \to (\Upsilon, \mathcal{Y}), \end{aligned}$$

taking elements of them event space Ω equipped with the appropriate σ-algebra, $\mathcal{F}$, into the spaces Ξ, Υ with their respective σ-algebras $\mathcal{X}, \mathcal{Y}$.

For each of these mappings we define the respective measures, $\mathbb{P}_X$ and $\mathbb{P}_Y$,

$$\begin{aligned} \mathbb{P}_X(A) &= \mu(X^{-1}(A)), \text{ for any } A \in \mathcal{X}, \\ \mathbb{P}_Y(B) &= \mu(Y^{-1}(B)), \text{ for any } B \in \mathcal{Y}. \end{aligned}$$

Consider further a mapping $g : \Xi \to \Xi$, resulting into a new random variable $V = g(X)$,

$$V : (\Omega, \mathcal{F}) \to (\Xi, \mathcal{V}),$$

inducing a measure,

$$\mathbb{P}_V(C) = \mu(V^{-1}(C)), \text{ for any } C \in \mathcal{V}.$$

Let us also define the product measures,

$$\begin{aligned} \mathbb{P} &= \mathbb{P}_X \times \mathbb{P}_Y, \\ \mathbb{Q} &= \mathbb{P}_V \times \mathbb{P}_Y. \end{aligned}$$

The learner and adversary are playing a game in which the learner wishes to find the hypothesis, h from a class $\mathcal{H}$, that minimizes the risk, $\mathcal{R}_{\mathbb{P}}\big[\ell(h(X), Y)\big]$. Each hypothesis $h \in \mathcal{H}$ being a function mapping $\Xi \to \Upsilon$, where $\Xi \subseteq \mathbb{R}^d$ represents the domain of features and Υ the domain of response variables. Conversely, the adversary's goal is to find an adversarial policy, g, which maximizes the risk, $\mathcal{R}_{\mathbb{P}}\big[\ell(h(g(X)), Y)\big]$. The only requirement imposed here is that the adversary can only tamper with the input X up to a certain level ζ.

In this work we will focus on the adversary's problem, however we fell insightful to introduce the learner's problem as well as this will present the necessary context for the adversary's problem.

A. The learner's problem

In the learning problem, the learner wishes to find the best hypothesis in a hypothesis class, $h \in \mathcal{H}$. In classification context such a hypothesis can be simply termed classifier.

The learner proceeds by finding the hypothesis that minimizes a certain risk, i.e.,

$$\inf_{h \in \mathcal{H}} \mathcal{R}_{\mathbb{P}}\big[\ell(h(X), Y)\big], \tag{1}$$

where $\ell : \Upsilon \times \Upsilon \to \mathbb{R}_+$ is a loss function and $\mathcal{R}_{\mathbb{P}}\big[\cdot\big]$ is a functional quantifying the risk. Often, the risk is taken to be the expected value associated with $\mathbb{P}$, leading to, the following problem,

$$\inf_{h \in \mathcal{H}} \mathbb{E}_{\mathbb{P}}\big[\ell(h(X), Y)\big] = \inf_{h \in \mathcal{H}} \int_{\Xi \times \Upsilon} \ell(h(x), y) d\mathbb{P}(x, y). \tag{2}$$

As the learner does not know the distribution $\mathbb{P}$, but has access to samples $S = ((x_1, y_1), \ldots (x_n, y_n))$, the problem is often solved empirically, by means of empirical risk minimization (ERM) framework that results in,

$$\inf_{h \in \mathcal{H}} \sum_{i=1}^{n} \ell(h(x_i), y_i). \tag{3}$$

Such empirical approaches are fundamental in statistical learning, providing a practical way to estimate the best hypothesis from data.

B. The adversary's problem

The adversary's goal is to find an algorithm, or attack, $g : \Xi \to \Xi$, not necessarily linear, that tampers with the input such that it increases the loss relative to the unbiased data. This attack results in a new random map $V = g(X)$, and induces the measures $\mathbb{P}_V$ and $\mathbb{Q}$ already introduced in the previous section.

More precisely, the adversary aims to maximize the loss while tampering with the input X up to a certain level ζ,

$$\begin{aligned} \sup_{g \in \mathcal{G}} \quad & \mathcal{R}_{\mathbb{P}}\big[\ell(h(g(X)), Y)\big] \\ \text{s.t.} \quad & \|g(X) - X\| \leqslant \zeta. \end{aligned} \tag{4}$$

In which parameter ζ can be thought of as the power of the adversary.

Similarly to the learner, the adversary also does not know the distribution $\mathbb{P}$, but has access to sample points $S = ((x_1, y_1), \dots (x_n, y_n))$. The problem is then solved empirically, with the risk as the expected value associated with $\mathbb{P}$.

C. Adversary model under the lens of optimal transport

A related problem was proposed in an optimal transport context by Monge [34], whose goal was to find the measurable map $g : \Xi \to \Xi$, called transport map, that transforms one distribution $\mathbb{P}$ of mass into another $\mathbb{Q}$ (pushes $\mathbb{P}$ onto $\mathbb{Q}$) while minimizing a cost function. Later on, Kantorovich [35] proposed an alternative, relaxed formulation, which uses the concept of transportation plan by considering the set of joint distributions $\Pi(\mathbb{P}, \mathbb{Q})$ with marginals $\mathbb{P}$ and $\mathbb{Q}$.

Now, restricting to the space $\mathcal{M}(\Xi)$ of measures supported on Ξ with finite p-moment, $\int_\Xi |x|^p d\mathbb{P}(x) < \infty$, the p-Wasserstein, W_p for $p \geqslant 1$ is defined by,

$$W_p(\mathbb{P}, \mathbb{Q}) = \left(\inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\Xi \times \Xi} |x - v|^p d\pi(x, v) \right)^{\frac{1}{p}}. \tag{5}$$

Specifically for $p = 1$, the Wasserstein metric is also known as the Monge-Rubinstein [36], [37] metric, or the earth mover distance [36]. The usual adversarial constraint, as defined in (4), that the adversary can only tamper with the input X up to a certain level ζ, is equivalent to, finding $\mathbb{Q}$ [21], such that,

$$W_1(\mathbb{P}, \mathbb{Q}) \leqslant \zeta,$$

for a given ζ.

These ideas inspire the following reformulation of (4),

$$\begin{aligned} \sup_{\mathbb{Q}} \quad & \mathbb{E}_{\mathbb{Q}}\big[(\ell(h(V), Y)\big] \\ \text{s.t.} \quad & W_1(\mathbb{P}, \mathbb{Q}) \leqslant \zeta, \end{aligned} \tag{6}$$

i.e., the adversary is now seeking for the worst case distribution that tampers with the original data.

As neither the learner nor the adversary have access to $\mathbb{P}$ but instead they observe its empirical version denoted by $\widehat{\mathbb{P}}_n$ through the samples $S = ((x_1, y_1), \dots, (x_n, y_n))$. Hence we proceed by taking the Wasserstein norm in relation to the empirical distribution resulting in,

$$\begin{aligned} \sup_{\mathbb{Q}} \quad & \mathbb{E}_{\mathbb{Q}}\big[(\ell(h(V), Y)\big] \\ \text{s.t.} \quad & W_1(\widehat{\mathbb{P}}_n, \mathbb{Q}) \leqslant \zeta, \end{aligned} \tag{7}$$

where $\ell(h(V), Y)$ may be a convex loss function, resulting in a non-convex optimization problem [1].

Such a data-driven DRO has been explored with limited data using a Wasserstein uncertainty set, optimizing decisions for worst-case scenarios [30]. The authors have shown that under convexity assumptions, these complex problems can be simplified into convex or linear programs with strong performance guarantees. In our context, the problem is generally non-convex, since we allow for a broad class of loss functions naturally arising in such settings. Nevertheless, we extend [30] and reformulate the problem as data-driven Difference Convex (DC) program with linear constraints, and then apply DC programming tools to derive an iterative solution algorithm.

III. Main Results

We present our results in two main parts, the first focusing on establishing a relationship among different optimization problems related to the adversary problem and the other presenting and discussing a numerical algorithm to solve the proposed optimization.

A. Data-driven DRO

Our primary objective is to solve a more flexible and general formulation of an adversarial attack expressed as a non-convex DRO problem,

$$\begin{aligned} D^* := \sup_{\mathbb{Q}} \quad & \mathbb{E}_{\mathbb{Q}}\big[\ell(h(V), Y)\big] \\ \text{s.t.} \quad & W_1(\widehat{\mathbb{P}}_n, \mathbb{Q}) \leqslant \zeta. \end{aligned} \tag{8}$$

However, this infinite-dimensional problem is generally intractable and difficult to solve in practice. To overcome this challenge, we turn to a finite-dimensional, data-driven surrogate,

$$\begin{aligned} L^* := \sup_{\{v_i \in \Xi\}_{i=1}^n} \quad & \frac{1}{n} \sum_{i=1}^n \ell(h(v_i), y_i) \\ \text{s.t.} \quad & \frac{1}{n} \sum_{i=1}^n \|x_i - v_i\|_1 \leqslant \zeta. \end{aligned} \tag{9}$$

These two problems are related in that replacing the infinite-dimensional ambiguity set with its finite, empirical counterpart, the data-driven surrogate (9) becomes a tighter version of the original formulation (8). This tightening arises due to the problem's non-convexity (strong duality does not hold). The result is formalized by,

Proposition III.1. *Consider a convex, lower-semi continuous loss $\ell(h(\cdot), y)$ and a bounded, closed space $\Xi \subset \mathbb{R}^d$ and*

[1] *Well-posedness:* As a technical remark, although our formulation is similar to other problems in the DRO literature, it differs in the definition of the Wasserstein metric constraint, as in our case, $\mathbb{P}$ and $\mathbb{Q}$ are already joint distributions. Considering problem (7) and the definition in (5), a natural question involves whether the problem is well-defined regarding the constraint $W_1(\mathbb{P}, \mathbb{Q})$ and the existence of $\pi \in \Pi(\mathbb{P}, \mathbb{Q})$. The Gluing Lemma [37], stated in Appendix A, ensures the existence of such a coupling, confirming that the set is non-empty.

