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.2026.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 Adversarial attacks are often modeled as pointwise perturbations of individual samples, which can miss structured distributional effects and may waste perturbation budget on examples that are already misclassified. We study a data-driven Wasserstein attack model in which the adversary shifts the empirical distribution under a label-preserving transport budget. Starting from this formulation, we derive a finite-dimensional transport surrogate and an equivalent lifting that makes the role of transport couplings explicit. We then introduce an entropic regularization, obtaining a difference-of-convex formulation that penalizes attacks which mainly amplify the loss of already misclassified samples. This leads to *Partial Sinkhorn*, an iterative algorithm that combines convex-concave linearization with Sinkhorn-type updates, such that any limit point of a convergent subsequence is a KKT stationary point of the penalized problem. Experiments on synthetic and MNIST tasks show that the proposed method generates stronger attacks than FGSM under comparable perturbation budgets, particularly in the low-distortion regime. The framework also highlights links between adversarial attack, optimal transport, and distributionally robust control.

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

I. Introduction

Neural networks are vulnerable to adversarial perturbations, namely small but carefully designed input modifications that can induce misclassification [1]–[3]. Most attack models are pointwise, perturbing each sample independently under a norm budget [4]. While effective, this viewpoint may fail to capture structured shifts that act at the level of the data distribution.

In this work, we study adversarial attacks from a distributional perspective by modeling the adversary as shifting the empirical distribution under a label-preserving Wasserstein transport budget [5]–[11]. This viewpoint naturally connects adversarial attack generation with optimal transport and distributionally robust optimization [12]–[16]. At the same time, it leads to a nonconvex optimization problem whose direct solution is computationally challenging, and whose surrogate formulations may encourage undesirable behavior, such as spending perturbation budget on samples that are already misclassified rather than inducing new misclassifications.

To address these issues, we derive a finite-dimensional transport-based formulation of the adversary problem and introduce an entropic regularization that yields a difference-of-convex reformulation. Based on this structure, we propose the *Partial Sinkhorn* algorithm, which combines convex-concave linearization with Sinkhorn-type updates to compute KKT stationary solutions of the penalized problem. Numerical experiments on synthetic and real datasets show that the method generates stronger attacks than FGSM under comparable perturbation budgets, particularly in the low-distortion regime. Beyond adversarial learning, the framework also highlights connections between optimal transport, distributional robustness, and control-oriented optimization [17]–[21].

A. Related work

Research on adversarial robustness has increasingly emphasized the connection between adversarial training (AT) and distributionally robust optimization (DRO). In particular, AT can be interpreted as a specific instance of DRO [5], while Wasserstein-DRO formulations have been used to train neural networks against worst-case perturbations with provable robustness guarantees [6]. Related work also considers Wasserstein threat models tailored to natural image transformations, such as scaling and rotation, showing that

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

Wasserstein adversarial examples can degrade model accuracy and that AT can partially recover robustness [7].

Beyond pointwise perturbations, Adversarial Distributional Training (ADT) optimizes over adversarial distributions around natural samples through a min–max formulation with entropic regularization [8]. In this sense, ADT lies between classical AT, which searches for worst-case perturbations of individual samples, and DRO, which optimizes against worst-case distributions in a Wasserstein ball around the empirical law. Closely related ideas also appear in work connecting adversarial robustness, DRO, and optimal control, where regularized learning in deep neural networks is formulated through a calculus-of-variations perspective and interpreted as a discretized control problem [9]. A further related line studies Wasserstein-based distributionally robust classification models, including formulations based on conditional value-at-risk and links with maximum-margin classifiers and adversarial models [10]. Wasserstein-DRO has also been used to define attack models allowing non-uniform perturbations across inputs, together with first-order attack algorithms and asymptotic guarantees on adversarial accuracy [11].

More broadly, Wasserstein ambiguity sets are central in data-driven stochastic programming, where they yield tractable convex or linear reformulations and finite-sample guarantees under suitable convexity assumptions [14]. Recent advances include Sinkhorn-DRO, which studies entropically regularized Wasserstein ambiguity sets, strong duality, and stochastic mirror-descent methods for convex settings [22]. Our formulation is also closely related to distributionally robust control and model predictive control (MPC), where one optimizes against worst-case disturbance distributions within Wasserstein ambiguity sets. Representative examples include distributionally robust constrained control [17], data-driven Wasserstein distributionally robust stochastic control [18], partially observable extensions [19], Wasserstein/optimal-transport DRO MPC [20], and data-driven risk-constrained DRO MPC [21]. These connections motivate our entropically regularized Difference-of-Convex (DC) approach as a tool for nonconvex inner worst-case distribution problems arising in both adversarial learning and distributionally robust control.

B. Contribution

We study the problem of finding worst-case inference-time perturbations that induce *new* misclassifications under a transportation budget. We formulate this as a data-driven Wasserstein DRO problem and derive an entropically regularized DC formulation. The main contributions of this paper are as follows:

- *Transport-based DC reformulation:* We show that the resulting OT/DRO problem is a linearly constrained nonconvex program, and introduce an entropic regularization that yields a DC decomposition while discouraging over-optimization of already misclassified samples.
- *Partial Sinkhorn algorithm:* Exploiting this structure, we propose the *Partial Sinkhorn* algorithm and establish convergence of convergent subsequences to KKT stationary points of the penalized problem.
- *Numerical validation:* We demonstrate on synthetic and real datasets that the proposed method is computationally efficient and generates stronger attacks than standard baselines under comparable perturbation budgets.

The remainder of the paper is organized as follows. Section II introduces the problem setup. Section III presents the main results: the data-driven DRO formulation and the Partial Sinkhorn algorithm. Section IV reports numerical experiments.

II. Problem Formulation

We consider a supervised classification problem with features X and labels Y. The features take values in a space $\Xi \subset \mathbb{R}^d$, equipped with the Euclidean norm $\|\cdot\|$ and the associated Borel σ-algebra $\mathcal{B}(\Xi)$. We assume that Ξ is a compact subset of $\mathbb{R}^d$, (for the purposes of this section, it would in fact suffice to assume that Ξ is closed). Under this assumption, $(\Xi, \|\cdot\|)$ is complete and separable [23, Sec. 4.1], and hence a Polish space. The labels take values in a space Υ, equipped with the power-set σ-algebra 2^{Υ}. In particular, since Ξ is equipped with $\mathcal{B}(\Xi)$ and Υ with 2^{Υ}, the product space $\Xi \times \Upsilon$ is equipped with the product σ-algebra $\mathcal{B}(\Xi) \otimes 2^{\Upsilon}$.

To ensure that the feature–label pair (X, Y), and its adversarially perturbed counterpart (V, Y) are well defined as a collection of random variables, we work on an underlying probability space and impose the appropriate measurability conditions. Specifically, let $(\Omega, \mathcal{F}, \mu)$ be a probability space and let

$$X : (\Omega, \mathcal{F}) \to (\Xi, \mathcal{B}(\Xi)),$$
$$Y : (\Omega, \mathcal{F}) \to (\Upsilon, 2^{\Upsilon})$$

be random variables. We denote by $\mathbb{P} \in \mathcal{P}(\Xi \times \Upsilon)$ the joint distribution of (X, Y), by $\mathbb{Q} \in \mathcal{P}(\Xi \times \Upsilon)$ the joint distribution of (V, Y), and $\mathbb{P}_X$ the marginal distribution of X, with $\mathcal{P}(\Xi \times \Upsilon)$ being the space of probability measures on $(\Xi \times \Upsilon, \mathcal{B}(\Xi) \otimes 2^{\Upsilon})$.

