Algorithms and Theory for Multiple-Source Adaptation

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

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

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

Abstract

This work considers the multiple-source adaptation problem where the learner only has access to predictors and density estimations trained on source domains, but he has no access to all source data simultaneously. The goal is to combine source predictors to derive an accurate predictor for *any* unknown mixture target domain. We present the distribution-weighted combination solutions with strong theoretical guarantees for the general stochastic scenario under cross-entropy loss and other similar losses. Moreover, we give new algorithms for determining the solution for the cross-entropy loss and other losses. We report the results on a real-world dataset to show that our algorithm outperforms competing approaches by producing a single robust model that performs well on any target mixture distribution.

1 Introduction

In many modern applications, often the learner has access to information about several source domains, including accurate predictors possibly trained and made available by others, but no direct information about a target domain for which one wishes to achieve a good performance. The target domain can typically be viewed as a combination of the source domains, that is a mixture of their joint distributions, or it may be close to such mixtures. In addition, often the learner does not have access to all source data simultaneously, for legitimate reasons such as privacy, storage limitation, etc. Thus the learner cannot simply pool all source data together to learn a predictor. Here, we focus on the problem of multiple-source domain adaptation and ask how the learner can combine relatively accurate predictors available for each source domain to derive an accurate predictor for *any* new mixture target domain?

This is known as the *multiple-source adaption (MSA) problem* first formalized and analyzed theoretically by (10; 11) and later studied under various assumptions (5; 6; 8; 2; 7; 18). (10; 11) gave strong theoretical guarantees for a distribution-weighted combination for the MSA problem, but they did not provide any algorithmic solution. Furthermore, the solution they proposed could not be used for loss functions such as cross-entropy, which require a normalized predictor. Their work also assumed a deterministic scenario (non-stochastic) with the same labeling function for all source domains.

This work makes a number of novel contributions to the MSA problem. We give new normalized solutions with strong theoretical guarantees for the cross-entropy loss and other similar losses. Our guarantees hold even when the conditional probabilities for the source domains are distinct. A by-product of our analysis is the extension of the theoretical results of (10; 11) to the stochastic scenario, where there is a joint distribution over the input and output space.

Moreover, we give new algorithms for determining the distribution-weighted combination solution for the cross-entropy loss and other losses. We prove that the problem of determining that solution can be cast as a DC-programming (difference of convex) and prove explicit DC-decompositions. We also give experimental results on a benchmark dataset demonstrating that our distribution-weighted

Preprint. Work in progress.

combination solution is remarkably robust. Our algorithm outperforms competing approaches and performs well on any target mixture distribution.

2 Problem setup

Let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ the output space. We consider a multiple-source domain adaptation (MSA) problem in the general stochastic scenario where there is a distribution over the joint input-output space, $\mathcal{X} \times \mathcal{Y}$. This extends the deterministic scenario in (10; 11), where a target function mapping from $\mathcal{X}$ to $\mathcal{Y}$ is assumed. This extension is needed for the analysis of the most common and realistic learning setups in practice. We will identify a *domain* with a distribution over $\mathcal{X} \times \mathcal{Y}$ and consider the scenario where the learner has access to a predictor h_k, for each domain $\mathcal{D}_k$, $k = 1, \ldots, p$. We will assume that $\mathcal{X}$ and $\mathcal{Y}$ are discrete, but our theory can be straightforwardly extended to the continuous case with summations replaced by integrals in the proofs.

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,

$$h_k \colon \mathcal{X} \to \mathbb{R},\ L \colon \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+ \quad (R); \qquad h_k \colon \mathcal{X} \times \mathcal{Y} \to [0,1],\ L \colon [0,1] \to \mathbb{R}_+ \quad (P).$$

We abuse the notation and write $L(h, x, y)$ to denote the loss of a predictor h at point (x, y), that is $L(h(x), y)$ in the regression model, and $L(h(x, y))$ in the probability model. We will denote by $\mathcal{L}(\mathcal{D}, h)$ the expected loss of a predictor h on domain $\mathcal{D}$: $\mathcal{L}(\mathcal{D}, h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[L(h, x, y)\big]$. Much of our theory only assumes that L is convex and continuous. But, we will be particularly interested in the case where in the regression model, $L(h(x), y) = (h(x) - y)^2$ is the squared loss, and where in the probability model, $L(h(x, y)) = -\log h(x, y)$ is the cross-entropy loss (log-loss).

