DISTRIBUTIONALLY ROBUST DOMAIN ADAPTATION VIA OPTIMAL TRANSPORT

Akram S. Awad[1] *Shuchin Aeron*[2] *George K. Atia*[1,3]

[1] Department of Electrical and Computer Engineering, University of Central Florida, Orlando FL, USA
[2] Department of Electrical and Computer Engineering, Tufts University, Medford, Massachusetts
[3] Department of Computer Science, University of Central Florida, Orlando FL, USA
Email: {akram.awad, george.atia}@ucf.edu, shuchin@ece.tufts.edu

2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP) | 979-8-3503-7225-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/MLSP58920.2024.10734802

ABSTRACT

Distributionally Robust Optimization (DRO) mitigates the effect of distributional uncertainty in supervised learning by optimizing over an uncertainty ball of distributions, typically centered around the empirical distribution of the training sample. In this paper, we consider DRO for the problem of Unsupervised Domain Adaptation (UDA). In classical UDA, the goal is to adapt a model trained on a labeled source domain to a new, unlabeled target domain. Modifying classical DRO to UDA settings by enlarging the uncertainty radius around the source to include the target can lead to excessive regularization. To mitigate this, we propose to utilize Optimal Transport (OT) to transport the source domain to a vicinity of the target and construct the DRO problem around the transported samples, thereby ensuring a small uncertainty radius in DRO with high likelihood of including the true target. Our numerical experiments validate the superiority of our method over existing robust approaches.

Index Terms— Domain Adaptation, Optimal Transport, Distributional Robustness, Classification.

1. INTRODUCTION

Supervised machine learning approaches aim to estimate decision variables that minimize the population risk within some hypothesis space. Unfortunately, they often display poor out-of-sample performance given that the risk function is typically only accessible through a limited training sample, leading to an approximation of the population risk via its empirical average. Moreover, the training sample may not adequately represent the true underlying distribution, for example, the training data may be biased [1]. As a result, the true population risk may be underestimated, and in turn the decision variable may not generalize well on unseen samples. This phenomenon is referred to as the optimizer's curse [2, 3].

Distributionally Robust Optimization (DRO) has emerged as a framework to address challenges in optimization and decision-making under uncertainty [3–12]. It is seen as a potential remedy for the optimizer's curse and ensuing overfitting issues. DRO is concerned with finding the decision function that minimizes the worst-case expected loss by optimizing over an ambiguity (uncertainty) set of probability distributions.

The uncertainty sets in DRO can be constructed in different ways. One approach involves moment-based uncertainty sets, which comprise distributions that satisfy specific moment constraints [13]. However, sets that involve distributions sharing some low-order moments often fail to converge as the number of training samples approaches infinity [14]. To address this, another strategy is to construct a ball of distributions centered at the empirical distribution of the training sample using a discrepancy metric. This approach drives the radius to shrink to zero as the number of training samples increases. Various discrepancy metrics, such as the Kullback-Leibler divergence [6, 7], Maximum Mean discrepancy (MMD) [9, 10], and Wasserstein distance [11, 12], have been proposed for constructing such ambiguity sets.

In this paper, we consider DRO for the problem of Unsupervised Domain Adaptation (UDA). In UDA, one is given labeled data from a source domain and unlabeled data from a target domain. The UDA problem involves adapting the source domain decision function to this unlabeled target domain. This has been approached in various ways under different assumptions [15–22] within the realm of DA. However, these DA methods usually exhibit poor performance when tested on perturbed target data. In the DRO setup for DA, we aim to make this adaptation robust to perturbations around the target domain at test time.

However, extending the classical DRO approach with an ambiguity set centered around the empirical source distribution to include the target domain may result in an overly conservative (overly regularized) decision function if the target distribution deviates significantly from the source[1] [11]. This over-conservatism occurs because the resulting uncertainty set is too large, encompassing too many distributions that the

This work was supported by NSF under Award CCF-2106339 and DARPA under Agreement No. HR0011-24-9-0427.

[1] This is typically the case in domain adaption.

decision function must be robust against.

In this paper, we introduce a distributionally robust approach to effectively transfer knowledge from the source to the target domain, which ensures good generalization performance on unseen data from the target domain. We summarize our main contributions below.