Within this context, let $\mathcal{G}$ denote a class of maps $g : \Xi \to \Xi$. We model an adversarial attack by selecting a map $g \in \mathcal{G}$. Throughout, we assume that every $g \in \mathcal{G}$ is $\mathcal{B}(\Xi)$-measurable, so that the attacked feature

$$V := g(X)$$

is a well-defined random variable. We then have that

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

The learner and adversary are playing a game in which the learner seeks a hypothesis, where each $h \in \mathcal{H} : \Xi \to \Upsilon$, that minimizes the risk, $\mathcal{R}_{\mathbb{P}}\big[\ell(h(X), Y)\big]$. Conversely, the adversary seeks an adversarial map g that maximizes the risk $\mathcal{R}_{\mathbb{P}}\big[\ell(h(g(X)), Y)\big]$, subject to a perturbation budget ζ.

In this work, we focus on the adversary's problem. Nevertheless, it is useful to briefly review the learner's problem, as it provides the necessary context.

This article has been accepted for publication in IEEE Open Journal of Control Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCSYS.2026.3679905

A. The learner's problem

In the learning problem, the learner seeks the best hypothesis in a hypothesis class, $h \in H$. In a classification context, such a hypothesis can simply be called a 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,

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

As the learner does not know the distribution $\mathbb{P}$ but has access to samples $S = \{(x_i, y_i)\}_{i=1}^n$, the problem is often solved empirically through the empirical risk minimization (ERM) framework, which yields,

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

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

B. The adversary's problem

In the standard evasion model [1]–[3], [24]–[30], the adversary perturbs each realized input x within a budget ζ and seeks to maximize the expected loss:

$$\mathbb{E}_{\mathbb{P}}\big[\sup_{v \,\in\, \Xi \,:\, \|v - X\| \leqslant \zeta} \ell(h(v), Y)\big].$$

Conveniently, we parameterize such perturbations by an *attack map* $g : \Xi \to \Xi$ satisfying $\|g(x) - x\| \leqslant \zeta$ for all x in the support of $\mathbb{P}_X$ and set $V = g(X)$. With this parameterization, the adversary's problem can be written as

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

This constraint enforces that, for inputs drawn from the data distribution, the adversary's perturbation magnitude does not exceed ζ almost surely. Here, ζ can be interpreted as the adversary's power, and $h \in \mathcal{H}$ denotes the learner's hypothesis. We consider a *white-box* adversary with access to h (and hence to $\ell \circ h$) when solving the maximization.

Examples: Consider binary classification with a score function $h(x) = \mathrm{sign}(w^\top x + b)$. A standard choice of loss is the logistic loss $\ell(h(x), y) = \log(1 + \exp(-y\, h(x)))$ (or equivalently the cross-entropy loss). For the attacker, a representative perturbation model is an additive one, $g(x, \xi) = x + \xi$, where ξ belongs to a constraint set Ξ (e.g., $\|\xi\|_\infty \leqslant \zeta$ or $\|\xi\|_2 \leqslant \zeta$).

Similarly to the learner, the adversary also does not know the distribution $\mathbb{P}$, but has access to sample points $S = \{(x_i, y_i)\}_{i=1}^n$. 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 [31], whose goal was to find a measurable transport map $g : \Xi \to \Xi$ that pushes a source probability measure on Ξ onto a target probability measure on Ξ while minimizing a transportation cost. Later, Kantorovich [32] proposed a relaxed formulation in terms of transportation plans, which we recall next.

Recall that $(\Xi, \|\cdot\|)$ is a Polish space, since Ξ is closed in $\mathbb{R}^d$, endowed with its Borel σ-algebra $\mathcal{B}(\Xi)$. In this case, the Wasserstein distance and the associated moment classes are well-defined in the standard way.

Wasserstein distance on Ξ

To avoid ambiguity with the joint distributions on $\Xi \times \Upsilon$ introduced in the beginning of this section, we first define the Wasserstein distance for generic measures on the feature space Ξ. Let $\mathcal{P}_p(\Xi)$ denote the set of Borel probability measures on Ξ with finite p-th moment,

$$\int_{\Xi} \|x\|^p \, \mathrm{d}\mu(x) < \infty.$$

For $\mu, \nu \in \mathcal{P}_p(\Xi)$ and $p \geqslant 1$, the p-Wasserstein distance is

$$\Big(\inf_{\gamma \,\in\, \Pi(\mu, \nu)} \int_{\Xi \times \Xi} \|x - v\|^p \, \mathrm{d}\gamma(x, v)\Big)^{\frac{1}{p}}. \tag{5}$$

For $p = 1$, W_1 is also known as the Kantorovich–Rubinstein metric [33], [34] (or earth mover distance [33]).

Label-preserving Wasserstein distance

To model inference-time (evasion) attacks that perturb features but preserve labels, we compare the joint distributions, $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\Xi \times \Upsilon)$ of (X, Y) and (V, Y) using the label-preserving Wasserstein distance defined [6] by the cost

$$\tilde{c}\big((x, y), (v, y')\big) := \|x - v\|^p + \infty \, \mathbb{1}_{\{y \neq y'\}},$$

and metric,

$$W_p(\mathbb{P}, \mathbb{Q}) := \Big(\inf_{\gamma \in \Pi(\mathbb{P}, \mathbb{Q})} \int \tilde{c} \, \mathrm{d}\gamma\Big)^{1/p}.$$

This enforces $y = y'$ γ-a.s. (and thus $W_p(\mathbb{P}, \mathbb{Q}) < \infty$ only if $\mathbb{P}_Y = \mathbb{Q}_Y$). Under this label-preserving restriction we use, throughout the paper, the equivalent shorthand

$$W_p(\mathbb{P}, \mathbb{Q}) = \left(\inf_{\pi} \int_{\Xi \times \Upsilon \times \Xi} \|x - v\|^p \, \mathrm{d}\pi(x, y, v)\right)^{1/p},$$

where π ranges over measures on $\Xi \times \Upsilon \times \Xi$ with marginals $\pi_{(x,y)} = \mathbb{P}$ and $\pi_{(v,y)} = \mathbb{Q}$.

From feature-only attack budgets to a joint Wasserstein ambiguity set

Recall the map-based adversary model (4), in which the adversary selects a measurable attack map $g : \Xi \to \Xi$ (with labels fixed) and produces the perturbed input $V = g(X)$. Each such g induces an attacked joint distribution $\mathbb{Q}$ on $\Xi \times \Upsilon$. Under the label-preserving coupling between $\mathbb{P}$ and $\mathbb{Q}$, we have

$$W_p(\mathbb{P}, \mathbb{Q})^p \leqslant \mathbb{E}_{\mathbb{P}}\big[\|g(X) - X\|^p\big] \leqslant \zeta^p,$$

and hence $\mathbb{Q}$ lies in the label-preserving Wasserstein ball of radius ζ centered at $\mathbb{P}$.

Motivated by this observation, we relax the map-based formulation by optimizing directly over joint distributions (equivalently, label-preserving couplings) rather than over perturbation maps g. Specializing to $p = 1$, we arrive at the following distributional adversary model:

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

That is, the adversary seeks a worst-case *joint* distribution within a label-preserving transportation budget ζ from the data distribution.

