2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP) | 979-8-3503-2411-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/MLSP55844.2023.10286010

DISTRIBUTIONALLY ROBUST DOMAIN ADAPTATION

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

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

ABSTRACT

Domain Adaptation (DA) has recently received significant attention due to its potential to adapt a learning model across source and target domains with mismatched distributions. Since DA methods rely exclusively on the given source and target domain samples, they generally yield models that are vulnerable to noise and unable to adapt to unseen samples from the target domain, which calls for DA methods that guarantee the robustness and generalization of the learned models. In this paper, we propose DRDA, a distributionally robust domain adaptation method. DRDA leverages a distributionally robust optimization (DRO) framework to learn a robust decision function that minimizes the worst-case target domain risk by transferring knowledge from a given labeled source domain sample. We utilize the Maximum Mean Discrepancy (MMD) metric to construct an ambiguity set of distributions that provably contains the source and target domain distributions with high probability. Hence, the risk is shown to upper bound the out-of-sample target domain loss. Our experimental results demonstrate that our formulation outperforms existing robust learning approaches.

Index Terms— Domain Adaptation, robust learning, regression.

1. INTRODUCTION

The performance of conventionally trained machine learning models can significantly degrade when the distribution of the data at the time of inference is different from that at the time of training. Domain adaptation (DA) is concerned with adapting learning algorithms trained in a source domain using samples from a given distribution to a target domain where the test samples are drawn from a different distribution [1]. Given its ability to mitigate the distributional mismatch, DA has made significant strides in diverse application domains, including but not limited to computer vision [2–5], natural language processing [6–8], and regression analysis [9].

The key challenge underlying DA is to reduce the discrepancy between the source and target domain distributions, which has been tackled using a number of approaches. One main approach is *instance weighting* in which the source sample instances are re-weighted to minimize the distribution mismatch while learning a decision function [10, 11]. An alternative strategy is to find across-domain feature representations that simultaneously minimize the discrepancy between distributions and preserve intrinsic statistical and structural properties of the data [12–14]. The main shortcoming of the foregoing approaches is that the decision function they learn is often insufficiently robust to generalize to unseen samples from the target domain. This is largely because they minimize the discrepancy between the empirical distributions associated with the given source and target samples rather than the true population distributions. In turn, the learned decision function has propensity for unpredictable performance in the presence of noise or with out-of-sample data.

Distributionally Robust Optimization (DRO) is the problem of finding the optimal decision function that minimizes the worst-case risk over an uncertainty (or ambiguity) set of distributions. Several ways have been proposed in the literature to construct such ambiguity sets. One approach uses *moment-based* ambiguity sets, which include all distributions that satisfy certain statistical properties in the form of some moment constraints [15]. An alternative approach – the focus of this work – constructs *distance-based* ambiguity sets, which define a ball of distributions that are within a certain distance with respect to some discrepancy metric from an empirical distribution. A key result of the latter is that, if the ambiguity set is large enough to contain the true population distribution with high probability, then the worst-case risk gives a high-probability upper bound on the population risk.

Different discrepancy metrics have been used to construct the ambiguity set such as the Wasserstein distance [16, 17] and the Kullback–Leibler divergence [18, 19]. While this choice of metrics is motivated by a number of structural results that facilitate the solution to the DRO formulation, the resulting ambiguity sets have main drawbacks. The Kullback–Leibler divergence set contains only discrete distributions with the same finite support as the empirical distribution, which makes it unsuitable when the true population distribution is continuous. The Wasserstein ambiguity set is computationally expensive

This work was supported by NSF Award CCF-2106339, NSF CAREER Award CCF-1552497, and DOE Award DE-EE0009152.

and, more importantly, its radius has to scale with the data dimension to certify out-of-sample performance. To address these limitations, [20] defines the ambiguity set with respect to the Maximum Mean Discrepancy (MMD) [21], resulting in an optimization over embedding means of distributions. The MMD DRO averts the aforementioned drawbacks since the MMD ambiguity set contains both discrete and continuous distributions and its radius is independent of the data dimension. The unifying work of [22] introduces a wide range of kernel-based ambiguity sets and relaxes the assumptions on the loss function in the DRO formulation.

