A MINIMALIST APPROACH FOR DOMAIN ADAPTATION WITH OPTIMAL TRANSPORT

Arip Asadulaev
Artificial Intelligence Research Institute
ITMO University
aripasadulaev@itmo.ru

Vitaly Shutov
ITMO University
VK

Alexander Korotin
Artificial Intelligence Research Institute
Skoltech

Alexander Panfilov
Eberhard Karls Universität Tübingen

Vladislava Kontsevaya
MIPT

Andrey Filchenkov
GO AI Lab

ABSTRACT

We reveal an intriguing connection between adversarial attacks and cycle monotone maps, also known as optimal transport maps. Based on this finding, we developed a novel method named *source fiction* for semi-supervised optimal transport-based domain adaptation. We conduct experiments on various datasets and show that our method can notably improve the performance of the optimal transport solvers in domain adaptation.

1 INTRODUCTION

Optimal Transport (OT) is a powerful framework for solving mass-moving problems for probability distributions. It was successfully applied in mathematics (Ferradans et al., 2014), economics (Reich, 2013), and machine learning (Arjovsky et al., 2017; Mroueh, 2019; Solomon et al., 2015; Colombo et al., 2021), especially in domain adaptation (DA) problem (Courty et al., 2015; Perrot et al., 2016; Rakotomamonjy et al., 2020). Usually, the domain adaptation problem involves two domains: a fully labeled source domain denoted by $\mathbb{Q}$ and an unlabeled or partially labeled target domain denoted by $\mathbb{P}$. The goal is to make correct predictions on the unlabeled target domain samples while being trained on the source domain samples.

In contrast to deep domain adaptation techniques (Ganin & Lempitsky, 2015; Long et al., 2018; Gretton et al., 2012; Long et al., 2015; 2017), optimal transport methods are very fast, have low computational complexity, and offer theoretical guarantees (Redko et al., 2017) in domain adaptation. But, in practice, simple OT-based methods are provide lower domain adaptation accuracy. The goal of our paper is to provide simple techniques that can improve the accuracy of OT solvers in the domain adaptation problem. To achieve this, we firstly ask the question: *what is the main reason for the inaccuracy of the OT solvers on empirical datasets?*

It is known, that the main geometric property of the optimal transport maps is c-cyclical monotonicity (Villani, 2008, §5). In the context of domain adaptation, this means that optimal transport match source samples to nearby target samples in terms of the cost function. However, being close in terms of Euclidean cost doesn't always mean that the samples belong to the same class. This leads to inaccuracies in optimal transport for domain adaptation. Our method addresses this issue.

The core of our method is the idea to modify the target dataset to ensure that samples with the same labels in source and target are close in terms of Euclidean cost, satisfying cyclical monotonicity. To apply a such transformation we used our finding that adversarial attacks (Goodfellow et al., 2014) with a small enough perturbation size ϵ value provides a cyclically monotone map of the data.

More precisely, our approach incorporates several steps. Initially, we train the classifier f_θ on samples from the source domain. Then, we convert the **labeled target samples into the correctly classified ones by using the *inverted adversarial attack* on the classifier** f_θ. The *inverted adversarial attack* (Huang et al., 2021; Mao et al., 2021) is the same as the original targeted adversarial attack 1, but the target class $\mathbf{k}$ is set to the true class of the sample. **After that, we use optimal transport T_γ to approximate these perturbations**. Finally, we apply T_γ to the unlabeled target samples to map them into the correctly classified by f_θ.

In our experiments, we demonstrate that our approach enhances discrete optimal transports (Courty et al., 2015; Flamary et al., 2021) solvers' accuracy in a variety of domain adaptation tasks (§6.4).

Contribution: We prove that the FSGD adversarial attacks with small parameter ϵ are c-cyclical monotone transformations of the dataset (§3) with quadratic cost. Using this property, we propose a new algorithm that improves the performance of optimal transport solvers in a number of domain adaptation problems. (§6).

Societal Impact: Nowadays, the data is shared on separate devices and usually contains personal information, which is inefficient for data transmission and may violate data privacy. The authors of the (Liang et al., 2020) address a challenging domain adaptation setting without access to the source data for higher privacy. In our method, we adapt the source classifier to the target domain **without access to the source data**, using only the source classifier, which can be used in various scenarios to avoid privacy issues.

Structure: The paper is structured as follows: first, we give an explanation of the adversarial attack, optimal transport, and domain adaptation (§2). We then prove that adversarial attacks are cyclical monotone mappings (§3). In (§4) we propose our algorithm and give a description of why it works. Finally, in (§6) we show the results of our experiments and summarize them in (§7).

2 Background

2.1 Adversarial Attacks

Adversarial examples are samples that are similar to the true samples $D(\mathbf{x}, \mathbf{x}') \leq \epsilon$, but "fool" a selected classifier and tend it to make incorrect predictions $\operatorname{argmax} p(\mathbf{k} \mid \mathbf{x}') \neq \mathbf{k}_{\text{true}}$ (Szegedy et al., 2014). A large body of work on adversarial attacks exists (Papernot et al., 2017; Yuan et al., 2019; Schott et al., 2019; Xie et al., 2017), and the phenomenon of the vulnerability of machine learning models to adversarial attacks breeds a great deal of concern in learning scenarios.

In our paper, we consider one of the simplest adversarial attacks: Fast Sign Gradient Descent (FSGD) (Goodfellow et al., 2014). The iterative version of the targeted FSGD can be presented as:

$$\mathbf{x}'_0 = \mathbf{x}, \quad \mathbf{x'}_{i+1} = \operatorname{clip}_{\mathbf{x},\epsilon} \left\{ \mathbf{x'}_i - \alpha \operatorname{sign} \left(\nabla_{\mathbf{x}} L \left(\theta, \mathbf{x'}_i, \mathbf{k} \right) \right) \right\} \tag{1}$$

With sample $\mathbf{x}$, class label $\mathbf{k}$, and classifier f_θ, we can obtain adversarial examples using gradient descent perturbations, minimizing the loss L on a sample $\mathbf{x}$ with respect to some perturbation size ϵ.

2.2 Optimal Transport for Domain Adaptation

Optimal transport. Optimal transport (OT) is a simple framework for solving mass-moving problems for probability distributions (Ferradans et al., 2014; Reich, 2013; Arjovsky et al., 2017; Mroueh, 2019; Solomon et al., 2015; Colombo et al., 2021).

Formally optimal transport aims at finding a cost-effective mapping $T : X \to Y$ of the two probability measures $\mathbb{P} = \sum_{i=1}^{n} a_i \mathbf{x}_i$ and $\mathbb{Q} = \sum_{j=1}^{m} b_j \mathbf{y}_j$ with respect to the cost function $c : X \times Y \to \mathbb{R}_+$, where a_i and b_i are the values of the Dirac function at $\mathbf{x}_i$ and $\mathbf{y}_j$ correspondingly.

Monge's problem was the first example of the optimal transport problem (Villani, 2008, §3) and can be formally expressed as follows:

$$\inf_{T \# \mathbb{P} = \mathbb{Q}} \int_{\Omega_{\mathbb{P}}} c(\mathbf{x}, T(\mathbf{x})) \mathbb{P}(\mathbf{x}) d\mathbf{x} \tag{2}$$

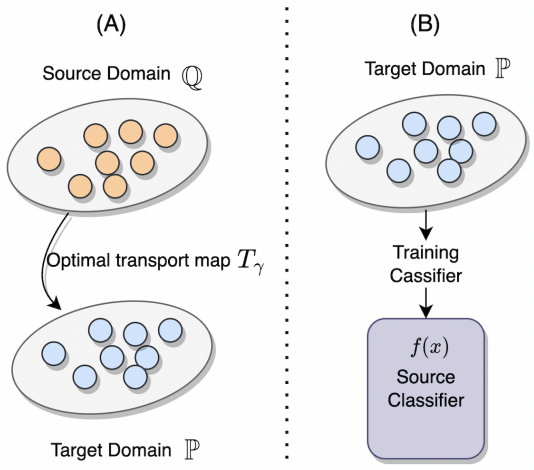
Monge's formulation of optimal transport aims at finding a map T where $T \# \mathbb{P} = \mathbb{Q}$ represents the mass preserving push forward operator. In Monge's formulation, for two given measures $\mathbb{P}$ and $\mathbb{Q}$, the existence of a transport map T is not only non-trivial but it also may not exist (Villani, 2008, §5.1).

