Probabilistic Multilevel Clustering via Composite Transportation Distance: Supplementary Material

Nhat Ho[⋆,†] **Viet Huynh**[⋆,‡] **Dinh Phung**[‡] **Michael I. Jordan**[†]

[†] University of California at Berkeley, Berkeley, USA [‡]Monash University, Australia

1 Finite mixtures with regularized composite transportation distance

In this section, we provide detailed analyses for obtaining updates with weights and atoms in Algorithm 1 to find the local solution of the objective function in Eq. (6), which optimizes finite mixtures with regularized composite transportation distance. To ease the presentation, we would like to remind this objective function, which is defined as follows

$$\min_{\boldsymbol{\omega}_K,\boldsymbol{\Theta}_K} \inf_{\boldsymbol{\pi}\in\Pi\left(\frac{1}{n}\mathbf{1}_n,\boldsymbol{\omega}_K\right)} \langle \boldsymbol{\pi}, \boldsymbol{M}\rangle - \lambda \mathbb{H}\left(\boldsymbol{\pi}\right)$$

where $\lambda > 0$ is a penalization term and $\mathbb{H}(\boldsymbol{\pi}) = -\sum_{i,j}\pi_{ij}\log\pi_{ij}$ is an entropy of $\boldsymbol{\pi} \in \Pi\left(\mathbf{1}_n/n, \boldsymbol{\omega}_K\right)$. Here, $P_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$ an empirical measure with respect to samples $X_1, \ldots, X_n$. Furthermore, $\boldsymbol{M} = (M_{ij})$ is a cost matrix such that $M_{ij} = -\log f(X_i|\theta_j)$ for $1 \leq i \leq n, 1 \leq j \leq K$ while $\Pi(\cdot,\cdot)$ is the set of transportation plans between $\mathbf{1}_n/n$ and $\boldsymbol{\omega}_K$.

1.1 Update weights

Our strategy for updating weights $\boldsymbol{\omega}_K$ in the above objective function relies on solving the following relaxation of that optimization problem

$$\inf_{\boldsymbol{\pi}\in\mathcal{S}_n} \langle \boldsymbol{\pi}, \boldsymbol{M}\rangle - \lambda \mathbb{H}\left(\boldsymbol{\pi}\right) \tag{15}$$

where $\mathcal{S}_n = \left\{\boldsymbol{\pi} :\ \sum_{j=1}^{K}\pi_{ij} = 1/n\right\}$. Invoking the Lagrangian multiplier for the constraint $\boldsymbol{\pi}\mathbf{1}_K = \frac{1}{n}\mathbf{1}_n$, the above objective function is equivalent to minimize the following function

$$\mathcal{F} = \sum_{i=1}^{n}\sum_{j=1}^{K}\pi_{ij}M_{ij} + \lambda\sum_{i=1}^{n}\sum_{j=1}^{K}\pi_{ij}\left(\log\pi_{ij} - 1\right) + \sum_{i=1}^{n}\kappa_i\left(\sum_{j=1}^{K}\pi_{ij} - \frac{1}{n}\right).$$

By taking the derivative of $\mathcal{F}$ with respect to π_{ij} and setting it to zero, the following equation holds

$$\frac{\partial\mathcal{F}}{\partial\pi_{ij}} = M_{ij} + \lambda\log\pi_{ij} + \kappa_i = 0.$$

The above equation leads to

$$\pi_{ij} = \exp\left(\frac{-M_{ij} - \kappa_i}{\lambda}\right) = \left(f(X_i|\theta_j)\right)^{1/\lambda}\exp\left(\frac{-\kappa_i}{\lambda}\right).$$

Invoking the condition $\sum_k \pi_{ik} = \frac{1}{n}$, we have

$$\exp\left(\frac{-\kappa_i}{\lambda}\right)\left(\sum_{j=1}^{K}\left(f(X_i|\theta_j)\right)^{1/\lambda}\right) = \frac{1}{n},$$

which suggests that

$$\exp\left(\frac{-\kappa_i}{\lambda}\right) = \frac{1}{n}\frac{1}{\sum_{j=1}^{K}\left(f(X_i|\theta_j)\right)^{1/\lambda}}.$$

Governed by the previous equations, we find that

$$\pi_{ij} = \frac{1}{n}\frac{\left(f(X_i|\theta_j)\right)^{1/\lambda}}{\sum_{j=1}^{K}\left(f(X_i|\theta_j)\right)^{1/\lambda}}. \tag{16}$$

Therefore, we can update the weight ω_j as

$$\omega_j = \sum_{i=1}^{n}\pi_{ij}$$

for any $1 \leq j \leq K$.

