Algorithms and Theory for Multiple-Source Adaptation

Judy Hoffman
CS Department UC Berkeley
Berkeley, CA 94720
jhoffman@eecs.berkeley.edu

Mehryar Mohri
Courant Institute and Google
New York, NY 10012
mohri@cims.nyu.edu

Ningshan Zhang
New York University
New York, NY 10012
nzhang@stern.nyu.edu

Abstract

We present a number of novel contributions to the multiple-source adaptation problem. We derive new normalized solutions with strong theoretical guarantees for the cross-entropy loss and other similar losses. We also provide new guarantees that hold in the case where the conditional probabilities for the source domains are distinct. Moreover, we give new algorithms for determining the distribution-weighted combination solution for the cross-entropy loss and other losses. We report the results of a series of experiments with real-world datasets. We find that our algorithm outperforms competing approaches by producing a single robust model that performs well on any target mixture distribution. Altogether, our theory, algorithms, and empirical results provide a full solution for the multiple-source adaptation problem with very practical benefits.

1 Introduction

In many modern applications, often the learner has access to information about several source domains, including accurate predictors possibly trained and made available by others, but no direct information about a target domain for which one wishes to achieve a good performance. The target domain can typically be viewed as a combination of the source domains, that is a mixture of their joint distributions, or it may be close to such mixtures. In addition, often the learner does not have access to all source data simultaneously, for legitimate reasons such as privacy or storage limitation. Thus, the learner cannot simply pool all source data together to learn a predictor.

Such problems arise commonly in speech recognition where different groups of speakers (domains) yield different acoustic models and the problem is to derive an accurate acoustic model for a broader population that may be viewed as a mixture of the source groups (Liao, 2013). In object recognition, multiple image databases exist, each with its own bias and labeled categories (Torralba and Efros, 2011), but the target application may contain images which most closely resemble only a subset of the available training data. Finally, in sentiment analysis, accurate predictors may be available for sub-domains such as TVs, laptops and CD players, each previously trained on labeled data, but no labeled data or predictor may be at the learner’s disposal for the more general category of electronics, which can be modeled as a mixture of the sub-domains (Blitzer et al., 2007; Dredze et al., 2008).

The problem of transfer from a single source to a known target domain (Ben-David, Blitzer, Crammer, and Pereira, 2006; Mansour, Mohri, and Rostamizadeh, 2009b; Cortes and Mohri, 2014; Cortes, Mohri, and Muñoz Medina, 2015), either through unsupervised adaptation techniques (Gong et al., 2012; Long et al., 2015; Ganin and Lempitsky, 2015; Tzeng et al., 2015), or via lightly supervised ones (some amount of labeled data from the target domain) (Saenko et al., 2010; Yang et al., 2007; Hoffman et al., 2013; Girshick et al., 2014), has been extensively investigated in the past. Here, we focus on the problem of multiple-source domain adaptation and ask how the learner can combine relatively accurate predictors available for each source domain to derive an accurate predictor for
any new mixture target domain? This is known as the multiple-source adaptation (MSA) problem
first formalized and analyzed theoretically by Mansour, Mohri, and Rostamizadeh (2008, 2009a)
and later studied for various applications such as object recognition (Hoffman et al., 2012; Gong
et al., 2013a,b). Recently, Zhang et al. (2015) studied a causal formulation of this problem for a
classification scenario, using the same combination rules as Mansour et al. (2008, 2009a). A closely
related problem to the MSA problem is that of domain generalization (Pan and Yang, 2010; Muandet
et al., 2013; Xu et al., 2014), where knowledge from an arbitrary number of related domains is
combined to perform well on a previously unseen domain. Appendix G includes a more detailed
discussion of previous work related to the MSA problem.

Mansour, Mohri, and Rostamizadeh (2008, 2009a) gave strong theoretical guarantees for a distribution-
weighted combination to address the MSA problem, but they did not provide an algorithmic solution
to determine that combination. Furthermore, the solution they proposed could not be used for loss
functions such as cross-entropy, which require a normalized predictor. Their work also assumed a
deterministic scenario (non-stochastic) with the same labeling function for all source domains.