Main contributions: As a methodological contribution, we introduce a novel distributionally robust approach for UDA termed Distributionally Robust Domain Adaptation via Optimal Transport (`DRDA-OT`). This approach aims to learn a robust and generalizable decision variable for the target domain through data transportation and optimization within an uncertainty set of distributions constructed around the transported data. As an empirical contribution, we provide evidence demonstrating the generalization and robustness of our decision variable on both synthetic and real-world datasets.

Our approach consists of two key stages. First, we employ Optimal Transport (OT) [23,24] techniques to estimate a mapping from the source to the target domain, effectively reducing the distributional mismatch while enabling the transfer of labeling information. Second, we utilize the transported data to derive a robust decision variable by optimizing within an uncertainty set of distributions defined with respect to the Wasserstein distance around the transported data. This allows us to construct an uncertainty set with a smaller radius than the one centered at the empirical source distribution.

1.1. Relevant Work

Previous studies have explored the application of DRO across diverse domains to enable knowledge transfer [25–27]. This line of work largely focuses on optimizing within an uncertainty set of conditional distributions, under some moment constraints. On the other hand, our approach focuses on optimizing within a set of joint distributions that maintain a certain distance from the empirical distribution of the transported data. In a recent study [28], a distributionally robust framework for UDA was proposed, leveraging the Joint Maximum Mean Discrepancy (JMMD) method [29]. Specifically, the authors introduce perturbations to the source data using an adversary defined with respect to a Wasserstein ball of distributions centered at the source data. More closely related is the work by [30], which proposes a DRO approach to DA. This method optimizes within an uncertainty set of distributions centered at a weighted empirical distribution. However, this approach has certain limitations. First, it relies on an importance weighting approach to determine the weighted center, assuming a common support between both domains, a requirement we do not impose in our method. Second, it assumes that the loss function is an element of the reproducing kernel Hilbert space (RKHS), which may be restrictive for many real-world applications. In contrast, our approach is applicable to any loss function and works for both regression and classification settings.

The rest of the paper is organized as follows. In Section 2, we review some preliminary background information. In Section 3, we introduce our approach, dubbed `DRDA-OT`. Section 4 presents numerical experiments on both synthetic and real data. We conclude in Section 5.

2. BACKGROUND

2.1. Notation

Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$ be the feature and label spaces, respectively. We define $\mathcal{P}$ to be the set of all probability measures defined on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. For any domain, e.g., $\mathcal{X}$, we use $\mathcal{P}(\mathcal{X})$ to denote the set of probability measures on $\mathcal{X}$.

Let $\zeta = (x, y) \in \mathcal{Z}$ denote the pair of data and label, respectively. We use $\hat{\mathbf{P}}_m(.) = \frac{1}{m}\sum_{j=1}^{m} \delta_{\zeta_j}(.)$ to represent the empirical distribution on the sample $\{\zeta_j\}_{j=1}^{m}$, where δ is the Dirac measure. We denote the marginal probability distribution on $\mathcal{X}$ by μ. The loss function associated with the decision variable θ is denoted by $l(\theta)$.

2.2. Unsupervised Domain Adaptation

Given *i.i.d.* m labeled source domain samples $D_s = \{\zeta_i^{(s)} = (x_i^{(s)}, y_i^{(s)})\}_{i=1}^{m}$ and n unlabeled target domain samples $D_t = \{x_i^{(t)}\}_{i=1}^{n}$ drawn form the source distribution $\mathbf{P_s}$ and target distribution $\mathbf{P_t}$, respectively, where $\mathbf{P_s}, \mathbf{P_t} \in \mathcal{P}(\mathcal{Z})$, the unsupervised domain adaptation (UDA) problem seeks to estimate θ such that the expected risk (loss) $\mathbb{E}^{\mathbf{P_t}}[l(\theta)]$ on the target domain is minimized for a given loss function l.

2.3. Optimal Transport

Let μ_s, μ_t be two distributions with finite second moments. Given a metric $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ and $p \geq 1$, *Monge's* formulation of the Optimal Transport (OT) problem is [31]

$$T_0 := \underset{T:T\#\mu_s=\mu_t}{\operatorname{argmin}} \mathbb{E}[c(x^{(s)}, T(x^{(s)}))^p], \tag{1}$$

where the notation $T_{\#}\mu_s = \mu_t$ means that, if $x^{(s)} \sim \mu_s$, then $T(x^{(s)}) \sim \mu_t$. The solution T_0 is called an *OT map*. In general, a solution may not exist. A sufficient condition for existence of an optimal map is that μ_s be absolutely continuous (a.c.) with respect to the Lebesgue measure [23]. A relaxed version of *Monge's* OT problem is given by the *Kantorovich* formulation [32]