1.2 Update atoms

Given the updates for weight $\boldsymbol{\omega}_K$ and the formulation of cost matrix $\boldsymbol{M}$, to obtain the update for atoms θ_j as $1 \leq j \leq K$, we optimize the following objective function

$$\min_{\boldsymbol{\Theta}_K} -\sum_{i=1}^{n}\sum_{j=1}^{K}\pi_{ij}\log f(X_i|\theta_j). \tag{17}$$

Since $f(x|\theta)$ is an exponential family distribution with natural parameter θ, we can represent it as

$$f(x|\theta) = h(x) \exp\left(\langle T(x), \theta \rangle - A(\theta)\right),$$

where $A(\theta)$ is the log-partition function which is convex. Plugging this formulation of $f(x|\theta)$ into the objective function (17) and taking the derivative with respect to θ_j, we obtain the following equation

$$\sum_{i=1}^{n} \pi_{ij} T(X_i) - \omega_j \nabla A(\theta_j) = 0. \tag{18}$$

Therefore, we can update atoms θ_j as the solution of the above equation for $1 \leq j \leq K$,

1.3 Proof for local convergence of Algorithm 1

Given the formulation of Algorithm 1, we would like to demonstrate its convergence to local solution of objective function (6) in Theorem 1.

Our proof of the theorem is straight-forward from the updates of weights and atoms via Lagrangian multipliers. In particular, we denote $\boldsymbol{\omega}_K^{(t)}$, $\boldsymbol{\Theta}_K^{(t)}$, and $\boldsymbol{\pi}^{(t)}$ as the update of weights, atoms, and transportation plan in step t of Algorithm 1 for $t \geq 0$. Additionally, let $\boldsymbol{M}^{(t)}$ be the cost matrix at step t, i.e., $\boldsymbol{M}_{ij}^{(t)} = -\log f(X_i|\theta_j^{(t)})$ for all i, j. Furthermore, we denote

$$g(\boldsymbol{\omega}_K, \boldsymbol{\Theta}_K) := \inf_{\boldsymbol{\pi} \in \Pi\left(\frac{1}{n}\mathbf{1}_n, \boldsymbol{\omega}_K\right)} \langle \boldsymbol{\pi}, \boldsymbol{M} \rangle - \lambda \mathbb{H}(\boldsymbol{\pi}).$$

Then, for any $t \geq 0$, it is clear that

$$\begin{aligned} g(\boldsymbol{\omega}_K^{(t)}, \boldsymbol{\Theta}_K^{(t)}) &= \inf_{\boldsymbol{\pi} \in \Pi\left(\frac{1}{n}\mathbf{1}_n, \boldsymbol{\omega}_K^{(t)}\right)} \left\langle \boldsymbol{\pi}, \boldsymbol{M}^{(t)} \right\rangle - \lambda \mathbb{H}(\boldsymbol{\pi}) \\ &\geq \inf_{\boldsymbol{\pi} \in \mathcal{S}_n} \left\langle \boldsymbol{\pi}, \boldsymbol{M}^{(t)} \right\rangle - \lambda \mathbb{H}(\boldsymbol{\pi}) \\ &\geq \left\langle \boldsymbol{\pi}^{(t+1)}, \boldsymbol{M}^{(t)} \right\rangle - \lambda \mathbb{H}\left(\boldsymbol{\pi}^{(t+1)}\right) \end{aligned}$$

where $\mathcal{S}_n = \left\{\boldsymbol{\pi} : \sum_{j=1}^{K} \pi_{ij} = 1/n \ \forall 1 \leq i \leq n\right\}$. Here, the first inequality is due to the fact that $\Pi\left(\frac{1}{n}\mathbf{1}_n, \boldsymbol{\omega}_K^{(t)}\right) \subset \mathcal{S}_n$ while the second inequality is due to (16) in subsection 1.1. According to the update of atoms in (18) in subsection 1.2, we have that

$$\begin{aligned} &\left\langle \boldsymbol{\pi}^{(t+1)}, \boldsymbol{M}^{(t)} \right\rangle - \lambda \mathbb{H}\left(\boldsymbol{\pi}^{(t+1)}\right) \\ &\quad \geq \left\langle \boldsymbol{\pi}^{(t+1)}, \boldsymbol{M}^{(t+1)} \right\rangle - \lambda \mathbb{H}\left(\boldsymbol{\pi}^{(t+1)}\right) \\ &\quad \geq \inf_{\boldsymbol{\pi} \in \Pi\left(\frac{1}{n}\mathbf{1}_n, \boldsymbol{\omega}_K^{(t+1)}\right)} \left\langle \boldsymbol{\pi}, \boldsymbol{M}^{(t+1)} \right\rangle - \lambda \mathbb{H}(\boldsymbol{\pi}) \\ &\quad = g(\boldsymbol{\omega}_K^{(t+1)}, \boldsymbol{\Theta}_K^{(t+1)}). \end{aligned}$$