This work makes a number of novel contributions to the MSA problem. We give new normalized
solutions with strong theoretical guarantees for the cross-entropy loss and other similar losses. Our
guarantees hold even when the conditional probabilities for the source domains are distinct. A
by-product of our analysis is the extension of the theoretical results of Mansour et al. (2008, 2009a)
to the stochastic scenario, where there is a joint distribution over the input and output space.

Moreover, we give new algorithms for determining the distribution-weighted combination solution
for the cross-entropy loss and other losses. We prove that the problem of determining that solution
can be cast as a DC-programming (difference of convex) and prove explicit DC-decompositions for
the cross-entropy loss and other losses. We also give a series of experimental results with several
datasets demonstrating that our distribution-weighted combination solution is remarkably robust. Our
algorithm outperforms competing approaches and performs well on any target mixture distribution.

Altogether, our theory, algorithms, and empirical results provide a full solution for the MSA problem
with very practical benefits.

2 Problem setup

Let \( \mathcal{X} \) denote the input space and \( \mathcal{Y} \) the output space. We consider a multiple-source domain
adaptation (MSA) problem in the general stochastic scenario where there is a distribution over the
joint input-output space \( \mathcal{X} \times \mathcal{Y} \). This is a more general setup than the deterministic scenario in
(Mansour et al., 2008, 2009a), where a target function mapping from \( \mathcal{X} \) to \( \mathcal{Y} \) is assumed. This
extension is needed for the analysis of the most common and realistic learning setups in practice. We
will assume that \( \mathcal{X} \) and \( \mathcal{Y} \) are discrete, but the predictors we consider can take real values. Our theory
can be straightforwardly extended to the continuous case with summations replaced by integrals in
the proofs. We will identify a domain with a distribution over \( \mathcal{X} \times \mathcal{Y} \) and consider the scenario where
the learner has access to a predictor \( h_k \), for each domain \( \mathcal{D}_k, k \in [p] = \{1, \ldots, p\} \).

We consider two types of predictor functions \( h_k \), and their associated loss functions \( L \) under the
regression model (R) and the probability model (P) respectively,

\[
\begin{align*}
    h_k : \mathcal{X} &\rightarrow \mathbb{R} \\
    h_k : \mathcal{X} \times \mathcal{Y} &\rightarrow [0, 1] \\
    L: \mathbb{R} \times \mathcal{Y} &\rightarrow \mathbb{R}_+ \quad \text{(R)} \\
    L: [0, 1] &\rightarrow \mathbb{R}_+ \quad \text{(P)}
\end{align*}
\]

We abuse the notation and write \( L(h, x, y) \) to denote the loss of a predictor \( h \) at point \((x, y)\), that is \( L(h(x), y) \) in the regression model, and \( L(h(x, y)) \) in the probability model. We will denote by \( \mathcal{L}(\mathcal{D}, h) \) the expected loss of a predictor \( h \) with respect to the distribution \( \mathcal{D} \):

\[
\mathcal{L}(\mathcal{D}, h) = \mathbb{E}_{(x, y) \sim \mathcal{D}} \left[ L(h, x, y) \right] = \sum_{(x, y) \in \mathcal{X} \times \mathcal{Y}} \mathcal{D}(x, y) L(h, x, y).
\]

Much of our theory only assumes that \( \mathcal{L} \) is convex and continuous. But, we will be particularly
interested in the case where, in the regression model, \( L(h(x), y) = (h(x) - y)^2 \) is the squared loss,
and where, in the probability model, \( L(h(x, y)) = -\log h(x, y) \) is the cross-entropy loss (log-loss).

