Recovery Bounds on Class-Based Optimal Transport: A Sum-of-Norms Regularization Framework

Arman Rahbar [1] **Ashkan Panahi** [1] **Morteza Haghir Chehreghani** [1] **Devdatt Dubhashi** [1] **Hamid Krim** [2]

Abstract

We develop a novel theoretical framework for understating Optimal Transport (OT) schemes respecting a class structure. For this purpose, we propose a convex OT program with a sum-of-norms regularization term, which provably recovers the underlying class structure under geometric assumptions. Furthermore, we derive an accelerated proximal algorithm with a closed-form projection and proximal operator scheme, thereby affording a more scalable algorithm for computing optimal transport plans. We provide a novel argument for the uniqueness of the optimum even in the absence of strong convexity. Our experiments show that the new regularizer not only results in a better preservation of the class structure in the data but also yields additional robustness to the data geometry, compared to previous regularizers.

1. Introduction

Optimal transport (OT) is a classical mathematical discipline for discovering a transport map from a source distribution to a target distribution with a minimum cost of the transport. It has recently been successfully used in various applications in computer vision, texture analysis, tomographic reconstruction and clustering, as documented in the recent surveys (Kolouri et al., 2017) and (Solomon, 2018). In many of these applications, OT exploits the geometry of the underlying spaces to effectively yield improved performance over the alternative of obviating it.

The main purpose of this paper is to provide a theoretical foundation for OT in such geometrically-aware conditions. We focus on a scenario where a common, potentially hidden class structure is present in both domains. Examples are abundant, such as (Long et al., 2022) and (Ott et al., 2022), where respecting the class structure can significantly improve the performance. In our setup, classes are associated with well separated regions of the data space, called *components*, on which the source and target distributions are supported. Each component is associated with a class and hence there is a correspondence between the components of the source and target domains. Such a model may become relevant after a suitable re-representation (embedding) of the data in a latent space.

Our central question is to identify conditions on the geometry of the components, under which OT can be performed in a polynomial time, with an additional property that each source sample is mapped to its corresponding component in the target domain. For this purpose, we introduce and study a two-stage procedure, where in the first stage the components and their association is recovered and in the second one, OT is exclusively performed over the corresponding component pairs. Any off-the-shelf OT method can be used in the second stage, and hence our results mainly pertain to the recovery condition in the first stage. For this, we introduce a novel convex optimization framework, combining two popular procedures: the celebrated Kantorovich relaxation scheme of OT and the well-known sum-of-norms (SON) formulation for vector clustering (Lindsten et al., 2011a; Hocking et al., 2011), used as a regularization. Accordingly, we show for the first time that sufficiently well-separated and associated components in the two domains can be recovered by a regularized OT scheme, which naturally enjoys a polynomial algorithm due to convexity. No such results, to the best of our knowledge, are known for other regularizers. We also experimentally show that our regularizer does not only yield a better class structure preservation, but also provides additional robustness compared to other class-based regularizers

Computational benefits: Despite a convex nature, the improvements of OT often come at a significant computational cost. We further argue that the SON regularization of OT also enjoys practical computational benefits by proposing a novel stochastic incremental algorithm. First, we construct an abstract stochastic framework that is based on a combina-

[1]Department of Computer Science and Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden [2]Electrical and Computer Engineering Department, North Carolina State University, Raleigh, NC 27606, USA. Correspondence to: Arman Rahbar <armanr@chalmers.se>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

tion of proximal and projection iterations, for which we give a generic proof of convergence at rate $O(1/\sqrt{T})$. Subsequently, we specialize this general scheme for our problem, which leads to an algorithm with computationally low-cost iterations. Beyond the proposed SON-based framework, our proximal scheme can be used to avoid the reported convergence difficulties of gradient-based methods (Patrascu and Necoara, 2018).

Summary of contributions: Our main contributions can be summarized as follows:
i. In section 2.1, we develop a theoretical framework for class-based OT, where we introduce the concept of multi-class recovery schemes.
ii. As an instance of a multi-class recovery scheme, we propose in Section 2.2 a new regularized formulation of OT that recovers a class structure typically arising in real-world problems.
iii. We derive the first rigorous results for recovering an OT plan that respects class structure, presented in section 3, with more details provided in the appendix (section A).
iv. We develop in Section 4 a general accelerated stochastic incremental proximal-projection optimization scheme, for which we give a proof of convergence at a rate $O(1/\sqrt{T})$ without a decaying step size. We specialize the general scheme with an explicit closed form of proximal operators and fast projections to yield a scalable stochastic incremental algorithm for computing our OT formulation.
v. In section 5 and further in the appendix (section C), we investigate the algorithm on several synthetic and benchmark data sets, and demonstrate the benefits of the new regularizer.
vi. In the appendix (section D), we develop a new proof for the uniqueness of the optimal solution of our convex formulation in spite of its non-strong convexity, with wide applicability in other model recovery studies.

1.1. Relation to Literature

OT, first proposed by Monge as an analysis problem (Monge, 1781), has become a classic topic in probability and statistics. A comprehensive introduction and theoretical framework can be found in (Villani, 2008; Santambrogio, 2015; Peyré et al., 2019). Theoretical works on extensions of OT have only recently received more attention. (Redko et al., 2017), for example, provides an analysis of OT in the context of domain adaptation. In (Gordaliza et al., 2019), OT is studied from the fairness point of view. To the best of our knowledge, there is no study on the geometrically-aware extensions (regularizations) of OT, as this paper offers.

Regularized OT: A case for exploring new regularizers was made in (Courty et al., 2017) in the context of domain adaptation applications. In (Blondel et al., 2018) the primal and dual formulations of OT are regularized with a strongly convex term, and the constraints are relaxed with smooth approximations. (Dessein et al., 2018) also propose a framework to solve discrete optimal transport problems with smooth convex regularization. Regularization is often introduced to promote sparsity. The L2 regularization for example yields a sparse plan. It has been used in a doubly stochastic scheme in (Seguy et al., 2018). These techniques are not tailored to the underlying class structure of the data, which may not be known in advance. Hidden class structure is addressed in (Courty et al., 2017) by the so-called Laplacian regularizer. However, no theoretical study is provided. We further illustrate the differences between the Laplacian regularizer and our framework in the experiments. Recently, a related attempt is made in (Asadulaev et al., 2022) to address class structures in the two domains. Our formulation differs from the *partial domain adaption* (Cao et al., 2018) and *robust domain adaption* (Balaji et al., 2020) settings which do not generally take the class structure into account.

Computational aspects: Much attention has focused on efficient computational and numerical algorithms for OT, and a monograph focusing on this topic has appeared in (Peyré et al., 2019). In (Guo et al., 2020a), an "accelerated primal-dual randomized coordinate descent (APDRCD)" algorithm is developed to solve the OT problem. An upper bound is also provided for the complexity of the algorithm and it is shown that it could be used for large-scale purposes. In (Courty et al., 2017), a generalized conditional gradient method is used to compute OT with the help of a couple of regularizers. Most notably Cuturi introduced an entropic regularizer and showed that its adoption with the Sinkhorn algorithm yields a fast computation of OT (Cuturi, 2013); a theoretical guarantee that the Sinkhorn iteration computes the approximation in near linear time was also provided by (Altschuler et al., 2017). Screenkhorn algorithm proposed in (Alaya et al., 2019) performs screening to eliminate (and accelerate) the solution of the Sinkhorn algorithm. Another computational breakthrough was achieved by (Genevay et al., 2016) who gave a stochastic incremental algorithm to solve the entropic regularized OT problem. Proximal method has been used in OT but not in a stochastic scheme. Examples include (Alvarez-Melis et al., 2018) that uses notions of sub-modularity and also (Papadakis et al., 2014). Unlike the above works, our algorithm is tailored to the class-based OT framework, and exploits proximal/ projection operators in a stochastic way.

2. Problem Formulation and SON-regularized Optimal Transport

2.1. Problem Formulation

Consider two probability measures μ, ν defined on data spaces $\mathcal{Y}^s, \mathcal{Y}^t$, respectively referred to as the source and the

target domains. Further, assume a non-negative function $d : \mathcal{Y}^s \times \mathcal{Y}^t \to \mathbb{R}_{\geq 0}$ evaluating the transport cost between the two domains. Classical Monge problem seeks a measurable transport map $T : \mathcal{Y}^s \to \mathcal{Y}^t$ such that the push-forward measure $T\sharp\mu$ of μ under T coincides with ν and the expected transport cost $\mathbb{E}\left[d(Y, T(Y))\right]$ is minimized, where $Y \sim \mu$.

We similarly define the class-based OT problem:

Definition 1. A *K-class structure* is a pair of mixtures of K probability measures:

$$\mu = \sum_{\alpha=1}^{K} p_\alpha \mu_\alpha, \quad \nu = \sum_{\beta=1}^{K} q_\beta \nu_\beta \tag{2.1}$$

on $\mathcal{Y}^s$ and $\mathcal{Y}^t$, respectively, with a one-to-one correspondence π between the components of the two domains (i.e. μ_α, ν_β). We define a solution to the $K-$class Monge problem as a transport map $T : \mathcal{Y}^s \to \mathcal{Y}^t$ such that:

1. For every corresponding pair $(\mu_\alpha, \nu_\beta) \in \pi$ we have $T\sharp\mu_\alpha = \nu_\beta$, i.e T transports any component to its corresponding component.
2. The expected transport cost $\mathbb{E}\left[d(Y, T(Y))\right]$ is minimized, where $Y \sim \mu$.

Note that a solution T of the multi-class Monge problem depends on the components and their association. We consider a case that these are not provided:

Definition 2. Given a transport cost function d, a *Monge recovery scheme* assigns to each pair (μ, ν) of source and target distributions, a $K-$class structure $((\mu_\alpha, p_\alpha)_\alpha, (\nu_\beta, q_\beta)_\beta, \pi)$ for some $K \geq 0$, and its corresponding $K-$class Monge solution T such that equation 2.1 holds true. A recovery scheme is said to recover a family $\mathcal{F}$ of multi-class structures if it assigns to any (μ, ν) of $\mathcal{F}$ its corresponding mixture components and their association in $\mathcal{F}$. This naturally requires the components and their association to be unique in $\mathcal{F}$.

The main purpose of this paper is to introduce a wide family $\mathcal{F}$ of mixture models and a particular Monge recovery scheme recovering it in a polynomial time with the problem size.

2.1.1. Disjoint Supports and Two-Stage Schemes

Our focus is on mixture models with disjoint supports. For this case, we make the following straightforward observation:

Proposition 1. Consider a multi-class structure $((\mu_\alpha, p_\alpha)_\alpha, (\nu_\beta, q_\beta)_\beta, \pi)$ and suppose that the supports $\mathscr{S}_\alpha$ of μ_α are disjoint, and the same property holds for the supports $\mathscr{T}_\beta$ of ν_β. Then the multi-class Monge solution comprises the restricted maps $T_\alpha : \mathcal{S}_\alpha \to \mathcal{T}_\beta$, for $(\alpha, \beta) \in \pi$, each being the solution of the conventional Monge problem for (μ_α, ν_β).

