MOT: Masked Optimal Transport for Partial Domain Adaptation (Supplementary Material)

You-Wei Luo Chuan-Xian Ren*
School of Mathematics, Sun Yat-Sen University, China
luoyw28@mail2.sysu.edu.cn, rchuanx@mail.sysu.edu.cn

Abstract

This supplementary material contains the proofs of theoretical results, implementation details for numerical experiments and illustrations of experiment datasets.

S.1. Proof of Theorem 1

Theorem 1 (Proxy) *Assume* $\mathrm{supp}(q_Y) \subseteq \mathrm{supp}(p_Y)$*, the following identities hold.*
(a) ***Kantorovich OT****:*

$$\mathrm{OT}_{\mathrm{mask}}(P^{w^*}, Q, \tilde{\mathbf{C}}) \;=\; \mathrm{OT}^{q_Y}_{\mathrm{cond}}(P^{w^*}, Q, \mathbf{C}).$$

(b) ***Sinkhorn OT****:*

$$S^{\lambda}_{\mathrm{mask}}(P^{w^*}, Q, \tilde{\mathbf{C}}) + \lambda H(Q_Y) \;=\; S^{\lambda, q_Y}_{\mathrm{cond}}(P^{w^*}, Q, \mathbf{C}).$$

(c) ***UOT****: there exists non-negative* $\alpha(\cdot)$ *on* $\mathcal{Y}$ *such that* $\mathrm{supp}(\alpha) = \mathrm{supp}(q_Y)$*,* $\sum_{y\in\mathcal{Y}} \alpha(y) = 1$ *and*

$$S^{\lambda,\beta}_{\mathrm{mask}}(P^{w^*}, Q, \tilde{\mathbf{C}}) + C_0(\alpha, Q_Y) \;=\; S^{\lambda,\beta,\alpha}_{\mathrm{cond}}(P^{w^*}, Q, \mathbf{C}),$$

where C_0 *is a constant depending only on* α *and* Q_Y*.*

Proof *For convenience, we first introduce some notations for proof. For finite sample setting, denote* $|\mathcal{Y}| = k$ *as class number,* n_l/m_l *as the sample size of* l*-th source/target class. Since* $\mathrm{supp}(q_Y) \subseteq \mathrm{supp}(p_Y)$*, we denote* $\mathrm{supp}(p_Y) = \mathcal{Y} = \{1, 2, \ldots, k\}$*,* $\mathrm{supp}(p_Y) = \{1, 2, \ldots, k_0\}$ *and* $n_0 = \sum_{l=1}^{k_0} n_l$*, where* $k_0 \leq k$ *is the number of shared classes and* n_0 *the sample size of shared classes on source domain. Without loss of generality, we denote the data matrix with cluster data as* $\mathbf{X}^s = [\mathbf{X}^s_1, \mathbf{X}^s_2, \ldots, \mathbf{X}^s_k] \in \mathbb{R}^{d\times n}$ *and* $\mathbf{X}^t = [\mathbf{X}^t_1, \mathbf{X}^t_2, \ldots, \mathbf{X}^t_{k_0}] \in \mathbb{R}^{d\times m}$*, where* d *is data dimension,* $\mathbf{X}^s_l \in \mathbb{R}^{d\times n_l}$ *and* $\mathbf{X}^t_l \in \mathbb{R}^{d\times m_l}$ *are the data matrix of* l*-th source class and* l*-th target class, respectively. Generally, for a matrix* $\mathbf{A}$*, let the uppercase letters* $\mathbf{A}_{ij}$ *denote the blocks of* $\mathbf{A}$ *and lowercase letters* a_{ij} *the entries of* $\mathbf{A}$*. Note that for the reweighted source we have*

$$p^{w^*}_X = \sum_{y\in\mathcal{Y}} p^{w^*}_y p_{X|y} = \sum_{y\in\mathcal{Y}} q_y p_{X|y} = \sum_{l=1}^{k_0} q_{Y=l} p_{X|l}, \tag{S.1}$$

which implies the proportions of outlier classes are 0 in reweighted distribution. Then a submatrix $\tilde{\mathbf{C}}^{\mathrm{sub}} \in \mathbb{R}^{n_0\times m}$ *of* $\tilde{\mathbf{C}}$*, which considers the cost between samples of shared classes, is defined as the first* n_0 *rows of* $\tilde{\mathbf{C}}$*. Now we begin to prove the main results.*

(1) Kantorvich OT.

*Corresponding Author.

Recall the masked Kantorvich OT is formulated as

$$\mathrm{OT}_{\mathrm{mask}}(P^{w^*}, Q, \tilde{\mathbf{C}}) = \min_{\mathbf{\Gamma} \in \Pi(P_X^{w^*}, Q_X)} \left\langle \mathbf{\Gamma}, \tilde{\mathbf{C}} \right\rangle_F .$$

Let the source distribution of shared classes be $r_X^{w^*} \in \mathbb{R}^{n_0}$, which consists of the first n_0 elements of p^{w^*}. Since the values of outlier classes' samples are 0 in $p_X^{w^*} \in \mathbb{R}^n$ as Eq. (S.1), there transport plan for outlier classes will be 0, i.e., $\gamma_{ij} = 0$ if $i > n_0$. Then the original problem boils down to the transportation between shared classes:

$$\mathrm{OT}_{\mathrm{mask}}(P^{w^*}, Q, \tilde{\mathbf{C}}) = \min_{\mathbf{\Gamma} \in \Pi(P_X^{w^*}, Q_X)} \left\langle \mathbf{\Gamma}, \tilde{\mathbf{C}} \right\rangle_F = \min_{\mathbf{\Gamma}^{\mathrm{sub}} \in \Pi(R_X^{w^*}, Q_X)} \left\langle \mathbf{\Gamma}^{\mathrm{sub}}, \tilde{\mathbf{C}}^{\mathrm{sub}} \right\rangle_F .$$