We will assume that each \( h_k \) is a relatively accurate predictor for the distribution \( \mathcal{D}_k \): there exists \( \epsilon > 0 \) such that \( \mathcal{L}(\mathcal{D}_k, h_k) \leq \epsilon \) for all \( k \in [p] \). We will also assume that the loss of the source
hypotheses \( h_k \) is bounded, that is \( L(h_k, x, y) \leq M \) for all \((x, y) \in \mathcal{X} \times \mathcal{Y} \) and all \( k \in [p] \).
In the MSA problem, the learner’s objective is to combine these predictors to design a predictor with small expected loss on a target domain that could be an arbitrary and unknown mixture of the source domains, the case we are particularly interested in, or even some other arbitrary distribution. It is worth emphasizing that the learner has no knowledge of the target domain.

How do we combine the $h_k$’s? Can we use a convex combination rule, $\sum_{k=1}^{p} \lambda_k h_k$, for some $\lambda \in \Delta$? In Appendix A (Lemmas 9 and 10) we show that no convex combination rule will perform well even in very simple MSA problems. These results generalize a previous lower bound of Mansour et al. (2008). Next, we show that the distribution-weighted combination rule is a suitable solution.

Extending the definition given by Mansour et al. (2008), we define the distribution-weighted combination of the functions $h_k, k \in [p]$ as follows. For any $\eta > 0, z \in \Delta$, and $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$h^\eta_z(x) = \sum_{k=1}^{p} \frac{z_k}{\sum_{k=1}^{p} z_k} D^1_k(x) + \frac{\eta}{p} \mathbb{I}(x) h_k(x),\quad (R) \quad (1)$$

$$h^\eta_z(x, y) = \sum_{k=1}^{p} \frac{z_k}{\sum_{j=1}^{p} z_j} D^1_k(x, y) + \frac{\eta}{p} \mathbb{I}(x, y) h_k(x, y),\quad (P) \quad (2)$$

where we denote by $D^1$ the marginal distribution over $\mathcal{X}$, for all $x \in \mathcal{X}$, $D^1_k(x) = \sum_{y \in \mathcal{Y}} D(x, y)$, and by $\mathbb{I}$ the uniform distribution over $\mathcal{X}$. This extension may seem technically straightforward in hindsight, but the form of the predictor was not immediately clear in the stochastic case.

### 3 Theoretical guarantees

In this section, we present a series of theoretical guarantees for distribution-weighted combinations with a suitable choice of the parameters $z$ and $\eta$, both for the regression model and for the probability model. We first give our main result for the general stochastic scenario. Next, for the probability model with cross-entropy loss, we introduce a normalized distribution weighted combination and prove that it benefits from strong theoretical guarantees.

Our theoretical results rely on a measure of divergence between two distributions. The one that naturally comes up in our analysis is the Rényi Divergence (Rényi, 1961). We will denote by $d_{\alpha}(\mathcal{D} \parallel \mathcal{D}') = e^{\alpha \beta_{\alpha}(\mathcal{D} \parallel \mathcal{D}')} \beta_{\alpha}(\mathcal{D} \parallel \mathcal{D}')$ the exponential of the $\alpha$-Rényi Divergence of two distributions $\mathcal{D}$ and $\mathcal{D}'$. See Appendix F for more details about the notion of Rénily Divergence.

#### 3.1 General guarantees for regression and probability models

Let $\mathcal{D}_T$ be an unknown target distribution. We will denote by $\mathcal{D}_T(\cdot|x)$ and $\mathcal{D}_k(\cdot|x)$ the conditional probability distribution on the target and the source domain $k$ respectively. We do not assume that the target and source conditional probabilities $\mathcal{D}_T(\cdot|x)$ and $\mathcal{D}_k(\cdot|x)$ coincide for all $k \in [p]$ and $x \in \mathcal{X}$. This is a significant extension of the MSA scenario with respect to the one considered by Mansour et al. (2009a), which assumed exactly the same labeling function $f$ for all source domains, in the deterministic scenario.