We will assume that each h_k is a relatively accurate predictor for the distribution $\mathcal{D}_k$: there exists $\epsilon > 0$ such that $\mathcal{L}(\mathcal{D}_k, h_k) \leq \epsilon$ for all $k \in [p]$. We will also assume that the loss of the source hypotheses h_k is bounded, that is $L(h_k, x, y) \leq M$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and all $k \in [p]$. When L is cross-entropy loss, we will further assume that source predictors are normalized for every x: $\sum_{y\in\mathcal{Y}} h_k(x, y) = 1,\ \forall x \in \mathcal{X},\ \forall k \in [p]$.

In the MSA problem, the learner's objective is to combine h_ks to design a predictor with small expected loss on a target domain that could be an arbitrary and unknown mixture of the source domains, or even some other arbitrary distribution. It is worth emphasizing that the leaner has no access to all source data simultaneously, and the leaner has no knowledge of the target domain.

Our solution extends the result of (10). We define the distribution-weighted combination of the functions h_k as follows. For any $z \in \Delta$, $\eta > 0$, and $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

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

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

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

3 Theoretical analysis

In this section, we present theoretical analyses of the general MSA setting under stochastic scenario.

Our theoretical results rely on the measure of divergence between distributions. The one that naturally comes up in our analysis is the *Rényi Divergence* (12). We will denote by $\mathsf{d}_\alpha(\mathcal{D} \parallel \mathcal{D}') = e^{\mathsf{D}_\alpha(\mathcal{D}\|\mathcal{D}')}$ the exponential of the α-Rényi Divergence of two distributions $\mathcal{D}$ and $\mathcal{D}'$. More details of the Rényi Divergence are given in Appendix D.

We first assume the target distribution $\mathcal{D}_T$ is an unknown mixture of source distributions, such that $\mathcal{D}_T^1 \in \mathcal{D}^1 = \{\sum_{k=1}^p \lambda_k \mathcal{D}_k^1 : \lambda \in \Delta\}$ in the regression model (R), or $\mathcal{D}_T \in \mathcal{D} = \{\sum_{k=1}^p \lambda_k \mathcal{D}_k \colon \lambda \in \Delta\}$ in the probability model (P). We will denote by $\mathcal{D}_T(\cdot|x)$ and $\mathcal{D}_k(\cdot|x)$ the conditional probability distribution on the target and the source domain respectively. Given the same input x, $\mathcal{D}_T(\cdot|x), \mathcal{D}_k(\cdot|x), k \in [p]$ are not necessarily the same. This is a novel extension that was not

discussed in (11), where in the deterministic scenario, exactly the same labeling function f is assumed for all source domains.

Our theoretical results depend on the following quantity: for some choice of $\alpha > 1$, define ϵ_T by

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

When the average divergence is small, α can be chosen to be very large and ϵ_T is close to ϵ. We have theoretical guarantees for the distribution-weighted combination rules as follows.

Theorem 1. *For any $\delta > 0$, there exists $\eta > 0$ and $z \in \Delta$ such that the following inequalities hold for any $\alpha > 1$:*

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

The proof is given in Appendix A. The learning guarantees for the regression and the probability model are slightly different, since the definitions of the distribution-weighted combinations are different for the two models. Theorem 1 shows the existence of $\eta > 0$ and a mixture weight $z \in \Delta$ with a remarkable property: in the regression model (R), for any target distribution $\mathcal{D}_T$ whose conditional probability $\mathcal{D}_T(\cdot|x)$ is on average not too far away from $\mathcal{D}_k(\cdot|x)$ for any $k \in [p]$, and $\mathcal{D}_T^1 \in \mathcal{D}^1$, the loss of h_z^η on $\mathcal{D}_T$ is small. It is even more remarkable that, in the probability model (P), the loss of h_z^η is at most ϵ on any target distribution $\mathcal{D}_T \in \mathcal{D}$. Thus, h_z^η is a robust hypothesis with favorable property for any such target distribution $\mathcal{D}_T$.

