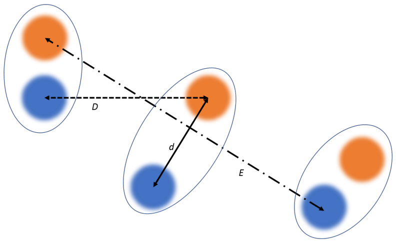


Figure 1: An example of three pairs of Gaussian clusters in the source (blue) and target (red) domains. The maximum distance d between associated (paired) centers, the minimum distance D between unassociated centers and the maximum distance E of centers between two domains are respectively shown by solid, dashed and dash-dotted lines.

7 Class Based Regularization

Asymptotic analysis for Gaussian Mixtures: We start with a simplified probabilistic result for Gaussian mixtures in an asymptotic scenario, reflecting the main underlying intuitions of our analysis. We shortly present a more extensive study for finite and deterministic cases, which is also used for proving the first result:

Theorem 6. *Suppose in each of K domains, that an equal number m of random real vectors are drawn from each of K individual Gaussian distributions, leading to a total number of $n = mK$ samples. The Gaussian distributions in the source and target domains are respectively centered at $\boldsymbol{\theta}^s_\alpha, \boldsymbol{\theta}^t_\alpha$ for $\alpha = 1, 2, \ldots, K$, and all have uncorrelated entries with equal variance ω^2. Squared ℓ_2 distance is used, $d(\mathbf{y}_1, \mathbf{y}_2) = \|\mathbf{y}_1 - \mathbf{y}_2\|_2^2$. With a probability higher than $1 - {}^1\!/n^{10}$ the solution of equation 2.2 with a suitable choice of λ classifies the samples of each Gaussian distribution, and associate the α^{th} distributions of the two domains for each $\alpha \in [K]$ if*

$$\frac{D^2 - d^2}{K\sqrt{K}} \geq C\sqrt{E^2 + \omega^2}\log(nK), \tag{7.1}$$

for some universal constant C, where $D = \min_{\alpha \neq \beta} \|\boldsymbol{\theta}^s_\alpha - \boldsymbol{\theta}^t_\beta\|$, $d = \max_\alpha \|\boldsymbol{\theta}^s_\alpha - \boldsymbol{\theta}^t_\alpha\|$ and $E = \max_{\alpha,\beta} \|\boldsymbol{\theta}^s_\alpha - \boldsymbol{\theta}^t_\beta\|$.

Fig. 1 clarifies in a simple example the geometric meaning of the concepts used in the above result. As seen, the left hand side of the condition in equation 7.1, requires the associated clusters to be substantially closer to each other than the other clusters. Moreover, the right hand side of equation 7.1 requires the distances to remain relatively bounded. An example of this situation is when $E, D, d \sim \sigma K\sqrt{K}\log(nK)$, i.e. all three grow proportionally with the number of samples, with a suitable proportion between them.

Deterministic Guarantee: Now, we present an extended deterministic result that is used to prove theorem 6. We use a setting inspired by the stochastic block model. For simplicity, we describe here a model in which the data points in the source and target domains are each partitioned into K parts with equal size m. We respectively denote the partitions in the source and target domains by $\{\mathscr{S}_\alpha\}$, $\{\mathscr{T}_\beta\}$. The total number of points in each domain is $n = mK$ (these assumptions are relaxed in Appendix B). Further, $\mathscr{S}_\alpha$ is paired with Δ_α for every $\alpha \in [K]$. We investigate that the plan obtained by solving equation 2.2 consists of blocks, recovering both the sets of clusters $\{\mathscr{S}_\alpha\}$, $\{\mathscr{T}_\beta\}$ and their association. For this, we ensure that X_{ij} remains zero for the i^{th} data point in the source domain and j^{th} data point in the target domain, belonging to unassociated clusters. Accordingly, we require the *ideal solution* to be the one with $X_{i,j} = X_{\alpha,\beta}$ for $i \in \mathscr{S}_\alpha$ and $j \in \mathscr{T}_\beta$, where $X_{\alpha,\beta}$ are constants satisfying $X_{\alpha,\beta} = 0$ for $\beta \neq \alpha$.

For simplicity, we take $S_{j,j'} = 1$ everywhere and study two cases where $R_{i,i'} = 1$ holds true either everywhere (no kernel) or for i, i' belonging to the same cluster and $R_{i,i'} = 0$ otherwise (perfect kernels in the source domain). The general case is presented in Appendix B. Introducing an indicator variable R, the first case is referred to by $R = 0$ and the second one by $R = 1$. Note also that we assume the optimization in equation 2.2 to be feasible for our ideal solution, which requires for every $i, i' \in \mathscr{S}_\alpha$ and $j, j' \in \mathscr{T}_\alpha$ that $\mu_i = \mu_{i'} = \nu_j = \nu_{j'}$. In Appendix B, we treat the general infeasible cases by considering a relaxation of equation 2.2.

In the context of recovery by the Kantorovich relaxation, a key concept is cyclical monotonicity [Villani, 2008], which we slightly modify and state below:

Definition 7. *We say that a set of coefficients $D_{\alpha,\alpha'}$ for $\alpha, \alpha' \in [K]$ satisfies the $\delta-$strong cyclical monotonicity condition* *if for each simple loop $\alpha_1 \to \alpha_2 \to \ldots \to \alpha_k \to \alpha_{k+1} = \alpha_1$ with length $k > 1$ we have*

$$\sum_{l=1}^{k} D_{\alpha_l \alpha_{l+1}} > \sum_{l=1}^{k} D_{\alpha_l \alpha_l} + k\delta. \tag{7.2}$$

Compared to the standard notion of cyclic monotonicity, we introduce a constant $\delta \geq 0$ in the right hand side of equation 7.2, which can be nonzero only when $(D_{\alpha,\beta})$ has a discrete or discontinuous nature. We apply this condition to the average distance of clusters given by $D_{\alpha,\beta} = \frac{1}{m^2} \sum\limits_{i \in \mathscr{S}_\alpha, j \in \mathscr{T}_\beta} D_{i,j}$.

We denote by Δ the maximum of the values ${\|\mathbf{d}_i - \mathbf{d}_{i'}\|}/{\sqrt{n}}$ and ${\left\|\mathbf{d}^j - \mathbf{d}^{j'}\right\|}/{\sqrt{n}}$ where source points i, i' and target points j, j' belong to the same cluster and we remind that $\mathbf{d}_i, \mathbf{d}^j$ respectively refer to the rows and columns of $\mathbf{D}$. We also define $\omega_\alpha := \sum\limits_{i \in \mathscr{S}_\alpha} \mu_i = \sum\limits_{j \in D_\alpha} \nu_j$ and then take $T_{\alpha,\beta} = \sum\limits_{\gamma \in [K]} \left(\frac{R\omega_\alpha}{\sqrt{\omega_\alpha^2 + \omega_\gamma^2}} + \frac{\omega_\beta}{\sqrt{\omega_\beta^2 + \omega_\gamma^2}} \right) - \frac{1+R}{\sqrt{2}}$. Finally, we define

$$\Lambda_{\alpha,\beta} = \left(T_{\alpha,\beta} + \frac{\omega_\alpha + R\omega_\beta}{\sqrt{\omega_\beta^2 + \omega_\beta^2}} \right)^{-1},$$

and take Λ as its maximum over $\alpha \neq \beta$. Accordingly, we obtain the following result:

Theorem 8. *Suppose that $(D_{\alpha,\beta})$ is $\delta-$strongly cyclical monotone. Take λ such that $\Delta \leq \lambda\sqrt{m/K}$. Then, the solution of equation 2.2 is given by $X_{ij} = X_{\alpha,\beta}$ for $i \in \mathscr{S}_\alpha$ and $j \in \mathscr{T}_\beta$ satisfying one of the following two conditions:*

1. *We have $X_{\alpha,\beta} = {\omega_\alpha}/{m^2}\delta_{\beta,\alpha}$ if $\Delta\sqrt{K} \leq \lambda\sqrt{m} \leq \Lambda\delta$*
2. *Otherwise, we have $\delta \sum\limits_{\beta \neq \alpha} X_{\alpha,\beta} \leq \lambda(1+R)\sqrt{m} \sum\limits_{\alpha \neq \alpha'} \sqrt{\omega_\alpha^2 + \omega_{\alpha'}^2}$.*

Furthermore, the solution is unique in part 1 if all inequalities are strict.

Proof. Proof can be found in section 9. □

The first part of theorem 8 establishes ideal recovery under the condition that the "effective cluster diameter" Δ is relatively smaller than $\Lambda\delta$. The second part gives an upper bound on the error $\sum\limits_{\beta \neq \alpha} X_{\alpha,\beta}$. Note that Δ is always smaller with $R = 1$ compared to $R = 0$, making the conditions less restrictive. This reflects the intuitive fact that introducing kernels simplifies the estimation process.

Proof of Theorem 6: Based on theorem 8, we present a sketch of the proof for theorem 6. Under the assumptions of theorem 6, we directly verify that $\delta = D^2 - d^2$ is a valid choice. Moreover $\Lambda = \Lambda_{\alpha,\beta} = {\sqrt{2}}/{K(1+R)}$. Finally, we may conclude by Chernoff bound that with a probability exceeding $1 - {1}/{n^{10}}$ (the power 10 is arbitrary) we have $\Delta = O(\sqrt{E^2 + \omega^2}\log(nK))$. Replacing these expression in the first part of theorem 8 gives us the result.