Here, we propose a distributionally robust framework for domain adaptation. The main objective of our formulation is to learn a robust and generalized regression function that generalizes well on a target domain given a labeled and an unlabeled sample from the source and target domains, respectively. Some previous works have considered the use of DRO across domains for DA [23–26]. There, the search is over an uncertainty set of probabilistic (conditional) mappings from input to output subject to moment constraints. In sharp contrast, here we define an uncertainty set of joint distributions within a given distance from a weighted empirical source domain distribution with respect to the MMD metric, with the main goal of establishing guarantees on out-of-sample performance. Our work makes three contributions. First, we formulate a robust DA problem, dubbed Distributionally Robust Domain Adaptation (DRDA), to learn a robust regression model that guarantees the out-of-sample performance in the target domain. Second, we construct an MMD ambiguity set and prove that it contains the source and target domain distributions with high probability, thereby ensuring the generalization of the learned model. Third, we develop a solution methodology to the formulated DRDA problem leveraging a DRO formulation under an additional common assumption on the loss function.

2. BACKGROUND

2.1. Notation

Let $\mathcal{X} \subseteq \mathbb{R}^N$ be a topological input space, and $\mathcal{P}$ the set of all probability measures defined on $\mathcal{X}$. Let l_h be a loss function associated with a decision function h, and $d : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ a distance metric between probability measures. We denote the pair (x, y) by ζ. We use $\hat{\mathbf{P}}(.) = \frac{1}{m}\sum_{j=1}^{m} \delta_{\zeta_j}(.)$ to represent the empirical distribution on the sample $\{\zeta_j\}_{j=1}^{m}$, where δ is the Dirac measure. Let $\mathcal{H}$ be a Reproducing Kernel Hilbert Space (RKHS) associated with a characteristic kernel k, $\phi : \mathbb{R}^N \to \mathcal{H}$ is the corresponding feature map, $\langle ., . \rangle_{\mathcal{H}}$ is the inner product on $\mathcal{H}$, and $\|.\|_{\mathcal{H}}$ is the induced norm. We define $\mu_{\mathbf{P}} = \mathbb{E}^{\mathbf{P}}[\phi(x)]$ to be the embedding mean of the probability measure $\mathbf{P}$, where $\mu_{\mathbf{P}} \in \mathcal{H}$ and $\mathbb{E}^{\mathbf{P}}[.]$ denotes the expectation with respect to measure $\mathbf{P}$. The kernel k is assumed to be characteristic, thus every probability measure $\mathbf{P} \in \mathcal{P}$ is embedded as a unique element in $\mathcal{H}$ [27]. Hence, the embedding mean $\mu_{\mathbf{P}}$ is an injective map.

2.2. Domain Adaptation (DA)

In the *unsupervised* DA setting, we are given a labeled source domain sample $D_s = \{x_i, y_i\}_{i=1}^{n_s}$ and an unlabeled target domain sample $D_t = \{x_i\}_{i=1}^{n_t}$ drawn from two different distributions, where n_s and n_t are the source and target domain sample sizes, respectively. In DA, one seeks to find a decision function h that minimizes the target domain risk $R_t = \mathbb{E}^{\mathbf{P_t}}[l_h]$. Since the labeling information is unavailable for the target domain sample, DA transfers knowledge from the source domain to improve the performance of the learned decision function in the target domain.

Covariate shift and density ratio: Under the covariate shift assumption, it is assumed that the conditional distributions of the labels given the features are similar across domains, but the marginals are different [11]. We define the density ratio between the target and source domain distributions as $w(\zeta) = \frac{\mathbf{P_t}(\zeta)}{\mathbf{P_s}(\zeta)}$, where $\mathbf{P_t}$ is absolutely continuous with respect to $\mathbf{P_s}$. Also, let us denote by $\hat{\mathbf{P}}_{t'}(.) = \frac{1}{n_s}\sum_{i=1}^{n_s} w(\zeta_i)\delta_{\zeta_i}(.)$ the empirical weighted source domain distribution.

The Maximum Mean Discrepancy (MMD) distance, denoted by d_m, between two probability distributions $\mathbf{Q}$ and $\mathbf{P}$ is defined as

$$\begin{aligned} d_m(\mathbf{Q}, \mathbf{P}) &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}^{\mathbf{Q}}[f(x)] - \mathbb{E}^{\mathbf{P}}[f(x)] \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} < f, \mu_{\mathbf{Q}} - \mu_{\mathbf{P}} >_{\mathcal{H}} = \|\mu_{\mathbf{Q}} - \mu_{\mathbf{P}}\|_{\mathcal{H}} . \end{aligned} \tag{1}$$

Since the embedding mean is an injective map when the kernel associated with the RKHS is characteristic, the MMD metric can measure the distance between distributions by finding the distance between their embedding means.