Kantorovich proposed the relaxation of Monge's problem2 and presented the formulation in which a solution always exists (Villani, 2008, §5.1). The Kantorovich problem aims to find a joint distribution over the $\mathbb{P}$ and the $\mathbb{Q}$ that determines how the mass is allocated. To find an optimal solution, it is necessary to build the cost matrix for all $\mathbf{x} \in X$ and $\mathbf{y} \in Y$ samples:

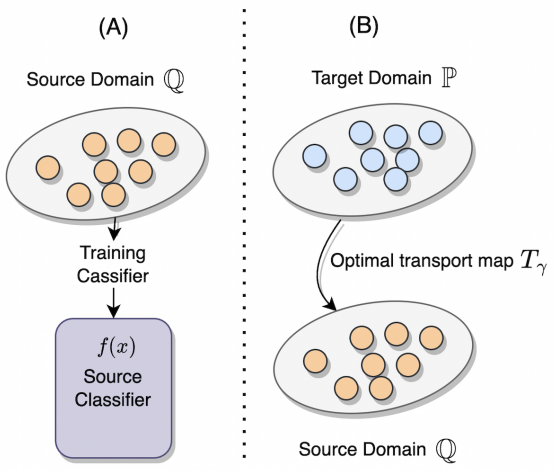
$$M_{XY} \stackrel{\text{def}}{=} \left[c \left(\mathbf{x}_i, \mathbf{y}_j \right)^p \right]_{ij} \tag{3}$$

Having the cost matrix M_{XY}, the optimization goal is to find the optimal coupling γ, that minimizes the displacement cost between two probability measures $\mathbb{P}$ and $\mathbb{Q}$

$$W_p^p(\mathbb{P}, \mathbb{Q}) = \min_{\gamma \in U(a,b)} \langle \gamma, M_{XY} \rangle \tag{4}$$



(a) Label transfer for domain adaptation. (A) Match labeled source samples $\mathbb{Q}$ with unlabeled target samples $\mathbb{P}$ by optimal transport T_γ. (B) Train or fine-tune a classifier f_θ on target samples $\mathbb{P}$ labeled by the T_γ map.



(b) Classifier domain adaptation. (A) Train classifier on the labeled source domain $\mathbb{Q}$. (B) Transform unlabeled target domain samples $\mathbb{P}$ into the source domain samples $\mathbb{Q}$, to increase the accuracy of f_θ on target domain $\mathbb{P}$.

with the constraints to the coupling $\gamma \in U(a, b)$ such that:

$$U(a,b) \stackrel{\text{def}}{=} \left\{\gamma \in \mathbb{R}_+^{n \times m} \mid \gamma \mathbf{1}_m = a, \gamma^\top \mathbf{1}_n = b\right\} \tag{5}$$

The infimum of this optimization problem induces the Wasserstein distance, and coupling γ gives us a non-bijective map between probability measures $\mathbb{P}$ and $\mathbb{Q}$.

For differentiable optimal transport, the Sinkhorn algorithm (Cuturi, 2013) was proposed. The Sinkhorn is based on the matrix-vector multiplication operations and can be combined with various regularizations like group lasso regularization (L1L2) and Laplacian regularization (L1LP) (Courty et al., 2015).

To make the optimal transport applicable to the out-of-sample mapping, a linear optimal transport mapping estimator (OTLin) was proposed (Perrot et al., 2016). OTLin jointly computes the Kantorovich coupling γ equation 4 and maps T linked to the original Monge problem equation 2.

Labels Transfer. Optimal transport can be a simple solution for the domain adaptation problem (Courty et al., 2015; Perrot et al., 2016; Rakotomamonjy et al., 2020). Usually, optimal transport is used to map labeled source samples in $\mathbb{Q}$ to the unlabeled or partly labeled samples in the target $\mathbb{P}$. We can name this process *labels transfer*, see Figure 1a.

In this setting, optimal transport provides a match between the labeled source samples and the unlabeled ones in the target. After matching, we set the target samples' labels equal to the source samples' corresponding labels.

To solve domain adaptation in this setting, linear programming-based solvers (Nash, 2000) are usually used. For example, the Earth Mover's Distance (EMD) solver is actively used in various domain adaptation scenarios (Courty et al., 2015; Flamary et al., 2021).

Classifier Domain Adaptation. Alternately, domain adaptation can be solved by adapting the source classifier f_θ to the target domain $\mathbb{P}$ (Ben-David et al., 2010a;b; Germain et al., 2013). In this scenario, a mapping function transforms the target domain $\mathbb{P}$ samples to "look like" the source domain $\mathbb{Q}$ samples after a classifier is trained on the labeled source samples (Ben-David et al., 2010a;b; Germain et al., 2013), see Figure 1b.

Deep distance-based algorithms (Gretton et al., 2012; Long et al., 2015; 2017) or adversarial-based algorithms (Ganin & Lempitsky, 2015; Long et al., 2018) are common solutions to this problem. With the help of these techniques, a feature extractor is learning to bring target domain samples closer to the source domain samples in the latent space. These techniques exhibit high accuracy but require time-consuming computations and changes to the source classifier's architecture.

2.2.1 CYCLICAL MONOTONICITY

The main geometric property of the optimal transport maps is c-cyclical monotonicity. Formally the map is c-cyclical monotone if for all points $\mathbf{x}_0 \dots \mathbf{x}_n$, $\mathbf{y}_0 \dots \mathbf{y}_n$, $n \in \mathbb{N}$, and every permutation σ holds:

$$\sum_{n=1}^{N} c\left(\mathbf{x}_n, \mathbf{y}_n\right) \leq \sum_{n=1}^{N} c\left(\mathbf{x}_n, \mathbf{y}_{\sigma(n)}\right) \tag{6}$$

The term "cyclical" refers to the fact that it suffices to test this property for cyclical permutations $\sum_{n=1}^{N} c\left(\mathbf{x}_n, \mathbf{y}_n\right) \leq \sum_{n=1}^{N} c\left(\mathbf{x}_n, \mathbf{y}_{n+1}\right)$, because every permutation σ is a composition of disjoint cycles.

A c-cyclical monotone map cannot be improved in terms of the cost function c (Villani, 2008, §5). Recently was showed that a generalization of c-cyclical monotonicity from the Monge-Kantorovich problem with two marginals gives rise to a sufficient condition for optimality also in the multi-marginal version of that problem (Griessler, 2018).

Let's consider the domain adaptation problem with optimal transport. Suppose we have two domains with samples $\mathbf{x}_0 \dots \mathbf{x}_n$ and $\mathbf{y}_0 \dots \mathbf{y}_n$ correspondingly. Then, following Eq. 6, optimal transport will match the x_n samples to the y_n samples that are close to each other in terms of the cost c. But closeness in terms of the cost c, doesn't always mean that these samples are from the same class (Li et al., 2019). This leads to the inaccuracy of optimal transport for domain adaptation.

3 ADVERSARIAL ATTACKS ARE CYCLICAL MONOTONE MAPS

In our paper, we consider the FSGD (Goodfellow et al., 2014) 1 attack. With mild assumptions, we prove that it is a cyclical monotone transformation of the data w.r.t. the quadratic cost $c(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2$.

