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

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

any new mixture target domain? This is known as the *multiple-source adaption (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,

$$\begin{array}{llll} h_k \colon \mathcal{X} \to \mathbb{R} & & L \colon \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+ & (R) \\ h_k \colon \mathcal{X} \times \mathcal{Y} \to [0,1] & & L \colon [0,1] \to \mathbb{R}_+ & (P) \end{array}$$

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 $\mathcal{D}$:

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

Much of our theory only assumes that 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_ks? 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 $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$h_z^\eta(x) = \sum_{k=1}^p \frac{z_k \mathcal{D}_k^1(x) + \eta \frac{\mathcal{U}^1(x)}{p}}{\sum_{k=1}^p z_k \mathcal{D}_k^1(x) + \eta \mathcal{U}^1(x)} h_k(x), \qquad (R) \qquad (1)$$

$$h_z^\eta(x, y) = \sum_{k=1}^p \frac{z_k \mathcal{D}_k(x, y) + \eta \frac{\mathcal{U}(x,y)}{p}}{\sum_{j=1}^p z_j \mathcal{D}_j(x, y) + \eta \mathcal{U}(x, y)} h_k(x, y), \qquad (P) \qquad (2)$$

where we denote by $\mathcal{D}^1$ the marginal distribution over $\mathcal{X}$, for all $x \in \mathcal{X}$, $\mathcal{D}^1(x) = \sum_{y \in \mathcal{Y}} \mathcal{D}(x, y)$, and by $\mathcal{U}^1$ 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 η, 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 $\mathsf{d}_\alpha(\mathcal{D} \parallel \mathcal{D}') = e^{\mathsf{D}_\alpha(\mathcal{D} \| \mathcal{D}')}$ the exponential of the α-Rényi Divergence of two distributions $\mathcal{D}$ and $\mathcal{D}'$. See Appendix F for more details about the notion of Rényi 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 $\mathcal{D}_T^1 \in \mathcal{D}^1 = \{\sum_{k=1}^p \lambda_k \mathcal{D}_k^1 : \lambda \in \Delta\}$ 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 ϵ_T by

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

ϵ_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 $\mathcal{D}_k^1$, and the maximum taken over $k \in [p]$. When the target conditional is close to all source ones, α can be chosen to be very large and ϵ_T is close to ϵ. In particular, when the conditional probabilities coincide, for $\alpha = +\infty$, we have $\epsilon_T = \epsilon$.

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 $\mathcal{D}_T$ that is a mixture of source distributions:*

$$\mathcal{L}(\mathcal{D}_T, h_z^\eta) \leq \epsilon_T + \delta, \qquad (R)$$
$$\mathcal{L}(\mathcal{D}_T, h_z^\eta) \leq \epsilon + \delta. \qquad (P)$$

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 $\mathcal{D}_T$ whose conditional $\mathcal{D}_T(\cdot|x)$ is on average not too far away from $\mathcal{D}_k(\cdot|x)$ for any $k \in [p]$, and $\mathcal{D}_T^1 \in \mathcal{D}^1$, the loss of h_z^η on $\mathcal{D}_T$ is small. It is even more remarkable that, in the probability model (P), the loss of h_z^η is at most ϵ on any target distribution $\mathcal{D}_T \in \mathcal{D}$. Thus, h_z^η is a robust hypothesis with favorable property for any such target distribution $\mathcal{D}_T$.

We now present a more general result, Theorem 2, that relaxes the assumptions under the regression model that all possible target distributions $\mathcal{D}_T$ 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 $\mathcal{Q}(\cdot|x)$ defined for all $x \in \mathcal{X}$, define $\epsilon_\alpha(\mathcal{Q})$ as follows:

$$\epsilon_\alpha(\mathcal{Q}) = \max_{k \in [p]} \left[\mathop{\mathbb{E}}_{x \sim \mathcal{D}_k^1} \left[\mathsf{d}_\alpha \left(\mathcal{Q}(\cdot|x) \,\|\, \mathcal{D}_k(\cdot|x) \right)^{\alpha - 1} \right] \right]^{\frac{1}{\alpha}} \epsilon^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}.$$

Thus, $\epsilon_\alpha(\mathcal{Q})$ depends on the maximal expected α-Rényi divergence between $\mathcal{Q}(\cdot|x)$ and $\mathcal{D}_k(\cdot|x)$, and $\epsilon_\alpha(\mathcal{Q}) = \epsilon_T$ when $\mathcal{Q}(\cdot|x) = \mathcal{D}_T(\cdot|x)$. When there exists $\mathcal{Q}(\cdot|x)$ such that the expected α-Rényi divergence is small for all $k \in [p]$, then $\epsilon_\alpha(\mathcal{Q})$ is close to ϵ for $\alpha = +\infty$. In addition, we will use the following definitions: $\mathcal{D}_{k,\mathcal{Q}}(x, y) = \mathcal{D}_k^1(x)\mathcal{Q}(y|x)$ and $\mathcal{D}_{P,\mathcal{Q}} = \left\{ \sum_{k=1}^p \lambda_k \mathcal{D}_{k,\mathcal{Q}} : \lambda \in \Delta \right\}$.

Theorem 2 (Regression model)**.** *Fix a conditional probability distribution $\mathcal{Q}(\cdot|x)$ defined for all $x \in \mathcal{X}$. 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 $\mathcal{D}_T$:*

$$\mathcal{L}(\mathcal{D}_T, h_z^\eta) \leq \left[\left(\epsilon_\alpha(\mathcal{Q}) + \delta \right) \mathsf{d}_\beta(\mathcal{D}_T \,\|\, \mathcal{D}_{P,\mathcal{Q}}) \right]^{\frac{\beta-1}{\beta}} M^{\frac{1}{\beta}}.$$

The learning guarantee of Theorem 2 depends on the Rényi divergence between the conditional probabilities of the source and target domains and a fixed *pivot* $\mathcal{Q}(\cdot|x)$. In particular, when there exists a pivot $\mathcal{Q}(\cdot|x)$ that is close to $\mathcal{D}_T(\cdot|x)$ and $\mathcal{D}_k(\cdot|x)$, for all $k \in [p]$, then the guarantee is significant. One candidate for such a pivot is a conditional probability distribution $\mathcal{Q}(\cdot|x)$ minimizing $\epsilon_\alpha(\mathcal{Q})$.

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 $\mathcal{D}_T(\cdot|x) = \mathcal{D}_k(\cdot|x)$, $\forall k \in [p]$ and therefore $\mathsf{d}_\alpha \left(\mathcal{D}_T(\cdot|x) \,\|\, \mathcal{D}_k(\cdot|x) \right) = 1$. Setting $\alpha = +\infty$, we recover the main result of Mansour et al. (2008).

Corollary 3. *Assume that the conditional probability distributions $\mathcal{D}_k(\cdot|x)$ do not depend on k. Then, for any $\delta > 0$, there exist $\eta > 0$ and $z \in \Delta$ such that $\mathcal{L}(\mathcal{D}_\lambda, h_z^\eta) \leq \epsilon + \delta$ 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_z^η is at most $\epsilon + \delta$, that is arbitrarily close to ϵ. h_z^η 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 $\mathcal{D}_k$ are not directly available to the learner, and instead estimates $\widehat{\mathcal{D}}_k$ have been derived

from data, and further to the case where the target distribution $\mathcal{D}_T$ is not a mixture of source distributions. We will denote by $\widehat{h}_z^\eta$ the distribution-weighted combination rule based on the estimates $\widehat{\mathcal{D}}_k$. Our learning guarantee for $\widehat{h}_z^\eta$ depends on the Rényi divergence of $\widehat{\mathcal{D}}_k$ and $\mathcal{D}_k$, as well as the Rényi divergence of $\mathcal{D}_T$ and the family of mixtures of source distributions.

Corollary 4. *For any $\delta > 0$, there exist $\eta > 0$ and $z \in \Delta$, such that the following inequality holds for any $\alpha > 1$ and arbitrary target distribution $\mathcal{D}_T$:*

$$\mathcal{L}(\mathcal{D}_T, \widehat{h}_z^\eta) \leq \left[(\widehat{\epsilon} + \delta)\, \mathsf{d}_\alpha(\mathcal{D}_T \parallel \widehat{\mathcal{D}})\right]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}},$$

where $\widehat{\epsilon} = \max_{k \in [p]} \left[\epsilon\, \mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)\right]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}$, and $\widehat{\mathcal{D}} = \left\{\sum_{k=1}^p \lambda_k \widehat{\mathcal{D}}_k : \lambda \in \Delta\right\}$.

Corollary 4 shows that there exists a predictor $\widehat{h}_z^\eta$ based on the estimate distributions $\widehat{\mathcal{D}}_k$ that is $\widehat{\epsilon}$-accurate with respect to any target distribution $\mathcal{D}_T$ whose Rényi divergence with respect to the family $\widehat{\mathcal{D}}$ is not too large ($\mathsf{d}_\alpha(\mathcal{D}_T \parallel \widehat{\mathcal{D}})$ close to 1). Furthermore, $\widehat{\epsilon}$ is close to ϵ, provided that $\widehat{\mathcal{D}}_k$s are good estimates of $\mathcal{D}_k$s (that is $\mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)$ close to 1). The proof is given in Appendix B.

3.2 Guarantees for the probability model with the cross-entropy loss

Here, we discuss the important special case where L coincides with the cross-entropy loss in the probability model, and present a guarantee for a normalized distribution-weighted combination solution. This analysis is a complement to Theorem 1, which only holds for the unnormalized hypothesis $h_z^\eta(x, y)$.

The cross-entropy loss assumes normalized hypotheses. Thus, here, we assume that the source functions are normalized for every x: $\sum_{y \in \mathcal{Y}} h_k(x, y) = 1,\ \forall x \in \mathcal{X}, \forall k \in [p]$. For any $\eta > 0$ and $z \in \Delta$, we define a normalized weighted combination $\overline{h}_z^\eta(x, y)$ that is based on distribution-weighted combination $h_z^\eta(x, y)$ defined by (2):

$$\overline{h}_z^\eta(x, y) = \frac{h_z^\eta(x, y)}{\sum_{y \in \mathcal{Y}} h_z^\eta(x, y)}.$$

We will first assume the conditional probability distributions $\mathcal{D}_k(\cdot|x)$ do not depend on k.

Theorem 5. *Assume that there exists $\mu > 0$ such that $\mathcal{D}_k(x, y) \geq \mu \mathcal{U}(x, y)$ for all $k \in [p]$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Then, for any $\delta > 0$, there exist $\eta > 0$ and $z \in \Delta$ such that $\mathcal{L}(\mathcal{D}_\lambda, \overline{h}_z^\eta) \leq \epsilon + \delta$ for any mixture parameter $\lambda \in \Delta$.*

Theorem 5 provides a strong guarantee that is the analogue of Corollary 3 for normalized distribution-weighted combinations. The theorem can also be extended to the case of arbitrary target distributions and estimated densities. When the conditional probabilities are distinct across the source domains, we propose a marginal distribution-weighted combination rule, which is already normalized. We can directly apply Theorem 1 to that solution and achieve favorable guarantees. More details are presented in Appendix C.

These results are non-trivial and important, as they provide a guarantee for an accurate and robust predictor for a commonly used loss function, the cross-entropy loss.

