Appendix

A Existence of a good hypothesis

We define the set $\mathcal{F}$ of consistent target functions:

$$\mathcal{F} = \{g : \forall i \in \{1, ..., k\}, \mathcal{L}(h_i, p_i, g) \leq \epsilon\} \tag{24}$$

In this section, we will show that for any target function $f \in \mathcal{F}$ there is a distribution weighted combining rule h_w^η that has a loss of at most ϵ with respect to any mixture p_T, as given in the paper by Mansour, Mohri, and Rostamizadeh [11].

Proposition 1. *Let U denote the uniform distribution over X. Furthermore, for any $\eta \geq 0$ and $w \in \Delta$, let h_w^η be the function defined in Definition 1. Then, for any distribution p, $\mathcal{L}(h_w^\eta, p, f)$ is continuous.*

Theorem 2. *(Brouwer Fixed Point Theorem) For any compact and convex non-empty set $A \subset R^n$ and any continuous function $f : A \to A$, there is a point $x \in A$ such that $f(x) = x$.*

We first show that there exists a distribution weighted combining rule h_w^η for which the losses $\mathcal{L}(h_w^\eta, p_i, f)$ are all nearly the same.

Lemma 1. *For any target function $f \in \mathcal{F}$ and any $\eta, \eta' \geq 0$, there exists $w \in \Delta$ with $w_i \neq 0$ for all $i \in \{1, ..., k\}$, such that the following holds:*

$$\mathcal{L}(h_w^\eta, p_i, f) = \gamma + \eta' - \frac{\eta'}{w_i k} \leq \gamma + \eta' \quad \forall i \in \{1, ..., k\} \tag{25}$$

Where:

$$\gamma = \sum_{j=1}^{k} w_j \mathcal{L}(h_w^\eta, p_j, f) \tag{26}$$

Proof. Consider the mapping $\psi : \Delta \to \Delta$ defined by:

$$[\psi(w)]_i = \frac{w_i \mathcal{L}(h_w^\eta, p_i, f) + \eta'/k}{\sum_{j=1}^{k} w_j \mathcal{L}(h_w^\eta, p_j, f) + \eta'} \tag{27}$$

By Proposition 1, ψ is continuous. Thus, by Theorem 2, there exists $w \in \Delta$ such that $\psi(w) = w$. This implies that:

$$w_i = \frac{w_i \mathcal{L}(h_w^\eta, p_i, f) + \eta'/k}{\sum_{j=1}^{k} w_j \mathcal{L}(h_w^\eta, p_j, f) + \eta'}$$

Since $\eta' > 0$, $w_i > 0$ for all $i \in \{1, ..., k\}$, we can divide by w_i:

$$\begin{aligned}
\mathcal{L}(h_w^\eta, p_i, f) &= \sum_{j=1}^{k} w_j \mathcal{L}(h_w^\eta, p_j, f) + \eta' - \frac{\eta'}{w_i k} \\
&= \gamma + \eta' - \frac{\eta'}{w_i k} && \textit{by equation (26).} \\
&\leq \gamma + \eta'
\end{aligned}$$

□

Lemma 2. *For any target function $f \in \mathcal{F}$ and any $\eta, \eta' \geq 0$, there exists $w \in \Delta$ such that $\mathcal{L}(h_w^\eta, p_\lambda, f) \leq \epsilon + \eta M + \eta'$ for any $\lambda \in \Delta$.*

Proof. Let w be the parameter guaranteed in Lemma 1. Then:

$$\begin{aligned}
\forall i \in \{1, ..., k\} \;\; \mathcal{L}(h_w^\eta, p_i, f) &= \gamma + \eta' - \frac{\eta'}{w_i k} && \textit{By Lemma 1.} \\
&= \sum_{j=1}^{k} w_j \mathcal{L}(h_w^\eta, p_j, f) + \eta' - \frac{\eta'}{w_i k} && \textit{By equation (26).} \\
&\leq \sum_{j=1}^{k} w_j \mathcal{L}(h_w^\eta, p_j, f) + \eta' && \eta', w_i, k \geq 0.
\end{aligned}$$

Consider the mixture $p_w = \sum_{i=1}^{k} w_i p_i$ for the guaranteed w. On the one hand, by definition of $\mathcal{L}$,

$$\mathcal{L}(h_w^\eta, p_w, f) = \sum_{x \in X} \sum_{i=1}^{k} w_i L(h_w^\eta, f) p_i(x) = \sum_{i=1}^{k} w_i \mathcal{L}(h_w^\eta, p_i, f) = \gamma \tag{28}$$