Empirical (data-driven) ambiguity set

As previously highlighted, neither the learner nor the adversary has access to $\mathbb{P}$; instead they observe i.i.d. samples $S = \{(x_i, y_i)\}_{i=1}^n$ and form the empirical joint distribution $\widehat{\mathbb{P}}_n$. We therefore replace $\mathbb{P}$ by $\widehat{\mathbb{P}}_n$ in the ambiguity set,

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

Without any further assumptions on the hypothesis classes and loss functions, this formulation generally leads to a non-convex problem, since we allow broad hypothesis classes and losses (e.g., $\ell(h(\cdot), \cdot)$ induced by nonlinear models), even when ℓ is convex in its first argument.

Data-driven Wasserstein DRO estimators admit statistical performance guarantees under mild regularity assumptions. In particular, the robust certificate and the corresponding data-driven solution are known to be asymptotically consistent as $n \to \infty$ [14].

Connection to Wasserstein DRO

Data-driven DRO with Wasserstein ambiguity sets has been studied extensively, including finite-sample performance guarantees under convexity assumptions [14]. Our setting departs from the classical convex DRO regime because the adversary's objective induced by modern hypothesis classes is typically non-convex. In the sequel, we develop tractable relaxations leading to a difference-of-convex (DC) program with linear constraints, and derive an iterative algorithm tailored to the resulting label-preserving, entropically-regularized transport structure.

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 adversary's problem

For a fixed classifier $h \in \mathcal{H}$, the adversary solves the inner maximization over an ambiguity set around the (empirical) data distribution. Accordingly, (8) is the *adversary's problem* (the inner "sup" in a min–max/DRO formulation), rather than a standalone distributionally robust optimization problem, since the outer "inf" over h is not part of (8),

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

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,

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

To state the relationship between (8) and the finite-dimensional surrogate (9) precisely, we introduce the standing regularity assumption used below. For any fixed $h \in \mathcal{H}$ and, for each label $y \in \Upsilon$, define $\phi_y(\cdot)$ such that,

$$\varphi_y(v) := \ell(h(v), y), \qquad \text{for all } v \in \Xi.$$

Standing Assumption III.1. *For each $y \in \Upsilon$, φ_y is bounded and upper semicontinuous on Ξ.*

This assumption together with the compactness of Ξ, ensures the regularity needed for the min–max interchange. In particular, each φ_y is Borel measurable and integrable under any probability measure on Ξ.

When gradients are used (e.g., in Algorithm 1), we additionally assume that φ_y is differentiable in v and we use the standard notation $\nabla_v \varphi_y(v)$.

Proposition III.1. *The distributional adversary problem (8) and its finite-dimensional surrogate (9) satisfy*

$$L^\star \leqslant D^\star.$$

Remark III.1 (Proof outline)**.** *The proof follows arguments analogous to those in [14] together with results in [35]. The full proof is given in Appendix A.*

As shown in Section IV, directly maximizing the surrogate objective in (9) may yield undesirable attacks. The optimization tends to increase the misclassification error of samples that are already misclassified, increasing their loss rather than using the perturbation budget to induce *new* misclassifications. This phenomenon is reflected in the gradient profile of the binary cross-entropy loss (Section IV), which favors pushing misclassified samples further from the decision boundary over moving correctly classified samples across it.

For this reason, we introduce the transport-based formulation below, which allows us to incorporate an additional penalization term and better steer the attack toward enlarging

This article has been accepted for publication in IEEE Open Journal of Control Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCSYS.2026.3679905

the set of misclassified samples:

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

where P_i denotes a transport plan mapping the sample x_i to the perturbed point v_i, $\langle A, B\rangle_F = \sum_i \sum_j A_{ij}B_{ij}$ denotes the Frobenius inner product, and C is a cost matrix consistent with the Wasserstein metric.

This formulation follows from the Kantorovich characterization of Wasserstein-ball constraints, which represents ball membership through optimal-transport couplings between the empirical joint law and an attacked joint law. In our setting, (10) is equivalent to the surrogate problem (9), as stated in the following proposition.

Proposition III.2. *Let x and v be nonnegative and normalized, with $\sum_i x_i = \sum_j v_j = 1$, and let C be a cost matrix satisfying $C_{ij} = 2\mathbb{1}_{\{i\neq j\}}$. Then problem (10) is equivalent to problem (9), that is,*

$$K^\star = L^\star.$$

Remark III.2 (Proof outline)**.** *The proof reduces the transport constraint to a discrete Kantorovich problem and relates it to a total-variation formulation. The full proof is given in Appendix B.*

We emphasize that (10) is a transport-based reformulation of the finite-dimensional surrogate problem (9), not of the original distributional adversary problem (8). The role of Proposition III.1 is to relate (8) and (9), whereas Proposition III.2 establishes the equivalence between (9) and (10). Moreover, the coupling underlying (8) is distribution-level, whereas the matrices P_i in (10) are samplewise transport variables used only in the reformulation of the surrogate problem (9).

While formulation (10) is exact, directly optimizing over transport couplings is computationally demanding. We therefore adopt an entropic regularization of the transport plan, which smooths the objective and enables efficient iterative scaling updates. A key advantage of formulation (10) is that it naturally accommodates an entropic penalization term. This term mitigates the algorithm's tendency to keep perturbing samples that are already misclassified merely to further increase their loss, and instead encourages perturbations that induce new misclassifications.

In particular, letting $K^\star$ denote the optimal value of the exact inner transport problem (10) we define $D_\lambda^\star$ as the optimal value of its entropically regularized counterpart. This yields the following entropically penalized inner problem,

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

Remark III.3. *Combining Proposition III.2 with the definitions of $K^\star$ (exact inner problem) and $D_\lambda^\star$ (entropically regularized surrogate), we have that,*

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

An adversary who solves the penalized problem (11) instead of the original one (8) may yield a less effective attack, but with greater computational efficiency. Here, by *effectiveness* we mean attack success (e.g., achieved misclassification rate or adversarial loss) under the same perturbation budget. The entropic regularization smooths the inner problem and enables efficient Sinkhorn-type scaling iterations, but can bias the solution away from the sharp optimum of the unregularized formulation (recovering it as $\lambda \downarrow 0$). As we shall see in the numerical experiments section, despite this potential loss in effectiveness, the adversary still shows enhanced adversarial power compared to FGSM[1] [3], achieving higher misclassification with less need to tamper with already misclassified samples.

B. Partial Sinkhorn Optimization Algorithm

For the remainder of this section, we additionally assume:

Assumption III.2. *For each $y \in \Upsilon$, the map $v \mapsto \varphi_y(v) = \ell(h(v), y)$ is convex on Ξ.*

This assumption restricts the analysis below to settings in which the composition between ℓ and h gives rise to a function that is convex in the adversarial variable, such as linear or affine models with convex loss. Notice that in the adversary's problem (11), the objective function would thus be concave (due to the negation), but since this is a

[1]FGSM is a 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,

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

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,

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

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

This article has been accepted for publication in IEEE Open Journal of Control Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCSYS.2026.3679905

minimization problem, it will be a nonconvex optimization program.

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 KKT stationary point, as stated in Proposition III.3.

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

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

where each f_i and g_i is convex on C, with C being the feasible set

$$C := \Big\{ (v_{1:n}, P_{1:n}) : v_i \in \Xi,\ P_i \in \mathbb{R}_+^{d \times d},\ i = 1, \dots, n \Big\}.$$

Note that if Ξ is convex (e.g., $\Xi = [0,1]^d$), then C is convex.