4 Algorithms

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^η 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^η 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^η:*

$$\forall k \in [p], \quad \mathcal{L}(\mathcal{D}_k, h_z^\eta) \leq \sum_{j=1}^{p} z_j \mathcal{L}(\mathcal{D}_j, h_z^\eta) + \eta'. \tag{3}$$

Lemma 6 indicates that for the solution z, h_z^η 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: $\min_{z\in\Delta} \max_{k\in[p]} \mathcal{L}(\mathcal{D}_k, h_z^\eta) - \mathcal{L}(\mathcal{D}_z, h_z^\eta)$, which can be equivalently formulated as the following optimization problem:

$$\min_{z\in\Delta, \gamma\in\mathbb{R}} \gamma \quad \text{s.t. } \mathcal{L}(\mathcal{D}_k, h_z^\eta) - \mathcal{L}(\mathcal{D}_z, h_z^\eta) \leq \gamma, \ \forall k \in [p]. \tag{4}$$

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^η 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:

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

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

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

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

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

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

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

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

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):

$$\min_{z\in\Delta, \gamma\in\mathbb{R}} \gamma \quad \text{s.t.} \big(u_k(z) - v_k(z) \leq \gamma\big) \wedge \big(-z_k \leq 0\big) \wedge \big(\textstyle\sum_{k=1}^p z_k - 1 = 0\big), \quad \forall k \in [p]. \tag{5}$$

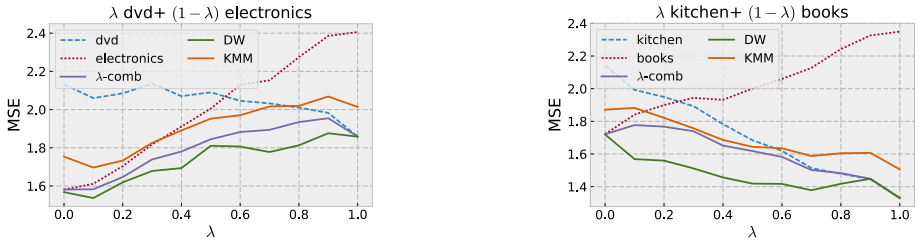


Figure 1: MSE sentiment analysis under mixture of two domains: (a) (*left figure*) `dvd` and `electronics`; (b) (*right figure*) `kitchen` and `books`.

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

$$z_{t+1} \in \operatorname*{argmin}_{z,\gamma\in\mathbb{R}} \gamma \tag{6}$$
$$\text{s.t. } \big(u_k(z) - v_k(z_t) - (z - z_t)\nabla v_k(z_t) \le \gamma\big) \wedge \big(-z_k \le 0\big) \wedge \big(\textstyle\sum_{k=1}^p z_k - 1 = 0\big), \quad \forall k \in [p],$$

where $z_0 \in \Delta$ is an arbitrary starting value. Then, $(z_t)_t$ 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 $\mathcal{L}(\mathcal{D}_k, h_z^\eta) - \mathcal{L}(\mathcal{D}_z, h_z^\eta) \le \gamma$, for all $k \in [p]$ for an arbitrarily small value of γ. Since $\mathcal{L}(\mathcal{D}_z, h_z^\eta) = \sum_{k=1}^p z_k \mathcal{L}(\mathcal{D}_k, h_z^\eta)$ is a weighted average of the expected losses $\mathcal{L}(\mathcal{D}_k, h_z^\eta)$, $k \in [p]$, the solution γ cannot be negative. Furthermore, by Lemma 6, a parameter z verifying that inequality exists for any $\gamma > 0$. Thus, the global solution γ 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 $\mathcal{D}_k$ are not readily available to the learner. However, Corollary 4 extends the learning guarantees of our solution to the case where an estimate $\widehat{\mathcal{D}}_k$ is used in lieu of the ideal distribution $\mathcal{D}_k$. Thus, we used standard density estimation methods to derive an estimate $\widehat{\mathcal{D}}_k$ 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 $\widehat{h}_z^\eta$ 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).

Table 1: MSE on the sentiment analysis dataset of source-only baselines for each domain, K,D, B,E, the uniform weighted predictor unif, KMM, and the distribution-weighted method DW based on the learned z. DW outperforms all competing baselines.

	Test Data										
	K	D	B	E	KD	BE	DBE	KBE	KDB	KDB	KDBE
K	1.46±0.08	2.20±0.14	2.29±0.13	1.69±0.12	1.83±0.08	1.99±0.10	2.06±0.07	1.81±0.07	1.78±0.07	1.98±0.06	1.91±0.06
D	2.12±0.08	1.78±0.08	2.12±0.08	2.10±0.07	1.95±0.07	2.11±0.07	2.00±0.06	2.11±0.06	2.00±0.06	2.01±0.06	2.03±0.06
B	2.18±0.11	2.01±0.09	1.73±0.12	2.24±0.07	2.10±0.09	1.99±0.08	1.99±0.05	2.05±0.06	2.14±0.06	1.98±0.06	2.04±0.05
E	1.69±0.09	2.31±0.12	2.40±0.11	1.50±0.06	2.00±0.09	1.95±0.07	2.07±0.06	1.86±0.04	1.84±0.06	2.14±0.06	1.98±0.05
unif	1.62±0.05	1.84±0.09	1.86±0.09	1.62±0.07	1.73±0.06	1.74±0.07	1.77±0.05	1.70±0.05	1.69±0.04	1.77±0.04	1.74±0.04
KMM	1.63±0.15	2.07±0.12	1.93±0.17	1.69±0.12	1.83±0.07	1.82±0.07	1.89±0.07	1.75±0.07	1.78±0.06	1.86±0.09	1.82±0.06
DW(ours)	**1.45±0.08**	**1.78±0.08**	**1.72±0.12**	**1.49±0.06**	**1.62±0.07**	**1.61±0.08**	**1.66±0.05**	**1.56±0.04**	**1.58±0.05**	**1.65±0.04**	**1.61±0.04**

Table 2: **Digit** dataset statistics.

	SVHN	MNIST	USPS
			
# train images	73,257	60,000	7,291
# test images	26,032	10,000	2,007
image size	32x32	28x28	16x16
color	rgb	gray	gray

Table 3: **Digit** dataset accuracy.

	Test Data							
	svhn	mnist	usps	mu	su	sm	smu	mean
CNN-s	**92.3**	66.9	65.6	66.7	90.4	85.2	84.2	78.8
CNN-m	15.7	**99.2**	79.7	96.0	20.3	38.9	41.0	55.8
CNN-u	16.7	62.3	**96.6**	68.1	22.5	29.4	32.9	46.9
CNN-unif	75.7	91.3	92.2	91.4	76.9	80.0	80.7	84.0
DW (ours)	91.4	98.8	95.6	98.3	**91.7**	**93.5**	**93.6**	**94.7**
CNN-joint	90.9	99.1	96.0	**98.6**	91.3	93.2	93.3	94.6

We compared our method (DW) against each source hypothesis, h_k. We also computed a privileged baseline using the oracle λ mixing parameter, λ-comb: $\sum_{k=1}^{p} \lambda_k h_k$. λ-comb is of course not accessible in practice since the target mixture λ is not known to the user. We also compared against a previously proposed domain adaptation algorithm (Huang et al., 2006) known as KMM. It is important to note that the KMM model requires access to the unlabeled target data during adaptation and learns a new predictor for every target domain, while DW does not use any target data. Thus KMM operates in a favorable learning setting when compared to our solution.

We first considered the same test scenario as in (Mansour et al., 2008), where the target is a mixture of two source domains. The plots of Figures 1a and 1b report the results of our experiments. They show that our distribution-weighted predictor DW outperforms all baseline predictors despite the privileged learning scenarios of λ-comb and KMM. We also compared our results with the *weighted predictor* used in the empirical studies by Mansour et al. (2008), which is not a realistic solution since it is using the unknown target mixture λ as z to compute h_z. Nevertheless, we observed that the performance of this "cheating" solution almost always coincides with that of our DW algorithm and thus did not include it in our plots and tables to avoid confusion.

Next, we compared the performance of DW with accessible baseline predictors on various target mixtures. Since λ is not known in practice, we replaced λ-comb with the uniform combination of all hypotheses (unif), $\sum_{k=1}^{p} h_k/p$. Table 1 reports the mean and standard deviations of MSE over 10 repetitions. Each column corresponds to a different target test data source. Our distribution-weighted method DW outperforms all baseline predictors across all test domains. Observe that, even when the target is a single source domain, our method successfully outperforms the predictor which is trained and tested on the same domain. Results on more target mixtures are available in Appendix E.

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, $h_{\texttt{unif}} = \sum_{k=1}^{p} h_k/p$. As a privileged baseline, we also trained a model on all source data combined, $h_{\texttt{joint}}$. 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 DW was computed with h_ks, density estimates, and our learned weighting, z. Our baselines then consists of the classifiers from h_k, $h_{\texttt{unif}}$, $h_{\texttt{joint}}$, and DW.

Table 4: ***Office*** dataset accuracy: We report accuracy across six possible test domains. We show performance all baselines: CNN-`a,w,d`, CNN-`unif`, `DW` based on the learned z, and the jointly trained model CNN-`joint`. `DW` outperforms all competing models.

	Test Data							
	`amazon`	`webcam`	`dslr`	`aw`	`ad`	`wd`	`awd`	mean
CNN-`a`	**75.7 ± 0.3**	53.8 ± 0.7	53.4 ± 1.3	71.4 ± 0.3	73.5 ± 0.2	53.6 ± 0.8	69.9 ± 0.3	64.5 ± 0.6
CNN-`w`	45.3 ± 0.5	91.1 ± 0.8	91.7 ± 1.2	54.4 ± 0.5	50.0 ± 0.5	91.3 ± 0.8	57.5 ± 0.4	68.8 ± 0.7
CNN-`d`	50.4 ± 0.4	89.6 ± 0.9	90.9 ± 0.8	58.3 ± 0.4	54.6 ± 0.4	90.0 ± 0.7	61.0 ± 0.4	70.7 ± 0.6
CNN-`unif`	69.7 ± 0.3	93.1 ± 0.6	93.2 ± 0.9	74.4 ± 0.4	72.1 ± 0.3	93.1 ± 0.5	75.9 ± 0.3	81.6 ± 0.5
`DW` (ours)	75.2 ± 0.4	**93.7 ± 0.6**	**94.0 ± 1.0**	**78.9 ± 0.4**	**77.2 ± 0.4**	**93.8 ± 0.6**	**80.2 ± 0.3**	**84.7 ± 0.5**
CNN-`joint`	72.1 ± 0.3	93.7 ± 0.5	93.7 ± 0.5	76.4 ± 0.4	76.4 ± 0.4	93.7 ± 0.5	79.3 ± 0.4	83.6 ± 0.4

We began our study with a generalization of digit recognition task, which consists of three digit recognition datasets: Google Street View House Numbers (SVHN), MNIST, and USPS. Dataset statistics as well as example images can be found in Table 2. We trained the ConvNet (or CNN) architecture following Taigman et al. (2017) as our source models and joint model. We used the second fully-connected layer's output as our features for density estimation, and the output from the softmax score layer as our predictors. We used the full training sets per domain to learn the source model and densities. Note, these steps are completely isolated from one another and may be performed by unique entities and in parallel. Finally, for our DC-programming algorithm we used a small subset of 200 real image-label pairs from each domain to learn the parameter z.

Our next experiment used the standard visual adaptation *Office* dataset, which has 3 domains: `amazon`, `webcam`, and `dslr`. The dataset contains 31 recognition categories of objects commonly found in an office environment. There are 4,110 images total with 2,817 from `amazon`, 795 from `webcam`, and 498 from `dslr`.

We followed the standard protocol from Saenko et al. (2010), whereby 20 labeled examples are available for training from the `amazon` domain and 8 labeled examples are available from both the `webcam` and `dslr` domains. The remaining examples from each domain are used for testing. We used the AlexNet Krizhevsky et al. (2012) ConvNet (CNN) architecture, and used the output from the softmax score layer as our base predictors, pre-trained on ImageNet and used fc7 activations as our features for density estimation Donahue et al. (2014).