Kernel Mean Matching (KMM) [11] is concerned with finding the weights $w(x) \geq 0$ such that the MMD distance between the weighted source and target probability measures is minimized. Thus, the KMM problem is defined as follows

$$\begin{aligned} &\min_{w(x)} \|\mathbb{E}^{\mathbf{P_s}}[w(x)\phi(x)] - \mathbb{E}^{\mathbf{P_t}}[\phi(x)]\|_{\mathcal{H}} \\ &\text{subject to} \quad w(x) \geq 0,\ \mathbb{E}^{\mathbf{P_s}}[w(x)] = 1 . \end{aligned} \tag{2}$$

For a given source D_s and target D_t domain samples, one can define the empirical KMM as follows:

$$\begin{aligned} &\min_{w(x_i), 1 \leq i \leq n_s} \left\| \frac{1}{n_s}\sum_{i=1}^{n_s} w(x_i)\phi(x_i) - \frac{1}{n_t}\sum_{i=1}^{n_t} \phi(x_i) \right\|_{\mathcal{H}}^2 \\ &\text{s.t.} \quad w(x_i) \in [0, B], \left| \sum_{i=1}^{n_s} w(x_i) - n_s \right| \leq n_s c, \quad c > 0 . \end{aligned} \tag{3}$$

2.3. Distributionally Robust Optimization (DRO)

Let $\zeta_1, \ldots, \zeta_n$ be an i.i.d. sample drawn from a probability distribution $\mathbf{P} \in \mathcal{P}$. The Distributionally Robust Optimization

(DRO) problem is defined as

$$\hat{J}_n = \inf_h \sup_{\mathbf{Q}\in\Omega} \mathbb{E}^{\mathbf{Q}}[l_h]\,, \tag{4}$$

where $\Omega = \{\mathbf{Q}|d(\mathbf{Q},\hat{\mathbf{P}}) \leq \epsilon\}$ is the ambiguity set. The DRO problem (4) finds the learning model h that minimizes the worst-case risk. Specifically, it optimizes over all distributions in the ambiguity set, i.e., that are within a distance ϵ from the empirical distribution $\hat{\mathbf{P}}$. A key challenge in the DRO problem is to construct an ambiguity set that contains the true population distribution with high confidence. Formally, if we ensure that $\mathbf{P} \in \Omega$ with high probability, then fixing any model h, $\mathbb{E}^{\mathbf{P}}[l_h] \leq \sup_{\mathbf{Q}\in\Omega} \mathbb{E}^{\mathbf{Q}}[l_h]$ with high probability, i.e. it provides a high probability bound on the true population risk.

3. DISTRIBUTIONALLY ROBUST DOMAIN ADAPTATION (DRDA)

In this section, we present the problem setup, formulate the DRDA problem, and establish our main theoretical result.

We are given two samples D_s and D_t in the source and target domains, respectively, where the labels of D_t are not available. The samples are drawn from probability measures $\mathbf{P_s}$ and $\mathbf{P_t}$, respectively. The goal of DRDA is to learn a hypothesis $h : \mathcal{X} \to \mathbb{R}$ that simultaneously minimizes the target domain risk $R_t = \mathbb{E}^{\mathbf{P_t}}[l_h]$, and generalizes well on any unseen sample from the target domain distribution $\mathbf{P_t}$.

Towards this goal, we seek to solve the following DRO problem

$$\hat{J}_{n_t} = \inf_h \sup_{\mathbf{Q}\in\Omega_t} \mathbb{E}^{\mathbf{Q}}[l_h], \tag{5}$$

where the ambiguity set $\Omega_t := \{\mathbf{Q}|d(\mathbf{Q},\hat{\mathbf{P}}_t) \leq \epsilon_t\}$ is centered at the empirical target domain probability measure $\hat{\mathbf{P}}_t(.) = \frac{1}{n_t}\sum_{i=1}^{n_t} \delta_{\zeta_i}(.)$. The DRO formulation in (5) finds the decision function h that minimizes the worst-case target domain risk. In principle, if we ensure that $\mathbf{P_t} \in \Omega_t$ with high confidence, then $\mathbb{E}^{\mathbf{P_t}}[l_h] \leq \sup_{\mathbf{Q}\in\Omega_t} \mathbb{E}^{\mathbf{Q}}[l_h]$ with high probability.