$\Upsilon \subset \mathbb{R}$, then, problems (8) and (9) are related such that,

$$L^* \leqslant D^*.$$

Proof:
The proof explores duality principles and the structure of the Wasserstein metric and explores the fact that the problem is separable and can be cast as a point-wise optimization problem. The arguments used here follow closely those used by [30] together with results presented in [38] but differing from the former work in that we consider a convex $\ell(h(\cdot), y)$. For the full proof see Appendix B. ■

However, as we shall see later in the manuscript, while optimizing (9) naturally leads to an increased loss, the process may counterintuitively emphasize samples already misclassified raising their loss without actually increasing the number of misclassified instances, opposing the adversary's goal of generating new misclassifications. To address this, we introduce a transport-based formulation that allows incorporating a penalization term, which guides the optimization toward genuinely expanding misclassification rather than merely amplifying the loss.

For this reason, we build upon Proposition III.1 and lift the finite-dimensional surrogate (9) by introducing a transport plan, P_i, mapping x_i to v_i, resulting in,

$$\begin{aligned} K^* := \sup_{\substack{\{v_i \in \Xi\}_{i=1}^n, \\ \{P_i \in [0,1]^{d\times d}\}_{i=1}^n}} \quad & \frac{1}{n}\sum_{i=1}^{n} \ell(h(v_i), y_i) \\ \text{s.t.} \quad & \frac{1}{n}\sum_{i=1}^{n} \langle P_i, C\rangle_F \leqslant \zeta \\ & P_i \cdot \mathbf{1} = x_i, \forall i = 1, \dots, n \\ & P_i^T \cdot \mathbf{1} = v_i, \forall i = 1, \dots, n, \end{aligned} \tag{10}$$

where $\langle A, B\rangle_F = \sum_i \sum_j A_{ij} B_{ij}$ is the Frobenius inner product and C is an appropriate cost matrix, related to the Wasserstein metric.

For the problem at hand, this lifted formulation (10) is equivalent to the original surrogate (9) as formalized by,

Proposition III.2. *Let x and v be non-negative and normalized, $\sum_i x_i = \sum_j v_j = 1$ and let C be a cost matrix such that $C_{ij} = 2 \cdot \mathbb{1}_{i \neq j}$, then, solving (10) is equivalent to solving (9), that is,*

$$K^* = L^*.$$

Proof:
The proof exploits the fact that the transport constraint in our specific data-driven DRO reduces itself to a discrete Kantorovic transport problem, and shows the equivalence of the total-variation and the Kantorovich problem. It is worth noting that the result does not hold in the continuous setting, as there exists extremal points in $d\pi$ which are not concentrated on any graph [39]. For the full proof see Appendix C. ■

$$\begin{aligned} D_\lambda^* := \inf_{\substack{\{v_i \in \Xi\}_{i=1}^n, \\ \{P_i \in [0,1]^{d\times d}\}_{i=1}^n}} \quad & \frac{1}{n}\sum_{i=1}^{n} -\ell(h(v_i), y_i) + \\ & \frac{\lambda}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{k=1}^{d} P_{i_{jk}} \log P_{i_{jk}} \\ \text{s.t.} \quad & \frac{1}{n}\sum_{i=1}^{n} \langle P_i, C\rangle_F \leqslant \zeta \\ & P_i \cdot \mathbf{1} = x_i, \forall i = 1, \dots, n \\ & P_i^T \cdot \mathbf{1} = v_i, \forall i = 1, \dots, n. \end{aligned} \tag{11}$$

The advantage of (10) is that it naturally allows us to introduce an entropic penalization term which we use to control the algorithm's tendency to tamper with already misclassified samples as further tampering those prove to be better in increasing the loss function, but not necessarily increase misclassification.

Remark III.1. *Combining the result of both previous propositions we have that,*

$$K^* \leqslant D^* \leqslant D_\lambda^*.$$

An adversary who solves the penalized problem (11) instead of the original one (8) ends up with a less effective attack, but with greater computational efficiency. As we shall see in the numerical experiment session, despite this lower effectiveness, the adversary still shows enhanced adversarial power, that is, achieves a higher misclassification with less need to tamper with the data when compared to the FGSM[2] [3].

B. Partial Sinkhorn Optimization Algorithm

Having established the link between (8) and (11), we now focus on developing an algorithmic solution for the latter, whose non-convexity renders it challenging to optimize directly. Accordingly, in this subsection we derive an iterative algorithm (Algorithm 1) to solve the non-convex penalized formulation (11) and provide a formal convergence guarantee to a stationary point, as stated in Proposition III.3.

First note that (11) is a Difference Convex (DC) problem,

$$\begin{aligned} \inf_{x \in C} \quad & f_0(x) - g_0(x) \\ \text{s.t.} \quad & f_i(x) - g_i(x) \leqslant 0 \\ & i = 1, \dots, n, \end{aligned} \tag{12}$$

[2] FGSM is an one-step adversarial attack that perturbs an input sample in the direction of the gradient of the loss function to maximize the model's prediction error by solving,

$$\begin{aligned} \operatorname*{argsup}_{\delta} \quad & \ell(h(x+\delta), y) \\ \text{s.t.} \quad & \|\delta\|_\infty \leqslant \varepsilon, \end{aligned}$$

in which ϵ controls the magnitude of the perturbation. Given a model with loss function $\ell(h(\cdot), y)$, the adversarial example v is generated in one step,

$$v = x + \varepsilon \cdot \operatorname{sign}(\nabla_x \ell(h(x), y)),$$

This small, simple and carefully crafted change can mislead the model while remaining nearly imperceptible to the learner.

with,

$$
\begin{aligned}
x &= [v_1, \dots, v_n, P_1, \dots, P_n], \\
f_0(x) &= \frac{\lambda}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \sum_{k=1}^{d} P_{i_{jk}} \log P_{i_{jk}}, \\
g_0(x) &= \frac{1}{n} \sum_{i=1}^{n} \ell(h(v_i), y_i), \\
f_i(x) &= P_i \cdot \mathbf{1} - x_i, \forall i = 1, \dots n \\
f_{n+i}(x) &= P_i^T \cdot \mathbf{1} - v_i, \forall i = 1, \dots n \\
f_{2n+1}(x) &= \frac{1}{n} \sum_{i=1}^{n} \langle P_i, C \rangle_F - \zeta
\end{aligned}
$$

$$g_i(x) = 0, \forall i = 1, \dots n$$

and $C = (\Xi \times [0,1]_{d \times d})^n$ is a convex set.

The class of DC functions is very broad, for instance, $\mathcal{C}^2$ functions can be expressed as a difference of convex functions [40]. This kind of problem has been addressed in the literature and solved using DC algorithms (DCA) [41], [42]. Alternatively, one can use the Convex Concave Procedure (CCP) which finds a local optimal [43] of (12). CCP can be seen as a version of DCA that linearizes the concave functions instead of solving a dual problem as DCA does. That is, starting with an initial value x_0, the CCP algorithm iteratively solves,

$$
\begin{aligned}
x^{(k+1)} \leftarrow \underset{x \in C}{\operatorname{argmin}} \quad & f_0(x) - \big(g_0(x^{(k)}) + \\
& \qquad \nabla_{g_0}^T(x^{(k)})(x - x^{(k)})\big) \\
\text{s.t.} \quad & f_i(x) - \big(g_i(x^{(k)}) + \\
& \qquad \nabla_{g_i}^T(x^{(k)})(x - x^{(k)})\big) \leqslant 0 \\
& \forall i = 1, \dots, n,
\end{aligned}
\tag{13}
$$

until a tolerance is achieved.

We proceed in a CCP fashion by iteratively solving the following optimization problem, with $\{v_i^{(k+1)}\}_{i=1}^n$ being the argument that solves,

$$
\begin{aligned}
\min_{\substack{\{v_i \in [0,1]^d\}_{i=1}^n, \\ \{P_i \in [0,1]^{d \times d}\}_{i=1}^n}} \quad & \frac{\lambda}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \sum_{k=1}^{d} P_{i_{jk}} \log P_{i_{jk}} \\
& - \frac{1}{n} \sum_{i=1}^{n} \nabla_{\ell(h((v_i^{(k)}),y_i)}^T (v_i^{(k)})(v_i - v_i^{(k)}) \\
\text{s.t.} \quad & \frac{1}{n} \sum_{i=1}^{n} \langle P_i, C \rangle_F \leqslant \zeta \\
& P_i \cdot 1 = x_i, \forall i = 1, \dots, n \\
& P_i^T \cdot 1 = v_i, \forall i = 1, \dots, n,
\end{aligned}
\tag{14}
$$

where we established $\Xi = [0,1]^d$.

