# Supplementary Material
# Wasserstein Barycenter for Multi-Source Domain Adaptation

Eduardo Fernandes Montesuma
Universidade Federal do Ceará
Fortaleza, Brazil
eduardomontesuma@alu.ufc.br

Fred Maurice Ngolè Mboula
Université Paris-Saclay, Institut LIST, CEA
F-91120 Palaiseau, France
fred-maurice.ngole-mboula@cea.fr

In this file, we present some details that were left out from the main text in favor of the space limit. This material is divided into two sections: Section A, containing the proof of Theorem 1, and Section B, detailing the experiments.

## A. Theorem Proofs

For the proof of Theorem 1 three lemmas need to be stated. Lemma 1, which bounds the difference of the error functional on target and source between any two hypothesis $h, h'$ using the Wasserstein distance, Lemma 2, which presents the uniform convergence bound for the Wasserstein distance, and Lemma 3, which presents the concentration inequality for the error functional. The proof of these lemmas is omitted, but can found in the papers of their respctive authors.

**Lemma 1.** *(Due to [4]) Let $\mu_s, \mu_t \in \mathcal{P}(\mathcal{X})$ be two probability measures on $\mathcal{X} \subset \mathbb{R}^d$. Assume that the cost function $c(\mathbf{x}, \mathbf{y}) = ||\phi(\mathbf{x}) - \phi(\mathbf{y})||_{\mathcal{H}_{k_\ell}}$, where $\mathcal{H}_{k_\ell}$ is a Reproducing Kernel Hilbert Space (RKHS) with associated kernel $k_\ell : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ induced by $\phi : \mathcal{X} \rightarrow \mathcal{H}_{k_\ell}$ and $k_\ell(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathcal{H}_{k_\ell}}$. Assume further that the loss function $\ell_{h,f} : x \rightarrow \ell(h(x), f(x))$ is convex, symmetric, bounded, obeys the triangle inequality and has the parametric form $|h(x) - f(x)|^q$ for some $q > 0$. Assume also that $k_\ell$ is square-root integrable w.r.t. both $\mu_s$ and $\mu_t$, for all $\mu_s, \mu_t \in \mathcal{P}_p(\mathcal{X})$, where $\mathcal{X}$ is separable and $0 \leq k_\ell(\mathbf{x}, \mathbf{y}) \leq K, \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$. Then, the following holds,*

$$\epsilon_t(h, h') \leq \epsilon_s(h, h') + W_1(\mu_s, \mu_t),$$

*for every hypothesis $h, h' \in \mathcal{H}_{k_\ell}$.*

**Lemma 2.** *(Due to [2]) Let $\mu$ be a probability measure in $\mathbb{R}^d$, so that for some $\alpha > 0$ we have that $\int_{\mathbb{R}^d} e^{\alpha ||\mathbf{x}||^2} d\mu < \infty$ and $\hat{\mu} = \dfrac{1}{n} \sum_{i=1}^n \delta_{\mathbf{x}_i}$ be its associated empirical measure defined on a sample of independent variables $\{\mathbf{x}_i\}_{i=1}^n$ drawn from $\mu$. Then, for any $d' > d$ and $\xi' < \sqrt{2}$ there exists a constant $n_0$ depending on $d'$ and some square exponential moment of $\mu$ such that for any $\epsilon > 0$ and $n \geq n_0 max(\epsilon^{-(d'+2)}, 1)$,*

$$\mathbb{P}[W_1(\hat{\mu}, \mu) > \epsilon] \leq exp\bigg(-\frac{\xi'}{2} n \epsilon^2\bigg),$$