On the other hands, note that the transport plan between inter-class sample pair will be 0 since, i.e., $\gamma_{ij}^{\mathrm{sub}} = 0$ if $y_i^s \neq y_j^t$, since otherwise the overall transport cost will be infinity and the problem will not be well-defined. It implies the plan $\mathbf{\Gamma}^{\mathrm{sub}}$ under masked cost admits a block diagonal structure, then we have

$$\begin{aligned}
\mathrm{OT}_{\mathrm{mask}}(P^{w^*}, Q, \tilde{\mathbf{C}}) &= \min_{\mathbf{\Gamma}^{\mathrm{sub}} \in \Pi(R_X^{w^*}, Q_X)} \left\langle \mathbf{\Gamma}^{\mathrm{sub}}, \tilde{\mathbf{C}}^{\mathrm{sub}} \right\rangle_F \\
&= \min_{\mathbf{\Gamma}^{\mathrm{sub}} \in \Pi(R_X^{w^*}, Q_X)} \left\langle \begin{bmatrix} \mathbf{\Gamma}_{11}^{\mathrm{sub}} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{\Gamma}_{k_0 k_0}^{\mathrm{sub}} \end{bmatrix}, \begin{bmatrix} \tilde{\mathbf{C}}_{11}^{\mathrm{sub}} & \cdots & \infty \\ \vdots & \ddots & \vdots \\ \infty & \cdots & \tilde{\mathbf{C}}_{k_0 k_0}^{\mathrm{sub}} \end{bmatrix} \right\rangle_F \\
&= \min_{\mathbf{\Gamma}^{\mathrm{sub}} \in \Pi(R_X^{w^*}, Q_X)} \sum_{l=1}^{k_0} \left\langle \mathbf{\Gamma}_{ll}^{\mathrm{sub}}, \tilde{\mathbf{C}}_{ll}^{\mathrm{sub}} \right\rangle_F \\
&= \sum_{l=1}^{k_0} \min_{\mathbf{\Gamma}_{ll}^{\mathrm{sub}} \in \Pi(q_{Y=l} R_{X|l}^{w^*}, q_{Y=l} Q_{X|l})} \left\langle \mathbf{\Gamma}_{ll}^{\mathrm{sub}}, \tilde{\mathbf{C}}_{ll}^{\mathrm{sub}} \right\rangle_F \\
&= \sum_{l=1}^{k_0} \min_{\frac{\mathbf{\Gamma}_{ll}^{\mathrm{sub}}}{q_{Y=l}} \in \Pi(R_{X|l}^{w^*}, Q_{X|l})} q_{Y=l} \left\langle \frac{\mathbf{\Gamma}_{ll}^{\mathrm{sub}}}{q_{Y=l}}, \tilde{\mathbf{C}}_{ll}^{\mathrm{sub}} \right\rangle_F \\
\left(\bar{\mathbf{\Gamma}}_{ll}^{\mathrm{sub}} \triangleq \frac{\mathbf{\Gamma}_{ll}^{\mathrm{sub}}}{q_{Y=l}} \right) \quad &= \sum_{l=1}^{k_0} q_{Y=l} \min_{\bar{\mathbf{\Gamma}}_{ll}^{\mathrm{sub}} \in \Pi(R_{X|l}^{w^*}, Q_{X|l})} \left\langle \bar{\mathbf{\Gamma}}_{ll}^{\mathrm{sub}}, \tilde{\mathbf{C}}_{ll}^{\mathrm{sub}} \right\rangle_F \\
&= \sum_{l=1}^{k_0} q_{Y=l} \mathrm{OT}(R_{X|l}^{w^*}, Q_{X|l}, \tilde{\mathbf{C}}_{ll}^{\mathrm{sub}}) \\
&= \sum_{l=1}^{k_0} q_{Y=l} \mathrm{OT}(P_{X|l}^{w^*}, Q_{X|l}, \mathbf{C}_{ll}^{\mathrm{sub}}) \qquad \text{(S.2)} \\
&= \mathrm{OT}_{\mathrm{cond}}^{q_Y}(P^{w^*}, Q, \mathbf{C}),
\end{aligned}$$

where $\bar{\mathbf{\Gamma}}_{ll}^{\mathrm{sub}}$ and $\tilde{\mathbf{C}}_{ll}^{\mathrm{sub}}$ are $n_l \times m_l$ blocks of l-th class, Eq. (S.2) holds since $R_{X|l}^{w^*} = P_{X|l}^{w^*}$ for shared classes and $\tilde{\mathbf{C}}_{ll}^{\mathrm{sub}} = \mathbf{C}_{ll}^{\mathrm{sub}}$ for intra-class sample pairs.

(2) Sinkhorn OT.