$$\gamma_0 = \arg \inf_{\gamma \in \Pi(\mu_s, \mu_t)} \mathbb{E}^{\gamma}[c(x^{(s)}, x^{(t)})^p], \tag{2}$$

where $\Pi(\mu_s, \mu_t)$ is the set of probability measures on $\mathcal{X} \times \mathcal{X}$ with marginals μ_s and μ_t. The solution γ_0 is called the optimal coupling. The associated optimal value

$$W_p(\mu_s, \mu_t) = \inf_{\gamma \in \Pi(\mu_s, \mu_t)} \mathbb{E}^{\gamma}[c(x^{(s)}, x^{(t)})^p]^{1/p}, \tag{3}$$

is referred to in the literature as the Wasserstein-p distance with respect to c [23], with the typical choice for c being the Euclidean distance.

2.4. Distributionally Robust Optimization (DRO)

For some $\varepsilon > 0$, let $\Omega = \{\mathbf{Q} \in \mathcal{P}(\mathcal{Z}) | \mathsf{d}(\mathbf{Q}, \mathbf{P}_0) \leq \varepsilon\}$ be an uncertainty set that is defined with respect to a discrepancy measure d. The set Ω includes all probability distributions that are ε-distant from the center distribution $\mathbf{P}_0$, which is typically chosen to be the empirical distribution $\hat{\mathbf{P}}_n$ supported on $\{\zeta_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbf{P}$. Formally, the (DRO) problem is defined as

$$\inf_{\theta} \sup_{\mathbf{Q} \in \Omega} \mathbb{E}^{\mathbf{Q}}[l(\theta)] \,. \tag{4}$$

A main challenge in DRO is selecting ε such that Ω contains the true distribution $\mathbf{P}$ with high probability, thus providing a high-probability upper bound on the true risk [11]. In other words, for any fixed model θ, we need to ensure that $\mathbb{E}^{\mathbf{P}}[l(\theta)] \leq \sup_{\mathbf{Q} \in \Omega} \mathbb{E}^{\mathbf{Q}}[l(\theta)]$ with high confidence. This is achieved by guaranteeing that $\mathbf{P} \in \Omega$ with high probability.

3. DISTRIBUTIONALLY ROBUST UDA VIA OPTIMAL TRANSPORT

Given a labeled source domain sample $\{\zeta_i^{(s)} = (x_i^{(s)}, y_i^{(s)})\}_{i=1}^m$ and unlabeled target domain sample $\{x_i^{(t)}\}_{i=1}^n$, drawn *i.i.d.* from source and target distributions $\mathbf{P}_s$ and $\mathbf{P}_t$, respectively, we seek to learn a robust decision variable θ that minimizes the target domain risk and generalizes well on unseen target domain samples.

Since the target domain samples are unlabeled, it is not feasible to directly apply the DRO formulation (4) with respect to the target domain samples, as minimizing the worst-case risk requires labeled samples. An alternative approach is to construct the ambiguity set in the DRO formulation (4) around the empirical source domain distribution. However, this can pose a significant challenge when aiming for solutions that generalize effectively to the target domain. Specifically, ensuring that $\mathbf{P}_t \in \Omega$ with high confidence may result in a large ambiguity set, particularly if $\mathbf{P}_t$ is substantially different from $\mathbf{P}_s$. Consequently, this could lead to an overly regularized decision function θ.

To address this issue and ensure favorable out-of-sample performance on the target domain, we propose constructing an alternative uncertainty set. This set is designed to enable generalization to unseen samples from the true $\mathbf{P}_t$ with a potentially much smaller radius, thereby ensuring a moderately regularized model.

We leverage the labeled source domain samples to transfer knowledge, specifically the label information, to the target domain. Our approach assumes the existence of a transformation T that maps the source data to the target domain while preserving the labeling information across domains, as expressed by $\mathbf{P}_s(y|x^{(s)}) = \mathbf{P}_t(y|T(x^{(s)}))$ [21]. Following [21], we also make the assumption that T corresponds to an OT map (assuming that both source and target measures are a.c.) with respect to the squared Euclidean cost. As we shall see, this assumption enables the transfer of information across domains.