However, a key difficulty for establishing such guarantee is that the labels for the given target domain sample D_t are not available. Therefore, we make use of the labeled source domain sample D_s to learn a hypothesis h that achieves the desired two-fold objective. Hence, we propose the following DRO problem

$$\hat{J}_{n_{t'}} = \inf_h \sup_{\mathbf{Q}\in\Omega_{t'}} \mathbb{E}^{\mathbf{Q}}[l_h] \tag{6}$$

where $\Omega_{t'} = \{\mathbf{Q}|d(\mathbf{Q},\hat{\mathbf{P}}_{t'}) \leq \epsilon_{t'}\}$, recalling that $\hat{\mathbf{P}}_{t'}$ is the empirical weighted source domain probability. The introduced set $\Omega_{t'}$ can be viewed as a transferred version of a source domain ambiguity set Ω_s (that is centered at the empirical source domain distribution $\hat{\mathbf{P}}_s$). However, how do we set the radius $\epsilon_{t'}$ to ensure that $\mathbf{P_t} \in \Omega_{t'}$ with high confidence? We answer this question by establishing the following result.

Theorem 1. *Let $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be a positive definite kernel on the space $\mathcal{X}$ with $\|\phi(x)\|_{\mathcal{H}} \leq \sqrt{M}, \forall x \in \mathcal{X}$. Let $w(x) \in [0, B]$, and $0 < \delta < 1$. Then, with probability at least $1-\delta$,*

$$d_m(\mathbf{P_t},\hat{\mathbf{P}}_{t'}) \leq B\sqrt{\frac{M}{n_s}}\left(1+\sqrt{2\log\frac{2}{\delta}}\right). \tag{7}$$

Theorem 1 establishes a high probability upper bound on the distance between the empirical source and true target domain probabilities and its proof is deferred to an extended version of this work. By setting $\epsilon_{t'}$ to the RHS of (7), we can guarantee that $\mathbf{P_t} \in \Omega_{t'}$ with probability at least $1-\delta$, and in turn that $\mathbb{E}^{\mathbf{P_t}}[l_h] \leq \sup_{\mathbf{Q}\in\Omega_{t'}} \mathbb{E}^{\mathbf{Q}}[l_h]$.

Given the primal robust domain adaptation problem in (6), the value of $\hat{J}_{n_{t'}}$ depends on the density ratio $w(x)$, since we are optimizing over all distributions that lie in the ambiguity set $\Omega_{t'}$. The main problem is to find the decision function h corresponding to the worst-case distribution such that the discrepancy between the source and target domain probability measures is minimized. We propose to achieve this objective in two different ways: (i) we can estimate w by solving the KMM (2) then use (6) to estimate the decision function h; (ii) we can optimize jointly over the decision function h and the density ratio w, and since $\sup_{\mathbf{Q}\in\Omega_{t'}} \mathbb{E}^{\mathbf{Q}}[l_h] \leq \sup_{\mathbf{Q}\in\Omega_{t'}} \mathbb{E}^{\mathbf{Q}}[l_h] + d_m^2(\mathbf{P}_{t'},\mathbf{P}_t)$, we can instead solve

$$\inf_{h,w} \sup_{\mathbf{Q}\in\Omega_{t'}} \mathbb{E}^{\mathbf{Q}}[l_h] + \beta d_m^2(\mathbf{P}_{t'},\mathbf{P}_t)\,, \tag{8}$$

where $\beta \geq 0$ is introduced to control the domain adaptation component.

Optimizing jointly over h and w is desirable since the parameters that control the learned density ratio and the decision function are not independent [28]. More specifically, the source domain sample is common in both the estimation of the weights w and for learning the decision function h. Hence, optimizing first over w (using KMM) then over h using the DRO formulation in (6) would yield a sub-optimal solution. We refer to the formulation in (8) as Distributionally Robust Domain Adaptation (DRDA). The first term in the DRDA formulation accounts for the generalization of the learned model on the target domain, while the second regularizing term is to minimize the discrepancy between the two domain distributions.

