Annals of Mathematics and Artificial Intelligence manuscript No.
(will be inserted by the editor)

Multiple-Source Adaptation Theory and Algorithms – Addendum*

Judy Hoffman · Mehryar Mohri · Ningshan Zhang

Received: TBD / Accepted: TBD

Abstract In this note, we present some key results complementing a previous manuscript (Hoffman et al., 2021) dealing with the problem of multiple-source adaptation, a key learning problem in applications. In particular, we extend the theoretical results presented for the *probability model* to the case where estimated distributions are used, first by giving a guarantee that depends on the Rényi divergence of the target distribution and the family of mixtures of estimated distributions, next by generalizing that to a result that only depends on the Rényi divergence with respect to the family of mixtures of the exact source distributions.

Keywords domain adaptation · multiple-source adaptation · Rényi divergence · transfer learning.

1 Introduction

In (Hoffman et al., 2021), we presented a general theoretical and algorithmic analysis of the problem of multiple-source adaptation, a key learning problem in applications. This note is complementing that work with some key theoretical results for the probability model.

We first briefly recapitulate the learning scenario we consider. Let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ the output space. We consider a multiple-source domain adaptation (MSA) problem in the general stochastic scenario where there is a distribution over the joint input-output space $\mathcal{X} \times \mathcal{Y}$. We assume that $\mathcal{X}$ and $\mathcal{Y}$ are discrete, but the predictors we consider can take real values. Our theory can be straightforwardly extended to the continuous case with summations replaced by integrals in the proofs. We identify a domain with a distribution

* In the original version of this paper, the authors were mistakenly listed in non-alphabetic order.

Judy Hoffman
School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA 30332
E-mail: judy@gatech.edu

Mehryar Mohri
Google Research, New York, NY 10012
E-mail: mohri@cims.nyu.edu

Ningshan Zhang
New York University, New York, NY 10012
E-mail: nzhang@stern.nyu.edu

$\mathcal{D}_k$ over $\mathcal{X} \times \mathcal{Y}$, and assume that the learner admits access to the true, or, more likely, to an estimated distribution for each domain.

We further assume that the learner has access to a predictor h_k for each domain $\mathcal{D}_k$, $k \in [p] = \{1, \ldots, p\}$. We consider two types of predictor functions h_k, and their associated loss functions ℓ under the regression model (R) and the probability model (P) respectively:

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

We abuse the notation and write $\ell(h, x, y)$ to denote the loss of a predictor h at point (x, y), that is $\ell(h(x), y)$ in the regression model, and $\ell(h(x, y))$ in the probability model. In this note, we are particularly interested in the probability model. We denote by $\mathcal{L}(\mathcal{D}, h)$ the expected loss of a predictor h with respect to the distribution $\mathcal{D}$:

$$\mathcal{L}(\mathcal{D}, h) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \big[\ell(h, x, y)\big]$$

We assume that ℓ is convex, continuous, and bounded. We will assume that each h_k is a relatively accurate predictor for the distribution $\mathcal{D}_k$: there exists $\epsilon > 0$ such that $\mathcal{L}(\mathcal{D}_k, h_k) \leq \epsilon$ for all $k \in [p]$. We will also assume that the loss of the source predictor h_k is bounded, that is $\ell(h_k, x, y) \leq M$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and all $k \in [p]$.

The learner's objective in the MSA problem is to combine these predictors to design a predictor with small expected loss on a target domain $\mathcal{D}_T$. The target distribution may be an arbitrary and unknown mixture of the source domains: $\mathcal{D}_T \in \{\mathcal{D}_\lambda\colon \mathcal{D}_\lambda = \sum_{k=1}^p \lambda_k \mathcal{D}_k, \lambda \in \Delta\}$, where $\mathcal{D}_\lambda$ is a mixture of source domains with mixture parameter $\lambda \in \Delta$, and where Δ is the simplex of dimension p, $\Delta = \{(\lambda_1, \cdots, \lambda_p)\colon \sum_{k=1}^p \lambda_k = 1, \lambda_k \geq 0, \forall k \in [p]\}$. We are also interested in the case where $\mathcal{D}_T$ is some arbitrary distribution that is not necessarily a mixture of source domains. Let us emphasize that the learner has no knowledge of the target domain, including whether the target domain is a mixture of source domains.