On the other hand:

$$\begin{aligned}
\mathcal{L}(h_w^\eta, p_w, f) &= \sum_{x\in X} L(h_w^\eta, f) p_w(x) \\
&\leq \sum_{x\in X} p_w(x) \sum_{i=1}^{k} \frac{w_i p_i(x) + \eta\frac{U(x)}{k}}{\sum_{j=1}^{k}(w_j p_j(x)) + \eta U(x)} L(h_i, f) && \textit{By convexity of } L \\
&= \sum_{x\in X} \frac{p_w(x)}{p_w(x) + \eta U(x)} \left(\sum_{i=1}^{k} (w_i p_i(x) + \eta\frac{U(x)}{k}) L(h_i, f) \right) \\
&\leq \sum_{x\in X} \left(\sum_{i=1}^{k} w_i p_i(x) L(h_i, f) \right) + \sum_{x\in X} \eta U(x) M \\
&= \sum_{i=1}^{k} w_i \mathcal{L}(h_i, p_i, f) + \eta M && \textit{By the definition of } \mathcal{L}(h, p, f) \\
&\leq \sum_{i=1}^{k} w_i \epsilon + \eta M \\
&\leq \epsilon + \eta M
\end{aligned}$$

Therefore:

$$\gamma = \sum_{i=1}^{k} w_i \mathcal{L}(h_w^\eta, p_i, f) = \mathcal{L}(h_w^\eta, p_w, f) \leq \epsilon + \eta M$$

To complete the proof, note that the following inequality holds for any mixture p_λ:

$$\mathcal{L}(h_w^\eta, p_\lambda, f) = \sum_{i=1}^{k} \lambda_i \mathcal{L}(h_w^\eta, p_i, f) \leq \gamma + \eta' \qquad \textit{By Lemma 1}$$

□

By Setting $\eta = \delta/2M$ and $\eta' = \delta/2$ we can derive Theorem 1.

B Selected properties of Rényi divergence

The properties were taken from [16].

Table 3: Special cases in the Rényi divergence family

α	***Definition***	***Notes***
$\alpha \to 0$	$-\log q(\{p > 0\})$	Not a divergence
$\alpha \to 1$	$E_{x\sim p(x)}[\log \frac{p(x)}{q(x)}]$	KL divergence
$\alpha = \frac{1}{2}$	$-2\log\left(1 - \mathrm{Hel}^2(p\|\|q)\right)$	Rényi divergence symmetric in its arguments.
$\alpha = 2$	$-\log\left(1 - \chi^2(p\|\|q)\right)$	Correlated to the χ^2 divergence
$\alpha \to \infty$	$\log\max\left(\frac{p}{q}\right)$	Worst-case regret

Theorem 3. *(Positivity): For any order $\alpha \in [0, \infty]$: $D_\alpha(p||q) \geq 0$, and $D_\alpha(p||q) = 0 \iff p = q$*

Theorem 4. *(Convexity): For any order $\alpha \in [0, 1]$ Rényi divergence is jointly convex in its arguments. That is, for any two pairs of probability distributions (p_0, q_0) and (p_1, q_1), and any $0 < \lambda < 1$:*

$$D_\alpha((1-\lambda)p_0 + \lambda p_1||(1-\lambda)q_0 + \lambda q_1) \leq (1-\lambda)D_\alpha(p_0||q_0) + \lambda D_\alpha(p_1||q_1) \quad (29)$$

For any order $\alpha \in [0, \infty]$ Rényi divergence is convex in its second argument. That is, for any probability distributions p, q_0 and q_1:

$$D_\alpha(p||(1-\lambda)q_0 + \lambda q_1) \leq (1-\lambda)D_\alpha(p||q_0) + \lambda D_\alpha(p||q_1) \quad (30)$$

Theorem 5. *(Continuity in the Order): The Rényi divergence is continuous in α on $A = \{\alpha \in [0, \infty]|0 \leq \alpha \leq 1$ or $D_\alpha(p||q) < \infty\}$.*

Theorem 6. *(Monotonicity) [9]: Rényi divergence, extended to negative α, is continuous and non-decreasing on $\alpha \in \{\alpha : -\infty < D_\alpha < +\infty\}$.*

Lemma 3. *The Skew Symmetry property:*

- *For any* $\alpha \in (-\infty, \infty), \alpha \notin \{0, 1\}$

$$D_\alpha(p||q) = \frac{\alpha}{1-\alpha} D_{1-\alpha}(q||p)$$
$$D_{-\infty}(p||q) = -D_\infty(q||p)$$

- *For any* $\alpha \in (-\infty, \infty), \alpha \notin \{0, 1\}$

$$D_\alpha(p||q) \leq \frac{\alpha}{1-\alpha} D_{1-\alpha}(p||q)$$

Many other properties described in [16], [9].