Let $\mathcal{D}_T$ be a mixture of source distributions, such that $\mathcal{D}^1_T \in \mathcal{D} = \{\sum_{k=1}^{p} \lambda_k \mathcal{D}^1_k; \lambda \in \Delta\}$ in the regression model, or $\mathcal{D}_T \in \mathcal{D} = \{\sum_{k=1}^{p} \lambda_k \mathcal{D}_k; \lambda \in \Delta\}$ in the probability model. We also assume that under the regression model, all possible target distributions $\mathcal{D}_T$ admit the same (unknown) conditional probability distribution.

Fix $\alpha > 1$ and define $\epsilon_T$ by

$$\epsilon_T = \max_{k \in [p]} \left[ \mathbb{E}_{x \sim \mathcal{D}^1_k} \left[ d_{\alpha}(\mathcal{D}_T(\cdot|x) \parallel \mathcal{D}_k(\cdot|x))^{\alpha-1} \right] \right]^{\frac{1}{\alpha-1}} \epsilon^{\frac{1}{\alpha-1}} M^\frac{1}{\alpha}. $$

$\epsilon_T$ depends on the maximal expected Rényi divergence between the target conditional probability distribution $\mathcal{D}_T(\cdot|x)$ and the source ones $\mathcal{D}_k(\cdot|x), \forall k \in [p]$, with the expectation taken over the source marginal distribution $\mathcal{D}^1_k$, and the maximum taken over $k \in [p]$. When the target conditional is close to all source ones, $\alpha$ can be chosen to be very large and $\epsilon_T$ is close to $\epsilon$. In particular, when the conditional probabilities coincide, for $\alpha = +\infty$, we have $\epsilon_T = \epsilon$. 

3
Theorem 1. For any $\delta > 0$, there exist $\eta > 0$ and $z \in \Delta$ such that the following inequalities hold for any $\alpha > 1$ and any target distribution $D_T$ that is a mixture of source distributions:

$$L(D_T, h^\eta_T) \leq \epsilon_T + \delta,$$

$$L(D_T, h^{\eta}_T) \leq \epsilon + \delta. \quad (R)$$

As discussed later, the proof of more general results (Theorem 2 and Theorem 14) is given in Appendix B. The learning guarantees for the regression and the probability model are slightly different, since the definitions of the distribution-weighted combinations are different for the two models. Theorem 1 shows the existence of $\eta > 0$ and a mixture weight $z \in \Delta$ with a remarkable property: in the regression model (R), for any target distribution $D_T$ whose conditional $D_k(\cdot|x)$ is on average not too far away from $D_k(\cdot|x)$ for any $k \in [p]$, and $D_T^\bot \in D^1$, the loss of $h^\eta_T$ on $D_T$ is small. It is even more remarkable that, in the probability model (P), the loss of $h^\eta_T$ is at most $\epsilon$ on any target distribution $D_T \in D$. Thus, $h^\eta_T$ is a robust hypothesis with favorable property for any such target distribution $D_T$.

We now present a more general result, Theorem 2, that relaxes the assumptions under the regression model that all possible target distributions $D_T$ admit the same conditional probability distribution, and that the target’s marginal distribution is a mixture of source ones. In Appendix B, we show that Theorem 2 coincides with Theorem 1 under those assumptions. In Appendix B, we further give a more general result than Theorem 1 under the probability model (Theorem 14).

To present this more general result, we first introduce some additional notation. Given a conditional probability distribution $Q(\cdot|x)$ defined for all $x \in X$, define $\epsilon_\alpha(\cdot)$ as follows:

$$\epsilon_\alpha(\cdot) = \max_{k \in [p]} \left[ \frac{1}{x \sim D^*_k} \left( \sum_{x \sim D^*_k} \left| d_\alpha(Q(\cdot|x) \parallel D_k(\cdot|x)) \right|^{\alpha-1} \right)^{\frac{1}{\alpha}} \right].$$