This is equivalent so solving the following,

$$
\begin{aligned}
\min_{\substack{\{v_i \in [0,1]^d\}_{i=1}^n, \\ \{P_i \in [0,1]^{d \times d}\}_{i=1}^n}} \quad & \frac{\lambda}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \sum_{k=1}^{d} P_{i_{jk}} \log P_{i_{jk}} \\
& - \frac{1}{n} \sum_{i=1}^{n} \nabla_{\ell(h((v_i^{(k)}),y_i)}^T (v_i^{(k)}) v_i \\
\text{s.t.} \quad & \frac{1}{n} \sum_{i=1}^{n} \langle P_i, C \rangle_F \leqslant \zeta \\
& P_i \cdot 1 = x_i, \forall i = 1, \dots, n \\
& P_i^T \cdot 1 = v_i, \forall i = 1, \dots, n.
\end{aligned}
\tag{15}
$$

To illustrate the need for the penalization term, consider for a moment the non-penalized problem, that is, take $\lambda = 0$,

$$
\begin{aligned}
\min_{\substack{\{v_i \in [0,1]^d\}_{i=1}^n, \\ \{P_i \in [0,1]^{d \times d}\}_{i=1}^n}} \quad & - \frac{1}{n} \sum_{i=1}^{n} \nabla_{\ell(h((v_i^{(k)}),y_i)}^T (v_i^{(k)}) v_i \\
\text{s.t.} \quad & \frac{1}{n} \sum_{i=1}^{n} \langle P_i, C \rangle_F \leqslant \zeta \\
& P_i \cdot 1 = x_i, \forall i = 1, \dots, n \\
& P_i^T \cdot 1 = v_i, \forall i = 1, \dots, n,
\end{aligned}
\tag{16}
$$

This is a Linear Program (LP) that can be efficiently solved using standard convex optimization solvers. When employing convex, smooth, and differentiable loss functions, such as Binary Cross Entropy, the gradient naturally amplifies as the loss increases, pushing the adversary to adjust in the direction that increases loss. While this behavior aligns with the attacker's goal of promoting misclassification, it can undermine the attacker's effectiveness as the optimization tends to push already misclassified samples deeper into the wrong class, merely increasing the expected loss rather than expanding the set of misclassified instances. This behavior focuses on maximizing loss magnitude instead of achieving true misclassification, thereby undermining the adversary's core goal. We illustrate and discuss this unintended effect in detail using low-dimensional synthetic datasets in Section A.

The entropic penalization term allows us to regulate the adversarial deviation from the original samples, counteracting excessive drift and maintaining meaningful perturbations. A further advantage of this formulation is its computational efficiency: it can be solved by leveraging ideas from the Sinkhorn-Knopp matrix scaling algorithm [44], [45] and its projected variant [23], thereby avoiding the need to solve a constrained optimization problem directly. The classical Sinkhorn algorithm iteratively rescales the rows and columns of a transport matrix so that their sums match the prescribed source and target marginals, effectively producing a balanced (doubly stochastic) transport plan.

In our case, we modify the classical Sinkhorn iterations by performing only partial rescaling to match the source distribution while optimizing over the target, resulting in a new method that we term the *Partial Sinkhorn* algorithm.

Algorithm 1 results from applying the iterative CCP procedure combined with the Partial Sinkhorn algorithm to maximize the loss. Its convergence to a stationary point of the penalized formulation is established in the following proposition,

Proposition III.3. *The limit of any convergent subsequence of $\{v_i^{(k)}\}_{k=0}^{\infty}$ generated by Algorithm 1, which iteratively solves (15), is a stationary point $\{v_i^*\}_{i=1}^n$ of problem (11).*

Proof:
The algorithm follows by formulating and solving the dual problem and its convergence leverages Zangwill's global convergence theorem. The convergence properties of CCP have been briefly discussed in [43] and explored in detail in [46]. In the following, we draw upon concepts from these previous works, particularly Zangwill's global convergence theorem [47], which serves as the foundation of the proof provided in [46]. The full proof is on Appendix D. ■

Algorithm 1 Partial Sinkhorn CCP procedure

```
function CCP($\ell$, $h$, $C$, $\lambda$, $v_1^{(0)}, \dots, v_n^{(0)}$)
    $l \leftarrow 0$
    repeat
        $\beta_i \leftarrow -\lambda \nabla_{\ell(h((v_i^{(l)}),y_i)}(v_i^{(l)})$
        $\alpha_i \leftarrow \log \frac{1}{n}$
        $\gamma \leftarrow 1$
        repeat
            for $i = 1, \dots, n$ do
                for $j = 1, \dots, d$ do
                    $\alpha_{i_j} \leftarrow \log\left(\sum_{k=1}^{d} e^{-\beta_{i_k} - \gamma C_{jk} - 1}\right)$
                            $- \log x_{i_j}$
                    $P_{i_{jk}} \leftarrow e^{-\alpha_{i_j} - \beta_{i_k} - \gamma C_{jk} - 1}$
                end for

            end for
            $\mathcal{L}_\gamma \leftarrow \sum_{i=1}^{n} \sum_{j=1}^{d} \sum_{k=1}^{d} C_{jk} P_{i_{jk}} - n\zeta$
            $\mathcal{L}_{\gamma\gamma} \leftarrow -\sum_{i=1}^{n} \sum_{j=1}^{d} \sum_{k=1}^{d} C_{jk}^2 P_{i_{jk}}$
            $\gamma \leftarrow \gamma - \frac{\mathcal{L}_\gamma}{\mathcal{L}_{\gamma\gamma}}$
        until stopping criteria
        $v_i^{(l+1)} \leftarrow P_i^T \cdot 1$
    until stopping criteria
    return $\{P_i\}_{i=1}^n, \{v_i^{(l+1)}\}_{i=1}^n$
end function
```

Remark III.2 (Interpretation)**.** *The proposed algorithm is similar to the Projected Sinkhorn algorithm [23], the α_i step rescales the rows of $e^{\gamma C - 1}$ to sum up to x_i while γ ensures the budget constraint is satisfied. However, unlike the Projected Sinkhorn version, in which β_i rescales the columns of $e^{\gamma C - 1}$ to sum up to v_i, in the present algorithm β_i rescales the columns of $e^{\gamma C - 1}$ while approaching an extremal, in our case, the minimum, as the problem is a linear program.*

IV. Numerical Examples

We applied the methodology to generate adversarial data and perform adversarial training considering different models and datasets ranging from synthetic to real datasets.

A. Synthetic Data

We began by examining synthetic data to analyze the algorithm and visualize its behavior in low dimensions. Figure 2 shows the results of simulations generating linearly separable classes (blue and red) in 2D space. We sampled 1000 data points and fitted a linear model by minimizing Binary Cross Entropy (BCE) loss with a sigmoid layer, using batch Stochastic Gradient Descent across multiple epochs. We continued by generating DRO and penalized DRO adversarial samples considering that the adversary could disturb the sample up to norm 0.1.

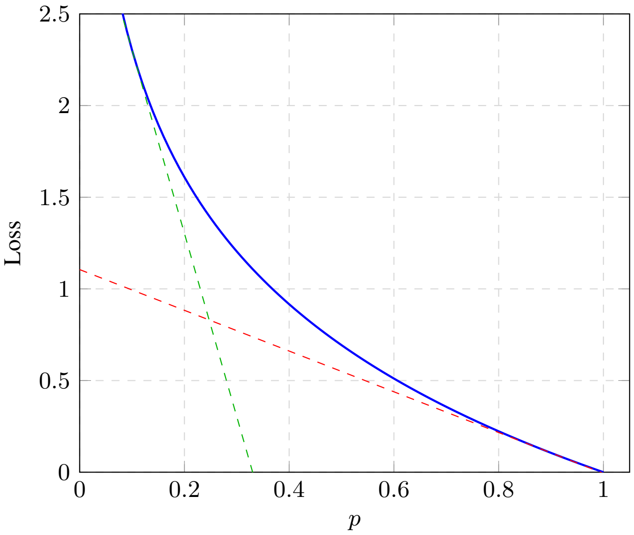
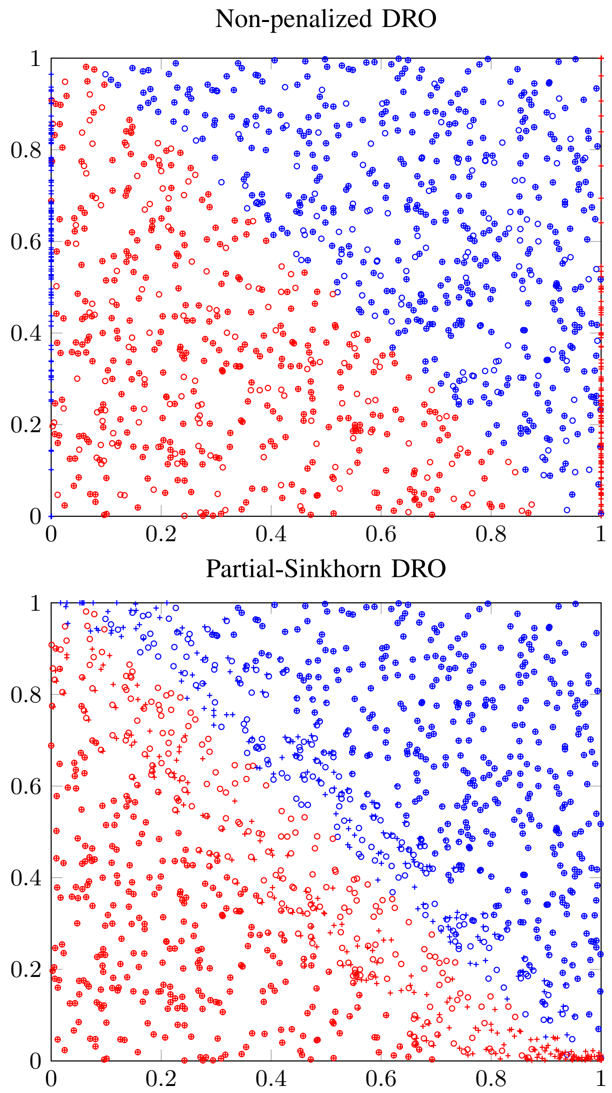


Figure 1: BCE loss and gradient effect: derivative when correctly classifying (dashed-red) and when incorrectly classifying (dashed-green).