We continue by taking $x := [v_1, \dots, v_n, P_1, \dots, P_n]$, defining $g_i(x) \equiv 0$ for all constraint indices, and f_0, g_0, f_i as follows.

$$\begin{aligned} f_0(x) &= \frac{\lambda}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \sum_{k=1}^{d} P_{i_{jk}} \log P_{i_{jk}}, \\ g_0(x) &= \frac{1}{n} \sum_{i=1}^{n} \ell(h(v_i), y_i), \\ f_i(x) &= P_i \cdot \mathbf{1} - x_i, \forall i = 1, \dots n, \\ f_{n+i}(x) &= -P_i \cdot \mathbf{1} + x_i, \forall i = 1, \dots n, \\ f_{2n+i}(x) &= P_i^\top \cdot \mathbf{1} - v_i, \forall i = 1, \dots n, \\ f_{3n+i}(x) &= -P_i^\top \cdot \mathbf{1} + v_i, \forall i = 1, \dots n, \\ f_{4n+1}(x) &= \frac{1}{n} \sum_{i=1}^{n} \langle P_i, C \rangle_F - \zeta. \end{aligned}$$

This kind of problem has been addressed in the literature and solved using DC algorithms (DCA) [36], [37]. Alternatively, one can use the Convex Concave Procedure (CCP) which finds a local optimum of (12). CCP was introduced in [38] and we adopt the standard convexification/linearization template and implementation details as presented in [39]. To apply the CCP linearization, we use the differentiability assumption already stated, so that $\nabla_x g_0(x)$ (and, in particular, $\nabla_{v_i} \varphi_{y_i}(v_i)$) is well defined on C. Starting from an initial value x_0, the CCP algorithm iteratively solves,

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

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,

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

In the theoretical analysis, we assume that $\Xi \subset \mathbb{R}^d$ is compact. In the numerical implementation, we further restrict to $\Xi = [0,1]^d$ (after feature normalization), which allows the updates to be written as projected steps.

Solving (14) is equivalent to solving the following,

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

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

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

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

This article has been accepted for publication in IEEE Open Journal of Control Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCSYS.2026.3679905

core goal. We illustrate and discuss this unintended effect in detail using low-dimensional synthetic datasets in Section IV.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 [40], [41] and its projected variant [7], 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 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 first-order KKT 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)*, satisfies the first-order KKT conditions of problem* (11)*.*

Remark III.4 (Proof outline)**.** *The algorithm is obtained by solving the dual problem, with the updates of* α*,* γ*, and* β *enforcing the primal constraints, the transport-budget constraint, and the nonnegativity conditions, respectively. The convergence analysis uses Zangwill's global convergence theorem together with standard CCP arguments. The full proof is given in Appendix C.*

Remark III.5 (Stopping criteria/step-size)**.** *The inner loop terminates when* $|\gamma^{(t+1)} - \gamma^{(t)}|$ *falls below a prescribed tolerance or a maximum number of iterations is reached, in line with Sinkhorn-type algorithms [7]. The outer loop terminates when the improvement in the objective falls below a threshold or a maximum number of iterations is reached, following the standard CCP criterion [39].*

The update of γ *uses the safeguarded Newton step of [7], with backtracking to ensure nonnegativity. If needed, the parameter* η *is progressively decreased until a nonnegative* γ *is obtained, as implemented in steps 19–21 of Algorithm 1.*

Remark III.6 (Interpretation)**.** *The proposed algorithm is closely related to the Projected Sinkhorn method [7]. In particular, the updates operate on the Gibbs kernel induced by the cost matrix* C*: the transport variables are updated in Algorithm 1 (line 14) in exponential form*

$$P_{ijk} \propto \exp\big(-\alpha_{ij} - \beta_{ik} - \gamma\, C_{jk}\big),$$

Algorithm 1 Partial Sinkhorn CCP procedure

1: **function** CCP(ℓ, h, C, λ, $v_1^{(0)}, \dots, v_n^{(0)}$)
2: $l \leftarrow 0$
3: $\tau \leftarrow 1/\lambda$
4: **repeat**
5: $\beta_i \leftarrow -\tau \nabla_v \ell(h((v_i^{(l)}), y_i))$
6: $\alpha_i \leftarrow \log \frac{1}{n}$
7: $\gamma \leftarrow 1$
8: **repeat**
9: **for** $i = 1, \dots, n$ **do**
10: **for** $j = 1, \dots, d$ **do**
11: $\alpha_{i_j} \leftarrow \log\big(\sum_{k=1}^{d} e^{-\beta_{i_k} - \gamma C_{jk} - 1}\big)$
12: $\quad - \log x_{i_j}$
13: $P_{i_{jk}} \leftarrow e^{-\alpha_{i_j} - \beta_{i_k} - \gamma C_{jk} - 1}$
14: **end for**
15: **end for**
16: $\mathcal{L}_\gamma \leftarrow \sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{k=1}^{d} C_{jk} P_{i_{jk}} - n\zeta$
17: $\mathcal{L}_{\gamma\gamma} \leftarrow -\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{k=1}^{d} C_{jk}^2 P_{i_{jk}}$
18: $\eta \leftarrow 1$
19: **while** $\gamma - \eta \frac{\mathcal{L}_\gamma}{\mathcal{L}_{\gamma\gamma}} < 0$ **do**
20: $\eta \leftarrow \eta/2$
21: **end while**
22: $\gamma \leftarrow \gamma - \eta \frac{\mathcal{L}_\gamma}{\mathcal{L}_{\gamma\gamma}}$
23: **until** stopping criteria are satisfied
24: $v_i^{(l+1)} \leftarrow P_i^\top \cdot 1$
25: **until** stopping criteria are satisfied
26: **return** $\{P_i\}_{i=1}^n, \{v_i^{(l+1)}\}_{i=1}^n$
27: **end function**

so that, up to the scaling variables (α_i, β_i)*, one has the usual kernel* $\exp(-\gamma C)$*. The* α_i *update rescales rows to satisfy the prescribed row-sum constraints (matching* x_i*), while the scalar* γ *is updated to enforce the transport-budget constraint. Unlike standard Projected Sinkhorn where both marginals are fixed and* β_i *rescales columns to match a prescribed target marginal, here the column scaling via* β_i *is coupled with the outer CCP step: the induced targets* v_i *are updated iteratively so as to (locally) minimize the linearized objective in the inner problem.*

We provide numerical evidence on the computational requirements of Algorithm 1 in Section IV. The inner loop (steps 9–23) is regarded as the preferred approach for solving optimal transport problems [41], as exact solutions are expensive at moderate scales, while entropic regularization enables computations that are orders of magnitude faster. In contrast, the impact of the outer loop on scalability is harder to quantify, as it depends on the number of linearization steps (step 5) required and is therefore problem-dependent.

This article has been accepted for publication in IEEE Open Journal of Control Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCSYS.2026.3679905

IV. Numerical Examples

We evaluate the proposed attack on synthetic and real datasets. The synthetic experiments illustrate the effect of the entropic penalization, while the MNIST experiments compare Partial-Sinkhorn with FGSM under a common perturbation metric[2].

A. Synthetic Data

We began by examining synthetic data in order to analyze the algorithm and visualize its behavior in low dimensions. Figure 2 shows simulations on linearly separable classes (blue and red) in $\mathbb{R}^2$. We sampled 1000 data points and fitted a linear model by minimizing the binary cross-entropy (BCE) loss with a sigmoid output layer using batch stochastic gradient descent over multiple epochs. We then generated DRO and penalized-DRO adversarial samples, assuming that the adversary could perturb each sample by at most 0.1 in norm.