8 Extension of Theorem 8

We consider the analysis of our proposed method for general kernel coefficient and cluster sizes. Hence, we respectively consider two partitions $\{C_\alpha\}$, $\{D_\beta\}$ of $[n], [m]$ with the same number of parts K. We denote the cardinalities of C_α and D_β by n_α and m_β, respectively. Further, we consider a permutation π on $[K]$ as the target of OT. Also, we address infeasibility by consider the following optimization:

$$\begin{gathered}\min_{\mathbf{X} \in \mathbb{R}_{\geq 0}^{n \times n}} \langle \mathbf{D}, \mathbf{X} \rangle + \\ \lambda \left(\sum_{i,i'} R_{i,i'} \|\mathbf{x}_i - \mathbf{x}_{i'}\|_2 + \sum_{j,j'} S_{j,j'} \|\mathbf{x}^j - \mathbf{x}^{j'}\|_2 \right) \\ + \frac{\theta}{2} \left(\|\mathbf{X}\mathbf{1} - \boldsymbol{\mu}\|_2^2 + \left\|\mathbf{X}^T\mathbf{1} - \boldsymbol{\nu}\right\|_2^2 \right)\end{gathered} \tag{8.1}$$

where $\theta > 0$ is a design parameter and we remind that $\mathbf{x}_i = (X_{i,j})_j$, $\mathbf{x}^j = (X_{i,j})_i$, and $R_{i,i'}$ and $S_{j,j'}$ are positive kernel coefficients. Now, we introduce few intermediate optimizations to carry out the analysis. Define the following more general characteristic optimization:

$$\begin{gathered}\min_{X_{\alpha,\beta}\geq 0} \sum_{\alpha,\beta} n_\alpha m_\beta X_{\alpha,\beta} D_{\alpha,\beta} + \\ \lambda \left(\sum_{\alpha,\alpha'} R_{\alpha,\alpha'} \|\mathbf{x}_\alpha - \mathbf{x}_{\alpha'}\|_M + \sum_{\beta,\beta'} S_{\beta,\beta'} \|\mathbf{x}^\beta - \mathbf{x}^{\beta'}\|_N \right) \\ + \frac{\theta}{2} \left(\sum_\alpha n_\alpha \left(\mathbf{a}_M^T \mathbf{x}_\alpha - \mu_\alpha\right)^2 + \sum_\beta m_\beta \left(\mathbf{a}_N^T \mathbf{x}^\beta - \nu^\beta\right)^2 \right)\end{gathered} \tag{8.2}$$

where

$$R_{\alpha,\alpha'} = \sum_{i\in C_\alpha,\ i'\in C_{\alpha'}} R_{i,i'}, \quad S_{\beta,\beta'} = \sum_{j\in D_\beta,\ j'\in D_{\beta'}} S_{j,j'}$$

$$D_{\alpha,\beta} = \frac{\sum\limits_{i\in C_\alpha, j\in D_\beta} D_{i,j}}{n_\alpha m_\beta},\ \mu_\alpha = \frac{\sum\limits_{i\in C_\alpha} \mu_i}{n_\alpha},\ \nu^\beta = \frac{\sum\limits_{j\in D_\beta} \nu_j}{m_\beta},$$

$\|\mathbf{x}\|_O = \sqrt{\mathbf{x}^T O \mathbf{x}}$, N, M are diagonal matrices with n_α m_β as diagonals, resepectively, $\mathbf{x}_\alpha = (X_{\alpha,\beta})_\beta$ and $\mathbf{x}^\beta = (X_{\alpha,\beta})_\alpha$, and $\mathbf{a}_M = (m_\alpha)_\alpha$, $\mathbf{a}_N = (n_\alpha)_\alpha$.

Further, define the ideal optimization:

$$\begin{gathered}\min_{Y_{\alpha,\beta}\geq 0} \sum_{\alpha,\beta} Y_{\alpha,\beta} D_{\alpha,\beta} \\ \text{s.t} \\ q_\beta:\ \mathbf{1}^T \mathbf{y}^\beta = \sigma^\beta, p_\alpha : \mathbf{1}^T \mathbf{y}_\alpha = \sigma_\alpha\end{gathered} \tag{8.3}$$

where $\sigma_\alpha = (n_\alpha \mu_\alpha + m_{\pi(\alpha)} \nu^{\pi(\alpha)})/2$, $\sigma^\beta = \sigma_{\pi^{-1}(\beta)} = (n_{\pi^{-1}(\beta)} \mu_{\pi^{-1}(\beta)} + m_\beta \nu^\beta)/2$, and $\{p_\alpha\}$, $\{q_\beta\}$ are dual variables. Also, define $\delta_\alpha = (\mu_\alpha n_\alpha - m_{\pi(\alpha)} \nu^{\pi(\alpha)})/2$, $\delta^\beta = -\delta_{\pi^{-1}(\beta)} = (m_\beta \nu^\beta - n_{\pi^{-1}(\beta)} \mu_{\pi^{-1}(\beta)})/2$ and $\boldsymbol{\delta} = (\delta_\alpha)$. Finally, take

$$R_{i,\alpha} = \sum_{i'\in C_\alpha} R_{i,i'}, \quad S_{j,\beta} = \sum_{j'\in D_\beta} S_{j,j'}$$

In case $\boldsymbol{\delta} = 0$, we may also define the tight characteristic optimization:

$$\begin{gathered}\min_{X_{\alpha,\beta}\geq 0} \sum_{\alpha,\beta} n_\alpha m_\beta X_{\alpha,\beta} D_{\alpha,\beta} + \\ \lambda \left(\sum_{\alpha,\alpha'} R_{\alpha,\alpha'} \|\mathbf{x}_\alpha - \mathbf{x}_{\alpha'}\|_M + \sum_{\beta,\beta'} S_{\beta,\beta'} \|\mathbf{x}^\beta - \mathbf{x}^{\beta'}\|_N \right) \\ \text{s.t.} \\ \mathbf{a}_M^T \mathbf{x}_\alpha = \mu_\alpha, \quad \mathbf{a}_N^T \mathbf{x}^\beta = \nu^\beta\end{gathered} \tag{8.4}$$

which in this case, coincides with equation 8.2 when $\theta = \infty$. We also define

$$T_\alpha = \frac{1}{n_\alpha m_{\pi(\alpha)}} \sum_{\alpha' \neq \alpha} \frac{R_{\alpha,\alpha'} \frac{\sigma_\alpha}{n_\alpha m_{\pi(\alpha)}}}{\sqrt{\left(\frac{\sigma_\alpha}{n_\alpha m_{\pi(\alpha)}}\right)^2 + \left(\frac{\sigma_{\alpha'}}{n_{\alpha'} m_{\pi(\alpha')}}\right)^2}},$$

$$U^\beta = \frac{1}{n_{\pi^{-1}(\beta)} m_\beta} \times$$

$$\sum_{\beta' \neq \beta} \frac{S_{\beta,\beta'} \frac{\sigma^\beta}{n_{\pi^{-1}(\beta)} m_\beta}}{\sqrt{\left(\frac{\sigma^\beta}{n_{\pi^{-1}(\beta)} m_\beta}\right)^2 + \left(\frac{\sigma^{\beta'}}{n_{\pi^{-1}(\beta')} m_{\beta'}}\right)^2}}$$

$$\Lambda_{\alpha,\beta} = \Bigg(T_\alpha + U^\beta +$$

$$\frac{1}{n_\alpha m_\beta} \frac{R_{\alpha,\pi^{-1}(\beta)} \frac{\sigma^\beta}{n_{\pi^{-1}(\beta)} m_\beta} + S_{\beta,\pi(\alpha)} \frac{\sigma_\alpha}{n_\alpha m_{\pi(\alpha)}}}{\sqrt{\left(\frac{\sigma_\alpha}{n_\alpha m_{\pi(\alpha)}}\right)^2 + \left(\frac{\sigma^{\beta'}}{n_{\pi^{-1}(\beta')} m_{\beta'}}\right)^2}} \Bigg)^{-1} \tag{8.5}$$

Then, we have the following more general result:

Theorem 9.

1. *Suppose that $\tilde{D}_{\alpha,\alpha'} = D_{\alpha,\pi(\alpha')}$ satisfies the strong cyclical monotonicity condition, where for each simple loop $i_1 \to i_2 \to \ldots \to i_k \to i_{k+1} = i_1$ with length $k > 1$ we have*

$$\sum_{l=1}^{k} \tilde{D}_{i_l i_{l+1}} \geq \sum_{l=1}^{k} \tilde{D}_{i_l i_l} + k\delta. \tag{8.6}$$

   Then, the solution $X_{\alpha,\beta}$ of the characteristic optimization in equation 8.2 satisfies one of the following:

   (a) If $\boldsymbol{\delta} = 0$ and $\theta = \infty$, we have $X_{\alpha,\beta} = \delta_{\beta,\pi(\alpha)} \frac{\sigma_\alpha}{m_\alpha n_\beta}$ with $\delta_{\cdot,\cdot}$ being the Kronecker index, if

$$\lambda \leq \delta \max_{\alpha \neq \beta} \Lambda_{\alpha,\beta}$$

   (b) Otherwise,

$$\delta \sum_{\beta \neq \pi(\alpha)} X_{\alpha,\beta} \leq$$

$$\lambda \sum_{\alpha \neq \alpha'} \Bigg(\frac{R_{\alpha,\alpha'}}{n_\alpha n_{\alpha'}} \sqrt{\frac{n_{\alpha'}^2 \sigma_\alpha^2}{m_{\pi(\alpha)}} + \frac{n_\alpha^2 \sigma_{\alpha'}^2}{m_{\pi(\alpha')}}}$$

$$+ \frac{S_{\pi(\alpha),\pi(\alpha')}}{m_{\pi(\alpha)} m_{\pi(\alpha')}} \sqrt{\frac{m_{\pi(\alpha')}^2 \sigma_\alpha^2}{n_\alpha} + \frac{m_{\pi(\alpha)}^2 \sigma_{\alpha'}^2}{n_{\alpha'}}} \Bigg)$$

$$+ \frac{\theta}{2} \left(\sum_\alpha \frac{\delta_\alpha^2}{n_\alpha} + \sum_\alpha \frac{\delta_\alpha^2}{m_{\pi(\alpha)}} \right) + \frac{\Delta_1^2 n}{\theta} + \Delta_0 \left(\|\boldsymbol{\delta}\|_1 - \|\boldsymbol{\delta}\|_\infty \right)$$

   where

$$\Delta_0 = \max_{\alpha,\alpha'} \left| 2\tilde{D}_{\alpha,\alpha'} - \tilde{D}_{\alpha,\alpha} - \tilde{D}_{\alpha',\alpha'} \right|,$$

