A Discriminative Technique for Multiple-Source Adaptation

Corinna Cortes [1] Mehryar Mohri [1 2] Ananda Theertha Suresh [1] Ningshan Zhang [3]

Abstract

We present a new discriminative technique for the multiple-source adaptation (MSA) problem. Unlike previous work, which relies on density estimation for each source domain, our solution only requires conditional probabilities that can be straightforwardly accurately estimated from unlabeled data from the source domains. We give a detailed analysis of our new technique, including general guarantees based on Rényi divergences, and learning bounds when conditional Maxent is used for estimating conditional probabilities for a point to belong to a source domain. We show that these guarantees compare favorably to those that can be derived for the generative solution, using kernel density estimation. Our experiments with real-world applications further demonstrate that our new discriminative MSA algorithm outperforms the previous generative solution as well as other domain adaptation baselines.

1. Introduction

Learning algorithms are applied to an increasingly broad array of problems. For some tasks, large amounts of labeled data are available to train very accurate predictors. But, for most new problems or domains, no such supervised information is at the learner's disposal. Furthermore, labeling data is costly since it typically requires human inspection and agreements between multiple expert labelers. Can we leverage past predictors learned for various domains and combine them to devise an accurate one for a new task? Can we provide guarantees for such combined predictors? How should we define that combined predictor? These are some of the challenges of *multiple-source domain adaptation*.

The problem of domain adaptation from multiple sources admits distinct instances defined by the type of source information available to the learner, the number of source domains, and the amount of labeled and unlabeled data available from the target domain (Mansour et al., 2008; 2009a; Hoffman et al., 2018; Pan and Yang, 2010; Muandet et al., 2013; Xu et al., 2014; Hoffman et al., 2012; Gong et al., 2013a;b; Zhang et al., 2015; Ganin et al., 2016; Tzeng et al., 2015; Motiian et al., 2017b;a; Wang et al., 2019b; Konstantinov and Lampert, 2019; Liu et al., 2015; Saito et al., 2019; Wang et al., 2019a). The specific instance we are considering is one where the learner has access to multiple source domains and where, for each domain, they only have at their disposal a predictor trained for that domain and some amount of unlabeled data. No other information about the source domains, in particular no labeled data is available. The target domain or distribution is unknown but it is assumed to be in the convex hull of the source distributions, or relatively close to that. The *multiple-source adaptation (MSA) problem* consists of combining relatively accurate predictors available for each source domain to derive an accurate predictor for *any* such new mixture target domain. This problem was first theoretically studied by Mansour et al. (2008; 2009a) and subsequently by Hoffman et al. (2018; 2021), who further provided an efficient algorithm for this problem and reported the results of a series of experiments with that algorithm and favorable comparisons with alternative solutions.

As pointed out by these authors, this problem arises in a variety of different contexts. In speech recognition, each domain may correspond to a different group of speakers and an acoustic model learned for each domain may be available. Here, the problem consists of devising a general recognizer for a broader population, a mixture of the source domains (Liao, 2013). Similarly, in object recognition, there may be accurate models trained on different image databases and the goal is to come up with an accurate predictor for a general domain, which is likely to be close to a mixture of these sources (Torralba and Efros, 2011). A similar situation often appears in sentiment analysis and various other natural language processing problems where accurate predictors are available for some source domains such as TVs, laptops and CD players, each previously trained on labeled data, but no labeled data or predictor is at hand for the broader category of electronics, which can be viewed as a mixture of the sub-domains (Blitzer et al., 2007; Dredze et al., 2008).

[1] Google Research, New York, NY; [2] Courant Institute of Mathematical Sciences, New York, NY; [3] Hudson River Trading, New York, NY. Correspondence to: Ananda Theertha Suresh <theertha@google.com>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

An additional motivation for this setting of multiple-source adaptation is that often the learner does not have access to labeled data from various domains for legitimate reasons such as privacy or storage limitation. This may be for example labeled data from various hospitals, each obeying strict regulations and privacy rules. But, a predictor trained on the labeled data from each hospital may be available. Similarly, a speech recognition system trained on data from some group may be available but the many hours of source labeled data used to train that model may not be accessible anymore, due to the very large amount of disk space it requires. Thus, in many cases, the learner cannot simply merge all source labeled data to learn a predictor.

Main contributions. In Section 3, we present a new *discriminative technique* for the MSA problem, Previous work showed that a *distribution-weighted combination* of source predictors benefited from favorable theoretical guarantees (Mansour et al., 2008; 2009a; Hoffman et al., 2018; 2021). However, that *generative solution* requires an accurate density estimation for each source domain, which, in general, is a difficult problem. Instead, our solution only needs conditional probabilities, which is easier to accurately estimate from unlabeled data from the source domains. We also describe an efficient DC-programming optimization algorithm for determining the solution of our discriminative technique, which is somewhat similar to but distinct from that of previous work, since it requires a new DC-decomposition.

In Section 4, we give a new and detailed theoretical analysis of our technique, starting with new general guarantees that depend on the Rényi divergences between the target distribution and mixtures of the true source distributions, instead of mixtures of estimates of those distributions (Section 3). We then present finite sample learning bounds for our new discriminative solution when conditional Maxent is used for estimating conditional probabilities. We also give a new and careful analysis of the previous generative solution, when using kernel density estimation, including the first finite sample generalization bound for that technique. We show that the theoretical guarantees for our discriminative solution compare favorably to those derived for the generative solution in several ways. While we benefit from some of the analysis in previous work (Hoffman et al., 2018; 2021), our main proofs and techniques are new and non-trivial.

We further report the results of several experiments with our discriminative algorithm both with a synthetic dataset and several real-world applications (Section 5). Our results demonstrate that, in all tasks, our new solution outperforms the previous work's generative solution, which had been shown itself to surpass empirically the accuracy of other domain adaptation baselines (Hoffman et al., 2018). They also indicate that our discriminative technique requires fewer samples to achieve a high accuracy than the previous solution, which matches our theoretical analysis.

Related work. There is a very broad literature dealing with single-source and multiple-source adaptation with distinct scenarios. Here, we briefly discuss the most related previous work, in addition to (Mansour et al., 2008; 2009a; Hoffman et al., 2018), and defer a more extensive discussion to Appendix A. The idea of using a domain classifier to combine domain-specific predictors has been suggested in the past. Jacobs et al. (1991) and Nowlan and Hinton (1991) considered an adaptive mixture of experts model, where there are multiple expert networks, as well as a gating network to determine which expert to use for each input. The learning method consists of jointly training the individual expert networks and the gating network. In our scenario, no labeled data is available, expert networks are pre-trained separately from the gating network, and our gating network admits a specific structure. Hoffman et al. (2012) learned a domain classifier via SVM on all source data combined, and predicted on new test points with the weighted sum of domain classifier's scores and domain-specific predictors. Such linear combinations were later shown by Hoffman et al. (2018) to perform poorly in some cases and not to benefit from strong guarantees. More recently, Xu et al. (2018) deployed multi-way adversarial training to multiple source domains to obtain a domain discriminator, and also used a weighted sum of discriminator's scores and domain-specific predictors to make predictions. Zhao et al. (2018) considered a scenario where labeled samples are available, unlike our scenario, and learned a domain classifier to approximate the discrepancy term in a MSA generalization bound, and proposed the MDAN model to minimize the bound.

We start with a description of the learning scenario we consider and the introduction of notation and definitions relevant to our analysis (Section 2).

2. Learning Scenario

We consider the MSA problem in the general stochastic scenario studied by Hoffman et al. (2018) and adopt the same notation.

Let $\mathcal{X}$ denote the input space, $\mathcal{Y}$ the output space. We will identify a *domain* with a distribution over $\mathcal{X} \times \mathcal{Y}$. There are p source domains $\mathcal{D}_1, \ldots, \mathcal{D}_p$. As in previous work, we adopt the assumption that the domains share a common conditional probability $\mathcal{D}(\cdot|x)$ and thus $\mathcal{D}_k(x, y) = \mathcal{D}_k(x)\mathcal{D}(y|x)$, for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $k \in [p]$. This is a natural assumption in many common machine learning tasks. For example, in image classification, the label of a picture as a *dog* may not depend much on whether the picture is from a personal collection or a more general dataset. Nevertheless, as discussed in Hoffman et al. (2018), this condition can be relaxed and, here too, all our results can be similarly extended to a more general case where the conditional probabilities vary across domains. Since not all k conditional probabilities are equally accurate on the single x, better target accuracy can be obtained by combining the $\mathcal{D}_k(x)$s in an x-dependent way.

For each domain $\mathcal{D}_k$, $k \in [p]$, the learner has access to some unlabeled data drawn i.i.d. from the marginal distribution $\mathcal{D}_k$ over $\mathcal{X}$, as well as to a predictor h_k. 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:

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

In the probability model, the predictors are assumed to be normalized: $\sum_{y \in \mathcal{Y}} h(x, y) = 1$ for all $x \in \mathcal{X}$. We will denote by $\mathcal{L}(\mathcal{D}, h)$ the expected loss of a predictor h with respect to the distribution $\mathcal{D}$:

$$\begin{aligned} \mathcal{L}(\mathcal{D}, h) &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \big[\ell(h(x), y)\big] \quad \textit{(R)},\\ \mathcal{L}(\mathcal{D}, h) &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \big[\ell(h(x, y))\big] \quad \textit{(P)}. \end{aligned}$$

Our theoretical results are general and only assume that the loss function ℓ is convex, continuous. But, in the regression model, we will be particularly interested in the squared loss $\ell(h(x), y) = (h(x) - y)^2$ and, in the probability model, the cross-entropy loss (or log-loss) $\ell(h(x, y)) = -\log h(x, y)$. We will also assume that each source predictor h_k is ϵ-accurate on its domain for some $\epsilon > 0$, that is, $\forall k \in [p], \mathcal{L}(\mathcal{D}_k, h_k) \leq \epsilon$. Our assumption that the loss of h_k is bounded, implies that $\ell(h_k(x), y) \leq M$ or $\ell(h_k(x, y)) \leq M$, for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $k \in [p]$.

Let $\Delta = \{\lambda = (\lambda_1, \ldots, \lambda_p)\colon \sum_{k=1}^{p} \lambda_k = 1, \lambda_k \geq 0\}$ denote the simplex in $\mathbb{R}^p$, and let $\mathcal{D} = \{\mathcal{D}_\lambda\colon \mathcal{D}_\lambda = \sum_{k=1}^{p} \lambda_k \mathcal{D}_k, \lambda \in \Delta\}$ be the family of all mixtures of the source domains, that is the convex hull of $\mathcal{D}_k$s.

Since not all k source predictors are necessarily equally accurate on the single input x, better target accuracy can be obtained by combining the $h_k(x)$s dependent on x. The MSA problem for the learner is exactly how to combine these source predictors h_k to design a predictor h with small expected loss for any unknown target domain $\mathcal{D}_T$ that is an element of $\mathcal{D}$, or any unknown distribution $\mathcal{D}_T$ close to $\mathcal{D}$.

Our theoretical guarantees are presented in terms of *Rényi divergences*, a broad family of divergences between distributions generalizing the relative entropy. The Rényi Divergence is parameterized by $\alpha \in [0, +\infty]$ and denoted by D_α. The α-Rényi Divergence between two distributions $\mathcal{P}$ and $\mathcal{Q}$ is defined by:

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

where, for $\alpha \in \{0, 1, +\infty\}$, the expression is defined by taking the limit (Arndt, 2004). For $\alpha = 1$, the Rényi divergence coincides with the relative entropy. We will denote by $\mathsf{d}_\alpha(\mathcal{P} \parallel \mathcal{Q})$ the exponential of $\mathsf{D}_\alpha(\mathcal{P} \parallel \mathcal{Q})$:

$$\mathsf{d}_\alpha(\mathcal{P} \parallel \mathcal{Q}) = \Bigg[\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \frac{\mathcal{P}^\alpha(x, y)}{\mathcal{Q}^{\alpha-1}(x, y)}\Bigg]^{\frac{1}{\alpha-1}}.$$

Appendix B provides more background on the definition and the main properties of Rényi divergences.

In the following, to alleviate the notation, we abusively denote the marginal distribution of a distribution $\mathcal{D}_k$ defined over $\mathcal{X} \times \mathcal{Y}$ in the same way and rely on the arguments for disambiguation, e.g. $\mathcal{D}_k(x)$ vs. $\mathcal{D}_k(x, y)$.

3. Discriminative MSA solution

In this section we present our new solution for the MSA problem and give an efficient algorithm for determining its parameter. But first we describe the previous solution.

3.1. Previous Generative Technique

In previous work, it was shown that, in general, standard convex combinations of source predictors can perform poorly (Mansour et al., 2008; 2009a; Hoffman et al., 2018): in some problems, even when the source predictors have zero loss, no convex combination can achieve a loss below some constant for a uniform mixture of the source distributions. Instead, a *distribution-weighted* solution was proposed to the MSA problem. That solution relies on density estimates $\widehat{\mathcal{D}}_k$ for the marginal distributions $x \mapsto \mathcal{D}_k(x)$, which are obtained via techniques such as kernel density estimation, for each source domain $k \in [p]$ independently.

Given such estimates, the solution is defined as follows in the regression and probability models, for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$:

$$\widehat{h}_z(x) = \sum_{k=1}^{p} \frac{z_k \widehat{\mathcal{D}}_k(x)}{\sum_{j=1}^{p} z_j \widehat{\mathcal{D}}_j(x)} h_k(x), \tag{1}$$

$$\widehat{h}_z(x, y) = \sum_{k=1}^{p} \frac{z_k \widehat{\mathcal{D}}_k(x)}{\sum_{j=1}^{p} z_j \widehat{\mathcal{D}}_j(x)} h_k(x, y), \tag{2}$$

with $z \in \Delta$ is a parameter determined via an optimization problem such that h_z admits the same loss for all $\mathcal{D}_k$. We are assuming here that the estimates verify $\widehat{\mathcal{D}}_k(x) > 0$ for all $x \in \mathcal{X}$ and therefore that the denominators are positive. Otherwise, a small positive number $\eta > 0$ can be added to the denominators of the solutions, as in previous work. We are adopting this assumption only to simplify the presentation. For the probability model, the joint estimates $\widehat{\mathcal{D}}_k(x, y)$ used in (Hoffman et al., 2018) can be equivalently replaced by marginal ones $\widehat{\mathcal{D}}_k(x)$ since all domain distributions share the same conditional probabilities.

Since this previous work relies on density estimation, we will refer to it as a *generative solution to the MSA problem*, in short, GMSA. The technique benefits from the following general guarantee (Hoffman et al., 2018), where we extend the Rényi divergences to divergences between a distribution $\mathcal{D}$ and a set of distributions $\mathcal{D}$ and write $\mathsf{D}_\alpha(\mathcal{D} \parallel \mathcal{D}) = \min_{\mathcal{D} \in \mathcal{D}} \mathsf{D}_\alpha(\mathcal{D} \parallel \mathcal{D})$.

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

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

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

The bound depends on the quality of the density estimates via the Rényi divergence between $\widehat{\mathcal{D}}_k$ and $\mathcal{D}_k$, for each $k \in [p]$, and the closeness of the target distribution $\mathcal{D}_T$ to the mixture family $\widehat{\mathcal{D}}$, a bound we elaborate on further in Appendix C.1 and express in terms of the closeness of the target distribution $\mathcal{D}_T$ to the true family $\mathcal{D}$. For $\alpha = +\infty$, for $\mathcal{D}_T$ close to $\widehat{\mathcal{D}}$ and accurate estimates of $\mathcal{D}_k$, $\mathsf{d}_\alpha(\mathcal{D}_T \parallel \widehat{\mathcal{D}})$ and $\mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)$ are close to one and the upper bound is as a result close to ϵ. That is, with good density estimates, the error of h_z is no worse than that of the source predictors h_ks. However, obtaining good density estimators is a difficult problem and in general requires large amounts of data. In the following section, we provide a new and less data-demanding solution based on conditional probabilities.

3.2. New Discriminative Technique

Let $\mathcal{D}$ denote the distribution over $\mathcal{X}$ defined by $\mathcal{D}(x) = \frac{1}{p} \sum_{k=1}^p \mathcal{D}_k(x)$. We will assume and can enforce that $\mathcal{D}$ is the distribution according to which we can expect to receive unlabeled samples from the p sources to train our discriminator. We will denote by $\mathcal{Q}$ the distribution over $\mathcal{X} \times [p]$ defined by $\mathcal{Q}(x, k) = \frac{1}{p} \mathcal{D}_k(x)$, whose $\mathcal{X}$-marginal coincides with $\mathcal{D}$: $\mathcal{Q}(x) = \mathcal{D}(x)$.