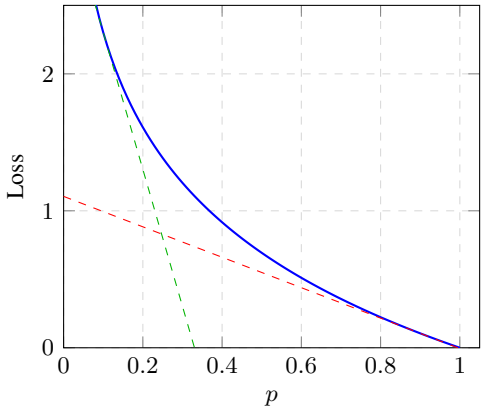

Figure 1: BCE loss gradient: correct classification (dashed red) vs. misclassification (dashed green).

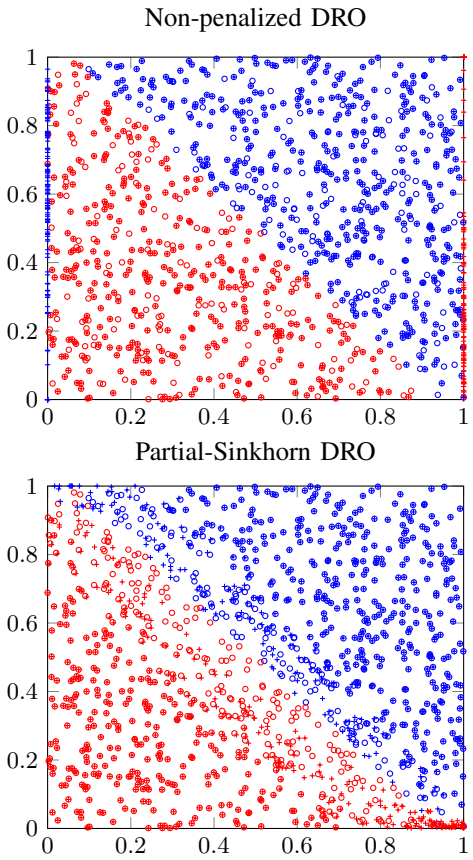

Figure 2: Effect of entropic penalization on a 2D linearly separable binary dataset: blue/red ∘ denote class 0/1 samples, and blue/red + denote the corresponding adversarial samples. Overlap indicates unchanged samples.

Because classification losses such as BCE can exhibit larger gradients for poorly classified samples than for correctly classified ones (see Figure 1), the iterative algorithm tends to prioritize updates on already misclassified samples. As a result, those samples are pushed further toward the boundary of the feasible set, while less attention is allocated to samples that are still correctly classified. This behavior is counterproductive, since the adversary seeks to increase the number of misclassifications rather than merely amplify the loss.

This effect is visible in the top image of Figure 2. To counter this effect, 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.

B. Real Datasets

We continue our numerical studies by applying the same methodology on the MNIST dataset. 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 the classification accuracy on an out-of-sample adversarial test set with the average absolute deviation,

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}|x_{ij}-v_{ij}|,$$

which quantifies the perturbation introduced by the adversary. This metric is directly proportional to the adversarial power ζ, but provides a fairer basis for comparing attacks of different types. FGSM is naturally parameterized by an ℓ_∞-budget, whereas our method is constrained through a transport-based budget. For this reason, we compare the attacks through the resulting average absolute deviation, which provides a common and more fair empirical perturbation metric across methods. Lower values therefore indicate less tampering with the data.

1) Linear binary classifier

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). In this case, the assumptions

[2]Code available at https://github.com/f2cf2e10/dro

This article has been accepted for publication in IEEE Open Journal of Control Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCSYS.2026.3679905

introduced in Section III are satisfied. 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.

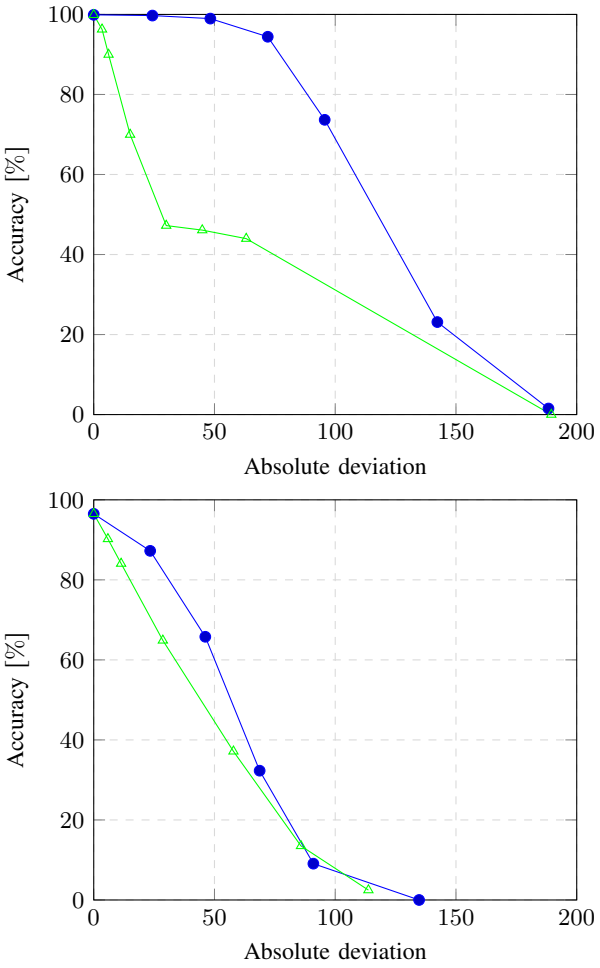

Figure 3: Out-of-sample accuracy of linear classifiers under adversarial perturbations: FGSM (blue) vs. Partial-Sinkhorn (green). Top: MNIST 0/1; bottom: MNIST 3/8.

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.

Since execution time depends strongly on hardware and implementation details, we use iteration counts as a more implementation-independent proxy for computational effort. The proposed method required on average 7 outer CCP iterations and 56 inner Sinkhorn iterations. For reference, with a batch size of 64, the average runtime was about 394 ms per inner loop, measured in a Docker container on a virtual machine on an AMD-EPYC Milan server using a single vCPU core and no GPU acceleration.

V. Conclusion

In this work, we formulated adversarial attacks as a data-driven distributionally robust optimization problem under Wasserstein ambiguity. We showed that the resulting nonconvex problem admits an entropically regularized DC reformulation, and we developed the Partial Sinkhorn algorithm to compute stationary solutions efficiently. Experiments on synthetic and real datasets show that the proposed method generates effective adversarial attacks under comparable perturbation budgets. Beyond adversarial learning, the proposed framework also highlights connections between optimal transport, distributional robustness, and control-oriented optimization.

Acknowledgments

We are grateful to the anonymous reviewers for their careful and thorough scrutiny, which has substantially improved the manuscript.

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] 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.