To cover the realistic cases in applications, we can further extend this result to the case where the distributions $\mathcal{D}_k$ are not directly available to the learner, and instead estimates $\widehat{\mathcal{D}}_k$ have been derived from data, and further to the case where the target distribution $\mathcal{D}_T$ is not a mixture of source distributions. The details are given in Appendix A.

Finally, when L coincides with the cross-entropy loss in the probability model, we propose a normalized distribution-weighted combination solution:

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

Its analysis is a complement to Theorem 1, which only works for the unnormalized hypothesis $h_z^\eta(x, y)$, and is provided in Appendix B.

4 Algorithms

We have shown that, for both the regression and the probability model, there exists a vector z defining a distribution-weighted combination hypothesis h_z^η that admits very favorable guarantees. But how we find a such z? This is a key question in the MSA problem which was not addressed by (10; 11): no algorithm was previously reported to determine the mixture parameter z (even for the deterministic scenario). Here, we give an algorithm for determining that parameter z.

Theorem 1 shows that the hypothesis h_z^η based on the mixture parameter $z \in \Delta$ benefits from a strong generalization guarantee. A key step in proving Theorem 1 is to show the existence of z such that for any $\eta, \eta' > 0$, $\forall k \in [p]$, $\mathcal{L}(\mathcal{D}_k, h_z^\eta) \le \mathcal{L}(\mathcal{D}_z, h_z^\eta) + \eta'$, where $\mathcal{D}_z = \sum_{k=1}^p z_k \mathcal{D}_k$. Thus, our problem consists of finding a parameter z verifying this property. This, can be equivalently formulated as the following optimization problem:

$$\min_{z\in\Delta, \gamma\in\mathbb{R}} \gamma \qquad \text{s.t. } \mathcal{L}(\mathcal{D}_k, h_z^\eta) - \mathcal{L}(\mathcal{D}_z, h_z^\eta) \le \gamma, \ \forall k \in [p]. \tag{1}$$

We give a DC-decomposition (difference of convex decomposition) of the objective for both models in Appendix C, such that $\mathcal{L}(\mathcal{D}_k, h_z^\eta) - \mathcal{L}(\mathcal{D}_z, h_z^\eta) = u_k(z) - v_k(z)$ for some convex functions u_k, v_k. Thus Problem (1) becomes the following variational form of a DC-programming problem (15; 16; 14):

$$\min_{z\in\Delta, \gamma\in\mathbb{R}} \gamma \qquad \text{s.t.} \big(u_k(z) - v_k(z) \le \gamma\big) \wedge \big(-z_k \le 0\big) \wedge \Big(\sum_{k=1}^p z_k - 1 = 0\Big), \quad \forall k \in [p]. \tag{2}$$

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

$$\begin{aligned} &z_{t+1} \in \operatorname{argmin}_{z,\gamma\in\mathbb{R}} \ \gamma \\ &\text{s.t. } \big(u_k(z) - v_k(z_t) - (z - z_t)\nabla v_k(z_t) \le \gamma\big) \wedge \big(-z_k \le 0\big) \wedge \Big(\sum_{k=1}^p z_k - 1 = 0\Big), \quad \forall k \in [p], \end{aligned} \tag{3}$$

Table 1: ***Office*** dataset accuracy: We report accuracy across six possible test domains. We show performance all baselines: CNN-`a,w,d`, CNN-`unif`, `DW` based on the learned z, and the jointly trained model CNN-`joint`. `DW` outperforms all competing models.