We report the performance of our algorithm and that of baselines on the digit recognition dataset in Table 3, and report the performance on the *Office* dataset in Table 4. On both datasets, we evaluated on various test distributions: each individual domain, the combination of each two domains and the fully combined set. When the test distribution equals one of the source distributions, our distribution-weighted classifier successfully outperforms (`webcam,dslr`) or maintains the performance of the classifier which is trained and tested on the same domain. For the more realistic scenario where the target domain is a mixture of any two or all three source domains, the performance of our method is comparable or marginally superior to that of the jointly trained network, despite the fact that we do not retrain any network parameters in our method and that we only use a small number of per-domain examples to learn the distribution weights – an optimization which may be solved on a single CPU in a matter of seconds for this problem. This again demonstrates the robustness of our distribution-weighted combined classifier to a varying target domain.

6 Conclusion

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

C. Arndt. *Information Measures: Information and its Description in Science and Engineering*. Signals and Communication Technology. Springer Verlag, 2004.

S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *NIPS*, pages 137–144, 2006.

G. Blanchard, G. Lee, and C. Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *NIPS*, pages 2178–2186, 2011.

J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, pages 440–447, 2007.

C. Cortes and M. Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theor. Comput. Sci.*, 519:103–126, 2014.

C. Cortes, M. Mohri, and A. Muñoz Medina. Adaptation algorithm and theory based on generalized discrepancy. In *KDD*, pages 169–178, 2015.

T. M. Cover and J. M. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006.

K. Crammer, M. J. Kearns, and J. Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9(Aug):1757–1774, 2008.

J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, volume 32, pages 647–655, 2014.

M. Dredze, K. Crammer, and F. Pereira. Confidence-weighted linear classification. In *ICML*, volume 307, pages 264–271, 2008.

L. Duan, I. W. Tsang, D. Xu, and T. Chua. Domain adaptation from multiple sources via auxiliary classifiers. In *ICML*, volume 382, pages 289–296, 2009.

L. Duan, D. Xu, and I. W. Tsang. Domain adaptation from multiple sources: A domain-dependent regularization approach. *IEEE Transactions on Neural Networks and Learning Systems*, 23(3): 504–518, 2012.

Y. Ganin and V. S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, volume 37, pages 1180–1189, 2015.

R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014.

B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, pages 2066–2073, 2012.

B. Gong, K. Grauman, and F. Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *ICML*, volume 28, pages 222–230, 2013a.

B. Gong, K. Grauman, and F. Sha. Reshaping visual datasets for domain adaptation. In *NIPS*, pages 1286–1294, 2013b.

J. Hoffman, B. Kulis, T. Darrell, and K. Saenko. Discovering latent domains for multisource domain adaptation. In *ECCV*, volume 7573, pages 702–715, 2012.

J. Hoffman, E. Rodner, J. Donahue, K. Saenko, and T. Darrell. Efficient learning of domain-invariant image representations. In *ICLR*, 2013.

J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In *NIPS*, pages 601–608, 2006.

A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba. Undoing the damage of dataset bias. In *ECCV*, volume 7572, pages 158–171, 2012.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.

H. Liao. Speaker adaptation of context dependent deep neural networks. In *ICASSP*, pages 7947–7951, 2013.

M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, volume 37, pages 97–105, 2015.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation with multiple sources. In *NIPS*, pages 1041–1048, 2008.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Multiple source adaptation and the Rényi divergence. In *UAI*, pages 367–374, 2009a.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT*, 2009b.

K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *ICML*, volume 28, pages 10–18, 2013.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10): 1345–1359, 2010.

Z. Pei, Z. Cao, M. Long, and J. Wang. Multi-adversarial domain adaptation. In *AAAI*, pages 3934–3941, 2018.

A. Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 547–561, 1961.

B. Roark, R. Sproat, C. Allauzen, M. Riley, J. Sorensen, and T. Tai. The opengrm open-source finite-state grammar software libraries. In *ACL (System Demonstrations)*, pages 61–66, 2012.

K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, volume 6314, pages 213–226, 2010.

B. K. Sriperumbudur and G. R. G. Lanckriet. A proof of convergence of the concave-convex procedure using Zangwill's theory. *Neural Computation*, 24(6):1391–1407, 2012.

Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. In *ICLR*, 2017.

P. D. Tao and L. T. H. An. Convex analysis approach to DC programming: theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22(1):289–355, 1997.

P. D. Tao and L. T. H. An. A DC optimization algorithm for solving the trust-region subproblem. *SIAM Journal on Optimization*, 8(2):476–505, 1998.

A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR*, pages 1521–1528, 2011.

E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *ICCV*, pages 4068–4076, 2015.

Z. Xu, W. Li, L. Niu, and D. Xu. Exploiting low-rank structure from latent domains for domain generalization. In *ECCV*, volume 8691, pages 628–643, 2014.

J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive svms. In *ACM Multimedia*, pages 188–197, 2007.

A. L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4):915–936, 2003.

K. Zhang, M. Gong, and B. Schölkopf. Multi-source domain adaptation: A causal view. In *AAAI*, pages 3150–3157, 2015.

A Lower bounds for convex combination rules

In this section, we give lower bounds for convex combination rule, for both squared loss and cross-entropy loss. For any $\alpha \in \Delta$, we define the convex combination rule for the regression and the probability model as follows:

$$g_\alpha(x) = \sum_{k=1}^{p} \alpha_k h_k(x), \qquad (R) \tag{7}$$

$$g_\alpha(x,y) = \sum_{k=1}^{p} \alpha_k h_k(x,y). \qquad (P) \tag{8}$$

Lemma 9 (*Regression model*, squared loss)**.** *There is a mixture adaptation problem for which the expected squared loss of g_α is $\frac{1}{4}$.*

Proof. Let $\mathcal{X} = \{a, b\}$, and $\mathcal{Y} = \{0, 1\}$. Consider $\mathcal{D}_0(x,y) = 1_{x=a,y=0}$, $h_0(x) = 0$, and $\mathcal{D}_1(x,y) = 1_{x=b,y=1}$, $h_1(x) = 1$. Consider the target distribution $\mathcal{D}_T = \frac{1}{2}\mathcal{D}_0 + \frac{1}{2}\mathcal{D}_1$. Then, for any convex combination rule $g_\alpha = \alpha h_0 + (1-\alpha)h_1 = 1 - \alpha$,

$$\left(\frac{1}{2}\right)^2 = \left(\frac{1}{2}\alpha + \frac{1}{2}(1-\alpha)\right)^2 = \left(\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \mathcal{D}_T(x,y)|g_\alpha(x) - y|\right)^2$$
$$\leq \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \mathcal{D}_T(x,y)(g_\alpha(x) - y)^2 = \mathcal{L}(\mathcal{D}_T, g_\alpha).$$

□

Note that the hypotheses h_0 and h_1 have *zero* error on their own domain, i.e. $\epsilon = 0$. However, *no* convex combination rule will perform well on the target distribution $\mathcal{D}_T$.

Lemma 10 (*Probability Model*, cross-entropy loss)**.** *There is a mixture adaptation problem for which the expected cross-entropy loss of g_α is* $\log(p)$.

Proof. Let $\mathcal{X} = \{x_1, \ldots, x_k\}$, and $\mathcal{Y} = \{y_1, \ldots, y_k\}$. Consider $\mathcal{D}_k(x,y) = 1_{x=x_k,y=y_k}$, and $h_k(x,y) = 1_{y=y_k}$. Consider the largest cross-entropy loss of g_α on any target mixture $\mathcal{D}_\lambda(x,y)$:

$$\max_{\lambda\in\Delta} \mathcal{L}(\mathcal{D}_\lambda, g_\alpha) = \max_{\lambda\in\Delta} \sum_{k=1}^{p} -\lambda_k \log(\alpha_k) = \max_{k\in[p]} \left[-\log(\alpha_k)\right].$$

Choosing $\alpha \in \Delta$ to minimize that adversarial loss gives

$$\min_{\alpha\in\Delta} \max_{k\in[p]} \left[-\log(\alpha_k)\right] = \log(p).$$

Therefore any convex combination rule g_α incurs at least a loss of $\log(p)$. □

Again, the base hypotheses h_ks have *zero* error on their own domain, yet there is no convex combination rule that is robust against arbitrary target mixture.

B Theoretical analysis for the stochastic scenario

In this section, we give a series of theoretical results for the general stochastic scenario with their full proofs. We will separate the proofs for the regression model (Appendix B.1) and the probability model (Appendix B.3), since the definitions of the distribution weighted combination are different in the two models.

B.1 Regression model

The proofs for the regression model (R) are presented in the following order: we first assume the conditional probabilities are the same across source domains, and prove Lemma 6; using that, we prove Corollary 3 and Corollary 4. Finally, we relax the assumption of same conditionals, and prove Theorem 2, which is a stronger version of Theorem 1.

Our proofs make use of the following Fixed-Point Theorem of Brouwer.

Theorem 11. *For any compact and convex non-empty set $C \subset \mathbb{R}^p$ and any continuous function $f: C \to C$, there is a point $x \in C$ such that $f(x) = x$.*

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^η:*

$$\forall k \in [p], \quad \mathcal{L}(\mathcal{D}_k, h_z^\eta) \leq \sum_{j=1}^{p} z_j \mathcal{L}(\mathcal{D}_j, h_z^\eta) + \eta'. \tag{9}$$

Proof. Consider the mapping $\Phi: \Delta \to \Delta$ defined for all $z \in \Delta$ by

$$[\Phi(z)]_k = \frac{z_k \, \mathcal{L}(\mathcal{D}_k, h_z^\eta) + \frac{\eta'}{p}}{\sum_{j=1}^{p} z_j \mathcal{L}(\mathcal{D}_j, h_z^\eta) + \eta'}.$$

Φ is continuous since $\mathcal{L}(\mathcal{D}_k, h_z^\eta)$ is a continuous function of z and since the denominator is positive ($\eta' > 0$). Thus, by Brouwer's Fixed Point Theorem, there exists $z \in \Delta$ such that $\Phi(z) = z$. For that z, we can write

$$z_k = \frac{z_k \, \mathcal{L}(\mathcal{D}_k, h_z^\eta) + \frac{\eta'}{p}}{\sum_{j=1}^{p} z_j \mathcal{L}(\mathcal{D}_j, h_z^\eta) + \eta'},$$

for all $k \in [p]$. Since η' is positive, we must have $z_k \neq 0$ for all k. Dividing both sides by z_k gives $\mathcal{L}(\mathcal{D}_k, h_z^\eta) = \sum_{j=1}^{p} z_j \mathcal{L}(\mathcal{D}_j, h_z^\eta) + \eta' - \frac{\eta'}{p z_k} \leq \sum_{j=1}^{p} z_j \mathcal{L}(\mathcal{D}_j, h_z^\eta) + \eta'$, which completes the proof. □

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

Proof. We first upper bound, for an arbitrary $z \in \Delta$, the expected loss of h_z^η with respect to the mixture distribution $\mathcal{D}_z$ defined using the same z, that is $\mathcal{L}(\mathcal{D}_z, h_z^\eta) = \sum_{k=1}^{p} z_k \mathcal{L}(\mathcal{D}_k, h_z^\eta)$. By definition of h_z^η and $\mathcal{D}_z$, we can write

$$\begin{aligned}
\mathcal{L}(\mathcal{D}_z, h_z^\eta) &= \sum_{(x,y)} \mathcal{D}_z(x,y) L(h_z^\eta(x), y) \\
&= \sum_{(x,y)} \mathcal{D}_z(x,y) L\left(\sum_{k=1}^{p} \frac{z_k \mathcal{D}_k^1(x) + \eta \frac{\mathcal{U}^1(x)}{p}}{\mathcal{D}_z^1(x) + \eta \mathcal{U}^1(x)} h_k(x), y \right).
\end{aligned}$$