$$\Delta_1 = \frac{\Delta_0 + \max_\alpha |\tilde{D}_{\alpha,\alpha}|}{2}$$

2. *The solution of equation 8.1 is given by $X_{ij} = X_{\alpha,\beta}$ if there exist positive constants a, c, d such that $2a + c + d \leq 1$ and for all $i, i' \in C_\alpha$ and $j, j' \in D_\beta$,*

$$\sqrt{\sum_{j \in [m]} \left(D_{ij} - D_{i'j}\right)^2} \leq 2a n_\alpha \lambda R_{i,i'},$$

$$\sqrt{\sum_{i\in[n]}\left(D_{ij}-D_{ij'}\right)^2}\leq 2am_\beta\lambda S_{j,j'}$$

$$|\mu_i-\mu_{i'}|\leq\frac{c\lambda n_\alpha R_{i,i'}}{\theta\sqrt{m}},\quad |\nu_j-\nu_{j'}|\leq\frac{c\lambda m_\beta S_{j,j'}}{\theta\sqrt{n}}$$

$$\sqrt{\left(\sum_{\alpha'\neq\alpha}\frac{R_{i,\alpha'}-R_{i',\alpha}}{\sqrt{m_\alpha+m_{\alpha'}}}\right)^2+\sum_{\alpha'\neq\alpha}\left(\frac{R_{i,\alpha'}-R_{i',\alpha}}{\sqrt{m_\alpha+m_{\alpha'}}}\right)^2}\leq$$
$$dn_\alpha R_{i,i'}$$

$$\sqrt{\left(\sum_{\beta'\neq\beta}\frac{S_{j,\beta'}-S_{j',\beta}}{\sqrt{n_\beta+n_{\beta'}}}\right)^2+\sum_{\alpha'\neq\alpha}\left(\frac{S_{j,\beta'}-S_{j',\beta}}{\sqrt{n_\beta+n_{\beta'}}}\right)^2}\leq$$
$$dm_\beta S_{j,j'}$$

Proof. Denote the optimal value of equation 8.3 and equation 8.2 by C_0 and C_1, respectively. Also, notice that since $\tilde{D}_{\alpha,\alpha'}$ satisfies the strong cyclical monotonicity condition, $Y_{\alpha,\beta}=\delta_{\beta,\pi(\alpha)}\sigma_\alpha$ is the solution of equation 8.3 and there exist dual variables p_α, q_β such that

$$D_{\alpha,\beta}-p_\alpha-q_\beta\begin{cases}=0 & \beta=\pi(\alpha)\\ \geq\delta & \beta\neq\pi(\alpha)\end{cases}$$

Moreover,

$$C_0=\sum_\alpha\sigma_\alpha p_\alpha+\sum_\beta\sigma^\beta q_\beta$$

For part 1.a, we note that under the given conditions, the solution $X_{\alpha,\beta}$ of equation 8.2 coincides with that of equation 8.4. Now, we show that $X'_{\alpha,\beta}=\frac{Y_{\alpha,\beta}}{n_\alpha m_\beta}=\frac{\delta_{\beta,\pi(\alpha)}\sigma_\alpha}{n_\alpha m_\beta}$ satisfies with the dual parameters p'_α, q'_β, the optimality condition of equation 8.4, which can be written as

$$(D_{\alpha,\beta}-p'_\alpha-q'_\beta)n_\alpha m_\beta+\lambda A_{\alpha,\beta}\begin{cases}=0 & \beta=\pi(\alpha)\\ \geq 0 & \beta\neq\pi(\alpha)\end{cases}$$

where $A_{\alpha,\beta}$ is the partial derivative at $X'_{\alpha,\beta}$ of the SON term w.r.t $X_{\alpha,\beta}$. By direct calculation, we observe that

$$A_{\alpha,\beta}=\begin{cases}n_\alpha m_\beta(T_\alpha+U^\beta) & \beta=\pi(\alpha)\\ -\frac{R_{\alpha,\pi^{-1}(\beta)}\frac{\sigma^\beta}{n_{\pi^{-1}(\beta)}m_\beta}+S_{\beta,\pi(\alpha)}\frac{\sigma_\alpha}{n_\alpha m_{\pi(\alpha)}}}{\sqrt{\left(\frac{\sigma_\alpha}{n_\alpha m_{\pi(\alpha)}}\right)^2+\left(\frac{\sigma^{\beta'}}{n_{\pi^{-1}(\beta')}m_{\beta'}}\right)^2}} & \beta\neq\pi(\alpha)\end{cases}$$

It is now simple to check that under the given assumption, taking $p'_\alpha=p_\alpha+\lambda T_\alpha$ and $q'_\beta=q_\beta+\lambda U^\beta$ will satisfy the optimality conditions.

For part 1.b, we note that for the solution $X_{\alpha,\beta}$ of equation 8.2,

$$C_1=F\left(\{X_{\alpha,\beta}\}\right)\geq\sum_{\alpha,\beta}n_\alpha m_\beta X_{\alpha,\beta}D_{\alpha,\beta}+$$

$$\frac{\theta}{2}\left(\sum_\alpha n_\alpha\left(\mathbf{a}_M^T\mathbf{x}_\alpha-\mu_\alpha\right)^2+\sum_\beta m_\beta\left(\mathbf{a}_N^T\mathbf{x}^\beta-\nu^\beta\right)^2\right)$$

$$=\sum_{\alpha,\beta}n_\alpha m_\beta X_{\alpha,\beta}\left(D_{\alpha,\beta}-p_\alpha-q_\beta\right)+\sum_\alpha p_\alpha\sigma_\alpha+\sum_\beta\sigma^\beta q_\beta$$

$$+\sum_\alpha(\mathbf{a}_M^T\mathbf{x}_\alpha-\mu_\alpha)p_\alpha n_\alpha+\sum_\beta(\mathbf{a}_N^T\mathbf{x}^\beta-\nu^\beta)q_\beta m_\beta+$$

$$\sum_{\alpha}(\mu_\alpha n_\alpha - \sigma_\alpha)p_\alpha + \sum_{\beta}(\nu^\beta m_\beta - \sigma^\beta)q_\beta$$

$$+\frac{\theta}{2}\left(\sum_{\alpha} n_\alpha\left(\mathbf{a}_M^T\mathbf{x}_\alpha - \mu_\alpha\right)^2 + \sum_{\beta} m_\beta\left(\mathbf{a}_N^T\mathbf{x}^\beta - \nu^\beta\right)^2\right)$$

$$\geq \delta\sum_{\beta\neq\pi(\alpha)} X_{\alpha,\beta} + C_0 + \sum_{\alpha} p_\alpha\delta_\alpha + \sum_{\beta}\delta^\beta q_\beta$$

$$-\frac{1}{2\theta}\left(\sum_{\alpha} p_\alpha^2 n_\alpha + \sum_{\beta} q_\beta^2 m_\beta\right),$$

where $F(.)$ denotes the objective function in equation 8.2. On the other hand for $X'_{\alpha,\beta} = \frac{Y_{\alpha,\beta}}{n_\alpha m_\beta} = \frac{\delta_{\beta,\pi(\alpha)}\sigma_\alpha}{n_\alpha m_\beta}$, we have that

$$C_1 \leq F(\{X'_{\alpha,\beta}\}) = C_0 +$$

$$\lambda\sum_{\alpha\neq\alpha'}\Bigg(\frac{R_{\alpha,\alpha'}}{n_\alpha n_{\alpha'}}\sqrt{\frac{n_{\alpha'}^2\sigma_\alpha^2}{m_{\pi(\alpha)}} + \frac{n_\alpha^2\sigma_{\alpha'}^2}{m_{\pi(\alpha')}}} +$$

$$\frac{S_{\pi(\alpha),\pi(\alpha')}}{m_{\pi(\alpha)}m_{\pi(\alpha')}}\sqrt{\frac{m_{\pi(\alpha')}^2\sigma_\alpha^2}{n_\alpha} + \frac{m_{\pi(\alpha)}^2\sigma_{\alpha'}^2}{n_{\alpha'}}}\Bigg)$$

$$+\frac{\theta}{2}\left(\sum_{\alpha}\frac{\delta_\alpha^2}{n_\alpha} + \sum_{\alpha}\frac{\delta_\alpha^2}{m_{\pi(\alpha)}}\right)$$

We conclude that

$$\delta\sum_{\beta\neq\pi(\alpha)} X_{\alpha,\beta} \leq$$

$$\lambda\sum_{\alpha\neq\alpha'}\Bigg(\frac{R_{\alpha,\alpha'}}{n_\alpha n_{\alpha'}}\sqrt{\frac{n_{\alpha'}^2\sigma_\alpha^2}{m_{\pi(\alpha)}} + \frac{n_\alpha^2\sigma_{\alpha'}^2}{m_{\pi(\alpha')}}}$$

$$+\frac{S_{\pi(\alpha),\pi(\alpha')}}{m_{\pi(\alpha)}m_{\pi(\alpha')}}\sqrt{\frac{m_{\pi(\alpha')}^2\sigma_\alpha^2}{n_\alpha} + \frac{m_{\pi(\alpha)}^2\sigma_{\alpha'}^2}{n_{\alpha'}}}\Bigg)$$

$$+\frac{\theta}{2}\left(\sum_{\alpha}\frac{\delta_\alpha^2}{n_\alpha} + \sum_{\alpha}\frac{\delta_\alpha^2}{m_{\pi(\alpha)}}\right) + \frac{1}{2\theta}\left(\sum_{\alpha} p_\alpha^2 n_\alpha + \sum_{\beta} q_\beta^2 m_\beta\right) -$$

$$\sum_{\alpha} p_\alpha\delta_\alpha - \sum_{\beta}\delta^\beta q_\beta$$

Lemma 10 gives the result in part 1.