Our new solution relies on estimates $\widehat{\mathcal{Q}}(k|x)$ of the conditional probabilities $\mathcal{Q}(k|x)$ for each domain $k \in [p]$, that is the probability that point x belongs to source k. Given such estimates, our new solution to the MSA problem is defined as follows in the regression and probability models, for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$:

$$\widehat{g}_z(x) = \sum_{k=1}^p \frac{z_k \widehat{\mathcal{Q}}(k|x)}{\sum_{j=1}^p z_j \widehat{\mathcal{Q}}(j|x)} h_k(x), \tag{3}$$

$$\widehat{g}_z(x, y) = \sum_{k=1}^p \frac{z_k \widehat{\mathcal{Q}}(k|x)}{\sum_{j=1}^p z_j \widehat{\mathcal{Q}}(j|x)} h_k(x, y), \tag{4}$$

with $z \in \Delta$ being a parameter determined via an optimization problem. As for the GMSA solution, we are assuming here that the estimates verify $\widehat{\mathcal{Q}}(k|x) > 0$ for all $x \in \mathcal{X}$ and therefore that the denominators are positive. Otherwise, a small positive number $\eta > 0$ can be added to the denominators of the solutions, as in previous work. We are adopting this assumption only to simplify the presentation. Note that in the probability model, $\widehat{g}_z(x, y)$ is normalized since h_ks are normalized: $\sum_{y \in \mathcal{Y}} g_z(x, y) = 1$ for all $x \in \mathcal{X}$.

Since our solution relies on estimates of conditional probabilities of domain membership, we will refer to it as a *discriminative solution to the MSA problem*, DMSA in short.

Observe that, by the Bayes' formula, the conditional probability estimates $\widehat{\mathcal{Q}}(k|x)$ induce density estimates $\widehat{\mathcal{D}}_k(x)$ of the marginal distributions $x \mapsto \mathcal{D}_k(x)$:

$$\widehat{\mathcal{D}}_k(x) = \frac{\widehat{\mathcal{Q}}(k|x)\mathcal{D}(x)}{\widehat{\mathcal{Q}}(k)} \tag{5}$$

where $\widehat{\mathcal{Q}}(k) = \sum_{x \in \mathcal{X}} \widehat{\mathcal{Q}}(k|x)\mathcal{D}(x)$. For an exact estimate, that is $\widehat{\mathcal{Q}}(k|x) = \mathcal{Q}(k|x)$, the formula holds with $\widehat{\mathcal{Q}}(k) = \sum_{x \in \mathcal{X}} \mathcal{Q}(x, k) = \frac{1}{p}$. In light of this observation, we can establish the following connection between the GMSA and DMSA solutions.

Proposition 1. *Let $\widehat{h}_z$ be the GMSA solution using the estimates $\widehat{\mathcal{D}}_k$ defined in (5). Then, for any $z \in \Delta$, we have $\widehat{h}_z = \widehat{g}_{z'}$ with $z'_k = \frac{z_k / \widehat{\mathcal{Q}}(k)}{\sum_{j=1}^p z_j / \widehat{\mathcal{Q}}(j)}$, for all $k \in [p]$.*

Proof. First consider the regression model. By definition of the GMSA solution, we can write:

$$\begin{aligned}
\widehat{h}_z(x) &= \sum_{k=1}^p \frac{z_k \frac{\widehat{\mathcal{Q}}(k|x)\mathcal{D}(x)}{\widehat{\mathcal{Q}}(k)}}{\sum_{j=1}^p z_j \frac{\widehat{\mathcal{Q}}(j|x)\mathcal{D}(x)}{\widehat{\mathcal{Q}}(j)}} h_k(x) \\
&= \sum_{k=1}^p \frac{\frac{z_k}{\widehat{\mathcal{Q}}(k)} \widehat{\mathcal{Q}}(k|x)}{\sum_{j=1}^p \frac{z_j}{\widehat{\mathcal{Q}}(j)} \widehat{\mathcal{Q}}(j|x)} h_k(x) = g_{z'}(x).
\end{aligned}$$

The probability model's proof is syntactically the same. □

In view of this result, the DMSA technique benefits from a guarantee similar to GMSA (Theorem 1), where for DMSA the density estimates are based on the conditional probability estimates $\widehat{\mathcal{Q}}(k|x)$.

Theorem 2. *For any $\delta > 0$, there exists a $z \in \Delta$ such that the following inequality holds for any $\alpha > 1$ and arbitrary target distribution $\mathcal{D}_T$:*

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

where $\widehat{\epsilon} = \max_{k \in [p]} \Big[\epsilon \, \mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)\Big]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}$, and $\widehat{\mathcal{D}} = \Big\{\sum_{k=1}^p \lambda_k \widehat{\mathcal{D}}_k \colon \lambda \in \Delta\Big\}$, with $\widehat{\mathcal{D}}_k(x, y) = \frac{\widehat{\mathcal{Q}}(k|x)\mathcal{D}(x,y)}{\widehat{\mathcal{Q}}(k)}$.

3.3. Optimization Algorithm

By Proposition 1, to determine the parameter z' guaranteeing the bound of Theorem 2 for $\widehat{g}_{z'}$, it suffices to determine the parameter z that yields the guarantee of Theorem 1 for $\widehat{h}_z$, when using the estimates $\widehat{\mathcal{D}}_k = \frac{\widehat{\mathcal{Q}}(k|x)\mathcal{D}(x)}{\widehat{\mathcal{Q}}(k)}$. As shown by (Hoffman et al., 2018), the parameter z is the one for which $\widehat{h}_z$ admits the same loss for all source domains, that is $\mathcal{L}(\widehat{\mathcal{D}}_k, \widehat{h}_z) = \mathcal{L}(\widehat{\mathcal{D}}_{k'}, \widehat{h}_z)$ for all $k, k' \in [p]$, where $\widehat{\mathcal{D}}_k$ is the joint distribution derived from $\widehat{\mathcal{D}}_k$: $\widehat{\mathcal{D}}_k(x, y) = \widehat{\mathcal{D}}_k(x)\mathcal{D}(y|x) = \frac{\widehat{\mathcal{Q}}(k|x)\mathcal{D}(x,y)}{\widehat{\mathcal{Q}}(k)}$, with $\mathcal{D}(x, y) = \frac{1}{p}\sum_{k=1}^{p} \mathcal{D}_k(x, y)$. Note, $\widehat{\mathcal{D}}(x, y)$ is abusively denoted the same way as $\widehat{\mathcal{D}}(x)$ to avoid the introduction of additional notation, but the difference in arguments should suffice to help distinguish the two distributions.

Thus, using $\widehat{g}_{z'} = \widehat{h}_z$, to find z, and subsequently z', it suffices to solve the following optimization problem in z:

$$\min_{z \in \Delta} \max_{k \in [p]} \quad \mathcal{L}(\widehat{\mathcal{D}}_k, \widehat{g}_{z'}) - \mathcal{L}(\widehat{\mathcal{D}}_z, \widehat{g}_{z'}), \tag{6}$$

where $z'_k = \frac{z_k / \widehat{\mathcal{Q}}(k)}{\sum_{j=1}^{p} z_j / \widehat{\mathcal{Q}}(j)}$ and $\widehat{\mathcal{D}}_z = \sum_{k=1}^{p} z_k \widehat{\mathcal{D}}_k$. As in previous work, this problem can be cast as a DC-programming (difference-of-convex) problem and solved using the DC algorithm (Tao and An, 1997; 1998; Sriperumbudur and Lanckriet, 2012). However, we need to derive a new DC-decomposition here, both for the regression and the probability model, since the objective is distinct from that of previous work. A detailed description of that DC-decomposition and its proofs, as well as other details of the algorithm are given in Appendix D.

4. Learning Guarantees

In this section, we prove favorable learning guarantees for the predictor $\widehat{g}_z$ returned by `DMSA`, when using conditional maximum entropy to derive domain estimates $\mathcal{Q}(k|x)$. We first extend Theorem 1 and present a general theoretical guarantee which holds for `DMSA` and `GMSA` (Section 4.1). Next, in Section 4.2, we give a generalization bound for conditional Maxent and use that to prove learning guarantees for `DMSA`. We then analyze `GMSA` using kernel density estimation (Section 4.3), and show that `DMSA` benefits from significantly more favorable learning guarantees than `GMSA`.

4.1. General Guarantee

Theorem 1 gives a guarantee in terms of the Rényi divergence of $\mathcal{D}_T$ and $\widehat{\mathcal{D}}$, which depends on the empirical estimates. Instead, we derive a bound in terms of the Rényi divergence of $\mathcal{D}_T$ and $\mathcal{D}$ and, as with Theorem 1, the Rényi divergences between the distributions $\mathcal{D}_k$ and their estimates $\widehat{\mathcal{D}}_k$.

To do so, we use an inequality that can be viewed as a triangle inequality result for Rényi divergences (Hoffman et al., 2021).

Proposition 2. *Let $\mathcal{P}$, $\mathcal{Q}$, $\mathcal{R}$ be three distributions on $\mathcal{X} \times \mathcal{Y}$. Then, for any $\gamma \in (0, 1)$ and any $\alpha > \gamma$, the following inequality holds:*

$$\Big[\mathsf{d}_\alpha(\mathcal{P} \parallel \mathcal{Q})\Big]^{\alpha-1} \leq \Big[\mathsf{d}_{\frac{\alpha}{\gamma}}(\mathcal{P} \parallel \mathcal{R})\Big]^{\alpha-\gamma} \Big[\mathsf{d}_{\frac{\alpha-\gamma}{1-\gamma}}(\mathcal{R} \parallel \mathcal{Q})\Big]^{\alpha-1}.$$

The proof is given in Appendix B. This result is used in combination with Theorem 1 to establish the following.

Theorem 3. *For any $\delta > 0$, there exists $z \in \Delta$ such that the following inequality holds for any $\alpha > 1$ and arbitrary target distribution $\mathcal{D}_T$:*

$$\mathcal{L}(\mathcal{D}_T, \widehat{g}_z) \leq [(\widehat{\epsilon} + \delta)\, \widehat{\mathsf{d}}']^{\frac{\alpha-1}{\alpha}} [\mathsf{d}_{2\alpha}(\mathcal{D}_T \parallel \mathcal{D})]^{\frac{2\alpha-1}{2\alpha}} M^{\frac{1}{\alpha}},$$

where $\widehat{\epsilon} = (\epsilon \widehat{\mathsf{d}})^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}$, $\widehat{\mathsf{d}} = \max_{k \in [p]} \mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)$, and $\widehat{\mathsf{d}}' = \max_{k \in [p]} \mathsf{d}_{2\alpha-1}(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k)$, with $\widehat{\mathcal{D}}_k = \frac{\widehat{\mathcal{Q}}(k|x)\mathcal{D}(x)}{\widehat{\mathcal{Q}}(k)}$.

The proof is given in Appendix C.1. The theorem holds similarly for `GMSA` with $\widehat{\mathcal{D}}_k$ a direct estimate of $\mathcal{D}_k$ (Theorem 9, Appendix E.2). This provides a strong performance guarantee for `GMSA` or `DMSA` when the target distribution $\mathcal{D}_T$ is close to the family of mixtures of the source distributions $\mathcal{D}_k$, and when $\widehat{\mathcal{D}}_k$ is a good estimate of $\mathcal{D}_k$.

4.2. Conditional Maxent

The distribution $\mathcal{D} = \frac{1}{p}\sum_{k=1}^{p} \mathcal{D}_k$ over $\mathcal{X} \times \mathcal{Y}$ naturally induces the distribution $\mathcal{Q}$ over $\mathcal{X} \times [p]$ defined for all (x, k) by $\mathcal{Q}(x, k) = \frac{1}{p} D_k(x)$. Let $S = ((x_1, k_1), \ldots, (x_m, k_m))$ be a sample of m labeled points drawn i.i.d. from $\mathcal{Q}$.

Let $\Phi\colon \mathcal{X} \times [p] \to \mathbb{R}^N$ be a feature mapping with bounded norm, $\|\Phi\| \leq r$, for some $r > 0$. Then, the optimization problem defining the solution of conditional Maxent (or multinomial logistic regression) with the feature mapping Φ is given by

$$\min_{w \in \mathbb{R}^N} \mu \|w\|^2 - \frac{1}{m}\sum_{i=1}^{m} \log \mathsf{p}_w[k_i | x_i], \tag{7}$$

where p_w is defined by $\mathsf{p}_w[k|x] = \frac{1}{Z(x)} \exp(w \cdot \Phi(x, k))$, with $Z(x) = \sum_{k \in [p]} \exp(w \cdot \Phi(x, k))$, and where $\mu \geq 0$ is a regularization parameter. Then, conditional Maxent benefits from the following theoretical guarantee.

Theorem 4. *Let $\widehat{w}$ be the solution of problem (7) and w^* the population solution of the conditional Maxent optimization problem:*

$$w^* = \operatorname*{argmin}_{w \in \mathbb{R}^N} \mu \|w\|^2 - \mathop{\mathbb{E}}_{(x,k) \sim \mathcal{Q}} \Big[\log \mathsf{p}_w[k|x]\Big].$$

Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $(x, k) \in \mathcal{X} \times [p]$, the following inequality holds:

$$\Big|\log \mathsf{p}_{\widehat{w}}[k|x] - \log \mathsf{p}_{w^*}[k|x]\Big| \leq \frac{2\sqrt{2}r^2}{\mu\sqrt{m}}\Big[1 + \sqrt{\log(1/\delta)}\Big].$$

The theorem shows that the pointwise log-loss of the conditional Maxent solution $\mathsf{p}_{\widehat{w}}$ is close to that of the best-in-class p_{w^*} modulo a term in $O(1/\sqrt{m})$ that does not depend on the dimension of the feature space. The proof is given in Appendix C.2.

4.3. Comparison of the Guarantees for DMSA and GMSA

We now use Theorem 3 and the bound of Theorem 4 to give a theoretical guarantee for DMSA used with conditional Maxent. We show that it is more favorable than a guarantee for GMSA using kernel density estimation.

Theorem 5 (DMSA)**.** *There exists $z \in \Delta$ such that for any $\delta > 0$, with probability at least $1 - \delta$ the following inequality holds DMSA used with conditional Maxent, for an arbitrary target mixture $\mathcal{D}_T$:*

$$\mathcal{L}(\mathcal{D}_T, \widehat{g}_z) \leq \epsilon\, p\, e^{\frac{6\sqrt{2}r^2}{\mu\sqrt{m}}\left[1+\sqrt{\log(1/\delta)}\right]}\, \mathsf{d}^*\, \mathsf{d}'^*,$$

$$\textit{with} \quad \mathsf{d}^* = \sup_{x\in\mathcal{X}} \mathsf{d}_\infty\left(\mathcal{Q}^*[\cdot|x] \parallel \mathcal{Q}(\cdot|x)\right)$$

$$\mathsf{d}'^* = \sup_{x\in\mathcal{X}} \mathsf{d}^2_\infty\left(\mathcal{Q}(\cdot|x) \parallel \mathcal{Q}^*[\cdot|x]\right),$$

where $\mathcal{Q}^*(\cdot|x) = \mathsf{p}_{w^*}[\cdot|x]$ is the population solution of conditional Maxent problem (statement of Theorem 4).

The proof is given in Appendix C.3. It is based on a new and careful analysis of the Rényi divergences, leveraging the guarantee of Theorem 4. More refined versions of these results with alternative Rényi divergence parameters and with expectations instead of suprema in the definitions of d^* and d'^* are presented in that same appendix. The theorem shows that the expected error of DMSA with conditional Maxent is close to ϵ modulo a factor that varies as $e^{1/\sqrt{m}}$, where m is the size of the total unlabeled sample received from all p sources, and factors $\mathcal{Q}^*$ and $\mathcal{Q}'^*$ that measure how closely conditional Maxent can approximate the true conditional probabilities with infinite samples.

Next, we prove learning guarantees for GMSA with densities estimated via kernel density estimation (KDE). We assume that the same i.i.d. sample $S = ((x_1, k_1), \ldots, (x_m, k_m))$ as with conditional Maxent is used. Here, the points labeled with k are used for estimating $\mathcal{D}_k$ via KDE. Since the sample is drawn from $\mathcal{Q}$ with $\mathcal{Q}(x, k) = \frac{1}{p}\mathcal{D}_k$, the number of samples points m_k labeled with k is very close to $\frac{m}{p}$. $\widehat{\mathcal{D}}_k$ is learned from m_k samples, via KDE with a normalized kernel function $K_\sigma(\cdot, \cdot)$ that satisfies $\int_{x\in\mathcal{X}} K_\sigma(x, x')\, dx = 1$ for all $x' \in \mathcal{X}$.