By convexity of L, this implies that

$$\begin{aligned}
\mathcal{L}(\mathcal{D}_z, h_z^\eta) &\le \sum_{(x,y)} \mathcal{D}_z(x,y) \sum_{k=1}^p \frac{z_k \mathcal{D}_k^1(x) + \eta \frac{\mathcal{U}^1(x)}{p}}{\mathcal{D}_z^1(x) + \eta \mathcal{U}^1(x)} L\big(h_k(x), y\big) \\
&\le \sum_{(x,y)} \mathcal{D}_z(y|x) \mathcal{D}_z^1(x) \sum_{k=1}^p \frac{z_k \mathcal{D}_k^1(x) + \eta \frac{\mathcal{U}^1(x)}{p}}{\mathcal{D}_z^1(x) + \eta \mathcal{U}^1(x)} L\big(h_k(x), y\big) \\
&\le \sum_{(x,y)} \mathcal{D}_z(y|x) \sum_{k=1}^p \left(z_k \mathcal{D}_k^1(x) + \eta \frac{\mathcal{U}^1(x)}{p} \right) L\big(h_k(x), y\big).
\end{aligned}$$

Next, observe that $\mathcal{D}_z(y|x) = \sum_{k=1}^p \frac{z_k \mathcal{D}_k^1(x)}{\mathcal{D}_z^1(x)} \mathcal{D}_k(y|x) = \mathcal{D}_k(y|x)$ for any $k \in [p]$ since by assumption $\mathcal{D}_k(y|x)$ does not depend on k. Thus,

$$\begin{aligned}
\mathcal{L}(\mathcal{D}_z, h_z^\eta) &\le \sum_{(x,y)} \mathcal{D}_z(y|x) \sum_{k=1}^p \left(z_k \mathcal{D}_k^1(x) + \eta \frac{\mathcal{U}^1(x)}{p} \right) L\big(h_k(x), y\big) \\
&= \sum_{(x,y)} \sum_{k=1}^p \left(z_k \mathcal{D}_k(x,y) + \eta \mathcal{D}_k(y|x) \frac{\mathcal{U}^1(x)}{p} \right) L\big(h_k(x), y\big) \\
&= \sum_{k=1}^p z_k \mathcal{L}(\mathcal{D}_k, h_k) + \frac{\eta}{p} \sum_{k=1}^p \sum_{(x,y)} \mathcal{D}_k(y|x) \mathcal{U}^1(x) L\big(h_k(x), y\big) \\
&\le \sum_{k=1}^p z_k \mathcal{L}(\mathcal{D}_k, h_k) + \eta M \le \sum_{k=1}^p z_k \epsilon + \eta M = \epsilon + \eta M.
\end{aligned}$$

Now, choose $z \in \Delta$ as in the statement of Lemma 6. Then, the following holds for any mixture distribution $\mathcal{D}_\lambda$:

$$\begin{aligned}
\mathcal{L}(\mathcal{D}_\lambda, h_z^\eta) &= \sum_{k=1}^p \lambda_k \mathcal{L}(\mathcal{D}_k, h_z^\eta) \le \sum_{k=1}^p \lambda_k \big(\mathcal{L}(\mathcal{D}_z, h_z^\eta) + \eta' \big) \\
&= \mathcal{L}(\mathcal{D}_z, h_z^\eta) + \eta' \le \epsilon + \eta M + \eta'.
\end{aligned}$$

Setting $\eta = \frac{\delta}{2M}$ and $\eta' = \frac{\delta}{2}$ concludes the proof. □

Next, we extend to the case where the target distribution is arbitrary, that is, the target distribution is not necessarily a mixture of source distributions.

Corollary 12. *For any $\delta > 0$, there exist $\eta > 0$ and $z \in \Delta$, such that the following inequality holds for any $\alpha > 1$ and arbitrary target distribution $\mathcal{D}_T$:*

$$\mathcal{L}(\mathcal{D}_T, h_z^\eta) \le \Big[(\epsilon + \delta)\, \mathsf{d}_\alpha(\mathcal{D}_T \,\|\, \mathcal{D}) \Big]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}.$$

Proof. For any hypothesis $h: \mathcal{X} \to \mathcal{Y}$ and any distribution $\mathcal{D}$, by Hölder's inequality, the following holds:

$$\begin{aligned}
\mathcal{L}(\mathcal{D}_T, h) &= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \mathcal{D}_T(x,y) L(h(x), y) \\
&= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \left[\frac{\mathcal{D}_T(x,y)}{\mathcal{D}(x,y)^{\frac{\alpha-1}{\alpha}}} \right] \left[\mathcal{D}(x,y)^{\frac{\alpha-1}{\alpha}} L(h(x), y) \right] \\
&\le \left[\sum_{(x,y)} \frac{\mathcal{D}_T(x,y)^\alpha}{\mathcal{D}(x,y)^{\alpha-1}} \right]^{\frac{1}{\alpha}} \left[\sum_{(x,y)} \mathcal{D}(x,y) L(h(x), y)^{\frac{\alpha}{\alpha-1}} \right]^{\frac{\alpha-1}{\alpha}}.
\end{aligned}$$

Thus, by definition of d_α, for any h such that $L(h(x), y) \leq M$ for all (x, y), we can write

$$\begin{aligned}
\mathcal{L}(\mathcal{D}_T, h) &\leq \mathsf{d}_\alpha(\mathcal{D}_T \parallel \mathcal{D})^{\frac{\alpha-1}{\alpha}} \left[\sum_{(x,y)} \mathcal{D}(x,y) L(h(x), y)^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}} \\
&= \mathsf{d}_\alpha(\mathcal{D}_T \parallel \mathcal{D})^{\frac{\alpha-1}{\alpha}} \left[\sum_{(x,y)} \mathcal{D}(x,y) L(h(x), y) L(h(x), y)^{\frac{1}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}} \\
&\leq \mathsf{d}_\alpha(\mathcal{D}_T \parallel \mathcal{D})^{\frac{\alpha-1}{\alpha}} \left[\sum_{(x,y)} \mathcal{D}(x,y) L(h(x), y) M^{\frac{1}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}} \\
&\leq \left[\mathsf{d}_\alpha(\mathcal{D}_T \parallel \mathcal{D})\, \mathcal{L}(\mathcal{D}, h)\right]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}.
\end{aligned}$$

Now, by Corollary 3, there exist $z \in \Delta$ and $\eta > 0$ such that $\mathcal{L}(\mathcal{D}, h_z^\eta) \leq \epsilon + \delta$ for any mixture distribution $\mathcal{D} \in \mathcal{D}$. Thus, in view of the previous inequality, we can write,for any $\mathcal{D} \in \mathcal{D}$,

$$\mathcal{L}(\mathcal{D}_T, h_z^\eta) \leq \left[(\epsilon + \delta)\, \mathsf{d}_\alpha(\mathcal{D}_T \parallel \mathcal{D})\right]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}.$$

Taking the infimum of the right-hand side over all $\mathcal{D} \in \mathcal{D}$ completes the proof. □

Corollary 4. *For any $\delta > 0$, there exist $\eta > 0$ and $z \in \Delta$, such that the following inequality holds for any $\alpha > 1$ and arbitrary target distribution $\mathcal{D}_T$:*

$$\mathcal{L}(\mathcal{D}_T, \widehat{h}_z^\eta) \leq \left[(\widehat{\epsilon} + \delta)\, \mathsf{d}_\alpha(\mathcal{D}_T \parallel \widehat{\mathcal{D}})\right]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}},$$

where $\widehat{\epsilon} = \max_{k \in [p]} \left[\epsilon\, \mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)\right]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}$, and $\widehat{\mathcal{D}} = \left\{\sum_{k=1}^p \lambda_k \widehat{\mathcal{D}}_k : \lambda \in \Delta\right\}$.

Proof. By the first part of the proof of Corollary 12, for any $k \in [p]$ and $\alpha > 1$, the following inequality holds:

$$\begin{aligned}
\mathcal{L}(\widehat{\mathcal{D}}_k, h_k) &\leq \left[\mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)\, \mathcal{L}(\mathcal{D}_k, h_k)\right]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}} \\
&\leq \left[\epsilon\, \mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)\right]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}} \leq \widehat{\epsilon}.
\end{aligned}$$

We can now apply the result of Corollary 12 (with $\widehat{\epsilon}$ instead of ϵ and $\widehat{\mathcal{D}}_k$ instead of $\mathcal{D}_k$). In view that, there exist $\eta > 0$ and $z \in \Delta$ such that

$$\mathcal{L}(\mathcal{D}_T, \widehat{h}_z^\eta) \leq \left[(\widehat{\epsilon} + \delta)\, \mathsf{d}_\alpha(\mathcal{D}_T \parallel \widehat{\mathcal{D}})\right]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}},$$

for any distribution $\widehat{\mathcal{D}}$ in the family $\widehat{\mathcal{D}}$. Taking the infimum over all $\widehat{\mathcal{D}}$ in $\widehat{\mathcal{D}}$ completes the proof. □

Corollary 4 uses Rényi divergence in both directions: $\mathsf{d}_\alpha(\mathcal{D}_T \parallel \widehat{\mathcal{D}})$ requires $\mathrm{Supp}(\mathcal{D}_T) \subseteq \mathrm{Supp}(\widehat{\mathcal{D}})$, and $\mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)$ requires $\mathrm{Supp}(\widehat{\mathcal{D}}_k) \subseteq \mathrm{Supp}(\mathcal{D}_k), k \in [p]$. In our experiments in Section 5, we used a bigram language model for sentiment analysis, and kernel density estimation with a Gaussian kernel for object recognition. Both density estimation methods fulfill these requirements.

Finally we prove our main result Theorem 1 under the regression model (R). We do so by proving a more general result, Theorem 2, and showing that it will coincide with Theorem 1 under the assumption that $\mathcal{D}_T^1 \in \mathcal{D}^1$ and $\mathcal{D}_T(\cdot|x)$ coincides for all $\mathcal{D}_T$.

Theorem 2 (Regression model)**.** *Fix a conditional probability distribution $\mathcal{Q}(\cdot|x)$ defined for all $x \in \mathcal{X}$. 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 $\mathcal{D}_T$:*

$$\mathcal{L}(\mathcal{D}_T, h_z^\eta) \leq \left[\left(\epsilon_\alpha(\mathcal{Q}) + \delta\right) \mathsf{d}_\beta(\mathcal{D}_T \parallel \mathcal{D}_{P,\mathcal{Q}})\right]^{\frac{\beta-1}{\beta}} M^{\frac{1}{\beta}}. \tag{10}$$

Proof. For any $k \in [p]$, by Hölder's inequality, the following holds:

$$\begin{aligned}
\mathcal{L}(\mathcal{D}_{k,\mathcal{Q}}, h_k) &= \sum_{x,y} \mathcal{D}_k^1(x)\mathcal{Q}(y|x)L(h_k, x, y) \\
&= \sum_x \mathcal{D}_k^1(x) \sum_y \left[\frac{\mathcal{Q}(y|x)}{\mathcal{D}_k(y|x)^{\frac{\alpha-1}{\alpha}}}\right]\left[\mathcal{D}_k(y|x)^{\frac{\alpha-1}{\alpha}} L(h_k, x, y)\right] \\
&\leq \sum_x \mathcal{D}_k^1(x)\mathsf{d}_\alpha(x; \mathcal{Q}, k)^{\frac{\alpha-1}{\alpha}}\Big[\sum_y \mathcal{D}_k(y|x)L(h_k, x, y)^{\frac{\alpha}{\alpha-1}}\Big]^{\frac{\alpha-1}{\alpha}},
\end{aligned}$$

where, for simplicity, we write $\mathsf{d}_\alpha(x; \mathcal{Q}, k) = \mathsf{d}_\alpha\left(\mathcal{Q}(\cdot|x) \parallel \mathcal{D}_k(\cdot|x)\right)$. Using the boundedness of the loss and Hölder's inequality again, we can write

$$\begin{aligned}
\mathcal{L}(\mathcal{D}_{k,\mathcal{Q}}, h_k) &\leq \sum_x \mathcal{D}_k^1(x)^{\frac{1}{\alpha}}\mathsf{d}_\alpha(x; \mathcal{Q}, k)^{\frac{\alpha-1}{\alpha}}\left[\sum_y \mathcal{D}_k(x, y)L(h_k, x, y)\right]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}} \\
&\leq \left[\sum_x \mathcal{D}_k^1(x)\mathsf{d}_\alpha(x; \mathcal{Q}, k)^{\alpha-1}\right]^{\frac{1}{\alpha}}\left[\sum_{x,y} \mathcal{D}_k(x, y)L(h_k, x, y)\right]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}} \\
&\leq \Big[\mathop{\mathbb{E}}_{\mathcal{D}_k^1}[\mathsf{d}_\alpha(x; \mathcal{Q}, k)^{\alpha-1}]\Big]^{\frac{1}{\alpha}} \epsilon^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}} \leq \epsilon_\alpha(\mathcal{Q}).
\end{aligned}$$