Given the data samples from the source and the target domains we can estimate T at the source samples in two steps, similar to [21]. First, we utilize the *Kantorovich* relaxation (2) to estimate the optimal coupling γ. Given the data in the source and target domains, $\{x_i^{(s)}\}_{i=1}^m$ and $\{x_i^{(t)}\}_{i=1}^n$, respectively, we estimate the optimal coupling as

$$\hat{\gamma}_{m,n} = \underset{\gamma \in \Pi(\hat{\mu}_m^{(s)}, \hat{\mu}_n^{(t)})}{\operatorname{argmin}} \mathbb{E}^{\gamma}[\|x^{(s)} - x^{(t)}\|_2^2] \tag{5}$$

where $\hat{\mu}_m^{(s)}$ and $\hat{\mu}_n^{(t)}$ denote the empirical marginal source and target domain distributions, respectively, based on the given samples. Second, we utilize barycentric projection [33] to estimate T. Given the estimated coupling, $\hat{\gamma}_{m,n}$, denote the estimate of the transformation by $T_{\hat{\gamma}_{m,n}}$, which is defined as the conditional mean of $x^{(t)}$ given $x^{(s)}$ under $\hat{\gamma}_{m,n}$,

$$\tilde{z}_i = T_{\hat{\gamma}_{m,n}}(x_i^{(s)}) = m \sum_j \hat{\gamma}_{m,n}(i,j) x_j^{(t)} \,. \tag{6}$$

Since we have assumed that $\mathbf{P}_s(y|x^{(s)}) = \mathbf{P}_t(y|T(x^{(s)}))$, we can transfer the source domain's labels to the transported data $\tilde{z}_i$, resulting in the transported labeled data $\{\zeta_i^*\}_{i=1}^m = \{\tilde{z}_i, y_i^s\}_{i=1}^m$.

We denote by $\tilde{\mathbf{P}}_m^{(t)}$ the empirical transported distribution supported on $\{\zeta_i^*\}_{i=1}^m$. We use $\tilde{\mathbf{P}}_m^{(t)}$ as a center for the desired ambiguity set, which we define with respect to a Wasserstein-p distance. Consequently, we can readily define our `DRDA-OT` program as

$$\inf_{\theta} \sup_{\mathbf{Q} \in \Omega} \mathbb{E}^{\mathbf{Q}}[l(\theta)] \,, \tag{7}$$

where

$$\Omega = \{\mathbf{Q} \in \mathcal{P}(\mathcal{Z}) | W_p(\mathbf{Q}, \tilde{\mathbf{P}}_m^{(t)}) \leq \varepsilon\}$$

is the ambiguity set. Since Ω contains all joint distributions over the feature and label space, we define the cost function as $c(\zeta, \zeta') = \|x - x'\| + \frac{\kappa}{2} L(y, y')$, where L is some metric measuring label discrepancy. A description of our `DRDA-OT` approach is provided in Algorithm 1.

A key aspect in `DRDA-OT` is to select the radius ε large enough such that $\mathbf{P}_t \in \Omega$ with high confidence. This ensures that $\mathbb{E}^{\mathbf{P}_t}[l(\theta)] \leq \sup_{\mathbf{Q} \in \Omega} \mathbb{E}^{\mathbf{Q}}[l(\theta)]$ holds with high probability, similar to what is shown in [34] for the DRO problem.

4. NUMERICAL EXPERIMENTS

In this section, we consider the special case of logistic regression to evaluate the performance of `DRDA-OT`.

Algorithm 1 DRDA−OT

Input: Labeled source domain data $\{x_i^{(s)}, y_i^{(s)}\}_{i=1}^m$, unlabeled target domain data $\{x_i^{(t)}\}_{i=1}^n$. Uncertainty set radius = ε
Output: Decision variable θ.

1: Obtain the optimal empirical coupling $\hat{\gamma}_{m,n}$ between the features of the source and target data by solving (5).

2: Use the barycentric projection defined in (6) to find the transported source features $\{\tilde{z}_i\}_{i=1}^m$.

3: Construct the new transported labeled data set $\{\zeta_i^*\}_{i=1}^m = \{\tilde{z}_i, y_i^{(s)}\}_{i=1}^m$.

4: Find the optimal θ by solving the DRDA−OT in (7) by optimizing over a set of distributions centered at the empirical transported distribution $\tilde{\mathbf{P}}_m^{(t)}$.