Theorem 6 (GMSA)**.** *There exists $z \in \Delta$ such that, for any $\delta > 0$, with probability at least $1 - \delta$ the following inequality holds for GMSA used KDE, for an arbitrary target mixture $\mathcal{D}_T$:*

$$\mathcal{L}(\mathcal{D}_T, \widehat{h}_z) \leq \epsilon^{\frac{1}{4}} M^{\frac{3}{4}} e^{\frac{6\kappa}{\sqrt{2(m/p)}}\sqrt{\log p + \log(1/\delta)}}\mathsf{d}^*\mathsf{d}'^*,$$

with $\kappa = \max_{x,x',x''\in\mathcal{X}} \frac{K_\sigma(x,x')}{K_\sigma(x,x'')}$, and

$$\mathsf{d}^* = \max_{k\in[p]} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_k} \left[\mathsf{d}_{+\infty}\big(K_\sigma(\cdot, x) \parallel \mathcal{D}_k\big)\right],$$

$$\mathsf{d}'^* = \max_{k\in[p]} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_k} \left[\mathsf{d}_{+\infty}\big(\mathcal{D}_k \parallel K_\sigma(\cdot, x)\big)\right].$$

The proof is given in Appendix E.2. More refined versions of these results with alternative Rényi divergences are presented in that same appendix. In comparison with the guarantee for DMSA, the bound for GMSA admits a worse dependency on ϵ. Furthermore, while the dependency of the learning bound of DMSA on the sample size is of the form $O(e^{1/\sqrt{m}})$ and thus decreases as a function of the full sample size m, that of GMSA is of the form $O(e^{1/\sqrt{m/p}})$ and only decreases as a function of the per-domain sample size. This further reflects the benefit of our discriminative solution since the estimation of the conditional probabilities is based on conditional Maxent trained on the full sample. Finally, the bound of GMSA depends on κ, a ratio that can be unbounded for Gaussian kernels commonly used for KDE.

The generalization guarantees for DMSA (Theorem 7) depends on two critical terms that measure the divergence between the population solution of conditional Maxent and the true domain classifier $\mathcal{Q}(\cdot|x)$:

$$\mathsf{d}_{+\infty}\big(\mathcal{Q}^*(\cdot|x) \parallel \mathcal{Q}(\cdot|x)\big) \quad \text{and} \quad \mathsf{d}_{+\infty}\big(\mathcal{Q}(\cdot|x) \parallel \mathcal{Q}^*(\cdot|x)\big).$$

When the feature mapping for conditional Maxent is sufficiently rich, for example when it is the reproducing kernel Hilbert space (RKHS) associated to a Gaussian kernel, one can expect the two divergences to be close to one. The generalization guarantees for GMSA (Theorem 10) also depend on two divergence terms:

$$\mathsf{d}_{+\infty}\big(K_\sigma(\cdot, x) \parallel \mathcal{D}_k\big) \quad \text{and} \quad \mathsf{d}_{+\infty}\big(\mathcal{D}_k \parallel K_\sigma(\cdot, x)\big).$$

Compared to learning a domain classifier $\widehat{\mathcal{Q}}(\cdot|x)$, it is more difficult to chose a good density kernel $K_\sigma(\cdot, \cdot)$ to ensure that the divergence between marginal distributions is small, which shows another benefit of DMSA.

The next section further illustrates the more advantageous sample complexity of the DMSA algorithm and shows that, in addition to the theoretical advantages discussed in this section, it also benefits from more favorable empirical results.

Table 1. MSE on the sentiment analysis dataset. Single source baselines, K, D, B, E, the uniform combination unif, GMSA, and DMSA.

	Sentiment Analysis Test Data										
	K	D	B	E	KD	BE	DBE	KBE	KDB	KDB	KDBE
K	1.42±0.10	2.20±0.15	2.35±0.16	1.67±0.12	1.81±0.07	2.01±0.10	2.07±0.08	1.81±0.06	1.76±0.06	1.99±0.06	1.91±0.05
D	2.09±0.08	1.77±0.08	2.13±0.10	2.10±0.08	1.93±0.07	2.11±0.07	2.00±0.06	2.11±0.06	1.99±0.06	2.00±0.06	2.02±0.05
B	2.16±0.13	1.98±0.10	1.71±0.12	2.21±0.07	2.07±0.11	1.96±0.07	1.97±0.06	2.03±0.06	2.12±0.07	1.95±0.08	2.02±0.06
E	1.65±0.09	2.35±0.11	2.45±0.14	1.50±0.07	2.00±0.09	1.97±0.09	2.10±0.08	1.86±0.05	1.83±0.07	2.15±0.07	1.99±0.06
unif	1.50±0.06	**1.75±0.09**	1.79±0.10	1.53±0.07	1.63±0.06	1.66±0.08	1.69±0.06	1.61±0.05	1.60±0.05	1.68±0.05	1.65±0.05
GMSA	1.42±0.10	1.88±0.11	1.80±0.10	1.51±0.07	1.65±0.08	1.66±0.07	1.73±0.05	1.58±0.04	1.60±0.05	1.70±0.04	1.65±0.04
DMSA (ours)	**1.42±0.08**	1.76±0.07	**1.70±0.11**	**1.46±0.07**	**1.59±0.06**	**1.58±0.07**	**1.64±0.05**	**1.53±0.04**	**1.55±0.04**	**1.63±0.04**	**1.59±0.04**

5. Experiments

We experimented with our DMSA technique on the same datasets as those used in (Hoffman et al., 2018), as well as with the UCI adult dataset, and compared its performance with several baselines, including GMSA. Since Hoffman et al. (2018) already showed that GMSA empirically outperforms alternative MSA solutions, in this section, we mainly focus on demonstrating improvements over GMSA under the same experimental setups.

Sentiment analysis. To evaluate the DMSA solution under the regression model, we used the sentiment analysis dataset (Blitzer et al., 2007), which 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 adopted the same training procedure and hyper-parameters as those used by Hoffman et al. (2018) to obtain base predictors: first define a vocabulary of 2,500 words that occur at least twice in each of the four domains, then use this vocabulary to define word-count feature vectors for every review text, and finally train base predictors for each domain using support vector regression. We used the same word-count features to train the domain classifier via logistic regression. We randomly split the 2,000 samples per domain into 1,600 train and 400 test samples for each domain, and learn the base predictors, domain classifier, density estimations, and parameter z for both MSA solutions on all available training samples. We repeated the process 10 times, and report the mean and standard deviation of the mean squared error on various target test mixtures in Table 1.

We compared our technique, DMSA, against each source predictor, h_k, the uniform combination of the source predictors (unif), $\frac{1}{p}\sum_{k=1}^{p} h_k$, and GMSA with kernel density estimation. Each column in Table 1 corresponds to a different target test mixture, as indicated by the column name: four single domains, and uniform mixtures of two, three, and four domains, respectively. Our distribution-weighted method DMSA outperforms all baseline predictors across almost all test domains. Observe that, even when the target is a single source domain, such as K, B, E, our method can still outperform the predictor which is trained and tested on the same domain, showing the benefits of ensembles. Moreover, DMSA improves upon GMSA by a wide margin on all test mixtures, which demonstrates the advantage of using a domain classifier over estimated densities in the distribution-weighted combination.

Digit dataset. To evaluate the DMSA solution under the probability model, we considered a digit recognition task consisting of three datasets: Google Street View House Numbers (SVHN), MNIST, and USPS. For each individual domain, we trained a convolutional neural network (CNN) with the same setup as in (Hoffman et al., 2018), and used the output from the softmax score layer as our base predictors h_k. Furthermore, for every input image, we extracted the last layer before softmax from each of the base networks and concatenated them to obtain the feature vector for training the domain classifier. We used the full training sets per domain to train the source model, and used $6{,}000$ samples per domain to learn the domain classifier. Finally, for our DC-programming algorithm, we used a $1{,}000$ image-label pairs from each domain, thus a total of $3{,}000$ labeled pairs to learn the parameter z.

We compared our DMSA algorithm against each source predictor (h_k), the uniform combination, unif, a network jointly trained on all source data combined, joint, and GMSA with kernel density estimation. Since the training and testing datasets are fixed, we simply report the numbers from the original GMSA paper. We measured the performance of these baselines on each of the three test datasets, on combinations of two test datasets, and on all test datasets combined. The results are reported in Table 2. Once again, DMSA outperforms all baselines on all test mixtures, and when the target is a single test domain, DMSA admits a comparable performance to the predictor that is trained and tested on the same domain. And, as in the sentiment analysis experiments, DMSA outperforms GMSA by a wide margin on most of the test domains. For example, on SVHN test data, the improvement is 0.9%, which is larger than 0.5%, the standard deviation estimate on the test data.

We also report empirical results for the adversarial domain adaptation method of Zhao et al. (2018) in Table 2. Let us emphasize that the learning scenario for this algorithm does not match ours: this algorithm makes use of labeled data from source domains, as well as unlabeled data from a fixed

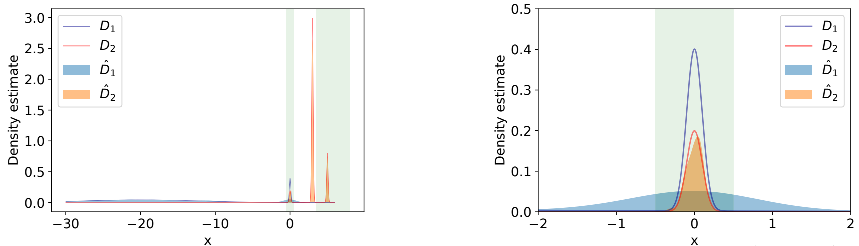

Figure 1. Left: Illustration of the true densities and kernel density estimates for GMSA for domains $\mathcal{D}_1$ and $\mathcal{D}_2$ with 1000 samples. The labeling function $f(x) = -1$ in the green regions and 1 otherwise. Right: Same estimates zoomed in at $x = 0$.

Table 2. Digit Dataset Accuracy. DMSA outperforms each single-source domain model, unif, joint, and most importantly GMSA, on various target mixtures.

	Digits 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
CNN-joint	90.9	99.1	96.0	98.6	91.3	93.2	93.3	94.6
adv-mu	91.5	98.5	95.7	98.1	91.8	93.5	93.6	94.7
adv-su	91.6	98.5	95.7	98.0	91.9	93.5	93.6	94.7
adv-sm	91.8	98.3	95.3	97.8	92.1	93.6	93.7	94.7
GMSA	91.4	98.8	95.6	98.3	91.7	93.5	93.6	94.7
DMSA (ours)	92.3	99.2	96.6	**98.8**	**92.6**	**94.2**	**94.3**	**95.4**

target domain. In other words, the algorithm makes use of more information than what is available in our scenario or accessible to DMSA. Nevertheless, we are including these results for reference.

For a target domain formed by the union of two out of the three domains svhn, mnist, or usps, that is a target domain defined as sm, su, or mu, we trained adv-target-domain, where we used unlabeled data from the target domain and labeled examples from all the three source domains smu. For these experiments, we used the entire training data from source domains and the entire unlabeled training data from target domains. We used the neural architecture and the discriminator used by Zhao et al. (2018). The results show that, while the adv-target-domain algorithm (Zhao et al., 2018) is making use of more information, its performance is inferior to that of GMSA and DMSA, even for the specific target distribution it is trained for and that it has therefore extra information about.

In Tables 5 and 6 in Appendix F, we report additional experimental results with the digits dataset for the scenario where the target domain is close to being a mixture of the source domains but where it may not necessarily be such a mixture, a scenario not covered by Hoffman et al. (2018). These experiments also demonstrate a consistently strong performance of DMSA.

To illustrate the efficiency of DMSA we further tested DMSA and GMSA on the digits dataset when only a small amount of data is available for domain adaptation. We plotted the performance of both algorithms as a function of m, the number of samples per domain, see Figure 2. As expected, DMSA consistently outperforms GMSA, especially in the small sample regime, thus matching our theoretical analysis that DMSA can succeed with fewer samples.

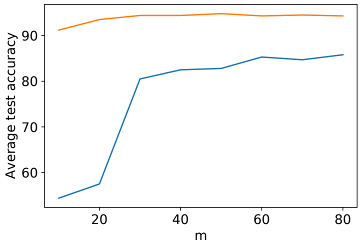

Figure 2. Average test accuracy of GMSA (blue) and DMSA (orange) on the digits dataset as a function of the number of samples used in domain adaptation.

Adult dataset. We also experimented with the UCI adult dataset (Blake, 1998), which contains 32,561 training samples with numerical and categorical features, each representing a person. The task consists of predicting if the person's income exceeds $50,000. Following (Mohri et al., 2019), we split the dataset into two domains, the doctorate Doc domain and non-doctorate NDoc domain and used categorical features for training linear classification models. We froze these models and experimented with the MSA methods GMSA and DMSA. Here, we repeatedly sampled 400 training samples from each domain for training, keeping the test set fixed.

Table 3. Linear models for adult dataset. The experiments are averaged over 100 runs.

Test data	Doc	NDoc	Doc-NDoc
GMSA	70.2 ± 1.2	76.4 ± 1.6	73.3 ± 0.8
DMSA	70.0 ± 0.8	$\mathbf{80.5} \pm \mathbf{0.5}$	$\mathbf{75.3} \pm \mathbf{0.4}$

The results are reported in Table 3. DMSA achieves a higher accuracy than GMSA on the NDoc domain and also in the

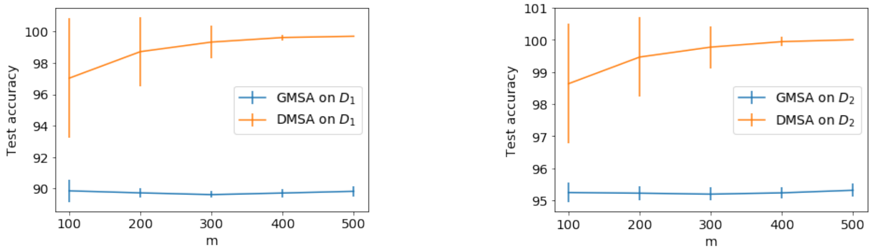

Figure 3. Comparison of GMSA and DMSA on the synthetic dataset. DMSA performs better than GMSA on both domains and thus on any convex combination. The experiments are averaged over 10 runs; error bars show one standard deviation.

average of two domains. The difference in performance is not statistically significant for the Doc domain as it has very few test samples.

Office dataset. We also carried out experiments on the visual adaptation office dataset (Saenko et al., 2010). The Office dataset is composed of 3 domains: amazon, dslr, and webcam. The amazon domain consists of 2817 images, dslr 498, and webcam 795 images. We divided the dataset into two splits following (Saenko et al., 2010). For the training data, we used 20 samples per category for amazon and 8 for both dslr and webcam. We used the rest of the samples as test data. We extracted the penultimate layer output from ResNet50 architecture (He et al., 2015) pre-trained on ImageNet and trained logistic regression models as base classifiers using this pretrained feature. The results are shown in Table 4. DMSA outperforms GMSA in all three domains and thus any convex combination. The differences for amazon and webcam is less than a standard deviation, however, we observe the advantage of DMSA over GMSA consistently. Similarly, DMSA achieves a higher accuracy than ResNet-unif, especially in the amazon domain, for which its performance matches that of a model specifically trained for that domain.

Table 4. Office Dataset Accuracy. The experiments are averaged over 10 runs.