Lemma 3.1 (cyclical monotonicity of small perturbations of a dataset.)**.** *Let $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^D$ be a dataset of N distinct samples. Let $\mathbf{x}'_1, \dots, \mathbf{x}'_N$ be its $\leq \epsilon$-perturbation, i.e. $\|\mathbf{x}_n - \mathbf{x}'_n\| \leq \epsilon$ for all $n = 1, 2, \dots, N$. Assume that $\epsilon \leq \frac{1}{2} \min_{n_1, n_2} \|\mathbf{x}_{n_1} - \mathbf{x}_{n_2}\|$, i.e. the perturbation does not exceed $\frac{1}{2}$ of the minimal pairwise distance between samples. Then for all K and N it holds:*

$$\sum_{k=1}^{K} \frac{1}{2}\|\mathbf{x}_{n_k} - \mathbf{x}'_{n_k}\|^2 \leq \sum_{k=1}^{K} \frac{1}{2}\|\mathbf{x}_{n_k} - \mathbf{x}'_{n_{k+1}}\|^2 \tag{7}$$

i.e. set $(\mathbf{x}_1, \mathbf{x}'_1), \dots, (\mathbf{x}_N, \mathbf{x}'_N)$ or, equivalently, the map $\mathbf{x}_k \mapsto \mathbf{x}_{k'}$ is cyclical monotone.

Proof. Due to triangle inequality for $\|\cdot\|$, we have

$$\|\mathbf{x}_{n_k} - \mathbf{x}'_{n_{k+1}}\| \geq \underbrace{\|\mathbf{x}_{n_k} - \mathbf{x}_{n_{k+1}}\|}_{\geq 2\epsilon} - \underbrace{\|\mathbf{x}_{n_{k+1}} - \mathbf{x}'_{n_{k+1}}\|}_{\leq \epsilon} = \epsilon. \tag{8}$$

Taking the square of both sides and summing equation 8 for $k = 1, 2, \dots, K$ yields

$$\sum_{k=1}^{K} \|\mathbf{x}_{n_k} - \mathbf{x}'_{n_{k+1}}\|^2 \geq \sum_{k=1}^{K} \epsilon^2 = K\epsilon^2. \tag{9}$$

Due to the assumptions of the lemma, the following inequality holds true:

$$\sum_{k=1}^{K} \|\mathbf{x}_{n_k} - \mathbf{x}'_{n_k}\|^2 \leq \sum_{k=1}^{K} \epsilon^2 \leq K\epsilon^2. \tag{10}$$

We combine equation 9 and equation 10 to obtain

$$\sum_{k=1}^{K} \|\mathbf{x}_{n_k} - \mathbf{x}'_{n_k}\|^2 \leq \sum_{k=1}^{K} \|\mathbf{x}_{n_k} - \mathbf{x}'_{n_{k+1}}\|^2,$$

which is equivalent to:

$$\sum_{k=1}^{K} c(\mathbf{x}_{n_k}, \mathbf{x}'_{n_k}) \leq \sum_{k=1}^{K} c(\mathbf{x}_{n_k}, \mathbf{x}'_{n_{k+1}}), \tag{11}$$

and yield c-cyclical monotone w.r.t. quadratic cost $c(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2$. □

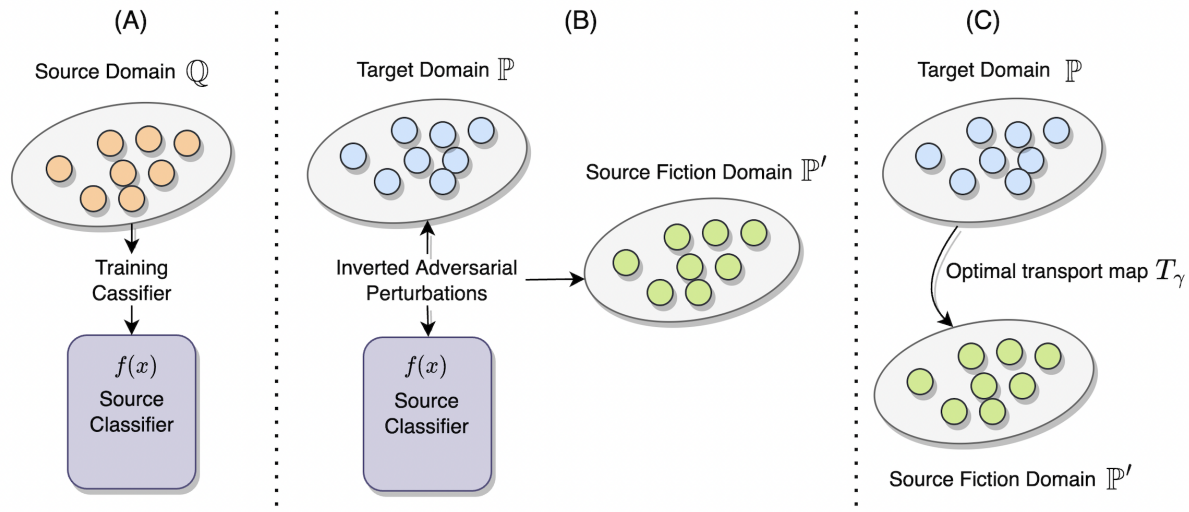


Figure 2: An illustration of the proposed pipeline. Our method takes three steps: (A) Pretrain a source classifier f_θ on the source data $\mathbb{Q}$. (B) Attack the source classifier with labeled target samples $\mathbb{P}$ to generate examples that classifiers correctly with high confidence $\mathbb{P}'$. (C) Use optimal transport T_γ to align the unlabeled target samples and label them using f_θ.

Adversarial attacks are small ϵ perturbations of the dataset samples, we immediately obtain:

Corollary 3.2. *Let $x_1, \dots, x_N \in \mathbb{R}^D$ be a dataset of N distinct samples. Then any adversarial attack $\mathbf{x}_n \mapsto \mathbf{x}'_n$ on the dataset with $\epsilon \leq \frac{1}{2} \min_{n_1, n_2} \|\mathbf{x}_{n_1} - \mathbf{x}_{n_2}\|$ is c-cyclical monotone.*

Corollary 3.2 suggests that FSGD attacks with a small parameter ϵ are c-cyclical monotone maps, i.e., optimal transport. In the next section, we present our algorithm that connects optimal transport and adversarial attacks for domain adaptation.

4 Proposed Method

In this section, we propose a new algorithm that generates a new domain by the *inverse adversarial attack.* Before introducing the algorithm, we outline the motivation for using an adversarial attack to generate a *source fiction* domain.

4.1 Motivation

While optimal transport maps are c-cyclical monotone, i.e., exhibit a specific structure of the map, thus, transportation $X \to Y$ via optimal transport maps with euclidean cost might not be applied to some problems (Li et al., 2019; Asadulaev et al., 2022), see (Courty et al., 2015, Figure 3) for counter-examples. Optimal transport applications for mass moving assume closeness of target $\mathbb{P}$ and source $\mathbb{Q}$ distributions (Lee et al., 2019). A cyclical monotone map with quadratic cost may not accurately capture the class-wise structure between domains.

For instance, in classifier adaptation settings (Figure 1b), samples from different domains but with the same labels are not always the closest if cost c is Euclidean (Li et al., 2019). As a result, the optimal transport aims to find a map that transforms the target sample into the incorrect class in the source domain. Actually, the label transfer setting has the same issue 1a.

However, if all samples $\mathbf{x}_n$ and $\mathbf{y}_n$ from the two domains are cyclically monotone with respect to quadratic cost and, at the same time, their labels are equal for each n 6, then optimal transport can accurately map between these domains and preserve a class-wise structure. Based on this, we can conclude that it would be the best possible scenario to have a source domain, where for each target sample $\mathbf{x}_n$, we have the closest sample $\mathbf{y}_n$ from the same class.