Recall the masked Sinkhorn OT is formulated as

$$S_{\mathrm{mask}}^{\lambda}(P^{w^*}, Q, \tilde{\mathbf{C}}) = \min_{\mathbf{\Gamma} \in \Pi(P_X^{w^*}, Q_X)} \left\langle \mathbf{\Gamma}, \tilde{\mathbf{C}} \right\rangle_F + \lambda \left\langle \mathbf{\Gamma}, \ln \mathbf{\Gamma} \right\rangle_F .$$

Similarly, we have

$$
\begin{aligned}
& S^{\lambda}_{\text{mask}}(P^{w^*}, Q, \tilde{\mathbf{C}}) \\
=& \min_{\mathbf{\Gamma}\in\Pi(P_X^{w^*}, Q_X)} \left\langle \mathbf{\Gamma}, \tilde{\mathbf{C}} \right\rangle_F + \lambda \left\langle \mathbf{\Gamma}, \ln \mathbf{\Gamma} \right\rangle_F \\
=& \min_{\mathbf{\Gamma}^{\text{sub}}\in\Pi(R_X^{w^*}, Q_X)} \left\langle \mathbf{\Gamma}^{\text{sub}}, \tilde{\mathbf{C}}^{\text{sub}} \right\rangle_F + \lambda \left\langle \mathbf{\Gamma}^{\text{sub}}, \ln \mathbf{\Gamma}^{\text{sub}} \right\rangle_F \\
=& \min_{\mathbf{\Gamma}^{\text{sub}}\in\Pi(R_X^{w^*}, Q_X)} \left\langle \begin{bmatrix} \mathbf{\Gamma}^{\text{sub}}_{11} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{\Gamma}^{\text{sub}}_{k_0 k_0} \end{bmatrix}, \begin{bmatrix} \tilde{\mathbf{C}}^{\text{sub}}_{11} & \cdots & \infty \\ \vdots & \ddots & \vdots \\ \infty & \cdots & \tilde{\mathbf{C}}^{\text{sub}}_{k_0 k_0} \end{bmatrix} + \lambda \ln \begin{bmatrix} \mathbf{\Gamma}^{\text{sub}}_{11} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{\Gamma}^{\text{sub}}_{k_0 k_0} \end{bmatrix} \right\rangle_F \\
=& \min_{\mathbf{\Gamma}^{\text{sub}}\in\Pi(R_X^{w^*}, Q_X)} \sum_{l=1}^{k_0} \left\langle \mathbf{\Gamma}^{\text{sub}}_{ll}, \tilde{\mathbf{C}}^{\text{sub}}_{ll} \right\rangle_F + \lambda \left\langle \mathbf{\Gamma}^{\text{sub}}_{ll}, \ln \mathbf{\Gamma}^{\text{sub}}_{ll} \right\rangle_F \\
=& \sum_{l=1}^{k_0} \min_{\mathbf{\Gamma}^{\text{sub}}_{ll}\in\Pi(q_{Y=l} R^{w^*}_{X|l}, q_{Y=l} Q_{X|l})} \left\langle \mathbf{\Gamma}^{\text{sub}}_{ll}, \tilde{\mathbf{C}}^{\text{sub}}_{ll} \right\rangle_F + \lambda \left\langle \mathbf{\Gamma}^{\text{sub}}_{ll}, \ln \mathbf{\Gamma}^{\text{sub}}_{ll} \right\rangle_F \\
=& \sum_{l=1}^{k_0} \min_{\frac{\mathbf{\Gamma}^{\text{sub}}_{ll}}{q_{Y=l}}\in\Pi(R^{w^*}_{X|l}, Q_{X|l})} q_{Y=l} \left[\left\langle \frac{\mathbf{\Gamma}^{\text{sub}}_{ll}}{q_{Y=l}}, \tilde{\mathbf{C}}^{\text{sub}}_{ll} \right\rangle_F + \lambda \left\langle \frac{\mathbf{\Gamma}^{\text{sub}}_{ll}}{q_{Y=l}}, \ln \mathbf{\Gamma}^{\text{sub}}_{ll} \right\rangle_F \right] \\
=& \sum_{l=1}^{k_0} \min_{\frac{\mathbf{\Gamma}^{\text{sub}}_{ll}}{q_{Y=l}}\in\Pi(R^{w^*}_{X|l}, Q_{X|l})} q_{Y=l} \left[\left\langle \frac{\mathbf{\Gamma}^{\text{sub}}_{ll}}{q_{Y=l}}, \tilde{\mathbf{C}}^{\text{sub}}_{ll} \right\rangle_F + \lambda \left\langle \frac{\mathbf{\Gamma}^{\text{sub}}_{ll}}{q_{Y=l}}, \ln \frac{\mathbf{\Gamma}^{\text{sub}}_{ll}}{q_{Y=l}} \right\rangle_F + \lambda \left\langle \frac{\mathbf{\Gamma}^{\text{sub}}_{ll}}{q_{Y=l}}, (\ln q_{Y=l}) \mathbf{1}_{n_l \times m_l} \right\rangle_F \right] \\
=& \sum_{l=1}^{k_0} q_{Y=l} \left[\min_{\bar{\mathbf{\Gamma}}^{\text{sub}}_{ll}\in\Pi(R^{w^*}_{X|l}, Q_{X|l})} \left\langle \bar{\mathbf{\Gamma}}^{\text{sub}}_{ll}, \tilde{\mathbf{C}}^{\text{sub}}_{ll} \right\rangle_F + \lambda \left\langle \bar{\mathbf{\Gamma}}^{\text{sub}}_{ll}, \ln \bar{\mathbf{\Gamma}}^{\text{sub}}_{ll} \right\rangle_F + \lambda \ln q_{Y=l} \left\langle \bar{\mathbf{\Gamma}}^{\text{sub}}_{ll}, \mathbf{1}_{n_l \times m_l} \right\rangle_F \right] \\
=& \sum_{l=1}^{k_0} q_{Y=l} \left[\min_{\bar{\mathbf{\Gamma}}^{\text{sub}}_{ll}\in\Pi(R^{w^*}_{X|l}, Q_{X|l})} \left\langle \bar{\mathbf{\Gamma}}^{\text{sub}}_{ll}, \tilde{\mathbf{C}}^{\text{sub}}_{ll} \right\rangle_F + \lambda \left\langle \bar{\mathbf{\Gamma}}^{\text{sub}}_{ll}, \ln \bar{\mathbf{\Gamma}}^{\text{sub}}_{ll} \right\rangle_F + \lambda \ln q_{Y=l} \right] \\
=& \sum_{l=1}^{k_0} q_{Y=l} S^{\lambda}(R^{w^*}_{X|l}, Q_{X|l}, \tilde{\mathbf{C}}^{\text{sub}}_{ll}) + \lambda \sum_{l=1}^{k_0} q_{Y=l} \ln q_{Y=l} \\
=& \sum_{l=1}^{k_0} q_{Y=l} S^{\lambda}(P^{w^*}_{X|l}, Q_{X|l}, \tilde{\mathbf{C}}^{\text{sub}}_{ll}) - \lambda H(Q_Y) \\
=& \, \mathrm{S}^{\lambda, q_Y}_{\text{cond}}(P^{w^*}, Q, \mathbf{C}) - \lambda H(Q_Y).
\end{aligned}
$$