Thus, $\epsilon_\alpha(\cdot)$ depends on the maximal expected $\alpha$-Rényi divergence between $Q(\cdot|x)$ and $D_k(\cdot|x)$, and $\epsilon_\alpha(\cdot) = \epsilon_T$ when $Q(\cdot|x) \parallel D_T(\cdot|x)$. When there exists $Q(\cdot|x)$ such that the expected $\alpha$-Rényi divergence is small for all $k \in [p]$, then $\epsilon_\alpha(\cdot)$ is close to $\epsilon$ for $\alpha = +\infty$. In addition, we will use the following definitions: $D_{k,0}(x,y) = D_k^1(x)Q(y|x) + D_{P,0} = \{ \sum_{k=1}^p \lambda_k D_{k,0} \mid \lambda \in \Delta \}$.

**Theorem 2 (Regression model).** Fix a conditional probability distribution $Q(\cdot|x)$ defined for all $x \in X$. Then, for any $\delta > 0$, there exist $\eta > 0$ and $z \in \Delta$ such that the following inequality holds for any $\alpha, \beta > 1$ and any target distribution $D_T$:

$$L(D_T, h^\eta_T) \leq \left( (\epsilon_\alpha(\cdot) + \delta) d_\beta(D_T \parallel D_{P,0}) \right)^{\frac{\beta-1}{eta}} M^{\frac{\beta}{\beta}}. \quad (R)$$

The learning guarantee of Theorem 1 depends on the Rényi divergence between the conditional probabilities of the source and target domains and a fixed pivot $Q(\cdot|x)$. In particular, when there exists a pivot $Q(\cdot|x)$ that is close to $D_T(\cdot|x)$ and $D_k(\cdot|x)$, for all $k \in [p]$, then the guarantee is significant. One candidate for such a pivot is a conditional probability distribution $Q(\cdot|x)$ minimizing $\epsilon_\alpha(\cdot)$.

In many learning tasks, it is reasonable to assume that the conditional probability of the output labels is the same across source domains. For example, a dog picture represents a dog regardless of whether the picture belongs to an individual’s personal collection or to a broader database of pictures from multiple individuals. This is a straightforward extension of the assumption adopted by Mansour et al. (2008) in the deterministic scenario, where exactly the same labeling function $f$ is assumed for all source domains. In that case, we have $D_T(\cdot|x) = D_k(\cdot|x), \forall k \in [p]$ and therefore $d_\alpha(D_T(\cdot|x) \parallel D_k(\cdot|x)) = 1$. Setting $\alpha = +\infty$, we recover the main result of Mansour et al. (2008).

**Corollary 3.** Assume that the conditional probability distributions $D_k(\cdot|x)$ do not depend on $k$. Then, for any $\delta > 0$, there exist $\eta > 0$ and $z \in \Delta$ such that $L(D_T, h^\eta_T) \leq \epsilon + \delta$ for any mixture parameter $\lambda \in \Delta$.

Corollary 3 shows the existence of a parameter $\eta > 0$ and a mixture weight $z \in \Delta$ with a remarkable property: for any $\delta > 0$, regardless of which mixture weight $\lambda \in \Delta$ defines the target distribution, the loss of $h^\eta_T$ is at most $\epsilon + \delta$, that is arbitrarily close to $\epsilon$. $h^\eta_T$ is therefore a robust hypothesis with a favorable property for any mixture target distribution.

To cover the realistic cases in applications, we further extend this result to the case where the distributions $D_k$ are not directly available to the learner, and instead estimates $\hat{D}_k$ have been derived.
We have shown that, for both the regression and the probability model, there exists a vector \( z \) defining a distribution-weighted combination hypothesis \( h^z \) that admits very favorable guarantees. But how can we find a such \( z \)? This is a key question in the MSA problem which was not addressed by Mansour et al. (2008, 2009a): no algorithm was previously reported to determine the mixture parameter \( z \), even for the deterministic scenario. Here, we give an algorithm for determining that vector \( z \).