2 Probability model – Complementary results

In previous work, the following general theorem was shown in the probability model, for which no assumption was made about the conditional probabilities of the source distributions.

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

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

In practice, however, the source distributions $\mathcal{D}_k$ are not known. Instead, we need to resort to estimates $\widehat{\mathcal{D}}_k$ of the source distributions that can be derived using various density estimation methods such as kernel density estimation. The following provides a guarantee for this more realistic scenario that depends on the maximum Rényi divergence between an estimate and a true source distributions.

Theorem 2 (Distinct conditionals; arbitrary target; distribution estimates) *For any $\delta > 0$, there exist $\eta > 0$ and $z \in \Delta$ such that the following inequality holds for any $\alpha > 1$ and arbitrary target distribution $\mathcal{D}_T$:*

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

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* $\widehat{\mathcal{D}} = \Big\{ \sum_{k=1}^p \lambda_k \widehat{\mathcal{D}}_k \colon \lambda \in \Delta \Big\}$.

Proof For any predictor $k \in [p]$ and any $\beta > 1$, by Hölder's inequality, the following holds:

$$\begin{aligned}
\mathcal{L}(\widehat{\mathcal{D}}_k, h_k) &= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \widehat{\mathcal{D}}_k(x,y) \ell(h(x,y)) \\
&= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \left[\frac{\widehat{\mathcal{D}}_k(x,y)}{\mathcal{D}_k(x,y)^{\frac{\alpha-1}{\alpha}}} \right] \left[\mathcal{D}_k(x,y)^{\frac{\alpha-1}{\alpha}} \ell(h(x,y)) \right] \\
&\leq \left[\sum_{(x,y)} \frac{\widehat{\mathcal{D}}_k(x,y)^\alpha}{\mathcal{D}_k(x,y)^{\alpha-1}} \right]^{\frac{1}{\alpha}} \left[\sum_{(x,y)} \mathcal{D}_k(x,y) \ell(h(x,y))^{\frac{\alpha}{\alpha-1}} \right]^{\frac{\alpha-1}{\alpha}} \\
&= \mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)^{\frac{\alpha-1}{\alpha}} \left[\sum_{(x,y)} \mathcal{D}_k(x,y) \ell(h(x,y))^{\frac{\alpha}{\alpha-1}} \right]^{\frac{\alpha-1}{\alpha}} \\
&\leq \mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k)^{\frac{\alpha-1}{\alpha}} \left[\sum_{(x,y)} \mathcal{D}_k(x,y) \ell(h(x,y)) M^{\frac{1}{\alpha-1}} \right]^{\frac{\alpha-1}{\alpha}} \\
&= \Big[\mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k) \, \mathcal{L}(\mathcal{D}_k, h_k) \Big]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}} \\
&\leq \Big[\mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \parallel \mathcal{D}_k) \, \epsilon \Big]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}.
\end{aligned}$$

Then, using $\widehat{\epsilon}$ in lieu of ϵ in Theorem 1, for any $\delta > 0$, there exist $\eta > 0$ and $z \in \Delta$ such that the following inequality holds:

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

This completes the proof. □

This results depends on the divergence between the target distribution $\mathcal{D}_T$ and the family of mixtures of estimates $\widehat{\mathcal{D}}_k$s of the source distributions. Instead, we will present a guarantee that depends only that the divergence between $\mathcal{D}_T$ and the family of mixtures of the true source distributions $\mathcal{D}_k$s.

Theorem 3 (Distinct conditionals; arbitrary target; distribution estimates) *For any* $\delta > 0$*, there exist* $\eta > 0$ *and* $z \in \Delta$*, such that the following inequality holds for any* $\alpha, \beta > 1$, $\gamma \in (0,1)$ *and arbitrary target distribution* $\mathcal{D}_T$*:*

$$\mathcal{L}(\mathcal{D}_T, \widehat{h}_z^\eta) \leq \left[(\widehat{\epsilon} + \delta) \, \mathsf{d}_{\frac{\beta}{\gamma}}^{\frac{\beta-\gamma}{\beta-1}}(\mathcal{D}_T \parallel \mathcal{D}) \max_{k \in [p]} \mathsf{d}_{\frac{\beta-\gamma}{1-\gamma}}\big(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k\big) \right]^{\frac{\beta-1}{\beta}} M^{\frac{1}{\beta}}.$$

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* $\mathcal{D} = \Big\{ \sum_{k=1}^p \lambda_k \mathcal{D}_k \colon \lambda \in \Delta \Big\}$.