Test data	amazon	webcam	dslr
ResNet-amazon	$\mathbf{82.2 \pm 0.6}$	75.8 ± 1.3	77.6 ± 1.4
ResNet-webcam	63.3 ± 1.6	95.7 ± 1.0	95.7 ± 1.3
ResNet-dslr	64.6 ± 1.0	94.0 ± 0.7	95.8 ± 1.0
ResNet-unif	79.3 ± 0.6	96.7 ± 0.7	97.2 ± 0.6
GMSA	82.1 ± 0.4	96.8 ± 0.8	96.7 ± 0.6
DMSA	$\mathbf{82.2 \pm 0.4}$	$\mathbf{97.2 \pm 0.9}$	$\mathbf{97.4 \pm 0.4}$

Synthetic dataset. We finally conducted simulations on a small synthetic dataset to further illustrate the difference between GMSA and DMSA. We used the sklearn toolkit for these experiments. Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be Gaussian mixtures in one dimension defined as follows: $\mathcal{D}_1 = 0.9 \cdot N(-20, 8) + 0.1 \cdot N(0, 0.1)$ and $\mathcal{D}_2 = 0.75 \cdot N(3, 0.1) + 0.25 \cdot N(5, 0.1) + 0.05 \cdot N(0, 0.1)$, see Figure 1. The two domains are similar around 0 but are disjoint otherwise. Let the labeling function $f(x) = -1$ if $x \in [-0.5, 0.5] \cup [3.5, +\infty)$. The example is designed such that if their estimates are good, then both GMSA and DMSA would achieve close to 100% accuracy. We first sampled 1000 examples and trained a linear separator h_k for each domain k. compared GMSA and DMSA on this dataset. For GMSA, we trained kernel density estimators and chose the bandwidth based on a five-fold cross-validation. For DMSA, we trained a conditional Maxent threshold classifier. We first illustrate the kernel density estimate using 1000 samples in Figure 1. For $x \in [-0.5, 0.5]$, $\mathcal{D}_1(x) > \mathcal{D}_2(x)$, but the kernel density estimates satisfy $\widehat{\mathcal{D}}_2(x) \geq \widehat{\mathcal{D}}_1(x)$, which shows the limitations of kernel density estimation with a single bandwidth. On the other hand, DMSA selected a threshold around 0.3 for distinguishing between $\mathcal{D}_1$ and $\mathcal{D}_2$ and achieves accuracy around 100%. We varied the number of examples available for domain adaptation and compared GMSA and DMSA. For simplicity, we found the best z using exhaustive search for both GMSA and DMSA. The results show that DMSA consistently outperforms GMSA on both the domains and hence on all convex combinations, see Figure 3. The results also indicate that DMSA converges faster, in accordance with our theory.

6. Conclusion

We presented a new algorithm for the important problem of multiple-source adaptation, which commonly arises in applications. Our algorithm was shown to benefit from favorable theoretical guarantees and a superior empirical performance, compared to previous work. Moreover, our algorithm is practical: it is straightforward to train a multi-class classifier in the setting we described and our DC-programming solution is very efficient.

Providing a robust solution for the problem is particularly important for under-represented groups, whose data is not necessarily well-represented in the classifiers to be combined and trained on source data. Our solution demonstrates improved performance even in the cases where the target distribution is not included in the source distributions. We hope that continued efforts in this area will result in more equitable treatment of under-represented groups.

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 *Proceedings of NIPS*. MIT Press, 2007.

C. Blake. UCI repository of machine learning databases. *https://archive.ics.uci.edu/ml/index.php*, 1998.

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.

J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman. Learning bounds for domain adaptation. In *Proceedings of NIPS*. MIT Press, 2008.

C. Cortes and M. Mohri. Domain adaptation in regression. In *Proceedings of ALT*, pages 308–323, 2011.

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. M. Medina. Adaptation algorithm and theory based on generalized discrepancy. In *Proceedings of KDD*, pages 169–178, 2015.

C. Cortes, M. Mohri, and A. M. Medina. Adaptation based on generalized discrepancy. *J. Mach. Learn. Res.*, 20: 1:1–1:30, 2019.

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

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.

B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of ICCV*, pages 2960–2967, 2013.

C. Gan, T. Yang, and B. Gong. Learning attributes equals multi-source domain generalization. In *Proceedings of CVPR*, pages 87–97, 2016.

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

M. Ghifary, W. Bastiaan Kleijn, M. Zhang, and D. Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of ICCV*, pages 2551–2559, 2015.

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.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

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, M. Mohri, and N. Zhang. Algorithms and theory for multiple-source adaptation. In *Proceedings of NIPS*, pages 8246–8256, 2018.

J. Hoffman, M. Mohri, and N. Zhang. Multiple-source adaptation theory and algorithms. *Ann. Math. Artif. Intell.*, 89(3-4):237–270, 2021.

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

I.-H. Jhuo, D. Liu, D. Lee, and S.-F. Chang. Robust visual domain adaptation with low-rank reconstruction. In *Proceedings of CVPR*, pages 2168–2175. IEEE, 2012.

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.

D. Kifer, S. Ben-David, and J. Gehrke. Detecting change in data streams. *Proceedings of the 30th International Conference on Very Large Data Bases*, 2004.

N. Konstantinov and C. Lampert. Robust learning from untrusted sources. In *Proceedings of ICML*, pages 3488–3498, 2019.

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

H. Liu, M. Shao, and Y. Fu. Structure-preserved multi-source domain adaptation. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 1059–1064. IEEE, 2016.

J. Liu, J. Zhou, and X. Luo. Multiple source domain adaptation: A sharper bound using weighted Rademacher complexity. In *Technologies and Applications of Artificial Intelligence (TAAI), 2015 Conference on*, pages 546–553. IEEE, 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 *Proceedings of UAI*, pages 367–374, 2009a.

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

Y. Mansour, M. Mohri, J. Ro, A. T. Suresh, and K. Wu. A theory of multiple-source adaptation with limited target labeled data. In *Proceedings of AISTATS*, pages 2332–2340, 2021.

R. McDonald, M. Mohri, N. Silberman, D. Walker, and G. S. Mann. Efficient large-scale distributed training of conditional maximum entropy models. In *Proceedings of NIPS*, pages 1231–1239, 2009.

M. Mohri, G. Sivek, and A. T. Suresh. Agnostic federated learning. In *Proceedings of ICML*, pages 4615–4625. PMLR, 2019.

S. Motiian, Q. Jones, S. Iranmanesh, and G. Doretto. Few-shot adversarial domain adaptation. In *Proceedings of NIPS*, pages 6670–6680, 2017a.

S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of ICCV*, pages 5715–5725, 2017b.

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

S. J. Nowlan and G. E. Hinton. Evaluation of adaptive mixtures of competing experts. In *Proceedings of NIPS*, pages 774–780, 1991.

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.

X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang. Moment matching for multi-source domain adaptation. In *Proceedings of ICCV*, pages 1406–1415, 2019.

K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of ICCV*, pages 8050–8058, 2019.

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.

Q. Sun, R. Chattopadhyay, S. Panchanathan, and J. Ye. A two-stage weighting framework for multi-source domain adaptation. In *Proceedings of NIPS*, pages 505–513, 2011.

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 *Proceedings of ICCV*, pages 4068–4076, 2015.

T. Van Erven and P. Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

B. Wang, J. Mendez, M. Cai, and E. Eaton. Transfer learning via minimizing the performance gap between domains. In *Proceedings of NIPS*, pages 10645–10655, 2019a.

T. Wang, X. Zhang, L. Yuan, and J. Feng. Few-shot adaptive faster r-cnn. In *Proceedings of CVPR*, pages 7173–7182, 2019b.

J. Wen, R. Greiner, and D. Schuurmans. Domain aggregation networks for multi-source domain adaptation. *arXiv preprint arXiv:1909.05352*, 2019.

R. Xu, Z. Chen, W. Zuo, J. Yan, and L. Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *Proceedings of CVPR*, pages 3964–3973, 2018.

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.

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

H. Zhao, S. Zhang, G. Wu, J. M. Moura, J. P. Costeira, and G. J. Gordon. Adversarial multiple source domain adaptation. In *Proceedings of NIPS*, pages 8559–8570, 2018.

Contents of Appendix

A. Related and Previous Work on Multiple-Source Adaptation (MSA)

The general theoretical problem of adaptation from a single domain to a target domain has been studied in a series of publications in the last two decades or so (Kifer et al., 2004; Ben-David et al., 2007; Blitzer et al., 2008; Mansour et al., 2009b; Cortes and Mohri, 2011; 2014; Cortes et al., 2015; 2019). There are many distinct instances of adaptation problems.

Multiple-source adaptation extends the single-source single-target scenario, and has been extensively studied from various aspects. (Yang et al., 2007) proposed to learn a linear combination of pre-trained auxiliary classifiers using SVMs on labeled target data. (Duan et al., 2009; 2012) further assumed plenty of unlabeled target data to form a meaningful regularizer, and a small set of labeled target data for training. (Khosla et al., 2012; Blanchard et al., 2011) combined all the source data to jointly train a single predictor. (Pei et al., 2018; Zhao et al., 2018) extended single domain adversarial learning techniques to the multiple-source setting to extract domain-invariant features. (Ghifary et al., 2015) extended auto-encoders to the multi-task setting and minimized the sum of reconstruction errors across domains. (Peng et al., 2019) proposed to align moments of feature distribution across source and target domains. (Muandet et al., 2013) proposed Domain-Invariant Component Analysis to transform features onto a low dimensional subspace that minimizes the dissimilarity across domains.

(Zhang et al., 2015) adopted a causal view of MSA where label Y is the cause for features X, estimated the weights for combining source conditional probabilities ($\mathbb{P}_{X|Y}$), and proposed various ways to construct target predictor based on estimated weights. (Crammer et al., 2008) considered learning accurate models for each source domain, using “nearby” data of other domains. (Gong et al., 2012) ranked multiple source domains by how good can they adapt to a target domain. (Gong et al., 2013a) learned domain-invariant features by constructing multiple auxiliary tasks, and learning new feature representations from each auxiliary task. (Gong et al., 2013b) proposed to discover multiple latent domains by maximizing distinctiveness and learnability between latent domains. (Jhuo et al., 2012) transfered source samples into an intermediate representation such that each transformed source sample can be linearly reconstructed by target samples. Wen et al. (2019) adjusted the weight of each source domain during training based on discrepancy minimization theory. Fernando et al. (2013) considered aligning subspaces for visual domain adaptation. Liu et al. (2016) proposed to preserve the structure information from source domains via clustering. Gan et al. (2016) tackled the multiple-source adaptation problem via attributes possessing. Sun et al. (2011) considered a two-stage adaptation where in the first stage one combines weighted source data based on marginal probability, and in the second stage based conditional probability as well.

More recently, Mansour, Mohri, Ro, Suresh, and Wu (2021) presented a theoretical and algorithmic study of the multiple-source domain adaptation problem in the common scenario where the learner has access only to a limited amount of labeled target data, but where they have at their disposal a large amount of labeled data from multiple source domains. They showed that a new family of algorithms based on model selection ideas benefits from very favorable guarantees in this scenario and discussed some theoretical obstacles affecting some alternative techniques.

B. Rényi Divergences

The Rényi Divergence 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

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

where, for $\alpha \in \{0, 1, +\infty\}$, the expression is defined by taking the limit. For $\alpha = 1$, the Rényi divergence coincides with the relative entropy. For $\alpha = +\infty$, it coincides with $\log \sup_{x \in \mathcal{X}} \frac{\mathcal{D}(x)}{\mathcal{D}(x)}$. 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:

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

The following lemma from (Van Erven and Harremos, 2014) summarizes some useful properties of the Rényi divergence.

Lemma 1. *The Rényi divergence admits the following properties:*

1. $\mathsf{D}_\alpha(\mathcal{D} \parallel \mathcal{D})$ *is a non-decreasing function of* α.
2. $\mathsf{D}_\alpha(\mathcal{D} \parallel \mathcal{D})$ *is jointly convex in* $(\mathcal{D}, \mathcal{D})$ *for* $\alpha \in [0, 1]$.
3. $\mathsf{D}_\alpha(\mathcal{D} \parallel \mathcal{D})$ *is convex in* $\mathcal{D}$ *for* $\alpha \in [0, \infty]$.
4. $\mathsf{D}_\alpha(\mathcal{D} \parallel \mathcal{D})$ *is jointly quasi-convex in* $(\mathcal{D}, \mathcal{D})$ *for* $\alpha \in [0, \infty]$.

The following general *triangle inequality* for Rényi divergences is due to Hoffman et al. (2021). Here, we give the full proof for completeness.

Proposition 2. *Let $\mathcal{P}$, $\mathcal{Q}$, $\mathcal{R}$ be three distributions on $\mathcal{X} \times \mathcal{Y}$. Then, for any $\gamma \in (0, 1)$ and any $\alpha > \gamma$, the following inequality holds:*

$$\Big[\mathsf{d}_\alpha(\mathcal{P} \parallel \mathcal{Q})\Big]^{\alpha - 1} \leq \Big[\mathsf{d}_{\frac{\alpha}{\gamma}}(\mathcal{P} \parallel \mathcal{R})\Big]^{\alpha - \gamma} \Big[\mathsf{d}_{\frac{\alpha - \gamma}{1 - \gamma}}(\mathcal{R} \parallel \mathcal{Q})\Big]^{\alpha - 1}.$$

Proof. Fix $\gamma \in (0, 1)$. By Hölder's inequality, we can write:

$$\begin{aligned}
\Big[\mathsf{d}_\alpha(\mathcal{P} \parallel \mathcal{Q})\Big]^{\alpha - 1} &= \sum_x \frac{\mathcal{P}^\alpha(x, y)}{\mathcal{Q}^{\alpha - 1}(x, y)} = \sum_x \frac{\mathcal{P}^\alpha(x, y)}{\mathcal{R}^{\alpha - \gamma}(x, y)} \frac{\mathcal{R}^{\alpha - \gamma}(x, y)}{\mathcal{Q}^{\alpha - 1}(x, y)} \\
&\leq \left[\sum_x \left(\frac{\mathcal{P}^\alpha(x, y)}{\mathcal{R}^{\alpha - \gamma}(x, y)}\right)^{\frac{1}{\gamma}}\right]^\gamma \left[\sum_x \left(\frac{\mathcal{R}^{\alpha - \gamma}(x, y)}{\mathcal{Q}^{\alpha - 1}(x, y)}\right)^{\frac{1}{1 - \gamma}}\right]^{1 - \gamma} \\
&= \left[\sum_x \frac{\mathcal{P}^{\frac{\alpha}{\gamma}}(x, y)}{\mathcal{R}^{\frac{\alpha}{\gamma} - 1}(x, y)}\right]^\gamma \left[\sum_x \frac{\mathcal{R}^{\frac{\alpha - \gamma}{1 - \gamma}}(x, y)}{\mathcal{Q}^{\frac{\alpha - \gamma}{1 - \gamma} - 1}(x, y)}\right]^{1 - \gamma} \\
&= \Big[\mathsf{d}_{\frac{\alpha}{\gamma}}(\mathcal{P} \parallel \mathcal{R})\Big]^{\alpha - \gamma} \Big[\mathsf{d}_{\frac{\alpha - \gamma}{1 - \gamma}}(\mathcal{R} \parallel \mathcal{Q})\Big]^{\alpha - 1}.
\end{aligned}$$

This concludes the proof. □

C. DMSA Guarantees

C.1. General Guarantee

Theorem 1 gives a guarantee in terms of a Rényi divergence of $\mathcal{D}_T$ and $\widehat{\mathcal{D}}$. Using the triangle inequality result of Proposition 2, we can derive an upper bound in terms of a Rényi divergence of $\mathcal{D}_T$ and $\mathcal{D}$ instead and only Rényi divergences between the distributions $\mathcal{D}_k$ and their estimate $\widehat{\mathcal{D}}_k$.

Theorem 3. *For any $\delta > 0$, there exists $z \in \Delta$ such that the following inequality holds for any $\alpha > 1$ and arbitrary target distribution $\mathcal{D}_T$:*

$$\mathcal{L}(\mathcal{D}_T, \widehat{g}_z) \leq [(\widehat{\epsilon} + \delta)\, \widehat{\mathsf{d}}']^{\frac{\alpha-1}{\alpha}} [\mathsf{d}_{2\alpha}(\mathcal{D}_T \parallel \mathcal{D})]^{\frac{2\alpha-1}{2\alpha}} M^{\frac{1}{\alpha}},$$

where $\widehat{\epsilon} = (\epsilon \widehat{\mathsf{d}})^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}$, $\widehat{\mathsf{d}} = \max_{k \in [p]} \mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)$, and $\widehat{\mathsf{d}}' = \max_{k \in [p]} \mathsf{d}_{2\alpha-1}(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k)$, with $\widehat{\mathcal{D}}_k = \frac{\widehat{\mathcal{Q}}(k|x)\mathcal{D}(x)}{\widehat{\mathcal{Q}}(k)}$.