In this section, we give practical and efficient algorithms for finding the vector \( z \) in the important cases of the squared loss in the regression model, or the cross-entropy loss in the probability model, by leveraging the differentiability of the loss functions. We first show that \( z \) is the solution of a general optimization problem. Next, we give a DC-decomposition (difference of convex decomposition)
of the objective for both models, thereby proving an explicit DC-programming formulation of the problem. This leads to an efficient DC algorithm that is guaranteed to converge to a stationary point. Additionally, we show that it is straightforward to test if the solution obtained is the global optimum. While we are not proving that the local stationary point found by our algorithm is the global optimum, empirically, we observe that that is indeed the case.

4.1 Optimization problem

Theorem 1 shows that the hypothesis $h_z^n$ based on the mixture parameter $z$ benefits from a strong generalization guarantee. A key step in proving Theorem 1 is to show the following lemma.

**Lemma 6.** For any $\eta, \eta' > 0$, there exists $z \in \Delta$, with $z_k \neq 0$ for all $k \in [p]$, such that the following holds for the distribution-weighted combining rule $h_z^n$:

$$\forall k \in [p], \quad \mathcal{L}(\mathcal{D}_k, h_z^n) \leq \sum_{j=1}^{p} z_j \mathcal{L}(\mathcal{D}_j, h_z^n) + \eta'.$$

Lemma 6 indicates that for the solution $z$, $h_z^n$ has essentially the same loss on all source domains. Thus, our problem consists of finding a parameter $z$ verifying this property. This, in turn, can be formulated as a min-max problem: $\min_{z \in \Delta} \max_{k \in [p]} \mathcal{L}(\mathcal{D}_k, h_z^n) - \mathcal{L}(\mathcal{D}_z, h_z^n)$, which can be equivalently formulated as the following optimization problem:

$$\min_{z \in \Delta, \gamma \in \mathbb{R}} \gamma \quad \text{s.t.} \quad \mathcal{L}(\mathcal{D}_k, h_z^n) - \mathcal{L}(\mathcal{D}_z, h_z^n) \leq \gamma, \quad \forall k \in [p].$$

4.2 DC-decomposition

We provide explicit DC decompositions of the objective of Problem (4) for the regression model with the squared loss and for the probability model with the cross-entropy loss. The derivations are given in Appendix D. We first rewrite $h_z^n$ as the division of two affine functions for both the regression (R) and the probability (P) model, $h_z = J_z/K_z$, where we adopt the following definitions and notation:

$$J_z(x) = \sum_{k=1}^{p} z_k \mathcal{D}_k(x) h_k(x) + \frac{\eta}{p} \mathcal{U}_{\mathcal{M}}(x) h_k(x), \quad K_z(x) = \mathcal{D}_z(x) + \eta \mathcal{U}_{\mathcal{M}}(x), \quad (R)$$

$$J_z(x, y) = \sum_{k=1}^{p} z_k \mathcal{D}_k(x, y) h_k(x, y) + \frac{\eta}{p} \mathcal{U}(x, y) h_k(x, y), \quad K_z(x, y) = \mathcal{D}_z(x, y) + \eta \mathcal{U}(x, y). \quad (P)$$

**Proposition 7** (Regression model, squared loss). Let $L$ be the squared loss. Then, for any $k \in [p]$, $\mathcal{L}(\mathcal{D}_k, h_z^n) - \mathcal{L}(\mathcal{D}_z, h_z^n) = u_k(z) - v_k(z)$, where $u_k$ and $v_k$ are convex functions defined for all $z$ by

$$u_k(z) = \mathcal{L}(\mathcal{D}_k + \eta \mathcal{U}_{\mathcal{M}}(\mathcal{D}_k(\cdot|x), h_z^n)) - 2M \sum_x (\mathcal{D}_k^1 + \eta \mathcal{U}_{\mathcal{M}}^1)(x) \log K_z(x),$$

$$v_k(z) = \mathcal{L}(\mathcal{D}_z + \eta \mathcal{U}_{\mathcal{M}}(\mathcal{D}_z(\cdot|x), h_z^n)) - 2M \sum_x (\mathcal{D}_k^1 + \eta \mathcal{U}_{\mathcal{M}}^1)(x) \log K_z(x).$$