4. PROPOSED DRDA FORMULATION

In this section, we present a DRO-based formulation to solve the DRDA problem in (8) under the assumption that $l \in \mathcal{H}$. A similar assumption was made in [20] for the DRO problem. Since $\mu_{\mathbf{P}}$ is an injective map, we have that

$$\begin{aligned}\sup_{\mathbf{Q}\in\Omega_{t'}} \mathbb{E}^{\mathbf{Q}}[l_h] &\leq \sup_{\mu_{\mathbf{Q}}:\|\mu_{\mathbf{Q}}-\mu_{\hat{\mathbf{P}}_{t'}}\|_{\mathcal{H}}\leq\epsilon_{t'}} \langle\mu_{\mathbf{Q}}, l\rangle_{\mathcal{H}} \\ &= \mathbb{E}^{\hat{\mathbf{P}}_{t'}}[l_h] + \epsilon_{t'}\|l\|_{\mathcal{H}}\end{aligned} \tag{9}$$

The inequality in (9) is because not every element of the RKHS is an embedding mean of some probability measure. Therefore, we can rewrite the DRDA problem as

$$\inf_{h,w(x)} \sup_{\mathbf{Q}\in\Omega_{t'}} \mathbb{E}^{\mathbf{Q}}[l_h] + \beta d_m^2(\mathbf{P}_{t'}, \mathbf{P}_t) \leq \inf_{h,w(x)} \mathbb{E}^{\hat{\mathbf{P}}_{t'}}[l_h] + \epsilon_{t'}\|l\|_{\mathcal{H}} + \beta d_m^2(\mathbf{P}_{t'}, \mathbf{P}_t)\,.$$

Since we do not have access to the true underlying source and target domain distributions and only have samples from both domains, we make use of the empirical MMD distance $d_m(\hat{\mathbf{P}}_{t'}, \hat{\mathbf{P}}_t)$. In addition, the weights $w(x_i)$ are bounded by a constant B per the assumption in Lemma 1. Also, the weighted source domain probability must sum up to 1, i.e., $\int_{\mathcal{X}} w(x) d\mathbf{P}_{\mathbf{s}} = 1$, thus for the empirical one $\sum_{i=1}^{n_s} w(x_i) = n_s$. We can readily formulate the final DRDA problem as

$$\inf_{h,w} \frac{1}{n_s}\sum_{i=1}^{n_s} w(x_i) l_h(x_i) + \epsilon_{t'}\|l_h\|_{\mathcal{H}} + \beta d_m^2(\mathbf{P}_{t'}, \mathbf{P}_t) \quad \text{s.t.} \quad w(x_i)\in[0,B], \left|\sum_{i=1}^{n_s} w(x_i) - n_s\right| \leq n_s c, \quad c>0\,. \tag{10}$$

5. PROPOSED SOLUTION

In this section, we present our solution to the DRDA formulation in (10). We consider the RKHS $\mathcal{H}_\sigma$ induced by a Gaussian kernel $k_\sigma(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$ of bandwidth σ, and assume a quadratic loss function $l_h = (h(x)-g(x))^2$, where $g(x)$ is a labeling function. We use $\|.\|_\sigma$ and $\langle .,.\rangle_\sigma$ to denote the norm and inner product of the corresponding RKHS $\mathcal{H}_\sigma$, respectively. Therefore, we need to minimize the objective function

$$\mathbb{E}^{\hat{\mathbf{P}}_{t'}}[(h(x)-g(x))^2] + \epsilon_{t'}\|(h-g)^2\|_\sigma + \beta d_m^2(\hat{\mathbf{P}}_{t'}, \hat{\mathbf{P}}_t). \tag{11}$$

To bound the norm $\|(h-g)^2\|_\sigma$, we need the following lemma from [20, Theorem 4.1].

Lemma 2 ([20]). *If $h, g \in \mathcal{H}_\sigma$, that is, the RKHS corresponds to the Gaussian kernel, then $\|hg\|_{\frac{\sigma}{\sqrt{2}}} \leq \|h\|_\sigma \|g\|_\sigma$.*

Since $fg \in \mathcal{H}_{\frac{\sigma}{\sqrt{2}}}$, from the triangular inequality, it follows that $\|(h-g)^2\|_\sigma \leq \|h^2\|_{\frac{\sigma}{\sqrt{2}}} + \|g^2\|_{\frac{\sigma}{\sqrt{2}}} + 2\|h\|_\sigma\|g\|_\sigma$. Therefore, the objective function in (11) can be bounded by

$$\mathbb{E}^{\hat{\mathbf{P}}_{t'}}[(h(x)-g(x))^2] + \epsilon_{t'}(\|h^2\|_{\frac{\sigma}{\sqrt{2}}} + \|g^2\|_{\frac{\sigma}{\sqrt{2}}} + 2\|h\|_\sigma\|g\|_\sigma) + \beta d_m^2(\hat{\mathbf{P}}_{t'}, \hat{\mathbf{P}}_t) \tag{12}$$

