Supplementary material for the paper: Optimal Transport for Multi-source Domain Adaptation under Target Shift

In this Supplementary material we present proofs of the main theoretical results given in the main paper and provide additional empirical evaluations of the proposed method that were not included into the main paper due to the lack of space.

1 Proofs of theoretical results

1.1 Proof of Proposition 2

Proposition 2. *Let $\mathcal{H}$ denote the hypothesis space of predictors $h : \Omega \to \{0, 1\}$ and l be a convex loss function. Let $disc_l(P_S, P_T) = \max_{h,h' \in \mathcal{H}} |\epsilon_S(l(h, h')) - \epsilon_T(l(h, h'))|$ be the discrepancy distance [1] between two probability distributions P_S and P_T. Then, for any fixed $\boldsymbol{\alpha}$ the following holds for any $h \in \mathcal{H}$:*

$$\epsilon_T(h) \leq \epsilon_S^{\boldsymbol{\alpha}}(h) + |\pi_T - \sum_{j=1}^{N} \alpha_j \pi_S^j| disc_l(P_0, P_1) + \lambda,$$

where $\lambda = \min_{h \in \mathcal{H}} \epsilon_S^{\boldsymbol{\alpha}}(h) + \epsilon_T(h)$ represents the joint error between the combined source error and the target one.

Proof.

$$\begin{aligned}
\epsilon_T(h) &\leq \epsilon_T(h^*, f_T) + \epsilon_T(h, h^*) \\
&\leq \epsilon_T(h^*, f_T) + \epsilon_S^{\boldsymbol{\alpha}}(h, h^*) + |\epsilon_T(h, h^*) - \epsilon_S^{\boldsymbol{\alpha}}(h, h^*)| \qquad (1) \\
&\leq \epsilon_T(h^*, f_T) + \epsilon_S^{\boldsymbol{\alpha}}(h, h^*) + \max_{h,h' \in \mathcal{H}} |\epsilon_T(h, h') - \epsilon_S^{\boldsymbol{\alpha}}(h, h')| \\
&= \epsilon_T(h^*, f_T) + \epsilon_S^{\boldsymbol{\alpha}}(h, h^*) + \mathrm{disc}_l(P_S^{\boldsymbol{\alpha}}, P_T) \qquad (2) \\
&\leq \epsilon_S^{\boldsymbol{\alpha}}(h, f_S^{\boldsymbol{\alpha}}) + \epsilon_S^{\boldsymbol{\alpha}}(h^*, f_S^{\boldsymbol{\alpha}}) + \epsilon_T(h^*, f_T) + \mathrm{disc}_l(P_S^{\boldsymbol{\alpha}}, P_T) \\
&= \epsilon_S^{\boldsymbol{\alpha}}(h, f_S^{\boldsymbol{\alpha}}) + \mathrm{disc}_l(P_S^{\boldsymbol{\alpha}}, P_T) + \lambda. \qquad (3)
\end{aligned}$$

Here lines (1) and (2) are obtained due to the validity of the triangle inequality for the classification error function [2]. Regarding the disc_l discrepancy term, we obtain:

$$\begin{aligned}
\mathrm{disc}_l(P_S^{\boldsymbol{\alpha}}, P_T) &= \max_{h,h' \in \mathcal{H}} |\epsilon_T(h, h') - \epsilon_S^{\boldsymbol{\alpha}}(h, h')| \\
&= \max_{h,h' \in \mathcal{H}} |\mathbb{E}_{P_T}(h, h') - \mathbb{E}_{P_S^{\boldsymbol{\alpha}}}(h, h')| \\
&= \max_{h,h' \in \mathcal{H}} \left| \left(\sum_{j=1}^{N} \alpha_j (1 - \pi_S^j) - (1 - \pi_T) \right) \mathbb{E}_{P_0}(h, h') + \left(\sum_{j=1}^{N} \alpha_j \pi_S^j - \pi_T \right) \mathbb{E}_{P_1}(h, h') \right| \\
&= \max_{h,h' \in \mathcal{H}} \left| \left(\pi_T - \sum_{j=1}^{N} \alpha_j \pi_S^j \right) (\mathbb{E}_{P_0}(h, h') - \mathbb{E}_{P_1}(h, h')) \right| \\
&= |\pi_T - \sum_{j=1}^{N} \alpha_j \pi_S^j| \mathrm{disc}_l(P_0, P_1).
\end{aligned}$$

The final result is obtained by combining the last expression with (3) from the proof. □

We further note that this result can be made data dependent for predefined families of loss functions l such as 0-1 loss and ℓ_q loss often used in classification and regression, respectively. To this end, one may apply [1, Corollary 6 and 7] in order to replace the true distributions P_0 and P_1 by their empirical counterparts.