We can now apply the result of Corollary 12, with β instead of α, $\epsilon_\alpha(\mathcal{Q})$ instead of ϵ and $\mathcal{D}_{k,\mathcal{Q}}$ instead of $\mathcal{D}_k$. This completes the proof. □

When $\mathcal{D}_T^1 \in \mathcal{D}^1$, $\mathcal{D}_T^1(x)\mathcal{Q}(y|x) \in \mathcal{D}_{P,\mathcal{Q}}$ and we can write

$$\begin{aligned}
\mathsf{d}_\beta(\mathcal{D}_T \parallel \mathcal{D}_{P,\mathcal{Q}}) &\leq \left[\sum_{x,y} \frac{[\mathcal{D}_T^1(x)\mathcal{D}_T(y|x)]^\beta}{[\mathcal{D}_T^1(x)\mathcal{Q}(y|x)]^{\beta-1}}\right]^{\frac{1}{\beta-1}} \\
&= \Big[\sum_x \mathcal{D}_T^1(x) \sum_y \frac{[\mathcal{D}_T(y|x)]^\beta}{[\mathcal{Q}(y|x)]^{\beta-1}}\Big]^{\frac{1}{\beta-1}} \\
&= \Big[\mathop{\mathbb{E}}_{\mathcal{D}_T^1}\Big[\mathsf{d}_\beta\big(\mathcal{D}_T(\cdot|x) \parallel \mathcal{Q}(\cdot|x)\big)^{\beta-1}\Big]\Big]^{\frac{1}{\beta-1}}.
\end{aligned}$$

Applying this inequality to (10) yields

$$\begin{aligned}
\mathcal{L}(\mathcal{D}_T, h_z^\eta) &\leq \Big[\big(\epsilon_\alpha(\mathcal{Q}) + \delta\big)\,\mathsf{d}_\beta(\mathcal{D}_T \parallel \mathcal{D}_{P,\mathcal{Q}})\Big]^{\frac{\beta-1}{\beta}} M^{\frac{1}{\beta}} \\
&\leq \big(\epsilon_\alpha(\mathcal{Q}) + \delta\big)^{\frac{\beta-1}{\beta}}\Big[\mathop{\mathbb{E}}_{\mathcal{D}_T^1}\Big[\mathsf{d}_\beta\big(\mathcal{D}_T(\cdot|x) \parallel \mathcal{Q}(\cdot|x)\big)^{\beta-1}\Big]\Big]^{\frac{1}{\beta}} M^{\frac{1}{\beta}}. \qquad (11)
\end{aligned}$$

Notice that when the target distribution $\mathcal{D}_T$ is arbitrary but admits a fixed (and unknown) conditional probability distribution $\mathcal{D}_T(\cdot|x)$, we can set $\mathcal{Q}(\cdot|x) = \mathcal{D}_T(\cdot|x)$ in (11). We then have $\mathsf{d}_\beta(\mathcal{D}_T(\cdot|x) \parallel \mathcal{Q}(\cdot|x)) = 1$ for all $x \in \mathcal{X}$ and $\mathcal{L}(\mathcal{D}_T, h_z^\eta) \leq (\epsilon_\alpha(\mathcal{Q}) + \delta)^{\frac{\beta-1}{\beta}} M^{\frac{1}{\beta}}$. Thus, Theorem 2 coincides with the statement of Theorem 1 for the regression model by setting $\beta = +\infty$.

B.2 Choice of z

We have shown the existence of a robust solution h_z^η that works well for arbitrary target distribution $\mathcal{D}_T$. However, in the proof of Theorem 2, the choice of z depends on a fixed conditional probability distribution $\mathcal{Q}(\cdot|x)$. In practice, if the learner assumes that the conditional probability distribution $\mathcal{D}_k(\cdot|x)$ coincides, he can then set $\mathcal{Q}(\cdot|x) = \mathcal{D}_k(\cdot|x)$ and use $\mathcal{D}_k$ to solve the DC programming problem (4) for z. When the conditional probabilities are distinct, however, the learner needs to first come up with a choice of $\mathcal{Q}(\cdot|x)$, and then solve the DC programming problem (4) for z with $\mathcal{D}_{k,\mathcal{Q}}$ instead of $\mathcal{D}_k$, and the theoretical guarantees of h_z depend on $\mathcal{Q}(\cdot|x)$.

Can we find a robust solution z using only the original distributions $\mathcal{D}_k, \forall k \in [p]$, even when the conditional probability distributions $\mathcal{D}_k(\cdot|x)$ vary by k? The answer is yes. We now prove a

variant of Theorem 2 where the choice of z only depends on $\mathcal{D}_k, \forall k \in [p]$. This variant allows us to always use $\mathcal{D}_k$s in the DC programming formulation (4). In what follows, we denote by $\mathcal{D} = \{\sum_{k=1}^p \lambda_k \mathcal{D}_k, \lambda \in \Delta\}$, and $\mathcal{D}_{z,\mathcal{Q}}(x,y) = \sum_{k=1}^p z_k \mathcal{D}_{k,\mathcal{Q}}$.

Theorem 13. *Given any $\eta, \eta' > 0$, there exits $z \in \Delta$ such that the following holds for any $\lambda \in \Delta$:*

$$\mathcal{L}(\mathcal{D}_\lambda, h_z^\eta) \le \min_{\mathcal{Q}(\cdot|x)} \Bigg\{ \Big[\mathsf{d}_\alpha(\mathcal{D}_z \parallel \mathcal{D}_{z,\mathcal{Q}})\Big]^{\frac{\alpha-1}{\alpha}} \Big[\max_{k\in[p]} \mathsf{d}_\alpha(\mathcal{D}_{k,\mathcal{Q}} \parallel \mathcal{D}_k)\Big]^{\frac{(\alpha-1)^2}{\alpha^2}} M^{\frac{2\alpha-1}{\alpha^2}} \epsilon^{\frac{(\alpha-1)^2}{\alpha^2}} + \Big[\mathsf{d}_\alpha(\mathcal{D}_z \parallel \mathcal{D}_{z,\mathcal{Q}})\Big]^{\frac{\alpha-1}{\alpha}} M\eta^{\frac{\alpha-1}{\alpha}} + \eta' \Bigg\}. \tag{12}$$

When the conditional probability distributions $\mathcal{D}_k(\cdot|x)$ do not depend on k, (12) recovers the result of Corollary 3.

Furthermore, denote by $\mathcal{E}(\epsilon, \alpha, \eta, \eta')$ the above upper bound of $\mathcal{L}(\mathcal{D}_\lambda, h_z^\eta)$ for any $\lambda \in \Delta$. There exists $z \in \Delta$ such that for arbitrary target distribution $\mathcal{D}_T$,

$$\mathcal{L}(\mathcal{D}_T, h_z^\eta) \le \Big[\mathcal{E}(\epsilon, \alpha, \eta, \eta')\, \mathsf{d}_\alpha(\mathcal{D}_T \parallel \mathcal{D})\Big]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}.$$

Proof. Given any conditional probability distribution $\mathcal{Q}(\cdot|x)$, by the proof of Corollary 12, for any $z \in \Delta$,

$$\mathcal{L}(\mathcal{D}_z, h_z^\eta) \le \Big[\mathsf{d}_\alpha(\mathcal{D}_z \parallel \mathcal{D}_{z,\mathcal{Q}})\, \mathcal{L}(\mathcal{D}_{z,\mathcal{Q}}, h_z^\eta)\Big]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}. \tag{13}$$

By the proof of Corollary 3 and Corollary 12,

$$\begin{aligned}
\mathcal{L}(\mathcal{D}_{z,\mathcal{Q}}, h_z^\eta) &\le \sum_{k=1}^p z_k \mathcal{L}(\mathcal{D}_{k,\mathcal{Q}}, h_k) + \eta M \\
&\le \sum_{k=1}^p z_k \Big[\mathsf{d}_\alpha(\mathcal{D}_{k,\mathcal{Q}} \parallel \mathcal{D}_k)\, \mathcal{L}(\mathcal{D}_k, h_k)\Big]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}} + \eta M \\
&\le \mathsf{d}_\alpha(\mathcal{Q})\, M^{\frac{1}{\alpha}} \epsilon^{\frac{\alpha-1}{\alpha}} + \eta M,
\end{aligned}$$

where for simplicity we write $\mathsf{d}_\alpha(\mathcal{Q}) = \max_{k\in[p]} \mathsf{d}_\alpha(\mathcal{D}_{k,\mathcal{Q}} \parallel \mathcal{D}_k)^{\frac{\alpha-1}{\alpha}}$. Applying this inequality to (13) yields

$$\begin{aligned}
\mathcal{L}(\mathcal{D}_z, h_z^\eta) &\le \mathsf{d}_\alpha(\mathcal{D}_z \parallel \mathcal{D}_{z,\mathcal{Q}})^{\frac{\alpha-1}{\alpha}} \mathcal{L}(\mathcal{D}_{z,\mathcal{Q}}, h_z^\eta)^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}} \\
&\le \mathsf{d}_\alpha(\mathcal{D}_z \parallel \mathcal{D}_{z,\mathcal{Q}})^{\frac{\alpha-1}{\alpha}} \Big[\mathsf{d}_\alpha(\mathcal{Q})\, M^{\frac{1}{\alpha}} \epsilon^{\frac{\alpha-1}{\alpha}} + \eta M\Big]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}} \\
&\le \Big[\mathsf{d}_\alpha(\mathcal{D}_z \parallel \mathcal{D}_{z,\mathcal{Q}})\, \mathsf{d}_\alpha(\mathcal{Q})\Big]^{\frac{\alpha-1}{\alpha}} M^{\frac{2\alpha-1}{\alpha^2}} \epsilon^{\frac{(\alpha-1)^2}{\alpha^2}} \\
&\quad + \Big[\mathsf{d}_\alpha(\mathcal{D}_z \parallel \mathcal{D}_{z,\mathcal{Q}})\Big]^{\frac{\alpha-1}{\alpha}} M\eta^{\frac{\alpha-1}{\alpha}}.
\end{aligned}$$

Next, let $\mathcal{D}_\lambda$ be an arbitrary mixture of source domains, $\lambda \in \Delta$. Notice that Lemma 6 does not rely on any assumption of conditional probabilities, thus given fixed η, η', we can still find z such that $\mathcal{L}(\mathcal{D}_k, h_z^\eta) \le \mathcal{L}(\mathcal{D}_z, h_z^\eta) + \eta'$ for all $k \in [p]$, which implies that $\mathcal{L}(\mathcal{D}_\lambda, h_z^\eta) \le \mathcal{L}(\mathcal{D}_z, h_z^\eta) + \eta'$ for any $\lambda \in \Delta$. Thus, the choice of z only depends on $\mathcal{D}_k, \forall k \in [p]$. This proves (12).

When the conditional probability distributions $\mathcal{D}_k(\cdot|x)$ do not depend on k, let $\mathcal{Q}(\cdot|x) = \mathcal{D}_k(\cdot|x)$, thus $\mathsf{d}_\alpha(\mathcal{D}_z \parallel \mathcal{D}_{z,\mathcal{Q}}) = 1$ and $\mathsf{d}_\alpha(\mathcal{Q}) = 1$. Setting $\alpha = +\infty$ and choosing η, η' accordingly, we recover the result of Corollary 3.