*where $d'$ and $\xi'$ can be calculated explicitly.*

The consequence of Lemma 2 is that we may express $\epsilon$ in terms of $\delta$,

$$\epsilon = \sqrt{\frac{2}{n\xi'} \log(\tfrac{1}{\delta})} \tag{1}$$

**Lemma 3.** *(Due to [4]) Under the assumptions of Lemma 1, let $D$ be a sample of size $n$, where for each $j \in \{1, \cdots, N\}$, $\beta_j n$ points are drawn from $\mu_{s_j}$ and labelled according to $f_j$. Then, for any fixed $\alpha$, with probability $1 - \delta$ for all $h$ the following holds,*

$$\mathbb{P}\bigg[|\hat{\epsilon}_\alpha(h) - \epsilon_\alpha(h)| > \epsilon + \theta\bigg] \leq 2exp\Bigg(\frac{-\epsilon^2 n}{2K \sum_{j=1}^N \frac{\alpha_j^2}{\beta_j}}\Bigg). \tag{2}$$

*where $\theta = 2\sqrt{K/n} \sum_{j=1}^N \dfrac{\alpha_j}{\beta_j n \sqrt{\beta_j}}$*

The consequence of Lemma 3 is that we may express $\epsilon$ in Equation 2 as,

$$\epsilon = \sqrt{\frac{2K \sum_{j=1}^N \frac{\alpha_j^2}{\beta_j} \log(\frac{2}{\delta})}{n}}. \tag{3}$$

**Theorem 1.** *(Due to [4]) Let $\mathbf{X}_{s_j}$, $j \in \{1, \cdots, N\}$ and $\mathbf{X}_t$ be $N + 1$ samples of size $n_{s_j}$ and $n_t$ drawn i.i.d. from*

*$\mu_{s_j}$ and $\mu_t$ respectively. Let $\hat{\mu}_{s_j}$ and $\hat{\mu}_{s_t}$ be the respective empirical measures. If $\hat{h}_\alpha$ is the empirical minimizer of $\hat{\epsilon}_\alpha$ and $h_t^* = \underset{h\in\mathcal{H}}{\text{minimize}}\,\epsilon_t(h)$, then for any fixed $\alpha$ and $\delta \in (0,1)$, with probability at least $1-\delta$ (over the choice of samples),*

$$\begin{aligned}\epsilon_t(\hat{h}_\alpha) \leq{}& \epsilon_t(h_T^*) + c_1 \\ &+ 2\sum_{j=1}^{N}\alpha_j(W_1(\hat{\mu}_{s_j},\hat{\mu}_t)+\lambda_j+c_2),\end{aligned} \tag{4}$$

*where,*

$$\begin{aligned} c_1 &= 2\sqrt{\frac{2K\sum_{j=1}^{N}\dfrac{\alpha_j^2}{\beta_j}\log(2/\delta)}{n}} + 2\sqrt{\sum_{j=1}^{N}\frac{K\alpha_j}{n\beta_j}}, \\ c_2 &= \sqrt{\frac{2\log(\frac{1}{\delta})}{\xi'}}\left(\sqrt{\frac{1}{n_{s_j}}}+\sqrt{\frac{1}{n_t}}\right), \\ \lambda_j &= \underset{h\in\mathcal{H}}{\textit{minimize}}\;\epsilon_{s_j}(h)+\epsilon_t(h). \end{aligned}$$

*Proof.* As suggested by [4], the proof steps are similar to those of Theorem 4 of [1],

$$\begin{aligned}|\epsilon_\alpha(h)-\epsilon_t(h)| ={}& \left|\sum_{j=1}^{N}\alpha_j\epsilon_{s_j}(h)-\epsilon_t(h)\right|, \\ \leq{}& \sum_{j=1}^{N}\alpha_j|\epsilon_{s_j}(h)-\epsilon_t(h)|.\end{aligned} \tag{5}$$