C VRS experiments

All our experiments were conducted using PyTorch. Throughout the experiments we used K=50 samples for Monte Carlo (MC) approximation; trained the VAE models using the ADAM optimizer [7]; set the learning rate to 0.001 and the batch size to 128 for the training set and 32 for the test set. Our VAE model includes a total of 6 linear layers. The first 3 are the encoder layers, and the last 3 are the decoder layers. The dimension of the latent space is 50.

We suggest two perspectives to evaluate and compare the performance:

- **Quality of the decoded signal** - Reconstruction error, measured by Mean Square Error (MSE) and Cross-Entropy (CE).
- **Quality of the evidence approximation** - Maximizing the evidence log likelihood, $\log p(x)$; the higher the better.

In the following experiments, we used the MNIST, USPS, and SVHN datasets, all of which contain digit images. They all share 10 classes of digits. USPS dataset consists of 7,291 training images and 2,007 test images of size 16×16. MNIST dataset consists of 60,000 training images and 10,000 test images of size 28×28. SVHN is obtained from house numbers in Google Street View images. It has 73,257 training images and 26,032 test images of size 32×32.

First, we compared the learning curves of $\mathbf{VRS}_{\alpha_+,\alpha_-}$ with $\alpha_- \in \{-0.5, -2\}$, $\alpha_+ \in \{0.5, 2\}$ and $\mathbf{VR}_\alpha$ with $\alpha \in \{0.5, 2, 5\}$ over MNIST dataset. Fig. 4 demonstrates that $\mathbf{VRS}_{\alpha_+,\alpha_-}$ converged faster than $\mathbf{VR}_\alpha$ and the resulted loss value is smaller for both α values. Also, we can see that $\mathbf{VR}_{0.5}$ performs better than $\mathbf{VR}_2$, and $\mathbf{VR}_2$ performs better than $\mathbf{VR}_5$. This observation is in sync with the results reported in [9].

Second, we trained variational autoencder models using $\mathbf{VRS}_{\alpha_+,\alpha_-}$ with $(\alpha_-, \alpha_+) \in \{(-0.5, 0.5), (-2, 2), (-5, 5)\}$, and compared performances with models trained using the **VAE** (equivalent to $\alpha = 1$), $\mathbf{VR}_\alpha$ ($\alpha \in \{0.5, 2, 5\}$) and $\mathbf{VRLU}_\alpha$ ($\alpha \in \{-0.5, -2, -5\}$) methods. The models' performance was evaluated over digits datasets (MNIST, SVHN, USPS) with respect to reconstruction error and log-likelihood maximization.

Table 4 depicts the reconstruction errors, cross-entropy (CE) and mean squared error (MSE), and the log-likelihood for the different learning methods. We can see that both MSE and CE reconstruction errors of $\mathbf{VRS}_{\alpha_+,\alpha_-}$ are better than **VAE** reconstruction errors in all the datasets. Moreover, the $\mathbf{VRS}_{\alpha_+,\alpha_-}$ trained on USPS obtain the best CE error, The MSE results obtained by $\mathbf{VRS}_{\alpha_+,\alpha_-}$ are good (2nd in MNIST and USPS), but not the best. As for the log-likelihood in the different datasets, we can see that $\mathbf{VRS}_{\alpha_+,\alpha_-}$ provides good results (either 1st or 2nd best) in all datasets, compared to the other methods.

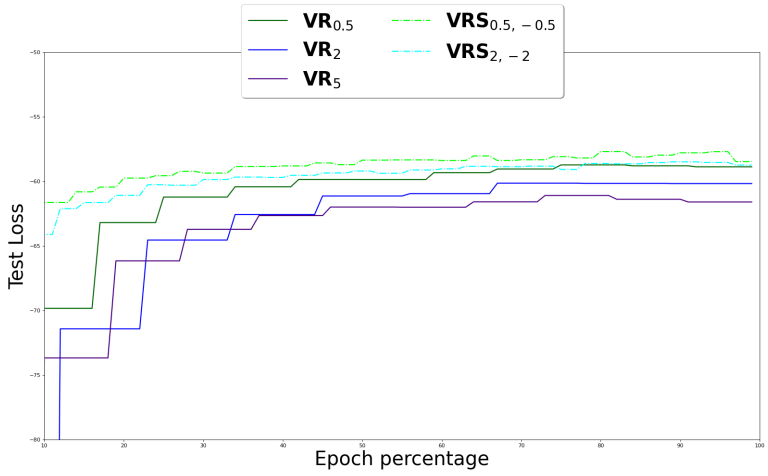

Fig. 4: Comparison between $\mathbf{VR}_\alpha$ and $\mathbf{VRS}_{\alpha_+,\alpha_-}$ learning curves over 'MNIST' dataset. Training with different values of α.