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

where $z_0 \in \Delta$ is an arbitrary starting value. Then, $(z_t)_t$ is guaranteed to converge to a local minimum of Problem (1) (17; 14). Problem (3) is a relatively simple optimization problem: $u_k(z)$ is a weighted sum of the negative logarithm of an affine function of z, plus a weighted sum of rational functions of z (squared loss), and all other terms appearing in the constraints are affine functions of z.

5 Experiments

We evaluate our DC-programming solution applied to real-world visual domain adaptation benchmark dataset *Office* (13), which has 3 domains: `amazon`, `webcam`, and `dslr`. We follow the standard protocol from (13), whereby 20, 8, and 8 labeled examples are available for training from the `amazon`, `webcam` and `dslr` domain respectively. The remaining examples from each domain are used for testing. We use the AlexNet (9) ConvNet (CNN) architecture, pre-trained on ImageNet.

The probability distributions $\mathcal{D}_k$ are not readily available. However, Corollary 6 enables us to use estimates $\widehat{\mathcal{D}}_k$ instead. We estimate $\widehat{\mathcal{D}}_k$ by kernel density estimation with fc7 activations (4) from AlexNet as features. For each individual domain, we use the output from the softmax score layer as our base predictors h_k. We then learn the weights z on a small subset of combined training samples, and obtain the distribution weighted predictor `DW` with h_ks, density estimates, and z.

We consider the uniformly weighted combination of source predictors, $h_{\texttt{unif}} = \sum_{k=1}^{p} h_k/p$. We also train a privileged baseline on all source data combined, $h_{\texttt{joint}}$, which is often not feasible if independent entities contribute classifiers and densities, but not full training datasets for privacy reasons. Thus this approach operates in a much more favorable learning setting than our solution.

We report the performance of our method (`DW`) and that of baselines (h_k, $h_{\texttt{unif}}$, $h_{\texttt{joint}}$) in Table 1. We evaluate on various test distributions: each individual domain, the combination of each two domains and the fully combined set. When the test distribution equals one of the source distributions, our distribution-weighted classifier successfully outperforms (`webcam,dslr`) or maintains performance of the classifier which is trained and tested on the same domain. For the more realistic scenario where the target domain is a mixture of any two or all three source domains, the performance of our method is comparable or marginally superior to that of the jointly trained network, despite the fact that we do not retrain any network parameters in our method and that we only use a small number of per-domain examples to learn the distribution weights – an optimization which may be solved on a single CPU in a matter of seconds for this problem. This again demonstrates the robustness of our distribution-weighted combined classifier to a varying target domain.

6 Conclusion

We presented practically applicable multiple-source domain adaptation algorithms for the cross-entropy loss and other similar losses. These algorithms benefit from very favorable theoretical guarantees that we extended to the stochastic setting. Our empirical results further demonstrate empirically their effectiveness and their importance in adaptation problems.

References

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

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

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

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

[5] 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.

[6] 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.

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

[8] 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.

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

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

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

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

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

[14] 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.

[15] 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.

[16] 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.

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

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

A Theoretical analysis for the stochastic scenario

In this section, we give a series of theoretical results for the general stochastic scenario with their full proofs. We will separate the proofs for the regression model (Appendix A.1) and the probability model (Appendix A.2), since the definitions of the distribution weighted combination are different in the two models.

A.1 Regression model

The proofs for the regression model (R) are presented in the following order: we first assume the conditional probabilities are the same across source domains, and prove Lemma 3; using that, we prove Corollary 4 and Corollary 6. Finally, we relax the assumption of same conditionals, and prove Theorem 7, which a stronger version of Theorem 1.