Now, let $h_j^* = \text{argmin}_{h\in\mathcal{H}}\epsilon_{s_j}(h)+\epsilon_t(h)$. Notice that we may rewrite the difference $|\epsilon_{s_j}(h)-\epsilon_t(h)|$ as,

$$\begin{aligned}|\epsilon_{s_j}(h)-\epsilon_t(h)| = |&\epsilon_{s_j}(h)-\epsilon_{s_j}(h,h_j^*) \\ &+\epsilon_{s_j}(h,h_j^*)-\epsilon_t(h,h_j^*) \\ &+\epsilon_t(h,h_j^*)-\epsilon_t(h)|,\end{aligned}$$

then, using the triangle inequality, it follows that,

$$\begin{aligned}|\epsilon_{s_j}(h)-\epsilon_t(h)| \leq{}& |\epsilon_{s_j}(h)-\epsilon_{s_j}(h,h_j^*)|+ \\ & |\epsilon_{s_j}(h,h_j^*)-\epsilon_t(h,h_j^*)|+ \\ & |\epsilon_t(h,h_j^*)-\epsilon_t(h)|.\end{aligned}$$

In this last equality, we may use again the triangle inequality to get the following bounds,

$$\begin{aligned}|\epsilon_{s_j}(h)-\epsilon_{s_j}(h,h_j^*)| &= |\epsilon_{s_j}(h,f_{s_j})-\epsilon_{s_j}(h,h_j^*)|, \\ &\leq |\epsilon_{s_j}(h_j^*)| = \epsilon_{s_j}(h_j^*),\end{aligned}$$

the same reasoning can be applied to the difference $|\epsilon_t(h,h_j)-\epsilon_t(h)|$. Plugging back these results into Equation 5,

$$\begin{aligned}|\epsilon_\alpha(h)-\epsilon_t(h)| \leq \sum_{j=1}^{N}&\alpha_j(\epsilon_{s_j}(h_j^*)+\epsilon_t(h_j^*) \\ &+|\epsilon_j(h,h_j^*)-\epsilon_t(h,h_j^*)|).\end{aligned}$$

In this last equation, one may notice that $\lambda_j = \epsilon_{s_j}(h_j^*)+\epsilon_t(h_j^*)$. Moreover, using Lemma 1 one has $|\epsilon_{s_j}(h,h_j^*)-\epsilon_t(h,h_j^*)| \leq W_1(\mu_{s_j},\mu_t)$, resulting in,

$$|\epsilon_\alpha(h)-\epsilon_t(h)| \leq \sum_{j=1}^{N}\alpha_j(\lambda_j+W_1(\mu_{s_j},\mu_t)),$$

$$\epsilon_t(h) \leq \epsilon_\alpha(h)+\sum_{j=1}^{N}\alpha_j(\lambda_j+W_1(\mu_{s_j},\mu_t)). \tag{6}$$

The bound we want to prove is achieved by bounding different terms in this equation for $h=\hat{h}$. First, we begin by noticing that by using the triangle inequality multiple times, one has,

$$W_1(\mu_{s_j},\mu_t) \leq W_1(\mu_{s_j},\hat{\mu}_{s_j})+W_1(\hat{\mu}_{s_j},\hat{\mu}_t)+W_1(\mu_t,\hat{\mu}_t),$$

now, we may bound each term $W_1(\mu,\hat{\mu})$ Lemma 2, especially through Equation 1,

$$\begin{aligned}W_1(\mu_{s_j},\hat{\mu}_{s_j}) &\leq \sqrt{\frac{2\log(\frac{1}{\delta})}{\xi'}}\sqrt{\frac{1}{n_{s_j}}}, \\ W_1(\mu_t,\hat{\mu}_t) &\leq \sqrt{\frac{2\log(\frac{1}{\delta})}{\xi'}}\sqrt{\frac{1}{n_t}},\end{aligned}$$

which ultimately leads to,

$$\begin{aligned}W_1(\mu_{s_j},\mu_t) \leq{}& W_1(\hat{\mu}_{s_j},\hat{\mu}_t)+ \\ & \sqrt{\frac{2\log(\frac{1}{\delta})}{\xi'}}\left(\sqrt{\frac{1}{n_{s_j}}}+\sqrt{\frac{1}{n_t}}\right).\end{aligned} \tag{7}$$