For such families of structures, the above observation suggests a two-stage recovery scheme: Given a pair of distributions (μ, ν), first recover the components $(\mu_\alpha), (\nu_\beta)$ and their association π. Next, solve the conventional Monge problem over each associated pair. We adopt this strategy, and as a rich theory already exists for the conventional Monge problem, we mainly focus on the first stage, i.e. recovering the components and their association.

2.2. Optimal Transport with SON

A major challenge in optimal transport is that only a finite number of samples from each distribution is given. Otherwise the distributions are unknown. Consider two finite sets $\{\mathbf{y}_i^s\}_{i=1}^m, \{\mathbf{y}_j^t\}_{j=1}^n$ of points, respectively sampled from μ and ν. Let $\mathbf{D} = (D_{ij} = d(\mathbf{y}_i^s, \mathbf{y}_j^t))$ be the $m \times n$ matrix of transport costs. We denote the i^{th} row and j^{th} column of $\mathbf{D}$ by $\mathbf{d}^i$ and $\mathbf{d}_j$, respectively. We let the positive probability masses μ_i, ν_j be respectively assigned to the data points $\mathbf{y}_i^s$ and $\mathbf{y}_j^t$. In this discrete setup, the Monge problem is often solved with a linear programming relaxation scheme known as the Kantorovich problem:

$$\min_{\mathbf{X} \in B(\boldsymbol{\mu}, \boldsymbol{\nu})} \langle \mathbf{D}, \mathbf{X} \rangle. \tag{2.2}$$

Here, the variable matrix $\mathbf{X} = (x_{i,j})$ is called the transport map and $B(\boldsymbol{\mu}, \boldsymbol{\nu}) = \{\mathbf{X} \in R^{m \times n}, \mathbf{X}\mathbf{1}_{n^s} = \boldsymbol{\mu}, \mathbf{X}^T\mathbf{1}_{n^t} = \boldsymbol{\nu}\}$ is the set of all discrete coupling distributions between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, respectively denoting the vectors of elements μ_i, ν_j. Moreover, $\langle \mathbf{D}, \mathbf{X} \rangle = \mathrm{Tr}(\mathbf{D}^T\mathbf{X}) = \sum_{i,j} X_{ij} D_{ij}$ is the Euclidean inner product of two matrices. In an ideal case, one hopes that the optimal solution for $\mathbf{X}$ become an assignment (permutation matrix) in which case it is seen to coincide with the solution of the Monge problem for the empirical distributions.

2.2.1. Multi-Class Recovery Scheme

Now, we introduce our multi-class recovery scheme by the following convex optimization problem:

$$\mathbf{X}^* = \arg \min_{\mathbf{X} \in B(\boldsymbol{\mu}, \boldsymbol{\nu})} \langle \mathbf{D}, \mathbf{X} \rangle + \lambda \left(\sum_{l,k} R_{l,k} \|\mathbf{x}_l - \mathbf{x}_k\|_2 + \sum_{l,k} S_{l,k} \|\mathbf{x}^l - \mathbf{x}^k\|_2 \right), \tag{2.3}$$

where $\mathbf{x}_l$ and $\mathbf{x}^k$ denote the (transpose of the) l^{th} row and k^{th} column of $\mathbf{X}$, respectively. Compared to equation 2.2, a regularization term with a tuning parameter $\lambda > 0$ is introduced, known as sum-of-norms (SON). SON is well-known

for its clustering properties and hence equation 2.3 combines the class discovery properties of SON with OT. The positive kernel coefficients $S_{l,k}, R_{l,k}$ are also introduced to incorporate class prior information.

The effect of the SON regularization in equation 2.3 is explained in (Lindsten et al., 2011b; Panahi et al., 2017). In short, it enforces many vanishing regularization terms (sparsity), hence yielding identical columns and identical rows in the solution. In other words, the resulting map $\mathbf{X}^*$ after a suitable permutation of rows and columns is a block matrix with constant values in each block. The block structure reflects the discovered components in the source and target domains. Under suitable conditions , the constraints of the Kanturovitch problem will further force many blocks to be zero. If each row and column will contain exactly one non-zero block (i.e. a block diagonal matrix under a correct order), the solution further reflects an assignment π between the components. In this way, equation 2.3 performs the first stage in the two-stage multi-class recovery scheme. This is made precise and proved in section 3.

The regularization parameter λ sets a desired balance between the cluster structure and the underlying transport problem. When $\lambda = 0$, equation 2.3 reduces to equation 2.2. For large values of λ, the SON regularization dominates the result. In a typical situation, $\lambda \to \infty$ results in all data in each domain assigned to the same cluster, hence a trivial transformation between single clusters in each domain. Smaller values of λ lead to a larger number of identified classes in the solution.

As mentioned, the kernel coefficients $R_{l,k}, S_{l,k}$ are related to the prior knowledge of the components. For example if a perfect knowledge of the components in the source domain is available, we may set $R_{l,k} = 0$ if $\mathbf{y}_l^s$ and $\mathbf{y}_k^s$ belong to different components (classes), otherwise set $R_{l,k} = k_s(\mathbf{y}_l^s, \mathbf{y}_k^s)$ for a suitable (differentiable) kernel k_s. On the target side where no class information is ordinarily provided, we may set $S_{l,k} = k_t(\mathbf{y}_l^t, \mathbf{y}_k^t)$ for a suitable kernel k_t of choice.

3. Main Theoretical Results

Geometric threshold: To provide the main geometric intuitions of our analysis, we start by a simplified case with well-separated components. Next, we will present a more extensive study, which is also used for proving the simplified result:

Theorem 3.1. *Consider a mixture of K components in each domain with m samples from each component, hence a total of $n = Km$ samples in each domain. Suppose that $(\alpha, \alpha) \in \pi$ for all α.*

1. *Consider the ℓ_2 distance, $d(\mathbf{y}_1, \mathbf{y}_2) = \|\mathbf{y}_1 - \mathbf{y}_2\|_2$ and assume that the components are supported on spheres of radius ω and centered on $\boldsymbol{\theta}_\alpha^s, \boldsymbol{\theta}_\alpha^t$ for $\alpha = 1, 2, \ldots, K$, in the source and target domains, respectively. With a suitable choice of λ, the solution of equation 2.3 recovers the $K-$class structure if:*

$$\frac{D - d - 2\omega}{K^{\frac{3}{2}}} \geq C\omega, \tag{3.1}$$

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

2. *Consider the ℓ_2^2 distance, $d(\mathbf{y}_1, \mathbf{y}_2) = \|\mathbf{y}_1 - \mathbf{y}_2\|_2^2$. Assume that the components are isotropic Gaussian with means $\boldsymbol{\theta}_\alpha^s, \boldsymbol{\theta}_\alpha^t$ for $\alpha = 1, 2, \ldots, K$, in the source and target domains, respectively. All variances are equal to ω^2. With a probability higher than $1 - 1/n^{10}$ the solution of equation 2.3 with a suitable choice of λ recovers the $K-$class structure if:*

$$\frac{D^2 - d^2}{K^{\frac{3}{2}}} \geq C\omega\sqrt{(E+\omega)^2 + 1}\log(nK), \tag{3.2}$$

where $E = \max_{\alpha,\beta} \|\boldsymbol{\theta}_\alpha^s - \boldsymbol{\theta}_\beta^t\|$.

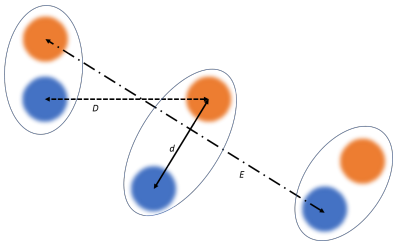


Figure 1. An example of three pairs of Gaussian clusters in the source (blue) and target (red) domains. The maximum distance d between associated (paired) centers, the minimum distance D between unassociated centers and the maximum distance E of centers between two domains are respectively shown by solid,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

Fig. 1 clarifies in a simple example the geometric meaning of the concepts used in the above result. As seen, the conditions in equation 3.2 and equation 3.1 require the associated clusters to be substantially closer to each other than the other clusters. As expected, there are differences between the ℓ_2 and ℓ_2^2 geometries. For example, the scale E of the problem only appears in the latter. However, in both cases we may bound the maximum number of resolvable components K by a purely geometric parameter that we refer to as *the geometric threshold*. In case 1), for example, defining $r = \frac{D-d-2\omega}{\omega}$, we observe that the geometric threshold is $O(r^{\frac{2}{3}})$. Part 2) has a similar interpretation.

Deterministic guarantee: Now, we present an extended deterministic result that is used to prove theorem 3.1. In many respects, the presented results are still simplified and

the most comprehensive analysis is postponed to Appendix B. For simplicity, K components with equal size (number of samples) m is assumed. The resulting data partitions in the source and target domains are respectively denoted by $\{\mathscr{S}_\alpha\}$, $\{\mathscr{T}_\beta\}$. The total number of points in each domain is $n = mK$. Further, $\mathscr{S}_\alpha$ is paired with $\{\mathscr{T}_\alpha\}$ for every $\alpha \in [K]$, i.e. $(\alpha, \alpha) \in \pi$. We investigate that the transport plan obtained by solving equation 2.3 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 ideal solution to be feasible for the optimization problem in equation 2.3, 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 2.3.

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 3. 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

$$\sum_{l=1}^{k} D_{\alpha_l \alpha_{l+1}} > \sum_{l=1}^{k} D_{\alpha_l \alpha_l} + k\delta. \tag{3.3}$$

Compared to the standard notion of cyclic monotonicity, we introduce a constant $\delta \geq 0$ on the right hand side of equation 3.3, 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 $D_{\alpha,\beta} = \frac{1}{m^2} \sum_{i \in \mathscr{S}_\alpha, j \in \mathscr{T}_\beta} D_{i,j}$.

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 $\omega_\alpha := \sum_{i \in \mathscr{S}_\alpha} \mu_i = \sum_{j \in \mathscr{T}_\alpha} \nu_j$ and then take $T_{\alpha,\beta} = \sum_{\gamma \in [K]} \left(\frac{R\omega_\alpha}{\sqrt{\omega_\alpha^2 + \omega_\gamma^2}} + \frac{\omega_\beta}{\sqrt{\omega_\beta^2 + \omega_\gamma^2}} \right) - \frac{1+R}{\sqrt{2}}$. 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 3.2. *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 2.3 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 $\delta \sum_{\beta \neq \alpha} X_{\alpha,\beta} \leq \lambda(1 + R)\sqrt{m} \sum_{\alpha \neq \alpha'} \sqrt{\omega_\alpha^2 + \omega_{\alpha'}^2}$.*

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

Proof. Proof can be found in section A.1. □

The first part of theorem 3.2 establishes ideal recovery. It can be understood in light of theorem 3.1, e.g., in part 1. The term Δ corresponds to ω in 3.1. It is the maximal cluster diameter in a geometry embedded by the vectors $\mathbf{d}_i, \mathbf{d}^j$[1]. On the other hand, δ corresponds to $D - d - 2\omega$ which is the gain of the assignment. The second part gives an upper bound on the error $\sum_{\beta \neq \alpha} X_{\alpha,\beta}$. 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 recovery.