Finally, by the proof of Corollary 12, for any $\lambda \in \Delta$,

$$\begin{aligned}
\mathcal{L}(\mathcal{D}_T, h_z^\eta) &\le \Big[\mathsf{d}_\alpha(\mathcal{D}_T \parallel \mathcal{D}_\lambda)\, \mathcal{L}(\mathcal{D}_\lambda, h_z^\eta)\Big]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}} \\
&\le \Big[\mathsf{d}_\alpha(\mathcal{D}_T \parallel \mathcal{D}_\lambda)\, \mathcal{E}(\epsilon, \alpha, \eta, \eta')\Big]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}.
\end{aligned}$$

Taking the infimum of the right-hand side over all $\mathcal{D}_\lambda \in \mathcal{D}$ completes the proof. □

The main difference between Theorem 2 and Theorem 13 is the dependency of z: in Theorem 2, z depends on a prefixed $\mathcal{Q}(\cdot|x)$, while in Theorem 13, z only depends on $\mathcal{D}_k$. The guarantees in Theorem 13 ensure that we can first use $\mathcal{D}_k, k \in [p]$ to find a solution z such that h_z^η admits essentially the same loss on all source domains. Then, the performance of h_z^η on arbitrary target distribution $\mathcal{D}_T$ relies on how close the source and target conditional probability distributions are to a *pivot* $\mathcal{Q}(\cdot|x)$, as well as on the divergence between $\mathcal{D}_T$ and $\mathcal{D}$.

B.3 Probability model

In this section, we first present a series of general theoretical results for the probability model (P) in the same order as in Appendix B.1. Many of the them are similar to those for the regression model, except that we do not assume anything about the conditional probabilities throughout the proofs. In several instances, the proofs are syntactically the same as their counterparts in the regression model (R). In such cases, we do not reproduce them.

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^η:*

$$\forall k \in [p], \quad \mathcal{L}(\mathcal{D}_k, h_z^\eta) \leq \sum_{j=1}^{p} z_j \mathcal{L}(\mathcal{D}_j, h_z^\eta) + \eta'. \tag{14}$$

Proof. The proof is syntactically the same as that for the regression model. □

Corollary 3. *For any $\delta > 0$, there exist $\eta > 0$ and $z \in \Delta$, such that $\mathcal{L}(\mathcal{D}_\lambda, h_z^\eta) \leq \epsilon + \delta$ for any mixture parameter $\lambda \in \Delta$.*

Proof. Modifying the proof of Corollary 3 for the regression model gives

$$\begin{aligned}
\mathcal{L}(\mathcal{D}_z, h_z^\eta) &= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \mathcal{D}_z(x,y) L(h_z^\eta(x,y)) \\
&= \sum_{(x,y)} \mathcal{D}_z(x,y) L\left(\sum_{k=1}^{p} \frac{z_k \mathcal{D}_k(x,y) + \eta \frac{\mathcal{U}(x,y)}{p}}{\mathcal{D}_z(x,y) + \eta \mathcal{U}(x,y)} h_k(x,y) \right).
\end{aligned}$$

By convexity of L, this implies that

$$\mathcal{L}(\mathcal{D}_z, h_z^\eta) \leq \sum_{(x,y)} \mathcal{D}_z(x,y) \sum_{k=1}^{p} \frac{z_k \mathcal{D}_k(x,y) + \eta \frac{\mathcal{U}(x,y)}{p}}{\mathcal{D}_z(x,y) + \eta \mathcal{U}(x,y)} L\big(h_k(x,y)\big).$$

Next, since $\frac{\mathcal{D}_z(x,y)}{\mathcal{D}_z(x,y)+\eta\mathcal{U}(x,y)} \leq 1$, the following holds:

$$\begin{aligned}
\mathcal{L}(\mathcal{D}_z, h_z^\eta) &\leq \sum_{(x,y)} \Big(\sum_{k=1}^{p} \big(z_k \mathcal{D}_k(x,y) + \tfrac{\eta \mathcal{U}(x,y)}{p} \big) L(h_k(x,y)) \Big) \\
&= \sum_{k=1}^{p} z_k \mathcal{L}(\mathcal{D}_k, h_k) + \frac{\eta}{p} \sum_{k=1}^{p} \mathcal{L}(\mathcal{U}, h_k) \\
&\leq \sum_{k=1}^{p} z_k \epsilon + \eta M = \epsilon + \eta M.
\end{aligned}$$

Now choose $z \in \Delta$ as in the statement of Lemma 4a.Then, the following holds for any mixture distribution $\mathcal{D}_\lambda$:

$$\begin{aligned}
\mathcal{L}(\mathcal{D}_\lambda, h_z^\eta) &= \sum_{k=1}^{p} \lambda_k \mathcal{L}(\mathcal{D}_k, h_z^\eta) \leq \sum_{k=1}^{p} \lambda_k (\mathcal{L}(\mathcal{D}_z, h_z^\eta) + \eta') \\
&= \mathcal{L}(\mathcal{D}_z, h_z^\eta) + \eta' \leq \epsilon + \eta M + \eta'.
\end{aligned}$$

Setting $\eta = \frac{\delta}{2M}$ and $\eta' = \frac{\delta}{2}$ concludes the proof.

□

Since we do not assume the conditional probabilities are the same across domains, we can directly prove the following theorem for the conditional probability model (P), which coincides with Theorem 1 when $\mathcal{D}_T \in \mathcal{D}$.

Theorem 14. *For any $\delta > 0$, there exist $\eta > 0$ and $z \in \Delta$, such that the following inequality holds for any $\alpha > 1$ and arbitrary target distribution $\mathcal{D}_T$:*

$$\mathcal{L}(\mathcal{D}_T, h_z^\eta) \leq \left[(\epsilon + \delta)\, \mathsf{d}_\alpha(\mathcal{D}_T \parallel \mathcal{D})\right]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}} \qquad (P).$$

Proof. The proof is syntactically the same as that of Corollary 12 for the regression model. □

Corollary 15. *Then, for any $\delta > 0$, there exist $\eta > 0$ and $z \in \Delta$, such that the following inequality holds for any $\alpha > 1$ and arbitrary target distribution $\mathcal{D}_T$:*

$$\mathcal{L}(\mathcal{D}_T, \widehat{h}_z^\eta) \leq \left[(\widehat{\epsilon} + \delta)\, \mathsf{d}_\alpha(\mathcal{D}_T \parallel \widehat{\mathcal{D}})\right]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}},$$

where $\widehat{\epsilon} = \max_{k\in[p]} \left[\epsilon\, \mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)\right]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}$, and $\widehat{\mathcal{D}} = \left\{\sum_{k=1}^p \lambda_k \widehat{\mathcal{D}}_k : \lambda \in \Delta\right\}$.

Proof. The proof is syntactically the same as that of Corollary 4 for the regression model. □

C Specific theoretical analysis for the cross-entropy loss

Next, we give a specific theoretical analysis for the case of the cross-entropy loss. This is needed since the cross-entropy loss assumes normalized hypotheses. Thus, we are giving guarantees for the performance of normalized distribution-weighted predictor.

We will first assume that the conditional probability of the output labels is the same for all source domains, that is, for any (x, y), $\mathcal{D}_k(y|x)$ is independent of k.

Theorem 5. *Assume that there exists $\mu > 0$ such that $\mathcal{D}_k(x, y) \geq \mu \mathcal{U}(x, y)$ for all $k \in [p]$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Then, for any $\delta > 0$, there exist $\eta > 0$ and $z \in \Delta$ such that $\mathcal{L}(\mathcal{D}_\lambda, \overline{h}_z^\eta) \leq \epsilon + \delta$ for any mixture parameter $\lambda \in \Delta$.*

Proof. By the proof of Corollary 3 for the probability model, for any mixture distribution $\mathcal{D}_\lambda$:

$$\mathcal{L}(\mathcal{D}_\lambda, h_z^\eta) \leq \epsilon + \eta M + \eta',$$

for some $\eta > 0, \eta' > 0$. For any $x \in \mathcal{X}$,

$$\begin{aligned}
\sum_{y \in \mathcal{Y}} h_z^\eta(x, y) &= \sum_{y \in \mathcal{Y}} \sum_{k=1}^p \frac{z_k \mathcal{D}_k(x, y) + \frac{\eta \mathcal{U}(x,y)}{p}}{\mathcal{D}_z(x, y) + \eta \mathcal{U}(x, y)} h_k(x, y) \\
&\leq \sum_{y \in \mathcal{Y}} \sum_{k=1}^p \frac{z_k \mathcal{D}_k(x, y) + \frac{\eta \mathcal{U}(x,y)}{p}}{\mathcal{D}_z(x, y)} h_k(x, y) \\
&= 1 + \eta \left[\frac{1}{p} \sum_{y \in \mathcal{Y}} \sum_{k=1}^p \frac{\mathcal{U}(x, y)}{\mathcal{D}_z(x, y)} h_k(x, y) \right]. \qquad (15)
\end{aligned}$$

By assumption, $\mathcal{D}_k(x, y) \geq \mu \mathcal{U}(x, y)$ for any (x, y). Therefore $\mathcal{D}_z(x, y) \geq \mu \mathcal{U}(x, y)$ for any $z \in \Delta$. Since $0 \leq h_k(x, y) \leq 1$, equation (15) is further upper bounded by

$$\sum_{y \in \mathcal{Y}} h_z^\eta(x, y) \leq 1 + \eta \left[\frac{1}{p} \sum_{y \in \mathcal{Y}} \sum_{k=1}^p \frac{\mathcal{U}(x, y)}{\mathcal{D}_z(x, y)} h_k(x, y) \right] \leq 1 + \frac{\eta |\mathcal{Y}|}{\mu}.$$

It follows that

$$\begin{aligned}
\mathcal{L}(\mathcal{D}_\lambda, \overline{h}_z^\eta) &= \mathcal{L}(\mathcal{D}_\lambda, h_z^\eta) + \mathop{\mathbb{E}}_{x \sim \mathcal{D}_\lambda} \left[\log \Big(\sum_{y \in \mathcal{Y}} h_z^\eta(x, y) \Big) \right] \leq \epsilon + \eta M + \eta' + \log \left(1 + \frac{\eta |\mathcal{Y}|}{\mu} \right) \\
&\leq \epsilon + \eta \left(M + \frac{|\mathcal{Y}|}{\mu} \right) + \eta'.
\end{aligned}$$

Setting $\eta = \frac{\delta}{2\left(M + \frac{|\mathcal{Y}|}{\mu}\right)}$ and $\eta' = \frac{\delta}{2}$ concludes the proof. □

The analysis above depends on the key assumption that the conditional distributions $\mathcal{D}_k(y|x)$ are independent of k. When this assumption does not hold, we can show that there is a lower bound of $\log(p)$ on the generalization error $\mathcal{L}(\mathcal{D}_\lambda, \overline{h}_z^\eta)$. However, this lower bound coincides with that of convex combination rule (Lemma 10). In that case, one can use the following marginal distribution-weighted combination instead:

$$\widetilde{h}_z^\eta(x, y) = \sum_{k=1}^p \frac{z_k \mathcal{D}_k^1(x) + \eta \frac{\mathcal{U}^1(x)}{p}}{\sum_{j=1}^p z_j \mathcal{D}_j^1(x) + \eta \mathcal{U}^1(x)} h_k(x, y), \qquad (16)$$

where $\mathcal{D}_k^1(x)$ is the marginal distribution over $\mathcal{X}$, $\mathcal{D}_k^1(x) = \sum_{y \in \mathcal{Y}} \mathcal{D}_k(x, y)$, and $\mathcal{U}^1(x)$ is a uniform distribution over $\mathcal{X}$. Observe that $\widetilde{h}_z^\eta(x, y)$ is already normalized.