Because losses normally used in classification, such as the BCE loss, exhibit increasingly larger gradients for poorly classified samples than to correctly classified ones (see Figure 1), the iterative algorithm prioritize updates on these misclassified samples, further adjusting them at each iteration. Consequently, these samples are pushed toward reaching a constraint, while the algorithm allocates less attention to other samples. This is counter productive as the adversary wishes to maximize misclassifications, not only the loss.

This result of this effect is visible in the top image of Figure 2. As highlighted in Section B, to counter this, we introduced a penalization term, resulting in the more stable behavior shown at the bottom of Figure 2, where we see that points close to the boundary just move over the boundary so that they are misclassified instead of being pushed to extreme values.


Non-penalized DRO

Partial-Sinkhorn DRO

Figure 2: Effect of entropic penalization illustrated in a synthetic dataset for a binary classifier, 2D, linearly separable: blue/red ○ (samples class 0/1), blue/red + (adversarial samples 0/1). Overlapping points are a result of the attacker not tampering with that specific sample.

B. Real Datasets

We continue our numerical studies by applying the same methodology on the MNIST data-set. We start by training a classifier h on a training set, and proceed by attacking the trained model on a separate test set. We then compare both accuracy, measured as the proportion of correctly classified instances on an out-of-sample adversarial test set, and the average absolute deviation, $\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}|x_{ij}-v_{ij}|$, from the sample set, a metric quantifying the perturbation introduced by the adversary in performing the attack. This metric is directly proportional to the adversarial power ζ but present itself as a more fair metric to compare attacks which are different in nature. Notably, lower values of this metric indicate less tampering with the data.

1) Linear binary classifier

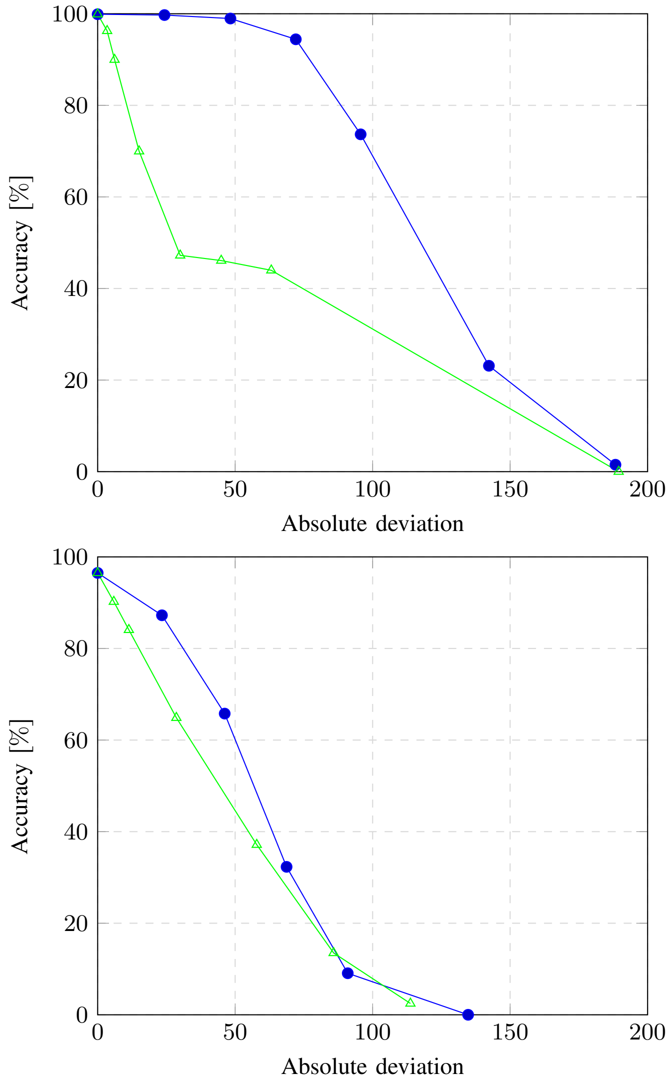


Figure 3: Accuracy of linear classifiers using out-of-sample adversarial tampered data considering non-adversarial training, FGSM (blue) and proposed Partial-Sinkhorn (green). Datasets: MNIST 0/1 (top), MNIST 3/8 (bottom). Figures obtained using logistic loss as surrogate.

Figure 3 shows the results of an experiment considering a linear classifier, $h(x) = a^T x + b$ using the logistic-loss (depicted in Figure 1). Note that in this case, the resulting problem is concave, as per our previous assumptions. We conducted a series of experiments on the MNIST dataset, focusing on distinguishing between two digit pairs: 0/1 and 3/8. The results (see Figure 3) compare our approach with FGSM. It can be observed that the proposed Partial-Sinkhorn method outperforms FGSM, achieving lower accuracy levels (the attacker's objective) for the same degree of data perturbation (measured by absolute deviation). This advantage is particularly pronounced for realistic adversaries, those with just enough power to induce misclassification without significantly distorting the image (that is, at lower levels of absolute deviation), while its performance remains comparable to other approaches when considering more powerful adversaries.

V. Conclusion

In this work, we introduced a new approach to understanding and defending against adversarial attacks by framing them as a problem of distributional shift, rather than just small input changes. Using tools from Optimal Transport and Distributionally Robust Optimization (DRO), we developed a method that considers how entire data distributions can be subtly shifted to fool a model.

This perspective leads to a more complex, non-convex optimization problem, which we addressed by introducing a relaxation and designing a new algorithm inspired by the Sinkhorn method. Our algorithm, called Partial Sinkhorn, comes with theoretical guarantees and is both efficient and practical.

We also showed that common loss functions used in adversarial settings can sometimes unintentionally focus on already misclassified data points, rather than creating new, effective adversarial examples. Our method overcomes this by encouraging perturbations that are more meaningful and more likely to cause true misclassifications.

Experiments on synthetic and real-world datasets confirmed that our approach improves model robustness compared to traditional adversarial training methods. Overall, this work provides a principled and effective way to design adversarial attacks and defenses by combining ideas from Optimal Transport and robust optimization, offering both theoretical insights and practical benefits.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2014.

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, (Cambridge, MA, USA), p. 2672–2680, MIT Press, 2014.

[3] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015.

[4] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317 – 331, 2018.

[5] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," *CoRR*, vol. abs/1810.00069, 2018.

[6] R. S. S. Kumar, D. R. O'Brien, K. Albert, S. Viljöen, and J. Snover, "Failure modes in machine learning systems," *CoRR*, vol. abs/1911.11034, 2019.

[7] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2805–2824, 2019.

[8] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang, "Recent advances in adversarial training for adversarial robustness," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21* (Z.-H. Zhou, ed.), pp. 4312–4321, International Joint Conferences on Artificial Intelligence Organization, 8 2021. Survey Track.

[9] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *CoRR*, vol. abs/1607.02533, 2016.

[10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2019.

[11] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 582–597, 2016.

[12] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. E. Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 7472–7482, PMLR, 09–15 Jun 2019.

[13] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu, "Improving adversarial robustness requires revisiting misclassified examples," in *International Conference on Learning Representations*, 2020.

[14] Y. Xing, Q. Song, and G. Cheng, "Why do artificially generated data help adversarial robustness," in *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 954–966, Curran Associates, Inc., 2022.

[15] T. Pang, M. Lin, X. Yang, J. Zhu, and S. Yan, "Robustness and accuracy could be reconcilable by (Proper) definition," in *Proceedings of the 39th International Conference on Machine Learning* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds.), vol. 162 of *Proceedings of Machine Learning Research*, pp. 17258–17277, PMLR, 17–23 Jul 2022.

[16] Z. Wang, T. Pang, C. Du, M. Lin, W. Liu, and S. Yan, "Better diffusion models further improve adversarial training," in *Proceedings of the 40th International Conference on Machine Learning* (A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, eds.), vol. 202 of *Proceedings of Machine Learning Research*, pp. 36246–36263, PMLR, 23–29 Jul 2023.

[17] S. Gowal, S.-A. Rebuffi, O. Wiles, F. Stimberg, D. Calian, and T. Mann, "Improving robustness using generated data," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, (Red Hook, NY, USA), Curran Associates Inc., 2024.