4. Stochastic Incremental Algorithms

4.1. Accelerated Proximal-Projection Scheme

An important advantage of the framework in equation 2.3 is the possibility of applying stochastic optimization techniques. Since the objective term includes a large number of non-smooth SON terms, our stochastic optimization avoids calculating the (sub)gradient or the proximal operator of the entire objective function, which is numerically infeasible for large-scale problems. Our algorithm is obtained by introducing the following "template function":

$$\phi_{\rho,\boldsymbol{\zeta},\boldsymbol{\eta}}(\mathbf{p}, \mathbf{q}) = \langle \mathbf{p}, \boldsymbol{\zeta} \rangle + \langle \mathbf{q}, \boldsymbol{\eta} \rangle + \rho\|\mathbf{p} - \mathbf{q}\|_2, \tag{4.1}$$

[1]Note that here $\mathbf{d}_i, \mathbf{d}^j$ are treated as a special embedding of the source and target points, respectively, which may be called the *distance embedding*. As seen, we do not assume any inherent geometry in the two domains and instead rely on the induced ℓ_2 geometry of the distance embedding.

and noting that the objective function in equation 2.3 can be written as

$$\sum_{l\neq k}\phi_{R_{l,k},\frac{1}{2(n-1)}\mathbf{d}_l,\frac{1}{2(n-1)}\mathbf{d}_k}(\mathbf{x}_l,\mathbf{x}_k)+ \sum_{l\neq k}\phi_{S_{l,k},\frac{1}{2(m-1)}\mathbf{d}^l,\frac{1}{2(m-1)}\mathbf{d}^k}(\mathbf{x}^l,\mathbf{x}^k), \tag{4.2}$$

with a total number of $P = m(m-1)+n(n-1)$ summands in the form of the template function. This places the problem in the setting of *finite sum* optimization problems (Bottou et al., 2018). However, there are two obstacles to the application of stochastic optimization techniques: First, the terms in (4.2) are not smooth, so gradient methods do not apply and second, equation 2.3 involves a fairly complex constraint. We address these issues in the following.

Non–smooth terms: We exploit the highly effective proximal methodology for optimizing non–smooth functions (Parikh and Boyd, 2016; Combettes and Pesquet, 2011) using a proximal operator. Defazio further gives a stochastic acceleration technique using proximal operators for unconstrained problems (Defazio, 2016). In addition to its fast convergence, the main advantage of this scheme is its potential constant step size convergence in contrast to the ordinary stochastic gradient approach. It unfortunately does not address constrained optimization problems.

Constrained optimization: Facing a constrained optimization problem, the calculation of the proximal operators over the feasible set is numerically intractable. However, we observe an appealing structure in the constraint which lends itself to a more efficient stochastic implementation: Recalling the definition of an $n-$dimensional standard simplex

$$S^{(n)} = \left\{\mathbf{x}=(x_i\geq 0)_{i=1}^n \mid \sum_i x_i = 1\right\},$$

we define the weighted cylinder-simplices $S_l(\mu) = \{\mathbf{X} \mid \mathbf{x}_l \in \mu S^{(n)}\}$ and $S^k(\nu) = \{\mathbf{X} \mid \mathbf{x}^k \in \nu S^{(m)}\}$ respectively corresponding to the l^{th} row and k^{th} column of $\mathbf{X}$ with weights $\mu,\nu \geq 0$. We then observe that the constraint set $B(\boldsymbol{\mu},\boldsymbol{\nu})$ is equal to $B(\boldsymbol{\mu},\boldsymbol{\nu}) = (\bigcap_{l=1}^m S_l(\mu_l)) \cap (\bigcap_{k=n}^m S^k(\nu_k))$, which is an intersection of $Q = m+n$ weighted cylinder-simplices.

In summary, the optimization problem in equation 2.3 can be written in the following abstract form:

$$\min_{x\in\mathbb{R}^D}\sum_{p=1}^P \phi_p(x) \quad \text{st} \quad x\in\bigcap_{q=1}^Q S_q, \tag{4.3}$$

where each term ϕ_p denotes a template function term in the objective and each set S_q is a weighted cylinder-simplex. The values of P, Q in equation 2.3 are given above and $D = mn$. (Bertsekas, 2011), (Wang and Bertsekas, 2016) and (Patrascu and Necoara, 2018) give general stochastic incremental schemes that combine gradient, proximal and projected schemes for optimizing such finite sum problems with convex constraints. These do not, however, use acceleration and their respective convergence is only guaranteed with a variable and vanishing step size, which is practically difficult to control and often yields extremely slow convergence.

Our proposed method: We herein jointly exploit the two ideas in (Defazio, 2016) and (Wang and Bertsekas, 2016) to obtain an accelerated proximal scheme for constrained framework in equation 2.3. Further, we shortly show in Lemma 4.2 that the proximal operator can be computed in closed form for our problem. Together with the projection to the simplex from (Condat, 2016; Duchi et al., 2008), this gives a stochastic incremental algorithm with much less costly iterations.

We extend the acceleration techniques of unconstrained optimization as in the Defazio's scheme (known as Point–SAGA) to the constrained setting. Point–SAGA utilizes individual "memory" vectors for each term in the objective function, which store a calculated subgradient of a selected term in every iteration. These vectors are subsequently used as an estimate of the subgradient at subsequent iterations. We extend this scheme by introducing similar memory vectors to constraints. Each memory vector $\mathbf{h}_m$ for a constraint S_m stores the last observed normal (separating) vector to S_m. At each iteration either an objective term ϕ_p or a constraint component S_q is considered by random selection. Accordingly, we propose the following rule for updating the solution:

$$\mathbf{x}_{t+1} = \begin{cases} \text{prox}_{\mu\phi_{p_t}}(\mathbf{x}_t+\mu\mathbf{g}_{p_t}), & \phi_{p_t} \text{ is selected} \\ \text{proj}_{S_{q_t}}(\mathbf{x}_t+\mu\mathbf{h}_{q_t}) & S_{q_t} \text{ is selected} \end{cases}, \tag{4.4}$$

where t is the iteration number, $\mu>0$ is the fixed step size and p_t, q_t denote the selected index in this iteration (only one of them exists). At each iteration, the corresponding memory vector to the selected term is also updated. Depending on the choice of ϕ_{p_t} or S_{q_t}, either $\mathbf{g}_{p_t} \leftarrow \mathbf{g}_{p_t}+\mathbf{a}_t$ or $\mathbf{h}_{q_t} \leftarrow \mathbf{h}_{q_t}+\mathbf{a}_t$, where

$$\mathbf{a}_t = \rho\frac{\mathbf{x}_t-\mathbf{x}_{t+1}}{\mu} - \alpha\left(\sum_n \mathbf{g}_n + \sum_m \mathbf{h}_m\right), \tag{4.5}$$

where $\rho\in(0\ 1)$ and $\alpha>0$ are design constants. The vector $\mathbf{a}_t$ consists of two parts: the first part $\rho\frac{\mathbf{x}_t-\mathbf{x}_{t+1}}{\mu}$ calculates a sub-gradient or a normal vector at point $\mathbf{x}_{t+1}$ corresponding to the selected term. The second term, the sum of the memory terms, implements acceleration. Our algorithm bears marked differences with Point-SAGA. While acceleration by the sum of memory vectors is also employed in Point-SAGA, it is moved in our scheme from the update

Algorithm 1 OT-SON

Input: Source and Target data: $\{\mathbf{y}_i^s\}_{i=1}^m, \{\mathbf{y}_j^t\}_{j=1}^n$, μ, $\rho \in (0\ 1)$ and $\alpha > 0$
Initialize $\mathbf{x} \in \mathbb{R}^D$, where $D = mn$
for $t = 1, 2, 3, \ldots$ **do**
  Randomly select objective or constraint and then a random index $p_t \in [P]$ or $q_t \in [Q]$
  if p_t is selected **then**
    $\mathbf{x}_{t+1} = \text{prox}_{\mu\phi_{p_t}}(\mathbf{x}_t + \mu\mathbf{g}_{p_t})$ which can be computed based on Theorem 4.2
    $\mathbf{g}_{p_t} \leftarrow \mathbf{g}_{p_t} + \mathbf{a}_t$, where $\mathbf{a}_t$ is given in equation 4.5
  else
    $\mathbf{x}_{t+1} = \text{proj}_{S_{q_t}}(\mathbf{x}_t + \mu\mathbf{h}_{q_t})$
    $\mathbf{h}_{q_t} \leftarrow \mathbf{h}_{q_t} + \mathbf{a}_t$
  end if
end for

rule of $\mathbf{x}_t$ to the update rule of $\mathbf{g}_t$. Also, the design parameters ρ and α are introduced to improve convergence. Similar to Point-SAGA we only need to calculate the sum of memory terms once in the beginning and later update it by simple manipulations. As we later employ initialization of the memory vectors by zero, the first summation trivially leads to zero. Algorithm 1 summarizes our optimization algorithm (OT-SON).

Convergence analysis: We show that for a generic convex optimization problem of the form in equation 4.3, the algorithmic scheme in section 4.1 converges with a guaranteed rate, under the following mild assumptions:

Assumption 1. The functions ϕ_p are $\beta-$Lipschitz.

Assumption 2. We require the monotone inclusion problem

$$\mathbf{0} \in \sum_{p=1}^{P} \partial\phi_p(\mathbf{x}) + \sum_{q=1}^{Q} \partial I_{S_q}(\mathbf{x}) \tag{4.6}$$

to have a solution at $\mathbf{x} = \mathbf{x}^*$ with a finite optimal value ϕ^* and $\mathbf{g}_p^* \in \partial\phi_p(\mathbf{x}^*)$ and $\mathbf{h}_q^* \in \partial I_{S_q}(\mathbf{x}^*)$ satisfying $\sum_p \mathbf{g}_p^* + \sum_q \mathbf{h}_q^* = \mathbf{0}$. Furthermore, we assume that

$$\sum_p \left\|\mathbf{g}_p^*\right\|^2 + \sum_q \left\|\mathbf{h}_q^*\right\|^2 = O(\beta^2 R) \tag{4.7}$$

where $R = P + Q$ is the total number of terms.

Here, $\partial\phi(\mathbf{x})$ and $\partial I_S(\mathbf{x})$ respectively denote the subdifferential of the function ϕ and the cone of normal vectors to the set S at $\mathbf{x}$. It is well-known that any solution to equation 4.6 is an optimal feasible solution to equation 4.3.

Assumption 3. We assume that the algorithm is initialized with $\mathbf{g}_p = \mathbf{h}_q = \mathbf{0}$.

Then, we can show the following result.

Theorem 4.1. *Suppose that Assumption 1-3 are satisfied. Then for $\alpha, \rho, \mu > 0$ and $\alpha < 2(1-\rho)$ the following holds true:*

1. *Defining $\bar{\mathbf{x}}_t = \frac{1}{t}\sum_{\tau=0}^{t-1} \mathbf{x}_\tau$ and $\eta = (1 + Q/P)(\|\mathbf{x}_0 - \mathbf{x}^*\|^2/\beta\mu + \beta\mu R)$ we have*

$$\mathbb{E}\left[\sum_p \phi_p(\bar{\mathbf{x}}_t)\right] - \phi^* \le c\beta P\left(\frac{\eta}{t} + \sqrt{\frac{\beta\mu\eta}{t}}\right),$$
$$\mathbb{E}\left[\sum_q \text{dist}^2(\bar{\mathbf{x}}_t, S_q)\right] \le c\beta\mu P\frac{\eta}{t}, \tag{4.8}$$

   where c is a constant depending on ρ, α and the underlying constant in equation 4.7.