One can modify Theorem 2 to obtain generalization guarantees for $\widetilde{h}_z^\eta$ under distinct conditional probabilities assumption. Let $\mathcal{D}_T(x, y)$, $\epsilon_\alpha(\mathcal{Q})$ and $\mathcal{D}_{P,\mathcal{Q}}$ be defined as before.

Theorem 16. *For any $\delta > 0$, there exist $\eta > 0$ and $z \in \Delta$ such that the following inequality holds for any $\alpha, \beta > 1$ and arbitrary target distribution $\mathcal{D}_T$:*

$$\mathcal{L}(\mathcal{D}_T, \widetilde{h}_z^{\eta}) \leq \left[\left(\epsilon_\alpha(\mathcal{Q}) + \delta\right) \mathsf{d}_\beta(\mathcal{D}_T \parallel \mathcal{D}_{P,\mathcal{Q}})\right]^{\frac{\beta-1}{\beta}} M^{\frac{1}{\beta}}.$$

Proof. The proof is syntactically the same as that of Theorem 2. □

Finally, we can extend Theorem 5 and Theorem 16 to the case where only estimate distributions $\widehat{\mathcal{D}}_k$s are available, and the predictor $\widetilde{\widehat{h}_z^{\eta}}$ and $\widehat{\widetilde{h}_z^{\eta}}$ based on the estimates $\widehat{\mathcal{D}}_k$ still admit favorable guarantees. The results and proofs are similar to proving Corollary 4 from Corollary 12 in the regression model, thus omitted here.

D DC-decomposition

In this section we give the full proofs for the DC-decompositions presented in Section 4.2.

D.1 Regression model

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

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

Proof. First, observe that $(h_z^\eta(x) - y)^2 = f_z(x,y) - g_z(x)$, where for every $(x,y) \in \mathcal{X} \times \mathcal{Y}$, f_z and g_z are convex functions defined for all z:

$$f_z(x,y) = (h_z^\eta(x) - y)^2 - 2M \log K_z(x),$$
$$g_z(x) = -2M \log K_z(x).$$

This is true because the Hessian matrix of f_z and g_z are

$$H_{f_z} = \frac{2}{K_z^2}\left[h_{D,z} h_{D,z}^T + \left(M - (y - h_z^\eta)^2\right) DD^T\right],$$
$$H_{g_z} = \frac{2M}{K_z^2} DD^T,$$

where $h_{D,z}$ is a p-dimensional vector defined as $[h_{D,z}]_k = \mathcal{D}_k(h_k + y - 2h_z^\eta)$ for $k \in [p]$, and $D = (\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_p)^T$. Using the fact that $M \geq (y - h_z^\eta)^2$, H_{f_z} and H_{g_z} are positive semidefinite matrices, therefore f_z, g_z are convex functions of z.

Thus, $u_k(z) = \sum_{(x,y)} (\mathcal{D}_k^1 + \eta \mathcal{U}^1)(x) \mathcal{D}_k(y|x) f_z(x,y)$ is convex. Similarly, we can write the second term of $v_k(z)$ as $\sum_x (\mathcal{D}_k^1 + \eta \mathcal{U}^1)(x) g_z(x)$, it is convex. Using the notation previously defined, we can write the first term of $v_k(z)$ as

$$\mathcal{L}(\mathcal{D}_z + \eta \mathcal{U}^1 \mathcal{D}_k(\cdot|x), h_z^\eta) = \sum_x \frac{J_z(x)^2}{K_z(x)} - 2\,\mathbb{E}(y|x) J_z(x) + \mathbb{E}(y^2|x) K_z(x).$$

The Hessian matrix of J_z^2/K_z is

$$\nabla_z^2 \left(\frac{J_z^2}{K_z}\right) = \frac{1}{K_z}(h_D - h_z^\eta D)(h_D - h_z^\eta D)^T$$

where $h_D = (h_1 \mathcal{D}_1, h_2 \mathcal{D}_2, \ldots, h_p \mathcal{D}_p)^T$ and $D = (\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_p)^T$. Thus J_z^2/K_z is convex. $-2\,\mathbb{E}(y|x) J_z(x) + \mathbb{E}(y^2|x) K_z(x)$ is an affine function of z and is therefore convex. Therefore the first term of $v_k(z)$ is convex, which completes the proof. □

D.2 Probability model

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

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

Proof. Using the notation previously introduced, we can now write

$$\begin{aligned}
&\mathcal{L}(\mathcal{D}_k, h_z^\eta) - \mathcal{L}(\mathcal{D}_z, h_z^\eta) \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_k}[-\log h_z^\eta(x,y)] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_z}[-\log h_z^\eta(x,y)] \\
&= \sum_{x,y} \big(\mathcal{D}_z(x,y) - \mathcal{D}_k(x,y)\big) \log\left[\frac{J_z(x,y)}{K_z(x,y)}\right] \\
&= \sum_{x,y} \big[K_z(x,y) - (\mathcal{D}_k(x,y) + \eta\,\mathcal{U}(x,y))\big] \log\left[\frac{J_z(x,y)}{K_z(x,y)}\right] \\
&= u_k(z) - v_k(z).
\end{aligned}$$

u_k is convex since $-\log J_z$ is convex as the composition of the convex function $-\log$ with an affine function. Similarly, $-\log K_z$ is convex, which shows that the second term in the expression of v_k is a convex function. The first term can be written in terms of the unnormalized relative entropy:[1]

$$\begin{aligned}
&\sum_{x,y} K_z(x,y) \log\left[\frac{K_z(x,y)}{J_z(x,y)}\right] \\
&= B(K_z \parallel J_z) + \sum_{(x,y)} (K_z - J_z)(x,y).
\end{aligned}$$

The unnormalized relative entropy $B(\cdot \parallel \cdot)$ is jointly convex (Cover and Thomas, 2006),[2] thus $B(K_z \parallel J_z)$ is convex as the composition of the unnormalized relative entropy with affine functions (for each of its two arguments). $(K_z - J_z)$ is an affine function of z and is therefore convex too. □

[1] The unnormalized relative entropy of P and Q is defined by $B(P \parallel Q) = \sum_{x,y} P(x,y) \log\left[\frac{P(x,y)}{Q(x,y)}\right] + \sum_{(x,y)} (Q(x,y) - P(x,y))$.

[2] To be precise, it can be shown that the relative entropy is jointly convex using the so-called log-sum inequality (Cover and Thomas, 2006). The same proof using the log-sum inequality can be used to show the joint convexity of the unnormalized relative entropy.

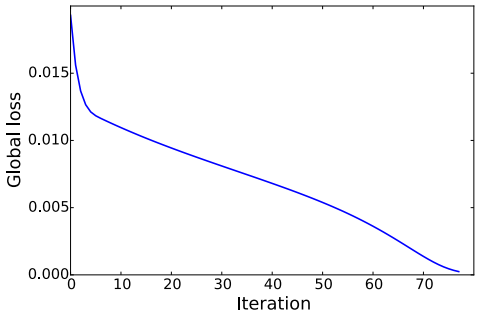
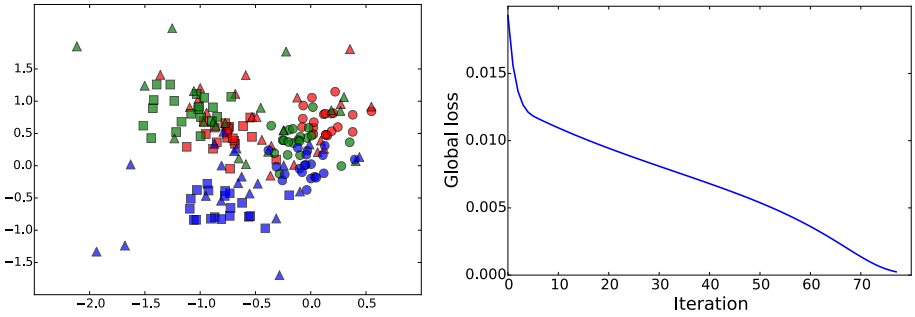


Figure 2: Synthetic global loss versus iteration for squared loss. Our solution converges to the global optimum of zero.



(a) (b)

Figure 3: (a) Artificial dataset for cross-entropy loss, with three domains (red, green and blue) and three categories (triangle, square, circle). (b) Artificial dataset global loss versus iteration for cross-entropy loss. We empirically find that our solution converges to the global optimum of zero.

E Additional experiment results

In this section we provide experiment results on artificial datasets to show that our global objective indeed approaches the known optimal of zero with DC-programming algorithm, for both squared loss and cross-entropy loss. We also provide details of our density estimation procedure on the real-world applications, as well as additional experiment results to show that our distribution-weighted predictor `DW` is robust across various test data mixtures.

E.1 Artificial dataset

We first evaluated our algorithm on synthetic datasets, for both squared loss and cross-entropy loss.

Consider the following multiple source domain study by Mansour et al. (2009a). Let g_1, g_2, g_3, g_4 denote the Gaussian distributions with means $(1, 1)$, $(-1, 1)$, $(-1, -1)$, and $(1, -1)$ and unit variance respectively. Each domain was generated as a uniform mixture of Gaussians: $\mathcal{D}_1$ from $\{g_1, g_2, g_3\}$ and $\mathcal{D}_2$ from $\{g_2, g_3, g_4\}$. The labeling function is $f(x_1, x_2) = x_1^2 + x_2^2$. We trained linear regressors for each domain to produce base hypotheses h_1 and h_2. Finally, as the true distribution is known for this artificial example, we directly use the Gaussian mixture density function to generate our $\mathcal{D}_k$s.

With this data source, we used our DC-programming solution to find the optimal mixing weights z. Figure 2 shows the global objective value (of Problem 4) vs number of iterations with the uniform initialization $z_0 = [1/2, 1/2]$. Here, the overall objective approaches 0.0, the known global minimum. To verify the robustness of the solution, we have experimented with various initial conditions and found that the solution converges to the global solution in each case.

We next evaluate our algorithm on cross-entropy loss. Here we generate the two-dimensional dataset shown in Figure 3a, which has three domains, denoted in the colors red, green, and blue, and three

Table 5: MSE on sentiment analysis dataset: target domain as various combinations of two domains.

	Test Data					
	KD	BE	KB	KE	DB	DE
K	1.83±0.08	1.99±0.10	1.87±0.08	1.57±0.06	2.25±0.08	1.94±0.10
D	1.95±0.07	2.11±0.07	2.12±0.07	2.11±0.05	1.95±0.06	1.94±0.06
B	2.10±0.09	1.99±0.08	1.96±0.07	2.21±0.06	1.87±0.07	2.13±0.05
E	2.00±0.09	1.95±0.07	2.05±0.05	1.60±0.05	2.36±0.07	1.91±0.07
unif	1.73±0.06	1.74±0.07	1.74±0.05	1.62±0.04	1.85±0.05	1.73±0.06
KMM	1.83±0.07	1.82±0.07	1.78±0.12	1.65±0.10	1.97±0.13	1.88±0.08
DW	**1.62±0.07**	**1.61±0.08**	**1.59±0.05**	**1.47±0.04**	**1.75±0.05**	**1.64±0.05**

Table 6: MSE on the sentiment analysis dataset: target domain as various mixture of four domains: $(\mathbf{0.4}, 0.2, 0.2, 0.2)$, $(0.2, \mathbf{0.4}, 0.2, 0.2)$, $(0.2, 0.2, \mathbf{0.4}, 0.2)$, $(0.2, 0.2, 0.2, \mathbf{0.4})$ of K, D, B, E respectively.