Governed by the above results, for any $t \geq 0$, the following holds

$$g(\boldsymbol{\omega}_K^{(t)}, \boldsymbol{\Theta}_K^{(t)}) \geq g(\boldsymbol{\omega}_K^{(t+1)}, \boldsymbol{\Theta}_K^{(t+1)}).$$

As a consequence, we achieve the conclusion of Theorem 1.

2 Regularized composite transportation barycenter

In this section, we provide a detailed algorithm for achieving local solution to regularized composite transportation barycenter in objective function in Eq. (8). To facilitate the discussion, we will remind the formulation of that objective function. In particular, the objective function with regularized composite transportation distance has the following formulation

$$\min_{\mathbf{w}_L, \boldsymbol{\Psi}_L} \sum_{j=1}^{J} a_j \min_{\boldsymbol{\pi}^j \in \Pi\left(\boldsymbol{\omega}_{K_j}^j, \mathbf{w}_L\right)} \left\langle \boldsymbol{\pi}^j, \boldsymbol{M}^j \right\rangle - \lambda \mathbb{H}\left(\boldsymbol{\pi}^j\right)$$

where $\boldsymbol{M}^j$ is the corresponding KL cost matrix between finite mixture probability distribution $P^j_{\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}}$ and $Q_{\mathbf{w}_L, \boldsymbol{\Psi}_L}$ for $1 \leq j \leq J$. Here, $\{a_j\}_{j=1}^J \in \Delta^J$ are given weights associated with the finite mixture probability distributions $\left\{P^j_{\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}}\right\}_{j=1}^{J}$.

As $f(x|\theta)$ is an exponential family, the cost matrix $\boldsymbol{M}^j = (M^j_{uv})$ has the following formulation

$$\begin{aligned} M^j_{uv} &= \mathrm{KL}(f(x|\psi_v), f(x|\theta^j_u)) \\ &= A(\theta^j_u) - A(\psi_v) - \left\langle \nabla A(\psi_v), (\theta^j_u - \psi_v) \right\rangle \end{aligned}$$

for all $1 \leq u, v \leq K$.

2.1 Update weights and atoms

Our procedure for updating weights $\mathbf{w}_K$ for the objective function of regularized composite transportation distance will be similar to Algorithm 1 in [1]. Therefore, we will only focus on the updates with atoms $\boldsymbol{\Psi}_L = (\psi_1, \ldots, \psi_L)$.

Given the updates of weights $\mathbf{w}_K$, we compute the optimal transportation plan $\boldsymbol{\pi}^j = (\pi^j_{uv})$ between $\boldsymbol{\omega}^j_{K_j}$ and $\mathbf{w}_L$ using Algorithm 3 in [1]. Then, to obtain the updates for $\boldsymbol{\Psi}_L$, we consider the following optimization problem

$$\min_{\boldsymbol{\Psi}_L} \sum_{j=1}^{J} a_j \sum_{u=1}^{K_j} \sum_{v=1}^{K_j} \pi^j_{uv} M^j_{uv}.$$

By taking the derivative of the above objective function with respect to ψ_v and setting it to 0, we achieve the following equation

$$\sum_{j=1}^{J}\sum_{u=1}^{K_j}\pi_{uv}^{j}\left\langle \nabla^2 A(\psi_v), \theta_u^j - \psi_v\right\rangle = 0.$$

One possible solution to the above equation is $\sum_{j=1}^{J}\sum_{u=1}^{K_j}\pi_{uv}^{j}(\theta_u^j - \psi_v) = 0$. This previous equation suggests that

$$\psi_v = \frac{\sum_{j=1}^{J}\sum_{u=1}^{K_j}\pi_{uv}^{j}\theta_u^j}{\sum_{j=1}^{J}\sum_{u=1}^{K_j}\pi_{uv}^{j}} \tag{19}$$

for all $1 \leq v \leq L$. Equipped with these updates for weights $\mathbf{w}_L$ and atoms $\mathbf{\Psi}_L$, we summarize the detail of an algorithm for determining the local solution of regularized composite transportation barycenter in Eq. (8) in Algorithm 3.