Table 4: Comparison between the log-likelihood of **VAE**, $\mathbf{VR}_{0.5}$, $\mathbf{VRLU}_{-0.5}$ and $\mathbf{VRS}_{0.5,-0.5}$, comparison between the mean square error of **VAE**, $\mathbf{VR}_2$, $\mathbf{VRLU}_{-2}$ and $\mathbf{VRS}_{2,-2}$ and comparison between the cross entropy error of **VAE**, $\mathbf{VR}_5$, $\mathbf{VRLU}_{-5}$ and $\mathbf{VRS}_{5,-5}$.

Log-Likelihood				
Datasets	**VAE**	$\mathbf{VR}_{0.5}$	$\mathbf{VRLU}_{-0.5}$	$\mathbf{VRS}_{0.5,-0.5}$
MNIST	-75.92	-56.03	**-55.58**	-56.10
USPS	-261.51	-251.53	-251.79	**-251.39**
SVHN	-481.09	-472.98	-478.47	**-472.37**

Mean Square Error				
Datasets	**VAE**	$\mathbf{VR}_2$	$\mathbf{VRLU}_{-2}$	$\mathbf{VRS}_{2,-2}$
MNIST	55.87	20.50	**17.11**	17.56
USPS	29.60	**7.38**	8.70	7.41
SVHN	34.09	**12.54**	20.31	23.43

Cross Entropy Error				
Datasets	**VAE**	$\mathbf{VR}_5$	$\mathbf{VRLU}_{-5}$	$\mathbf{VRS}_{5,-5}$
MNIST	1.24	**0.73**	0.75	0.78
USPS	4.29	4.18	4.31	**4.04**
SVHN	7.77	7.80	**7.53**	7.60

D MSA with estimated probabilities

D.1 An upper bound for loss with estimated probabilities

Corollary 1. *Let $\hat{p_i}$ be an estimation of the original domain distribution p_i. The following inequality holds for any $\alpha > 1$:*

$$\mathcal{L}(h_i, \hat{p_i}, f) \leq [d_\alpha(\hat{p_i}||p_i)\epsilon]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}} \tag{31}$$

Proof. [6] First, let us recall Holder's inequality:

Theorem 7. ***Holder's inequality:*** *For any s and t in the open interval $(1, \infty)$ with $\frac{1}{s} + \frac{1}{t} = 1$, and for $\{x_j\}$ and $\{y_j\}$ $j \in \{1, ..., k\}$ be sets of real numbers, we have:*

$$\sum_{j=1}^{n} |x_j y_j| \leq \left(\sum_{j=1}^{n} |x_j|^s \right)^{\frac{1}{s}} \left(\sum_{j=1}^{n} |y_j|^t \right)^{\frac{1}{t}} \tag{32}$$

For any hypothesis h and any distributions p, q, and for any $\alpha > 1$, the following holds:

$$\begin{aligned}
\mathcal{L}(h, q, f) &= \sum_{x \in X} q(x) L(h, f) \\
&= \sum_{x \in X} \left[\frac{q(x)}{p(x)^{\frac{\alpha-1}{\alpha}}} \right] p(x)^{\frac{\alpha-1}{\alpha}} L(h, f) \\
&\leq \left[\sum_{x \in X} \frac{q(x)^\alpha}{p(x)^{\alpha-1}} \right]^{\frac{1}{\alpha}} \left[\sum_{x \in X} p(x) L(h, f)^{\frac{\alpha}{\alpha-1}} \right]^{\frac{\alpha-1}{\alpha}} && \text{By Holder's inequality for } s = \alpha, t = \frac{\alpha}{\alpha-1} \\
&= \left(\left[\sum_{x \in X} \frac{q(x)^\alpha}{p(x)^{\alpha-1}} \right]^{\frac{1}{\alpha-1}} \right)^{\frac{\alpha-1}{\alpha}} \left[\sum_{x \in X} p(x) L(h, f) L(h, f)^{\frac{1}{\alpha-1}} \right]^{\frac{\alpha-1}{\alpha}} \\
&= (d_\alpha(q||p))^{\frac{\alpha-1}{\alpha}} \left[\sum_{x \in X} p(x) L(h, f) L(h, f)^{\frac{1}{\alpha-1}} \right]^{\frac{\alpha-1}{\alpha}} && \text{By Definition 4} \\
&\leq (d_\alpha(q||p))^{\frac{\alpha-1}{\alpha}} \left[\sum_{x \in X} p(x) L(h, f) M^{\frac{1}{\alpha-1}} \right]^{\frac{\alpha-1}{\alpha}} && \text{Since } M \geq |L(h, f)| \text{ and } \frac{1}{\alpha-1} > 0 \\
&= [d_\alpha(q||p) \mathcal{L}(h, p, f)]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}
\end{aligned}$$