	Test Data			
	KDBE	KDBE	KDBE	KDBE
K	1.78±0.05	1.94±0.10	1.96±0.08	1.84±0.07
D	2.02±0.10	1.98±0.10	2.06±0.11	2.05±0.09
B	2.01±0.12	2.01±0.14	1.94±0.14	2.06±0.11
E	1.93±0.08	2.04±0.10	2.08±0.10	1.89±0.08
unif	1.69±0.06	1.74±0.07	1.75±0.08	1.70±0.06
KMM	1.83±0.12	1.92±0.14	1.87±0.15	1.85±0.13
DW	**1.55±0.08**	**1.62±0.08**	**1.59±0.09**	**1.56±0.08**

categories, denoted as squares, circles, and triangles. Each domain is generated according to a Gaussian mixture model, one mixture per category, with random means. The means of each corresponding category across domains are related according to a random fixed orthonormal transformation. Finally, the covariance of each mixture is diagonal and fixed across categories. We choose covariance magnitudes of 0.05, 0.05, and 0.3 for the red, green, and blue domains, respectively. We then train a logistic regression classifier per domain to produce score functions, h_k. Finally, as the true distribution is known for this artificial example, we forgo density estimation and use the Gaussian mixture density function to generate our $\mathcal{D}_k$s.

With this data source, we use our DC-programming solution to find the optimal mixing weights, z. Since only each convex sub-problem is guaranteed to converge, Figure 3b reports this global loss vs iteration when initializing $z_0 = 1/p$, uniform weights. Here, the overall objective approaches 0.0, the known global minimum.To verify the robustness of the solution, we have experimented with various initial conditions and found the solution converges to the global solution from each case.

E.2 Sentiment analysis task for squared loss

We begin by detailing our density estimation method for the sentiment analysis experiment. We first used the same vocabulary defined for feature extraction to train a separate bigram statistical language model for each domain, using the OpenGrm library (Roark et al., 2012). Next, we randomly draw a sample set S_k of $10{,}000$ sentences from each bigram language model. We define $\widehat{\mathcal{D}}_k$ to be the empirical distribution of S_k, which is a very close estimate of marginal distribution of the language model, thus it is also a good estimate of $\mathcal{D}_k$. We approximate the label of a randomly generated sample x_i by taking the average of the h_ks: $y_i = \sum_{\{k: x_i \in S_k\}} h_k(x_i) / |\{k: x_i \in S_k\}|$. These randomly drawn samples were used to find the fixed-point z.

Note that we only use estimates of the marginal distributions (language models) to find z and do not use any labels. We use the original product review text and rating labels for testing. Their densities $\widehat{\mathcal{D}}_k$ were estimated by the bigram language models directly, therefore a close estimate of $\mathcal{D}_k$.

Next we compare DW to accessible predictors on various test mixture domains. Table 5 shows MSE on all combinations of two domains. Table 6, 7 reports MSE on additional test mixture domains. The first four target mixtures correspond to various orderings of $(0.4, 0.2, 0.2, 0.2)$. The next six target mixtures correspond to various orderings of $(0.3, 0.3, 0.2, 0.2)$. In column titles we bold the domain(s) with highest weight.

Table 7: MSE on the sentiment analysis dataset: target domain as various mixture of four domains: $(\mathbf{0.3}, \mathbf{0.3}, 0.2, 0.2)$, $(\mathbf{0.3}, 0.2, \mathbf{0.3}, 0.2)$, $(\mathbf{0.3}, 0.2, 0.2, \mathbf{0.3})$, $(0.2, \mathbf{0.3}, \mathbf{0.3}, 0.2)$, $(0.2, \mathbf{0.3}, 0.2, \mathbf{0.3})$, $(0.2, 0.2, \mathbf{0.3}, \mathbf{0.3})$ of `K, D, B, E` respectively.

	Test Data					
	KDBE	KDBE	KDBE	KDBE	KDBE	KDBE
K	1.86±0.10	1.87±0.07	1.79±0.08	1.96±0.10	1.89±0.10	1.89±0.08
D	2.01±0.13	2.05±0.12	2.04±0.12	2.03±0.12	2.02±0.13	2.06±0.12
B	2.01±0.15	1.98±0.14	2.05±0.13	1.98±0.15	2.04±0.14	2.01±0.13
E	2.00±0.10	2.01±0.09	1.91±0.08	2.08±0.10	1.97±0.08	1.99±0.08
unif	1.72±0.09	1.72±0.08	1.69±0.07	1.75±0.08	1.72±0.08	1.73±0.08
KMM	1.85±0.16	1.86±0.14	1.85±0.15	1.90±0.14	1.89±0.16	1.90±0.14
DW	**1.58±0.10**	**1.57±0.10**	**1.55±0.09**	**1.61±0.10**	**1.59±0.08**	**1.58±0.09**

In all these experiments, our distribution-weighted predictor `DW` outperforms all competing baselines: the source only baselines for each domain, `K, D, B, E`, a uniform weighted predictor `unif`, and `KMM`.

E.3 Recognition tasks for cross-entropy loss

Here, we describe our density estimation technique for the object recognition task.

To estimate the per domain densities, we first extract per image features using the in-domain ConvNet model, and then estimate the marginal distribution $\mathcal{D}_k^1(x)$ over the per domain collection of features, using non-parametric kernel density estimation with a Gaussian kernel and a cross-validated bandwidth parameter. We use estimated marginals $\widehat{\mathcal{D}^1}_k$ instead of estimated joint distributions $\widehat{\mathcal{D}}_k$, because when the conditional probabilities are the same across domains and when $\eta \to 0$, $h_z^\eta(x, y)$ converges to a normalized predictor $\widetilde{h}_z(x, y) = \sum_{k=1}^p \frac{z_k \mathcal{D}_k^1(x)}{\sum_{j=1}^p z_j \mathcal{D}_j^1(x)} h_k(x, y)$. Thus in our experiments, we approximate $\widehat{h}_z^\eta(x, y)$ with $\widetilde{\widehat{h}}(x, y)$ using our estimated marginal distributions $\widehat{\mathcal{D}^1}_k(x)$.

F Rényi Divergence

The Rényi Divergence measures the divergence between two distributions. It is parameterized by $\alpha \in [0, +\infty]$ and denoted by D_α. The α-Rényi Divergence of two distributions $\mathcal{D}$ and $\mathcal{D}'$ is defined by

$$\mathsf{D}_\alpha(\mathcal{D} \parallel \mathcal{D}') = \frac{1}{\alpha - 1} \log \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \mathcal{D}(x,y) \left[\frac{\mathcal{D}(x,y)}{\mathcal{D}'(x,y)} \right]^{\alpha - 1}, \tag{17}$$

where, for $\alpha \in \{0, 1, +\infty\}$, the expression is defined by taking the limit. It can be shown that the Rényi Divergence is always non-negative and that for any $\alpha > 0$, $\mathsf{D}_\alpha(\mathcal{D} \parallel \mathcal{D}') = 0$ iff $\mathcal{D} = \mathcal{D}'$ (Arndt, 2004). We will denote by $\mathsf{d}_\alpha(\mathcal{D} \parallel \mathcal{D}')$ the exponential:

$$\mathsf{d}_\alpha(\mathcal{D} \parallel \mathcal{D}') = e^{\mathsf{D}_\alpha(\mathcal{D} \parallel \mathcal{D}')} = \left[\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \frac{\mathcal{D}^\alpha(x,y)}{\mathcal{D}'^{\alpha-1}(x,y)} \right]^{\frac{1}{\alpha - 1}}. \tag{18}$$

The Rényi divergence (and $\mathsf{d}_\alpha(\mathcal{D} \parallel \mathcal{D}')$) is a non-decreasing function of α; in particular, the following inequality holds:

$$\mathsf{d}_\alpha(\mathcal{D} \parallel \mathcal{D}') \leq \mathsf{d}_\infty(\mathcal{D} \parallel \mathcal{D}') = \sup_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \left[\frac{\mathcal{D}(x,y)}{\mathcal{D}'(x,y)} \right]. \tag{19}$$

G Related work on multiple source adaptation (MSA)

We give an extensive discussion of related work on multiple source adaptation (MSA) problem here, and point out how our scenario and our results are distinct from most previous works.

The learning scenario that we consider is distinct from and is more challenging than the one considered by many other existing multiple source adaptation (MSA) studies:

1. We assume that the learner does not have access to and therefore cannot *combine all the source labeled data* together to jointly train a target predictor. This is a very realistic assumption with legitimate reasons, such as data privacy, storage limitation, etc. Instead, the learner is only given pre-trained models, density estimations from the source domains, and a small subset of combined source labeled data.
2. We are not given any target data, even unlabeled, and we are competing against any target mixture distribution. This is a significantly more difficult problem. Remarkably, the learner only needs to run our algorithm `DW` once to obtain a single predictor, which is guaranteed to perform well on *any* target mixture. This is also verified experimentally.

To our knowledge, there is no discussion of this learning scenario other than by Mansour et al. (2008, 2009a). On the contrary, most MSA algorithms require access to the full combined source data to jointly train a predictor. In addition, many MSA algorithms require a set of labeled or unlabeled data from the target domain, and the solution only performs well on that specific target domain. If the target domain changes, the learner has to rerun the algorithm to find a new solution. Besides, many MSA algorithms do not admit theoretical guarantees, while we provide a careful analysis and series of strong theoretical guarantees for our algorithm `DW`.

In what follows, we categorize and discuss previous works on MSA problem by their learning scenarios.

Combine source data. Khosla et al. (2012) considered a similar setting where the learner trains a single predictor for any target domain and where the learner has access to source data but not target data. However, Khosla et al. (2012) combine all the source data to jointly train the final predictor and a large set of combined data is necessary for a good predictor. Additionally, the solution of Khosla et al. (2012) only works for linear functions, a very limited family of hypotheses. `DW` does not combine all the source data, and works for hypotheses of any form. Blanchard et al. (2011) presented MSA algorithms with theoretical guarantees. However, it combines all source data and target data to learn a final predictor. This paper also makes the strong assumption that the source and the target domains are i.i.d. realizations of some distribution, and their learning guarantee is with respect to that distribution. `DW` makes no assumption about the relationship between the source domains. Hoffman et al. (2012) considered multiclass classification problem where the predicted label for a novel test point is determined by a weighted sum of probabilities of each category given that the test point comes from a particular source domain. The weights are the predicted probability that the test point belongs to each source domain, which are learned via SVM on all source data combined. Zhang et al. (2015) considered a causal view of MSA where label Y is the cause for features X, and learned a weighted combination of source conditional probabilities ($\mathbb{P}_{X|Y}$) by minimizing the maximum mean discrepancy (MMD) on the combined source data. Muandet et al. (2013) proposed Domain-Invariant Component Analysis (DICA) to transform features onto a low dimensional subspace by minimizing dissimilarity across multiple source domains, while preserving relationship between features and label. The projection is learned via a kernel-based optimization on all source data combined. Recently, Pei et al. (2018) extended domain adversarial learning techniques for the multiple source setting.

Use labeled target data. Duan et al. (2009, 2012) considered a somewhat similar problem where the learner leverages pre-trained predictors from the source domains to learn a good predictor on the target domain. However, they assume plenty of unlabeled target data to form a meaningful regularizer and they also assume a small set of labeled target data. Their solution directly depends on the labeled and unlabeled target data and is of course only useful for that specific target. Yang et al. (2007) also considered the problem of combining pre-trained classifiers from multiple auxiliary datasets to adapt to a target dataset, and the solution is to learn a good linear combinations of auxiliary classifiers using Adaptive SVMs on the labeled target data.

Others. Crammer et al. (2008) dealt with a problem with multiple sources distinct from domain adaptation where the sources have the same input distribution but can have different labels, modulo some disparity constraints. (See also discussion by Mansour et al. (2009a)). Gong et al. (2012) proposed a Rank of Domain (ROD) metric that ranks multiple source domains by how likely they are to adapt well to a target domain. Gong et al. (2013a) learned domain-invariant features by first constructing multiple auxiliary tasks based on landmarks within the source data, and then learning new feature representations from each auxiliary task. Gong et al. (2013b) proposed to discover multiple latent domains by maximizing two key properties, distinctiveness and learnability, between latent domains. Xu et al. (2014) also considered the problem of discovering latent domains, and proposed to learn exemplar-SVMs with low-rank structure.