Algorithm 3 Regularized composite transportation barycenter

Input: Finite mixture probability distributions $\left\{P^j_{\boldsymbol{\omega}^j_{K_j}, \mathbf{\Theta}^j_{K_j}}\right\}_{j=1}^{J}$, given weights $\{a_j\}_{j=1}^{J}$, and the regularized hyper-parameter $\lambda > 0$.

Output: Optimal weights $\mathbf{w}_L$ and atoms $\mathbf{\Psi}_L$.

Initialize weights $\{w_j\}_{j=1}^{L}$ and atoms $\{\psi_j\}_{j=1}^{L}$.

while not converged **do**

1. Update weights $\mathbf{w}_L$ as Algorithm 1 in [1].
2. Compute transportation plans $\boldsymbol{\pi}^j$ for $1 \leq j \leq J$ using Algorithm 3 in [1].
3. Update atoms $\mathbf{\Psi}_L$ as in Eq. (19).

end while

2.2 Local convergence of Algorithm 3

Given the formulation of Algorithm 3, the following theorem demonstrates that this algorithm determines the local solution of objective function (8)

Theorem 3. *The Algorithm 3 monotonically decreases the objective function* (8) *of regularized composite transportation barycenter until local convergence.*

The proof of Theorem 3 is a direct consequence of the updates with weights and atoms in the above subsection and can be argued in the similar fashion as that of Theorem 1; therefore, it is omitted.

3 Multilevel clustering with composite transportation distance

In this section, we provide detailed argument for the algorithm development to determine the local solutions of regularized multilevel composite transportation (MCT). To ease the presentation later, we would like to remind the objective function of this problem as well as all its important relevant notations. We start with the objective function in Eq. (12) as follows

$$\inf_{\boldsymbol{\omega}^j_{K_j}, \mathbf{\Theta}^j_{K_j}, \boldsymbol{\mathcal{Q}}} \sum_{j=1}^{J} \widehat{W}\left(P^j_{n_j}, P^j_{\boldsymbol{\omega}^j_{K_j}, \mathbf{\Theta}^j_{K_j}}\right) + \zeta \widehat{W}\left(\boldsymbol{P}, \boldsymbol{\mathcal{Q}}\right) - R\left(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{a}\right),$$

where $R\left(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{a}\right) = \lambda_l \sum_{j=1}^{J}\mathbb{H}(\boldsymbol{\pi}^j) + \zeta\lambda_a[\mathbb{H}(\boldsymbol{a}) + \lambda_g \sum_{j=1}^{J}\sum_{m=1}^{\boldsymbol{C}}\mathbb{H}(\boldsymbol{\tau}^{j,m})]$ is a combination of all regularized terms for the local and global clustering. Here, for the simplicity of our argument, we choose $\zeta = 1$ to derive our learning updates. In the above objective function, $\boldsymbol{P} = \frac{1}{J}\sum_{j=1}^{J}\delta_{P^j_{\boldsymbol{\omega}^j_{K_j}, \mathbf{\Theta}^j_{K_j}}}$ and $\boldsymbol{\mathcal{Q}} = \sum_{m=1}^{\boldsymbol{C}} b_m \delta_{Q^m_{\mathbf{w}^m_L, \mathbf{\Psi}^m_L}}$. We summarize below the notations for our algorithm development.

Variables of local clustering structures

- Local transportation plans for group j: $\boldsymbol{\pi}^j = \left\{\pi^j_{uv}\right\}_{u,v} \in \Pi(\frac{1}{n_j}\mathbf{1}_{n_j}, \boldsymbol{\omega}^j_{K_j})$ s.t. $\sum_v \pi^j_{uv} = \frac{1}{n_j}$, and $\sum_u \pi^j_{uv} = \omega^j_v$,
- Local atoms for local group $\mathbf{\Theta}^j_{K_j} = \left\{\theta^j_k\right\}_{k=1}^{K_j}$ and their local mixing weights $\boldsymbol{\omega}^j = \left\{\omega^j_k\right\}_{k=1}^{K_j}$.

Variables of assignment group to barycenter

- Global transportation plan $\boldsymbol{a} = (a_{jm}) \in \Pi(\frac{1}{J}\mathbf{1}_J, \boldsymbol{b})$ between $\boldsymbol{P}$ and $\boldsymbol{\mathcal{Q}}$.