1.2 Proof of Proposition 3

Proposition 3. *Assume that* $\forall i, \nexists \alpha \in \{\Delta_C | \alpha_i = 0, P_i = \sum_j \alpha_j P_j\}$. *Then, for any distribution* P_T, *the unique solution* π^* *minimizing*

$$\pi^\star = \underset{\pi \in \Delta_C}{\arg\min}\ W(P_S^\pi, P_T). \tag{4}$$

is given by π^T.

Proof. We first note that for any two probability distributions P_1 and P_2, $W(P_1, P_2) \geq 0$ and $W(P_1, P_2) = 0$ if and only if $P_1 = P_2$ as Wasserstein distance is a valid metric on the space of probability measures. In this case, when $\pi^* = \pi_T$, $W(P_S^{\pi^*}, P_T) = 0$ and thus $\pi^* = \pi_T$ is a feasible solution of the optimization problem given in (4). On the other hand, for a given solution $\tilde{\pi}$ such that $\exists i \in \{1, \ldots, C\} : \tilde{\pi}_i^* \neq \pi_i^T$, we have due to the assumption made in the statement of the proposition that $\nexists \alpha \in \Delta_C$ and $\nexists j \in \{1, \ldots, i-1, i+1, \ldots, C\} : P_j = \sum_{k \in \mathcal{A}: i \in \mathcal{A}} \alpha_k P_k$ and thus $W(P_S^{\tilde{\pi}^*}, P_T) > 0$. This last condition roughly means that none of the class distributions for classes $\{1, \ldots, i-1, i+1, \ldots, C\}$ can be expressed as a weighted sum involving class distribution P_i. Hence, $\pi^\star = \pi^T$ is the unique solution of the optimization problem (4). □

1.3 Proof of Proposition 4

For the sake of completeness, we recall that the considered optimization problem has the following form:

$$\underset{\mathbf{h}}{\arg\min} \quad \sum_{k=1}^{K} \lambda_k W_{\epsilon, C^{(k)}} \left((\mathbf{D}_2^{(k)} \mathbf{h})^T \delta_{\mathbf{X}^{(k)}}, \mu \right), \tag{5}$$

where the regularized Wasserstein distances can be expressed as

$$W_{\epsilon, C^{(k)}}(\mu^{(k)}, \mu) \overset{\text{def}}{=} \min_{\gamma^{(k)} \in \Pi(\mu^{(k)}, \mu)} \mathrm{KL}(\gamma^{(k)}, \zeta^{(k)}),$$

provided that $\zeta^{(k)} = \exp\left(-\frac{C^{(k)}}{\epsilon}\right)$ and with λ_k being convex coefficients ($\sum_k \lambda_k = 1$) accounting for the relative importance of each domain.

In order to solve it for K constraints related to the unknown proportions $\mathbf{h}$, we formulate the problem as a Bregman projection with prescribed row sum ($\forall k \ \ \mathbf{D}_1^{(k)} \gamma^{(k)} \mathbf{1}_n = \ \mathbf{h}$), i.e.,

$$\mathbf{h}^\star = \quad \underset{\mathbf{h}, \Gamma}{\arg\min} \sum_{k=1}^{K} \lambda_k \, \mathrm{KL}(\gamma^{(k)}, \bar{\gamma}^{(k)}) \ \ \text{s.t.} \quad \forall k \ \ \mathbf{D}_1^{(k)} \gamma^{(k)} \mathbf{1}_n = \mathbf{h}. \tag{6}$$

This problem admits a closed form solution that we establish in the following result.