Therefore, we have $\mathrm{S}^{\lambda, q_Y}_{\text{cond}}(P^{w^*}, Q, \mathbf{C}) = S^{\lambda}_{\text{mask}}(P^{w^*}, Q, \tilde{\mathbf{C}}) + \lambda H(Q_Y)$

(3) Unbalanced OT.

Recall the masked unbalanced OT is formulated as

$$
S^{\lambda,\beta}_{\text{mask}}(P^{w^*}, Q, \tilde{\mathbf{C}}) = \min_{\mathbf{\Gamma}\in\mathcal{M}_+(\mathbb{R}^{n\times m})} \langle \mathbf{\Gamma}, \tilde{\mathbf{C}} \rangle_F + \lambda \left\langle \mathbf{\Gamma}, \ln \mathbf{\Gamma} \right\rangle_F + \beta \left[D_\phi(\mathbf{\Gamma}_{P_X^{w^*}} \| P_X^{w^*}) + D_\phi(\mathbf{\Gamma}_{Q_X} \| Q_X) \right],
$$

where D_ϕ is KL divergence. The major difference between unbalanced OT and other OTs with marginal constraints is that the $\mathbf{\Gamma}$ is only required to be a distribution over $\mathbb{R}^{n\times m}$, i.e., $\mathbf{\Gamma} \in \mathcal{M}_+(\mathbb{R}^{n\times m})$ will satisfy that $\gamma_{ij} >= 0$ and $\sum_{ij} \gamma_{ij} = 1$. Since $\mathbf{\Gamma}$ is no longer a coupling of $(P^{w^*}, Q, \tilde{\mathbf{C}})$, it is necessary to consider whether the 0 transport plans for outlier classes still hold.

Note the KL penalty $D_\phi(\mathbf{\Gamma}_{P_X^{w^*}} \| P_X^{w^*})$ implies that $\mathbf{\Gamma}_{P_X^{w^*}}$ should be absolutely continuous with respect to $P_X^{w^*}$, since otherwise the penalty value will be infinity and the problem is not well-defined. Therefore, for the i-th source sample, if it belongs to outlier classes, the corresponding values in $P_X^{w^*}$ are 0 (i.e., $[P_X^{w^*}]_i = 0$), and the transport plan will also be 0 (i.e., $[\mathbf{\Gamma}_{P_X^{w^*}}]_i = \sum_j \gamma_{ij} = 0 \implies \gamma_{ij} = 0$). Therefore, the original problem can also be written as the transportation between shared classes, i.e.,

$$S^{\lambda,\beta}_{\text{mask}}(P^{w^*},Q,\tilde{\mathbf{C}}) = \min_{\mathbf{\Gamma}^{\text{sub}}\in\mathcal{M}_+(\mathbb{R}^{n_0\times m})} \langle \mathbf{\Gamma}^{\text{sub}}, \tilde{\mathbf{C}}^{\text{sub}}\rangle_F + \lambda \left\langle \mathbf{\Gamma}^{\text{sub}}, \ln \mathbf{\Gamma}^{\text{sub}}\right\rangle_F + \beta\left[D_\phi(\mathbf{\Gamma}^{\text{sub}}_{R^{w^*}_X}\|R^{w^*}_X) + D_\phi(\mathbf{\Gamma}^{\text{sub}}_{Q_X}\|Q_X)\right].$$

Let $\mathbf{\Gamma}^{\text{sub}^*}$ be the optimal solution for the objective above. Similarly, $\mathbf{\Gamma}^{\text{sub}^*}$ is also block-diagonal since the non-zero plan values for inter-class sample pairs $(\mathbf{x}^s_i,\mathbf{x}^t_j)$ will induce infinite transport cost with $\tilde{c}^{\text{sub}}_{ij}$. Then we consider the following coefficient

$$\alpha(l) = \sum_{ij}[\mathbf{\Gamma}^{\text{sub}^*}_{ll}]_{ij},$$

which represents the values assigned to the transportation between l-th source class and l-th target class. It is clear that $\alpha(\cdot)$ is non-negative on $\mathcal{Y}$ and satisfies that $\text{supp}(\alpha) = \text{supp}(q_Y)$. For simplicity, we denote the blocks of $\mathbf{\Gamma}^{\text{sub}}_{R^{w^*}_X}$ and $\mathbf{\Gamma}^{\text{sub}}_{Q_X}$ as