Proof. For $\alpha > 1$, by Proposition 2, choosing $\gamma = \frac{1}{2}$, the following holds for any $\lambda \in \Delta$:

$$\begin{aligned}
[\mathsf{d}_\alpha(\mathcal{D}_T \parallel \widehat{\mathcal{D}}_\lambda)]^{\alpha-1} &\leq [\mathsf{d}_{2\alpha}(\mathcal{D}_T \parallel \mathcal{D}_\lambda)]^{\alpha-\frac{1}{2}} [\mathsf{d}_{2\alpha-1}(\mathcal{D}_\lambda \parallel \widehat{\mathcal{D}}_\lambda)]^{\alpha-1} \\
&= [\mathsf{d}_{2\alpha}(\mathcal{D}_T \parallel \mathcal{D}_\lambda)]^{\alpha-\frac{1}{2}} [e^{\mathsf{D}_{2\alpha-1}(\mathcal{D}_\lambda \parallel \widehat{\mathcal{D}}_\lambda)}]^{\alpha-1} \\
&\leq [\mathsf{d}_{2\alpha}(\mathcal{D}_T \parallel \mathcal{D}_\lambda)]^{\alpha-\frac{1}{2}} [e^{\max_{k \in [p]} (\mathsf{D}_{2\alpha-1}(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k)}]^{\alpha-1} \\
&\qquad \text{(quasi-convexity of Rényi divergence (Lemma 1))} \\
&= [\mathsf{d}_{2\alpha}(\mathcal{D}_T \parallel \mathcal{D}_\lambda)]^{\alpha-\frac{1}{2}} [\max_{k \in [p]} \mathsf{d}_{2\alpha-1}(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k)]^{\alpha-1}. \qquad \text{(monotonicity of exp)}
\end{aligned}$$

Thus, by Theorem 1, for $\widehat{\epsilon} = \max_{k \in [p]} \Big[\epsilon\, \mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)\Big]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}$, for any $\lambda \in \Delta$, we have:

$$\begin{aligned}
\mathcal{L}(\mathcal{D}_T, \widehat{g}_z) \leq \Big[(\widehat{\epsilon} + \delta)\, \mathsf{d}_\alpha(\mathcal{D}_T \parallel \widehat{\mathcal{D}}_\lambda)\Big]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}} &= (\widehat{\epsilon} + \delta)^{\frac{\alpha-1}{\alpha}} \Big[\mathsf{d}_\alpha(\mathcal{D}_T \parallel \widehat{\mathcal{D}}_\lambda)\Big]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}} \\
&\leq (\widehat{\epsilon} + \delta)^{\frac{\alpha-1}{\alpha}} [\mathsf{d}_{2\alpha}(\mathcal{D}_T \parallel \mathcal{D}_\lambda)]^{\frac{2\alpha-1}{2\alpha}} [\max_{k \in [p]} \mathsf{d}_{2\alpha-1}(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k)]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}.
\end{aligned}$$

Taking the infimum of the right-hand side over $\lambda \in \Delta$ completes the proof. □

C.2. Conditional Maxent

Here, we prove a general pointwise guarantee for conditional Maxent that will be later used in the analysis of DMSA, when used with this algorithm (Appendix C.3).

Theorem 4. *Let $\widehat{w}$ be the solution of problem (7) and w^* the population solution of the conditional Maxent optimization problem:*

$$w^* = \operatorname*{argmin}_{w \in \mathbb{R}^N} \mu \|w\|^2 - \mathop{\mathbb{E}}_{(x,k) \sim \mathcal{Q}} \big[\log \mathsf{p}_w[k|x]\big].$$

Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $(x, k) \in \mathcal{X} \times [p]$, the following inequality holds:

$$\Big|\log \mathsf{p}_{\widehat{w}}[k|x] - \log \mathsf{p}_{w^*}[k|x]\Big| \leq \frac{2\sqrt{2}r^2}{\mu\sqrt{m}} \Big[1 + \sqrt{\log(1/\delta)}\Big].$$

Proof. By Theorem 2 of (McDonald et al., 2009), for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds:

$$\|\widehat{w} - w^*\| \leq \frac{r}{\mu\sqrt{m/2}} \Big[1 + \sqrt{\log 1/\delta}\Big].$$

Next, for any $(x, k) \in \mathcal{X} \times [p]$, observe that

$$\nabla_w \left[\log \mathsf{p}_w[k|x]\right] = \nabla_w \left[w \cdot \Phi(x,k) - \log \left[\sum_{j=1}^{p} e^{w \cdot \Phi(x,j)} \right] \right] = \nabla_w \left[\Phi(x,k) - \frac{\sum_{j=1}^{p} e^{w \cdot \Phi(x,j)} \Phi(x,j)}{\sum_{j=1}^{p} e^{w \cdot \Phi(x,j)}} \right]$$
$$= \mathop{\mathbb{E}}_{j \sim \mathsf{p}_w[\cdot|x]} \left[\Phi(x,k) - \Phi(x,j) \right].$$

Thus, the following upper bound holds: $\|\nabla_w \log \mathsf{p}_w[k|x]\| \leq \| \mathbb{E}_{j \sim \mathsf{p}_w[\cdot|x]}[\Phi(x,k) - \Phi(x,j)]\| \leq 2r$ for any $(x,k) \in \mathcal{X} \times [p]$. Therefore, by the $2r$-Lipschitzness of $w \mapsto \log \mathsf{p}_w[k|x]$ for any $(x,k) \in \mathcal{X} \times [p]$, with probability at least $1 - \delta$, the following inequality holds:

$$\Big| \log \mathsf{p}_{\widehat{w}}[k|x] - \log \mathsf{p}_{w^*}[k|x] \Big| \leq 2r \|\widehat{w} - w^*\| \leq \frac{2\sqrt{2} r^2}{\mu \sqrt{m}} \left[1 + \sqrt{\log(1/\delta)} \right],$$

which completes the proof. □

C.3. Guarantees for DMSA with Conditional Maxent

Theorem 5 (DMSA)**.** *There exists $z \in \Delta$ such that for any $\delta > 0$, with probability at least $1 - \delta$ the following inequality holds DMSA used with conditional Maxent, for an arbitrary target mixture $\mathcal{D}_T$:*

$$\mathcal{L}(\mathcal{D}_T, \widehat{g}_z) \leq \epsilon \, p \, e^{\frac{6\sqrt{2} r^2}{\mu \sqrt{m}} \left[1 + \sqrt{\log(1/\delta)}\right]} \, \mathsf{d}^* \, \mathsf{d}'^*,$$
$$\text{with} \quad \mathsf{d}^* = \sup_{x \in \mathcal{X}} \mathsf{d}_\infty \left(\mathcal{Q}^*[\cdot|x] \parallel \mathcal{Q}(\cdot|x) \right)$$
$$\mathsf{d}'^* = \sup_{x \in \mathcal{X}} \mathsf{d}_\infty^2 \left(\mathcal{Q}(\cdot|x) \parallel \mathcal{Q}^*[\cdot|x] \right),$$

where $\mathcal{Q}^*(\cdot|x) = \mathsf{p}_{w^*}[\cdot|x]$ is the population solution of conditional Maxent problem (statement of Theorem 4).

We give the proof for the following more general result.

Theorem 7. *There exists $z \in \Delta$ such that the following inequality holds for any $\alpha > 1$ and arbitrary target mixture $\mathcal{D}_T$:*

$$\mathcal{L}(\mathcal{D}_T, \widehat{g}_z) \leq \epsilon^{\frac{(\alpha-1)^2}{\alpha^2}} M^{\frac{2\alpha-1}{\alpha^2}} p^{\frac{(2\alpha-1)(\alpha+2)}{2\alpha^2}} e^{\frac{(12\alpha^2 - 11\alpha + 2)}{2\alpha^2} r \|w^* - \widehat{w}\|} \mathsf{d}_1(\alpha) \mathsf{d}_2(\alpha) \mathsf{d}_3(\alpha)$$
$$\text{with} \quad \mathsf{d}_1(\alpha) = \left[\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\mathsf{d}_{4\alpha-2}^{4\alpha-3} \Big(\mathcal{Q}(\cdot|x) \parallel \mathcal{Q}^*(\cdot|x) \Big) \right] \right]^{\frac{1}{4\alpha}}$$
$$\mathsf{d}_2(\alpha) = \left[\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\mathsf{d}_{2\alpha}^{2\alpha-1} \Big(\mathcal{Q}(\cdot|x) \parallel \mathcal{Q}^*[\cdot|x] \Big) \right] \right]^{\frac{1}{2\alpha}}$$
$$\mathsf{d}_3(\alpha) = \left[\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\mathsf{d}_{2\alpha-1}^{2\alpha-2} \Big(\mathcal{Q}^*(\cdot|x) \parallel \mathcal{Q}(\cdot|x) \Big) \right] \right]^{\frac{\alpha-1}{2\alpha^2}}.$$

Proof. The proof relies on the auxiliary Lemmas 2 and 3 proven below. In Theorem 3, the bound depends on $\max_{k \in [p]} \mathsf{d}_\alpha(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k)$ and $\max_{k \in [p]} \mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)$, for some Rényi parameter $\alpha > 1$. We will analyze these terms here for the DMSA solution, for which $\widehat{\mathcal{D}}_k(x) = \frac{\widehat{\mathcal{Q}}(k|x)\mathcal{D}(x)}{\widehat{\mathcal{Q}}(k)}$. Using this expression, for any $\alpha > 1$, we can write:

$$\begin{aligned}
\max_{k \in [p]} \Big[\mathsf{d}_\alpha(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k)\Big]^{\alpha-1} = \max_{k \in [p]} \left[\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \frac{\mathcal{D}_k^\alpha(x,y)}{\widehat{\mathcal{D}}_k^{\alpha-1}(x,y)}\right] &= \max_{k \in [p]} \left[\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \frac{\Big[\mathcal{D}_k(x)\mathcal{D}_k(y|x)\Big]^\alpha}{\Big[\widehat{\mathcal{D}}_k(x)\mathcal{D}_k(y|x)\Big]^{\alpha-1}}\right] \\
&= \max_{k \in [p]} \left[\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \mathcal{D}(y|x) \frac{\Big[\mathcal{D}_k(x)\Big]^\alpha}{\Big[\widehat{\mathcal{D}}_k(x)\Big]^{\alpha-1}}\right] \\
&= \max_{k \in [p]} \left[\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \mathcal{D}(y|x) \frac{\Big[\mathcal{Q}(k|x)\mathcal{D}(x)/(1/p)\Big]^\alpha}{\Big[\widehat{\mathcal{Q}}(k|x)\mathcal{D}(x)/\widehat{\mathcal{Q}}(k)\Big]^{\alpha-1}}\right] \\
&= \max_{k \in [p]} \left[\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \mathcal{D}(y|x)\mathcal{D}(x) p^\alpha \widehat{\mathcal{Q}}^{\alpha-1}(k) \frac{\mathcal{Q}^\alpha[k|x]}{\widehat{\mathcal{Q}}^{\alpha-1}[k|x]}\right] \\
&= \max_{k \in [p]} \left[\sum_{x \in \mathcal{X}} \mathcal{D}(x) p^\alpha \widehat{\mathcal{Q}}^{\alpha-1}(k) \frac{\mathcal{Q}^\alpha[k|x]}{\widehat{\mathcal{Q}}^{\alpha-1}[k|x]}\right].
\end{aligned}$$

Next, upper-bounding the maximum by a sum yields:

$$\begin{aligned}
\max_{k \in [p]} \Big[\mathsf{d}_\alpha(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k)\Big]^{\alpha-1} \leq \sum_{k \in [p]} \left[\sum_{x \in \mathcal{X}} \mathcal{D}(x) p^\alpha \widehat{\mathcal{Q}}^{\alpha-1}(k) \frac{\mathcal{Q}^\alpha[k|x]}{\widehat{\mathcal{Q}}^{\alpha-1}[k|x]}\right] &= \left[\sum_{x \in \mathcal{X}} \mathcal{D}(x) \sum_{k \in [p]} p^\alpha \widehat{\mathcal{Q}}^{\alpha-1}(k) \frac{\mathcal{Q}^\alpha[k|x]}{\widehat{\mathcal{Q}}^{\alpha-1}[k|x]}\right] \\
&\leq p^\alpha \left[\sum_{x \in \mathcal{X}} \mathcal{D}(x) \sum_{k \in [p]} \frac{\mathcal{Q}^\alpha[k|x]}{\widehat{\mathcal{Q}}^{\alpha-1}[k|x]}\right] \\
&= p^\alpha \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \Big[\mathsf{d}_\alpha^{\alpha-1}(\mathcal{Q}(\cdot|x) \parallel \widehat{\mathcal{Q}}(\cdot|x))\Big].
\end{aligned}$$

Thus, by Lemma 2, we have

$$\begin{aligned}
&\max_{k \in [p]} \Big[\mathsf{d}_\alpha(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k)\Big]^{\alpha-1} \leq p^\alpha e^{(2\alpha-1)r\|w^* - \widehat{w}\|} \left[\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \Big[\mathsf{d}_{2\alpha}^{2\alpha-1}\Big(\mathcal{Q}(\cdot|x) \parallel \mathcal{Q}^*[\cdot|x]\Big)\Big]\right]^{\frac{1}{2}}, \\
\text{and therefore} \quad &\max_{k \in [p]} \Big[\mathsf{d}_{2\alpha-1}(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k)\Big]^{2\alpha-2} \leq p^{2\alpha-1} e^{(4\alpha-3)r\|w^* - \widehat{w}\|} \left[\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \Big[\mathsf{d}_{4\alpha-2}^{4\alpha-3}\Big(\mathcal{Q}(\cdot|x) \parallel \mathcal{Q}^*[\cdot|x]\Big)\Big]\right]^{\frac{1}{2}},
\end{aligned}$$

an expression needed later. As for the previous analysis of the Rényi divergence, we can write for any $\alpha > 1$:

$$\begin{aligned}
\max_{k \in [p]} \Big[\mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)\Big]^{\alpha-1} = \max_{k \in [p]} \left[\sum_{(x,y)} \frac{\widehat{\mathcal{D}}_k^\alpha(x,y)}{\mathcal{D}_k^{\alpha-1}(x,y)}\right] &= \max_{k \in [p]} \left[\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \mathcal{D}(y|x) \frac{\Big(\widehat{\mathcal{Q}}(k|x)\mathcal{D}(x)/\widehat{\mathcal{Q}}(k)\Big)^\alpha}{\Big(\mathcal{Q}(k|x)\mathcal{D}(x)/\mathcal{Q}(k)\Big)^{\alpha-1}}\right] \\
&= \max_{k \in [p]} \left[\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \mathcal{D}(y|x)\mathcal{D}(x) \frac{1}{p^{\alpha-1}\widehat{\mathcal{Q}}^\alpha(k)} \frac{\widehat{\mathcal{Q}}_k^\alpha(x)}{\mathcal{Q}_k^{\alpha-1}(x)}\right] \\
&= \max_{k \in [p]} \left[\sum_{x \in \mathcal{X}} \mathcal{D}(x) \frac{1}{p^{\alpha-1}\widehat{\mathcal{Q}}^\alpha(k)} \frac{\widehat{\mathcal{Q}}_k^\alpha(x)}{\mathcal{Q}_k^{\alpha-1}(x)}\right].
\end{aligned}$$

Using the upper bound on $\frac{1}{\widehat{\mathcal{Q}}(k)}$ of Lemma 3 yields:

$$\begin{aligned}
\max_{k\in[p]} \Big[\mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)\Big]^{\alpha-1} &\leq p^{\frac{2\alpha-1}{\alpha-1}} \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\mathsf{d}_\alpha^{\alpha-1}(\mathcal{Q}(\cdot|x) \parallel \widehat{\mathcal{Q}}(\cdot|x))\Big]^{\frac{\alpha}{\alpha-1}} \max_{k\in[p]}\left[\sum_{x\in\mathcal{X}} \mathcal{D}(x)\frac{\widehat{\mathcal{Q}}_k^\alpha(x)}{\mathcal{Q}_k^{\alpha-1}(x)}\right]\\
&\leq p^{\frac{2\alpha-1}{\alpha-1}} \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\mathsf{d}_\alpha^{\alpha-1}(\mathcal{Q}(\cdot|x) \parallel \widehat{\mathcal{Q}}(\cdot|x))\Big]^{\frac{\alpha}{\alpha-1}} \sum_{k\in[p]}\left[\sum_{x\in\mathcal{X}} \mathcal{D}(x)\frac{\widehat{\mathcal{Q}}_k^\alpha(x)}{\mathcal{Q}_k^{\alpha-1}(x)}\right]\\
&\leq p^{\frac{2\alpha-1}{\alpha-1}} \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\mathsf{d}_\alpha^{\alpha-1}(\mathcal{Q}(\cdot|x) \parallel \widehat{\mathcal{Q}}(\cdot|x))\Big]^{\frac{\alpha}{\alpha-1}} \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\mathsf{d}_\alpha^{\alpha-1}\Big(\widehat{\mathcal{Q}}(x) \parallel \mathcal{Q}(x)\Big)\Big].
\end{aligned}$$