**Proposition 8** (Probability model, cross-entropy loss). Let $L$ be the cross-entropy loss. Then, for $k \in [p]$, $\mathcal{L}(\mathcal{D}_k, h_z^n) - \mathcal{L}(\mathcal{D}_z, h_z^n) = u_k(z) - v_k(z)$, where $u_k$ and $v_k$ are convex functions defined for all $z$ by

$$u_k(z) = -\sum_{x,y} [\mathcal{D}_k(x, y) + \eta \mathcal{U}(x, y)] \log J_z(x, y),$$

$$v_k(z) = \sum_{x,y} K_z(x, y) \log \left[ \frac{K_z(x, y)}{J_z(x, y)} \right] - [\mathcal{D}_k(x, y) + \eta \mathcal{U}(x, y)] \log K_z(x, y).$$

4.3 DC algorithm

Our DC decompositions prove that the optimization problem (4) can be cast as the following variational form of a DC-programming problem (Tao and An, 1997, 1998; Sriperumbudur and Lanckriet, 2012):

$$\min_{z \in \Delta, \gamma \in \mathbb{R}} \gamma \quad \text{s.t.} \quad (u_k(z) - v_k(z) \leq \gamma) \land (-z_k \leq 0) \land \left( \sum_{k=1}^{p} z_k - 1 = 0 \right), \quad \forall k \in [p].$$
We used the sentiment analysis dataset proposed by Mansour et al. (2008, 2009a), where the intersection of the four domains. These words were used to define feature vectors, where every sample was encoded by the number of occurrences of each word. We trained our base hypotheses using support vector regression with the same hyper-parameters as in (Mansour et al., 2008, 2009a).

The DC-programming algorithm works as follows. Let \((z_t)\), be the sequence defined by repeatedly solving the following convex optimization problem:

\[
\begin{align*}
    z_{t+1} &\in \arg\min_{z, \gamma \in \mathbb{R}} \gamma \\
    \text{s.t. } & (u_k(z) - v_k(z_t) - (z - z_t) \nabla v_k(z_t) \leq \gamma) \land (-z_k \leq 0) \land (\sum_{k=1}^{p} z_k - 1 = 0), \quad \forall k \in [p],
\end{align*}
\]

where \(z_0 \in \Delta\) is an arbitrary starting value. Then, \((z_t)\), is guaranteed to converge to a local minimum of Problem (4) (Yuille and Rangarajan, 2003; Sriperumbudur and Lanckriet, 2012). Note that Problem (6) is a relatively simple optimization problem: \(u_k(z)\) is a weighted sum of the negative logarithm of an affine function of \(z\), plus a weighted sum of rational functions of \(z\) (squared loss), and all other terms appearing in the constraints are affine functions of \(z\).

Problem (4) seeks a parameter \(z\) verifying \(\mathcal{L}((D_k, h^0_k)) - \mathcal{L}((D_k, h^0_k)) \leq \gamma\), for all \(k \in [p]\) for an arbitrarily small value of \(\gamma\). Since \(\mathcal{L}((D_k, h^0_k)) = \sum_{k=1}^{p} z_k \mathcal{L}_k(D_k, h^0_k)\) is a weighted average of the expected losses \(\mathcal{L}_k(D_k, h^0_k)\), \(k \in [p]\), the solution \(\gamma\) cannot be negative. Furthermore, by Lemma 6, a parameter \(z\) verifying that inequality exists for any \(\gamma > 0\). Thus, the global solution \(\gamma\) of Problem (4) must be close to zero. This provides us with a simple criterion for testing the global optimality of the solution \(z\) we obtain using a DC-programming algorithm with a starting parameter \(z_0\).