2. *Moreover, $\sum_{\tau=0}^{\infty} \mathbb{E}[\|\mathbf{x}_{\tau+1} - \mathbf{x}_\tau\|^2] \le c\frac{\mu\beta P}{P+Q}\eta$.*

Proof. The proof is given in section B. □

Comment: As expected, the results only depend on $\beta\mu$, except for the optimality gap being linearly proportional to β. Applying this technique to our problem of interest and assuming that m, n are of the same order, we observe that $P = O(n^2)$ and $Q = O(n)$. Since many terms in our objective function are for regularization, it is fair to consider the relative optimality gap obtained by driving the optimality gap to the number of objective terms. We observe that this quantity is controlled by η/t. We conclude that $t \sim \eta$ iterations is required to achieve a desired relative optimality gap. The total feasibilty gap is controlled by $\beta\mu P\eta$. If we take $\beta\mu \sim 1/n$, we obtain $\eta \sim n$ and the relative optimality gap vanishes in $O(n)$ iterations. Then, the total optimality gap and feasibility gap will vanish in $O(n^3)$ and $O(n^2)$ iterations, respectively. In the absence of the regularization terms, we may reorganize the objective to have only $O(n)$ terms. In this case, taking $\beta\mu \sim 1/\sqrt{n}$, we get $\eta \sim O(\sqrt{n})$ and we require $O(n^{\frac{3}{2}})$ and $O(n)$ iterations to control the total optimality and feasibility gaps. Compared to the results of (Guo et al., 2020b), which establishes convergence in $O(n^{\frac{5}{2}})$ our convergence rates are better. Moreover, the $O(n^2)$ dependence can be improved by reducing the number of terms in the objective function. It has been pointed out that in the sum of norms approach many terms may be redundant and only $O(n)$ terms corresponding to pre-selected pairs (i, j) can be sufficient.

4.2. Proximal Operator for the SON-Regularized Kantorovich Relaxation

We next show that we can explicitly compute the proximal operator for each term in (4.2):

Theorem 4.2. *The proximal operator of the template function $\phi_{\rho,\boldsymbol{\zeta},\boldsymbol{\eta}}$ is given by $\mathcal{T}_{\mu\rho}(\mathbf{p}-\mu\boldsymbol{\zeta}, \mathbf{q}-\mu\boldsymbol{\eta})$, where*

$$\mathcal{T}_\lambda(\mathbf{a},\mathbf{b}) = \left(\tfrac{\mathbf{a}+\mathbf{b}}{2} + \mathcal{T}_\lambda\left(\tfrac{\mathbf{a}-\mathbf{b}}{2}\right), \tfrac{\mathbf{a}+\mathbf{b}}{2} - \mathcal{T}_\lambda\left(\tfrac{\mathbf{a}-\mathbf{b}}{2}\right)\right) \quad (4.9)$$

and $\mathcal{T}_\lambda(\mathbf{c})$ is a thresholding operator given by $\frac{\|\mathbf{c}\|-\lambda}{\|\mathbf{c}\|}\mathbf{c}$ if $\|\mathbf{c}\| \geq \lambda$ and is zero otherwise.

Proof. Proof is found in section B.3. □

5. Experiments

We now investigate the effectiveness of our OT framework for the well-known domain adaptation problem which aims at improving the performance of a model in a target domain by utilizing the knowledge available in a different but related source domain. In the appendix (section C), we investigate the benefits of several other aspects of our framework, such as early stopping, class diversity, and unsupervised domain adaptation. We compare our method (OT-SON) with the other regularized optimal transport-based methods OT-l1l2, OT-lpl1 and OT-Sinkhorn, as developed and used in (Courty et al., 2017; Cuturi, 2013; Perrot et al., 2016).

5.1. Domain Adaptation with Real-World Datasets

5.1.1. MNIST AND USPS

In these experiments, we compare the different models on the real-world images of digits. For this, we consider the MNIST data as the source and the USPS data as the target. To further increase the difficulty of the problem, we use all 10 classes of the source (MNIST) data, and we discard some of the classes of the target (USPS) data. In our experiments, each object (image) is represented by 256 features. By discarding the different subsets from the USPS data, we consider several pairs of source and target datasets. i) real1: the USPS classes are $1, 2, 3, 5, 6, 7, 8$, ii) real2: the USPS classes are $0, 2, 4, 5, 6, 7, 9$, iii) real3: the USPS classes are $0, 1, 3, 5, 7, 9$, and iv) real4: the USPS classes are: $0, 1, 3, 4, 6, 8, 9$. We note that these settings where the number of classes is different between the domains are the typical cases in practice. Therefore, our class-specific OT approach is more suitable and robust to class imbalance, as it avoids splitting a class in one domain among multiple classes in another domain.

The transformed source samples are used to train a 1-nearest neighbor classifier. We then use this (parameter-free) classifier to estimate the class labels in the target data and then compute the respective accuracy. Table 1 shows the best accuracy results for different OT-based models when using different values of the regularization parameters λ_1 and λ_2 (i.e., $\lambda_1, \lambda_2 \in \{10^{-5}, ..., 10^3\}$)[2]. Specifically, for each OT method we use different values of regularization parameters and we report the best accuracy achieved by that method. We observe, i) OT-SON yields the highest accuracy scores, and ii) it is significantly more robust to variation of the regularization parameters, in comparison to the other methods. Moreover, the other methods are prone to yielding numerical errors for small regularizations.

model	real1	real2	real3	real4
OT-SON	**0.550**	**0.564**	**0.608**	**0.628**
OT-l1l2	0.421	0.507	0.500	0.621
OT-lpl1	0.457	0.521	0.516	0.592
OT-Sinkhorn	0.414	0.521	0.508	0.621

Table 1. Accuracy scores on MNIST and USPS.

5.1.2. CALTECH OFFICE

A commonly used dataset for domain adaptation is an object recognition dataset known as *Caltech Office* (Saenko et al., 2010; Griffin et al., 2006) which consists of four different domains: A (Amazon, 958 samples) , W (Webcam, 295 samples), C (Caltech, 1123 samples), and D (DSLR, 157 samples). These domains have 10 classes of objects in common that are represented with two sets of features: SURF (800 features) and DeCAF (4096 features). Similar to the setting in (Courty et al., 2017), we use all possible pairs of the four domains as source and target with SURF features, and we remove the classes 3, 5, and 7 from the target domain. Similar to the previous study, we assume imbalanced source and target classes in order to make the task more realistic. Table 2 compares the classification accuracies achieved by our method with the scores obtained by the three other methods. We calculate the accuracy similar to Section 5.1.1. In these experiments we use class information in the source domain together with Gaussian kernels. Specifically, we set $R_{l,k} = 0$ if y_l^s and y_k^s are not in the same classes and otherwise we set $R_{l,k} = \lambda \exp(-\|y_l^s - y_k^s\|^2)$ for different values of λ. We observe that our method yields the best results in seven cases. In other cases, our method is still competitive compared to the alternatives. In addition, we conclude that the different methods may perform differently on different datasets. Even though the setting of imbalanced classes is more important in practice and we have focused more on that, for the sake of completeness, we also compare our method with the three other methods in a setting where the classes are balanced. We use the same number of classes in the source (W) and target (C) domains. The accuracy results for the different methods are respectively: i) OT-SON: 0.260, ii) OT-l1l2: 0.244, iii) OT-lpl1: 0.247, and iv) OT-Sinkhorn: 0.243. We observe that even in this setting, our method yields the highest accuracy.

[2]For OT-l1l2 and OT-lpl1, λ_1 is the entropic regularization parameter and λ_2 is class regularization parameter.

S→T	OT-SON	OT-l1l2	OT-lpl1	OT-Sinkhorn
A→C	0.3552	0.3229	**0.3565**	0.3018
A→D	0.3274	0.3097	**0.3539**	0.3008
A→W	**0.3144**	0.2577	0.3092	0.2422
C→A	**0.4601**	0.3714	0.4120	0.3548
C→D	0.3982	0.3274	**0.4424**	0.3362
C→W	**0.3556**	0.2474	0.3144	0.2474
D→A	**0.3248**	0.2812	**0.3248**	0.2812
D→C	0.2720	0.2658	**0.2956**	0.2621
D→W	0.7216	**0.7628**	0.5876	0.7525
W→A	**0.2661**	0.2225	0.2616	0.2210
W→C	**0.2149**	0.1962	0.2124	0.2012
W→D	**0.8230**	0.7964	0.6637	0.8141

Table 2. Accuracy scores on Caltech Office (S: source, T: target).

6. Conclusion

We introduced a novel theoretical framework for OT in the presence of a multi-class structure in the two domains. Accordingly, we provided first theoretical guarantees for the recovery of the class structure, and developed constrained incremental algorithms which are generally suitable for non-smooth problems and enjoy theoretical convergence guarantees. Our experimental studies have substantiated the effectiveness of our proposed approach in different illustrative settings and datasets.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. We would like to thank the anonymous reviewers for their constructive comments.

References

Alaya, M. Z., Berar, M., Gasso, G., and Rakotomamonjy, A. (2019). Screening sinkhorn algorithm for regularized optimal transport. In Wallach, H., Larochelle, H., Beygelzimer, A., d Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 12169–12179. Curran Associates, Inc.

Altschuler, J., Weed, J., and Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 1964–1974.

Alvarez-Melis, D., Jaakkola, T., and Jegelka, S. (2018). Structured optimal transport. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1771–1780. PMLR.

Asadulaev, A., Korotin, A., Egiazarian, V., and Burnaev, E. (2022). Neural optimal transport with general cost functionals. *arXiv preprint arXiv:2205.15403*.

Balaji, Y., Chellappa, R., and Feizi, S. (2020). Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems (NeurIPS) 2020*, abs/2010.05862.

Bertsekas, D. (2011). Incremental proximal methods for large scale convex optimization. *Math. Program.*, 129(163).

Blondel, M., Seguy, V., and Rolet, A. (2018). Smooth and sparse optimal transport. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 880–889. PMLR.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *Siam Reviews*, 60(2):223–311.

Cao, Z., Ma, L., Long, M., and Wang, J. (2018). Partial adversarial domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Chang, H. and Yeung, D. (2008). Robust path-based spectral clustering. *Pattern Recognition*, 41(1):191–203.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F. (2018). Scaling algorithms for unbalanced optimal transport problems. *Math. Comput.*, 87(314):2563–2609.

Combettes, P. L. and Pesquet, J.-C. (2011). Proximal splitting methods in signal processing. In Bauschke, H. H., Burachik, R. S., Combettes, P. L., Elser, V., Luke, D. R., and Wolkowicz, H., editors, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer, New York.