Variables of global clustering structures

- Partial global transportation plans between local measure $P_{\boldsymbol{\omega}^j_{K_j}, \mathbf{\Theta}^j_{K_j}}$ and global measure $Q^m_{\mathbf{w}^m, \Psi^m_k}$: $\boldsymbol{\tau}^{j,m} = \left\{\tau^{j,m}_{kl}\right\}_{k,l}$ where $\sum_l \tau^{j,m}_{kl} = \omega^j_k$ and $\sum_l \tau^{j,m}_{kl} = w^m_l$ for all $1 \leq j \leq J$ and $1 \leq m \leq \boldsymbol{C}$.
- Global atoms for global measure $\Psi^m_L = \{\psi^m_l\}_{l=1}^{L}$. and global mixing weights $\mathbf{w}^m_L = \{w^m_l\}_{l=1}^{L}$where $w^m_l = \sum_k \tau^{jm}_{kl}$ for any j.

3.1 Local clustering updates

As being mentioned in the main text, to obtain updates for local weights $\boldsymbol{\omega}^j_{K_j}$ and local atoms $\mathbf{\Theta}^j_{K_j}$, we solve the following regularized composite transportation barycenter problem

$$\inf_{\boldsymbol{\omega}^j_{K_j},\boldsymbol{\Theta}^j_{K_j}} \widehat{W}\left(P^j_{n_j}, P^j_{\boldsymbol{\omega}^j_{K_j},\boldsymbol{\Theta}^j_{K_j}}\right) - \lambda_l \mathbb{H}(\boldsymbol{\pi}^j)$$
$$+ \sum_{m=1}^{\boldsymbol{C}} a_{jm} \widehat{W}(P^j_{\boldsymbol{\omega}^j_{K_j},\boldsymbol{\Theta}^j_{K_j}}, Q^m_{\mathbf{w}^m_L,\Psi^m_L}) - \lambda_g \sum_{m=1}^{K} \mathbb{H}(\boldsymbol{\tau}^{j,m}).$$

The above objective function can be rewritten as

$$\inf_{\boldsymbol{\omega}^j_{K_j},\boldsymbol{\Theta}^j_{K_j}} \inf_{\boldsymbol{\pi}^j \in \Pi(\frac{1}{n_j}\mathbf{1}_{n_j}, \boldsymbol{\omega}^j_{K_j})} \left\langle \boldsymbol{\pi}^j, \boldsymbol{M}^j \right\rangle - \lambda_l \mathbb{H}(\boldsymbol{\pi}^j)$$
$$+ \sum_{m=1}^{\boldsymbol{C}} a_{jm} \inf_{\boldsymbol{\tau}^{j,m} \in \Pi(\boldsymbol{\omega}^j_{K_j}, \mathbf{w}^m_L)} \left\langle \boldsymbol{\tau}^{j,m}, \boldsymbol{\gamma}^{j,m} \right\rangle$$
$$- \lambda_g \sum_{m=1}^{K} \mathbb{H}(\boldsymbol{\tau}^{j,m}) \quad (20)$$

where $\boldsymbol{M}^j$ is the cost matrix between $P^j_{n_j}$ and $P^j_{\boldsymbol{\omega}^j_{K_j},\boldsymbol{\Theta}^j_{K_j}}$ that has a formulation as

$$[M^j]_{uv} = -\log f(X_{j,u}|\theta^j_v),$$

for $1 \le u \le n_j$ and $1 \le v \le K_j$. Additionally, the cost matrix $\boldsymbol{\gamma}^{j,m}$ has the following formulation

$$\gamma^{j,m}_{kl} = A\left(\theta^j_k\right) - A\left(\psi^m_l\right) - \left\langle \nabla A\left(\psi^m_l\right), \left(\theta^j_k - \psi^m_l\right) \right\rangle.$$

Update local weights: The idea for obtaining the local solutions of above objective function is similar to that in Section 1.

Update local atoms: Given the updates for local weight $\boldsymbol{\omega}^j_{K_j}$, to obtain the update equation for local atoms $\boldsymbol{\Theta}^j_{K_j}$, we optimize the following objective function

$$\min_{\boldsymbol{\Theta}^j_{K_j}} - \sum_{u=1}^{n_j} \sum_{v=1}^{K_j} \pi^j_{uv} \log f(X_{j,u}|\theta^j_v) + \sum_{m=1}^{\boldsymbol{C}} a_{jm} \sum_{v,l} \tau^{j,m}_{vl} \gamma^{j,m}_{kl}. \quad (21)$$