[6] The proof is based on a similar corollary proven in [5].

For each $i \in \{1, ..., k\}$, by setting $p := p_i, q := \hat{p}_i$ and $h := h_i$, we will get that:

$$\begin{aligned}\mathcal{L}(h_i, \hat{p}_i, f) &\leq [d_\alpha(\hat{p}_i||p_i)\mathcal{L}(h_i, p_i, f)]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}} \\ &\leq [d_\alpha(\hat{p}_i||p_i)\epsilon]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}} \qquad \textit{We assume that } \forall i \in \{1, ..., k\} \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \mathcal{L}(h_i, p_i, f) \leq \epsilon\end{aligned}$$

□

Corollary 1 provides us an upper bound of the loss using the estimated distributions $\hat{p}_i$. When $\hat{p}_i \to p_i$, $d_\alpha(\hat{p}_i||p_i) \to 1$ and we will remain with $\epsilon^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}$. We will set $M = 1$, since we use the loss function $L(h, f) = -\log(h(x, f(x)))$ as the cross-entropy loss (log-loss). Thus, when $\hat{p}_i \to p_i$, $[d_\alpha(\hat{p}_i||p_i)\epsilon]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}} \to \epsilon^{\frac{\alpha-1}{\alpha}}$. Let us mark $\forall i \in \{1, ..., k\}$ $[d_\alpha(\hat{p}_i||p_i)\epsilon]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}} = \epsilon_i^*$.

D.2 A lower bound for loss with estimated probabilities

By performing the calculation from the previous section using $\alpha < 1$, we can derive a lower bound for $\mathcal{L}(h_i, \hat{p}_i, f)$. The confirmation of whether approximated probabilities lead to excessive error can be accomplished by utilizing this lower limit. If the lower bound value is significantly high, it indicates that our approximation is insufficient. Conversely, if the lower bound value is small, it signifies that our approximation is reliable. By employing both upper and lower bounds, we can attain a more precise estimation of the loss.

Corollary 2. *Let $\hat{p}_i$ be an estimation of the original domain distribution p_i. The following inequality holds for any $\alpha < 1$:*

$$\mathcal{L}(h_i, \hat{p}_i, f) \geq (d_\alpha(\hat{p}_i||p_i))^{\frac{\alpha-1}{\alpha}} \psi \tag{33}$$

Where $\psi = \left[\sum_{x \in X} p_i(x) L(h_i, f)^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}$

Proof. First, we will prove for $0 < \alpha < 1$, and then for $\alpha < 0$.

Theorem 8. *(Generalization of Holder's inequality) [1]: Let $0 < s < 1$ and $t \in R$ with $\frac{1}{s} + \frac{1}{t} = 1$, and for $\{x_j\}$ and $\{y_j\}$ $j \in \{1, ..., n\}$ be sets of real numbers, we have:*

$$\sum_{j=1}^{n} |x_j y_j| \geq \left(\sum_{j=1}^{n} |x_j|^s\right)^{\frac{1}{s}} \left(\sum_{j=1}^{n} |y_j|^t\right)^{\frac{1}{t}} \tag{34}$$