For part 2, notice that the optimality condition of $X_{\alpha,\beta}$ yields

$$n_\alpha m_\beta D_{\alpha,\beta} + \lambda\sum_{\alpha'\neq\alpha} R_{\alpha,\alpha'} m_\beta(\mathbf{z}_{\alpha,\alpha'})_\beta + \lambda\sum_{\beta'\neq\beta} S_{\beta,\beta'} n_\alpha(\mathbf{z}^{\beta,\beta'})_\alpha$$

$$+\theta n_\alpha m_\beta(\mathbf{a}_M^T\mathbf{x}_\alpha - \mu_\alpha) + \theta m_\beta n_\alpha(\mathbf{a}_N^T\mathbf{x}^\beta - \nu^\beta) = 0$$

where

$$\mathbf{z}_{\alpha,\alpha'} = \frac{\mathbf{x}_\alpha - \mathbf{x}_{\alpha'}}{\|\mathbf{x}_\alpha - \mathbf{x}_{\alpha'}\|_M},\ \mathbf{z}^{\beta,\beta'} = \frac{\mathbf{x}^\beta - \mathbf{x}^{\beta'}}{\|\mathbf{x}^\beta - \mathbf{x}^{\beta'}\|_N}$$

Define for $i, i' \in C_\alpha$ and $j, j' \in D_\beta$

$$(\mathbf{z}_{i,i'})_j = \frac{1}{2\lambda n_\alpha R_{i,i'}}\Bigg(-D_{ij} + D_{i'j} - \frac{\sum\limits_{j'\in D_\beta} D_{ij'}}{m_\beta} + \frac{\sum\limits_{j'\in D_\beta} D_{i'j'}}{m_\beta} - 2\theta\mu_i + 2\theta\mu_{i'}\Bigg)$$

$$-\frac{1}{n_\alpha R_{i,i'}}\sum_{\alpha'\neq\alpha}\left(R_{i,\alpha'} - R_{i',\alpha'}\right)(\mathbf{z}_{\alpha,\alpha'})_\beta$$

$$(\mathbf{z}^{j,j'})_i = \frac{1}{2\lambda m_\beta S_{j,j'}}\Bigg(-D_{ij} + D_{ij'} - \frac{\sum\limits_{i'\in C_\alpha} D_{i'j}}{n_\alpha} + \frac{\sum\limits_{i'\in C_\alpha} D_{i'j'}}{n_\alpha} - 2\theta\nu_j + 2\theta\nu_{j'}\Bigg)$$

$$-\frac{1}{m_\beta S_{j,j'}}\sum_{\beta'\neq\beta}\left(S_{j,\beta'} - S_{j',\beta'}\right)(\mathbf{z}^{\beta,\beta'})_\alpha$$

Also for $i \in C_\alpha, i' \in C_{\alpha'}$ and $j \in D_\beta, j' \in D_{\beta'}$, where $\alpha \neq \alpha'$ and $\beta \neq \beta'$, take $(\mathbf{z}_{ii'})_j = (\mathbf{z}_{\alpha,\alpha'})_\beta$, $(\mathbf{z}^{jj'})_i = (\mathbf{z}^{\beta,\beta'})_\alpha$. Then, it simple to check that $X_{ij} = X_{\alpha,\beta}$ satisfies the optimality conditions of equation 8.1 under conditions of the theorem and noticing that by the root-means-square and arithmetic mean (RMS-AM) inequality, we also have

$$\sqrt{\sum_{\beta\in[K]} m_\beta \left(\frac{\sum\limits_{j\in D_\beta}\left(D_{ij} - D_{i',j}\right)}{m_\beta}\right)^2} \leq 2a\lambda n_\alpha R_{i,i'}$$

$$\sqrt{\sum_{\alpha\in[K]} n_\alpha \left(\frac{\sum\limits_{i\in C_\alpha}\left(D_{ij} - D_{ij'}\right)}{n_\alpha}\right)^2} \leq 2a\lambda m_\beta S_{j,j'}$$

□

Lemma 10. *Suppose that the ideal optimization in equation 8.3 has a solution where $X_{\alpha,\pi(\alpha)} > 0$ holds for every α. For every $\boldsymbol{\delta} = (\delta_\alpha)_\alpha$ satisfying $\mathbf{1}^T\boldsymbol{\delta} = 0$ and any choice of the optimal dual parameters $\{p_\alpha, q_\beta\}$ we have that*

$$\sum_\alpha p_\alpha\delta_\alpha + \sum_\beta q_\beta\delta^\beta \leq \Delta_0\left(\|\boldsymbol{\delta}\|_1 - \|\boldsymbol{\delta}\|_\infty\right)$$

where $\delta^\beta = -\delta_{\pi^{-1}(\beta)}$. As a result in this case, equation 8.3 has optimal dual parameters $\{p_\alpha, q_\beta\}$ satisfying

$$|p_\alpha| \leq \Delta_1,\ |q_\beta| \leq \Delta_1$$

Proof. Denote the minimum value of $X_{\alpha,\pi(\alpha)}$ by ϵ. Without loss of generality, we assume that $\|\boldsymbol{\delta}\|_1 - \|\boldsymbol{\delta}\|_\infty \leq \epsilon$. Take $\alpha_0 \in \arg\min\limits_\alpha |\delta_\alpha|$. Hence, $\|\boldsymbol{\delta}\|_1 - \|\boldsymbol{\delta}\|_\infty = \sum\limits_{\alpha\neq\alpha_0}|\delta_\alpha|$.

Denote the optimal value of equation 8.3 by C_0. From the strong duality theorem we have that

$$C_0 = \sum_\alpha p_\alpha\sigma_\alpha + \sum_\beta q_\beta\sigma^\beta$$

Take

$$
\begin{aligned}
C_1 = \min_{Y_{\alpha,\beta} \geq 0} \sum_{\alpha,\beta} Y_{\alpha,\beta} D_{\alpha,\beta} \\
\text{s.t} \\
\mathbf{1}^T \mathbf{y}^\beta = \sigma^\beta + \delta^\beta, \mathbf{1}^T \mathbf{y}_\alpha = \sigma_\alpha + \delta_\alpha
\end{aligned} \tag{8.7}
$$

We notice that $\{p_\alpha, q_\beta\}$ are feasible dual vectors for equation 8.7. Hence, from the weak duality theorem we have

$$
\begin{aligned}
C_1 &\geq \sum_\alpha p_\alpha(\sigma_\alpha + \delta_\alpha) + \sum_\beta q_\beta(\sigma^\beta + \delta^\beta) \\
&= C_0 + \sum_\alpha p_\alpha \delta_\alpha + \sum_\beta q_\beta \delta^\beta
\end{aligned}
$$

Now take the solution

$$
Y'_{\alpha,\beta} = Y_{\alpha,\beta} \begin{cases} -|\delta_\alpha| & \alpha \neq \alpha_0,\ \beta = \pi(\alpha) \\ -\sum\limits_{\alpha \neq \alpha_0} |\delta_\alpha| & \alpha = \alpha_0,\ \beta = \pi(\alpha_0) \\ +(\delta^\beta)_+ & \alpha = \alpha_0,\ \beta \neq \pi(\alpha_0) \\ +(\delta_\alpha)_+ & \alpha \neq \alpha_0,\ \beta = \pi(\alpha_0) \\ +0 & \text{Otherwise} \end{cases}
$$

It is simple to check that $Y'_{\alpha,\beta}$ is feasible in equation 8.7. Moreover, we have

$$
\begin{aligned}
C_1 \leq \sum_{\alpha,\beta} Y'_{\alpha,\beta} D_{\alpha,\beta} = C_0 + \\
\sum_{\alpha \neq \alpha_0} \big(2D_{\alpha_0 \pi(\alpha)}(\delta_\alpha)_+ + 2D_{\alpha\alpha_0}(\delta_\alpha)_- - \\
(D_{\alpha,\alpha} + D_{\alpha_0,\alpha_0})|\delta_\alpha|\big) \\
\leq C_0 + \Delta_0 \sum_{\alpha \neq \alpha_0} |\delta_\alpha|
\end{aligned}
$$

We conclude that

$$
\sum_\alpha p_\alpha \delta_\alpha + \sum_\beta q_\beta \delta^\beta \leq \Delta_0 \sum_{\alpha \neq \alpha_0} |\delta_\alpha|
$$

which proves the first part. Now, notice that for any pair (α_1, α_2) of distinct indices, taking $\delta_{\alpha_1} = 1$ and $\delta_{\alpha_1} = -1$ gives

$$
p_{\alpha_1} - p_{\alpha_2} - q_{\alpha_1} + q_{\alpha_2} \leq \Delta_0
$$

switching α_1, α_2 yield

$$
|p_{\alpha_1} - p_{\alpha_2} - q_{\alpha_1} + q_{\alpha_2}| \leq \Delta_0
$$

Now, notice that from the optimality of equation 8.3 we have $p_\alpha + q_\alpha = D_{\alpha,\alpha}$, which leads to

$$
2|p_{\alpha_1} - p_{\alpha_2}| \leq \Delta_0 + |D_{\alpha_1,\alpha_1} - D_{\alpha_2,\alpha_2}|
$$

which yield

$$
\left|\left(p_{\alpha_1} + \frac{D_{\alpha_1,\alpha_1}}{2}\right) - \left(p_{\alpha_2} + \frac{D_{\alpha_2,\alpha_2}}{2}\right)\right| \leq \Delta_0
$$

The result is obtained by noticing that the set of optimal dual solutions is invariant under shift, i.e. $p_i + \lambda$ and $q_i - \lambda$ are also solutions for any $\lambda \in \mathbb{R}$. Hence, we may take λ such that

$$
\left|p_\alpha + \frac{D_{\alpha,\alpha}}{2}\right| \leq \frac{\Delta_0}{2}
$$

and hence

$$
\left|q_\alpha - \frac{D_{\alpha,\alpha}}{2}\right| \leq \frac{\Delta_0}{2}
$$

Triangle inequality gives the result.

□

9 Proof of Theorem 8

The first claim that $X_{ij} = X_{\alpha,\beta}$ follows by specializing part 2 of Theorem 9 for the conditions of Theorem 8 with $a = 1/2, b = c = 0$: $\theta = \infty$ and inside clusters we have $R_{i,i'} = S_{j,j'=1} = 1$, $R_{i,\alpha'} = R_{i',\alpha}$ and $S_{j,\beta'} = R_{j',\beta}$. Moreover $n_\alpha = m_\beta = m$.