4.2 Algorithm

Algorithm 1 Algorithm for domain adaptation with *source fiction*

Input: Classifier f_θ, optimal transport T_γ, source samples Y, labeled X_l and unlabeled target samples X, perturbations size ϵ
Initialize: $\epsilon \leq \frac{1}{2} \min_{n_1,n_2} \|\mathbf{x}_{n_1} - \mathbf{x}_{n_2}\|$ for $\mathbf{x}_n$ in labeled X_l
Pretrain classifier f_θ on the source samples Y.
$Y \leftarrow \emptyset$
for $\mathbf{x}, \mathbf{k} \in X_l$ **do**
  $\mathbf{x}' \leftarrow \mathbf{x}$
  for some iterations **do**
    $\mathbf{x}' \leftarrow \text{clip}_{\mathbf{x},\epsilon} \{\mathbf{x}'_n - \alpha \operatorname{sign} (\nabla_{\mathbf{x}} L (\theta, \mathbf{x}', \mathbf{k}_{\text{true}}))\}$
  end for
  Add $\mathbf{x}'$ to the dataset X'
end for
Find a map T_γ using $X_l \to X'$.
Apply the map T_γ for $X \to X'$.
Apply the classifier f_θ to the output of the T_γ to label all target samples X.
return Labeled target samples.

To correctly capture the class-wise structure during domain adaptation via optimal transport, we present a novel algorithm that maps the target to the domain named *source fiction*. The *source fiction* domain differs from the original source, but the source classifier f_θ correctly classified it. Simultaneously, mapping from the target to the *source fiction* by optimal transport with Euclidean cost maintains the class-wise structure.

To generate the *source fiction*, we propose to use the cyclical monotone transformation of the target domain. Because if we transform the target using a cyclical monotone map, then the corresponding sample in the *source fiction* will have the same label.

As we proved, adversarial attacks are c-cyclical monotone transformations over the dataset 3.2. With minor changes, the adversarial attack can turn any sample into an accurately classified one in the same way as it fools the classifier (Huang et al., 2021; Mao et al., 2021). To generate correctly classified target samples, we use the inverted FSGD adversarial attack equation 1 on the source classifier f_θ, with the target label equal to the true class of the given target sample. Such an attack adds to the image features of the class it really belongs to (Ilyas et al., 2019).

By inverse adversarial attack we obtain a new domain $\mathbb{P}'$ with samples $X \subset \mathbb{R}^D$. Following the corollary 3.2, to obtain cyclical monotonicity, we set the size of the perturbation $\epsilon \leq \frac{1}{2} \min_{n_1,n_2} \|\mathbf{x}_{n_1} - \mathbf{x}_{n_2}\|$ for all $\mathbf{x}_n$ in target distribution. As a result, for each labeled sample $\mathbf{x}_n$ in the target distribution, we receive a corresponding sample $\mathbf{x}'_n$. While the new domain $\mathbb{P}'$ is a cyclical monotone transformation of the $\mathbb{P}$, we have a low quadratic cost between each target sample $\mathbf{x}_n$ and its corresponding sample $\mathbf{x}'_n$ in the new domain. As a result, we can apply optimal transport T_γ to approximate this transformation and save a class-wise structure.

So, instead of using the source domain, **we create a new domain using the gradients of the source classifier over the labeled target samples.** To use an inverse adversarial attack, some labeled samples must be in the target domain. We can therefore describe our method as semi-supervised.

Finally, we apply the map T_γ to all unlabeled target samples to adapt a source classifier f_θ to the target data or train a new classifier f_ϕ on the labeled target domain samples. In Figure 2 and Algorithm 1, we displayed the pipeline for the domain adaptation using the source fiction $\mathbb{P}'$ domain.

5 Related Work

5.1 Adversarial Attacks

Previously, the various properties of adversarial examples were studied (Petrov & Hospedales, 2019; Papernot et al., 2016; Ilyas et al., 2019). Applications of adversarial examples for model accuracy improvements were also proposed (Xie et al., 2019; Yang et al., 2020). The connection between optimal transport and adversarial examples was studied in the context of robustness problems (Pydi & Jog, 2020; Bouniot et al., 2021; Song et al., 2018).

Formerly, the connection between optimal transport and adversarial examples was studied in the context of robustness problems (Pydi & Jog, 2020; Bouniot et al., 2021). The authors (Wong et al., 2019) suggested the Wasserstein

Method	M / S	S / M	M / U	M / MM
EMD	21.2 ±3	68.7±3	79.2±2	56.1±3
EMD(*sf*)	**23.0±3**	**86.3±3**	**83.1±2**	**62.7±2**
OTLin	21.8±4	69.9±4	84.1±7	62.3±1
OTLin(*sf*)	**25.5±4**	**88.4±4**	**89.3±6**	**64.5±3**
Sinkh	21.8±4	68.8±2	82.1±7	55.7±12
Sinkh(*sf*)	**25.5±4**	**86.2±4**	**83.8±6**	**62.9±4**
SinkhLp	21.8±4	68.8±6	84.8±16	55.7±19
SinkhLp(*sf*)	**25.5±4**	**86.3±7**	**88.3±19**	**63.0±27**
SinkhL2	21.8±4	68.8±4	84.8±2	55.7±4
SinkhL2(*sf*)	**25.5±4**	**86.3±2**	**88.3±2**	**63.0±2**

(a) Digits dataset domains.

Method	A / S	S/ A	A / W	W / A	S / W	W/ S
EMD	38.4±3	9.3±5	45.2±3	45.6±5	13.6±3	36.7±3
EMD(*sf*)	**56.8±3**	**29.7±4**	**64.9±3**	**73.9±4**	**40.1±3**	**60.1±2**
OTLin	37.1±3	11.0±3	38.7±3	47.5±3	6.2±3	39.6±4
OTLin(*sf*)	**58.5±3**	**29.8±3**	**65.2±3**	**74.4±3**	**40.1±3**	**63.1±5**
Sinkh	38.0±3	10.1±4	44.7±6	45.5±3	13.1±7	37.2±3
Sinkh(*sf*)	**57.0±3**	**31.0±4**	**65.2±7**	**73.9±3**	**39.9±4**	**60.0±2**
SinkhLp	38.1±6	10.4±8	45.2±7	45.3±5	13.1±7	37.2±3
SinkhLp(*sf*)	**57.2±6**	**31.0±11**	**65.2±8**	**74.0±5**	**40.1±5**	**60.1±4**
SinkhL2	38.1±4	10.4±7	45.0±4	45.3±6	13.1±6	37.2±3
SinkhL2(*sf*)	**57.2±4**	**31.0±7**	**65.2±4**	**74.0±6**	**40.1±5**	**60.1±4**

(b) Modern Office-31 dataset domains.

Table 1: Accuracy↑ of the different optimal transport-based domain adaptation algorithms in the latent space of ResNet50 model

adversarial attack with Sinkhorn iterations. This algorithm allows one to find adversarial perturbations with respect to the Wasserstein ball. To the best of our knowledge, we propose the first method that connects adversarial attacks and optimal transport for domain adaptation.

5.2 Domain Adaptation

Another method of preserving the class-wise structure during mapping with optimal transport is to use a cost function c that is suitable for the problem at hand. The problem is that the cost is often unknown. Inverse optimal transport (Li et al., 2019; Liu et al., 2019) algorithms were proposed as a solution to this issue. It was shown that the cost function that preserves the underlying data structure during mapping can be reconstructed using the given mapping between data distributions.

For example, it was shown that the cost function could be approximated by the neural network (Liu et al., 2019). To train the cost function, it is necessary to solve the transport problem using the Sinkhorn algorithm (Cuturi, 2013) at every optimization step (Liu et al., 2019). Inverse optimal transport methods are hardly scalable and have not yet been applied to solve domain adaptation problems.

The connection between optimal transport and deep networks was proposed for unsupervised DA (Damodaran et al., 2018) and transfer learning (Li et al., 2020). In domain adaptation with label and target shift problems, optimal transport methods align probability distributions between a few domains (Redko et al., 2019; Rakotomamonjy et al., 2020).

We suggest a new problem setting for domain adaptation with optimal transport in contrast to all of these approaches. In general, all varieties of optimal transport solvers can be connected to our method. In our evaluations, we demonstrate that our method enhances the performance of a number of transport solvers.