In addition, $d_m^2(\hat{\mathbf{P}}_{t'}, \hat{\mathbf{P}}_t)$ can be written as

$$d_m^2(\hat{\mathbf{P}}_{t'}, \hat{\mathbf{P}}_t) = \frac{1}{n_s^2} w\mathbf{K}_\sigma^s w - \frac{2}{n_t n_s} w^T \mathbf{K}_\sigma^{s,t}\mathbf{1}_{n_t} + \text{const}\,, \tag{13}$$

and the first term in (12) as

$$\mathbb{E}^{\hat{\mathbf{P}}_{t'}}[(h(x)-y)^2] = \frac{1}{n_s}(\mathbf{K}_\sigma^s\alpha - y_s)^T\mathbf{W}(\mathbf{K}_\sigma^s\alpha - y_s)\,, \tag{14}$$

where the matrix $\mathbf{K}_\sigma^s$ has the elements $k_\sigma(x_i,x_j), i,j = 1,\ldots,n_s$, the matrix $\mathbf{K}_\sigma^{s,t}$ has the elements $k_\sigma(x_i,x_j), i = 1,\ldots,n_s, j = 1,\ldots,n_t$, $\mathbf{W} = \text{diag}(w(x_1),\ldots,w(x_{n_s}))$ and $y = (y_1,\ldots,y_{n_s})$, where $\text{diag}(.)$ returns a diagonal matrix of its vector argument. Since $h \in \mathcal{H}_\sigma$, then by the representer Theorem [29], we have the expansion $h = \sum_{i=1}^{n_s}\alpha_i\phi(x_i)$. Using the bound in (12), the DRDA problem in (10) can be expressed as

$$\begin{aligned}\inf_{\alpha,w} &(\mathbf{K}_\sigma^s\alpha - y_s)^T\mathbf{W}(\mathbf{K}_\sigma^s\alpha - y_s) + \lambda\,\text{Tr}((\mathbf{D}_\alpha\mathbf{K}^s_{\frac{\sigma}{\sqrt{2}}})^4) \\ &+\lambda\alpha^T\mathbf{K}_\sigma^s\alpha + \beta(\frac{1}{n_s^2}w\mathbf{K}_\sigma^s w - \frac{2}{n_t n_s}w^T\mathbf{K}_\sigma^{s,t}\mathbf{1}_{n_t}) \\ &\text{s.t.} \quad w(x_i)\in[0,B] \\ &\left|\sum_{i=1}^{n_s} w(x_i) - n_s\right| \leq n_s c, \quad c>0\end{aligned} \tag{15}$$

where $\alpha = (\alpha_1,\ldots,\alpha_{n_s})^T$, $\mathbf{D}_\alpha = \text{diag}(\alpha_1,\ldots,\alpha_{n_s})$ and $\text{Tr}(.)$ is the trace operator.

6. GENERALIZATION BOUND

In this section, we derive a generalization bound on the true (population) target domain risk $R_t = \mathbb{E}^{\mathbf{P_t}}[l_h] = \mathbb{E}^{\mathbf{P_t}}[(h(x)-g(x))^2]$ in terms of the empirical source domain risk $\hat{R}_s = \mathbb{E}^{\hat{\mathbf{P}}_\mathbf{s}}[(h(x)-g(x))^2]$. This bound is in the same spirit of [20, Theorem 4.3], which was derived for the original DRO problem (without domain adaptation).

Theorem 3. *Let the labeling function g satisfy $\|g\|_\sigma \leq \eta$. Therefore, for any $\delta > 0$, with probability $1-\delta$, the following holds for all functions h satisfying that $\|h\|_\sigma \leq \eta$:*

$$R_t \leq B\hat{R}_s + 4\eta^2 B\sqrt{\frac{M}{n_s}}\left(1 + \sqrt{2\log\frac{2}{\delta}}\right)\,. \tag{16}$$

We note that the RHS of (16) is inversely proportional to the source and target domain samples sizes n_s, and directly proportional to B. Therefore, for large sample size (i.e., $n_s \to \infty$), we have that $R_t \leq B\hat{R}_s$, i.e., depends on B, which is indicative of the degree of discrepancy between both domains.

Table 1: Different least-square methods ($\mathbf{W}$ in W-RLS and W-DRO is estimated using the KMM formulation in (2)).