Since $f(x|\theta)$ is an exponential family distribution with natural parameter θ, we can represent it as

$$f(x|\theta) = h(x) \exp\left(\langle T(x), \theta \rangle - A(\theta)\right),$$

where $A(\theta)$ is the log-partition function which is convex. Given that formulation of $f(x|\theta)$, our objective function (17) is equivalent to minimize the following objective function

$$\mathcal{F}_{\text{local}} = - \sum_{u=1}^{n_j} \sum_{v=1}^{K_j} \pi^j_{uv} \left(\left\langle T\left(X_{j,u}\right), \theta^j_v \right\rangle - A\left(\theta^j_v\right) \right)$$
$$+ \sum_{m=1}^{\boldsymbol{C}} a_{jm} \sum_{v,l} \tau^{j,m}_{kl} \Big[A\left(\theta^j_v\right) - A\left(\psi^m_l\right)$$
$$- \left\langle \nabla A\left(\psi^m_l\right), \left(\theta^j_v - \psi^m_l\right) \right\rangle \Big].$$

By direct computation, $\mathcal{F}_{\text{local}}$ has the following partial derivative with respect to θ^j_v

$$\frac{\partial \mathcal{F}_{\text{local}}}{\partial \theta^j_v} = - \sum_{u=1}^{n_j} \pi^j_{uv} \left(T(X_{j,u}) - \nabla A\left(\theta^j_v\right) \right)$$
$$+ \sum_{m=1}^{\boldsymbol{C}} a_{jm} \sum_{l=1}^{L} \tau^{j,m}_{vl} \left[\nabla A\left(\theta^j_v\right) - \nabla A\left(\psi^m_l\right) \right]$$
$$= - \sum_{u=1}^{n_j} \pi^j_{uv} T(X_{j,u}) + \omega^j_v \nabla A\left(\theta^j_v\right)$$
$$+ \sum_{m=1}^{C} a_{jm} \sum_{l=1}^{L} \tau^{j,m}_{vl} \left(\nabla A\left(\theta^j_v\right) - \nabla A\left(\psi^m_l\right) \right),$$

where in the last equality, we use the identity $\sum_{u=1}^{n_j} \pi^j_{uv} = \omega^j_v$.

Given the above partial derivatives, we can update the atoms θ^j_v to be the solution of the following equation

$$\nabla A\left(\theta^j_v\right) = \frac{\sum_{m=1}^{C} a_{jm} \sum_{l=1}^{L} \tau^{j,m}_{vl} \nabla A\left(\psi^m_l\right) + \sum_{u=1}^{n_j} \pi^j_{uv} T(X_{j,u})}{\sum_{m=1}^{C} a_{jm} \sum_{l=1}^{L} \tau^{j,m}_{vl} + \omega^j_v} \quad (22)$$

3.2 Computing global transportation plan

Given the updates for local weights $\boldsymbol{\omega}^j_{K_j}$ and local atoms $\boldsymbol{\Theta}^j_{K_j}$ for $1 \le j \le J$, we now develop an update on for global transportation plan $\boldsymbol{a} = (a_{jm})$ between $\boldsymbol{P}$ and $\boldsymbol{Q}$. Our strategy for the update relies on solving the following objective function

$$\inf_{\boldsymbol{a}} \sum_{j,m} a_{jm} \widehat{W}\left(P^j_{\boldsymbol{\omega}^j_{K_j},\boldsymbol{\Theta}^j_{K_j}}, Q^m_{\mathbf{w}^m_L,\boldsymbol{\Psi}^m_L}\right) - \lambda_a \mathbb{H}\left(\boldsymbol{a}\right).$$

where $\boldsymbol{a}$ in the above infimum satisfies the constraint $\boldsymbol{a}\mathbf{1}_K = \frac{1}{J}\mathbf{1}_n$. By means of Lagrangian multiplier, the above objective function can be rewritten as

$$\mathcal{F}_{\text{global}} = \sum_{j,m} a_{jm} \widehat{W}\left(P^j_{\boldsymbol{\omega}^j_{K_j},\boldsymbol{\Theta}^j_{K_j}}, Q^m_{\mathbf{w}^m_L,\boldsymbol{\Psi}^m_L}\right)$$
$$- \lambda_a \mathbb{H}\left(\boldsymbol{a}\right) + \kappa_a \sum_{j=1}^{J} \left(\sum_{m=1}^{\boldsymbol{C}} a_{jm} - \frac{1}{J} \right),$$

The function $\mathcal{F}_{\text{global}}$ has the partial derivative with respect to a_{jm} as follows

$$\frac{\partial \mathcal{F}_{\text{global}}}{\partial a_{jm}} = \widehat{W}\left(P^j_{\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}}, Q^m_{\mathbf{w}^m_L, \boldsymbol{\Psi}^m_L}\right) + \lambda_a \log a_{jm} + \kappa_a$$
$$= \sum_{k,l} \tau^{j,m}_{kl}\gamma^{j,m}_{kl} + \lambda_a \ln a_{jm} + \kappa_a.$$