Proof By (Hoffman et al., 2021)[Lemma 8], for any $\widehat{\mathcal{D}}_\lambda = \sum_{k=1}^{p} \lambda_k \widehat{\mathcal{D}}_k$, $\lambda \in \Delta$,

$$\begin{aligned}
\mathsf{d}_{\beta}^{\beta-1}(\mathcal{D}_T \parallel \widehat{\mathcal{D}}_\lambda) &\leq \mathsf{d}_{\frac{\beta}{\gamma}}^{\beta-\gamma}(\mathcal{D}_T \parallel \mathcal{D}_\lambda)\, \mathsf{d}_{\frac{\beta-\gamma}{1-\gamma}}^{\beta-1}\big(\mathcal{D}_\lambda \parallel \widehat{\mathcal{D}}_\lambda\big)\\
&= \mathsf{d}_{\frac{\beta}{\gamma}}^{\beta-\gamma}(\mathcal{D}_T \parallel \mathcal{D}_\lambda) \exp\Big\{(\beta-1)\mathsf{D}_{\frac{\beta-\gamma}{1-\gamma}}\big(\mathcal{D}_\lambda \parallel \widehat{\mathcal{D}}_\lambda\big)\Big\}\\
&\leq \mathsf{d}_{\frac{\beta}{\gamma}}^{\beta-\gamma}(\mathcal{D}_T \parallel \mathcal{D}_\lambda) \exp\Big\{(\beta-1)\max_{k\in[p]} \mathsf{D}_{\frac{\beta-\gamma}{1-\gamma}}\big(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k\big)\Big\}\\
&\qquad\qquad \text{(joint quasi-convexity wrt first argument of } \mathsf{D}_{\frac{\beta-\gamma}{1-\gamma}}\text{)}\\
&= \mathsf{d}_{\frac{\beta}{\gamma}}^{\beta-\gamma}(\mathcal{D}_T \parallel \mathcal{D}_\lambda) \max_{k\in[p]} \mathsf{d}_{\frac{\beta-\gamma}{1-\gamma}}^{\beta-1}\big(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k\big).
\end{aligned}$$

Plugging in this inequality in the inequality of Theorem 2 yields:

$$\mathcal{L}(\mathcal{D}_T, \widehat{h}_z^{\eta}) \leq \left[(\widehat{\epsilon} + \delta)\, \mathsf{d}_{\frac{\beta}{\gamma}}^{\frac{\beta-\gamma}{\beta-1}}(\mathcal{D}_T \parallel \mathcal{D}_\lambda) \max_{k\in[p]} \mathsf{d}_{\frac{\beta-\gamma}{1-\gamma}}\big(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k\big)\right]^{\frac{\beta-1}{\beta}} M^{\frac{1}{\beta}}.$$

Taking the infimum over $\lambda \in \Delta$ of the right-hand side gives

$$\mathcal{L}(\mathcal{D}_T, \widehat{h}_z^{\eta}) \leq \left[(\widehat{\epsilon} + \delta)\, \mathsf{d}_{\frac{\beta}{\gamma}}^{\frac{\beta-\gamma}{\beta-1}}(\mathcal{D}_T \parallel \mathcal{D}) \max_{k\in[p]} \mathsf{d}_{\frac{\beta-\gamma}{1-\gamma}}\big(\mathcal{D}_k \parallel \widehat{\mathcal{D}}_k\big)\right]^{\frac{\beta-1}{\beta}} M^{\frac{1}{\beta}}.$$

This concludes the proof. □

3 Conclusion

The theoretical results presented in this note complement those given in (Hoffman et al., 2021) and overall provide an exhaustive analysis of the problem of multiple-source adaptation and algorithmic solutions. The proof concepts and tools introduced are likely to be useful in the analysis of other related problems.

Declarations

Conflict of Interest: The authors declare that they have no conflict of interest.

References

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