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

Theorem 2. *For any compact and convex non-empty set $C \subset \mathbb{R}^p$ and any continuous function $f: C \to C$, there is a point $x \in C$ such that $f(x) = x$.*

Lemma 3. *For any $\eta, \eta' > 0$, there exists $z \in \Delta$, with $z_k \neq 0$ for all $k \in [p]$, such that the following holds for the distribution-weighted combining rule h_z^η:*

$$\forall k \in [p], \quad \mathcal{L}(\mathcal{D}_k, h_z^\eta) \leq \sum_{j=1}^{p} z_j \mathcal{L}(\mathcal{D}_j, h_z^\eta) + \eta'. \tag{4}$$

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

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

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

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

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

Corollary 4. *Assume the conditional probability $\mathcal{D}_k(y|x)$ does not depend on k. Let $\mathcal{D}_\lambda$ be an arbitrary mixture of source domains, $\lambda \in \Delta$. For any $\delta > 0$, there exists $\eta > 0$ and $z \in \Delta$, such that $\mathcal{L}(\mathcal{D}_\lambda, h_z^\eta) \leq \epsilon + \delta$.*

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

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

By convexity of L, this implies that

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

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

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

Now, choose $z \in \Delta$ as in the statement of Lemma 3. Then, the following holds for any mixture distribution $\mathcal{D}_\lambda$:

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

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

Next, we introduce a useful Corollary and give its proof.

Corollary 5. *Let $\mathcal{D}_T$ be an arbitrary target distribution. For any $\delta > 0$, there exists $\eta > 0$ and $z \in \Delta$, such that the following inequality holds for any $\alpha > 1$:*

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

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

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

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

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

Now, by Corollary 4, there exists $z \in \Delta$ and $\eta > 0$ such that $\mathcal{L}(\mathcal{D}, h_z^\eta) \leq \epsilon + \delta$ for any mixture distribution $\mathcal{D} \in \mathcal{D}$. Thus, in view of the previous inequality, we can write,for any $\mathcal{D} \in \mathcal{D}$,

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

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

Corollary 6. *Let $\mathcal{D}_T$ be an arbitrary target distribution. Then, for any $\delta > 0$, there exists $\eta > 0$ and $z \in \Delta$, such that the following inequality holds for any $\alpha > 1$:*

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

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

Proof. By the first part of the proof of Corollary 5, for any $k \in [p]$ and $\alpha > 1$, the following inequality holds:

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

We can now apply the result of Corollary 5 (with $\widehat{\epsilon}$ instead of ϵ and $\widehat{\mathcal{D}}_k$ instead of $\mathcal{D}_k$). In view that, there exists $\eta > 0$ and $z \in \Delta$ such that

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

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

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

Corollary 6 used Rényi divergence in both directions: $\mathsf{d}_\alpha(\mathcal{D}_T \,\|\, \widehat{\mathcal{D}})$ requires $\mathrm{Supp}(\mathcal{D}_T) \subseteq \mathrm{Supp}(\widehat{\mathcal{D}})$, and $\mathsf{d}_\alpha(\widehat{\mathcal{D}}_k \,\|\, \mathcal{D}_k)$ requires $\mathrm{Supp}(\widehat{\mathcal{D}}_k) \subseteq \mathrm{Supp}(\mathcal{D}_k), k \in [p]$. In our experiments in Section 5, we used bigram language model for sentiment analysis, and kernel density estimation with a Gaussian kernel for object recognition. Both density estimation methods fulfill these requirements.

Finally we prove our main result Theorem 1 under the regression model (R). We first prove a stronger version for Theorem 1, next we show that it will coincide with Theorem 1 under the assumption that $\mathcal{D}_T^1 \in \mathcal{D}^1$.

Theorem 7. *Let $\mathcal{D}_T$ be an arbitrary target distribution. Then, for any $\delta > 0$, there exists $\eta > 0$ and $z \in \Delta$ such that the following inequality holds for any $\alpha > 1$:*

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

where

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

and $\mathcal{D}_{k,T}(x,y) = \mathcal{D}_k^1(x)\mathcal{D}_T(y|x)$, $\mathcal{D}_{P,T} = \left\{\sum_{k=1}^p \lambda_k \mathcal{D}_{k,T}, \lambda \in \Delta\right\}$.