Setting the above derivative to 0 and invoking the constraint that $\sum_{m=1}^{\boldsymbol{C}} a_{jm} = \frac{1}{J}$, we find that

$$a_{jm} = \frac{1}{J}\frac{\exp\left(-\sum_{k,l}\tau^{j,m}_{kl}\gamma^{j,m}_{kl}\right)^{1/\lambda_a}}{\sum_m \exp\left(-\sum_{k,l}\tau^{j,m}_{kl}\gamma^{j,m}_{kl}\right)^{1/\lambda_a}} \tag{23}$$

for $1 \leq j \leq J$ and $1 \leq m \leq L$.

3.3 Global clustering updates

Given the updates with local weights and atoms as well as the global transportation plan, we are now ready to develop an update for global weights $\mathbf{w}^m_L$ and global atoms $\boldsymbol{\Psi}^m_L$ for $1 \leq m \leq \boldsymbol{C}$. In particular, the objective function for updating these global parameters are as follows

$$\min_{\{\mathbf{w}^m_L\},\{\boldsymbol{\Psi}^m_L\}} \sum_{j=1}^{J}\sum_{m=1}^{\boldsymbol{C}} a_{jm}\widehat{W}(P^j_{\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}}, Q^m_{\mathbf{w}^m_L, \boldsymbol{\Psi}^m_L})$$
$$- \lambda_g \sum_{j=1}^{J}\sum_{m=1}^{\boldsymbol{C}} \mathbb{H}(\boldsymbol{\tau}^{j,m})$$

The above objective function can be rewritten as

$$\min_{\{\mathbf{w}^m_L\},\{\boldsymbol{\Psi}^m_L\}} \sum_{j=1}^{J}\sum_{m=1}^{\boldsymbol{C}} a_{jm} \inf_{\boldsymbol{\tau}^{j,m}\in\Pi(\boldsymbol{\omega}^j_{K_j}, \mathbf{w}^m_L)} \Big\langle \boldsymbol{\tau}^{j,m}, \boldsymbol{\gamma}^{j,m}\Big\rangle$$
$$+ \lambda_g \sum_{k,l}\tau^{j,m}_{kl}\left(\log\tau^{j,m}_{kl} - 1\right)\Big)$$

Given the above objective function, for each m, to update the global weights $\mathbf{w}^m_L$ and global atoms $\boldsymbol{\Psi}^m_L$, we consider the following composite transportation barycenter

$$\min_{\mathbf{w}^m_L, \boldsymbol{\Psi}^m_L} \sum_{j=1}^{J} a_{jm} \inf_{\boldsymbol{\tau}^{j,m}\in\Pi(\boldsymbol{\omega}^j_{K_j}, \mathbf{w}^m_L)} \Big\langle \boldsymbol{\tau}^{j,m}, \boldsymbol{\gamma}^{j,m}\Big\rangle$$
$$+ \lambda_g \sum_{k,l}\tau^{j,m}_{kl}\left(\log\tau^{j,m}_{kl} - 1\right)\Big).$$

Update global weights: Given the above objective function, the idea for updating the global weights $\mathbf{w}^m_L$ is similar to Algorithm 1 in [1].

Update partial transportation plans: Once global weights are obtained, we can use Algorithm 3 in [1] to update the optimal partial transportation plans $\boldsymbol{\tau}^{j,m}$ between local measure $P_{\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}}$ and global measure $Q^m_{\mathbf{w}^m, \Psi^m_k}$.

Update global atoms: With the updates for the global weight $\mathbf{w}^m_L$, to obtain the update equation for global atoms $\boldsymbol{\Psi}^m_L$, we minimize the following objective function

$$\mathcal{F}_{\text{p-global}} = \sum_{j=1}^{J} a_{jm}\sum_{k,l}\tau^{j,m}_{kl}\gamma^{j,m}_{kl}$$
$$= \sum_{j=1}^{J} a_{jm}\sum_{k,l}\tau^{j,m}_{kl}\Big(A\left(\theta^j_k\right) - A\left(\psi^m_l\right)$$
$$- \Big\langle \nabla A\left(\psi^m_l\right), \theta^j_k - \psi^m_l\Big\rangle\Big).$$