$$\mathbf{\Gamma}^{\text{sub}}_{R^{w^*}_X} = \begin{bmatrix}\mathbf{o}^s_1\\ \vdots \\ \mathbf{o}^s_{k_0}\end{bmatrix}\in\mathbb{R}^{n_0}, \qquad \mathbf{\Gamma}^{\text{sub}}_{Q_X} = \begin{bmatrix}\mathbf{o}^t_1\\ \vdots \\ \mathbf{o}^t_{k_0}\end{bmatrix}\in\mathbb{R}^{m}$$

where $\mathbf{o}^s_l\in\mathbb{R}^{n_l}$ and $\mathbf{o}^t_l\in\mathbb{R}^{m_l}$. Then we have

$$\begin{aligned}
&S^{\lambda,\beta}_{\text{mask}}(P^{w^*},Q,\tilde{\mathbf{C}})\\
&= \min_{\mathbf{\Gamma}^{\text{sub}}\in\mathcal{M}_+(\mathbb{R}^{n_0\times m})} \left\langle \mathbf{\Gamma}^{\text{sub}}, \tilde{\mathbf{C}}^{\text{sub}}\right\rangle_F + \lambda \left\langle \mathbf{\Gamma}^{\text{sub}}, \ln \mathbf{\Gamma}^{\text{sub}}\right\rangle_F + \beta\left[D_\phi(\mathbf{\Gamma}^{\text{sub}}_{R^{w^*}_X}\|R^{w^*}_X) + D_\phi(\mathbf{\Gamma}^{\text{sub}}_{Q_X}\|Q_X)\right]\\
&= \min_{\mathbf{\Gamma}^{\text{sub}}\in\mathcal{M}_+(\mathbb{R}^{n_0\times m})} \left\langle \mathbf{\Gamma}^{\text{sub}}, \tilde{\mathbf{C}}^{\text{sub}}\right\rangle_F + \lambda \left\langle \mathbf{\Gamma}^{\text{sub}}, \ln \mathbf{\Gamma}^{\text{sub}}\right\rangle_F + \beta\left[\left\langle \mathbf{\Gamma}^{\text{sub}}_{R^{w^*}_X}, \ln\frac{\mathbf{\Gamma}^{\text{sub}}_{R^{w^*}_X}}{R^{w^*}_X}\right\rangle_F + \left\langle \mathbf{\Gamma}^{\text{sub}}_{Q_X}, \ln\frac{\mathbf{\Gamma}^{\text{sub}}_{Q_X}}{Q_X}\right\rangle_F\right]\\
&= \sum_{l=1}^{k_0}\min_{\mathbf{\Gamma}^{\text{sub}}_{ll}\in\alpha(l)\mathcal{M}_+(\mathbb{R}^{n_l\times m_l})} \left\langle \mathbf{\Gamma}^{\text{sub}}_{ll}, \tilde{\mathbf{C}}^{\text{sub}}_{ll}\right\rangle_F + \lambda \left\langle \mathbf{\Gamma}^{\text{sub}}_{ll}, \ln \mathbf{\Gamma}^{\text{sub}}_{ll}\right\rangle_F + \beta\left[\left\langle \mathbf{o}^s_l, \ln\frac{\mathbf{o}^s_l}{r^{w^*}_{Y=l}R^{w^*}_{X|l}}\right\rangle_F + \left\langle \mathbf{o}^t_l, \ln\frac{\mathbf{o}^t_l}{q_{Y=l}Q_{X|l}}\right\rangle_F\right]
\end{aligned} \tag{S.3}$$

$$\begin{aligned}
&= \sum_{l=1}^{k_0}\min_{\frac{\mathbf{\Gamma}^{\text{sub}}_{ll}}{\alpha(l)}\in\mathcal{M}_+(\mathbb{R}^{n_l\times m_l})} \alpha(l)\Bigg[\left\langle \frac{\mathbf{\Gamma}^{\text{sub}}_{ll}}{\alpha(l)}, \tilde{\mathbf{C}}^{\text{sub}}_{ll}\right\rangle_F + \lambda\left\langle \frac{\mathbf{\Gamma}^{\text{sub}}_{ll}}{\alpha(l)}, \ln \mathbf{\Gamma}^{\text{sub}}_{ll}\right\rangle_F\\
&\qquad\qquad + \beta\left\langle \frac{\mathbf{o}^s_l}{\alpha(l)}, \ln\frac{\mathbf{o}^s_l}{q_{Y=l}R^{w^*}_{X|l}}\right\rangle_F + \beta\left\langle \frac{\mathbf{o}^t_l}{\alpha(l)}, \ln\frac{\mathbf{o}^t_l}{q_{Y=l}Q_{X|l}}\right\rangle_F\Bigg]\\
&= \sum_{l=1}^{k_0}\min_{\bar{\mathbf{\Gamma}}^{\text{sub}}_{ll}\in\mathcal{M}_+(\mathbb{R}^{n_l\times m_l})} \alpha(l)\Bigg[\left\langle \bar{\mathbf{\Gamma}}^{\text{sub}}_{ll}, \tilde{\mathbf{C}}^{\text{sub}}_{ll}\right\rangle_F + \lambda\left\langle \bar{\mathbf{\Gamma}}^{\text{sub}}_{ll}, \ln \bar{\mathbf{\Gamma}}^{\text{sub}}_{ll}\right\rangle_F + \lambda\ln\alpha(l)\\
&\qquad\qquad + \beta\left\langle \frac{\mathbf{o}^s_l}{\alpha(l)}, \ln\frac{\mathbf{o}^s_l}{q_{Y=l}R^{w^*}_{X|l}}\right\rangle_F + \beta\left\langle \frac{\mathbf{o}^t_l}{\alpha(l)}, \ln\frac{\mathbf{o}^t_l}{q_{Y=l}Q_{X|l}}\right\rangle_F\Bigg],
\end{aligned} \tag{S.4}$$