Thus, by Lemma 2, we have

$$\begin{aligned}
\max_{k\in[p]} \Big[\mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)\Big]^{\alpha-1} &\leq p^{\frac{2\alpha-1}{\alpha-1}} e^{\frac{\alpha(2\alpha-1)}{\alpha-1}r\|w^*-\widehat{w}\|} \left[\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\mathsf{d}_{2\alpha}^{2\alpha-1}\Big(\mathcal{Q}(\cdot|x) \parallel \mathcal{Q}^*[\cdot|x]\Big)\Big]\right]^{\frac{\alpha}{2\alpha-2}}\\
&\quad e^{(2\alpha-1)r\|\widehat{w}-w^*\|}\left[\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\mathsf{d}_{2\alpha-1}^{2\alpha-2}\Big(\mathcal{Q}^*(\cdot|x) \parallel \mathcal{Q}(\cdot|x)\Big)\Big]\right]^{\frac{1}{2}}\\
&= p^{\frac{2\alpha-1}{\alpha-1}} e^{\frac{(2\alpha-1)^2)}{\alpha-1}r\|w^*-\widehat{w}\|} \left[\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\mathsf{d}_{2\alpha}^{2\alpha-1}\Big(\mathcal{Q}(\cdot|x) \parallel \mathcal{Q}^*[\cdot|x]\Big)\Big]\right]^{\frac{\alpha}{2\alpha-2}}\\
&\quad \left[\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\mathsf{d}_{2\alpha-1}^{2\alpha-2}\Big(\mathcal{Q}^*(\cdot|x) \parallel \mathcal{Q}(\cdot|x)\Big)\Big]\right]^{\frac{1}{2}}.
\end{aligned}$$

Plugging these inequalities into the bound of Theorem 3 yields:

$$\begin{aligned}
\mathcal{L}(\mathcal{D}_T, \widehat{g}_z) &\leq \epsilon^{\frac{(\alpha-1)^2}{\alpha^2}} M^{\frac{2\alpha-1}{\alpha^2}} \Big[\max_{k\in[p]} \mathsf{d}_{2\alpha-1}(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k)\Big]^{\frac{\alpha-1}{\alpha}} \Big[\max_{k\in[p]} \mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)\Big]^{\frac{(\alpha-1)^2}{\alpha^2}}\\
&\leq \epsilon^{\frac{(\alpha-1)^2}{\alpha^2}} M^{\frac{2\alpha-1}{\alpha^2}} \left[p^{\frac{2\alpha-1}{2\alpha}} e^{\frac{4\alpha-3}{2\alpha}r\|w^*-\widehat{w}\|}\left[\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\mathsf{d}_{4\alpha-2}^{4\alpha-3}\Big(\mathcal{Q}(\cdot|x) \parallel \mathcal{Q}^*(\cdot|x)\Big)\Big]\right]^{\frac{1}{4\alpha}}\right]\\
&\quad \left[p^{\frac{2\alpha-1}{\alpha^2}} e^{\frac{(2\alpha-1)^2}{\alpha^2}r\|w^*-\widehat{w}\|}\left[\mathbb{E}\Big[\mathsf{d}_{2\alpha}^{2\alpha-1}\Big(\mathcal{Q}(\cdot|x) \parallel \mathcal{Q}^*(\cdot|x)\Big)\Big]\right]^{\frac{1}{2\alpha}}\left[\mathbb{E}\Big[\mathsf{d}_{2\alpha-1}^{2\alpha-2}\Big(\mathcal{Q}^*(\cdot|x) \parallel \mathcal{Q}(\cdot|x)\Big)\Big]\right]^{\frac{\alpha-1}{2\alpha^2}}\right]\\
&= \epsilon^{\frac{(\alpha-1)^2}{\alpha^2}} M^{\frac{2\alpha-1}{\alpha^2}} p^{\frac{(2\alpha-1)(\alpha+2)}{2\alpha^2}} e^{\frac{(12\alpha^2-11\alpha+2)}{2\alpha^2}r\|w^*-\widehat{w}\|}\mathsf{d}_1(\alpha)\mathsf{d}_2(\alpha)\mathsf{d}_3(\alpha)
\end{aligned}$$

$$\begin{aligned}
\text{with}\quad \mathsf{d}_1(\alpha) &= \left[\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\mathsf{d}_{4\alpha-2}^{4\alpha-3}\Big(\mathcal{Q}(\cdot|x) \parallel \mathcal{Q}^*(\cdot|x)\Big)\Big]\right]^{\frac{1}{4\alpha}}\\
\mathsf{d}_2(\alpha) &= \left[\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\mathsf{d}_{2\alpha}^{2\alpha-1}\Big(\mathcal{Q}(\cdot|x) \parallel \mathcal{Q}^*[\cdot|x]\Big)\Big]\right]^{\frac{1}{2\alpha}}\\
\mathsf{d}_3(\alpha) &= \left[\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\mathsf{d}_{2\alpha-1}^{2\alpha-2}\Big(\mathcal{Q}^*(\cdot|x) \parallel \mathcal{Q}(\cdot|x)\Big)\Big]\right]^{\frac{\alpha-1}{2\alpha^2}},
\end{aligned}$$

which completes the proof. □

Lemma 2. *For any $\alpha > 1$ and $k \in [p]$, the following inequalities hold for the expected Rényi divergences between $\mathcal{Q}$ and $\widehat{\mathcal{Q}}$:*

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\mathsf{d}_\alpha^{\alpha-1}(\mathcal{Q}(\cdot|x) \parallel \widehat{\mathcal{Q}}(\cdot|x))\Big] &\leq e^{(2\alpha-1)r\|w^*-\widehat{w}\|}\left[\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\mathsf{d}_{2\alpha}^{2\alpha-1}\Big(\mathcal{Q}(\cdot|x) \parallel \mathcal{Q}^*[\cdot|x]\Big)\Big]\right]^{\frac{1}{2}}\\
\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\mathsf{d}_\alpha^{\alpha-1}\Big(\widehat{\mathcal{Q}}(\cdot|x) \parallel \mathcal{Q}(\cdot|x)\Big)\Big] &\leq e^{(2\alpha-1)r\|\widehat{w}-w^*\|}\left[\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\mathsf{d}_{2\alpha-1}^{2\alpha-2}\Big(\mathcal{Q}^*(\cdot|x) \parallel \mathcal{Q}(\cdot|x)\Big)\Big]\right]^{\frac{1}{2}},
\end{aligned}$$

where $\mathcal{Q}^*(\cdot|x) = \mathsf{p}_{w^*}[\cdot|x]$, and $\widehat{\mathcal{Q}}(\cdot|x) = \mathsf{p}_{\widehat{w}}[\cdot|x]$, the population and empirical solution of conditional Maxent problem (7), respectively.

Proof. By Proposition 2, we can write for any $\gamma \in (0,1)$, $\gamma < \alpha$:

$$\begin{aligned}
\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \Big[\mathsf{d}_\alpha^{\alpha-1}(\mathcal{Q}(\cdot|x) \parallel \widehat{\mathcal{Q}}(\cdot|x))\Big] &= \sum_{x \in \mathcal{X}} \mathcal{D}(x)\, \mathsf{d}_\alpha^{\alpha-1}(\mathcal{Q}(\cdot|x) \parallel \widehat{\mathcal{Q}}(\cdot|x)) \\
&\leq \sum_{x \in \mathcal{X}} \mathcal{D}(x) \left[\sum_{k=1}^{p} \frac{\mathcal{Q}_k^{\frac{\alpha}{\gamma}}[k|x]}{\mathcal{Q}_k^{*\frac{\alpha}{\gamma}-1}[k|x]}\right]^{\gamma} \left[\sum_{k=1}^{p} \frac{\mathcal{Q}_k^{*\frac{\alpha-\gamma}{1-\gamma}}[k|x]}{\widehat{\mathcal{Q}}_k^{\frac{\alpha-\gamma}{1-\gamma}-1}[k|x]}\right]^{1-\gamma} \\
&= \sum_{x} \left[\mathcal{D}(x) \sum_{k=1}^{p} \frac{\mathcal{Q}_k^{\frac{\alpha}{\gamma}}[k|x]}{\mathcal{Q}_k^{*\frac{\alpha}{\gamma}-1}[k|x]}\right]^{\gamma} \left[\mathcal{D}(x) \sum_{k=1}^{p} \frac{\mathcal{Q}_k^{*\frac{\alpha-\gamma}{1-\gamma}}[k|x]}{\widehat{\mathcal{Q}}_k^{\frac{\alpha-\gamma}{1-\gamma}-1}[k|x]}\right]^{1-\gamma} \\
&\leq \left[\sum_{x} \mathcal{D}(x) \sum_{k=1}^{p} \frac{\mathcal{Q}_k^{\frac{\alpha}{\gamma}}[k|x]}{\mathcal{Q}_k^{*\frac{\alpha}{\gamma}-1}[k|x]}\right]^{\gamma} \left[\sum_{x} \mathcal{D}(x) \sum_{k=1}^{p} \frac{\mathcal{Q}_k^{*\frac{\alpha-\gamma}{1-\gamma}}[k|x]}{\widehat{\mathcal{Q}}_k^{\frac{\alpha-\gamma}{1-\gamma}-1}[k|x]}\right]^{1-\gamma} \\
&\hspace{20em} \text{(Hölder's inequality)} \\
&= \left[\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\mathsf{d}_{\frac{\alpha}{\gamma}}^{\frac{\alpha}{\gamma}-1}\Big(\mathcal{Q}(\cdot|x) \parallel \mathcal{Q}^*[\cdot|x]\Big)\right]\right]^{\gamma} \left[\mathop{\mathbb{E}}_{(x,k) \sim \mathcal{D} \times \mathcal{Q}^*} \left[\frac{\mathcal{Q}^*[k|x]}{\widehat{\mathcal{Q}}(k|x)}\right]^{\frac{\alpha-\gamma}{1-\gamma}}\right]^{1-\gamma} \\
&\leq \left[e^{(\frac{\alpha-\gamma}{1-\gamma})2r\|w^* - \widehat{w}\|}\right]^{1-\gamma} \left[\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\mathsf{d}_{\frac{\alpha}{\gamma}}^{\frac{\alpha}{\gamma}-1}\Big(\mathcal{Q}(\cdot|x) \parallel \mathcal{Q}^*[\cdot|x]\Big)\right]\right]^{\gamma}. \qquad \text{(Theorem 4)}
\end{aligned}$$

Choosing $\gamma = \frac{1}{2}$ gives

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \Big[\mathsf{d}_\alpha^{\alpha-\gamma}(\mathcal{Q}(\cdot|x) \parallel \widehat{\mathcal{Q}}(\cdot|x))\Big] \leq \left[e^{(2\alpha-1)r\|w^* - \widehat{w}\|}\right] \left[\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\mathsf{d}_{2\alpha}^{2\alpha-1}\Big(\mathcal{Q}(\cdot|x) \parallel \mathcal{Q}^*[\cdot|x]\Big)\right]\right]^{\frac{1}{2}}.$$

Similarly, we can write:

$$\begin{aligned}
\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\mathsf{d}_\alpha^{\alpha-1}\Big(\widehat{\mathcal{Q}}(\cdot|x) \parallel \mathcal{Q}(\cdot|x)\Big)\right] &\leq \sum_{x \in \mathcal{X}} \left[\mathcal{D}(x) \sum_{k \in [p]} \frac{\widehat{\mathcal{Q}}_k^{\frac{\alpha}{\gamma}}(x)}{\mathcal{Q}_k^{*\frac{\alpha}{\gamma}-1}(x)}\right]^{\gamma} \left[\mathcal{D}(x) \sum_{k \in [p]} \frac{\mathcal{Q}_k^{*\frac{\alpha-\gamma}{1-\gamma}}(x)}{\mathcal{Q}_k^{\frac{\alpha-\gamma}{1-\gamma}-1}(x)}\right]^{1-\gamma} \\
&\leq \left[\sum_{x \in \mathcal{D}} \mathcal{D}(x) \sum_{k \in [p]} \frac{\widehat{\mathcal{Q}}_k^{\frac{\alpha}{\gamma}}(x)}{\mathcal{Q}_k^{*\frac{\alpha}{\gamma}-1}(x)}\right]^{\gamma} \left[\sum_{x \in \mathcal{D}} \mathcal{D}(x) \sum_{k \in [p]} \frac{\mathcal{Q}_k^{*\frac{\alpha-\gamma}{1-\gamma}}(x)}{\mathcal{Q}_k^{\frac{\alpha-\gamma}{1-\gamma}-1}(x)}\right]^{1-\gamma} \qquad \text{(Hölder's ineq.)} \\
&= \mathop{\mathbb{E}}_{(x,k) \sim \mathcal{D} \times \widehat{\mathcal{Q}}} \left[\left[\frac{\widehat{\mathcal{Q}}(k|x)}{\mathcal{Q}^*(k|x)}\right]^{\frac{\alpha}{\gamma}-1}\right]^{\gamma} \left[\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\mathsf{d}_{\frac{\alpha-\gamma}{1-\gamma}}^{\frac{\alpha-\gamma}{1-\gamma}-1}\Big(\mathcal{Q}^*(\cdot|x) \parallel \mathcal{Q}(\cdot|x)\Big)\right]\right]^{1-\gamma} \\
&\leq \left[e^{(\alpha-\gamma)2r\|\widehat{w} - w^*\|}\right] \left[\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\mathsf{d}_{\frac{\alpha-\gamma}{1-\gamma}}^{\frac{\alpha-\gamma}{1-\gamma}-1}\Big(\mathcal{Q}^*(\cdot|x) \parallel \mathcal{Q}(\cdot|x)\Big)\right]\right]^{1-\gamma}. \qquad \text{(Theorem 4)}
\end{aligned}$$

Choosing $\gamma = \frac{1}{2}$ gives

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\mathsf{d}_\alpha^{\alpha-1}\Big(\widehat{\mathcal{Q}}(\cdot|x) \parallel \mathcal{Q}(\cdot|x)\Big)\right] \leq \left[e^{(2\alpha-1)r\|\widehat{w} - w^*\|}\right] \left[\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\mathsf{d}_{2\alpha-1}^{2\alpha-2}\Big(\mathcal{Q}^*(\cdot|x) \parallel \mathcal{Q}(\cdot|x)\Big)\right]\right]^{\frac{1}{2}},$$

which completes the proof. □

Lemma 3. *For any $\alpha > 1$ and $k \in [p]$, the following inequality holds:*

$$\frac{1}{\widehat{\mathcal{Q}}(k)} \leq p^{\frac{\alpha}{\alpha-1}} \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\mathsf{d}_\alpha^{\alpha-1}\left(\mathcal{Q}(\cdot|x) \parallel \widehat{\mathcal{Q}}(\cdot|x)\right)\right]^{\frac{1}{\alpha-1}}.$$

Proof. Observe that, for any $k \in [p]$, we have:

$$\begin{aligned}
\mathcal{Q}(k) = \sum_{x \in \mathcal{X}} \widehat{\mathcal{Q}}(k|x)\mathcal{D}(x) &= \sum_{x \in \mathcal{X}} \left[\frac{\mathcal{Q}(k|x)}{\widehat{\mathcal{Q}}^{\frac{\alpha-1}{\alpha}}(k|x)} \mathcal{D}^{\frac{1}{\alpha}}(x) \right] \left[\widehat{\mathcal{Q}}^{\frac{\alpha-1}{\alpha}}(k|x) \mathcal{D}^{\frac{\alpha-1}{\alpha}}(x) \right] \\
&\leq \left[\sum_{x \in \mathcal{X}} \frac{\mathcal{Q}^{\alpha}(k|x)}{\widehat{\mathcal{Q}}^{\alpha-1}(k|x)} \mathcal{D}(x) \right]^{\frac{1}{\alpha}} \left[\sum_{x \in \mathcal{X}} \widehat{\mathcal{Q}}(k|x)\mathcal{D}(x) \right]^{\frac{\alpha-1}{\alpha}} && \text{(Hölder's ineq.)} \\
&= \left[\sum_{x \in \mathcal{X}} \frac{\mathcal{Q}^{\alpha}(k|x)}{\widehat{\mathcal{Q}}^{\alpha-1}(k|x)} \mathcal{D}(x) \right]^{\frac{1}{\alpha}} \widehat{\mathcal{Q}}^{\frac{\alpha-1}{\alpha}}(k).
\end{aligned}$$

Thus, for any $k \in [p]$, we can write:

$$\begin{aligned}
\frac{1}{\widehat{\mathcal{Q}}(k)} \leq \left[\frac{1}{\mathcal{Q}(k)} \right]^{\frac{\alpha}{\alpha-1}} \left[\sum_{x \in \mathcal{X}} \frac{\mathcal{Q}^{\alpha}(k|x)}{\widehat{\mathcal{Q}}^{\alpha-1}(k|x)} \mathcal{D}(x) \right]^{\frac{1}{\alpha-1}} &= p^{\frac{\alpha}{\alpha-1}} \left[\sum_{x \in \mathcal{X}} \frac{\mathcal{Q}^{\alpha}(k|x)}{\widehat{\mathcal{Q}}^{\alpha-1}(k|x)} \mathcal{D}(x) \right]^{\frac{1}{\alpha-1}} && (\mathcal{Q}(k) = \tfrac{1}{p}) \\
&\leq p^{\frac{\alpha}{\alpha-1}} \max_{k \in [p]} \left[\sum_{x \in \mathcal{X}} \frac{\mathcal{Q}^{\alpha}(k|x)}{\widehat{\mathcal{Q}}^{\alpha-1}(k|x)} \mathcal{D}(x) \right]^{\frac{1}{\alpha-1}} \\
&\leq p^{\frac{\alpha}{\alpha-1}} \sum_{k \in [p]} \left[\sum_{x \in \mathcal{X}} \frac{\mathcal{Q}^{\alpha}(k|x)}{\widehat{\mathcal{Q}}^{\alpha-1}(k|x)} \mathcal{D}(x) \right]^{\frac{1}{\alpha-1}} \\
&= p^{\frac{\alpha}{\alpha-1}} \left[\sum_{x \in \mathcal{X}} \left(\sum_{k \in [p]} \frac{\mathcal{Q}^{\alpha}(k|x)}{\widehat{\mathcal{Q}}^{\alpha-1}(k|x)} \right) \mathcal{D}(x) \right]^{\frac{1}{\alpha-1}} \\
&= p^{\frac{\alpha}{\alpha-1}} \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\mathsf{d}_{\alpha}^{\alpha-1} \left(\mathcal{Q}(\cdot|x) \parallel \widehat{\mathcal{Q}}(\cdot|x) \right) \right]^{\frac{1}{\alpha-1}},
\end{aligned}$$

which completes the proof. □

D. DMSA Optimization Algorithm

Here we give a DC-decomposition for the DMSA optimization problem both in the regression model with the squared loss and the probability model with the cross-entropy loss. We then describe the DC algorithm based on these decompositions.