Let's set $0 < \alpha < 1$, $s = \alpha$ and $t = \frac{\alpha}{\alpha-1}$. For any hypothesis h and any distributions p, q, the following holds:

$$\begin{aligned}
\mathcal{L}(h,q,f) &= \sum_{x\in X} q(x)L(h,f) \\
&= \sum_{x\in X} \left[\frac{q(x)}{p(x)^{\frac{\alpha-1}{\alpha}}}\right] p(x)^{\frac{\alpha-1}{\alpha}} L(h,f) \\
&\geq \left[\sum_{x\in X} \frac{q(x)^\alpha}{p(x)^{\alpha-1}}\right]^{\frac{1}{\alpha}} \left[\sum_{x\in X} p(x)L(h,f)^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}} && \textit{By the generalization of} \\
& && \textit{Holder's inequality for} \\
& && s=\alpha, t=\frac{\alpha}{\alpha-1} \\
&= \left(\left[\sum_{x\in X} \frac{q(x)^\alpha}{p(x)^{\alpha-1}}\right]^{\frac{1}{\alpha-1}}\right)^{\frac{\alpha-1}{\alpha}} \left[\sum_{x\in X} p(x)L(h,f)^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}} \\
&= \left(d_\alpha(q||p)\right)^{\frac{\alpha-1}{\alpha}} \left[\sum_{x\in X} p(x)L(h,f)^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}} && \textit{By Definition 4}
\end{aligned}$$

Next, let's set $\alpha < 0$, $t = \alpha$ and $s = \frac{\alpha}{\alpha-1}$ (notice that $\alpha < 0 \rightarrow 0 < s < 1$). For any hypothesis h and any distributions p, q, the following holds:

$$\begin{aligned}
\mathcal{L}(h,q,f) &= \sum_{x\in X} q(x)L(h,f) \\
&= \sum_{x\in X} \left[\frac{q(x)}{p(x)^{\frac{\alpha-1}{\alpha}}}\right] p(x)^{\frac{\alpha-1}{\alpha}} L(h,f) \\
&= \sum_{x\in X} p(x)^{\frac{\alpha-1}{\alpha}} L(h,f) \left[\frac{q(x)}{p(x)^{\frac{\alpha-1}{\alpha}}}\right] \\
&\geq \left[\sum_{x\in X} p(x)L(h,f)^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}} \left[\sum_{x\in X} \frac{q(x)^\alpha}{p(x)^{\alpha-1}}\right]^{\frac{1}{\alpha}} && \textit{By the generalization of} \\
& && \textit{Holder's inequality for} \\
& && t=\alpha, s=\frac{\alpha}{\alpha-1}
\end{aligned}$$

$$= \left[\sum_{x\in X} \frac{q(x)^\alpha}{p(x)^{\alpha-1}}\right]^{\frac{1}{\alpha}} \left[\sum_{x\in X} p(x)L(h,f)^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}$$

$$= \left(\left[\sum_{x\in X} \frac{q(x)^\alpha}{p(x)^{\alpha-1}}\right]^{\frac{1}{\alpha-1}}\right)^{\frac{\alpha-1}{\alpha}} \left[\sum_{x\in X} p(x)L(h,f)^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}$$

$$= (d_\alpha(q||p))^{\frac{\alpha-1}{\alpha}} \left[\sum_{x\in X} p(x)L(h,f)^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}} \qquad \textit{By Definition 4}$$

For each $i \in \{1, ..., k\}$, by setting $p := p_i, q := \hat{p}_i$ and $h := h_i$, we will get that:

$$\mathcal{L}(h_i, \hat{p}_i, f) \geq (d_\alpha(\hat{p}_i||p_i))^{\frac{\alpha-1}{\alpha}} \left[\sum_{x\in X} p_i(x)L(h_i,f)^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}} = (d_\alpha(\hat{p}_i||p_i))^{\frac{\alpha-1}{\alpha}}\ \psi$$

□

We would like to argue that the value of $\psi = \left[\sum_{x\in X} p_i(x)L(h_i,f)^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}$ can be ignored when investigating the loss bound. Recall that we assume $L(h_i, f) \leq M$. We set $M = 1$ (as we mentioned before), hence, we are left with $\left(\sum_{x\in X} p_i(x)\right)^{\frac{\alpha-1}{\alpha}}$. Since p_i is a distribution, the sum is equal to 1.

D.3 Bound for the estimated good hypothesis

Theorem 9. *Let p_T be an arbitrary target distribution. For any $\delta > 0$, there exists $\eta > 0$ and $w \in \Delta$, such that the following inequality holds for any $\alpha > 1$ and any mixture parameter λ:*

$$\mathcal{L}(h^\eta_w, p_T, f) \leq [(\epsilon + \delta)d_\alpha(p_T||p_\lambda)]^{\frac{\alpha-1}{\alpha}}\ M^{\frac{1}{\alpha}} \tag{35}$$

Proof. Let $\delta > 0$. In the proof for Corollary 1 we showed that for any hypothesis h and any distributions p, q, and for any $\alpha > 1$, the following holds:

$$\mathcal{L}(h, q, f) \leq [d_\alpha(q||p)\mathcal{L}(h, p, f)]^{\frac{\alpha-1}{\alpha}}\ M^{\frac{1}{\alpha}} \tag{36}$$