4.1. Logistic Regression

In this scenario, we consider the logloss function , i.e., $l(\theta) = l_\theta(x, y) = \log(1 + \exp(-y\langle\theta, x\rangle))$, where $x \in \mathbb{R}^d$ and $y \in \{+1, -1\}$. By leveraging the strong duality result in [34] for robust logistic regression (without DA), we can show that the DRDA−OT formulation (7) associated with the logloss function, the cost function $c(\zeta, \zeta') = \|x - x'\| + \frac{\kappa}{2}|y - y'|$, and an ambiguity set defined with respect to Wassertein-1 distance, admits the following tractable reformulation

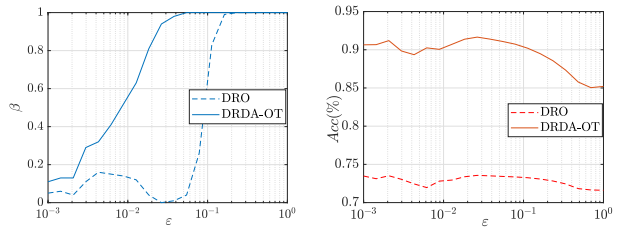
$$\inf_{\theta} \sup_{\mathbf{Q}\in\Omega} \mathbb{E}^{\mathbf{Q}}[l(\theta)] = \begin{cases} \min\limits_{\theta,\lambda,s_i} & \lambda\varepsilon + \frac{1}{m}\sum\limits_{i=1}^m s_i \\ \text{s.t.} & l_\theta(\tilde{z}_i, y_i) \le s_i, \forall i \le m \\ & l_\theta(\tilde{z}_i, y_i) - \lambda\kappa \le s_i, \forall i \le m \\ & \|\theta\|_* \le \lambda. \end{cases} \tag{8}$$

Here, $\|\cdot\|_*$ represents the dual norm of $\|\cdot\|$.

4.2. DRDA-OT for Logistic Regression

Out-of-sample Performance: In this experiment, we examine the out-of-sample guarantees provided by the DRDA−OT formulation (7) with respect to the target domain distribution. We use a similar example as in [34]. We assume that the features of the source $x^{(s)} \in \mathbb{R}^{10}$ follow a multivariate normal distribution and the conditional distribution of $y^{(s)}$ is modeled as $\mathbf{P}_s(y^{(s)}|x^{(s)}) = \frac{1}{1+\exp(-y^{(s)}\langle\theta_0, x^{(s)}\rangle)}$, with $\theta_0 = (10, \ldots, 0)$. The target domain samples are generated by rotating each two consecutive features of the source domain samples independently by δ degrees, i.e., we rotate the features $\{(1,2), (3,4), \ldots, (9,10)\}$. Specifically, we use the rotational matrix $R(\delta) = \begin{pmatrix} \cos(\delta) & -\sin(\delta) \\ \sin(\delta) & \cos(\delta) \end{pmatrix}$, and choose $\delta = 45$ degrees.

Our experimental analysis involves 100 runs. In each run, we generate 50 samples from the source and target domains for training, and 10^4 from the target distribution for testing (to obtain a reliable estimate of the true target domain risk). For each run, we utilize the training data to estimate the decision variable θ, and then evaluate it using the test data. Specifically, we record the average number of runs for which the test risk is smaller than or equal to the worst-case training risk, and the Accuracy (Acc) averaged over these runs. We compare the solutions obtained from DRDA−OT (7) with those from source DRO (4) (centered at the empirical source distribution). Fig. 1a and 1b show the fraction β of runs where the target risk is smaller than or equal to the worst case training risk, and the Acc of the test data as a function of the radius ε. As shown, our DRDA−OT approach requires a significantly smaller radius than source DRO, which underscores the superiority of our approach in generalizing to unseen samples from the target domain while avoiding unduly conservatism. Moreover, it can be observed that Acc improves significantly when using the transported data.



(a) (b)

Fig. 1: a) Fraction of runs where the test risk is smaller than the worst case risk as a function of the radius. b) Accuracy (Acc) as a function of the radius.

Effect of Sample Size: In this part, we study the effect of the source and target domain sample sizes on the radius of the ambiguity set in our DRDA−OT formulation (7). Using the same data as in the first experiment, we generate samples of varying sizes, $\{25, 50, 75, 100\}$, from the source and target domains, and solve for the decision variable. Our experiment consists of 100 runs, and we test our approach on 10^4 target domain samples. Fig. 2 illustrates the fraction β of runs for which the target risk is upper bounded by the worst-case risk as a function of the radius ε, averaged over all runs. As expected, for $\beta > 0.9$ (high confidence), the radius decreases as the sample size increases.