Proposition 3 (Regression model)**.** *Let ℓ be the squared loss. Then, for any $k \in [p]$, $\mathcal{L}(\widehat{\mathcal{D}}_k, g_{z'}) - \mathcal{L}(\widehat{\mathcal{D}}_z, g_{z'}) = u_k(z) - v_k(z)$, where u_k and v_k are convex functions defined for all z by*

$$u_k(z) = \mathcal{L}(\widehat{\mathcal{D}}_k, g_{z'}) - 2M\left[\sum_{x \in \mathcal{X}} \widehat{\mathcal{D}}_k(x) \log \widehat{\mathcal{Q}}_z(x)\right],$$

$$v_k(z) = \mathcal{L}(\widehat{\mathcal{D}}_z, g_{z'}) - 2M\left[\sum_{x \in \mathcal{X}} \widehat{\mathcal{D}}_k(x) \log \widehat{\mathcal{Q}}_z(x)\right],$$

where $z'_k = \frac{z_k / \widehat{\mathcal{Q}}(k)}{\sum_{j=1}^p z_j / \widehat{\mathcal{Q}}(j)}$, $\widehat{\mathcal{D}}_z = \sum_{k=1}^p z_k \widehat{\mathcal{D}}_k$, and $\widehat{\mathcal{Q}}_z(x) = \sum_{j=1}^p z_j \widehat{\mathcal{Q}}(j|x)$.

Proof. First, notice that $g_{z'}(x) = g_{\overline{z}}$, where $\overline{z}_k = z_k / \widehat{\mathcal{Q}}(k)$, since in the expression of $g_{z'}(x)$ we can divide the numerator and the denominator by $\sum_{j=1}^p z_j / \widehat{\mathcal{Q}}(j)$.

Next, observe that $(g_{\overline{z}}(x) - y)^2 = F_{\overline{z}}(x, y) - G_{\overline{z}}(x)$, where, for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, F_z and G_z are functions defined for all $z \in \Delta$ by

$$F_z(x, y) = (g_z(x) - y)^2 - 2M \log \widehat{\mathcal{Q}}_z(x) \quad \text{and} \quad G_z(x) = -2M \log \widehat{\mathcal{Q}}_z(x).$$

We will show that $F_z(x, y)$ and $G_z(x)$ are convex functions of z. Since composition with an affine function preserves convexity, this will show that $F_{\overline{z}}(x, y)$ and $G_{\overline{z}}(x)$ are also convex functions of z. The convexity of $F_z(x, y)$ and $G_z(x)$ holds since their Hessians with respect to z are positive semi-definite:

$$H_{F_z(x,y)} = \frac{2}{\widehat{\mathcal{Q}}_z^2(x)} \left[h_{d,z}(x) h_{d,z}^\top(x) + \left(M - (y - g_z(x))^2\right) D(x) D^\top(x)\right],$$

$$H_{G_z(x)} = \frac{2M}{\widehat{\mathcal{Q}}_z^2(x)} D(x) D(x)^\top,$$

where $h_{d,z}(x)$ is the p-dimensional vector defined as $[h_{d,z}]_k(x) = \widehat{\mathcal{Q}}(k|x)\,(h_k(x) + y - 2g_z(x))$ for $k \in [p]$, and $D(x) = (\widehat{\mathcal{Q}}(1|x), \ldots, \widehat{\mathcal{Q}}(p|x))^\top$. Using the fact that $M \geq (y - g_z(x, y))^2$, $H_{F_z(x,y)}$ and $H_{G_z(x,y)}$ are positive semi-definite matrices, and thus F_z and G_z are convex functions of z for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

$u_k(z)$ is a convex function of z, since it can be expressed as an expectation of $F_{\overline{z}}$:

$$u_k(z) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \widehat{\mathcal{D}}_k(x, y) \Big[(y - g_{\overline{z}}(x))^2 - 2M \log \widehat{\mathcal{Q}}_{\overline{z}}(x)\Big] = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \widehat{\mathcal{D}}_k(x, y) F_{\overline{z}}(x, y).$$

Next, denote by $j_z(x) = \sum_{k=1}^p z_k \widehat{\mathcal{Q}}(k|x) h_k(x)$ and $k_z(x) = \widehat{\mathcal{Q}}_z(x) = \sum_{k=1}^p z_k \widehat{\mathcal{Q}}(k|x)$. By definition of $\widehat{g}_z$, we have $\widehat{g}_{\overline{z}}(x) = j_{\overline{z}}(x) / k_{\overline{z}}(x)$.

Similarly, we can write the second term of $v_k(z)$ as $\sum_{x \in \mathcal{X}} \widehat{\mathcal{D}}_k(x) G_{\overline{z}}(x)$, which is a convex function of z as an expectation of $G_{\overline{z}}$. Using the notation previously introduced, to analyze the $v_k(z)$, notice that we can write

$$\begin{aligned}
&\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \widehat{\mathcal{D}}_z(x, y) \Big[y - \frac{j_{\overline{z}}(x)}{k_{\overline{z}}(x)}\Big]^2 \\
&= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \sum_{k=1}^p z_k \frac{\widehat{\mathcal{Q}}(k|x) \mathcal{D}(x, y)}{\widehat{\mathcal{Q}}(k)} \Big[y - \frac{j_{\overline{z}}(x)}{k_{\overline{z}}(x)}\Big]^2 \\
&= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \mathcal{D}(x, y) \Big(\frac{j_{\overline{z}}(x)^2}{k_{\overline{z}}(x)} - 2y j_{\overline{z}}(x) + y^2 k_{\overline{z}}(x)\Big).
\end{aligned}$$

The Hessian matrix of $j_z(x)^2/k_z(x)$ with respect to z is

$$\nabla_z^2\Big(\frac{j_z^2(x)}{k_z(x)}\Big) = \frac{1}{k_z(x)}(h_D(x) - g_z(x)D(x))(h_D(x) - g_z(x)D(x))^\top$$

where $h_D(x) = (h_1(x)\widehat{\mathcal{Q}}(1|x), \ldots, h_p(x)\widehat{\mathcal{Q}}(p|x))^\top$ and $D(x) = (\widehat{\mathcal{Q}}(1|x), \ldots, \widehat{\mathcal{Q}}(p|x))^\top$. Thus, $j_z(x)^2/k_z(x)$ is convex and so is $j_{\overline{z}}(x)^2/k_{\overline{z}}(x)$, by composition with an affine function. $-2yj_{\overline{z}}(x) + y^2k_{\overline{z}}(x)$ is an affine function of z and is therefore convex. Thus, the first term of $v_k(z)$ is also a convex function of z, which completes the proof. □

Proposition 4 (Probability model)**.** *Let ℓ be the cross-entropy loss. Then, for $k \in [p]$, $\mathcal{L}(\widehat{\mathcal{D}}_k, g_{z'}) - \mathcal{L}(\widehat{\mathcal{D}}_z, g_{z'}) = u_k(z) - v_k(z)$, where u_k and v_k are convex functions defined for all z by*

$$\begin{aligned}
u_k(z) &= \sum_{(x,y)\in\mathcal{Y}\times\mathcal{Y}} -\widehat{\mathcal{D}}_k(x,y)\log\left[\sum_{k=1}^{p} z'_k\widehat{\mathcal{Q}}(k|y)h_k(x,y)\right]\\
v_k(z) &= \mathcal{L}(\widehat{\mathcal{D}}_z, g_{z'}) - \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \widehat{\mathcal{D}}_k(x,y)\log\mathcal{Q}_{z'}(x),
\end{aligned}$$

where $z'_k = \frac{z_k/\widehat{\mathcal{Q}}(k)}{\sum_{j=1}^p z_j/\widehat{\mathcal{Q}}(j)}$, $\widehat{\mathcal{D}}_z = \sum_{k=1}^p z_k\widehat{\mathcal{D}}_k$, and $\widehat{\mathcal{Q}}_z(x) = \sum_{j=1}^p z_j\widehat{\mathcal{Q}}(j|x)$.

Proof. Let j_z and k_z be defined for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$ and $z \in \Delta$ by $j_z(x,y) = \sum_{k=1}^p z_k\widehat{\mathcal{Q}}(k|x)h_k(x,y)$, and $k_z(x) = \widehat{\mathcal{Q}}_z(x)$. By definition, $g_{\overline{z}}(x,y) = j_{\overline{z}}(x,y)/k_{\overline{z}}(x)$. We can write

$$\begin{aligned}
&\mathcal{L}(\widehat{\mathcal{D}}_k, g_{\overline{z}}) - \mathcal{L}(\widehat{\mathcal{D}}_z, g_{\overline{z}})\\
&= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \big(\widehat{\mathcal{D}}_z(x,y) - \widehat{\mathcal{D}}_k(x,y)\big)\log\left[\frac{j_{\overline{z}}(x,y)}{k_{\overline{z}}(x)}\right]\\
&= \left[\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} -\widehat{\mathcal{D}}_k(x,y)\log j_{\overline{z}}(x,y)\right] - \left[\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \widehat{\mathcal{D}}_z(x,y)\log\left[\frac{k_{\overline{z}}(x)}{j_{\overline{z}}(x,y)}\right] - \widehat{\mathcal{D}}_k(x,y)\log k_{\overline{z}}(x)\right]\\
&= u_k(z) - v_k(z).
\end{aligned}$$

u_k is convex since $-\log j_{\overline{z}}$ is convex as the composition of the convex function $-\log$ with an affine function. Similarly, $-\log k_{\overline{z}}$ is convex, which shows that the second term in the expression of v_k is a convex function.

Observe that we can write:

$$\frac{k_{\overline{z}}(x)}{j_{\overline{z}}(x,y)} = \frac{\sum_{k=1}^p \overline{z}_k\widehat{\mathcal{Q}}(k|x)}{\sum_{k=1}^p \overline{z}_k\widehat{\mathcal{Q}}(k|x)h_k(x,y)} = \frac{\sum_{k=1}^p \overline{z}_k\widehat{\mathcal{Q}}(k|x)\mathcal{D}(x,y)}{\sum_{k=1}^p \overline{z}_k\widehat{\mathcal{Q}}(k|x)h_k(x,y)\mathcal{D}(x,y)} = \frac{\sum_{k=1}^p z_k\widehat{\mathcal{D}}_k(x,y)}{\sum_{k=1}^p z_k\widehat{\mathcal{D}}_k(x,y)h_k(x,y)} = \frac{K_z(x,y)}{J_z(x,y)}$$

where $J_{\overline{z}}(x,y) = \sum_{k=1}^p \overline{z}_k\widehat{\mathcal{D}}_k(x,y)h_k(x,y)$, and $K_z(x,y) = \widehat{\mathcal{D}}_z(x,y)$. Thus, the first term of v_k can be written in terms of the unnormalized relative entropy $\mathsf{B}(\cdot \parallel \cdot)$ as follows:

$$\begin{aligned}
\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \widehat{\mathcal{D}}_z(x,y)\log\left[\frac{k_{\overline{z}}(x)}{j_{\overline{z}}(x,y)}\right] &= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} K_z(x,y)\log\left[\frac{K_z(x,y)}{J_z(x,y)}\right]\\
&= \mathsf{B}(K_z \parallel J_z) + \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} (K_z - J_z)(x,y).
\end{aligned}$$

The rest of the proof follows from (Hoffman et al., 2018): The unnormalized relative entropy $\mathsf{B}(\cdot \parallel \cdot)$ is jointly convex, thus $\mathsf{B}(K_z \parallel J_z)$ is convex; $(K_z - J_z)$ is an affine function of z and is therefore convex too. □

Given the DC decompositions from Proposition 3 and 4, one can cast the min-max optimization problem (6) into the following variational form of a DC-programming problem (Tao and An, 1997; 1998; Sriperumbudur and Lanckriet, 2012):

$$\begin{aligned} & \min_{z\in\Delta,\gamma\in\mathbb{R}} \gamma \\ \text{s.t.} \quad & \big(u_k(z) - v_k(z) \leq \gamma\big) \wedge \big(-z_k \leq 0\big), \quad \forall k \in [p], \\ & \textstyle\sum_{k=1}^{p} z_k - 1 = 0. \end{aligned} \tag{8}$$

The DC-programming algorithm works by repeatedly solving the following convex optimization problem:

$$\begin{aligned} z_{t+1} \in \operatorname*{argmin}_{z,\gamma\in\mathbb{R}} \; & \gamma \\ \text{s.t.} \;\; & u_k(z) - v_k(z_t) - (z - z_t)\nabla v_k(z_t) \leq \gamma \\ & -z_k \leq 0, \; \textstyle\sum_{k=1}^{p} z_k - 1 = 0, \quad \forall k \in [p], \end{aligned} \tag{9}$$

where $z_0 \in \Delta$ is an arbitrary starting value, and $(z_t)_t$ denotes the sequence of solutions. Then, $(z_t)_t$ is guaranteed to converge to a local minimum of problem (6) (Sriperumbudur and Lanckriet, 2012). This leads to an efficient DC algorithm that guarantees convergence to a stationary point. Furthermore, since the minimal objective value of (6) is zero, it is straightforward to check the global optimality of a solution z. In our experiments, we have found the result of the DC algorithm to be almost always optimal.

E. Guarantees for GMSA

E.1. Convergence Results for Kernel Density Estimation

In this section, we show that the true marginal distribution $\mathcal{D}$ can be closely approximated via kernel density estimation (KDE), where the quality of approximation depends on the choice of the kernel function $K_\sigma(\cdot,\cdot)$.

Kernel density estimation (KDE) is a widely used nonparametric method for estimating densities. Let $K_\sigma(\cdot,\cdot) \geq 0$ be a normalized kernel function that satisfies $\int_{x\in\mathcal{X}} K_\sigma(x,x')dx = 1$ for all $x' \in \mathcal{X}$, where σ is the bandwidth parameter. A well-known kernel function is the Gaussian kernel: $K_\sigma(x,x') = \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^d \exp\Big\{-\frac{\|x-x'\|^2}{2\sigma^2}\Big\}$, where d is the dimension of the input space $\mathcal{X} \subseteq \mathbb{R}^d$. Let $S_n = \{x_1,\ldots,x_n\}$ be a sample of size n drawn from the true distribution $\mathcal{D}$. Then, the kernel density estimation based on the sample S_n is defined by $\widehat{\mathcal{D}}_{S_n}(\cdot) = \frac{1}{n}\sum_{i=1}^n K_\sigma(\cdot,x_i)$. With a slight abuse of notation, we adopt the shorthand $\mathcal{D}_{S_\infty}(\cdot) = \mathbb{E}_{x\sim\mathcal{D}}[K_\sigma(\cdot,x)]$, the kernel density estimation based on the entire population.