Hence, for $q = p_T$, $p = p_\lambda$ and $h = h^\eta_w$ we will get that:

$$\mathcal{L}(h^\eta_w, p_T, f) \leq [d_\alpha(p_T||p_\lambda)\mathcal{L}(h^\eta_w, p_\lambda, f)]^{\frac{\alpha-1}{\alpha}}\ M^{\frac{1}{\alpha}} \tag{37}$$

By Theorem 1, given $\delta > 0$, there exist $\eta > 0$ and $w \in \Delta$ such that $\mathcal{L}(h^\eta_w, p_\lambda, f) \leq \epsilon + \delta$ for any mixture parameter λ. Therefore:

$$\mathcal{L}(h^\eta_w, p_T, f) \leq [d_\alpha(p_T||p_\lambda)(\epsilon + \delta)]^{\frac{\alpha-1}{\alpha}}\ M^{\frac{1}{\alpha}} \tag{38}$$

□

Corollary 3. *Let p_T be an arbitrary target distribution. For any $\delta > 0$, there exists $\eta > 0$ and $w \in \Delta$, such that the following inequality holds for any $\alpha > 1$ and any mixture parameter λ:*

$$\mathcal{L}(\hat{h}^{\eta}_{w}, p_T, f) \leq [(\epsilon^* + \delta) d_{\alpha}(p_T || \hat{p}_{\lambda})]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}} \tag{39}$$

Where $\hat{p}_{\lambda} = \sum_{i=1}^{k} \lambda_i \hat{p}_i(x)$ (for $\lambda \in \Delta$) and $\hat{h}^{\eta}_{w}$ is our good hypothesis defined in Definition 1 but calculated with the estimated probabilities $\hat{p}_i$.

Proof. In Corollary 1 we showed that $\forall i \in \{1, ..., k\}$ and for any $\alpha > 1$:

$$\mathcal{L}(h_i, \hat{p}_i, f) \leq [d_{\alpha}(\hat{p}_i || p_i)\epsilon]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}} \leq \epsilon^*_i \tag{40}$$

Let's set ϵ^* such that:

$$\epsilon^* = \max_{i=1}^{k} \{\epsilon^*_i\} \tag{41}$$

Overall we got the following:

1. $\forall i \in \{1, ..., k\}$: $\mathcal{L}(h_i, \hat{p}_i, f) \leq \epsilon^*$
2. $\hat{h}^{\eta}_{w}(x, y) = \sum_{i=1}^{k} \frac{w_i \hat{p}_i(x) + \eta \frac{U(x)}{k}}{\sum_{j=1}^{k} (w_j \hat{p}_j(x)) + \eta U(x)} h_i(x, y)$

We can repeat the proof of Theorem 9 with ϵ^* instead of ϵ, $\hat{p}_i$ instead of p_i and $\hat{h}^{\eta}_{w}$ instead of h^{η}_{w}.

□

E Results

This section provides the results of the complete VRS-MSA experiment, where we evaluate and compare the performance of three variational inference models in the multiple source adaptation scenario: **VAE**, $\mathbf{VR}_{\alpha}$, and $\mathbf{VRS}_{\alpha+,\alpha-}$.

Table 5: Digit Dataset Accuracy (s - SVHN, m - MNIST and u - USPS). Previous results were taken from [2].