6 Experiments

In this section, we test our method on two types of datasets (§6.1). The goal of our experiments is to demonstrate that our method improves the performance of fundamental discrete optimal transport algorithms. Besides, we compare different deep domain adaptation baselines (§6.2). A discussion on empirical complexity is presented in (§6.4).

Additionally, we conduct an ablation study on the ϵ parameter to show the stability of our method in different settings of the inverse adversarial attack(§6.5). The code is written in *PyTorch* framework and will be made public. We give more experiments with different backbone networks and additional domains in Appendix.

6.1 Datasets

Digits: We evaluated our method on Digits datasets MNIST (LeCun & Cortes, 2010), USPS (Hull, 1994), SVNH (Netzer et al., 2011), and MNIST-M (Ganin & Lempitsky, 2015). Each dataset consists of 10 classes of digit images with different numbers of train and test samples. As a pre-processing step, we resized the images to the (32×32) pixel size. No augmentation was used. **Modern Office-31:** Besides the Digits dataset that consists of only ten classes in each domain, we tested our method on the Modern Office-31 dataset (Ringwald & Stiefelhagen, 2021) with 31 classes per domain. The Modern Office-31 dataset is one of the most extensive and diverse datasets for DA. The dataset consists of three domains: Amazon (A), Synthetic (S), and Webcam (W). In comparison to the Digits and original Office-31 dataset (Saenko et al., 2010), this dataset includes synthetic $\rightarrow$ real tasks, which is problematic. No augmentation was used. Additionally, we included the DLSR (D) domain from the original Office-31 to estimate our algorithm properly, see the results in Appendix

6.2 BASELINES

Discrete Optimal Transport: We tested several optimal transport solvers in semi-supervised domain adaptation settings: EMD (Courty et al., 2015), Sinkhorn (Cuturi, 2013) (Sinkh in figures) (Cuturi, 2013), SinkhornL2, SinkhornLp, and OTLin (Perrot et al., 2016).

Most of these algorithms are presented in the POT framework (Flamary et al., 2021), which provides state-of-the-art optimal transport solvers for domain adaptation. For experiments, we used quadratic cost $c(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2$ for each algorithm. Regularization size equals 4.0 for Sinkhorn, SinkhLp, SinkhL2, and OTLin; all other hyperparameters are equal to the default, presented in POT.

6.3 SETTINGS

For the source domain classifier, we trained ResNet50 (He et al., 2016) to achieve $90+$ accuracy on the test set of each domain in Digits and Modern Office-31. The classifier was trained using Adam (Ruder, 2016) optimizer with a $1e-3$ learning rate. The size of latent space before the output layer was equal to 2048.

After training, **we applied domain adaptation by moving mass in the latent space of the source classifier**. For discrete optimal transport baselines, the available labels were used as a penalty to the transport plan by building a cost matrix M with $M(i, j) = 0$ when $\mathbf{x}_i$ and $\mathbf{y}_j$ labels are equal and $+\infty$ if not (Courty et al., 2016; Yan et al., 2018)

To create a *source fiction*, we used 50 steps FSGD with ϵ equal to 0.45. We found that this value allows us to achieve strong perturbations and, at the same time, satisfies the proposed bound on all domains. The results with 10 labeled samples in target domain are presented in Table 1a, 1b. **All values in the tables are averaged over the 10 runs with randomly chosen sets of labeled samples in the target domain.** By **bold** we denote highest accuracy in all tables. By *(sf)* we denote a *source fiction*.

6.4 RESULTS

Discussion Our method demonstrates improvement for all adaptation tasks. The simplest EMD method is less accurate than other methods, and OTLin accuracy is slightly higher for all domains. The Sinkhorn algorithm with group lasso and Laplacian regularizations did not provide notable improvements over standard Sinkhorn, see Tables 1a and 1b. The results with only three known labels in class are presented in Appendix.

In our settings, optimal transport can find a map between a target and a source and, at the same time, save discriminability, i.e., class-wise structure. In our pipeline, optimal transport maps the unlabeled samples to the perturbed samples from the same class in the *source fiction*.

In practice, discrete optimal transport techniques are susceptible to regularization terms (Courty et al., 2015; Dessein et al., 2018; Paty & Cuturi, 2020) and require special scaling (Meng et al., 2021), but our method demonstrates stability in all domains and tested solvers. We present the adaptation results in the appendix with only three labels per class.

Empirical Complexity Our method improves the accuracy of discrete optimal transport methods. It is a significant accomplishment because discrete optimal transport methods are simple and fast technique for domain adaptation in comparison to deep neural network-based approaches. Often, deep domain adaptation methods typically solve the

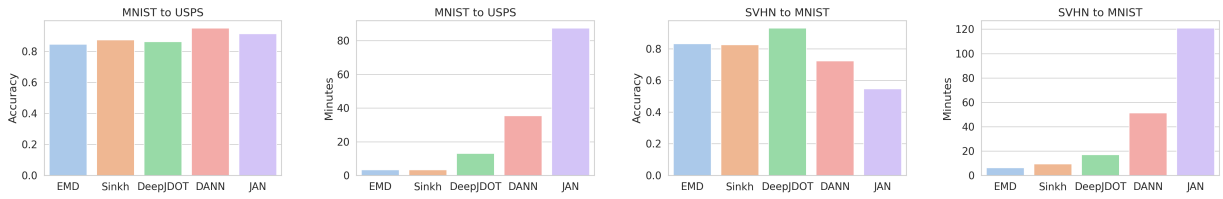


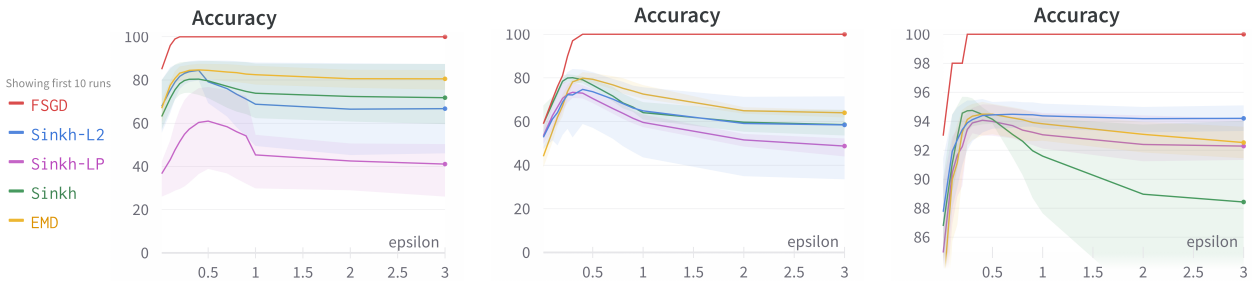
Figure 3: The amount of time (in minutes) required to achieve high accuracy for various DA methods. Our method was used in conjunction with the EMD and Sinkh methods.

difficult min-max optimization problem. (Ganin & Lempitsky, 2015; Long et al., 2018; 2015; 2017; Gretton et al., 2012), which require hours of training.

In practice, discrete optimal transport methods like EMD, Sinkhorn, and OTLin take a few minutes to solve domain adaptation on GPU GeForce GTX-1080 (12 GB). The creation of the source fiction domain also takes less than a minute. Total time required for domain adaptation using our method and DANN (Ganin & Lempitsky, 2015), DeepJDOT (Courty

et al., 2017) and JAN (Long et al., 2017) methods is presented in Figure 3. In the appendix, we present the additional comparison with deep adaptation methods, see Tables 6 5.

6.5 Ablation Study on Perturbation Size



(a) MNIST → MMNIST (b) SVHN → MNIST (c) USPS → MNIST

Figure 4: Results of ablation on ϵ parameter for MNIST → MNIST-M (left) and SVHN → MNIST (right) datasets. FSGD denotes how accurately the source domain model classifies *source fiction* samples obtained with the corresponding parameter ϵ. The accuracy is higher than in 1a because, for these experiments, we used 100 labeled samples in the target domain classes.