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

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

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.
[14] 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.
[15] H. Rahimian and S. Mehrotra, "Distributionally robust optimization: A review," 2019.
[16] H. Rahimian and S. Mehrotra, "Frameworks and results in distributionally robust optimization," *Open Journal of Mathematical Optimization*, vol. 3, p. 1–85, July 2022.
[17] B. P. G. Van Parys, D. Kuhn, P. J. Goulart, and M. Morari, "Distributionally robust control of constrained stochastic systems," *IEEE Transactions on Automatic Control*, vol. 61, no. 2, pp. 430–442, 2016.
[18] I. Yang, "Wasserstein distributionally robust stochastic control: A data-driven approach," *IEEE Transactions on Automatic Control*, vol. 66, no. 8, pp. 3863–3870, 2021.
[19] A. Hakobyan and I. Yang, "Wasserstein distributionally robust control of partially observable linear stochastic systems," *IEEE Transactions on Automatic Control*, vol. 69, no. 9, pp. 6121–6136, 2024.
[20] F. Wu, M. E. Villanueva, and B. Houska, "Ambiguity tube mpc," *Automatica*, vol. 146, p. 110648, 2022.
[21] A. Zolanvari and A. Cherukuri, "Data-driven distributionally robust iterative risk-constrained model predictive control," in *Proc. European Control Conference (ECC)*, pp. 1578–1583, 2022.
[22] J. Wang, R. Gao, and Y. Xie, "Sinkhorn distributionally robust optimization," *Operations Research*, vol. 0, no. 0, p. null, 0.
[23] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*. New York: John Wiley & Sons, 2nd ed., 1999.
[24] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317 – 331, 2018.
[25] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," *CoRR*, vol. abs/1810.00069, 2018.
[26] 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.
[27] 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.
[28] 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.
[29] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *CoRR*, vol. abs/1607.02533, 2016.
[30] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2019.
[31] G. Monge, *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale, 1781.
[32] L. Kantorovich and G. S. Rubinstein, "On a space of totally additive functions," *Vestnik Leningrad. Univ*, vol. 13, pp. 52–59, 1958.
[33] S. Kolouri, S. Park, M. Thorpe, D. Slepčev, and G. K. Rohde, "Transport-based analysis, modeling, and learning from signal and data distributions," 2016.
[34] C. Villani, *Optimal transport – Old and new*, vol. 338, pp. xxii+973. 01 2008.
[35] R. Rockafellar and R. J.-B. Wets, *Variational Analysis*. Heidelberg, Berlin, New York: Springer Verlag, 1998.
[36] P. D. Tao and E. B. Souad, "Algorithms for solving a class of nonconvex optimization problems. methods of subgradients," *Northholland Mathematics Studies*, vol. 129, pp. 249–271, 1986.
[37] 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.
[38] A. L. Yuille and A. Rangarajan, "The concave-convex procedure (cccp)," in *Advances in Neural Information Processing Systems* (T. Dietterich, S. Becker, and Z. Ghahramani, eds.), vol. 14, MIT Press, 2001.
[39] T. Lipp and S. Boyd, "Variations and extension of the convex–concave procedure," *Optimization and Engineering*, vol. 17, pp. 263–287, Jun 2016.
[40] R. Sinkhorn and P. Knopp, "Concerning nonnegative matrices and doubly stochastic matrices," *Pacific Journal of Mathematics*, vol. 21, pp. 343–348, 1967.
[41] 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.
[42] W. Rudin, *Real and Complex Analysis*. New York: McGraw-Hill, 3 ed., 1987.
[43] M. Sion, "On general minimax theorems," *Pacific Journal of Mathematics*, vol. 8, no. 1, pp. 171–176, 1958.
[44] 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.
[45] 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.
[46] 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. Proof of Proposition III.1

Proof:
Recall that $D^\star$ denotes the optimal value of problem (8), while $L^\star$ denotes the optimal value of its finite-dimensional surrogate (9). We begin from the distributional problem and derive the surrogate formulation. The only non-exact step in the argument occurs in the passage from (31) to (32).

Let us start by defining for each $y \in \Upsilon$ and a fixed $h \in \mathcal{H}$ the function $\varphi_y(v) := \ell(h(v), y)$, $v \in \Xi$.

The optimization problem (6) in integral form becomes,

$$\begin{array}{ll} \sup\limits_{\mathbb{Q} \in \mathcal{P}(\Xi \times \Upsilon)} & \int_{\Xi \times \Upsilon} \varphi_y(v) \mathrm{d}\mathbb{Q}(v, y) \\ \text{s.t.} & \inf\limits_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\Sigma} \|x - v\| \mathrm{d}\pi(x, y, v) \leqslant \zeta, \end{array} \tag{17}$$

which can be simplified to,

$$\begin{array}{ll} \sup\limits_{\mathbb{Q} \in \mathcal{P}(\Xi \times \Upsilon), \pi \in \Pi(\mathbb{P}, \mathbb{Q})} & \int_{\Xi \times \Upsilon} \varphi_y(v) \mathrm{d}\mathbb{Q}(v, y) \\ \text{s.t.} & \int_{\Sigma} \|x - v\| \mathrm{d}\pi(x, y, v) \leqslant \zeta, \end{array} \tag{18}$$

where the integrals in the constraints are taken in the domain $\Sigma := \Xi \times \Upsilon \times \Xi$ and $\mathcal{P}(\Xi \times \Upsilon)$ denotes the set of Borel *probability* measures on $\Xi \times \Upsilon$. Since Ξ is compact and Υ is finite, $\Xi \times \Upsilon$ is compact; hence $\mathcal{P}(\Xi \times \Upsilon)$ is tight and weakly compact (by Prokhorov's theorem).

The equivalence between (17) and (18) follows from the definition of the Wasserstein constraint: $W_1(\mathbb{P}, \mathbb{Q}) \leqslant \zeta$ holds if and only if there exists a coupling $\pi \in \Pi(\mathbb{P}, \mathbb{Q})$ such that

$$\int_{\Sigma} \|x - v\|_1 \, \mathrm{d}\pi(x, y, v) \leqslant \zeta.$$

Therefore we may equivalently optimize jointly over $\mathbb{Q}$ and an admissible coupling π as in (18).

Since neither the learner nor the adversary has access to $\mathbb{P}$, we replace it by the empirical distribution $\widehat{\mathbb{P}}_n$ constructed from

This article has been accepted for publication in IEEE Open Journal of Control Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCSYS.2026.3679905

the samples $S = ((x_1, y_1), \dots, (x_n, y_n))$. The Wasserstein constraint is therefore imposed with respect to $\widehat{\mathbb{P}}_n$, which yields

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

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 [42], (19) becomes,

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

Here $\mathcal{M} := \mathcal{P}(\Xi)$ since each $\mathbb{Q}_i$ is a probability measure on Ξ.

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

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

Denote the bracketed objective by $\mathcal{L}(\{\mathbb{Q}_i\}, \lambda)$. Since Ξ is compact, $\mathcal{M}$ (and hence $\mathcal{M}^n$) is convex and weakly compact in the weak topology (Prokhorov). Moreover, $\mathcal{L}(\cdot, \lambda)$ is linear (hence concave) in $\{\mathbb{Q}_i\}$ and $\mathcal{L}(\{\mathbb{Q}_i\}, \cdot)$ is affine (hence convex) in λ, and the required (upper/lower) semicontinuity holds by Assumption III.1. Therefore, Sion's minimax theorem [43] applies and yields

$$\inf_{\lambda \geqslant 0} \sup_{\{\mathbb{Q}_i\} \in \mathcal{M}^n} \mathcal{L}(\{\mathbb{Q}_i\}, \lambda) = \sup_{\{\mathbb{Q}_i\} \in \mathcal{M}^n} \inf_{\lambda \geqslant 0} \mathcal{L}(\{\mathbb{Q}_i\}, \lambda),$$

which justifies interchanging $\inf_{\lambda \geqslant 0}$ and $\sup_{\{\mathbb{Q}_i\} \in \mathcal{M}^n}$ and results in,

