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:

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

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 $\mathcal{D}_T$:*

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

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 $\lambda \in \Delta$:

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

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:

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

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:

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

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

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

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:

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

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 $\mathcal{D}_T$:*

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

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

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 $\mathcal{D}_T$:*

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

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:

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

Thus, by Lemma 2, we have

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

Plugging these inequalities into the bound of Theorem 3 yields:

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

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

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 $\widehat{\mathcal{Q}}$:*

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

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}}\left[\mathsf{d}_{\alpha}^{\alpha-1}(\mathcal{Q}(\cdot|x) \parallel \widehat{\mathcal{Q}}(\cdot|x))\right] &= \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

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

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 $[h_{d,z}]_k(x) = \widehat{\mathcal{Q}}(k|x) \left(h_k(x) + y - 2g_z(x) \right)$ 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: $K_\sigma(x,x') = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^d \exp\left\{-\frac{\|x-x'\|^2}{2\sigma^2}\right\}$, 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:

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

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**