[18] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma, "Adversarial classification," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, (New York, NY, USA), p. 99–108, Association for Computing Machinery, 2004.

[19] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?," in *AsiaCCS* (F.-C. Lin, D.-T. Lee, B.-S. P. Lin, S. Shieh, and S. Jajodia, eds.), pp. 16–25, ACM, 2006.

[20] J. Rony, E. Granger, M. Pedersoli, and I. Ben Ayed, "Augmented lagrangian adversarial attacks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7738–7747, October 2021.

[21] M. Staib and S. Jegelka, "Distributionally robust deep learning as a generalization of adversarial training," in *NIPS Machine Learning and Computer Security Workshop*, 2017.

[22] A. Sinha, H. Namkoong, and J. Duchi, "Certifiable distributional robustness with principled adversarial training," in *International Conference on Learning Representations*, 2018.

[23] E. Wong, F. Schmidt, and Z. Kolter, "Wasserstein adversarial examples via projected Sinkhorn iterations," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 6808–6817, PMLR, 09–15 Jun 2019.

[24] Y. Dong, Z. Deng, T. Pang, J. Zhu, and H. Su, "Adversarial distributional training for robust deep learning," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 8270–8283, Curran Associates, Inc., 2020.

[25] C. A. García Trillos and N. García Trillos, "On the regularized risk of distributionally robust learning over deep neural networks," *Research in the Mathematical Sciences*, vol. 9, p. 54, Aug 2022.

[26] N. Ho-Nguyen and S. J. Wright, "Adversarial classification via distributional robustness with wasserstein ambiguity," *Mathematical Programming*, vol. 198, pp. 1411–1447, Apr 2023.

[27] X. Bai, G. He, Y. Jiang, and J. Obloj, “Wasserstein distributional robustness of neural networks,” in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[28] S. Shafieezadeh-Abadeh, P. M. Esfahani, and D. Kuhn, “Distributionally robust logistic regression,” in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS’15, (Cambridge, MA, USA), p. 1576–1584, MIT Press, 2015.

[29] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh, *Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning*, ch. 6, pp. 130–166.

[30] P. Mohajerin Esfahani and D. Kuhn, “Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations,” *Mathematical Programming*, vol. 171, pp. 115–166, Sep 2018.

[31] H. Rahimian and S. Mehrotra, “Distributionally robust optimization: A review,” 2019.

[32] H. Rahimian and S. Mehrotra, “Frameworks and results in distributionally robust optimization,” *Open Journal of Mathematical Optimization*, vol. 3, p. 1–85, July 2022.

[33] J. Wang, R. Gao, and Y. Xie, “Sinkhorn distributionally robust optimization,” *Operations Research*, vol. 0, no. 0, p. null, 0.

[34] G. Monge, *Mémoire sur la théorie des déblais et des remblais*. De l’Imprimerie Royale, 1781.

[35] L. Kantorovich and G. S. Rubinstein, “On a space of totally additive functions,” *Vestnik Leningrad. Univ*, vol. 13, pp. 52–59, 1958.

[36] S. Kolouri, S. Park, M. Thorpe, D. Slepčev, and G. K. Rohde, “Transport-based analysis, modeling, and learning from signal and data distributions,” 2016.

[37] C. Villani, *Optimal transport – Old and new*, vol. 338, pp. xxii+973. 01 2008.

[38] R. Rockafellar and R. J.-B. Wets, *Variational Analysis*. Heidelberg, Berlin, New York: Springer Verlag, 1998.

[39] C. Villani, *Topics in Optimal Transportation*. Graduate studies in mathematics, American Mathematical Society, 2003.

[40] P. Hartman, “On functions representable as a difference of convex functions,” *Pacific Journal of Mathematics*, vol. 9, pp. 707–713, 1959.

[41] P. D. Tao and E. B. Souad, “Algorithms for solving a class of nonconvex optimization problems. methods of subgradients,” *North-holland Mathematics Studies*, vol. 129, pp. 249–271, 1986.

[42] P. D. Tao and E. B. Souad, *Duality in D.C. (Difference of Convex functions) Optimization. Subgradient Methods*, pp. 277–293. Basel: Birkhäuser Basel, 1988.

[43] T. Lipp and S. Boyd, “Variations and extension of the convex–concave procedure,” *Optimization and Engineering*, vol. 17, pp. 263–287, Jun 2016.

[44] R. Sinkhorn and P. Knopp, “Concerning nonnegative matrices and doubly stochastic matrices,” *Pacific Journal of Mathematics*, vol. 21, pp. 343–348, 1967.

[45] M. Cuturi, “Sinkhorn distances: Lightspeed computation of optimal transport,” in *Advances in Neural Information Processing Systems* (C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, eds.), vol. 26, Curran Associates, Inc., 2013.

[46] G. Lanckriet and B. K. Sriperumbudur, “On the convergence of the concave-convex procedure,” in *Advances in Neural Information Processing Systems* (Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, eds.), vol. 22, Curran Associates, Inc., 2009.

[47] W. I. Zangwill, *Nonlinear programming; a unified approach, by Willard I. Zangwill*. Prentice-Hall international series in management, Englewood Cliffs, N.J: Prentice-Hall, 1969.

[48] W. Rudin, *Real and Complex Analysis*. New York: McGraw-Hill, 3 ed., 1987.

[49] M. Sion, “On general minimax theorems,” *Pacific Journal of Mathematics*, vol. 8, no. 1, pp. 171–176, 1958.

[50] A. Gunawardana and W. Byrne, “Convergence theorems for generalized alternating minimization procedures,” *Journal of Machine Learning Research*, vol. 6, no. 69, pp. 2049–2073, 2005.

Appendix

A. Well-posedness

Lemma 1 (Gluing Lemma [37]). *Let* $(\mathcal{X}_i, \mu_i)$, $i = 1, 2, 3$, *be Polish probability spaces. If* (X_1, X_2) *is a coupling of* (μ_1, μ_2) *and* (Y_2, Y_3) *is a coupling of* (μ_2, μ_3), *then one can construct a triple of random variables* (Z_1, Z_2, Z_3) *such that* (Z_1, Z_2) *has the same law as* (X_1, X_2) *and* (Z_2, Z_3) *has the same law as* (Y_2, Y_3).

B. Proof of Proposition III.1

Proof:
The optimization problem (6) in integral form becomes,

$$\begin{aligned} &\sup_{\mathbb{Q}\in\mathcal{M}(\Xi)} && \int_{\Xi\times\Upsilon} (\ell_y \circ h)(v)d\mathbb{Q}(v,y) \\ &\text{s.t.} && \inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})} \int_{\Xi\times\Upsilon\times\Xi} \|x-v\| d\pi(x,y,v) \leqslant \zeta, \end{aligned} \tag{17}$$

which can be simplified to,

$$\begin{aligned} &\sup_{\mathbb{Q}\in\mathcal{M}(\Xi),\, \pi\in\Pi(\mathbb{P},\mathbb{Q})} && \int_{\Xi\times\Upsilon} (\ell_y \circ h)(v)d\mathbb{Q}(v,y) \\ &\text{s.t.} && \int_{\Xi\times\Upsilon\times\Xi} \|x-v\| d\pi(x,y,v) \leqslant \zeta, \end{aligned} \tag{18}$$

where $\mathcal{M}(\Xi)$ the space of measures supported on Ξ with finite 1-moment, that is, $\int_\Xi |\cdot| d\mathbb{P}(\cdot) < \infty$. The equivalence holds since in (17) we seek a $\mathbb{Q}$ for which there exists a π such that the constraint in (17) holds. This is equivalent to optizmizing with respect to both as in (18).

As, neither the learner nor the adversary have access to $\mathbb{P}$ but instead they observe $\widehat{\mathbb{P}}_n$ through the samples $S = ((x_1, y_1), \ldots, (x_n, y_n))$, the Wasserstein norm is taken in relation to the empirical distribution, leading to,

$$\begin{aligned} &\sup_{\mathbb{Q}\in\mathcal{M}(\Xi),\, \pi\in\Pi(\widehat{\mathbb{P}}_n,\mathbb{Q})} && \int_{\Xi\times\Upsilon} (\ell_y \circ h)(v)d\mathbb{Q}(v,y) \\ &\text{s.t.} && \int_{\Xi\times\Upsilon\times\Xi} \|x-v\| d\pi(x,y,v) \leqslant \zeta. \end{aligned} \tag{19}$$

Now, we construct the joint distribution π by the marginal $\widehat{\mathbb{P}}_n$ and the conditional distribution $\mathbb{Q}_i(A) = \mathbb{Q}(A|X = x_i, Y = y_i)$, for any $A \in \Xi$,

$$\pi(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{Q}_i(A) \times \delta_{x_i,y_i}, \tag{20}$$

where δ_{x_i,y_i} is the Dirac mass on (x_i, y_i).