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

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

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

This can alternatively be written as,

$$\sup_{V_i \in \mathcal{M}} \quad \int_\Omega \Big(\varphi_{y_i}(V_i(\omega)) - \lambda \|V_i(\omega) - x_i\| \Big) \mathrm{d}\mu(\omega). \tag{25}$$

By [35, Thrm. 14.60], 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 $V_i \in \mathcal{M}$, where $\mathcal{M}$ 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,

$$\int_\Omega \|V(\omega)\| \mathrm{d}\pi(\omega) = \int_A \|V_1(\omega)\| \mathrm{d}\pi(\omega) + \int_{A^c} \|V_0(\omega)\| \mathrm{d}\pi(\omega),$$

is $< \infty$, hence, it is decomposable, and the theorem applies, leading to,

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

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,

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

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

$$\sup_{v_i \in \Xi} \quad \Big(\varphi_{y_i}(v_i) - \lambda \|v_i - x_i\| \Big). \tag{28}$$

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

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

or, in epigraphic form,

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

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

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

For each i, let

$$F_i(v_i, z_i) := \varphi_{y_i}(v_i) - \langle z_i, v_i - x_i \rangle.$$

Then the constraint in (31) can be written as

$$\sup_{v_i \in \Xi} \min_{\|z_i\|_* \leqslant \lambda} F_i(v_i, z_i) \leqslant \gamma_i.$$

By weak duality,

$$\sup_{v_i \in \Xi} \min_{\|z_i\|_* \leqslant \lambda} F_i(v_i, z_i) \;\leqslant\; \min_{\|z_i\|_* \leqslant \lambda} \sup_{v_i \in \Xi} F_i(v_i, z_i).$$

Therefore, replacing the left-hand side by the right-hand side yields the more conservative formulation (32). This is precisely the step that yields the inequality $L^\star \leqslant D^\star$.

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

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

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

Dualizing the constraint in (33) with multipliers $\alpha_i,\ \beta_i$

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

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

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

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

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

which is equivalent to,

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

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,

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

Because the variables separate across i, we may write the supremum jointly over $\{u_i, v_i\}_{i=1}^n$, which yields

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

For fixed u_i and v_i, the inner infimum over z_i is finite only if

$$u_i + \frac{v_i - x_i}{n} = 0, \quad \text{for all } i = 1, \dots, n,$$

since otherwise the linear term in z_i is unbounded below. Hence,

$$u_i = \frac{x_i - v_i}{n}, \quad \text{for all } i = 1, \dots, n,$$

and substituting this identity into the constraint $\sum_{i=1}^n \|u_i\| \leqslant \zeta$ gives

$$\frac{1}{n} \sum_{i=1}^{n} \|x_i - v_i\| \leqslant \zeta.$$

Therefore, (39) reduces to,

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

Problem (40) is exactly the finite-dimensional surrogate (9), whose optimal value is $L^\star$. ■

B. 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. Using [35, Thrm. 14.60], we show that the problem of finding the measure is equivalent to a pointwise optimization problem, which can be interpreted as a discrete-to-discrete transport problem.

This article has been accepted for publication in IEEE Open Journal of Control Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCSYS.2026.3679905

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,

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

See [34] for the discrete optimal transport formulation in which any coupling between two weighted discrete probability measures can be represented by a nonnegative matrix $P \geqslant 0$ satisfying $P\mathbf{1} = x$ and $P^\top \mathbf{1} = v$, which yields the following Kantorovich problem,

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

for an appropriate cost matrix C.

In this case, the optimal transport consists in finding an optimal matching between the points in x and the target points v.

Let C satisfy $C_{ij} = 2\mathbb{1}_{i\neq j}$, that is,

$$C_{ij} = \begin{cases} 2, i \neq j \\ 0, \text{otherwise.} \end{cases}$$

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

$$\begin{aligned} \min_{P} \quad & 2\sum_i \sum_j P_{ij}\mathbb{1}_{i\neq j} \\ \text{s.t.} \quad & P \cdot 1 = x \\ & P^T \cdot 1 = v, \end{aligned} \tag{42}$$

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 $\sum_i \sum_j P_{ij}\mathbb{1}_{i\neq j} = 1 - \sum_i P_{ii}$, as $\sum_{ij} P_{ij} = 1$.

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,

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

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,

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

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

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

and (40) is equivalent to (Note that here we define for each $y \in \Upsilon$ and a fixed $h \in \mathcal{H}$ the function $\varphi_y(v) := \ell(h(v), y)$, $v \in \Xi$)

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

Consider now the problem

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

we have that,

$$K^\star = M^\star,$$

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).

■

C. 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.

This article has been accepted for publication in IEEE Open Journal of Control Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCSYS.2026.3679905

Algorithm

For algebraic convenience, we rescale the objective by $1/\lambda$ with $\lambda > 0$, i.e., we use the inverse-temperature parameter $\tau := 1/\lambda$ multiplying the loss, and work with the equivalent formulation,

$$\begin{aligned} D^\star_\lambda := \inf_{\substack{\{v_i \in \Xi\}_{i=1}^n, \\ \{P_i \in [0,1]^{d\times d}\}_{i=1}^n}} & \frac{1}{\lambda n}\sum_{i=1}^n -\ell(h(v_i), y_i) + \\ & \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^d\sum_{k=1}^d P_{i_{jk}}\log P_{i_{jk}} \\ \text{s.t.} \quad & \frac{1}{n}\sum_{i=1}^n \langle P_i, C\rangle_F \leqslant \zeta \\ & P_i\cdot \mathbf{1} = x_i, \forall i = 1,\dots,n \\ & P_i^\top \cdot \mathbf{1} = v_i, \forall i = 1,\dots,n. \end{aligned} \tag{47}$$

Let us start by deriving the dual formulation of (47). First, we look at the Lagrangian,

$$\begin{aligned} \mathcal{L}(v_i, P_i, \alpha_i, \beta_i, \gamma) = & \tau\sum_{i=1}^n -\nabla^T_{(\ell_{y_i}\circ h)}(v_i^{(k)})v_i + \\ & \sum_{i=1}^n\sum_{j=1}^d\sum_{k=1}^d P_{i_{jk}}\log P_{i_{jk}} + \\ & \gamma\Big(\sum_{i=1}^n\sum_{j=1}^d\sum_{k=1}^d P_{i_{jk}}C_{jk} - n\zeta\Big) + \\ & \sum_{i=1}^n \alpha_i^T(P_i\cdot 1 - x_i) + \\ & \sum_{i=1}^n \beta_i^T(P_i^T\cdot 1 - v_i), \end{aligned} \tag{48}$$

in which the primal and dual variables are given by,

- *Primal variables:* for $i = 1,\dots,n$, the perturbed samples $v_i \in \Xi$ and the associated transport plans $P_i \in \mathbb{R}_+^{d\times d}$ (or $\mathbb{R}_+^{n\times n}$, depending on the discretization) satisfying the marginal constraints
$$P_i\mathbf{1} = x_i, \qquad P_i^\top \mathbf{1} = v_i.$$
- *Dual variables:* Lagrange multipliers $\alpha_i, \beta_i \in \mathbb{R}^d$ for the marginal constraints, and the regularization/budget multiplier $\lambda > 0$ (equivalently $\tau := 1/\lambda > 0$ in standard entropic-OT notation) as well as $\gamma \geqslant 0$ for the budget constraint.