where Eq. (S.3) holds since $\mathbf{\Gamma}^{\text{sub}^*}$ is block diagonal, which implies the minimization problem can be divided into k_0 sub-problems and the mass assigned to l-th class is $\alpha(l)$. Note that for the KL terms, we have

$$\begin{aligned}
&\left\langle \frac{\mathbf{o}^s_l}{\alpha(l)}, \ln\frac{\mathbf{o}^s_l}{q_{Y=l}R^{w^*}_{X|l}}\right\rangle_F + \left\langle \frac{\mathbf{o}^t_l}{\alpha(l)}, \ln\frac{\mathbf{o}^t_l}{q_{Y=l}Q_{X|l}}\right\rangle_F\\
&= \left\langle \frac{\mathbf{o}^s_l}{\alpha(l)}, \ln\frac{\mathbf{o}^s_l}{\alpha(l)R^{w^*}_{X|l}}\right\rangle_F + \left\langle \frac{\mathbf{o}^s_l}{\alpha(l)}, \ln\frac{\alpha(l)\mathbf{1}_{n_l}}{q_{Y=l}}\right\rangle_F + \left\langle \frac{\mathbf{o}^t_l}{\alpha(l)}, \ln\frac{\mathbf{o}^t_l}{\alpha(l)Q_{X|l}}\right\rangle_F + \left\langle \frac{\mathbf{o}^t_l}{\alpha(l)}, \ln\frac{\alpha(l)\mathbf{1}_{m_l}}{q_{Y=l}}\right\rangle_F.
\end{aligned} \tag{S.5}$$

Denote $\bar{\mathbf{o}}^s_l = \frac{\mathbf{o}^s_l}{\alpha(l)}\in$ and $\bar{\mathbf{o}}^t_l = \frac{\mathbf{o}^t_l}{\alpha(l)}$, then $\sum_i[\bar{\mathbf{o}}^s_l]_i = \sum_i[\bar{\mathbf{o}}^t_l]_i = 1$ since the mass assigned to l-th class is $\alpha(l)$. Then Eq. (S.5)

can be further written as

$$
\begin{aligned}
&\left\langle \frac{\mathbf{o}_l^s}{\alpha(l)}, \ln \frac{\mathbf{o}_l^s}{q_{Y=l} R_{X|l}^{w^*}} \right\rangle_F + \left\langle \frac{\mathbf{o}_l^t}{\alpha(l)}, \ln \frac{\mathbf{o}_l^t}{q_{Y=l} Q_{X|l}} \right\rangle_F \\
&= \left\langle \bar{\mathbf{o}}_l^s, \ln \frac{\bar{\mathbf{o}}_l^s}{R_{X|l}^{w^*}} \right\rangle_F + \ln \frac{\alpha(l)}{q_{Y=l}} \left\langle \bar{\mathbf{o}}_l^s, \mathbf{1}_{n_l} \right\rangle_F + \left\langle \bar{\mathbf{o}}_l^t, \ln \frac{\bar{\mathbf{o}}_l^t}{Q_{X|l}} \right\rangle_F + \ln \frac{\alpha(l)}{q_{Y=l}} \left\langle \bar{\mathbf{o}}_l^t, \mathbf{1}_{m_l} \right\rangle_F \\
&= \left\langle \bar{\mathbf{o}}_l^s, \ln \frac{\bar{\mathbf{o}}_l^s}{R_{X|l}^{w^*}} \right\rangle_F + \ln \frac{\alpha(l)}{q_{Y=l}} + \left\langle \bar{\mathbf{o}}_l^t, \ln \frac{\bar{\mathbf{o}}_l^t}{Q_{X|l}} \right\rangle_F + \ln \frac{\alpha(l)}{q_{Y=l}}
\end{aligned} \tag{S.6}
$$

Finally, by substituting KL terms in Eq. (S.6) into main proof Eq. (S.4), we have