Proposition 4. *The solution of the projection defined in Equation 6 is given by:*

$$\forall k, \gamma^{(k)} = \mathrm{diag}\left(\frac{\mathbf{D}_2^{(k)} \mathbf{h}}{\bar{\gamma}^{(k)} \mathbf{1}_n} \right) \bar{\gamma}^{(k)}, \mathbf{h} = \Pi_{k=1}^K (\mathbf{D}_1^{(k)} (\bar{\gamma}^{(k)} \mathbf{1}_n))^{\lambda_k}.$$

Proof. We follow a similar line of reasoning as [3, Proposition (2)]. We write the following optimization problem with a collection of Lagrange multipliers denoted as $(\mathbf{u}_k \in \mathbb{R}^C)_{k=1,\ldots,K}$ in vector form.

$$\mathcal{L}(\Gamma, (\mathbf{u}_k)_{k=1,\ldots,K}, \mathbf{h}) = \sum_{k=1}^{K} \lambda_k < \gamma^{(k)}, \log \frac{\gamma^{(k)}}{\bar{\gamma}^{(k)}} - \mathbf{1} > + \sum_{k=1}^{K} \mathbf{u}_k^T (\mathbf{D}_1^{(k)} \gamma^{(k)} \mathbf{1}_n - \mathbf{h}). \tag{7}$$

We now compute the derivative w.r.t. $\gamma^{(k)}$, $\mathbf{u}_k$ and $\mathbf{h}$:

$$\frac{\partial \mathcal{L}(.)}{\partial \gamma^{(k)}} = \lambda_k \log \frac{\gamma^{(k)}}{\bar{\gamma}^{(k)}} + \mathbf{D}_1^{(k)} \mathbf{u}_k \mathbf{1}_n^T, \ \ \forall k \tag{8}$$

$$\frac{\partial \mathcal{L}(.)}{\partial \mathbf{u}_k} = \mathbf{D}_1^{(k)} \gamma^{(k)} \mathbf{1}_n - \mathbf{h}, \ \ \forall k \tag{9}$$

$$\frac{\partial \mathcal{L}(.)}{\partial \mathbf{h}} = \sum_{k=1}^{K} \mathbf{u}_k. \tag{10}$$

Setting the first equation to zero leads to

$$\lambda_k \log \tfrac{\gamma^{(k)}}{\bar{\gamma}^{(k)}} + \mathbf{D}_1^{(k)} \mathbf{u}_k \mathbf{1}_n^T = 0, \tag{11}$$

$$\log \tfrac{\gamma^{(k)}}{\bar{\gamma}^{(k)}} = -\tfrac{\mathbf{D}_1^{(k)} \mathbf{u}_k \mathbf{1}_n^T}{\lambda_k}, \tag{12}$$

$$\gamma^{(k)} = \exp\left(-\tfrac{\mathbf{D}_1^{(k)} \mathbf{u}_k \mathbf{1}_n^T}{\lambda_k}\right) \odot \bar{\gamma}^{(k)}, \tag{13}$$

with $\odot$ the Hadamard product. Finally, by multiplying the two terms by $\mathbf{1}_n$, we get:

$$\gamma^{(k)} \mathbf{1}_n = \left(\exp\left(-\frac{\mathbf{u}_k \mathbf{1}_n^T}{\lambda_k}\right) \odot \bar{\gamma}^{(k)}\right) \mathbf{1}_n. \tag{14}$$

Using the optimality condition of equation two of the previous system, we know that $\mathbf{D}_1^{(k)} \gamma^{(k)} \mathbf{1}_n = \mathbf{h}$ or $\gamma^{(k)} \mathbf{1}_n = \mathbf{D}_2^{(k)} \mathbf{h}$ and subsequently

$$\exp\left(-\frac{\mathbf{u}_k \mathbf{1}_n^T}{\lambda_k}\right) = \operatorname{diag}\left(\frac{\mathbf{D}_2^{(k)} \mathbf{h}}{\bar{\gamma}^{(k)} \mathbf{1}_n}\right). \tag{15}$$

Plugging this expression into the first equation, we obtain:

$$\gamma_k = \operatorname{diag}\left(\frac{\mathbf{D}_2^{(k)} \mathbf{h}}{\bar{\gamma}^{(k)} \mathbf{1}_n}\right) \bar{\gamma}^{(k)}, \tag{16}$$

that is the first element of the solution of the projection. We sum over k the first optimality equation, and we get:

$$\sum_{k=1}^{K} \lambda_k \log \frac{\gamma^{(k)}}{\bar{\gamma}^{(k)}} + \sum_k \mathbf{D}_1^{(k)} \mathbf{u}_k \mathbf{1}_n^T = 0. \tag{17}$$

Setting the third question to zero leads to $\sum_{k=1}^{K} \mathbf{u}_k = 0$. Because of the specific structure of $\mathbf{D}_1^{(k)}$, we also have $\sum_{k=1}^{K} \mathbf{D}_1^{(k)} \mathbf{u}_k = 0$. Therefore, we obtain:

$$\sum_{k=1}^{K} \lambda_k \log \frac{\gamma^{(k)}}{\bar{\gamma}^{(k)}} = 0, \tag{18}$$

$$\tag{19}$$

or equivalently

$$\prod_{k=1}^{K} \left(\frac{\gamma^{(k)}}{\bar{\gamma}^{(k)}}\right)^{\lambda_k} = \mathbf{1}_C. \tag{20}$$