Robustness to Attacks: We test the robustness of the solution of our DRDA−OT formulation against PGD attacks [35]. For evaluation, we use the well-known digit datasets [36, 37], specifically considering the adaptation scenario USPS → MNIST. We focus on binary classification of two different digits, sampling 100 samples from each domain to construct the transported data and solve for DRDA−OT. We evaluate our solution on an independent dataset from the target domain, perturbing the test images using a PGD attack with different L_2-norm attack levels. We compare the DRDA−OT solution with source DRO, Empirical Risk Minimization (ERM) without adaptation (using the source data for training), and ERM with adaptation (using the transported data for training). Fig.

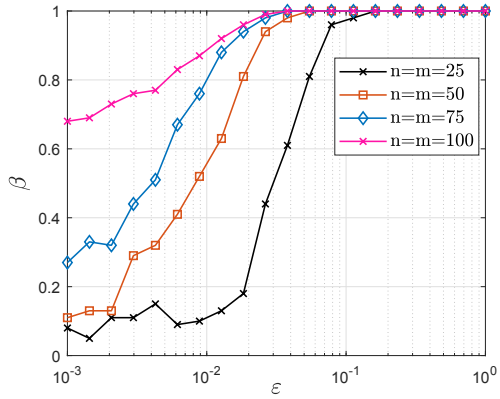


Fig. 2: Fraction of runs where the test risk is smaller than the worst-case risk as a function of the radius for different sample sizes for the DRDA-OT.

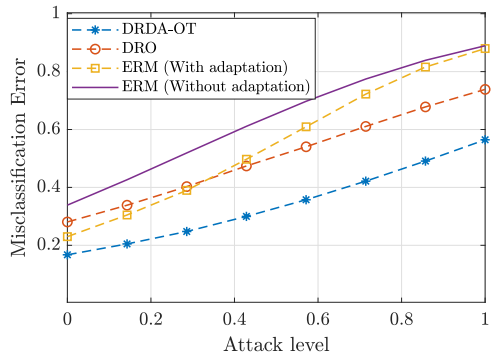


Fig. 3: Misclassification error as a function of the PGD attack level for different classification approaches.

3 shows the misclassification error of DRDA-OT, DRO, ERM (without adaptation), and ERM (with adaptation) as a function of the attack level, averaged over 10 runs. It is evident that the DRDA-OT solution achieves the dual objective of being robust to attacks in the target domain and achieving the lowest misclassification error, indicating the success of our DA approach.

5. CONCLUSION

Current methods for domain adaptation often fail to produce decision models that remain robust to disturbances and exhibit strong generalization capabilities when confronted with unseen data from the target domain. To address this challenge, we developed a robust approach to domain adaptation called Distributionally Robust Domain Adaptation via Optimal Transport (DRDA-OT). Our formulation accounts for both domain shift and uncertainty in the target domain. We utilized optimal transport and barycentric projection to construct the transported data, subsequently transferring the source domain labels. We then constructed an uncertainty set of distributions centered at the empirical distribution of the transported data with respect to the Wasserstein distance. This approach ensures that the uncertainty set is large enough to include the target distribution with high confidence, thereby guaranteeing that the target risk is upper bounded by the worst-case risk, but without being overly conservative. Our numerical experiments demonstrated the out-of-sample performance of our approach. DRDA-OT required a much smaller radius to achieve generalization on the target domain compared to other methods. Additionally, we showed the robustness of our approach to PGD attacks in comparison to ERM and standard DRO approaches.

6. REFERENCES

[1] Christopher Bishop and Nasser Nasrabadi, *Pattern recognition and machine learning*, vol. 4, Springer, 2006.

[2] James E Smith and Robert L Winkler, "The optimizer's curse: Skepticism and postdecision surprise in decision analysis," *Management Science*, vol. 52, no. 3, pp. 311–322, 2006.

[3] Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh, "Wasserstein distributionally robust optimization: Theory and applications in machine learning," in *Operations Research & Management Science in the Age of Analytics*, pp. 130–166. Informs, 2019.

[4] Giuseppe Calafiore and L. El Ghaoui, "On distributionally robust chance-constrained linear programs," *Journal of Optimization Theory and Applications*, vol. 130, pp. 1–22, 2006.