Condat, L. (2016). Fast projection onto the simplex and the l1 ball. *Math. Program.*, 158(1-2):575–585.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2017). Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(9):1853–1865.

Cuturi, M. (2013). Sinkhorn distances:lightspeed computation of optimal transport. In *Adv. in Neural Information Processing Systems*, pages 2292–2300.

Defazio, A. (2016). A simple practical accelerated method for finite sums. *Advances in Neural Information Processing Systems 29 (NIPS 2016)*.

Dessein, A., Papadakis, N., and Rouas, J.-L. (2018). Regularized optimal transport and the rot mover's distance. *J. Mach. Learn. Res.*, 19(1):590–642.

Duchi, J. C., Shalev-Shwartz, S., Singer, Y., and Chandra, T. (2008). Efficient projections onto the l_1-ball for learning in high dimensions. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, pages 272–279.

Genevay, A., Cuturi, M., Peyre, G., and Bach, F. (2016). Stochastic optimization for large-scale optimal transport. In *Adv. in Neural Information Processing Systems*.

Gordaliza, P., Del Barrio, E., Fabrice, G., and Loubes, J.-M. (2019). Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pages 2357–2365. PMLR.

Griffin, G., Holub, A., and Perona, P. (2006). Caltech256 image dataset.

Guo, W., Ho, N., and Jordan, M. (2020a). Fast algorithms for computational optimal transport and wasserstein barycenter. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2088–2097. PMLR.

Guo, W., Ho, N., and Jordan, M. (2020b). Fast algorithms for computational optimal transport and wasserstein barycenter. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2088–2097. PMLR.

Hocking, T., Vert, J., Bach, F. R., and Joulin, A. (2011). Clusterpath: an algorithm for clustering using convex fusion penalties. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 745–752.

Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D., and Rohde, G. K. (2017). Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Process. Mag.*, 34(4):43–59.

Lindsten, F., Ohlsson, H., and Ljung, L. (2011a). Clustering using sum-of-norms regularization: With application to particle filter output computation. In *IEEE Statistical Signal Pro- cessing Workshop (SSP)*.

Lindsten, F., Ohlsson, H., and Ljung, L. (2011b). Clustering using sum-of-norms regularization: With application to particle filter output computation.

Long, T., Sun, Y., Gao, J., Hu, Y., and Yin, B. (2022). Video domain adaptation based on optimal transport in grassmann manifolds. *Information Sciences*, 594:151–162.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Ott, F., Rügamer, D., Heublein, L., Bischl, B., and Mutschler, C. (2022). Domain adaptation for time-series classification to mitigate covariate shift. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5934–5943.

Panahi, A., Dubhashi, D., Johansson, F. D., and Bhattacharyya, C. (2017). Clustering by sum of norms: Stochastic incremental algorithm, convergence and cluster recovery. In *International Conference on Machine Learning*, pages 2769–2777.

Papadakis, N., Peyré, G., and Oudet, E. (2014). Optimal transport with proximal splitting. *SIAM Journal on Imaging Sciences*, 7(1):212–238.

Parikh, N. and Boyd, S. (2016). Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231.

Patrascu, A. and Necoara, I. (2018). Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization. *Journal of Machine Learning Research*, 18(198):1–42.

Perrot, M., Courty, N., Flamary, R., and Habrard, A. (2016). Mapping estimation for discrete optimal transport. In *Advances in Neural Information Processing Systems*, pages 4197–4205.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Redko, I., Habrard, A., and Sebban, M. (2017). Theoretical analysis of domain adaptation with optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 737–753. Springer.

Saenko, K., Kulis, B., Fritz, M., and Darrell, T. (2010). Adapting visual category models to new domains. In Daniilidis, K., Maragos, P., and Paragios, N., editors, *Computer Vision – ECCV 2010*, pages 213–226, Berlin, Heidelberg. Springer Berlin Heidelberg.

Santambrogio, F. (2015). *Optimal transport for applied mathematicians*. Birkäuser, NY.

Schmidt, M., Le Roux, N., and Bach, F. (2017). Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112.

Seguy, V., Damodaran, B. B., Flamary, R., Courty, N., Rolet, A., and Blondel, M. (2018). Large scale optimal transport and mapping estimation. In *International Conference on Learning Representations*.

Solomon, J. (2018). Optimal transport on discrete domains. *CoRR*, abs/1801.07745.

Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

Wang, M. and Bertsekas, D. (2016). Stochastic first-order methods with random constraint projection. *SIAM J. Optimization*, 26(1):681–717.

A. Extension of Theorem 3.2

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:

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

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:

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

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:

$$\begin{aligned}&\min_{Y_{\alpha,\beta}\geq 0} \sum_{\alpha,\beta} Y_{\alpha,\beta} D_{\alpha,\beta} \\ &\text{s.t} \\ &q_\beta:\ \mathbf{1}^T\mathbf{y}^\beta = \sigma^\beta, p_\alpha : \mathbf{1}^T\mathbf{y}_\alpha = \sigma_\alpha\end{aligned} \tag{A.3}$$

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:

$$\begin{gathered}\min_{X_{\alpha,\beta}\geq 0} \sum_{\alpha,\beta} n_\alpha m_\beta X_{\alpha,\beta} D_{\alpha,\beta} + \\ \lambda \left(\sum_{\alpha,\alpha'} R_{\alpha,\alpha'} \|\mathbf{x}_\alpha - \mathbf{x}_{\alpha'}\|_M + \sum_{\beta,\beta'} S_{\beta,\beta'} \|\mathbf{x}^\beta - \mathbf{x}^{\beta'}\|_N \right) \\ \text{s.t.} \\ \mathbf{a}_M^T \mathbf{x}_\alpha = \mu_\alpha, \quad \mathbf{a}_N^T \mathbf{x}^\beta = \nu^\beta \end{gathered} \tag{A.4}$$

which in this case, coincides with equation A.2 when $\theta = \infty$. We also define

$$\begin{gathered} T_\alpha = \frac{1}{n_\alpha m_{\pi(\alpha)}} \sum_{\alpha' \neq \alpha} \frac{R_{\alpha,\alpha'} \frac{\sigma_\alpha}{n_\alpha m_{\pi(\alpha)}}}{\sqrt{\left(\frac{\sigma_\alpha}{n_\alpha m_{\pi(\alpha)}}\right)^2 + \left(\frac{\sigma_{\alpha'}}{n_{\alpha'} m_{\pi(\alpha')}}\right)^2}}, \\ U^\beta = \frac{1}{n_{\pi^{-1}(\beta)} m_\beta} \times \\ \sum_{\beta' \neq \beta} \frac{S_{\beta,\beta'} \frac{\sigma^\beta}{n_{\pi^{-1}(\beta)} m_\beta}}{\sqrt{\left(\frac{\sigma^\beta}{n_{\pi^{-1}(\beta)} m_\beta}\right)^2 + \left(\frac{\sigma^{\beta'}}{n_{\pi^{-1}(\beta')} m_{\beta'}}\right)^2}} \\ \Lambda_{\alpha,\beta} = \Bigg(T_\alpha + U^\beta + \\ \frac{1}{n_\alpha m_\beta} \frac{R_{\alpha,\pi^{-1}(\beta)} \frac{\sigma^\beta}{n_{\pi^{-1}(\beta)} m_\beta} + S_{\beta,\pi(\alpha)} \frac{\sigma_\alpha}{n_\alpha m_{\pi(\alpha)}}}{\sqrt{\left(\frac{\sigma_\alpha}{n_\alpha m_{\pi(\alpha)}}\right)^2 + \left(\frac{\sigma^{\beta'}}{n_{\pi^{-1}(\beta')} m_{\beta'}}\right)^2}} \Bigg)^{-1} \end{gathered} \tag{A.5}$$

Then, we have the following more general result:

Theorem A.1.

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*

$$\sum_{l=1}^{k} \tilde{D}_{i_l i_{l+1}} \geq \sum_{l=1}^{k} \tilde{D}_{i_l i_l} + k\delta. \tag{A.6}$$

Then, the solution $X_{\alpha,\beta}$ of the characteristic optimization in equation A.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 $\delta_{.,.}$ being the Kronecker index, if*

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

   (b) *Otherwise,*

$$\begin{gathered} \delta \sum_{\beta \neq \pi(\alpha)} X_{\alpha,\beta} \leq \\ \lambda \sum_{\alpha \neq \alpha'} \Bigg(\frac{R_{\alpha,\alpha'}}{n_\alpha n_{\alpha'}} \sqrt{\frac{n_{\alpha'}^2 \sigma_\alpha^2}{m_{\pi(\alpha)}} + \frac{n_\alpha^2 \sigma_{\alpha'}^2}{m_{\pi(\alpha')}}} \\ + \frac{S_{\pi(\alpha),\pi(\alpha')}}{m_{\pi(\alpha)} m_{\pi(\alpha')}} \sqrt{\frac{m_{\pi(\alpha')}^2 \sigma_\alpha^2}{n_\alpha} + \frac{m_{\pi(\alpha)}^2 \sigma_{\alpha'}^2}{n_{\alpha'}}} \Bigg) \\ + \frac{\theta}{2} \left(\sum_\alpha \frac{\delta_\alpha^2}{n_\alpha} + \sum_\alpha \frac{\delta_\alpha^2}{m_{\pi(\alpha)}} \right) + \frac{\Delta_1^2 n}{\theta} + \Delta_0 \left(\|\boldsymbol{\delta}\|_1 - \|\boldsymbol{\delta}\|_\infty \right) \end{gathered}$$

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 A.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 $j, j' \in D_\beta$,*

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

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

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

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

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

Proof. Denote the optimal value of equation A.3 and equation A.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 A.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 A.2 coincides with that of equation A.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 A.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 A.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 A.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

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

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

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

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

We conclude that

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

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

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

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

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

Lemma A.2 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 $j, j' \in D_\beta$

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

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

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 A.1 under conditions of the theorem and noticing that by the root-means-square and arithmetic mean (RMS-AM) inequality, we also have

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

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

□

Lemma A.2. *Suppose that the ideal optimization in equation A.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 A.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 $\alpha_0 \in \arg\min_\alpha |\delta_\alpha|$. Hence, $\|\boldsymbol{\delta}\|_1 - \|\boldsymbol{\delta}\|_\infty = \sum_{\alpha \neq \alpha_0} |\delta_\alpha|$.

Denote the optimal value of equation A.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

$$\begin{gathered} C_1 = \min_{Y_{\alpha,\beta} \geq 0} \sum_{\alpha,\beta} Y_{\alpha,\beta} D_{\alpha,\beta} \\ \text{s.t} \\ \mathbf{1}^T \mathbf{y}^\beta = \sigma^\beta + \delta^\beta, \mathbf{1}^T \mathbf{y}_\alpha = \sigma_\alpha + \delta_\alpha \end{gathered} \tag{A.7}$$

We notice that $\{p_\alpha, q_\beta\}$ are feasible dual vectors for equation A.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

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

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

□

A.1. Proof of Theorem 3.2

The first claim that $X_{ij} = X_{\alpha,\beta}$ follows by specializing part 2 of Theorem A.1 for the conditions of Theorem 3.2 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 $\theta = \infty, \boldsymbol{\delta} = \mathbf{0}$.