From Equation 16 we get

$$\prod_{k=1}^{K} \left(\frac{\mathbf{D}_2^{(k)} \mathbf{h}}{\bar{\gamma}^{(k)} \mathbf{1}_n}\right)^{\lambda_k} = \mathbf{1}_n \tag{21}$$

(22)

which is equivalent to

$$\prod_{k=1}^{K} \mathbf{h}^{\lambda_k} = \prod_{k=1}^{K} (\mathbf{D}_1^{(k)} \bar{\gamma}^{(k)})^{\lambda_k} \tag{23}$$

and finally, since $\sum_{k=1}^{K} \lambda_k = 1$, we get

$$\mathbf{h} = \prod_{k=1}^{K} (\mathbf{D}_1^{(k)} \bar{\gamma}^{(k)})^{\lambda_k} \tag{24}$$

which concludes the proof. □

3 Experimental results

In this section, we provide the details on the generative process of the synthetic data used in the main paper and present results of several other experiments that we could not include into the main paper due to lack of space.

3.1 Data set generation

In the main paper, we considered the multi-source scenario for which we generated a binary classification problem with the instances of each class were drawn from the Gaussian distributions $\mathcal{N}\left(\begin{pmatrix} -1 \\ 0 \end{pmatrix}, 0.3\boldsymbol{I}\right)$ and $\mathcal{N}\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}, 0.3\boldsymbol{I}\right)$, respectively.

3.2 Running time comparison

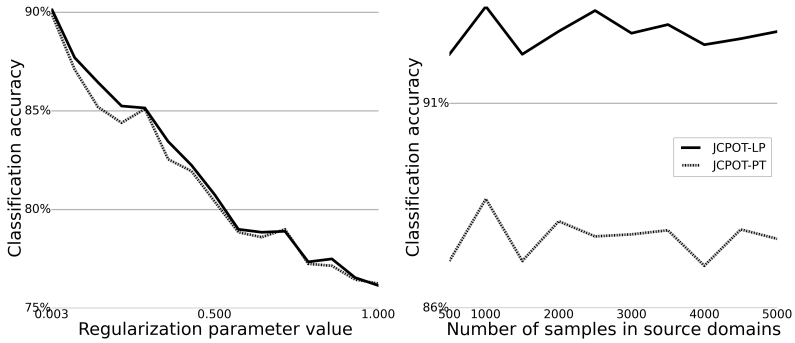
In Table 1, we give the running times of all the algorithms considered in the empirical evaluation of the main paper for the simulated data. From the results, we can see that `betaEM` is the less computationally demanding method. `MDAC Causal`, `JCPOT` method and `OTDA` share performances with the same order of magnitude. We note also that `betaKMM` is the most computationally heavy method.

	Number of source domains						
	2	5	8	11	14	17	20
`betaEM`	0.179	0.174	0.241	0.314	0.394	0.458	0.524
`betaKMM`	16.057	193.331	119.859	117.982	190.623	172.903	209.53
`MDAC Causal`	1.4130	1.5466	1.8402	2.1484	2.5962	3.2463	3.7972
`OTDA`	0.515	1.04	1.622	2.276	2.978	3.824	4.488
`JCPOT`	0.31	1.079	1.766	2.285	3.296	4.38	4.722

Table 1: Running times (in seconds) of all baselines considered in the main paper.

3.3 Sensitivity to hyper-parameters

Figures 1a and 1b illustrate the classification results obtained by `JCPOT` when varying the regularization parameter λ and the overall size of source samples in source domains, respectively in a setting with 4 source and 1 target domains. In the latter scenario, we vary the sample size by increasing it by 500 for the source domains (125 instances per domain) and by 200 for the target domain. From these figures, we observe that higher values of λ can lead to a decrease in the performance of our algorithm, while the source domains' sample size does not appear to have a high influence on the results.



(a) (b)

Figure 1: Performances of JCPOT obtained when varying (a) the regularization parameter λ; (b) the size of source and target domains samples.

References

[1] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT*, 2009.

[2] Koby Crammer, Michael J. Kearns, and Jennifer Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9:1757–1774, 2008.

[3] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.