[5] Hamed Rahimian and Sanjay Mehrotra, "Distributionally robust optimization: A review," *arXiv preprint arXiv:1908.05659*, 2019.

[6] Laurent El Ghaoui, Maksim Oks, and Francois Oustry, "Worst-case value-at-risk and robust portfolio optimization: A conic programming approach," *Operations research*, vol. 51, no. 4, pp. 543–556, 2003.

[7] Giuseppe C Calafiore, "Ambiguous risk measures and optimal robust portfolios," *SIAM Journal on Optimization*, vol. 18, no. 3, pp. 853–877, 2007.

[8] Henry Lam, "Robust sensitivity analysis for stochastic systems," *Mathematics of Operations Research*, vol. 41, no. 4, pp. 1248–1275, 2016.

[9] Matthew Staib and Stefanie Jegelka, "Distributionally robust optimization and generalization in kernel methods," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[10] Jia-Jie Zhu, Wittawat Jitkrittum, Moritz Diehl, and Bernhard Schölkopf, "Kernel distributionally robust optimization: Generalized duality theorem and stochastic approximation," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 280–288.

[11] Peyman Mohajerin Esfahani and Daniel Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, no. 1-2, pp. 115–166, 2018.

[12] Rui Gao and Anton Kleywegt, "Distributionally robust stochastic optimization with Wasserstein distance," *Mathematics of Operations Research*, vol. 48, no. 2, pp. 603–655, 2023.

[13] Erick Delage and Yinyu Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Operations Research*, vol. 58, no. 3, pp. 595–612, 2010.

[14] Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani, "Regularization via mass transportation," *Journal of Machine Learning Research*, vol. 20, no. 103, pp. 1–68, 2019.

[15] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2960–2967.

[16] Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin, "How transferable are neural networks in NLP applications?," *arXiv preprint arXiv:1603.06111*, 2016.

[17] Xiaochuang Han and Jacob Eisenstein, "Unsupervised domain adaptation of contextualized embeddings for sequence labeling," *arXiv preprint arXiv:1904.02817*, 2019.

[18] Ismail R Alkhouri, Akram S Awad, Connor Hatfield, and George K Atia, "A discriminative approach to unsupervised domain adaptation in coarse-to-fine classifiers," in *IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP)*, 2023.

[19] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[20] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu, "Transfer feature learning with joint distribution adaptation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2200–2207.

[21] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy, "Optimal transport for domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 9, pp. 1853–1865, 2017.

[22] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy, "Joint distribution optimal transportation for domain adaptation," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[23] Filippo Santambrogio, "Optimal transport for applied mathematicians," *Birkäuser, NY*, vol. 55, no. 58-63, pp. 94, 2015.

[24] Gabriel Peyré, Marco Cuturi, et al., "Computational optimal transport: With applications to data science," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

[25] Anqi Liu and Brian Ziebart, "Robust classification under sample selection bias," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[26] Xiangli Chen, Mathew Monfort, Anqi Liu, and Brian D Ziebart, "Robust covariate shift regression," in *Artificial Intelligence and Statistics*. PMLR, 2016, pp. 1270–1279.

[27] Anqi Liu, Rizal Fathony, and Brian D Ziebart, "Kernel robust bias-aware prediction under covariate shift," *arXiv preprint arXiv:1712.10050*, 2017.

[28] Yibin Wang and Haifeng Wang, "Distributionally robust unsupervised domain adaptation," *Journal of Computational and Applied Mathematics*, vol. 436, pp. 115369, 2024.

[29] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan, "Deep transfer learning with joint adaptation networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2208–2217.

[30] Akram S Awad and George K Atia, "Distributionally robust domain adaptation," in *IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP)*, 2023.

[31] G. Monge, "Mémoire sur la théorie des déblais et des remblais," *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[32] Leonid V Kantorovich, "On the translocation of masses," *Journal of Mathematical Sciences*, vol. 133, no. 4, pp. 1381–1382, 2006.

[33] Nabarun Deb, Promit Ghosal, and Bodhisattva Sen, "Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29736–29753, 2021.

[34] Soroosh Shafieezadeh Abadeh, Peyman M Mohajerin Esfahani, and Daniel Kuhn, "Distributionally robust logistic regression," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

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

[36] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[37] J.J. Hull, "A database for handwritten text recognition research," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 5, pp. 550–554, 1994.