Part 1 in the main text also is achieved by specializing part 1.a.: We will have $\sigma_\alpha = \omega_\alpha$, $n_\alpha = m_\beta = m$, $R_{\alpha,\alpha'} = m^2 R$ and $S_{\beta,\beta'} = m^2$.

Finally, part 2 is a result of 1.b. with $\theta = \infty, \boldsymbol{\delta} = \mathbf{0}$.

10 Proof of Theorem 4

To simplify the notation, we introduce $\phi_{P+q} = I_{S_q}$ for $q = 1, 2 \ldots, Q$, where I_S denotes the indicator function of a convex set S. It is well-known that the proximal operator of I_S coincides with the orthogonal projection operator onto S. Hence, we may simplify our algorithm to

$$\mathbf{x}_{t+1} = \operatorname{prox}\left(\mathbf{x}_t + \mu \mathbf{g}_{r_t}\right), \quad \mathbf{a}_t = \rho \frac{\mathbf{x}_t - \mathbf{x}_{t+1}}{\mu} - \alpha \bar{\mathbf{g}}_t,$$
$$\mathbf{g}_{r_t} \leftarrow \mathbf{g}_{r_t} + \mathbf{a}_t,$$

where we introduce $\mathbf{g}_{P+q} = \mathbf{h}_q$ for $q = 1, 2 \ldots, Q$ and denote $\bar{\mathbf{g}}_t = \sum\limits_{r=0}^{R} \mathbf{g}_r$ with $R = P + Q$. Moreover, r_t is equal to either p_t or $P + q_t$, depending on the random choice. We also define $\mathbf{x}^\dagger_{r,t} = \operatorname{prox}\left(\mathbf{x}_t + \mu \mathbf{g}_r\right)$ and hence $\mathbf{x}_{t+1} = \mathbf{x}^\dagger_{r,t}$.

To prove convergence, we adopt a so-called Lyaponov function approach. We introduce two non-negative functions L, M of the state variables, $\mathbf{x}$ and $\{\mathbf{g}_r\}$ such that

$$\mathbb{E}[L_{t+1}] - \mathbb{E}[L_t] + \mathbb{E}[M_t] \leq 0, \quad t = 1, 2, \ldots, \tag{10.1}$$

where L_t, M_t denote the values of L, M at the variables of the t^{th} iteration. Then, summing these inequalities up to an arbitrary time t gives

$$\mathbb{E}[L_t] - L_0 + \sum_{\tau=0}^{t-1} \mathbb{E}[M_\tau] \leq 0 \tag{10.2}$$

which by the non-negativity of L implies

$$\sum_{\tau=0}^{t-1} \mathbb{E}[M_\tau] \leq L_0. \tag{10.3}$$

In particular, we take

$$\begin{aligned} M_t = F_t &+ \tfrac{1-\rho}{2\mu(1+\rho)} \sum_r \left\|\mathbf{x}_t - \mathbf{x}^\dagger_{r,t}\right\|^2 \\ &+ \tfrac{\mu}{\rho}\left[\tfrac{2+\alpha}{2} - \tfrac{\alpha^2}{1-\rho}\right] \|\bar{\mathbf{g}}_t\|^2 \end{aligned} \tag{10.4}$$

where

$$F_t = \sum_{r=1}^{P} \left[\phi_r\left(\mathbf{x}^\dagger_{r,t}\right) - \phi_r(\mathbf{x}^*) - \langle \mathbf{g}^*_r, \mathbf{x}^\dagger_{r,t} - \mathbf{x}^* \rangle\right] \tag{10.5}$$

and $\mathbf{g}^*_r \in \partial\phi_r(\mathbf{x}^*)$ for $r \in [R]$ satisfy the monotone inclusion problem in equation 3.6 at $\mathbf{x}^*$, i.e $\sum\limits_r \mathbf{g}^*_r = 0$. Then, the non-negativity of each summand of F_t follows from the convexity of ϕ_r. The third term of M_t is also positive for $\alpha < \frac{1+\sqrt{17}}{4}(1-\rho)$. This establishes the non-negativity of M_t. We further define

$$\begin{aligned} \Gamma_t &= \sum_{r=1}^{R} \|\mathbf{g}_r - \mathbf{g}^*_r\|^2, \quad G_t = \|\mu\bar{\mathbf{g}}_t + \rho(\mathbf{x}_t - \mathbf{x}^*)\|^2, \\ D_t &= \|\mathbf{x}^* - \mathbf{x}_t\|^2 \end{aligned} \tag{10.6}$$

and take

$$L_t = \frac{R}{2\mu} D_t + \frac{1}{2\rho\mu\alpha} G_t + \frac{R\mu}{2\rho} \Gamma_t \tag{10.7}$$

10.1 Proof of Theorem Under equation 10.1

Let us first prove the theorem assuming equation 10.1 holds true for the given L, M. Then equation 10.3 also holds true and we conclude from the definition of M that

$$\begin{aligned} &\frac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{E}[F_\tau] \le \frac{L_0}{t}, \\ &\frac{1-\rho}{2t(1+\rho)}\sum_{\tau=0}^{t-1}\sum_{r}\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{x}^\dagger_{r,t}\right\|^2\right] \le \frac{\mu L_0}{t} \end{aligned} \tag{10.8}$$

Now from equation 10.5 and the triangle inequality, we conclude that

$$\begin{aligned} \mathbb{E}[F_\tau] \ge \mathbb{E}\left[\sum_{r=1}^{P}\phi_r(\mathbf{x}_\tau) - \phi_r(\mathbf{x}^*)\right] \\ -\sum_{r=1}^{P}\mathbb{E}\left|\phi_r\left(\mathbf{x}^\dagger_{r,\tau}\right) - \phi_r(\mathbf{x}_\tau)\right| - \\ \sum_{r=1}^{P}\mathbb{E}\left|\langle\mathbf{g}^*_r, \mathbf{x}^\dagger_{r,\tau} - \mathbf{x}_\tau\rangle\right| \end{aligned} \tag{10.9}$$

Further since the functions are $\beta-$Lipschitz, we observe that $\|\mathbf{g}^*_r\| \le \beta$. Hence,

$$\begin{aligned} &\sum_{r=1}^{P}\mathbb{E}\left|\langle\mathbf{g}^*_r, \mathbf{x}^\dagger_{r,\tau} - \mathbf{x}_\tau\rangle\right| \le \beta\sum_{r=1}^{P}\mathbb{E}\left\|\mathbf{x}^\dagger_{r,\tau} - \mathbf{x}_\tau\right\| \\ &\le \beta\sqrt{P}\sqrt{\sum_{r=1}^{P}\mathbb{E}\left\|\mathbf{x}^\dagger_{r,\tau} - \mathbf{x}_\tau\right\|^2} \end{aligned} \tag{10.10}$$

Similarly,

$$\begin{aligned} &\sum_{r=1}^{P}\mathbb{E}\left|\phi_r\left(\mathbf{x}^\dagger_{r,\tau}\right) - \phi_r(\mathbf{x}_\tau)\right| \le \beta\sum_{r=1}^{P}\mathbb{E}\left\|\mathbf{x}^\dagger_{r,\tau} - \mathbf{x}_\tau\right\| \\ &\le \beta\sqrt{P}\sqrt{\sum_{r=1}^{P}\mathbb{E}\left\|\mathbf{x}^\dagger_{r,\tau} - \mathbf{x}_\tau\right\|^2} \end{aligned} \tag{10.11}$$

We conclude that

$$\begin{aligned} \mathbb{E}\left[\sum_{r=1}^{P}\phi_r(\mathbf{x}_\tau) - \phi_r(\mathbf{x}^*)\right] \\ \le \mathbb{E}[F_\tau] + 2\beta\sqrt{P}\sqrt{\sum_{r=1}^{P}\mathbb{E}\left\|\mathbf{x}^\dagger_{r,\tau} - \mathbf{x}_\tau\right\|^2} \end{aligned} \tag{10.12}$$

and hence,

$$\begin{aligned} \tfrac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{E}\left[\sum_{r=1}^{P}\phi_r(\mathbf{x}_\tau) - \phi_r(\mathbf{x}^*)\right] \\ \le \tfrac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{E}[F_\tau] + \tfrac{2\beta\sqrt{P}}{t}\sum_{\tau=0}^{t-1}\sqrt{\sum_{r=1}^{P}\mathbb{E}\left\|\mathbf{x}^\dagger_{r,\tau} - \mathbf{x}_\tau\right\|^2} \\ \le \tfrac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{E}[F_\tau] + \\ 2\beta\sqrt{P}\sqrt{\tfrac{1}{t}\sum_{\tau=0}^{t-1}\sum_{r=1}^{P}\mathbb{E}\left\|\mathbf{x}^\dagger_{r,\tau} - \mathbf{x}_\tau\right\|^2} \end{aligned} \tag{10.13}$$

Using Jensen's inequality and equation 10.8, we obtain

$$\begin{aligned}&\mathbb{E}\left[\sum_{r=1}^{P}\phi_r(\bar{\mathbf{x}}_\tau)-\phi_r(\mathbf{x}^*)\right]\leq\\&\frac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{E}\left[\sum_{r=1}^{P}\phi_r(\mathbf{x}_\tau)-\phi_r(\mathbf{x}^*)\right]\leq\\&\frac{L_0}{t}+2\beta\sqrt{\frac{2P\mu L_0(1+\rho)}{t(1-\rho)}}\end{aligned}\tag{10.14}$$

This proves the first bound in part 1 noting that for a suitable constant c only depending on ρ,α and the constant in equation 3.7

$$L_0\leq c\beta P\lambda.\tag{10.15}$$

For the second bound in part 1 note that for $r=P+1,P+2,\dots,P+Q$, we have $\operatorname{dist}(\mathbf{x}_t,S_r)\leq\|\mathbf{x}_{r,t}^\dagger-\mathbf{x}_t\|$, since by definition $\mathbf{x}_{r,t}^\dagger\in S_r$. We conclude that

$$\sum_{q=1}^{Q}\operatorname{dist}^2(\bar{\mathbf{x}}_t,S_q)\leq\frac{1}{t}\sum_{\tau=0}^{t-1}\sum_{r=P+1}^{R}\|\mathbf{x}_{r,t}^\dagger-\mathbf{x}_t\|^2\leq\frac{\mu L_0}{t}\tag{10.16}$$

For part 2, note that

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2\right]=\mathbb{E}\left[\mathbb{E}\left[\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2\mid\mathbf{x}_t\right]\right]=\\&\mathbb{E}\left[\frac{1}{R}\sum_r\left\|\mathbf{x}_{r,t}^\dagger-\mathbf{x}_t\right\|^2\right]\end{aligned}$$

We conclude from equation 10.3 that

$$\sum_t\mathbb{E}\left[\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2\right]\leq\frac{\mu L_0}{R}\tag{10.17}$$

which completes the proof.