With this, and Tonelli’s theorem [48], (19) becomes,

$$\begin{aligned} &\sup_{\mathbb{Q}_i\in\mathcal{M}(\Xi)} && \frac{1}{n}\sum_{i=1}^{n}\int_\Xi (\ell_{y_i} \circ h)(v)d\mathbb{Q}_i(v) \\ &\text{s.t.} && \frac{1}{n}\sum_{i=1}^{n}\int_\Xi \|v-x_i\| d\mathbb{Q}_i(v) \leqslant \zeta. \end{aligned} \tag{21}$$

Dualizing the constraint in (21) with the Lagrange multiplier, the above is equivalent to,

$$\sup_{\{\mathbb{Q}_i\}_{i=1}^n\in\mathcal{M}(\Xi)} \inf_{\lambda\geqslant 0} \left\{ \lambda\zeta + \frac{1}{n}\sum_{i=1}^{n}\int_\Xi \Big((\ell_{y_i}\circ h)(v) - \lambda\|v-x_i\| \Big) d\mathbb{Q}_i(v) \right\} \tag{22}$$

and through duality (note that equality holds because of Sion's min-max theorem [49] - the objective function is linear on $\mathbb{Q}_i$ and λ) this is equivalent to,

$$\inf_{\lambda \geqslant 0} \sup_{\{\mathbb{Q}_i \in \mathcal{M}(\Xi)\}_{i=1}^n} \Big\{\lambda\zeta + \frac{1}{n}\sum_{i=1}^n \int_\Xi \Big((\ell_{y_i}\circ h)(v) - \lambda\|v - x_i\|\Big) d\mathbb{Q}_i(v)\Big\} \tag{23}$$

This problem is separable, that is, we can optimize for each $\mathbb{Q}_i$ independently,

$$\sup_{\mathbb{Q}_i \in \mathcal{M}(\Xi)} \int_\Xi \Big((\ell_{y_i}\circ h)(v) - \lambda\|v - x_i\|\Big) d\mathbb{Q}_i(v). \tag{24}$$

This can alternatively be written as,

$$\sup_{V_i \in \mathcal{M}(\Xi)} \int_\Omega \Big((\ell_{y_i}\circ h)(V_i(\omega)) - \lambda\|V_i(\omega) - x_i\|\Big) d\mu(\omega). \tag{25}$$

By Theorem 14.60 in [38], we can interchange the sup and the integral for spaces that are decomposable, i.e., spaces of measurable functions such that, for any function $V_0(\omega)$ in the space, any measurable set $A \in \Omega$, and any bounded measurable function $V_1(\omega)$ on A, the function,

$$V(\omega) = \begin{cases} V_1(\omega), \omega \in A, \\ V_0(\omega), \omega \notin A \end{cases}$$

also belongs to the space. Since $V_i \in \mathcal{M}(\Xi)$, where $\mathcal{M}(\Xi)$ is the space of measures supported on Ξ with finite 1-moment. The space of random variables with finite 1-moment is closed under such local patching, since,

$$\int_\Omega |V(\omega)| d\omega = \int_A |V_1(\omega)| d\omega + \int_{A^c} |V_0(\omega)| d\omega < \infty.$$

hence, it is decomposable, and the theorem applies, leading to,

$$\int_\Omega \sup_{v_i \in \Xi} \Big((\ell_{y_i}\circ h)(v_i) - \lambda\|v_i - x_i\|\Big) d\mu(\omega). \tag{26}$$

This means, that since the given distribution is discrete, we optimize over discrete masses too.

As the integrating function is independent of ω and μ is a probability measure, the above becomes,

$$\sup_{v_i \in \Xi} \Big((\ell_{y_i}\circ h)(v_i) - \lambda\|v_i - x_i\|\Big) \int_\Omega d\mu(\omega), \tag{27}$$

but μ is a probability measure, and we are left with,

$$\sup_{v_i \in \Xi} \Big((\ell_{y_i}\circ h)(v_i) - \lambda\|v_i - x_i\|\Big). \tag{28}$$

As a result returning to (23) this can be simplified to,

$$\inf_{\lambda \geqslant 0} \quad \lambda\zeta + \frac{1}{n}\sum_{i=1}^n \sup_{v_i \in \Xi} \Big((\ell_{y_i}\circ h)(v_i) - \lambda\|v - x_i\|\Big), \tag{29}$$

or, in epigraphic form,

$$\begin{aligned} \inf_{\lambda \geqslant 0, \gamma_i \in \mathbb{R}} \quad & \lambda\zeta + \frac{1}{n}\sum_{i=1}^n \gamma_i \\ \text{s.t.} \quad & \sup_{v_i \in \Xi} \Big((\ell_{y_i}\circ h)(v_i) - \lambda\|v_i - x_i\|\Big) \leqslant \gamma_i, \\ & \text{for all } i = 1, \dots n. \end{aligned} \tag{30}$$

From this point forward, we transition between the dual space and the primal space. Specifically, using the definition of the dual norm,

$$\begin{aligned} \|y\|_* = \sup_{x \in \Xi} \quad & \langle y, x\rangle \\ \text{s.t.} \quad & \|x\| \leqslant 1, \end{aligned}$$

for $y \in \Xi^*$, the dual space of Ξ.

With this, the problem in epigraph form can be expressed as follows,

$$\begin{aligned} \inf_{\lambda \geqslant 0, \gamma_i} \quad & \lambda\zeta + \frac{1}{n}\sum_{i=1}^n \gamma_i \\ \text{s.t.} \quad & \sup_{v_i \in \Xi} \Big((\ell_{y_i}\circ h)(v_i) - \max_{\|z_i\|_* \leqslant \lambda} \langle z_i, v_i - x_i\rangle\Big) \leqslant \gamma_i \\ & \text{for all } i = 1, \dots n. \end{aligned} \tag{31}$$

Since $\sup_{v_i} \min_{z_i} \geqslant \min_{z_i} \sup_{v_i}$ by weak duality and (32) is less constrained than (32) (equality holds under strong duality, but we cannot claim it here since we consider $(\ell_{y_i}\circ h)(\cdot)$ to be convex),

$$\begin{aligned} \inf_{\lambda \geqslant 0, \gamma_i} \quad & \lambda\zeta + \frac{1}{n}\sum_{i=1}^n \gamma_i \\ \text{s.t.} \quad & \min_{z_i} \sup_{v_i \in \Xi} \Big((\ell_{y_i}\circ h)(v_i) - \langle z_i, v_i - x_i\rangle\Big) \leqslant \gamma_i \\ & \|z_i\|_* \leqslant \lambda \\ & \text{for all} i = 1, \dots n, \end{aligned} \tag{32}$$

and can be simplified to (by redefining $z_i = -z_i$),

$$\begin{aligned} \inf_{\lambda, \gamma_i, z_i} \quad & \lambda\zeta + \frac{1}{n}\sum_{i=1}^n \gamma_i \\ \text{s.t.} \quad & \sup_{v_i \in \Xi} \Big((\ell_{y_i}\circ h)(v_i) + \langle z_i, v_i\rangle\Big) - \langle z_i, x_i\rangle \leqslant \gamma_i \\ & \|z_i\|_* \leqslant \lambda \\ & \text{for all } i = 1, \dots n. \end{aligned} \tag{33}$$

Dualizing the constraint in (33) with multipliers α_i, β_i

$$\sup_{\alpha_i, \beta_i} \inf_{\lambda, \gamma_i, z_i} \quad \lambda\zeta + \sum_{i=1}^n \Big(\alpha_i\Big(\sup_{v_i \in \Xi} \Big((\ell_{y_i}\circ h)(v_i) + \frac{\gamma_i}{n} + \langle z_i, v_i\rangle\Big) - \langle z_i, x_i\rangle - \gamma_i\Big) + \beta_i(\|z_i\|_* - \lambda)\Big). \tag{34}$$

This is equal to the following when substituting for the optimal λ and γ_i (as those are unconstrained problems),

$$\begin{aligned} \sup_{\beta_i \geqslant 0} \inf_{z_i} \quad & \sum_{i=1}^{n} \sup_{v_i \in \Xi} \Big(\frac{1}{n} (\ell_{y_i} \circ h)(v_i) + \langle z_i, \frac{v_i - x_i}{n} \rangle \Big) \\ & \qquad + \beta_i \|z_i\|_* \\ \text{s.t.} \quad & \sum_{i=1}^{n} \beta_i = \zeta. \end{aligned} \tag{35}$$

By the definition of the dual norm this is equivalent to,

$$\begin{aligned} \sup_{\beta_i \geqslant 0} \inf_{z_i} \quad & \sum_{i=1}^{n} \sup_{v_i \in \Xi} \Big(\frac{1}{n} (\ell_{y_i} \circ h)(v_i) + \langle z_i, \frac{v_i - x_i}{n} \rangle \Big) \\ & \qquad + \max_{\|u_i\| \leqslant \beta_i} \langle z_i, u_i \rangle \\ \text{s.t.} \quad & \sum_{i=1}^{n} \beta_i = \zeta, \end{aligned} \tag{36}$$

which is equivalent to,

$$\begin{aligned} \sup_{\beta_i \geqslant 0} \max_{\|u_i\| \leqslant \beta_i} \inf_{z_i} \quad & \sum_{i=1}^{n} \sup_{v_i \in \Xi} \Big\{ \frac{1}{n} (\ell_{y_i} \circ h)(v_i) + \\ & \qquad \langle z_i, u_i + \frac{v_i - x_i}{n} \rangle \Big\} \\ \text{s.t.} \quad & \sum_{i=1}^{n} \beta_i = \zeta. \end{aligned} \tag{37}$$