A.2. Proof of Theorem 3.1

Based on theorem 3.2, we present a sketch of the proof for theorem 3.1. Under the assumptions of theorem 3.1, we directly verify that $\Lambda = \Lambda_{\alpha,\beta} = \sqrt{2}/K(1+R)$. For part 1) $\delta = D - d - 2\omega,\ \Delta \leq 2\omega$ and for part 2) $\delta = D^2 - d^2$ are valid choices. Finally, for part 2) we may conclude by Chernoff bound that with a probability exceeding $1 - 1/n^{10}$ (the power 10 is arbitrary) we have $\Delta = O(\omega\sqrt{(E+\omega)^2 + 1}\log(nK))$. Replacing these expression in the first part of theorem 3.2 gives us the result.

B. Proof of Theorem 4.1

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 $\bar{\mathbf{g}}_t = \sum_{r=0}^{R} \mathbf{g}_r$ 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 Lyapunov function approach. We introduce two non-negative functions L, M of the state variables, $\mathbf{x}$ and $\{\mathbf{g}_r\}$ such that

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

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

$$\mathbb{E}[L_t] - L_0 + \sum_{\tau=0}^{t-1} \mathbb{E}[M_\tau] \leq 0 \tag{B.2}$$

which by the non-negativity of L implies

$$\sum_{\tau=0}^{t-1} \mathbb{E}[M_\tau] \leq L_0. \tag{B.3}$$

In particular, we take

$$M_t = F_t + \tfrac{1-\rho}{2\mu(1+\rho)} \sum_r \left\| \mathbf{x}_t - \mathbf{x}_{r,t}^\dagger \right\|^2 + \tfrac{\mu}{\rho} \left[\tfrac{2+\alpha}{2} - \tfrac{\alpha^2}{1-\rho} \right] \|\bar{\mathbf{g}}_t\|^2 \tag{B.4}$$

where

$$F_t = \sum_{r=1}^{P} \left[\phi_r \left(\mathbf{x}_{r,t}^\dagger \right) - \phi_r(\mathbf{x}^*) - \langle \mathbf{g}_r^*, \mathbf{x}_{r,t}^\dagger - \mathbf{x}^* \rangle \right] \tag{B.5}$$

and $\mathbf{g}_r^* \in \partial \phi_r(\mathbf{x}^*)$ for $r \in [R]$ satisfy the monotone inclusion problem in equation 4.6 at $\mathbf{x}^*$, i.e $\sum_r \mathbf{g}_r^* = 0$. 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

$$\begin{aligned} \Gamma_t &= \sum_{r=1}^{R} \|\mathbf{g}_r - \mathbf{g}_r^*\|^2, \quad G_t = \|\mu \bar{\mathbf{g}}_t + \rho(\mathbf{x}_t - \mathbf{x}^*)\|^2, \\ D_t &= \|\mathbf{x}^* - \mathbf{x}_t\|^2 \end{aligned} \tag{B.6}$$

and take

$$L_t = \frac{R}{2\mu} D_t + \frac{1}{2\rho\mu\alpha} G_t + \frac{R\mu}{2\rho} \Gamma_t \tag{B.7}$$

B.1. Proof of Theorem Under equation B.1

Let us first prove the theorem assuming equation B.1 holds true for the given L, M. Then equation B.3 also holds true and we conclude from the definition of M that

$$\begin{aligned} &\frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}[F_\tau] \leq \frac{L_0}{t}, \\ &\frac{1-\rho}{2t(1+\rho)} \sum_{\tau=0}^{t-1} \sum_r \mathbb{E} \left[\left\| \mathbf{x}_t - \mathbf{x}_{r,t}^\dagger \right\|^2 \right] \leq \frac{\mu L_0}{t} \end{aligned} \tag{B.8}$$

Now from equation B.5 and the triangle inequality, we conclude that

$$\begin{aligned} \mathbb{E}[F_\tau] &\geq \mathbb{E} \left[\sum_{r=1}^{P} \phi_r(\mathbf{x}_\tau) - \phi_r(\mathbf{x}^*) \right] \\ &- \sum_{r=1}^{P} \mathbb{E} \left| \phi_r \left(\mathbf{x}_{r,\tau}^\dagger \right) - \phi_r(\mathbf{x}_\tau) \right| - \\ &\sum_{r=1}^{P} \mathbb{E} \left| \langle \mathbf{g}_r^*, \mathbf{x}_{r,\tau}^\dagger - \mathbf{x}_\tau \rangle \right| \end{aligned} \tag{B.9}$$

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

$$\begin{aligned} &\sum_{r=1}^{P} \mathbb{E} \left| \langle \mathbf{g}_r^*, \mathbf{x}_{r,\tau}^\dagger - \mathbf{x}_\tau \rangle \right| \leq \beta \sum_{r=1}^{P} \mathbb{E} \left\| \mathbf{x}_{r,\tau}^\dagger - \mathbf{x}_\tau \right\| \\ &\leq \beta \sqrt{P} \sqrt{\sum_{r=1}^{P} \mathbb{E} \left\| \mathbf{x}_{r,\tau}^\dagger - \mathbf{x}_\tau \right\|^2} \end{aligned} \tag{B.10}$$

Similarly,

$$\sum_{r=1}^{P} \mathbb{E}\left|\phi_r\left(\mathbf{x}_{r,\tau}^{\dagger}\right) - \phi_r(\mathbf{x}_\tau)\right| \leq \beta \sum_{r=1}^{P} \mathbb{E}\left\|\mathbf{x}_{r,\tau}^{\dagger} - \mathbf{x}_\tau\right\| \\ \leq \beta\sqrt{P}\sqrt{\sum_{r=1}^{P} \mathbb{E}\left\|\mathbf{x}_{r,\tau}^{\dagger} - \mathbf{x}_\tau\right\|^2} \tag{B.11}$$

We conclude that

$$\mathbb{E}\left[\sum_{r=1}^{P} \phi_r(\mathbf{x}_\tau) - \phi_r(\mathbf{x}^*)\right] \\ \leq \mathbb{E}[F_\tau] + 2\beta\sqrt{P}\sqrt{\sum_{r=1}^{P} \mathbb{E}\left\|\mathbf{x}_{r,\tau}^{\dagger} - \mathbf{x}_\tau\right\|^2} \tag{B.12}$$

and hence,

$$\tfrac{1}{t}\sum_{\tau=0}^{t-1} \mathbb{E}\left[\sum_{r=1}^{P} \phi_r(\mathbf{x}_\tau) - \phi_r(\mathbf{x}^*)\right] \\ \leq \tfrac{1}{t}\sum_{\tau=0}^{t-1} \mathbb{E}[F_\tau] + \tfrac{2\beta\sqrt{P}}{t}\sum_{\tau=0}^{t-1}\sqrt{\sum_{r=1}^{P} \mathbb{E}\left\|\mathbf{x}_{r,\tau}^{\dagger} - \mathbf{x}_\tau\right\|^2} \\ \leq \tfrac{1}{t}\sum_{\tau=0}^{t-1} \mathbb{E}[F_\tau] + \\ 2\beta\sqrt{P}\sqrt{\tfrac{1}{t}\sum_{\tau=0}^{t-1}\sum_{r=1}^{P} \mathbb{E}\left\|\mathbf{x}_{r,\tau}^{\dagger} - \mathbf{x}_\tau\right\|^2} \tag{B.13}$$

Using Jensen's inequality and equation B.8, we obtain

$$\mathbb{E}\left[\sum_{r=1}^{P} \phi_r(\bar{\mathbf{x}}_\tau) - \phi_r(\mathbf{x}^*)\right] \leq \\ \frac{1}{t}\sum_{\tau=0}^{t-1} \mathbb{E}\left[\sum_{r=1}^{P} \phi_r(\mathbf{x}_\tau) - \phi_r(\mathbf{x}^*)\right] \leq \\ \frac{L_0}{t} + 2\beta\sqrt{\frac{2P\mu L_0(1+\rho)}{t(1-\rho)}} \tag{B.14}$$

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 4.7

$$L_0 \leq c\beta P\lambda. \tag{B.15}$$

For the second bound in part 1 note that for $r = P+1, P+2, \ldots, P+Q$, 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

$$\sum_{q=1}^{Q} \operatorname{dist}^2(\bar{\mathbf{x}}_t, S_q) \leq \frac{1}{t}\sum_{\tau=0}^{t-1}\sum_{r=P+1}^{R} \|\mathbf{x}_{r,t}^{\dagger} - \mathbf{x}_t\|^2 \leq \frac{\mu L_0}{t} \tag{B.16}$$

For part 2, note that

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

We conclude from equation B.3 that

$$\sum_t \mathbb{E}\left[\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2\right] \leq \frac{\mu L_0}{R} \tag{B.17}$$

which completes the proof.

B.2. Proof of equation B.1

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

Lemma B.1. *The average dynamics of G_t is given by:*

$$\begin{aligned}\mathbb{E}[G_{t+1}]-\mathbb{E}[G_t] &= -\mu^2\left[1-(1-\alpha)^2\right]\mathbb{E}\|\bar{\mathbf{g}}_t\|^2 \\ &+ 2\rho\mu\alpha\mathbb{E}\left\langle \mathbf{x}^*-\mathbf{x}_t, \bar{\mathbf{g}}_t\right\rangle\end{aligned} \tag{B.18}$$

Lemma B.2. *The average dynamics of Γ_t is given by:*

$$\begin{aligned}\mathbb{E}[\Gamma_{t+1}]-\mathbb{E}[\Gamma_t] &= \frac{1}{R}\sum_r \mathbb{E}\left\|\rho\frac{\mathbf{x}_t-\mathbf{x}^\dagger_{r,t}}{\mu}-\alpha\bar{\mathbf{g}}_t\right\|^2 \\ &+ \frac{2\rho}{R\mu}\sum_r \mathbb{E}\left\langle \mathbf{x}_t-\mathbf{x}^\dagger_{r,t}, \mathbf{g}_r-\mathbf{g}^*_r\right\rangle - \frac{2\alpha}{R}\|\bar{\mathbf{g}}_t\|^2\end{aligned} \tag{B.19}$$

Lemma B.3. *The average dynamics of D_t is given by:*

$$\begin{aligned}\mathbb{E}[D_{t+1}]-\mathbb{E}[D_t] &= -\frac{1}{R}\sum_r \mathbb{E}\|\mathbf{x}^\dagger_{r,t}-\mathbf{x}_t\|^2 \\ &+ \frac{2}{R}\sum_r \mathbb{E}\langle \mathbf{x}_t-\mathbf{x}^\dagger_{r,t}, \mathbf{x}^*-\mathbf{x}^\dagger_{r,t}\rangle\end{aligned} \tag{B.20}$$

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

Lemma B.4. *The following inequality holds at every time:*

$$F_t+\sum_{r=1}^{R}\left\langle \frac{\mathbf{x}_t-\mathbf{x}^\dagger_{r,t}}{\mu}+\mathbf{g}_r-\mathbf{g}^*_r, \mathbf{x}^*-\mathbf{x}^\dagger_{r,t}\right\rangle \leq 0 \tag{B.21}$$