Consider two samples S_n and S'_n that only differ by one point: $S_n = S_{n-1}\cup\{x_n\}$, $S'_n = S_{n-1}\cup\{x'_n\}$, where $x_n \neq x'_n$. Assume that for all such pairs of samples S_n, S'_n, we have $\mathsf{d}_\infty(\widehat{\mathcal{D}}_{S_n} \parallel \widehat{\mathcal{D}}_{S'_n}) \leq B_n$ for some positive constant B_n. Then, the following result holds, which depends on B_n and the choice of the kernel function (Hoffman et al., 2021)[Theorem 10]. Observe that we can choose $B_n = \kappa_n$.

Theorem 8. *For any $\delta > 0$, with probability at least $1-\delta$, each of the following two inequalities holds:*

$$\mathsf{d}_\alpha\big(\widehat{\mathcal{D}}_{S_n} \parallel \mathcal{D}\big) \leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\mathsf{d}_\alpha\big(K_\sigma(\cdot,x) \parallel \mathcal{D}\big)\Big] B_n^{\frac{\alpha}{\alpha-1}\sqrt{n\log\frac{1}{\delta}/2}}, \qquad \textit{for all } \alpha \in [1,2]\,,$$

$$\mathsf{d}_\alpha\big(\mathcal{D} \parallel \widehat{\mathcal{D}}_{S_n}\big) \leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\mathsf{d}_\alpha\big(\mathcal{D} \parallel K_\sigma(\cdot,x)\big)\Big] B_n^{\sqrt{n\log\frac{1}{\delta}/2}}, \qquad \textit{for all } \alpha \geq 1.$$

Theorem 8 shows that the Rényi divergence between $\widehat{\mathcal{D}}_{S_n}$ and $\mathcal{D}$ is upper bounded by the product of two terms: the first term is the expected pointwise divergence, or, more precisely, the expected Rényi divergence between the kernel function centered at x, $K_\sigma(\cdot,x)$, and the true distribution $\mathcal{D}$, with the expectation taken over $x\sim\mathcal{D}$. Thus, the first term is purely determined by the choice of the kernel function $K_\sigma(\cdot,\cdot)$. The second term is a polynomial function of $B_n^{\sqrt{n}}$. As shown by Hoffman et al. (2021)[Theorem 12], we have $B_n = 1 + O(\frac{1}{n})$ under mild conditions, which implies $B_n^{\sqrt{n}} \to 1$ as n increases, and thus the second term converges to 1. Therefore, as the sample size n goes to infinity, we have

$$\mathsf{d}_\alpha\big(\widehat{\mathcal{D}}_{S_n} \parallel \mathcal{D}\big) \leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\mathsf{d}_\alpha\big(K_\sigma(\cdot,x) \parallel \mathcal{D}\big)\Big] \qquad \text{for all } 1 \leq \alpha \leq 2, \tag{10}$$

$$\mathsf{d}_\alpha\big(\mathcal{D} \parallel \widehat{\mathcal{D}}_{S_n}\big) \leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\mathsf{d}_\alpha\big(\mathcal{D} \parallel K_\sigma(\cdot,x)\big)\Big] \qquad \text{for all } \alpha \geq 1. \tag{11}$$

Thus, the kernel density estimation is accurate, provided that the expected pointwise Rényi divergence is small with a suitably chosen kernel function $K_\sigma(\cdot,\cdot)$.

E.2. Guarantees for GMSA with Kernel Density Estimation

The following is an analogue of Theorem 3 for GMSA.

Theorem 9. *For any $\delta > 0$, there exists $z \in \Delta$ such that the following inequality holds for any $\alpha > 1$ and arbitrary target distribution $\mathcal{D}_T$:*

$$\mathcal{L}(\mathcal{D}_T, \widehat{h}_z) \leq \big[(\widehat{\epsilon} + \delta)\,\widehat{\mathsf{d}}'\big]^{\frac{\alpha-1}{\alpha}} \big[\mathsf{d}_{2\alpha}(\mathcal{D}_T \parallel \mathcal{D})\big]^{\frac{2\alpha-1}{2\alpha}} M^{\frac{1}{\alpha}},$$

where $\widehat{\epsilon} = (\epsilon\widehat{\mathsf{d}})^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}$, $\widehat{\mathsf{d}} = \max_{k\in[p]} \mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)$, and $\widehat{\mathsf{d}}' = \max_{k\in[p]} \mathsf{d}_{2\alpha-1}(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k)$.

Proof. By Theorem 1, there exists $z \in \Delta$ such that the following inequality holds for any $\alpha > 1$ and arbitrary target mixture $\mathcal{D}_T \in \mathcal{D}$:

$$\mathcal{L}(\mathcal{D}_T, \widehat{h}_z) \leq \widehat{\epsilon}^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}} \Big[\max_{k\in[p]} \mathsf{d}_{2\alpha-1}(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k)\Big],$$

where $\widehat{\epsilon} = \max_{k\in[p]} \Big[\epsilon\, \mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)\Big]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}$. The rest of the proof is identical to that of Theorem 3. □

Using this theorem and the previous results for KDE, we can show the following.

Theorem 6 (GMSA)**.** *There exists $z \in \Delta$ such that, for any $\delta > 0$, with probability at least $1 - \delta$ the following inequality holds for* GMSA *used KDE, for an arbitrary target mixture $\mathcal{D}_T$:*

$$\mathcal{L}(\mathcal{D}_T, \widehat{h}_z) \leq \epsilon^{\frac{1}{4}} M^{\frac{3}{4}} e^{\frac{6\kappa}{\sqrt{2(m/p)}}\sqrt{\log p + \log(1/\delta)}} \mathsf{d}^* \mathsf{d}'^*,$$

with $\kappa = \max_{x, x', x'' \in \mathcal{X}} \frac{K_\sigma(x,x')}{K_\sigma(x,x'')}$, and

$$\begin{aligned} \mathsf{d}^* &= \max_{k \in [p]} \mathop{\mathbb{E}}_{x \sim \mathcal{D}_k} [\mathsf{d}_{+\infty}\big(K_\sigma(\cdot, x) \parallel \mathcal{D}_k\big)], \\ \mathsf{d}'^* &= \max_{k \in [p]} \mathop{\mathbb{E}}_{x \sim \mathcal{D}_k} [\mathsf{d}_{+\infty}\big(\mathcal{D}_k \parallel K_\sigma(\cdot, x)\big)]. \end{aligned}$$

We will prove in fact the more general result below. Setting $\alpha = 2$ in the following theorem and upper bounding the α-Rényi divergences by the $+\infty$-Rényi divergences yields immediately the result of Theorem 6. The result assumes that the number of samples used in each domain for density estimation is (m/p).

Theorem 10 (GMSA)**.** *There exists $z \in \Delta$ such that, for any $\delta > 0$, with probability at least $1 - \delta$ the following inequality holds for any $\alpha \in (1, 2]$ and arbitrary target mixture $\mathcal{D}_T$:*

$$\mathcal{L}(\mathcal{D}_T, \widehat{h}_z) \leq \epsilon^{\frac{(\alpha-1)^2}{\alpha^2}} M^{\frac{2\alpha-1}{\alpha^2}} e^{2\kappa\left(2 + \frac{1}{\alpha-1}\right)\sqrt{\frac{\log \frac{p}{\delta}}{2(m/p)}}} \mathsf{d}^*(\alpha)\, \mathsf{d}'^*(\alpha),$$

with $\kappa = \max_{x, x', x'' \in \mathcal{X}} \frac{K_\sigma(x,x')}{K_\sigma(x,x'')}$, and

$$\begin{aligned} \mathsf{d}^*(\alpha) &= \max_{k \in [p]} \mathop{\mathbb{E}}_{x \sim \mathcal{D}_k} \Big[\mathsf{d}_\alpha\big(K_\sigma(\cdot, x) \parallel \mathcal{D}_k\big)\Big], \\ \mathsf{d}'^*(\alpha) &= \max_{k \in [p]} \mathop{\mathbb{E}}_{x \sim \mathcal{D}_k} \Big[\mathsf{d}_{2\alpha-1}\big(\mathcal{D}_k \parallel K_\sigma(\cdot, x)\big)\Big]. \end{aligned}$$

Proof. By Theorem 8, for any $\delta > 0$, with probability at least $1 - \delta$, each of the following two inequalities holds for all domains:

$$\begin{aligned} \mathsf{d}_\alpha\big(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k\big) &\leq \mathop{\mathbb{E}}_{x \sim \mathcal{D}_k} \Big[\mathsf{d}_\alpha\big(K_\sigma(\cdot, x) \parallel \mathcal{D}_k\big)\Big] \kappa_m^{\frac{\alpha}{\alpha-1}\sqrt{(m/p)\log\frac{p}{\delta}/2}} && \text{for all } 1 \leq \alpha \leq 2 \\ \mathsf{d}_\alpha\big(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k\big) &\leq \mathop{\mathbb{E}}_{x \sim \mathcal{D}_k} \Big[\mathsf{d}_\alpha\big(\mathcal{D}_k \parallel K_\sigma(\cdot, x)\big)\Big] \kappa_m^{\sqrt{(m/p)\log\frac{p}{\delta}/2}}, && \text{for all } \alpha \geq 1 \end{aligned}$$

with $\kappa_m = 1 + \frac{2}{(m/p)} \Big[\max_{x_i, x_j, x \in \mathcal{X}} \frac{K_\sigma(x, x_i)}{K_\sigma(x, x_j)}\Big]$. It follows that, for all $1 < \alpha \leq 2$,

$$\begin{aligned} \max_{k \in [p]} \mathsf{d}_\alpha\big(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k\big) &\leq \max_{k \in [p]} \mathop{\mathbb{E}}_{x \sim \mathcal{D}_k} \Big[\mathsf{d}_\alpha\big(K_\sigma(\cdot, x) \parallel \mathcal{D}_k\big)\Big] \kappa_m^{\frac{\alpha}{\alpha-1}\sqrt{(m/p)\log\frac{p}{\delta}/2}}, \\ \max_{k \in [p]} \mathsf{d}_\alpha\big(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k\big) &\leq \max_{k \in [p]} \mathop{\mathbb{E}}_{x \sim \mathcal{D}_k} \Big[\mathsf{d}_\alpha\big(\mathcal{D}_k \parallel K_\sigma(\cdot, x)\big)\Big] \kappa_m^{\sqrt{(m/p)\log\frac{p}{\delta}/2}}. \end{aligned}$$

Plugging in these inequalities into the bound of Theorem 9, for $1 < \alpha \leq 2$, we obtain the following:

$$\begin{aligned} \mathcal{L}(\mathcal{D}_T, \widehat{h}_z) &\leq \epsilon^{\frac{(\alpha-1)^2}{\alpha^2}} M^{\frac{2\alpha-1}{\alpha^2}} \Big[\max_{k \in [p]} \mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)\Big]^{\frac{(\alpha-1)^2}{(\alpha)^2}} \Big[\max_{k \in [p]} \mathsf{d}_{2\alpha-1}(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k)\Big]^{\frac{\alpha-1}{\alpha}} \\ &\leq \epsilon^{\frac{(\alpha-1)^2}{\alpha^2}} M^{\frac{2\alpha-1}{\alpha^2}} \Big[\max_{k \in [p]} \mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)\Big] \Big[\max_{k \in [p]} \mathsf{d}_{2\alpha-1}(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k)\Big] \\ &\qquad\qquad \text{(since } \mathsf{d}_\alpha(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k) \geq 1 \text{ and } \mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k) \geq 1\text{)} \\ &\leq \epsilon^{\frac{(\alpha-1)^2}{\alpha^2}} M^{\frac{2\alpha-1}{\alpha^2}} \kappa_m^{(2 + \frac{1}{\alpha-1})\sqrt{(m/p)\log\frac{p}{\delta}/2}} \mathsf{d}^*(\alpha)\, \mathsf{d}'^*(\alpha), \end{aligned}$$

with

$$\mathsf{d}^*(\alpha) = \max_{k \in [p]} \mathop{\mathbb{E}}_{x \sim \mathcal{D}_k} \Big[\mathsf{d}_\alpha\big(K_\sigma(\cdot, x) \parallel \mathcal{D}_k\big) \Big], \quad \mathsf{d}'^*(\alpha) = \max_{k \in [p]} \mathop{\mathbb{E}}_{x \sim \mathcal{D}_k} \Big[\mathsf{d}_{2\alpha-1}\big(\mathcal{D}_k \parallel K_\sigma(\cdot, x)\big) \Big].$$

The bound can be further simplified as follows:

$$\begin{aligned}
\mathcal{L}(\mathcal{D}_T, \widehat{h}_z) &\leq \epsilon^{\frac{(\alpha-1)^2}{\alpha^2}} M^{\frac{2\alpha-1}{\alpha^2}} \kappa_m^{\left(2+\frac{1}{\alpha-1}\right)\sqrt{(m/p)\log\frac{p}{\delta}/2}} \mathsf{d}^*(\alpha)\, \mathsf{d}'^*(\alpha) \\
&= \epsilon^{\frac{(\alpha-1)^2}{\alpha^2}} M^{\frac{2\alpha-1}{\alpha^2}} e^{\left(2+\frac{1}{\alpha-1}\right)\sqrt{(m/p)\log\frac{p}{\delta}/2}\,\log\left(1+\frac{2\kappa}{(m/p)}\right)} \mathsf{d}^*(\alpha)\, \mathsf{d}'^*(\alpha) \\
&\leq \epsilon^{\frac{(\alpha-1)^2}{\alpha^2}} M^{\frac{2\alpha-1}{\alpha^2}} e^{2\kappa\left(2+\frac{1}{\alpha-1}\right)\sqrt{\frac{\log\frac{p}{\delta}}{2(m/p)}}} \mathsf{d}^*(\alpha)\, \mathsf{d}'^*(\alpha),
\end{aligned}$$

which completes the proof. □

F. Additional Experiments

In this section, we report experimental results for the scenario where the target domain is close to being a mixture of the source domains but where it may not necessarily be such a mixture, a scenario not covered by (Hoffman et al., 2018).

We begin with the three datasets used in Section 5: Google Street View House Numbers (SVHN), MNIST, and USPS. For these experiments, the learner is only given access to feature vectors and base predictors for two of the three domains, and is asked to predict on all three domains combined. Thus, the target domain is not a mixture of the source domains, but is not too far away from that. Table 5 presents the accuracy on all test data combined, for various baselines: the base predictors, the uniform combination of two base predictors, and DMSA trained on two domains. DMSA outperforms unif in two of the three cases, and is very close to unif in the other case.

To further evaluate the performance of DMSA, we also increased the number of source domains by introducing two additional digit datasets: MNIST-M (MNIST digits superimposed on patches randomly extracted from color photos), and a synthetic dataset (for details for these two additional datasets, see http://yaroslav.ganin.net/). Again, we left out one domain and trained on the other four, and then tested on all domains combined. The results are given in Table 6. With more source domains, DMSA significantly outperforms other baselines in all cases. This robust performance of the algorithm on domains that are poorly represented or even unrepresented during training makes the algorithm a strong candidate for tackling fairness questions.

Table 5. Train on two domains and test on all domains combined. Column name ~~dom~~ means that the learner is given features and base predictors from all domains except from domain dom.

Train data	~~svhn~~	~~mnist~~	~~usps~~
CNN-svhn	-	84.2	84.2
CNN-mnist	41.0	-	41.0
CNN-usps	32.9	32.9	-
CNN-unif	**43.8**	85.1	90.9
DMSA	43.4	**85.4**	**93.3**

Table 6. Train on four domains and test on all domains combined. Column name ~~dom~~ means that the learner is given features and base predictors from all domains except from domain dom.

Train data	~~svhn~~	~~mnist~~	~~usps~~	~~mnistm~~	~~synth~~
CNN-svhn	-	78.0	78.0	78.0	78.0
CNN-mnist	43.5	-	43.5	43.5	43.5
CNN-usps	28.4	28.4	-	28.4	28.4
CNN-mnistm	59.4	59.4	59.4	-	59.4
CNN-synth	83.8	83.8	83.8	83.8	-
CNN-unif	77.0	91.7	90.3	87.7	77.2
DMSA	**91.1**	**93.5**	**94.0**	**89.8**	**92.4**