Taking the derivative of $\mathcal{F}_{\text{p-global}}$ with respect to ψ^m_l and setting it to zero, we find that

$$\frac{\partial \mathcal{F}_{\text{p-global}}}{\partial \psi^m_l} = \sum_{j=1}^{J} a_{jm}\sum_{k}\tau^{j,m}_{kl}\nabla^2 A\left(\psi^m_l\right)\left(\psi^m_l - \theta^j_k\right) = 0.$$

Since the log-partition function $A(\cdot)$ is convex, $\nabla^2 A(\psi_v)$ is a positive-semidefinite matrix. Therefore, we can choose $\sum_{j=1}^{J} a_{jm}\sum_k \tau^{j,m}_{kl}\left(\psi^m_l - \theta^j_k\right) = 0$, which means that

$$\psi^m_l = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K_j} a_{jm}\tau^{j,m}_{kl}\theta^j_k}{\sum_{j=1}^{J}\sum_{k=1}^{K_j} a_{jm}\tau^{j,m}_{kl}}.$$

3.4 Proof for local convergence of Algorithm 2

Equipped with the above updates with local and global parameters of regularized MCT, we are ready to demonstrate the convergence of Algorithm 2 to local solution of objective function (12) of regularized MCT in Theorem 2. To simplify the argument, we only provide proof sketch for this theorem.

In particular, we denote $\boldsymbol{\omega}^{j,(t)}_{K_j}$ and $\boldsymbol{\Theta}^{j,(t)}_{K_j}$ as the updates of local weights and local atoms in step t of Algorithm 2 for $t \geq 0$. Similarly, we denote $\mathbf{w}^{m,(t)}_L$ and $\boldsymbol{\Psi}^{m,(t)}_L$ as the updates of global weights and global atoms at step t. Furthermore, we denote

$$g(\{\boldsymbol{\omega}^{j,(t)}_{K_j}\}, \{\boldsymbol{\Theta}^{j,(t)}_{K_j}\}, \{\mathbf{w}^{m,(t)}_L\}, \{\boldsymbol{\Psi}^{m,(t)}_L\})$$
$$:= \sum_{j=1}^{J}\widehat{W}\left(P^j_{n_j}, P^j_{\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}}\right) + \zeta\widehat{W}\left(\boldsymbol{P}, \boldsymbol{\mathcal{Q}}\right)$$
$$- R\left(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{a}\right).$$

Then, according to local clustering updates step, we would have

$$g(\{\boldsymbol{\omega}_{K_j}^{j,(t)}\}, \{\boldsymbol{\Theta}_{K_j}^{j,(t)}\}, \{\mathbf{w}_L^{m,(t)}\}, \{\boldsymbol{\Psi}_L^{m,(t)}\}) \\ \geq g(\{\boldsymbol{\omega}_{K_j}^{j,(t+1)}\}, \{\boldsymbol{\Theta}_{K_j}^{j,(t+1)}\}, \{\mathbf{w}_L^{m,(t)}\}, \{\boldsymbol{\Psi}_L^{m,(t)}\}).$$

On the other hand, invoking the global clustering updates step, we achieve

$$g(\{\boldsymbol{\omega}_{K_j}^{j,(t+1)}\}, \{\boldsymbol{\Theta}_{K_j}^{j,(t+1)}\}, \{\mathbf{w}_L^{m,(t)}\}, \{\boldsymbol{\Psi}_L^{m,(t)}\}) \\ \geq g(\{\boldsymbol{\omega}_{K_j}^{j,(t+1)}\}, \{\boldsymbol{\Theta}_{K_j}^{j,(t+1)}\}, \{\mathbf{w}_L^{m,(t+1)}\}, \{\boldsymbol{\Psi}_L^{m,(t+1)}\})$$

Governed by the above results, for any $t \geq 0$, the following holds

$$g(\{\boldsymbol{\omega}_{K_j}^{j,(t)}\}, \{\boldsymbol{\Theta}_{K_j}^{j,(t)}\}, \{\mathbf{w}_L^{m,(t)}\}, \{\boldsymbol{\Psi}_L^{m,(t)}\}) \\ \geq g(\{\boldsymbol{\omega}_{K_j}^{j,(t+1)}\}, \{\boldsymbol{\Theta}_{K_j}^{j,(t+1)}\}, \{\mathbf{w}_L^{m,(t+1)}\}, \{\boldsymbol{\Psi}_L^{m,(t+1)}\}).$$

As a consequence, we achieve the conclusion of Theorem 2.

References

[1] M. Cuturi and A. Doucet. Fast computation of wasserstein barycenters. *Proceedings of the 31st International Conference on Machine Learning*, 2014.