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

$$\phi_r(\mathbf{x}^*) \geq \phi_r(\bar{\mathbf{x}}^\dagger_{r,t}) + \left\langle \frac{\mathbf{x}_t-\bar{\mathbf{x}}^\dagger_{r,t}}{\mu}+\mathbf{g}_r, \mathbf{x}^*-\bar{\mathbf{x}}^\dagger_{r,t}\right\rangle \tag{B.22}$$

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

$$F_t+\sum_{r=1}^{P}\left\langle \frac{\mathbf{x}_t-\mathbf{x}^\dagger_{r,t}}{\mu}+\mathbf{g}_r-\mathbf{g}^*_r, \mathbf{x}^*-\mathbf{x}^\dagger_{r,t}\right\rangle \leq 0 \tag{B.23}$$

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

$$\left\langle \frac{\mathbf{x}_t-\mathbf{x}^\dagger_{r,t}}{\mu}+\mathbf{g}_r, \mathbf{x}^*-\mathbf{x}^\dagger_{r,t}\right\rangle \leq 0 \tag{B.24}$$

Similarly, we obtain that

$$\langle \mathbf{g}^*_r, \bar{\mathbf{x}}^\dagger_{r,t}-\mathbf{x}_t\rangle \leq 0 \tag{B.25}$$

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

B.2.1. COMBINING BOUNDS

To obtain equation B.1, we combine the inequalities in the above four lemmas in the following way. Respectively multiplying equation B.20, equation B.18 and equation B.19 by $\frac{R}{2\mu}$, $\frac{1}{2\rho\mu\alpha}$ and $\frac{R\mu}{2\rho}$ and adding to equation B.21 and after straightforward calculations, we obtain:

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

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.

B.3. Proof of Theorem 4.2

The proximal operator of $\phi_{k,\boldsymbol{\zeta},\boldsymbol{\eta}}$ is defined as

$$\underset{(\mathbf{x},\mathbf{y})\in\mathbb{R}^m\times\mathbb{R}^m}{\operatorname{argmin}} \frac{1}{2\mu}\|\mathbf{x}-\mathbf{p}\|_2^2 + \frac{1}{2\mu}\|\mathbf{y}-\mathbf{q}\|_2^2 + \phi_{\rho,\boldsymbol{\zeta},\boldsymbol{\eta}}(\mathbf{x},\mathbf{y}). \tag{B.26}$$

We introduce a change of variables by $\mathbf{u} = (\mathbf{x}+\mathbf{y})/2$, $\mathbf{v} = (\mathbf{x}-\mathbf{y})/2$. First note that $\|a\|^2 + \|b\|^2 = (\|a+b\|^2 + \|a-b\|^2)/2$. Hence

$$\frac{1}{2\mu}\|\mathbf{x}-\mathbf{p}\|_2^2 + \frac{1}{2\mu}\|\mathbf{y}-\mathbf{q}\|_2^2 = \frac{1}{\mu}\left(\left\|\mathbf{u} - \frac{\mathbf{p}+\mathbf{q}}{2}\right\|^2 + \left\|\mathbf{v} - \frac{\mathbf{p}-\mathbf{q}}{2}\right\|^2\right)$$

Furthermore,

$$\langle\mathbf{x},\boldsymbol{\zeta}\rangle + \langle\mathbf{y},\boldsymbol{\eta}\rangle = \langle\mathbf{u}+\mathbf{v},\boldsymbol{\zeta}\rangle + \langle\mathbf{u}-\mathbf{v},\boldsymbol{\eta}\rangle = \langle\mathbf{u},\boldsymbol{\zeta}+\boldsymbol{\eta}\rangle + \langle\mathbf{v},\boldsymbol{\zeta}-\boldsymbol{\eta}\rangle$$

Hence equation B.26 can be written as

$$\underset{(\mathbf{u},\mathbf{v})\in\mathbb{R}^m\times\mathbb{R}^m}{\operatorname{argmin}} \frac{1}{\mu}\left(\left\|\mathbf{u} - \frac{\mathbf{p}+\mathbf{q}}{2}\right\|^2 + \left\|\mathbf{v} - \frac{\mathbf{p}-\mathbf{q}}{2}\right\|^2\right) + \tag{B.27}$$

$$\begin{aligned}&\langle\mathbf{u},\boldsymbol{\zeta}+\boldsymbol{\eta}\rangle + \langle\mathbf{v},\boldsymbol{\zeta}-\boldsymbol{\eta}\rangle + 2\rho\|\mathbf{v}\|_2 \\ &= \underset{(\mathbf{u},\mathbf{v})\in\mathbb{R}^m\times\mathbb{R}^m}{\operatorname{argmin}} \frac{1}{\mu}\left(\left\|\mathbf{u} - \frac{\mathbf{p}+\mathbf{q}-\mu\boldsymbol{\zeta}-\mu\boldsymbol{\eta}}{2}\right\|^2\right) + \\ &\frac{1}{\mu}\left(\left\|\mathbf{v} - \frac{\mathbf{p}-\mathbf{q}-\mu\boldsymbol{\zeta}+\mu\boldsymbol{\eta}}{2}\right\|^2\right) + 2\rho\|\mathbf{v}\|_2\end{aligned} \tag{B.28}$$

This is separable over $\mathbf{u}$ and $\mathbf{v}$, and can be analytically solved. We get

$$\mathbf{u} = \frac{\mathbf{p}+\mathbf{q}-\mu\boldsymbol{\zeta}-\mu\boldsymbol{\eta}}{2},$$
$$\mathbf{v} = \mathcal{T}_{\mu\rho}\left(\frac{\mathbf{p}-\mathbf{q}-\mu\boldsymbol{\zeta}+\mu\boldsymbol{\eta}}{2}\right)$$

The result is obtained by setting $\mathbf{x} = \mathbf{u}+\mathbf{v}, \mathbf{y} = \mathbf{u}-\mathbf{v}$.

C. Additional experiments

C.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.

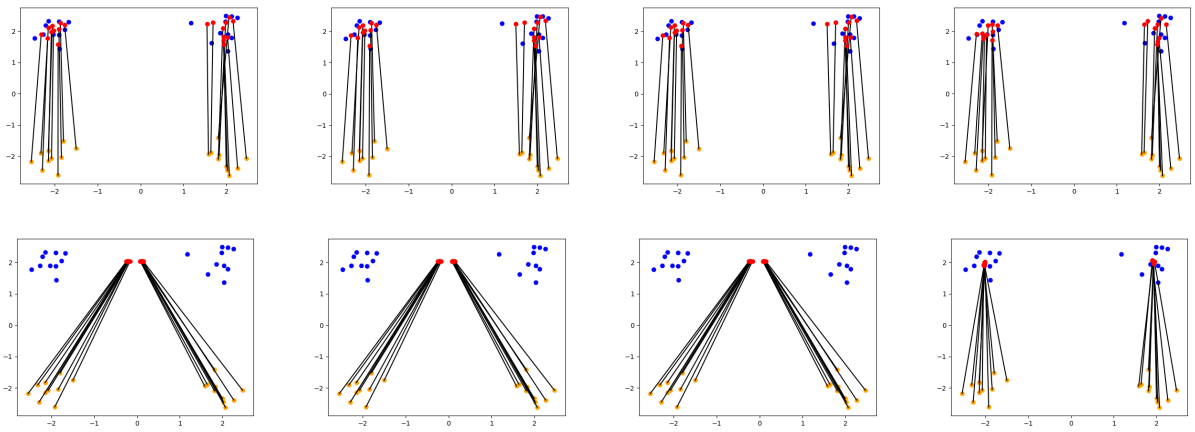


Figure 2. Illustration of different models on simple data, where the source and target domains have the same number of classes and similar distributions. The columns respectively correspond to OT-l1l2, OT-lpl1, OT-Sinkhorn and OT-SON. For each model, we illustrate the results for two different values 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.

C.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

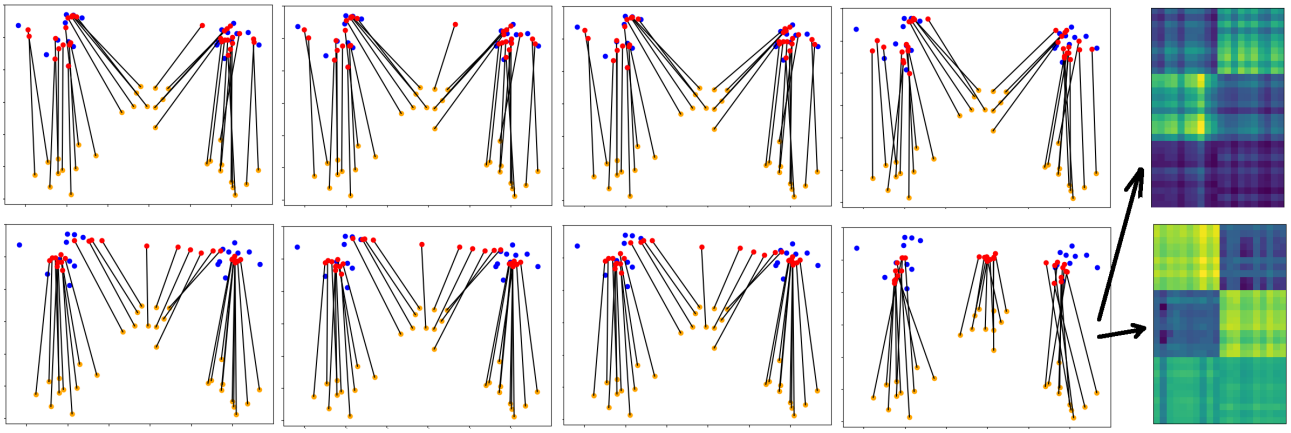
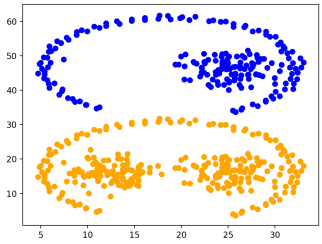
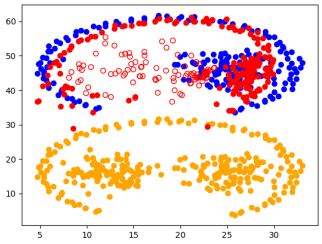


Figure 3. Illustration of different methods where the source and target domains have different number of classes. The first four columns respectively correspond to OT-l1l2, OT-lpl1, OT-Sinkhorn and OT-SON. Only OT-SON with an appropriate parameterization (the forth column and the second row) identifies 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.

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.

C.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.

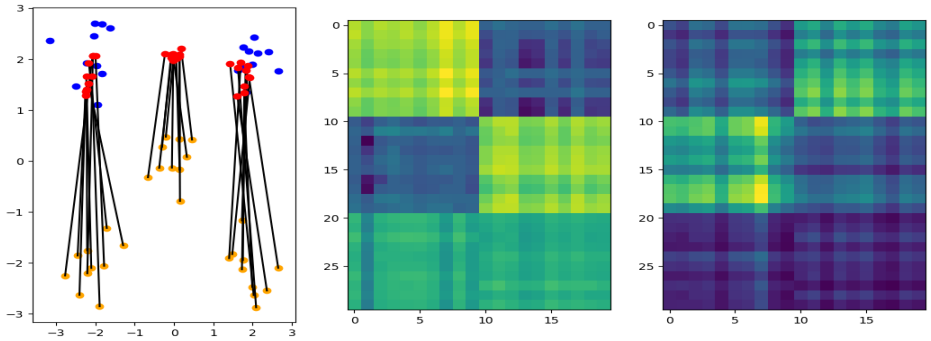


Figure 5. Unsupervised OT-SON, the OT-SON results and the consistency of transport costs and transport maps.