Proof. For any domain k, by Hölder's inequality, the following holds:

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

where, for simplicity, we write $\mathsf{d}_\alpha(x;T,k) = \mathsf{d}_\alpha\left(\mathcal{D}_T(\cdot|x) \parallel \mathcal{D}_k(\cdot|x)\right)$. Using the fact that the loss is bounded and Hölder's inequality again,

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

We can now apply the result of Corollary 5, with ϵ_T instead of ϵ and $\mathcal{D}_{k,T}$ instead of $\mathcal{D}_k$. This completes the proof. □

When $\mathcal{D}_T^1 \in \mathcal{D}^1$, $\mathcal{D}_T \in \mathcal{D}_{P,T}$, thus by the definition of Rényi divergence, $\mathsf{d}_\alpha(\mathcal{D}_T \parallel \mathcal{D}_{P,T}) = 1$. Theorem 7 coincides with Theorem 1 in this case.

A.2 Probability model

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

Lemma 3. *For any $\eta, \eta' > 0$, there exists $z \in \Delta$, with $z_k \neq 0$ for all $k \in [p]$, such that the following holds for the distribution-weighted combining rule h_z^η:*

$$\forall k \in [p], \quad \mathcal{L}(\mathcal{D}_k, h_z^\eta) \leq \sum_{j=1}^p z_j \mathcal{L}(\mathcal{D}_j, h_z^\eta) + \eta'. \tag{5}$$

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

Corollary 4. *For any $\delta > 0$, there exists $\eta > 0$ and $z \in \Delta$, such that $\mathcal{L}(\mathcal{D}_\lambda, h_z^\eta) \leq \epsilon + \delta$ for any mixture parameter $\lambda \in \Delta$.*

Proof. Modifying the proof of Corollary 4 for the regression model gives

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

By convexity of L, this implies that

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

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

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

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

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

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

□

Since we do not assume the conditional probabilities are the same across domains, we can directly prove Theorem 7 for the conditional probability model (P), which coincides with Theorem 1 when $\mathcal{D}_T \in \mathcal{D}$.

Theorem 7. *Let $\mathcal{D}_T$ be an arbitrary target distribution. For any $\delta > 0$, there exists $\eta > 0$ and $z \in \Delta$, such that the following inequality holds for any $\alpha > 1$:*

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

Proof. The proof is syntactically the same as that of Corollary 5 for the regression model. □

Corollary 6. *Let $\mathcal{D}_T$ be an arbitrary target distribution. Then, for any $\delta > 0$, there exists $\eta > 0$ and $z \in \Delta$, such that the following inequality holds for any $\alpha > 1$:*

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

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 $\widehat{\mathcal{D}} = \big\{\sum_{k=1}^p \lambda_k \widehat{\mathcal{D}}_k : \lambda \in \Delta\big\}$.

Proof. The proof is syntactically the same as that of Corollary 6 for the regression model. □

B Specific theoretical analysis for the cross-entropy loss

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

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

Theorem 8. *Assume there exists $\mu > 0$ such that $\mathcal{D}_k(x, y) \geq \mu \mathcal{U}(x, y)$ for all $k \in [p]$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Then, for any $\delta > 0$, there exists $\eta > 0$ and $z \in \Delta$, such that $\mathcal{L}(\mathcal{D}_\lambda, \overline{h}_z^\eta) \leq \epsilon + \delta$ for any mixture parameter $\lambda \in \Delta$.*

Proof. By the proof of Corollary 4 for the probability model, for any mixture distribution $\mathcal{D}_\lambda$:

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

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

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

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

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

It follows that

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

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

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

$$\widetilde{h}_z^\eta(x, y) = \sum_{k=1}^{p} \frac{z_k \mathcal{D}_k^1(x) + \eta \frac{\mathcal{U}^1(x)}{p}}{\sum_{j=1}^{p} z_j \mathcal{D}_j^1(x) + \eta \mathcal{U}^1(x)} h_k(x, y), \tag{6}$$

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

One can modify Theorem 7 to obtain generalization guarantees for $\widetilde{h}_z^\eta$ under distinct conditional probabilities assumption. Let $\mathcal{D}_T(x, y)$, ϵ_T and $\mathcal{D}_{P,T}$ be defined as before.