$$
\begin{aligned}
&S_{\text{mask}}^{\lambda,\beta}(P^{w^*}, Q, \tilde{\mathbf{C}}) \\
&= \sum_{l=1}^{k_0} \min_{\bar{\mathbf{\Gamma}}_{ll}^{\text{sub}} \in \mathcal{M}_+(\mathbb{R}^{n_l \times m_l})} \alpha(l) \Bigg[\left\langle \bar{\mathbf{\Gamma}}_{ll}^{\text{sub}}, \tilde{\mathbf{C}}_{ll}^{\text{sub}} \right\rangle_F + \lambda \left\langle \bar{\mathbf{\Gamma}}_{ll}^{\text{sub}}, \ln \bar{\mathbf{\Gamma}}_{ll}^{\text{sub}} \right\rangle_F + \lambda \ln \alpha(l) \\
&\qquad\qquad + \beta \left\langle \frac{\mathbf{o}_l^s}{\alpha(l)}, \ln \frac{\mathbf{o}_l^s}{q_{Y=l} R_{X|l}^{w^*}} \right\rangle_F + \beta \left\langle \frac{\mathbf{o}_l^t}{\alpha(l)}, \ln \frac{\mathbf{o}_l^t}{q_{Y=l} Q_{X|l}} \right\rangle_F \Bigg] \\
&= \sum_{l=1}^{k_0} \min_{\bar{\mathbf{\Gamma}}_{ll}^{\text{sub}} \in \mathcal{M}_+(\mathbb{R}^{n_l \times m_l})} \alpha(l) \Bigg[\left\langle \bar{\mathbf{\Gamma}}_{ll}^{\text{sub}}, \tilde{\mathbf{C}}_{ll}^{\text{sub}} \right\rangle_F + \lambda \left\langle \bar{\mathbf{\Gamma}}_{ll}^{\text{sub}}, \ln \bar{\mathbf{\Gamma}}_{ll}^{\text{sub}} \right\rangle_F + \lambda \ln \alpha(l) \\
&\qquad\qquad + \beta \left\langle \bar{\mathbf{o}}_l^s, \ln \frac{\bar{\mathbf{o}}_l^s}{R_{X|l}^{w^*}} \right\rangle_F + \beta \ln \frac{\alpha(l)}{q_{Y=l}} + \beta \left\langle \bar{\mathbf{o}}_l^t, \ln \frac{\bar{\mathbf{o}}_l^t}{Q_{X|l}} \right\rangle_F + \beta \ln \frac{\alpha(l)}{q_{Y=l}} \Bigg] \\
&= \sum_{l=1}^{k_0} \alpha(l) \Bigg[\min_{\bar{\mathbf{\Gamma}}_{ll}^{\text{sub}} \in \mathcal{M}_+(\mathbb{R}^{n_l \times m_l})} \left\langle \bar{\mathbf{\Gamma}}_{ll}^{\text{sub}}, \tilde{\mathbf{C}}_{ll}^{\text{sub}} \right\rangle_F + \lambda \left\langle \bar{\mathbf{\Gamma}}_{ll}^{\text{sub}}, \ln \bar{\mathbf{\Gamma}}_{ll}^{\text{sub}} \right\rangle_F + \beta \left\langle \bar{\mathbf{o}}_l^s, \ln \frac{\bar{\mathbf{o}}_l^s}{R_{X|l}^{w^*}} \right\rangle_F + \beta \left\langle \bar{\mathbf{o}}_l^t, \ln \frac{\bar{\mathbf{o}}_l^t}{Q_{X|l}} \right\rangle_F \Bigg] \\
&\qquad\qquad + \sum_{l=1}^{k_0} \left[\lambda \alpha(l) \ln \alpha(l) + 2\beta \ln \alpha(l) \frac{\alpha(l)}{q_{Y=l}} \right] \\
&= \sum_{l=1}^{k_0} \alpha(l) S^{\lambda,\beta}(P_{X|l}^{w^*}, Q_{X|l}, \tilde{\mathbf{C}}_{ll}^{\text{sub}}) - \lambda H(Q_Y) + 2\beta D_\phi(\alpha \| Q_Y) \\
&= S_{\text{cond}}^{\lambda,\beta,\alpha}(P^{w^*}, Q, \mathbf{C}) - \lambda H(Q_Y) + 2\beta D_\phi(\alpha \| Q_Y).
\end{aligned}
$$

Therefore, the non-negative $\alpha(\cdot)$ such that $S_{\text{cond}}^{\lambda,\beta,\alpha}(P^{w^*}, Q, \mathbf{C}) = S_{\text{mask}}^{\lambda,\beta}(P^{w^*}, Q, \tilde{\mathbf{C}}) + C_0(\alpha, Q_Y)$, where $C_0(\alpha, Q_Y) = \lambda H(Q_Y) - 2\beta D_\phi(\alpha \| Q_Y)$

S.2. Experiment Details and Additional Discussions

S.2.1. Implementation Details

The network-based model is implemented in PyTorch [11] platform. For network architectures, f_r consists of ResNet-50 [6] and two Fully-Connected (FC) layers ($\mathbb{R}^{2048} \to \mathbb{R}^{1024} \to \mathbb{R}^{512}$) with batch normalization, where the FC layers are activated by Leaky ReLU ($\alpha = 0.2$) and Tanh, respectively; f_c is a single FC layer ($\mathbb{R}^{512} \to \mathbb{R}^{|\mathcal{Y}|}$) with SoftMax activation. For optimization, we use batch gradient descent with Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$), where the learning rate is set as $1e$-3. Entropic parameter λ is empirically set as $1e$-2 in numerical experiments. To ensure the more accurate OT estimation with larger batch-size, we load the pretrained parameter on ImageNet for the ResNet-50 in representation learner f_r, and then froze them during the training. Therefore, the overall model is trained with batch gradient descent on Office-Home, Office-31, ImageCLEF and mini-batch gradient descent (batch size is 5k) on VisDA-2017. The importance weight w is estimated by BBSE algorithm [7] and updated on the fly. In training stage, we first warm up the model on source domain with risk $\mathbb{E}_P[\ell(f(\mathbf{x}^s), y^s)]$ for 20 epochs, and then train the model with full objective. Such a warm up will reduce the uncertainty induced by pseudo labels effectively. The overall training pipeline for full objective is summarized in Alg. S.1. Note that

Algorithm S.1 MOT-based Model for PDA

Input: labeled source data $\{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^n$ and unlabeled target data $\{(\mathbf{x}_i^t)\}_{i=1}^m$, training epochs $E_{\max}$, conditional alignment parameter η;

Output: representation learner f_r, task learner f_c;

Initialize f_r and f_c as neural networks;

for $i = 1, 2, \ldots, E_{\max}$ **do**

  Forward propagate $\{\mathbf{x}_i^s\}_{i=1}^{n_s}$ and $\{\mathbf{x}_i^t\}_{i=1}^{n_t}$ and obtain $\{(\mathbf{z}_i^s, \hat{y}_i^s)\}_{i=1}^{n_s}$ and $\{(\mathbf{z}_i^t, , \hat{y}_i^t)\}_{i=1}^{n_t}$;

  # Weight Estimation

  Estimate importance weight w on-the-fly with BBSE algorithm [7];