The second term on the right-hand-side of Equation 7 is exactly $c_2$, for which we get by bounding the Wasserstein distance in Equation 6. Now, following the proof presented for Theorem 4 of [1], we perform four changes on the term $\epsilon_\alpha(\hat{h}_\alpha)$, as follows,

$$\epsilon_\alpha(\hat{h}_\alpha) \overset{(1)}{\to} \hat{\epsilon}_\alpha(\hat{h}_\alpha) \overset{(2)}{\to} \hat{\epsilon}_\alpha(h_t^*) \overset{(3)}{\to} \epsilon_\alpha(h_t^*) \overset{(4)}{\to} \epsilon_t(h_t^*)$$

We justify each one of these steps bellow,

(1) This step is justified by Lemma 3, for which we have,

$$\epsilon_\alpha(\hat{h}_\alpha) \leq \hat{\epsilon}_\alpha(\hat{h}_\alpha) + \sqrt{\frac{2K\sum_{j=1}^{N}\frac{\alpha_j^2}{\beta_j}\log(\frac{2}{\delta})}{n}} + 2\sqrt{\frac{K}{n}}\sum_{j=1}^{N}\frac{\alpha_j}{\beta_j n\sqrt{\beta_j}}.$$

(2) This step is justified as $\hat{h}_\alpha$ is the minimizer of $\hat{\epsilon}_\alpha$, so,

$$\hat{\epsilon}_\alpha(\hat{h}_\alpha) \leq \hat{\epsilon}_\alpha(h_t^*)$$

(3) This step is justified by applying again Lemma 3,

$$\hat{\epsilon}_\alpha(h_t^*) \leq \epsilon_\alpha(h_t^*) + \sqrt{\frac{2K\sum_{j=1}^{N}\frac{\alpha_j^2}{\beta_j}\log(\frac{2}{\delta})}{n}} + 2\sqrt{\frac{K}{n}}\sum_{j=1}^{N}\frac{\alpha_j}{\beta_j n\sqrt{\beta_j}}.$$

(4) This step is justified by applying Equation 6 again, for $h = h_t^*$. Here we present it with $W_1(\mu_{s_j}, \mu_t)$ already substituted by $W_1(\hat{\mu}_{s_j}, \hat{\mu}_t)$, as we already justified this step.

$$\epsilon_\alpha(h_t^*) \leq \epsilon_t(h_t^*) + \sum_{j=1}^{N}\alpha_j(\lambda_j + W_1(\hat{\mu}_{s_j}, \hat{\mu}_t) + c_2)$$

Starting from Equation 6, and following each of the above mentioned steps proves the theorem.

□

## B. Experiments and Discussion

In this section we detail the experimental setup for the Wassertein Barycenter Transport (WBT) algorithm. To decide on the set of hyper-parameters, a 5-fold cross-validation procedure was used. Especially, WBT involves four hyper-parameters: (1) the barycenter regularization penalty $\epsilon_b$, (2) the transport (barycenter to target) regularization parameter $\epsilon$, (3) the class-based regularizer penalty $\eta$, and (4) the maximum number of iterations for the barycenter's convergence. Below we specify the range for each parameter,

- $\epsilon_b \in \{10^{-2}, 10^{-3}\}$,
- $\epsilon \in \{10^{-2}, 10^{-3}\}$,
- $\eta \in \{10^{-2}, 10^{-3}, 0\}$,
- $N_{max} \in \{1, 5, 10, 100\}$.

Furthermore, we used as stopping criteria for the WBT algorithm the squared norm of the barycenter displacement. More specifically, let $\mathbf{X}_b^{(k-1)}$ be the barycenter's support at iteration $k-1$, and $\mathbf{X}_b^{(k)}$ its support at the present iteration. The WBT algorithm stops if,

$$||\mathbf{X}_b^{(k)} - \mathbf{X}_b^{(k-1)}||_2^2 \leq \delta,$$

or if $k$ attains $N_{max}$. In our experiments we considered $\delta = 1$, as it is a very low value compared to the norm in the last equation, since the number of variables is high.

Before commenting on the importance of each parameter, we remark that for practical purposes of numerical accuracy, we normalize the cost matrix by its maximum value, that is, we use $\tilde{C} = C/max_{i,j}\, C_{ij}$. Such normalization was used for instance in [3], and justifies using such small values for the penalties.