We continue by solving for v_i and P_i, in a similar fashion to that of [7], [41],

$$\begin{aligned} \frac{\partial}{\partial P_{i_{jk}}}\mathcal{L}(v_i, P_i, \alpha_i, \beta_i, \gamma) = 0, \\ \frac{\partial}{\partial v_i}\mathcal{L}(v_i, P_i, \alpha_i, \beta_i, \gamma) = 0, \end{aligned} \tag{49}$$

which leads to,

$$\begin{aligned} P_{i_{jk}} &= e^{-\alpha_{i_j} - \beta_{i_k} - \gamma C_{jk} - 1}, \\ \beta_i &= -\tau\nabla_{(\ell_{y_i}\circ h)}(v_i^{(k)}). \end{aligned} \tag{50}$$

Now, substituting this on the Lagrangian, results in,

$$\mathcal{L}(\alpha_i, \gamma) = -\sum_{i=1}^n\sum_{j=1}^d\sum_{k=1}^d P_{i_{jk}} - \gamma n\zeta - \sum_{i=1}^n \alpha_i^T\cdot x_i. \tag{51}$$

Note that β_i is already known in (50), we just keep it on the RHS for readability purposes. The Lagrange dual formulation of (47) can be written as the following unconstrained maximization of a concave objective over the dual variables,

$$\begin{aligned} \max_{\substack{\{\alpha_i \in \mathbb{R}^d\}_{i=1}^n, \\ \gamma \in \mathbb{R}_+}} & -\sum_{i=1}^n\sum_{j=1}^d\sum_{k=1}^d P_{i_{jk}} - \gamma n\zeta - \sum_{i=1}^n \alpha_i^T\cdot x_i. \end{aligned} \tag{52}$$

This is an unconstrained concave maximization problem (equivalently, a convex minimization after a sign change). Taking the partial derivatives with respect to α_i and γ and equating them with 0, results in the unconstrained optima,

$$\alpha^\star_{i_j} = \Big(\log\big(\sum_{k=1}^d e^{-\beta_{i_k} - \gamma^\star C_{jk} - 1}\big) - \log x_{i_j}\Big), \tag{53}$$

where $\beta_i = -\tau\nabla_{(\ell_{y_i}\circ h)}(v_i^{(k)})$, while $\gamma^\star$ cannot be solved analytically but can be computed by means of Newton's second order procedure,

$$\begin{aligned} \gamma &= \gamma - \frac{\frac{\partial}{\partial\gamma}\mathcal{L}}{\frac{\partial^2}{\partial\gamma^2}\mathcal{L}} \\ \frac{\partial}{\partial\gamma}\mathcal{L} &= \sum_{i=1}^n\sum_{j=1}^d\sum_{k=1}^d C_{jk}P_{i_{jk}} - n\zeta \\ \frac{\partial^2}{\partial\gamma^2}\mathcal{L} &= -\sum_{i=1}^n\sum_{j=1}^d\sum_{k=1}^d C_{jk}^2 P_{i_{jk}}. \end{aligned} \tag{54}$$

Finally, the optimal solution for the primal problem is,

$$\begin{aligned} P^\star_{i_{jk}} &= e^{-\alpha^\star_{i_j} - \beta_{i_k} - \gamma^\star C_{jk} - 1}, \\ v_i^\star &= P_i^{\star^T}\cdot 1. \end{aligned} \tag{55}$$

Convergence

We continue by simplifying (47) 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,

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

This article has been accepted for publication in IEEE Open Journal of Control Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJCSYS.2026.3679905

and

$$\begin{aligned} \min_{\{P_i \in [0,1]^{d\times d}\}_{i=1}^n} \quad & \frac{1}{n}\sum_{i=1}^n -\nabla^T_{(\ell_{y_i}\circ h)}(v_i^{(k)})\cdot P_i^T\cdot 1 + \\ & \frac{\lambda}{n}\sum_{i=1}^n\sum_{j=1}^d\sum_{k=1}^d P_{i_{jk}}\log P_{i_{jk}} \\ \text{s.t.} \quad & \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^d\sum_{k=1}^d P_{i_{jk}}C_{jk} \leqslant \zeta \\ & P_i\cdot 1 = x_i, \\ & i=1,\dots,n. \end{aligned} \tag{57}$$

But (56) is a DC optimization problem such as the one in (12),

$$\begin{aligned} \min_{x\in C} \quad & u(x)-v(x) \\ \text{s.t.} \quad & f(x)\leqslant 0 \\ & g_i(x)=0 \\ & i=1,\dots,n, \end{aligned} \tag{58}$$

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

Consider now the objective function,

$$\phi(x) = u(x) - v(x). \tag{59}$$

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

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

This tells us that,

$$\phi(x^{(k+1)}) \leqslant \psi(x^{(k+1)}, x^{(k)}) \leqslant \psi(x^{(k)}, x^{(k)}) = \phi(x^{(k)}). \tag{60}$$

Let us denote our algorithm $\mathcal{A}$. This is a point-to-set map, that is, given an initial guess, $x^{(0)}$, it generate a sequence $\{x^{(k)}\}_{k=0}^\infty$ through the iteration,

$$\begin{aligned} x^{(k+1)} \in \underset{x\in C}{\operatorname{argmin}} \quad & u(x) - v(x^{(k)}) - (x-x^{(k)})^T\nabla v(x^{(k)}) \\ \text{s.t.} \quad & f(x)\leqslant 0 \\ & g_i(x)=0 \\ & i=1,\dots,n. \end{aligned} \tag{61}$$

A fixed point $x^\star$ of the map $\mathcal{A}$ is a point such that $x^\star \in \mathcal{A}(x^\star)$. That is, if the algorithm is initialized at $x^\star$, it remains at that point. We may now apply Zangwill's convergence theorem [44] to establish the main result, following a similar approach to [45].

Theorem. (Convergence Theorem [44]) *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, $\lim_{k\to\infty}\phi(x^{(k)}) = \phi(x^\star)$ for all limit points $x^\star$.

In our problem, all assumptions of the theorem are satisfied.

Assumption 1

First, we verify that Assumption 1. From the proof of Appendix B, for each i, the transport plan P_i belongs to the transport polytope

$$\mathcal{U}(a_i, b_i) := \{P \in \mathbb{R}_+^{d\times d} : P\mathbf{1} = a_i,\ P^\top\mathbf{1} = b_i\}.$$

Since $\mathcal{U}(a_i, b_i)$ is a closed and bounded subset of the finite-dimensional space $\mathbb{R}^{d\times d}$, it is compact. Therefore all iterates $\{P_i^{(k)}\}_{i=1}^n$ remain in a compact set.

Assumption 2

Let Γ denote the set of all fixed points, $x^\star$ of (61), i.e., all points $x^\star$ such that $x^\star \in \mathcal{A}(x^\star)$ with ϕ as in (59). Then condition 2b in the convergence theorem holds with equality by the definition of Γ, while condition 2a follows by the definition of Γ and the descent inequality (60).

Assumption 3

Closeness follows directly from Lemma 6 in [45], originally stablished in [46]

Convergence

Applying the convergence theorem, we conclude that any convergent subsequence $\{x^{(k)}\}_{k\geqslant 0}$ produced by $\mathcal{A}$ converges to a point in Γ, and that $\phi(x^{(k)}) \to \phi(x^\star)$ for every limit point $x^\star$.

It remains to relate the limit points of the algorithm to the original problem (58). This follows by verifying the KKT conditions. Let $x^\star$ be a limiting fixed point of (61). Then there exist Lagrange multipliers $\alpha^\star, \beta_1^\star, \dots, \beta_n^\star$ satisfying the KKT conditions,

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

These are the KKT conditions for the original problem (58). Hence, the limit points are first-order KKT points of the original problem. ■