  # Transport Assignment Learning

  Compute reweighted source p_X^w as Def. 1 and maksed kernel $\tilde{\mathcal{K}}$ as Eq. (10) with pseudo target labels $\{\hat{y}_i^t\}_{i=1}^{n_t}$;

  Compute transport plan $\tilde{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}} \mathcal{L}_{\text{cond}}$ for MOT according to Alg. 1;

  # Conditional Alignment and Risk Minimization

  Compute alignment loss $\mathcal{L}_{\text{cond}}(f_r, \tilde{\mathbf{\Gamma}})$ and risk loss $\mathcal{L}_{\text{risk}}(f_r, f_c)$ with transport plan $\tilde{\mathbf{\Gamma}}$ and barycenter mapping ψ;

  Update learners with overall loss $\mathcal{L}(f_r, f_c) = \mathcal{L}_{\text{risk}}(f_r, f_c) + \eta \mathcal{L}_{\text{cond}}(f_r, \tilde{\mathbf{\Gamma}})$:

$$f_r \leftarrow f_r - \lambda \nabla_{f_r} \mathcal{L}(f_r, f_c), \quad f_c \leftarrow f_c - \lambda \nabla_{f_c} \mathcal{L}(f_r, f_c)$$

end for

studying deep model-based implementation with mini-batch OT [3, 9] is also a meaningful direction. Compared with the shallow networks with larger batch-size, mini-batch OT algorithm ensures larger capacity of deep model.

S.2.2. Dataset Details

• **Office-Home** [14] contains 15k images from 4 domains with 65 classes, i.e., *Art* (**A**), *Clipart* (**C**), *Product* (**P**) and *Real-World* (**R**). In PDA setting, target domain consists of the first 25 classes (alphabetical order).

• **VisDA-2017** [12] contains 152k synthetic images from domain **S** and 55k real images from domain **R**. There are 12 classes, and we form target domain with the first 6 classes.

• **Office-31** [13] contains 4k images and 31 classes from *Amazon* (**A**), *Webcam* (**W**), *Dslr* (**D**). We follow standard protocol [1] to form target domain with 10 classes.

• **ImageCLEF** [2] contains 3 domains with 12 classes, i.e., *Caltech* (**C**), *ImageNet* (**I**), *Pascal* (**P**). We form target domain with the first 6 classes as protocol [8].

S.2.3. About Parameters

There are two major parameters for MOT model, i.e., entropic regularization parameter λ and relaxation parameter β for UOT. For sensitivity of parameters, the model performance is generally robust under different entropic parameter λ, while the larger λ (i.e., closer to original OT) may reduce the convergence speed of Sinkhorn iteration. For relaxation parameter β, its impact is related with the degree of label shift. When label shift is severer, β (i.e., penalty on marginals) should be smaller to reduce negative transfer. In this case, the impact of β will be significant, and vice versa.

S.2.4. About Barycenter Map

Note that the barycenter maps learned with original plan and entropy regularized plan are generally different. Empirically, we observed that intuitive strategy for increasing the sparsity of Sinkhorn plan is effective. For example, truncating the small values in plan $\mathbf{\Gamma}$ with threshold or contribution ratio can improve the proportion of accurate connection pairs in $\mathbf{\Gamma}$. Therefore, learning de-biased map and reducing the density of $\mathbf{\Gamma}$, e.g., low entropy regularization in Eq. (12), could be meaningful problems.

S.2.5. About Partial OT

For relaxation strategy, there is another methodology called partial OT (POT) [4, 5, 10]. For masked version of POT, the empirical modeling can be directly achieved by replacing the UOT-based relaxation model Eq. (8) with partial OT. But, the theoretical understanding of mask mechanism with partial OT needs an in-depth analysis, which could be a meaningful problem. We will provide preliminary discussions on masked partial OT.

References

[1] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial adversarial domain adaptation. In *ECCV*, pages 135–150, 2018. 6

[2] Barbara Caputo, Henning Müller, Jesus Martinez-Gomez, Mauricio Villegas, Burak Acar, Novi Patricia, Neda Marvasti, Suzan Üsküdarlı, Roberto Paredes, Miguel Cazorla, et al. Imageclef 2014: Overview and analysis of the results. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 192–211, 2014. 6

[3] Kilian Fatras, Younes Zine, Rémi Flamary, Remi Gribonval, and Nicolas Courty. Learning with minibatch wasserstein: asymptotic and gradient properties. In *AISTAT*, pages 2131–2141, 2020. 6

[4] Alessio Figalli. The optimal partial transport problem. *Archive for Rational Mechanics and Analysis*, 195(2):533–560, 2010. 6

[5] Xiang Gu, Yucheng Yang, Wei Zeng, Jian Sun, and Zongben Xu. Keypoint-guided optimal transport with applications in heterogeneous domain adaptation. In *NeurIPS*, 2022. 6

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[7] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *ICML*, pages 3122–3130. PMLR, 2018. 5, 6

[8] You-Wei Luo, Chuan-Xian Ren, Dao-Qing Dai, and Hong Yan. Unsupervised domain adaptation via discriminative manifold propagation. *IEEE TPAMI*, 44(3):1653–1669, 2022. 6

[9] Khai Nguyen, Dang Nguyen, Quoc Nguyen, Tung Pham, Hung Bui, Dinh Phung, Trung Le, and Nhat Ho. On transportation of mini-batches: A hierarchical approach. In *ICML*, page 34, 2022. 6

[10] Khai Nguyen, Dang Nguyen, Tung Pham, Nhat Ho, et al. Improving mini-batch optimal transport via partial transportation. In *ICML*, pages 16656–16690, 2022. 6

[11] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8026–8037, 2019. 5

[12] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017. 6

[13] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226, 2010. 6

[14] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017. 6