Models	**Test datasets**							
	s	m	u	mu	su	sm	smu	mean
CNN-s	92.3	66.9	65.6	66.7	90.4	85.2	84.2	78.8
CNN-m	15.7	99.2	79.7	96.0	20.3	38.9	41.0	55.8
CNN-u	16.7	62.3	96.6	68.1	22.5	29.4	32.9	46.9
CNN-unif	75.7	91.3	92.2	91.4	76.9	80.0	80.7	84.0
CNN-joint	90.9	99.1	96.0	98.6	91.3	93.2	93.3	94.6
GMSA	91.4	98.8	95.6	98.3	91.7	93.5	93.6	94.7
DMSA	92.3	**99.2**	**96.6**	**98.8**	92.6	94.2	94.3	95.4
VAE-MSA	72.1	97.7	94.6	96.0	92.3	95.7	95.7	92.0
VR_2-MSA	72.4	99.1	94.9	96.5	89.3	96.1	95.6	92.0
$VR_{0.5}$-MSA	70.0	99.1	95.1	96.5	89.2	96.1	95.7	91.7
$VRS_{2,-2}$-MSA	74.2	99.1	94.7	96.5	89.3	96.1	95.6	92.2
$VRS_{2,-0.5}$-MSA	71.5	98.9	95.7	96.5	87.5	95.9	95.6	91.6
$VRS_{0.5,-2}$-MSA	72.5	99.1	94.7	96.5	90.1	**96.1**	**95.7**	92.1
$VRS_{0.5,-0.5}$-MSA	76.0	99.1	94.6	96.5	89.4	95.8	95.4	92.4
VAE-SGD	93.8	99.0	94.6	98.3	93.8	95.2	95.2	95.7
VR_2-SGD	93.9	98.5	94.8	97.9	94.0	95.2	95.2	95.6
$VR_{0.5}$-SGD	93.7	99.0	94.8	98.3	93.8	95.2	95.2	95.7
$VRS_{2,-2}$-SGD	93.7	99.0	94.7	98.3	93.8	95.2	95.2	**95.7**
$VRS_{2,-0.5}$-SGD	93.9	98.4	95.0	97.8	94.0	95.2	95.1	95.6
$VRS_{0.5,-2}$-SGD	93.9	98.5	94.9	97.9	93.4	95.2	95.1	95.6
$VRS_{0.5,-0.5}$-SGD	**93.9**	98.4	94.9	97.8	**94.0**	95.2	95.2	95.6

Table 6: Office Dataset Accuracy (a - Amazon, w - Webcam, d - DSLR). Previous results were taken from [2].

Models	**Test datasets**							
	a	w	d	aw	ad	wd	awd	mean
Resnet-a	82.2	75.8	77.6	-	-	-	-	-
Resnet-w	63.3	95.7	95.7	-	-	-	-	-
Resnet-d	64.6	94.0	95.8	-	-	-	-	-
Resnet-unif	79.3	96.7	97.2	-	-	-	-	-
GMSA	82.1	96.8	96.7	-	-	-	-	-
DMSA	82.2	**97.2**	97.4	-	-	-	-	-
VAE-MSA	76.6	93.4	98.6	81.0	79.8	95.0	82.7	86.7
VR_2-MSA	76.0	94.1	98.2	80.5	79.0	95.2	82.4	86.5
$VR_{0.5}$-MSA	77.3	93.1	98.6	81.5	80.5	94.8	83.5	87.0
$VRS_{0.5,-2}$-MSA	69.0	93.0	**99.0**	74.6	72.6	94.9	77.0	82.9
$VRS_{2,-0.5}$-MSA	78.0	93.2	98.6	82.0	80.7	94.8	83.7	87.3
$VRS_{2,-2}$-MSA	81.6	92.2	98.6	84.5	84.0	94.3	86.0	88.7
$VRS_{0.5,-0.5}$-MSA	81.7	92.4	98.6	84.6	84.2	94.5	86.1	88.9
VAE-SGD	92.2	95.0	96.8	92.8	92.7	95.6	93.2	94.0
VR_2-SGD	92.2	95.0	96.8	92.8	92.7	95.6	93.1	94.0
$VR_{0.5}$-SGD	92.2	95.0	96.8	92.8	92.7	95.6	93.2	94.0
$VRS_{2,-2}$-SGD	92.2	95.0	96.8	92.7	92.7	95.6	93.1	94.0
$VRS_{2,-0.5}$-SGD	92.2	94.8	97.2	92.7	92.7	95.6	93.2	**94.1**
$VRS_{0.5,-2}$-SGD	92.2	95.0	96.8	92.8	92.7	95.6	93.1	94.0
$VRS_{0.5,-0.5}$-SGD	**92.2**	95.0	96.8	**92.8**	**92.7**	**95.6**	**93.2**	94.0

Tables 5 and 6 details the accuracy scores obtained from running the following models:

- CNN-s, CNN-m, CNN-u, and Resnet-a, Resnet-d, Resnet-w, each trained on the single source domain.
- CNN-unif and Resnet-unif, a classifier trained on a uniform combination of the source domains' data.
- CNN-joint, a global classifier trained on all the source domains' data combined.
- The GMSA model - a generative MSA model using the DC programming algorithm. To obtain the data distribution, GMSA used the last layer before softmax from each of the domains' classifiers.
- The DMSA model, which based on a discriminative technique using an estimate of the conditional probabilities (the probability that point x belongs to source i).

We compared the previous models to our combined variational inference with the DC programming model using the classic VAE model, the VR model with different positive α values, and the VRS model with different positive and negative

α values. The last part illustrates a combined model using VRS with Stochastic Gradient Descent instead of the DC algorithm.