In this section, we show the FSGD algorithm adaptation results with various ϵ values. We evaluated several transportation tasks: USPS → MNIST, MNIST → MNIST-M, and SVHN → MNIST.

For each task, we tested different values of ϵ (from 0.01 to 3) to create a *source fiction* samples and then fit optimal transport to find a map between target and constructed *source fiction*. With larger perturbations, the prediction of the classifier on perturbed samples becomes more accurate (see FSGD(red) curve in Figure 4).

Our results show that our method is not significantly sensitive to the size of perturbation; see Figure 4. The adaptation achieves the highest accuracy with a small value of ϵ. With perturbations a bit larger than the ϵ bound, most of the samples in *source fiction* are still c-cyclical monotone to the target, and the method still works.

The bound value of $\frac{1}{2}\min_{n_1,n_2}\|x_{n_1} - x_{n_2}\|$ is different for various datasets. For SVHN, this value is 0.74; for MNIST, it is 0.29; for USPS, it is equal to 0.85. These values were computed in the latent space of the ResNet50 classifier trained on the corresponding domain. When the ϵ value becomes larger than $\frac{1}{2}\min$ distance between samples, the accuracy of adaptation decreases because the transformation becomes less cyclical monotone.

7 Conclusion

We demonstrated that adversarial attacks are optimal transport maps over datasets and proposed an algorithm that modifies domain adaptation settings with optimal transport to bring target data closer to the perturbed target sample.

We conducted various experiments on multiple datasets and showed that our method improves various optimal transport baselines. Adaptation with source fiction improved accuracy by more than 10% in some domains for discrete optimal transport methods. Our method has a wide range of straightforward applications. For example, we plan to adapt our method to neural transport solvers in order to make it resistant to out-of-sample estimation. While optimal transport can solve domain adaptation problems with target shift and unbalanced classes (Redko et al., 2019; Rakotomamonjy et al., 2020), using source fiction is promising.

The main limitation of our approach is that it is necessary to have access to the labels in the target domain. To avoid the limitation of labels availability, we plan to use self-labeled techniques (Triguero et al., 2015). We expect our research to contribute to the development of less complicated domain adaptation techniques and open doors for the future application of cyclical monotonicity of adversarial attacks.

References

Brandon Amos, Lei Xu, and J. Zico Kolter. Input convex neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia,*

6-11 August 2017, volume 70 of *Proceedings of Machine Learning Research*, pp. 146–155. PMLR, 2017. URL http://proceedings.mlr.press/v70/amos17b.html.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Arip Asadulaev, Alexander Korotin, Vage Egiazarian, and Evgeny Burnaev. Neural optimal transport with general cost functionals. *arXiv preprint arXiv:2205.15403*, 2022.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175, 2010a. doi: 10.1007/s10994-009-5152-4. URL https://doi.org/10.1007/s10994-009-5152-4.

Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In Yee Whye Teh and D. Mike Titterington (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, volume 9 of *JMLR Proceedings*, pp. 129–136. JMLR.org, 2010b. URL http://proceedings.mlr.press/v9/david10a.html.

Quentin Bouniot, Romaric Audigier, and Angelique Loesch. Optimal transport as a defense against adversarial attacks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 5044–5051. IEEE, 2021.

Pierre Colombo, Guillaume Staerman, Chloe Clavel, and Pablo Piantanida. Automatic text evaluation through the lens of wasserstein barycenters. *arXiv preprint arXiv:2108.12463*, 2021.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *CoRR*, abs/1507.00504, 2015. URL http://arxiv.org/abs/1507.00504.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 3730–3739, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/0070d23b06b1486a538c0eaa45dd167a-Abstract.html.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 2292–2300, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 447–463, 2018.

Arnaud Dessein, Nicolas Papadakis, and Jean-Luc Rouas. Regularized optimal transport and the rot mover's distance. *The Journal of Machine Learning Research*, 19(1):590–642, 2018.

Jiaojiao Fan, Amirhossein Taghvaei, and Yongxin Chen. Scalable computations of wasserstein barycenter via input convex neural networks. *CoRR*, abs/2007.04462, 2020. URL https://arxiv.org/abs/2007.04462.

Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Zahdi Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In Francis R. Bach and David M. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 1180–1189. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/ganin15.html.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A pac-bayesian approach for domain adaptation with specialization to linear classifiers. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 738–746. JMLR.org, 2013. URL http://proceedings.mlr.press/v28/germain13.html.

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

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012. URL http://dl.acm.org/citation.cfm?id=2188410.

Claus Griessler. cyclical monotonicity as a sufficient criterion for optimality in the multimarginal monge–kantorovich problem. *Proceedings of the American Mathematical Society*, 146(11):4735–4740, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. *arXiv preprint arXiv:2101.04898*, 2021.

J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994. doi: 10.1109/34.291440.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *CoRR*, abs/1905.02175, 2019. URL http://arxiv.org/abs/1905.02175.

Alexander Korotin, Vage Egiazarian, Arip Asadulaev, and Evgeny Burnaev. Wasserstein-2 generative networks. *CoRR*, abs/1909.13082, 2019. URL http://arxiv.org/abs/1909.13082.

Alexander Korotin, Lingxiao Li, Aude Genevay, Justin Solomon, Alexander Filippov, and Evgeny Burnaev. Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark. *arXiv preprint arXiv:2106.01954*, 2021a.

Alexander Korotin, Lingxiao Li, Justin Solomon, and Evgeny Burnaev. Continuous wasserstein-2 barycenter estimation without minimax optimization. *CoRR*, abs/2102.01752, 2021b. URL https://arxiv.org/abs/2102.01752.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

John Lee, Max Dabagia, Eva L Dyer, and Christopher J Rozell. Hierarchical optimal transport for multimodal distribution alignment. *arXiv preprint arXiv:1906.11768*, 2019.

Ruilin Li, Xiaojing Ye, Haomin Zhou, and Hongyuan Zha. Learning to match via inverse optimal transport. *Journal of machine learning research*, 20, 2019.

Xuhong Li, Yves Grandvalet, Rémi Flamary, Nicolas Courty, and Dejing Dou. Representation transfer by optimal transport. *arXiv preprint arXiv:2007.06737*, 2020.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6028–6039. PMLR, 2020. URL http://proceedings.mlr.press/v119/liang20a.html.

Ruishan Liu, Akshay Balsubramani, and James Zou. Learning transport cost from subset correspondence. *arXiv preprint arXiv:1909.13203*, 2019.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In Francis R. Bach and David M. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 97–105. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/long15.html.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2208–2217. PMLR, 2017. URL http://proceedings.mlr.press/v70/long17a.html.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Conditional adversarial domain adaptation. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 1647–1657, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/ab88b15733f543179858600245108dd8-Abstract.html.

Ashok Vardhan Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason D. Lee. Optimal transport mapping via input convex neural networks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6672–6681. PMLR, 2020. URL http://proceedings.mlr.press/v119/makkuva20a.html.

Chengzhi Mao, Mia Chiquier, Hao Wang, Junfeng Yang, and Carl Vondrick. Adversarial attacks are reversible with natural supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 661–671, 2021.

Cheng Meng, Yuan Ke, Jingyi Zhang, Mengrui Zhang, Wenxuan Zhong, and Ping Ma. Large-scale optimal transport map estimation using projection pursuit. *arXiv preprint arXiv:2106.05838*, 2021.

Youssef Mroueh. Wasserstein style transfer. *arXiv preprint arXiv:1905.12828*, 2019.

John C Nash. The (dantzig) simplex method for linear programming. *Computing in Science & Engineering*, 2(1):29–31, 2000.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Quan Hoang Nhan Dam, Trung Le, Tu Dinh Nguyen, Hung Bui, and Dinh Phung. Threeplayer wasserstein gan via amortised duality. In *Proc. of the 28th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 2019.