## 5 Experiments

This section reports the results of our experiments with our DC-programming algorithm for finding a robust domain generalization solution when using squared loss and cross-entropy loss. We first evaluated our algorithm using an artificial dataset assuming known densities where we could compare our result to the global solution and found that indeed our global objective approached the known optimum of zero (see Appendix E for more details). Next, we evaluated our DC-programming solution applied to real-world datasets: a sentiment analysis dataset (Blitzer et al., 2007) with the squared loss, a visual domain adaptation benchmark dataset Office (Saenko et al., 2010), as well as a generalization of digit recognition task, with the cross-entropy loss.

For all real-world datasets, the probability distributions \(D_k\) are not readily available to the learner. However, Corollary 4 extends the learning guarantees of our solution to the case where an estimate \(\hat{D}_k\) is used in lieu of the ideal distribution \(D_k\). Thus, we used standard density estimation methods to derive an estimate \(\hat{D}_k\) for each \(k \in [p]\). While density estimation can be a difficult task in general, for our purpose, straightforward techniques were sufficient for our predictor \(h^0_k\) to achieve a high performance, since the approximate densities only serve to indicate the relative importance of each source domain. We give full details about our density estimation procedure in Appendix E.

### 5.1 Sentiment analysis task with the squared loss

We used the sentiment analysis dataset proposed by Blitzer et al. (2007) and used for multiple-source adaptation by Mansour et al. (2008, 2009a). This dataset consists of product review text and rating labels taken from four domains: books (B), dvd (D), electronics (E), and kitchen (K), with 2,000 samples for each domain. We defined a vocabulary of 2,500 words that occur at least twice in the intersection of the four domains. These words were used to define feature vectors, where every sample was encoded by the number of occurrences of each word. We trained our base hypotheses using support vector regression with the same hyper-parameters as in (Mansour et al., 2008, 2009a).
We considered two real-world domain adaptation tasks: a generalization of a digit recognition task and a standard visual adaptation Office dataset.

For each individual domain, we trained a convolutional neural network (CNN) and used the output from the softmax score layer as our base predictors $h_k$. We computed the uniformly weighted combination of source predictors, $h_{\text{unif}} = \sum_{k=1}^{p} h_k / p$. As a privileged baseline, we also trained a model on all source data combined, $h_{\text{joint}}$. Note, this approach is often not feasible if independent entities contribute classifiers and densities, but not full training datasets. Thus this approach is not consistent with our scenario, and it operates in a much more favorable learning setting than our solution. Finally, our distribution weighted predictor $D\hat{w}$ was computed with $h_k$, $h_{\text{unif}}$, $h_{\text{joint}}$, and $D\hat{w}$. 

### 5.2 Recognition tasks with the cross-entropy loss

We considered two real-world domain adaptation tasks: a generalization of a digit recognition task and a standard visual adaptation Office dataset.

For each individual domain, we trained a convolutional neural network (CNN) and used the output from the softmax score layer as our base predictors $h_k$. We computed the uniformly weighted combination of source predictors, $h_{\text{unif}} = \sum_{k=1}^{p} h_k / p$. As a privileged baseline, we also trained a model on all source data combined, $h_{\text{joint}}$. Note, this approach is often not feasible if independent entities contribute classifiers and densities, but not full training datasets. Thus this approach is not consistent with our scenario, and it operates in a much more favorable learning setting than our solution. Finally, our distribution weighted predictor $D\hat{w}$ was computed with $h_k$, $h_{\text{unif}}$, $h_{\text{joint}}$, and $D\hat{w}$.
We presented practically applicable multiple-source domain adaptation algorithms for the squared loss and the cross-entropy loss. Our algorithms benefit from a series of very favorable theoretical guarantees. Our results further demonstrate empirically their effectiveness and their importance in adaptation problems in practice.

Acknowledgments

We thank Cyril Allauzen for comments on a previous draft of this paper. This work was partly funded by NSF CCF-1535987 and NSF IIS-1618662.
References