10.2 Proof of equation 10.1

It remains to prove equation 10.1. We first state the following intermediate results that characterize the term $\mathbb{E}[L_t]-\mathbb{E}[L_{t-1}]$ in equation 10.1. They can be proven by direct substitution and hence the proofs are neglected.

Lemma 11. *The average dynamics of G_t is given by:*

$$\begin{aligned}\mathbb{E}[G_{t+1}]-\mathbb{E}[G_t]=&-\mu^2\left[1-(1-\alpha)^2\right]\mathbb{E}\|\bar{\mathbf{g}}_t\|^2\\&+2\rho\mu\alpha\mathbb{E}\left\langle\mathbf{x}^*-\mathbf{x}_t,\bar{\mathbf{g}}_t\right\rangle\end{aligned}\tag{10.18}$$

Lemma 12. *The average dynamics of Γ_t is given by:*

$$\begin{aligned}\mathbb{E}[\Gamma_{t+1}]-\mathbb{E}[\Gamma_t]=&\frac{1}{R}\sum_r\mathbb{E}\left\|\rho\frac{\mathbf{x}_t-\mathbf{x}_{r,t}^\dagger}{\mu}-\alpha\bar{\mathbf{g}}_t\right\|^2\\&+\frac{2\rho}{R\mu}\sum_r\mathbb{E}\left\langle\mathbf{x}_t-\mathbf{x}_{r,t}^\dagger,\mathbf{g}_r-\mathbf{g}_r^*\right\rangle-\frac{2\alpha}{R}\|\bar{\mathbf{g}}_t\|^2\end{aligned}\tag{10.19}$$

Lemma 13. *The average dynamics of D_t is given by:*

$$\begin{aligned}\mathbb{E}[D_{t+1}]-\mathbb{E}[D_t]=&-\frac{1}{R}\sum_r\mathbb{E}\|\mathbf{x}_{r,t}^\dagger-\mathbf{x}_t\|^2\\&+\frac{2}{R}\sum_r\mathbb{E}\langle\mathbf{x}_t-\mathbf{x}_{r,t}^\dagger,\mathbf{x}^*-\mathbf{x}_{r,t}^\dagger\rangle\end{aligned}\tag{10.20}$$

Now, we state a crucial inequality that connects the dynamics of L to M:

Lemma 14. *The following inequality holds at every time:*

$$F_t + \sum_{r=1}^{R} \left\langle \frac{\mathbf{x}_t - \mathbf{x}_{r,t}^{\dagger}}{\mu} + \mathbf{g}_r - \mathbf{g}_r^*, \mathbf{x}^* - \mathbf{x}_{r,t}^{\dagger} \right\rangle \leq 0 \tag{10.21}$$

Proof. From the definition of a proximal operator for $r = 1, 2, \ldots, P$, we have $\frac{\mathbf{x}_t - \bar{\mathbf{x}}_{r,t}^{\dagger}}{\mu} + \mathbf{g}_r \in \partial\phi_r(\bar{\mathbf{x}}_{r,t}^{\dagger})$. hence

$$\phi_r(\mathbf{x}^*) \geq \phi_r(\bar{\mathbf{x}}_{r,t}^{\dagger}) + \left\langle \frac{\mathbf{x}_t - \bar{\mathbf{x}}_{r,t}^{\dagger}}{\mu} + \mathbf{g}_r, \mathbf{x}^* - \bar{\mathbf{x}}_{r,t}^{\dagger} \right\rangle \tag{10.22}$$

Adding and subtracting the term $\langle \mathbf{g}_r^*, \bar{\mathbf{x}}_{r,t}^{\dagger} - \mathbf{x}_t \rangle$ and summing over $r \in [P]$ gives

$$F_t + \sum_{r=1}^{P} \left\langle \frac{\mathbf{x}_t - \mathbf{x}_{r,t}^{\dagger}}{\mu} + \mathbf{g}_r - \mathbf{g}_r^*, \mathbf{x}^* - \mathbf{x}_{r,t}^{\dagger} \right\rangle \leq 0 \tag{10.23}$$

Now, note that by the definition of a projection operator for $r = P+1, P+2, \ldots, R$ we have $\frac{\mathbf{x}_t - \bar{\mathbf{x}}_{r,t}^{\dagger}}{\mu} + \mathbf{g}_r$ is normal to S_r at $\bar{\mathbf{x}}_{r,t}^{\dagger}$. Since $\mathbf{x}^* \in S_r$, we have

$$\left\langle \frac{\mathbf{x}_t - \mathbf{x}_{r,t}^{\dagger}}{\mu} + \mathbf{g}_r, \mathbf{x}^* - \mathbf{x}_{r,t}^{\dagger} \right\rangle \leq 0 \tag{10.24}$$

Similarly, we obtain that

$$\langle \mathbf{g}_r^*, \bar{\mathbf{x}}_{r,t}^{\dagger} - \mathbf{x}_t \rangle \leq 0 \tag{10.25}$$

Summing (44) and (45) over $r = P+1, P+2, \ldots, R$ and adding to (43) gives the desired result. □

10.2.1 Combining Bounds

To obtain equation 10.1, we combine the inequalities in the above four lemmas in the following way. Respectively multiplying equation 10.20, equation 10.18 and equation 10.19 by $\frac{R}{2\mu}$, $\frac{1}{2\rho\mu\alpha}$ and $\frac{R\mu}{2\rho}$ and adding to equation 10.21 and after straightforward calculations, we obtain:

$$\begin{gathered}\mathbb{E}[L_{t+1}] - \mathbb{E}[L_t] + \frac{1}{2\mu}\sum_r \left\| \mathbf{x}_t - \mathbf{x}_{r,t}^{\dagger} \right\|^2 + \\ \mu \left[\frac{1-(1-\alpha)^2}{2\alpha\rho} + \frac{\alpha}{\rho} \right] \|\bar{\mathbf{g}}_t\|^2 \\ -\frac{1}{2\rho\mu} \sum_{k\in[K]} \left\| \rho\left(\mathbf{x} - \mathbf{x}_k^{\dagger}\right) - \mu\alpha\mathbf{g} \right\|^2 \leq 0.\end{gathered}$$

By invoking Jensen's inequality, we have,

$$\left\| \rho\left(\mathbf{x} - \mathbf{x}_k^{\dagger}\right) - \mu\alpha\mathbf{g} \right\|^2 \leq \frac{2\rho^2}{1+\rho}\left\| \mathbf{x} - \mathbf{x}_k^{\dagger} \right\|^2 + \frac{2\mu^2\alpha^2}{1-\rho}\|\mathbf{g}\|^2,$$

which yields the desired result.

11 Additional experiments

11.1 Impact of SON-Regularizer

We investigate the models on a simple dataset, shown in Fig. 2. We illustrate the behavior of each model with respect to two different values of its regularization parameter (low and high) respectively at the first and the second row (low: $\lambda_1 = 0.01, \lambda_2 = 0.0$, high: $\lambda_1 = 10, \lambda_2 = 5$). In the low setting, instead of $\lambda_2 = 0.0$ any other small value also yields consistent results. The source data, target data and transported source data are respectively shown by yellow, blue and red points. Each column of the sub-figures in Fig. 2 corresponds to a particular model performance respectively OT-l1l2, OT-lpl1, OT-Sinkhorn and OT-SON (our model). We observe that OT-SON yields stable and consistent results for different values of its parameters. Moreover, the data points transported by the proposed model are always informative providing a good representation of the underlying classes. Whereas, the other OT models are sensitive to the values of their regularization parameters and might thus transport the source data to somewhere in the middle of the actual target data, or away from the actual classes in the target domain.

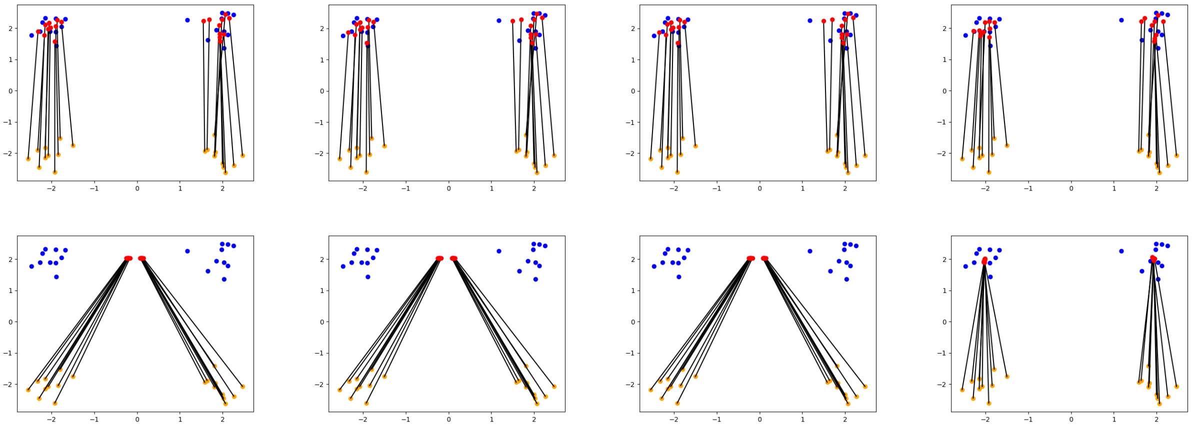


Figure 2: Illustration of different models on simple data, where the source and target domains have the same number of classes and similar distributions. The columns respectively correspond to OT-l1l2, OT-lpl1, OT-Sinkhorn and OT-SON. For each model, we illustrate the results for two different values of its regularization parameter. Among different models, OT-SON yields consistent, informative and stable transports for different regularization parameters.