Nicolas Papernot, Patrick D. McDaniel, and Ian J. Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *CoRR*, abs/1605.07277, 2016. URL http://arxiv.org/abs/1605.07277.

Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, AsiaCCS 2017, Abu Dhabi, United Arab Emirates, April 2-6, 2017*, pp. 506–519, 2017. doi: 10.1145/3052973.3053009. URL https://doi.org/10.1145/3052973.3053009.

François-Pierre Paty and Marco Cuturi. Regularized optimal transport is ground cost adversarial. In *International Conference on Machine Learning*, pp. 7532–7542. PMLR, 2020.

Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4197–4205, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/26f5bd4aa64fdadf96152ca6e6408068-Abstract.html.

Deyan Petrov and Timothy M. Hospedales. Measuring the transferability of adversarial examples. *CoRR*, abs/1907.06291, 2019. URL http://arxiv.org/abs/1907.06291.

Muni Sreenivas Pydi and Varun Jog. Adversarial risk via optimal transport and optimal couplings. In *International Conference on Machine Learning*, pp. 7814–7823. PMLR, 2020.

Alain Rakotomamonjy, Rémi Flamary, Gilles Gasso, Mokhtar Z Alaya, Maxime Berar, and Nicolas Courty. Optimal transport for conditional domain matching and label shift. *arXiv preprint arXiv:2006.08161*, 2020.

Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical analysis of domain adaptation with optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 737–753. Springer, 2017.

Ievgen Redko, Nicolas Courty, Rémi Flamary, and Devis Tuia. Optimal transport for multi-source domain adaptation under target shift. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 849–858. PMLR, 2019.

Sebastian Reich. A nonparametric ensemble transform method for bayesian inference. *SIAM Journal on Scientific Computing*, 35(4):A2013–A2024, 2013.

Tobias Ringwald and Rainer Stiefelhagen. Adaptiope: A modern benchmark for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 101–110, 2021.

Ralph Rockafellar. Characterization of the subdifferentials of convex functions. *Pacific Journal of Mathematics*, 17(3): 497–510, 1966.

Sebastian Ruder. An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747, 2016. URL http://arxiv.org/abs/1609.04747.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pp. 213–226. Springer, 2010.

Filippo Santambrogio. Optimal transport for applied mathematicians. calculus of variations, pdes and modeling. 2015. URL https://www.math.u-psud.fr/~filippo/OTAM-cvgmt.pdf.

Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on MNIST. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. URL https://openreview.net/forum?id=S1EHOsC9tX.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):1–11, 2015.

Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Improving the generalization of adversarial training with domain adaptation. *arXiv preprint arXiv:1810.00740*, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6199.

Amirhossein Taghvaei and Amin Jalali. 2-wasserstein approximation via restricted convex potentials with application to improved training for gans. *CoRR*, abs/1902.07197, 2019. URL http://arxiv.org/abs/1902.07197.

Anne-Marie Tousch and Christophe Renaudin. (yet) another domain adaptation library, 2020. URL https://github.com/criteo-research/pytorch-ada.

Isaac Triguero, Salvador García, and Francisco Herrera. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information systems*, 42(2):245–284, 2015.

C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. ISBN 9783540710509. URL https://books.google.ru/books?id=hV8o5R7_5tkC.

Eric Wong, Frank R. Schmidt, and J. Zico Kolter. Wasserstein adversarial examples via projected sinkhorn iterations. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 6808–6817, 2019. URL http://proceedings.mlr.press/v97/wong19a.html.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1369–1378, 2017.

Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L. Yuille, and Quoc V. Le. Adversarial examples improve image recognition. *CoRR*, abs/1911.09665, 2019. URL http://arxiv.org/abs/1911.09665.

Yuguang Yan, Wen Li, Hanrui Wu, Huaqing Min, Mingkui Tan, and Qingyao Wu. Semi-supervised optimal transport for heterogeneous domain adaptation. In *IJCAI*, volume 7, pp. 2969–2975, 2018.

Jihan Yang, Ruijia Xu, Ruiyu Li, Xiaojuan Qi, Xiaoyong Shen, Guanbin Li, and Liang Lin. An adversarial perturbation oriented domain adaptation approach for semantic segmentation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 12613–12620. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6952.

Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE Trans. Neural Netw. Learning Syst.*, 30(9):2805–2824, 2019. doi: 10.1109/TNNLS.2018.2886017. URL https://doi.org/10.1109/TNNLS.2018.2886017.

A Appendix

A.1 Additional background on OT

OT aims at finding a solution to transfer mass from one distribution to another with the least effort. Monge's problem was the first example of the OT problem and can be formally expressed as follows:

$$\inf_{T\#\mathbb{P}=\mathbb{Q}} \int_{\Omega_{\mathbb{P}}} c(\mathbf{x}, T(\mathbf{x}))\mathbb{P}(\mathbf{x})d\mathbf{x} \tag{12}$$

The Monge's formulation of OT aims at finding a mapping $T : \Omega_{\mathbb{P}} \to \Omega_{\mathbb{Q}}$ of the two probability measures $\mathbb{P}$ and $\mathbb{Q}$ and a cost function $c : \Omega_{\mathbb{P}} \times \Omega_{\mathbb{Q}} \to \mathbb{R}_+$, where $T\#\mathbb{P}_s = \mathbb{Q}_t$ represents the mass preserving push forward operator. In Monge's formulation, T cannot split the mass from a single point. The problem is that the mapping T may not even exist with such constraints.

To avoid this, Kantorovitch proposed a relaxation (Villani, 2008). Instead of obtaining a mapping, the goal is to seek a joint distribution over the source and the target that determines how the mass is allocated. For a given cost function $c : \Omega_{\mathbb{P}} \times \Omega_{\mathbb{Q}} \to \mathbb{R}_+$, the primal Kantorovitch formulation can be expressed as the following problem:

$$\min_{\gamma\in\gamma(\mathbb{P},\mathbb{Q})} \left\{ \int_{\Omega_{\mathbb{P}}\times\Omega_{\mathbb{Q}}} c(\mathbf{x},\mathbf{y})d\gamma(\mathbf{x},\mathbf{y}) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\gamma}[c(\mathbf{x},\mathbf{y})] \right\} \tag{13}$$

In primal Kantorovitch formulation, we look for a joint distribution γ with $\mathbb{P}$ and $\mathbb{P}$ as marginals that minimize the expected transportation cost. If the independent distribution $\gamma(\mathbf{x},\mathbf{y}) = \mathbb{P}(\mathbf{x})\mathbb{Q}(\mathbf{y})$ respects the constraints, linear program is convex and always has a solution for a semi-continuous c:

$$\gamma \in P(\Omega_{\mathbb{P}}, \Omega_{\mathbb{Q}}) : \int \gamma(\mathbf{x},\mathbf{y})d\mathbf{y} = \mathbb{P}(\mathbf{x}), \int \gamma(\mathbf{x},\mathbf{y})d\mathbf{x} = \mathbb{Q}(\mathbf{y}) \tag{14}$$

The primal Kantorovitch formulation can also be presented in dual form as stated by the Rockafellar—Fenchel theorem (Villani, 2008):

$$\max_{\phi\in C(\Omega_{\mathbb{P}}),\psi\in C(\Omega_{\mathbb{Q}})} \left\{ \int \phi d\mathbb{P} + \int \psi d\mathbb{Q} \mid \phi(\mathbf{x}) + \psi(\mathbf{y}) \leq c(\mathbf{x},\mathbf{y}) \right\} \tag{15}$$

After finding a solution to the transport problem, OT measures dissimilarity between the two distributions. This similarity is also called the Wasserstein distance (Villani, 2008):

$$W_p\left(\mathbb{P},\mathbb{Q}\right) = \min_{\gamma\in\gamma(\mathbb{P},\mathbb{Q})} \left\{ \int_{\Omega_{\mathbb{P}}\times\Omega_{\mathbb{Q}}} c(\mathbf{x},\mathbf{y})d\gamma(\mathbf{x},\mathbf{y}) \right\}^{\frac{1}{p}} \tag{16}$$

where $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^p$ and $p > 1$. The Wasserstein distance encodes the geometry of the space through the optimization problem and can be used on any distribution of mass.