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

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

Proof. The proof is syntactically the same as that of Theorem 7. □

Finally, we can extend Theorem 8 and Theorem 9 to the case where only estimate distributions $\widehat{\mathcal{D}}_k$s are available, and the predictor $\widehat{\overline{h_z^\eta}}$ and $\widehat{\overline{\ell_z^\eta}}$ based on the estimates $\widehat{\mathcal{D}}_k$ still admit favorable guarantees. The results and proofs are similar to proving Corollary 6 from Corollary 5 in the regression model, thus omitted here.

C DC-decomposition

In this section we give the full DC-decompositions and their proofs mentioned in Section 4.

Proposition 10. *Let L be the squared loss. Then, for any $k \in [p]$, $\mathcal{L}(\mathcal{D}_k, h_z^\eta) - \mathcal{L}(\mathcal{D}_z, h_z^\eta) = u_k(z) - v_k(z)$, where u_k and v_k are convex functions defined for all z by*

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

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

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

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

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

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

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

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

The Hessian matrix of J_z^2/K_z is

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

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

Proposition 11. *Let L be the cross-entropy loss. Then, for $k \in [p]$, $\mathcal{L}(\mathcal{D}_k, h_z^\eta) - \mathcal{L}(\mathcal{D}_z, h_z^\eta) = u_k(z) - v_k(z)$, where u_k and v_k are convex functions defined for all z by*

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

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

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

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

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

The unnormalized relative entropy of P and Q is defined by

$$B(P \parallel Q) = \sum_{x,y} P(x,y) \log\left[\frac{P(x,y)}{Q(x,y)}\right] + \sum_{(x,y)} (Q(x,y) - P(x,y)).$$

The unnormalized relative entropy $B(\cdot \parallel \cdot)$ is jointly convex (3),[1] thus $B(K_z \parallel J_z)$ is convex as the composition of the unnormalized relative entropy with affine functions (for each of its two arguments). $(K_z - J_z)$ is an affine function of z and is therefore convex too. □

[1] To be precise, it can be shown that the relative entropy is jointly convex using the so-called log-sum inequality (3). The same proof using the log-sum inequality can be used to show the joint convexity of the unnormalized relative entropy.

D Rényi Divergence

The Rényi Divergence measures the divergence between two distributions. The Rényi Divergence is parameterized by α and denoted by D_α. The α-Rényi Divergence of two distributions $\mathcal{D}$ and $\mathcal{D}'$ is defined by

$$\mathsf{D}_\alpha(\mathcal{D} \parallel \mathcal{D}') = \frac{1}{\alpha - 1} \log \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \mathcal{D}(x,y) \left[\frac{\mathcal{D}(x,y)}{\mathcal{D}'(x,y)} \right]^{\alpha - 1}. \tag{7}$$

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 $\mathcal{D} = \mathcal{D}'$, (see (1)). We will denote by $\mathsf{d}_\alpha(\mathcal{D} \parallel \mathcal{D}')$ the exponential:

$$\mathsf{d}_\alpha(\mathcal{D} \parallel \mathcal{D}') = e^{\mathsf{D}_\alpha(\mathcal{D} \parallel \mathcal{D}')} = \left[\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \frac{\mathcal{D}^\alpha(x,y)}{\mathcal{D}'^{\alpha-1}(x,y)} \right]^{\frac{1}{\alpha-1}}. \tag{8}$$

Rényi divergence (and $\mathsf{d}_\alpha(\mathcal{D} \parallel \mathcal{D}')$) is nondecreasing as a function of α, and

$$\mathsf{d}_\alpha(\mathcal{D} \parallel \mathcal{D}') \leq \mathsf{d}_\infty(\mathcal{D} \parallel \mathcal{D}') = \sup_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \left[\frac{\mathcal{D}(x,y)}{\mathcal{D}'(x,y)} \right]. \tag{9}$$