We next study the interesting case where the source and target domains do not include the same number of classes, as shown in Fig. 3. In this experiment, we assume that the source data contains three classes, whereas the target domain has only two classes. Using the same color code as in Fig.2, we see in Fig. 3 the target classes and the transported source classes to the target domain are shown in yellow, blue and red respectively corresponding to OT-l1l2, OT-lpl1, OT-Sinkhorn and OT-SON. We again illustrate the behavior of each model w.r.t. two different values of its regularization parameter (low and high) respectively at the first and the second row. We observe that among all different models, only OT-SON with an appropriate parameter is able to identify that the source and the target domains have different number of classes, and subsequently, matches the corresponding classes correctly. It maps the superfluous class to a space between the two matched classes. However, the other models assign the superfluous class to the two other classes and do not distinguish the presence of such an extra class in the source domain. This observation is consistent with the assumptions made in [Courty et al., 2017]. The unbalanced method in [Chizat et al., 2018] might be relevant but its use is unclear to us. In the last column of Fig. 3, the heat maps show the mapping cost among different source and target classes, and as well as the transport map obtained by our algorithm (OT-SON with a high regularization). We observe that the transport map respects the class structure.

11.2 Experiments on path-based data

In Fig. 4, we investigate the different OT-based domain adaptation models on a commonly-used synthetic dataset, wherein the three classes have diverse shapes and forms [Chang and Yeung, 2008]. In particular, we consider the case where one of the source classes is absent in the target domain. With the same number of classes in the source and target domains, the different models perform equally well. Fig. 4(a) shows a case where the source data (yellow points) and the target data (blue points), differ in the fact that the target data is missing the upper left Gaussian cloud of points appearing in the source data. Fig. 4(b) shows the two source and target datasets, as well as the transported data by our model (OT-SON). The transported data points are shown in red. We observe that our method avoids mapping the source data of the missing class to any of the present classes in the target domain. The points with white interior are those not assigned to any class in the target. This thus leads to a better prediction of the target data. In the table of Fig. 4(c), we compare the accuracy scores of different models on the target data, where our model yields the highest score.

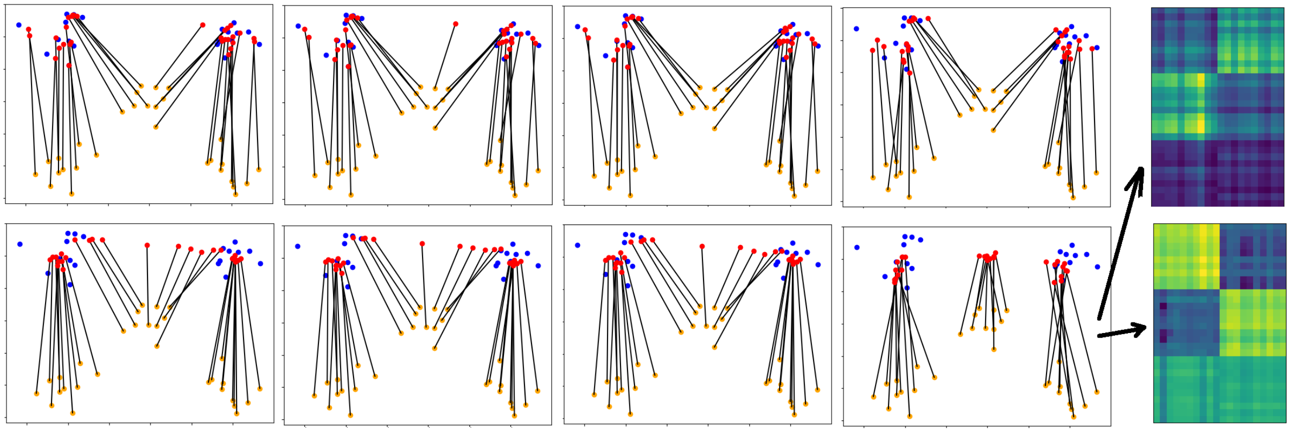
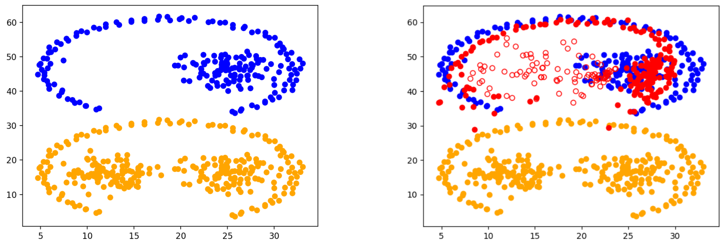


Figure 3: Illustration of different methods where the source and target domains have different number of classes. The first four columns respectively correspond to OT-l1l2, OT-lpl1, OT-Sinkhorn and OT-SON. Only OT-SON with an appropriate parameterization (the forth column and the second row) identifies the presence of a superfluous class in the source and handles it properly. The last column shows the consistency between the mapping costs and the transport map.



(a) Path-based data.

(b) Path-based transport.

model	acc.
OT-SON	0.852
OT-l1l2	0.788
OT-lpl1	0.759
OT-Sinkhorn	0.739
No reg	0.626

(c) Accuracy results.

Figure 4: Path-based source (yellow points) and target (blue points) datasets. Using OT-SON to transfer the path-based source data to the target domain (shown by red) yields the best results.

11.3 Unsupervised domain adaptation

In all prior experiments, we have assumed that the class labels of the source data are available. This setup is consistent with the study in [Courty et al., 2017]. We consequently evaluate in a side study the fully unsupervised setting, i.e., the case where no class label is available for the source or the target data. We consider the setting used in Fig. 3 with, this time, no given class labels. While the other methods fail for this task, the OT-SON with proper parameterization (i.e., the setting shown in the second row and the forth column) yields meaningful and consistent results. Fig. 5 shows the OT-SON results and the consistency of transport costs and transport maps computed by OT-SON.

11.4 Early stopping of the optimization

We study the early stopping of our optimization procedure. We use the data in Fig. 3 and investigate the results with different number of epochs. Here, we employ the OT-SON with proper parameterization, i.e., the results shown in the forth column and the second row for OT-SON in Fig. 3. In the experiments in Fig. 3 we performed the optimization with 20 epochs. Here, we study early stopping, i.e., we study the quality of results if we stop after a smaller number of epochs. According to the results in Fig. 6, we observe that even after a small number of epochs, we obtain reliable and stable results that represent well the ultimate solution. Such a property is very important in practice, as it can significantly reduce the heavy computations. Fig. 7 illustrates the transport maps for different number of epochs. The different transport maps at different number of epochs are consistent with the transport cost shown in the last row of Fig. 7.

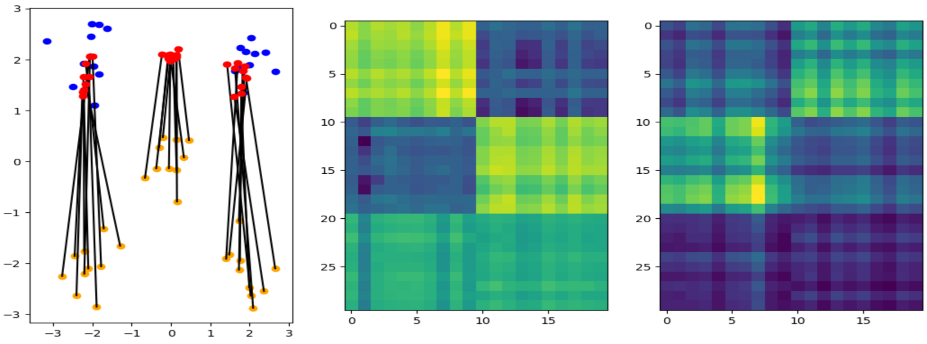


Figure 5: Unsupervised OT-SON, the OT-SON results and the consistency of transport costs and transport maps.

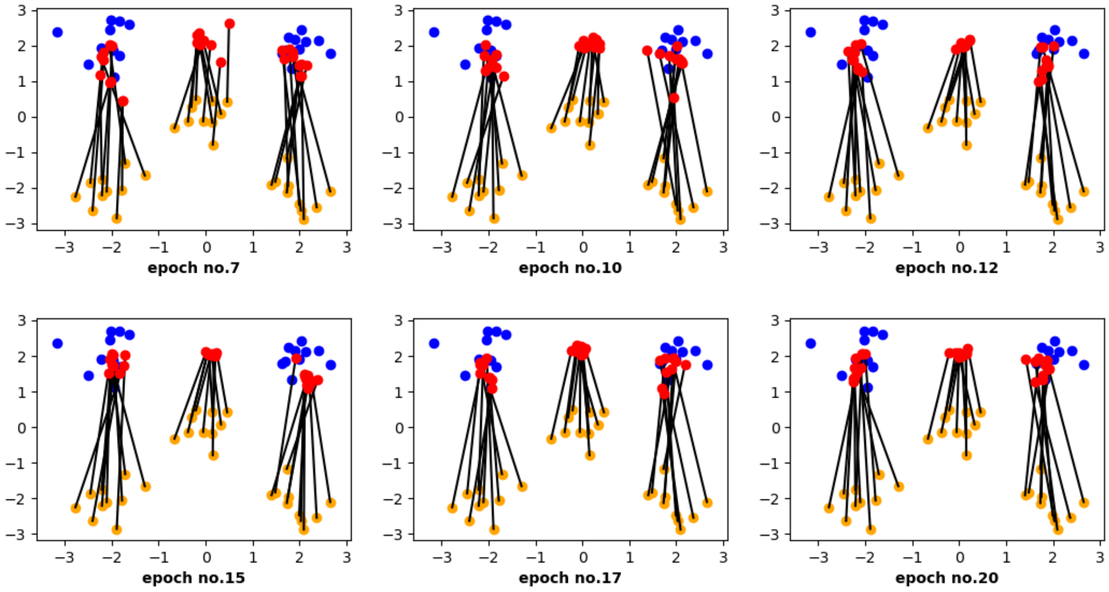


Figure 6: Early stopping of the optimization after a finite number of epochs. The results are very consistent and stable even of we stop the algorithm very early.