Recently, there has been a solid push to incorporate Input Convex Neural Networks (ICNNs) (Amos et al., 2017) in OT problems. According to Rockafellar's Theorem (Rockafellar, 1966), every cyclically monotone mapping g is contained in a sub-gradient of some convex function $f : X \to \mathbb{R}$. Furthermore, according to Brenier's Theorem (Theorem 1.22 of (Santambrogio, 2015)), these gradients uniquely solve the Monge problem equation 2. Following these theorems, a range of approaches explored ICNNs as parameterized convex potentials in dual Kantorovich problem (Taghvaei & Jalali, 2019; Makkuva et al., 2020).

Further development of this approach enabled the construction of the non-minimax Wasserstein-2 generative framework (Korotin et al., 2019) that can solve DA and Wasserstein-2 Barycenters estimation (Fan et al., 2020; Korotin et al., 2021b). Compared to discrete OT, neural methods provide generalized OT methods that can ensure out-of-sample estimates.

A.2 Additional experiments

In this section, we provide additional experiments using the *source fiction* domain for discrete OT solvers. In tables, **bold** denotes the results of discrete solvers with *source fiction* if this improves its accuracy compared to the standard settings.

We considered the range of fundamental gradient-based domain adaptation techniques. First of all, we compared our method to the prominent adversarial-based approach DANN (Ganin & Lempitsky, 2015), CDAN, CDAN-E (Long et al., 2018). We also tested the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) based domain adaptation techniques like DAN (Long et al., 2015) and JAN (Long et al., 2017). Additionally, we considered the Wasserstein distance-based method WDGRL (Shen et al., 2018). We used implementation for these methods proposed in ADA framework (Tousch & Renaudin, 2020). Digits dataset results with 3 (Table 2) known labels per class in target domain using ResNet50 classifier.

Following the benchmark results of the neural optimal transport algorithms benchmark (Korotin et al., 2021a), we choose the MM:R (Nhan Dam et al., 2019; Korotin et al., 2021a) method to apply domain adaptation.

We used Feed forward networks with three hidden layers $[64, 64, 32]$ as potential ϕ and transport map T in for MM:R neural optimal transport methods. Potentials are trained using Adam (Ruder, 2016) optimizer with lr equal to 1e-4. In total, generator was trained 300 epochs with Adam optimizer and lr equal to $1e-3$.

Method	MNIST SVHN	SVHN MNIST	MNIST USPS	USPS MNIST	MNIST M-MNIST
EMD	21.3	72.5	66.1	67.8	44.5
EMD(*sf*)	**23.1**	**83.5**	**82.6**	**86.5**	**54.7**
OTLin	21.8	73.4	67.4	68.8	45.0
OTLin(*sf*)	**23.9**	**85.3**	**86.3**	**86.9**	**55.1**
Sinkh	21.7	73.0	67.3	68.7	44.6
Sinkh(*sf*)	**23.7**	**85.0**	**82.6**	**86.8**	**54.8**
SinkhLp	21.7	73.4	67.3	68.8	45.0
SinkhLp(*sf*)	**23.7**	**85.2**	**86.3**	**86.9**	**54.9**
SinkhL2	21.7	73.4	67.3	68.8	45.0
SinkhL(*sf*)	**23.8**	**85.2**	**86.3**	**86.9**	**54.9**

Table 2: Accuracy of domain adaptation by optimal transport in the latent space of ResNet50 model with only 3 known labels for each class in the target domain on Digits datasets. The top part of the table represents semi-supervised settings for discrete OT methods, settings the bottom part presents results using *source fiction*.

Modern-Office dataset results with additional domain DLSR (D) (Table 3), with 10 known labels.

Additionally we tested our algorithm on the complicated CIFAR10-STL10 adaptation task. See results with ResNet18 source classifier in (Table 5).

Method	A D	D A	A S	S A	A W	W A	D S	S D	D W	W D	S W	W S
EMD	50.7	46.2	38.4	9.3	45.2	45.6	32.7	16.4	62.6	67.1	13.6	36.7
EMD(*sf*)	**70.9**	**72.5**	**56.8**	**29.7**	**64.9**	**73.9**	**56.6**	**47.3**	**75.7**	**75.1**	**40.1**	**60.1**
OTLin	45.8	48.0	37.1	11.0	38.7	47.5	36.5	4.1	60.9	61.8	6.2	39.6
OTLin(*sf*)	**71.3**	**73.6**	**58.5**	**29.8**	**65.2**	**74.4**	**59.8**	**47.3**	**76.6**	**75.1**	**40.1**	**63.1**
Sinkh	51.1	46.3	38.0	10.1	44.7	45.5	32.9	16.5	63.5	67.1	13.1	37.2
Sinkh(*sf*)	**70.6**	**72.7**	**57.0**	**31.0**	**65.2**	**73.9**	**56.7**	**47.3**	**77.4**	**74.8**	**39.9**	**60.0**
SinkhLp	51.1	46.7	38.1	10.4	45.2	45.3	33.0	16.5	63.8	68.3	13.1	37.2
SinkhLp(*sf*)	**70.6**	**72.8**	**57.2**	**31.0**	**65.2**	**74.0**	**56.8**	**47.3**	**77.4**	**75.5**	**40.1**	**60.1**
SinkhL2	51.1	46.7	38.1	10.4	45.0	45.3	33.0	16.5	63.8	68.3	13.1	37.2
SinkhL2(*sf*)	**70.6**	**72.8**	**57.2**	**31.0**	**65.2**	**74.0**	**56.8**	**47.3**	**77.4**	**75.5**	**40.1**	**60.1**

Table 3: Results of domain adaptation in the latent space of ResNet50 classifier on the Modern Office-31 dataset with the the additional DLSA(D) domain. 10 labels are known for each class in the target domain.

Method	MNIST SVHN	SVHN MNIST	MNIST USPS	MNIST M-MNIST
DANN	19.5	61.7	93.8	37.5
C-DAN	11.5	79.0	90.7	68.4
DAN	16.7	54.8	**95.0**	47.0
JAN	11.5	57.9	89.5	52.9
WDGRL	13.8	59.5	85.7	52.0
MM:R	20.3	**80.2**	78.0	63.8
MM:R(*sf*)	**21.5**	77.1	79.0	**70.3**

Table 4: Accuracy↑ of the deep DA methods and OT based neural method MM:R in the latent space of source classifier on Digits datasets with 10 known labels. We connected our source fiction method with the MM:R in this experiments

Method	CIFAR→STL	SLT → CIFAR
EMD	48.0	75.0
EMD(*sf*)	**51.0**	**76.2**
OTlin	48.1	74.1
OTlin(*sf*)	**51.0**	**76.2**
Sinkh	48.1	74.1
Sinkh(*sf*)	**51.0**	**76.2**
SinkhLp	48.0	74.1
SinkhLp(*sf*)	**51.0**	**76.2**
SinkhL2	48.0	74.1
SinkhL2(*sf*)	**51.0**	**76.2**

Table 5: Accuracy↑ on domain adaptation between CIFAR-10 and STL datasets with 10 known labels.

Method	Accuracy	Time (Minutes)
EMD(*sf*)	0.846	3.33
Sinkh(*sf*)	0.874	3.38
DeepJDOT	0.862	13.0
DANN	0.950	35.4
JAN	0.913	87.5

(a) MNIST to USPS.

Method	Accuracy	Time (Minutes)
EMD(*sf*)	0.83	6.23
Sinkh(*sf*)	0.82	9.41
DeepJDOT	0.93	17.08
DANN	0.72	51.31
JAN	0.54	120.98

(b) SVHN to MNIST.

Table 6: Accuracy↑ and required time for the different optimal transport-based domain adaptation algorithms