Method	Formulation
Regularized Least Squares (RLS)	$\min_\alpha \|\mathbf{K}_\sigma^s\alpha - y\|_2^2 + \lambda\alpha^T\mathbf{K}_\sigma^s\alpha$
Weighted RLS (W-RLS)	$\min_\alpha(\mathbf{K}_\sigma^s\alpha - y)^T\mathbf{W}(\mathbf{K}_\sigma^s\alpha - y) + \lambda\alpha^T\mathbf{K}_\sigma^s\alpha$
Weighted DRO (W-DRO)	(6)
DRDA	(15)

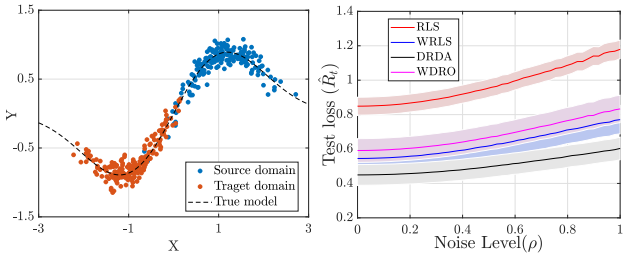
7. EXPERIMENTAL RESULTS

In this section, we verify the performance of the proposed approach. First, we generate data following the regression model $y = g(x)+n$, where $g(x) = k_\sigma(x,1)-k_\sigma(x,-1)$, n is drawn from a normal distribution with zero mean and variance 0.1^2, and the source and target domain samples follow the distributions $\mathcal{N}(1, 0.5^2)$ and $\mathcal{N}(-1, 0.6^2)$, respectively.

In our first experiment, we verify the robustness of our learned model to perturbations. We sample 50 source and target samples of size 100. For each instance, we learn a regression model and test it on an unseen target domain sample X_t of size 500 for different noise levels. A depiction of these samples along with the true model are shown in Figure 1a.

We perturb X_t with additive noise $\Delta \sim \mathcal{N}(0, \rho^2)$ with different noise levels ρ , i.e., $\hat{X}_t = X_t + \Delta$. We compute the test loss (risk) $\hat{R}_t$ for $\hat{X}_t$ for each noise level $\rho \in [0, 1]$ and report their average loss $\hat{\mathbf{R}}_t$ and the corresponding 95% interval. We compare the performance of DRDA to different least-square regression approaches (see Table 1). In the approach that we call weighted-DRO (W-DRO), we first solve for the weights w using (2), then optimize over the decision function h in (6), in contrast to the joint optimization in DRDA. The hyperparameters β and λ are set to 10 and 1.2, respectively.

Figure 1b demonstrates the test loss of the DRDA learned model for various noise levels in comparison to the least-square models. Our DRDA model achieves the lowest average loss for all noise levels due to the built-in robust domain adaptation capability along with the joint optimization over the weights and decision function in (8). To highlight the importance of the domain adaptation inherent in our framework, we also tested the standard DRO scheme, which only uses the source sample for training the model. We found DRO without domain adaptation to be considerably less robust than DRDA in this setting. For example, at $\rho = 0.8$, the DRO model yields an average test loss of 1.540 versus 0.556 for the DRDA model. Moreover, W-RLS and W-DRO, which first learn the weights then optimize over the model, underperform the DRDA model, underscoring the gain of jointly optimizing over w and h.



(a) (b)

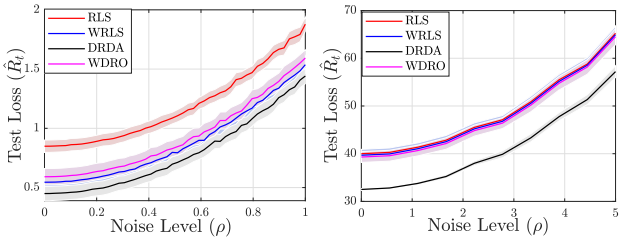
Fig. 1: (a) True model (dashed line), source (blue) and target (red) domain samples; (b) Test loss of different regression models as function of the noise level ρ.

For the second experiment, we demonstrate the effect of sample size on the target domain population risk. We sample source and target domain samples of different sizes. For each size, we use the source and target domain samples to learn the DRDA model. Table 2 shows the target domain risk as a function of the sample size. As expected, the risk decreases with the sample size, since training with a larger sample size (source and target) results in a model of higher accuracy.

Table 2: Population Risk for different sample sizes.

Sample size	50	100	150	200	250	300
Population risk	0.54	0.42	0.35	0.33	0.27	0.22

In our third experiment, we evaluate the performance of the proposed DRDA when the perturbations are added to the response. Specifically, Y_t is perturbed with additive noise Δ, i.e., $\hat{Y}_t = Y_t + \Delta$. Performance is evaluated on both synthetic and real-world data ('puma8nh' [11]) for which a sampling bias scheme is used to create the source and target datasets. As shown in Fig. 2, our approach outperforms other regression approaches as it achieves the lowest average loss at all noise levels.