11.5 Diverse classes in the source

We next study the case where two of the three source classes have the same label, as shown in Fig. 8. In the source data (shown by yellow), the left and the middle data clouds have the same class labels. This example shows why the transport based on only the pairwise distances between the source and target data is insufficient. In Fig. 8, the left plot corresponds to $\lambda_1 = \lambda_2 = 0$, the middle plot corresponds to $\lambda_1 = 10, \lambda_2 = 0.01$, and the right plot corresponds to $\lambda_1 = 100, \lambda_2 = 0.01$. We observe that the left plot (with $\lambda_1 = \lambda_2 = 0$) fails to perform a proper transport of the source data. On the other hand, with incorporating our proposed regularization, the two different classes (even-though one of them is diverse) are properly transported to the target domain. We observe this kind of transfer in both of the middle ($\lambda_1 = 10, \lambda_2 = 0.01$) and right ($\lambda_1 = 100, \lambda_2 = 0.01$) plots.

11.6 Fewer classes in the source

In the experiments of Fig. 3, we studied the case where the number of source classes is larger the number of target classes. Here, we consider an opposite setting: we assume two classes in the source and three classes in the target, as illustrated in Fig. 9. The source, target and transported data points are respectively shown by yellow, blue, and red. We use the same setting and parameters as in Fig. 3, i.e., the first row corresponds to low regularization and the second row to high

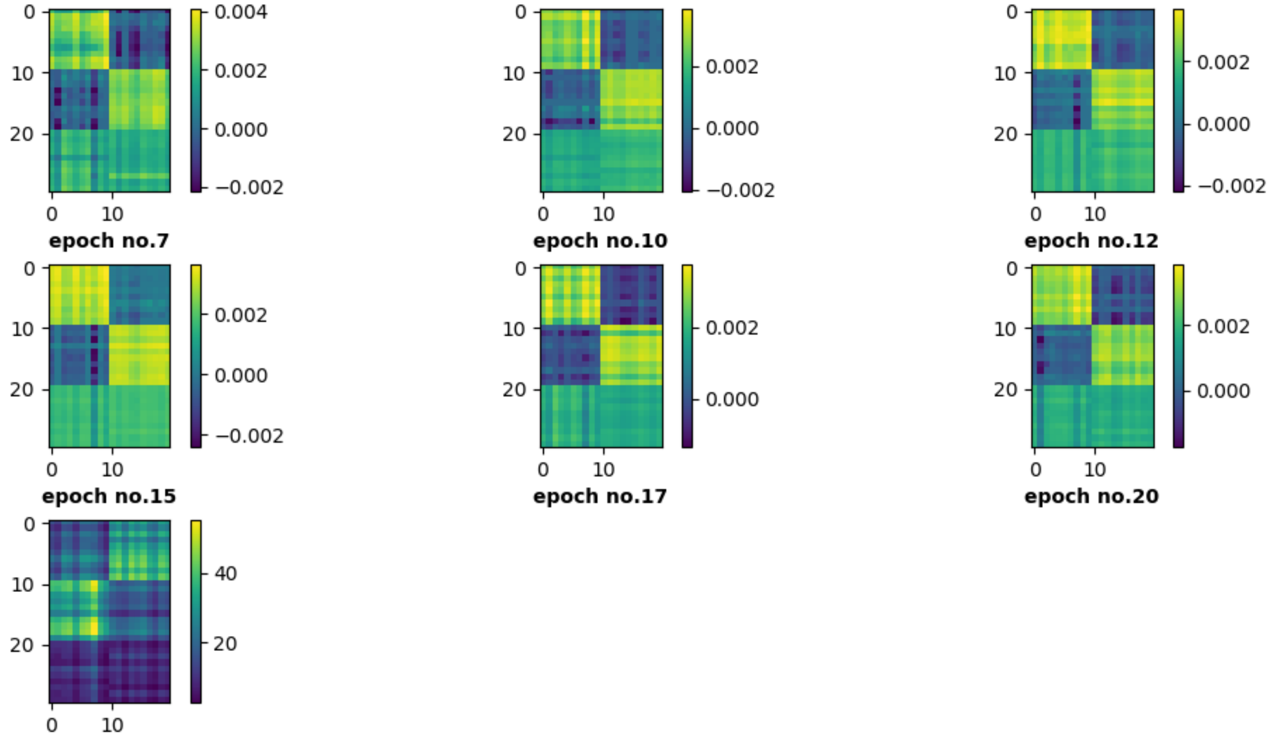


Figure 7: Consistency of the transport maps with the transport costs (shown at the last row) when using different finite number of epochs. Thus, early stopping can be useful for efficiency purposes.

regularization (low regularization: $\lambda_1 = 0.01, \lambda_2 = 0.0$, high regularization: $\lambda_1 = 10, \lambda_2 = 5$). We observe that similar to the results in Fig. 3, only OT-SON with high regularization prevents splitting the source data among all the three target classes. The last row in Fig. 9 indicates the consistency between the mapping costs and transport map for this setting (for OT-SON with high regularization).

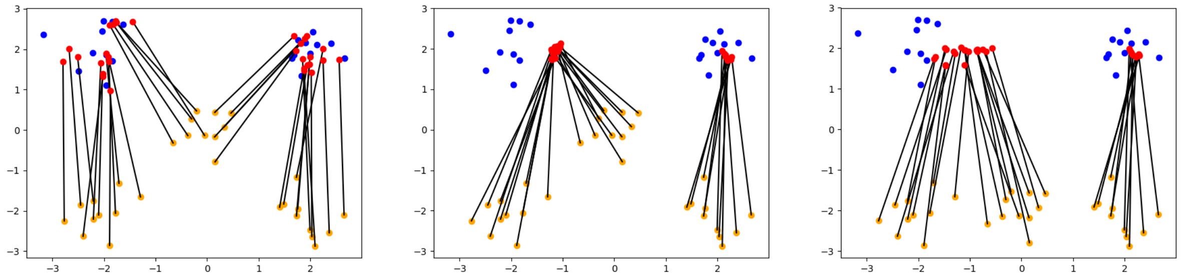


Figure 8: The impact of SON regularization when the class members are diverse. The plot on the left (where $\lambda_1 = \lambda_2 = 0$) performs transportation solely based on pairwise distances, thus fails to transfer the classes properly. Our SON regularization (either $\lambda_1 = 10, \lambda_2 = 0.01$ or $\lambda_1 = 100, \lambda_2 = 0.01$) improves the transportation by enforcing block-specific transfers.

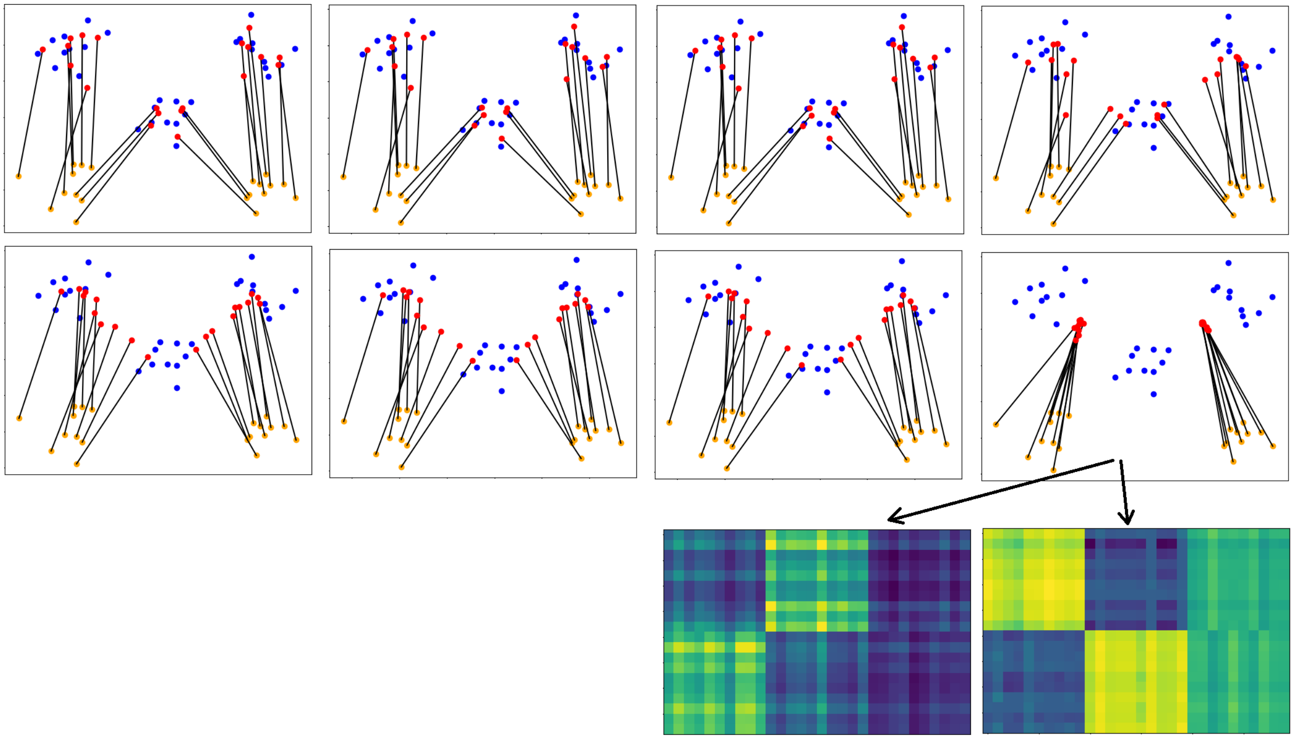


Figure 9: Performance of different methods when the source has two classes and the target consists of three classes. The columns in order represent OT-l1l2, OT-lpl1, OT-Sinkhorn and OT-SON. Among different methods, only OT-SON with high regularization prevents splitting the source data among all the three target classes. The last row shows the consistency between the mapping costs and the transport map for OT-SON with high regularization.