Moreover, the barycenter penalty $\epsilon_b$ is the most critic parameter, as it determines the quality of the intermediate built domain. Especially, if it is too high the intermediate domain is collapsed on the average of the various domains, while if its too low, the Sinkhorn algorithm suffers from numerical accuracy issues. In general, we found that for Music-Speech Discrimination (MSD), Music Genre Recognition (MGR), and Face Recognition, $\epsilon_b = 10^{-3}$ is the best value, while for Object Recognition $\epsilon_b = 10^{-2}$ is the better choice ($\epsilon_b = 10^{-3}$ for this latter task yields unstable results).

Additionally, we found out that usually a few iterations are sufficient for building a good intermediate domain, as the best results for the Caltech-Office dataset were achieved for $N_{max} = 1$. Surprisingly, this does not hold for the DSLR domain. This suggest that an improvement may be made in the stopping criteria of WBT. Table 1 shows a summary of the best found hyper-parameters.

## References

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

| Task | Domain | $\epsilon_b$ | $\epsilon$ | $\eta$ | $N_{max}$ | Accuracy |
|---|---|---|---|---|---|---|
| Face Recognition | PIE05 | $10^{-3}$ | $10^{-3}$ | $10^{-2}$ | 1 | $51.10 \pm 2.02$ |
|  | PIE07 | $10^{-3}$ | $10^{-3}$ | $10^{-3}$ | 5 | $80.66 \pm 2.65$ |
|  | PIE09 | $10^{-3}$ | $10^{-3}$ | $10^{-3}$ | 5 | $79.58 \pm 3.47$ |
|  | PIE29 | $10^{-3}$ | $10^{-3}$ | 0.00 | 5 | $66.74 \pm 4.78$ |
| Object Recognition | Amazon | $10^{-2}$ | $10^{-2}$ | $10^{-3}$ | 1 | $92.74 \pm 0.45$ |
|  | dslr | $10^{-2}$ | $10^{-2}$ | $10^{-2}$ | 100 | $95.87 \pm 1.43$ |
|  | webcam | $10^{-2}$ | $10^{-2}$ | $10^{-2}$ | 1 | $96.57 \pm 1.71$ |
|  | Caltech | $10^{-2}$ | $10^{-2}$ | $10^{-3}$ | 1 | $85.01 \pm 0.84$ |
| Music Genre Recognition | Buccaneer2 | $10^{-3}$ | $10^{-3}$ | $10^{-2}$ | 100 | $70.60 \pm 1.27$ |
|  | Destroyerengine | $10^{-3}$ | $10^{-3}$ | $10^{-2}$ | 100 | $83.05 \pm 0.97$ |
|  | F16 | $10^{-3}$ | $10^{-3}$ | $10^{-3}$ | 100 | $84.40 \pm 1.71$ |
|  | Factory2 | $10^{-3}$ | $10^{-3}$ | $10^{-3}$ | 100 | $90.17 \pm 0.46$ |
| Music-Speech Discrimination | Buccaneer2 | $10^{-3}$ | $10^{-3}$ | $10^{-3}$ | 100 | $96.27 \pm 1.60$ |
|  | Destroyerengine | $10^{-3}$ | $10^{-3}$ | $10^{-2}$ | 100 | $92.98 \pm 1.38$ |
|  | F16 | $10^{-3}$ | $10^{-3}$ | $10^{-3}$ | 100 | $94.93 \pm 0.68$ |
|  | Factory2 | $10^{-3}$ | $10^{-3}$ | 0.00 | 100 | $96.87 \pm 0.94$ |

Table 1. Cross-validation summary with best hyper-parameters values found, alongside their respective performance.

[2] François Bolley, Arnaud Guillin, and Cédric Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137(3-4):541–593, 2007.

[3] Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. In *Advances in Neural Information Processing Systems*, pages 4197–4205, 2016.

[4] Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical analysis of domain adaptation with optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 737–753. Springer, 2017.