Since u_i and β_i are related but the constraint $\|u_i\| \leqslant \beta_i$ and we are maximizing both, we opt to only keep u_i, leading to,

$$\begin{aligned} \sup_{u_i} \inf_{z_i} \quad & \sum_{i=1}^{n} \sup_{v_i \in \Xi} \frac{1}{n} (\ell_{y_i} \circ h)(v_i) + \langle z_i, u_i + \frac{v_i - x_i}{n} \rangle \\ \text{s.t.} \quad & \sum_{i=1}^{n} \|u_i\| \leqslant \zeta. \end{aligned} \tag{38}$$

Each term in the sum depends only on v_i the supremum over all v_i can be distributed across the sum, that is, the sum of the sup is equal to the sup of the sum. Furthermore, since the variable v_i does not affect the ordering of the inf over z_i, the sup over v_i can be commuted with the inf over z_i, allowing us to take the sup jointly over x_i and v_i while keeping the inf over z_i. As a result, (38) is equivalent to,

$$\begin{aligned} \sup_{u_i, v_i \in \Xi} \inf_{z_i} \quad & \sum_{i=1}^{n} \frac{1}{n} (\ell_{y_i} \circ h)(v_i) + \langle z_i, u_i + \frac{v_i - x_i}{n} \rangle \\ \text{s.t.} \quad & \sum_{i=1}^{n} \|u_i\| \leqslant \zeta, \end{aligned} \tag{39}$$

The inner inf is only bounded if $\langle z_i, u_i + \frac{v_i - x_i}{n} \rangle = 0$, in fact, only if $u_i + \frac{v_i - x_i}{n} = 0$. This concludes the proof, resulting in,

$$\begin{aligned} \max_{\{v_i \in \Xi\}_{i=1}^{n}} \quad & \frac{1}{n} \sum_{i=1}^{n} (\ell_{y_i} \circ h)(v_i) \\ \text{s.t.} \quad & \frac{1}{n} \sum_{i=1}^{n} \|x_i - v_i\| \leqslant \zeta. \end{aligned} \tag{40}$$

■

C. Proof of proposition III.2

Proof:
In the previous proof we have shown that the problem is separable and we can optimize for each $\mathbb{Q}_i$ independently. We continued by using Theorem 14.60 from [38] and in doing that show that the problem of finding the measure is equivalent to that of optimizing point-wise, deeming the problem a discrete-discrete transport problem, that is, one that transports a discrete distribution to another discrete distribution.

Given these observations, let us focus on one pair x, v (we will omit the index i in this case for simplicity in notation). If we have that these are non-negative and normalized, that is, $\sum x_i = \sum v_j = 1$, then we can interpret those as empirical distributions,

$$\widehat{\mathbb{P}} = \{\frac{1}{d}\delta_{x_i}\}_{i=1}^{d}, \quad \widehat{\mathbb{Q}} = \{\frac{1}{d}\delta_{v_j}\}_{j=1}^{d}.$$

As stated in [39], any measure $\pi(\widehat{\mathbb{P}}, \widehat{\mathbb{Q}})$ can be represented by a bistochastic $d \times d$ matrix P and in this case, the Kantorovich problem becomes,

$$\begin{aligned} \min_{P} \quad & \langle P, C \rangle_F \\ \text{s.t.} \quad & P \cdot 1 = x \\ & P^T \cdot 1 = v, \end{aligned} \tag{41}$$

for an appropriate cost matrix C.

This is a Linear Programming in the bounded convex set of bistochastic matrices. Borrowing from [39], we know that by Choquet's theorem that this problem admits a solution in the extremal points of the set of bistochastic matrices and by Birkhoff's theorem [39], we know that these extremal points are permutation matrices, $P^{(k)}$. Thus, the Kantorovich problem coincides with Monge's problem. In this case, the optimal transport consists in finding an optimal matching between the points in x and the target points v.

Consider, C such that $C_{ij} = \mathbb{1}_{i \neq j}$, that is,

$$C_{ij} = \begin{cases} 1, i \neq j \\ 0, \text{otherwise.} \end{cases}$$

In tis case, the Kantorovich problem, for given x and v, becomes,

$$\begin{aligned} \min_{P} \quad & \sum_{i} \sum_{j} P_{ij} \mathbb{1}_{i \neq j} \\ \text{s.t.} \quad & P \cdot 1 = x \\ & P^T \cdot 1 = v, \end{aligned} \tag{42}$$

which is equivalent to finding P that solves,

$$\begin{aligned} \max_{P} \quad & \sum_i P_{ii} \\ \text{s.t.} \quad & P \cdot 1 = x \\ & P^T \cdot 1 = v, \end{aligned} \tag{43}$$

since $\sum_i \sum_j P_{ij} \mathbb{1}_{i \neq j} = 1 - \sum_i P_{ii}$, because P is a doubly stochastic matrix.

We wish to maximize the diagonal of the transport plan matrix P, meaning to keep as much mass as possible unmoved. First, note that any admissible solution to this problem must satisfy the constraints, $\sum_j P_{ij} = x_i$ and $\sum_i P_{ij} = v_j$, which naturally lead to,

$$P_{ii} \leqslant x_i, \text{ and } P_{ii} \leqslant v_i, \text{ for all } i,$$

or, combined,

$$P_{ii} \leqslant \min(x_i, v_i),$$

since all values in P are positive.

Hence, the optimal is to choose,

$$P^*_{ii} = \min(x_i, v_i).$$

This solution is feasible (it satisfies the constraints) and given that P_{ii} is positive for any i, it holds that $\max \sum_i P_{ii} = \sum_i \max P_{ii}$. As a result the best solution is to choose the maximum value for each i, which is bounded by the $\min(x_i, v_i)$.

Note that, for any $a, b \geqslant 0$, $\min(a, b) = \frac{a+b-|b-a|}{2}$, which results in,

$$1 - \sum_i P^*_{ii} = 1 - \sum_i \frac{x_i + v_i - |v_i - x_i|}{2} = \frac{1}{2}\|v - x\|_1,$$

because x and v are such that $\sum_i x_i = \sum_i v_i = 1$. Hence,

$$\begin{aligned} \frac{1}{2}\|x - v\|_1 = \min_{P} \quad & \langle P, C \rangle_F \\ \text{s.t.} \quad & P \cdot 1 = x \\ & P^T \cdot 1 = v \end{aligned} \tag{44}$$

and (40) is equivalent to,

$$\begin{aligned} M^* = \max_{\{v_i \in \Xi\}_{i=1}^n} \quad & \frac{1}{n} \sum_{i=1}^n (\ell_{y_i} \circ h)(v_i) \\ \text{s.t.} \quad & \frac{1}{n} \sum_{i=1}^n \min_{P_i} \langle P_i, C \rangle_F \leqslant \frac{\zeta}{2} \\ & \qquad \text{s.t.} \quad P_i \cdot 1 = x_i \\ & \qquad\qquad P_i^T \cdot 1 = v_i. \\ & \text{for all } i = 1, \ldots, n \end{aligned} \tag{45}$$

Consider now the problem

$$\begin{aligned} K^* = \max_{\substack{\{v_i \in \Xi\}_{i=1}^n, \\ \{P_i \in [0,1]^{d \times d}\}_{i=1}^n}} \quad & \frac{1}{n} \sum_{i=1}^n (\ell_{y_i} \circ h)(v_i) \\ \text{s.t.} \quad & \frac{1}{n} \sum_{i=1}^n \langle P_i, C \rangle_F \leqslant \frac{\zeta}{2} \\ & P_i \cdot 1 = x_i \\ & P_i^T \cdot 1 = v_i. \\ & \text{for all } i = 1, \ldots, n \end{aligned} \tag{46}$$

we have that,

$$K^* = M^*,$$

since in (45) we seek the existence of P_i satisfying the constraint in (45), thus, we can equivalently maximize for it as in (46).

■

D. Proof of proposition III.3

Proof:

The proof is divided into two parts, one which shows that the optimization problem can be solved using a Sinkhorn-like algorithm and another which studies the convergence properties of the iterative procedure.