(a) (b)

Fig. 2: Test loss for (a) synthetic data and (b) 'puma8nh' dataset of different regression models as function of the noise level ρ.

8. CONCLUSION

Existing approaches to domain adaptation often fall short of yielding a decision model that is robust to perturbations and generalizes well to unseen target domain data. To address this limitation, we formulated a robust domain adaptation problem, dubbed Distributionally Robust Domain Adaptation (DRDA), that leverages a DRO framework. Our formulation simultaneously accounts for domain adaptation and the uncertainty in the target domain sample. Since the target domain labels are unavailable, we re-weight the source domain sample to minimize the discrepancy between the two domains. Also, we constructed an uncertainty set, centered at the empirical weighted source domain distribution, and prove that it contains the true target domain distribution with high probability. In turn, the worst-case risk gives a high probability upper bound on the true population risk, thereby providing a guarantee on the generalization of the learned model. Our experimental results demonstrate that the learned regression model outperforms existing least-square approaches both in terms of robustness to noise and generalization power.

9. REFERENCES

[1] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[2] Luis AM Pereira and Ricardo da Silva Torres, "Semi-supervised transfer subspace for domain adaptation," *Pattern Recognition*, vol. 75, pp. 235–249, 2018.

[3] Meng Yang, Lei Zhang, Xiangchu Feng, and David Zhang, "Fisher discrimination dictionary learning for sparse representation," in *International Conference on Computer Vision*, 2011, pp. 543–550.

[4] 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*, 2017, pp. 7167–7176.

[5] 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.

[6] 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.

[7] Han Guo, Ramakanth Pasunuru, and Mohit Bansal, "Multi-source domain adaptation for text classification via distancenet-bandits," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 7830–7838.

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

[9] Corinna Cortes and Mehryar Mohri, "Domain adaptation in regression," in *International Conference on Algorithmic Learning Theory*. Springer, 2011, pp. 308–323.

[10] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," *Advances in Neural Information Processing Systems*, vol. 20, 2007.

[11] Jiayuan Huang, Alexander Smola, Arthur Gretton, Karsten Borgwardt, and Bernhard Schölkopf, "Correcting sample selection bias by unlabeled data," in *Advances in Neural Information Processing Systems*, 2006, vol. 19, pp. 601–608.

[12] 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.

[13] Jing Zhang, Wanqing Li, and Philip Ogunbona, "Joint geometrical and statistical alignment for visual domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1859–1867.

[14] Mingsheng Long, Jianmin Wang, Jiaguang Sun, and S Yu Philip, "Domain invariant transfer kernel learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 6, pp. 1519–1532, 2014.

[15] 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.

[16] Georg Pflug and David Wozabal, "Ambiguity in portfolio selection," *Quantitative Finance*, vol. 7, no. 4, pp. 435–442, 2007.

[17] Rui Gao and Anton J Kleywegt, "Distributionally robust stochastic optimization with wasserstein distance," *arXiv preprint arXiv:1604.02199*, 2016.

[18] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen, "Robust solutions of optimization problems affected by uncertain probabilities," *Management Science*, vol. 59, no. 2, pp. 341–357, 2013.

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

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

[21] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[22] Jia-Jie Zhu, Wittawat Jitkrittum, Moritz Diehl, and Bernhard Schölkopf, "Kernel distributionally robust optimization," in *24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

[23] Anqi Liu and Brian Ziebart, "Robust classification under sample selection bias," *Advances in neural information processing systems*, vol. 27, 2014.

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

[25] Bahar Taskesen, Man-Chung Yue, Jose Blanchet, Daniel Kuhn, and Viet Anh Nguyen, "Sequential domain adaptation by synthesizing distributionally robust experts," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10162–10172.

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

[27] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al., "Kernel mean embedding of distributions: A review and beyond," *Foundations and Trends® in Machine Learning*, vol. 10, no. 1-2, pp. 1–141, 2017.

[28] Steffen Bickel, Michael Brückner, and Tobias Scheffer, "Discriminative learning under covariate shift.," *Journal of Machine Learning Research*, vol. 10, no. 9, 2009.

[29] Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola, "A generalized representer theorem," in *In Proceedings of the Annual Conference on Computational Learning Theory*, 2001, pp. 416–426.