Algorithm

Let us start by deriving the dual formulation of (11). First, we look at the Lagrangian,

$$\begin{aligned} \mathcal{L}(v_i, P_i, \alpha_i, \beta_i, \gamma) = & \lambda \sum_{i=1}^n -\nabla^T_{(\ell_{y_i} \circ h)}(v_i^{(k)}) v_i + \\ & \sum_{i=1}^n \sum_{j=1}^d \sum_{k=1}^d P_{i_{jk}} \log P_{i_{jk}} + \\ & \gamma \Big(\sum_{i=1}^n \sum_{j=1}^d \sum_{k=1}^d P_{i_{jk}} C_{jk} - n\zeta \Big) + \\ & \sum_{i=1}^n \alpha_i^T (P_i \cdot 1 - x_i) + \\ & \sum_{i=1}^n \beta_i^T (P_i^T \cdot 1 - v_i), \end{aligned} \tag{47}$$

and continue by solving for v_i and P_i, in a similar fashion to that of [23], [45],

$$\begin{aligned} \frac{\partial}{\partial P_{i_{jk}}} \mathcal{L}(v_i, P_i, \alpha_i, \beta_i, \gamma) = 0, \\ \frac{\partial}{\partial v_i} \mathcal{L}(v_i, P_i, \alpha_i, \beta_i, \gamma) = 0, \end{aligned} \tag{48}$$

which leads to,

$$\begin{aligned} P_{i_{jk}} &= e^{-\alpha_{i_j} - \beta_{i_k} - \gamma C_{jk} - 1}, \\ \beta_i &= -\lambda \nabla_{(\ell_{y_i} \circ h)}(v_i^{(k)}). \end{aligned} \tag{49}$$

Now, substituting this on the Lagrangian, results in,

$$\mathcal{L}(\alpha_i, \gamma) = -\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{k=1}^{d} P_{i_{jk}} - \gamma n\zeta - \sum_{i=1}^{n} \alpha_i^T \cdot x_i. \tag{50}$$

Note that β_i is already known in (49), we just keep it on the RHS for readability purposes. The dual formulation of (11) is,

$$\max_{\substack{\{\alpha_i \in \mathbb{R}^d\}_{i=1}^n,\\ \gamma \in \mathbb{R}_+}} \quad -\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{k=1}^{d} P_{i_{jk}} - \gamma n\zeta - \sum_{i=1}^{n} \alpha_i^T \cdot x_i. \tag{51}$$

Since this is an unconstrained, convex problem, we can proceed by solving it directly, taking the partial derivatives w.r.t. α_i and γ and solving for equality to 0, resulting in,

$$\begin{aligned} \alpha_{i_j}^* &= \Big(\log\big(\sum_{k=1}^{d} e^{-\beta_{i_k} - \gamma^* C_{jk} - 1}\big) - \log x_{i_j}\Big) \\ \beta_i &= -\lambda \nabla_{(\ell_{y_i} \circ h)}(v_i^{(k)}), \end{aligned} \tag{52}$$

and γ^* cannot be solved analytically but it is solvable using Newton's second order procedure,

$$\begin{aligned} \gamma &= \gamma - \frac{\frac{\partial}{\partial\gamma}\mathcal{L}}{\frac{\partial^2}{\partial\gamma^2}\mathcal{L}} \\ \frac{\partial}{\partial\gamma}\mathcal{L} &= \sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{k=1}^{d} C_{jk} P_{i_{jk}} - n\zeta \\ \frac{\partial^2}{\partial\gamma^2}\mathcal{L} &= -\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{k=1}^{d} C_{jk}^2 P_{i_{jk}}. \end{aligned} \tag{53}$$

Finally, the optimal solution for the primal problem is,

$$\begin{aligned} P_{i_{jk}}^* &= e^{-\alpha_{i_j}^* - \beta_{i_k} - \gamma^* C_{jk} - 1}, \\ v_i^* &= {P_i^*}^T \cdot 1. \end{aligned} \tag{54}$$

Convergence

We continue by simplifying (11) and (15). Note that we can remove $\{v_i \in [0,1]^d\}_{i=1}^n$ but still recover $v_i = P_i^T \cdot 1$ and get lower dimension problems,

$$\begin{aligned} \min_{\{P_i \in [0,1]^{d \times d}\}_{i=1}^n} \quad & \frac{1}{n}\sum_{i=1}^{n} -(\ell_{y_i} \circ h)(P_i^T \cdot 1) + \\ & \frac{\lambda}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{k=1}^{d} P_{i_{jk}} \log P_{i_{jk}} \\ \text{s.t.} \quad & \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{k=1}^{d} P_{i_{jk}} C_{jk} \leqslant \zeta \\ & P_i \cdot 1 = x_i, \\ & i = 1, \ldots, n, \end{aligned} \tag{55}$$

and

$$\begin{aligned} \min_{\{P_i \in [0,1]^{d \times d}\}_{i=1}^n} \quad & \frac{1}{n}\sum_{i=1}^{n} -\nabla_{(\ell_{y_i} \circ h)}^T(v_i^{(k)}) \cdot P_i^T \cdot 1 + \\ & \frac{\lambda}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{k=1}^{d} P_{i_{jk}} \log P_{i_{jk}} \\ \text{s.t.} \quad & \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{k=1}^{d} P_{i_{jk}} C_{jk} \leqslant \zeta \\ & P_i \cdot 1 = x_i, \\ & i = 1, \ldots, n. \end{aligned} \tag{56}$$

But (55) is a DC optimization problem such as the one in (12),

$$\begin{aligned} \min_{x \in C} \quad & u(x) - v(x) \\ \text{s.t.} \quad & f(x) \leqslant 0 \\ & g_i(x) = 0 \\ & i = 1, \ldots, n, \end{aligned} \tag{57}$$

in which u and v are convex functions and $f, g_1, \ldots, g_n$ are linear functions.

Consider now the objective function,

$$\phi(x) = u(x) - v(x). \tag{58}$$

Since v is convex,

$$v(x) \geqslant v(y) + (x - y)^T \nabla v(y),$$

then

$$\phi(x) \leqslant u(x) - v(y) - (x - y)^T \nabla v(y) := \psi(x, y).$$

As a result,

$$\phi(x^{(k+1)}) \leqslant \psi(x^{(k+1)}, x^{(k)}).$$

Furthermore, if $x^{(k+1)} \in \underset{x \in C}{\text{argmin}}\ \psi(x, x^{(k)})$,

$$\psi(x^{(k+1)}, x^{(k)}) \leqslant \psi(x^{(k)}, x^{(k)}).$$

This tells us that,

$$\phi(x^{(k+1)}) \leqslant \psi(x^{(k+1)}, x^{(k)}) \leqslant \psi(x^{(k)}, x^{(k)}) = \phi(x^{(k)}). \tag{59}$$

Let is call our algorithm $\mathcal{A}$. Note that this algorithm is a point-to-set, that is, an algorithm that maps a initial guess, $x^{(0)}$, into a sequence $\{x^{(k)}\}_{k=0}^{\infty}$ through the iteration,

$$\begin{aligned} x^{(k+1)} \in \underset{x \in C}{\text{argmin}} \quad & u(x) - v(x^{(k)}) - (x - x^{(k)})^T \nabla v(x^{(k)}) \\ \text{s.t.} \quad & f(x) \leqslant 0 \\ & g_i(x) = 0 \\ & i = 1, \ldots, n. \end{aligned} \tag{60}$$

Given the point to set map, we call a fixed point, x^*, of the map, $\mathcal{A}$, the point such that $x^* = \mathcal{A}(x^*)$. That is, it is a point in which if we start at the point the algorithm stays at the same point.

Given all these functions and observations we can use Zangwill's convergence theorem [47] to show our main result following a similar approach to [46].

Theorem. (Convergence Theorem [47]) *Let $\mathcal{A}$ be a point-to-set map that given a point $x^{(0)} \in X$ generates a sequence $\{x^{(k)}\}_{k=0}^{\infty}$ through the iteration $x^{(k+1)} \in \mathcal{A}(x^{(k)})$. Also let a solution set $\Gamma \in X$ be given. Suppose*

1) *All points $x^{(k)}$ are in a compact set $S \subset X$.*
2) *There is a continuous function $\phi : X \to \mathbb{R}$ such that:*
   a) $x \notin \Gamma \implies \phi(y) < \phi(x), \forall y \in \mathcal{A}(x),$
   b) $x \in \Gamma \implies \phi(y) \leqslant \phi(x), \forall y \in \mathcal{A}(x).$
3) *$\mathcal{A}$ is closed at x if $x \notin \Gamma$.*

Then the limit of any convergent subsequence of $\{x^{(k)}\}_{k=0}^{\infty}$ is in Γ. Furthermore, $\lim\limits_{k\to\infty} \phi(x^{(k)}) = \phi(x^*)$ for all limit points x^*.

In our problem, we satisfy all requirements:

Assumption 1

First, we show that assumption 1 in the above theorem holds. From proof C we know that $\{P_i \in [0,1]^{d\times d}\}_{i=1}^{n}$ is an element of the bounded convex set of bistochastic matrices, which implies that all points $\{P_i^{(k)}\}_{i=1}^{n}$ are in a compact space.

Assumption 2

Now, take Γ be the set of all fixed points, x^* of (60), i.e., all points x^* such that $x^* = \mathcal{A}(x^*)$ and ϕ as in (58), then 2b in the convergence theorem follows with equality by the definition of Γ and 2a follows by the defintion of Γ and the descent inequality (59).

Assumption 3

Closeness follows directly from Lemma 6 in [46] which has first been proposed in [50]

Convergence

With these we apply the convergence theorem and get that any convergent subsequence $\{x^{(k)}\}_{k=0}^{\infty}$ produced by $\mathcal{A}$ is in Γ and that $\lim\limits_{k\to\infty} \phi(x^{(k)}) = \phi(x^*)$ for all limit points x^*.

We would like however to relate the solution of the algorithm to the original problem (57) and this can be checked by verifying the KKT conditions. Let x^* be a fixed, limiting point in (60), then we know $\exists \alpha^*, \beta_1^*, \dots \beta_n^*$, Lagrange multipliers, that satisfy the KKT sufficient conditions,

$$\begin{aligned}\nabla u(x^*) - \nabla v(x^*) + \alpha^* \nabla f(x^*) + \sum_{i=1}^{n} {\beta_i^*}^T \cdot \nabla g_i(x^*) = 0 \\ f(x^*) \leqslant 0 \\ g_i(x^*) = 0 \\ \alpha^* \geqslant 0.\end{aligned}$$

These are also the KKT conditions for the original problem (57). Hence, the limit points x^* are also stationary points of the original problem. ■