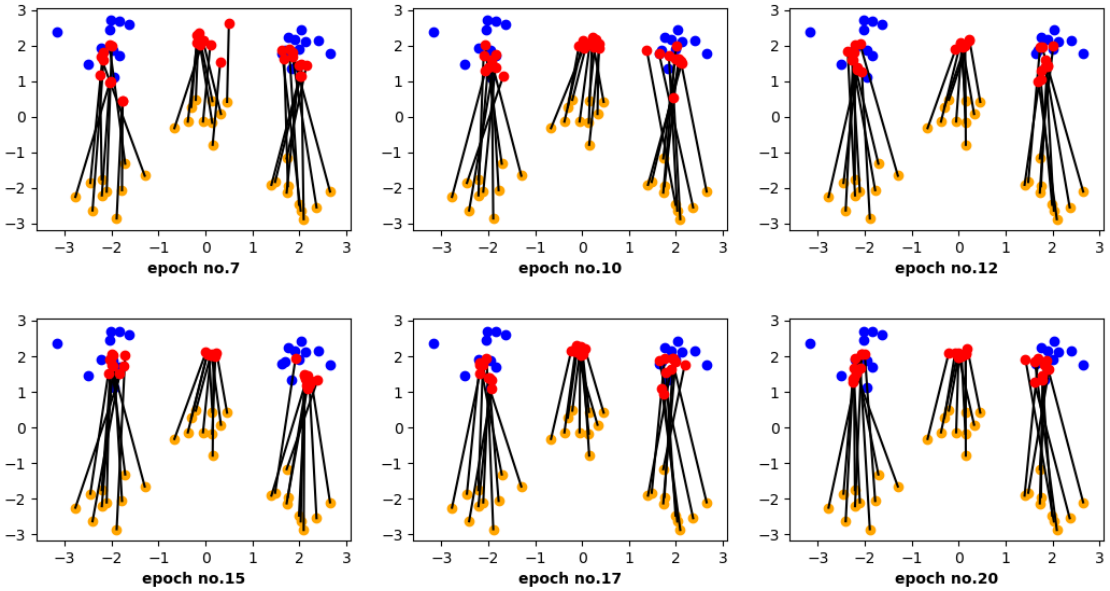


Figure 6. Early stopping of the optimization after a finite number of epochs. The results are very consistent and stable even of we stop the algorithm very early.

C.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.

C.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

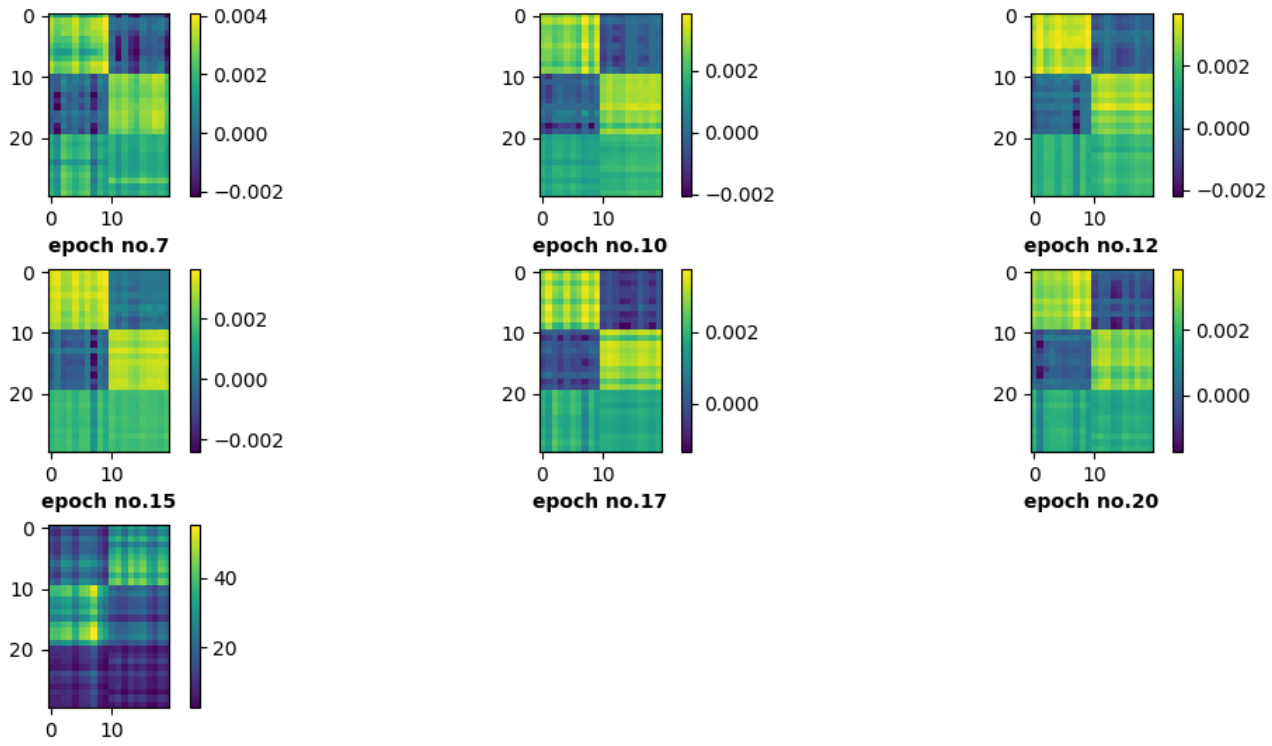


Figure 7. Consistency of the transport maps with the transport costs (shown at the last row) when using different finite number of epochs. Thus, early stopping can be useful for efficiency purposes.

$\lambda_1 = \lambda_2 = 0$, 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.

C.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 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).

D. Uniqueness

An elementary question concerning any optimization formulation, including the Kantorovich problem and its regularization in equation 2.3, is the uniqueness of their optimal solution, and a standard method for verifying uniqueness is to establish strong convexity of the objective function. Even though it is seen that the objective in (2.3) is not strongly convex, we are nevertheless able to identify conditions, under which the solution still remains unique. For this, we develop an alternative approach, which is not only useful in our framework, but may also be generically used in many similar problems including a

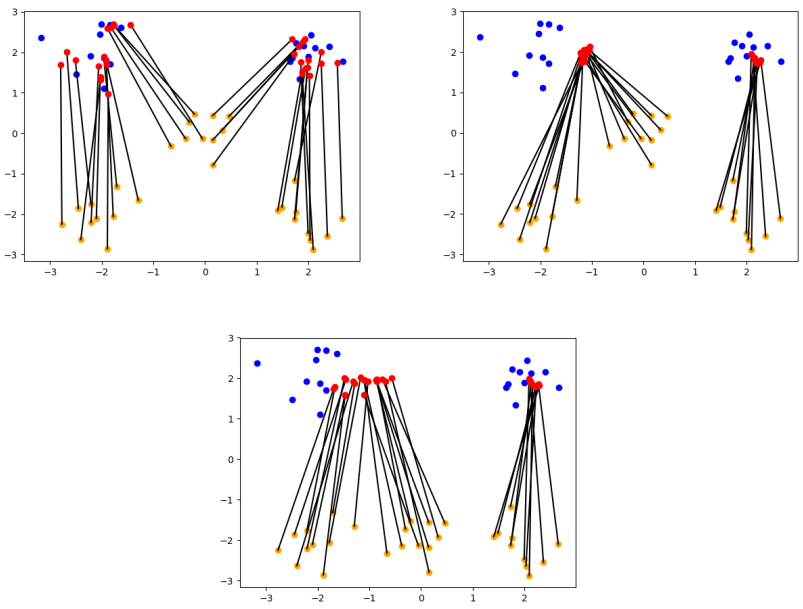


Figure 8. The impact of SON regularization when the class members are diverse. The plot on the left (where $\lambda_1 = \lambda_2 = 0$) performs transportation solely based on pairwise distances, thus fails to transfer the classes properly. Our SON regularization (either $\lambda_1 = 10, \lambda_2 =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

wide range of linear programming (LP) relaxation problems, and for this reason it is first presented. Our approach is based on the following definition:

Definition D.1. We call a (global) optimal solution $\mathbf{X}_0$ of a convex optimization problem

$$\min_{\mathbf{X}\in\mathscr{S}} \mathscr{F}(\mathbf{X}),$$

where $\mathscr{F}(.)$ is a convex function and $\mathscr{S}$ is a convex set, a **resistant optimal point** if adding a linear perturbation term $\langle\tilde{\mathbf{D}}, \mathbf{X}\rangle$ with sufficiently small coefficients in $\tilde{\mathbf{D}}$ to the objective leads to an arbitrarily small perturbation of the solution $\mathbf{X}_0$. In mathematical terms for any open neighborhood $\mathcal{N}$ of $\mathbf{X}_0$ there exists an open neighborhood $\mathcal{M}$ of $\tilde{\mathbf{D}}$ around $\tilde{\mathbf{D}} = \mathbf{0}$ such that

$$\forall\tilde{\mathbf{D}} \in \mathcal{M}, \quad \mathcal{N} \cap \arg\min_{\mathbf{X}\in\mathscr{S}} \mathscr{F}(\mathbf{X}) + \langle\tilde{\mathbf{D}}, \mathbf{X}\rangle \neq \emptyset.$$

Accordingly, we have the following result:

Theorem D.2. *A resistant optimal point of a convex optimization problem is its unique optimal point.*

Proof. Suppose that there exists a different optimal point $\mathbf{X}'$. Take $\mathbf{D}_0 = \frac{\mathbf{X}_0 - \mathbf{X}'}{\|\mathbf{X}_0 - \mathbf{X}'\|}$, $r = \|\mathbf{X}_0 - \mathbf{X}'\|$ and $\tilde{\mathbf{D}} = \epsilon\mathbf{D}_0$ for arbitrary $\epsilon > 0$. Further, define $\mathcal{N}$ as the ball of radius $\delta = r/2$ centered at $\mathbf{X}_0$. Note that for each $\mathbf{Y} \in \mathcal{N}$ we have

$$\mathscr{F}(\mathbf{Y}) + \langle\tilde{\mathbf{D}}, \mathbf{Y}\rangle \geq \mathscr{F}(\mathbf{X}_0) + \langle\tilde{\mathbf{D}}, \mathbf{Y}\rangle =$$

$$\mathscr{F}(\mathbf{X}') + \langle\tilde{\mathbf{D}}, \mathbf{X}'\rangle + \langle\tilde{\mathbf{D}}, (\mathbf{Y} - \mathbf{X}_0) + (\mathbf{X}_0 - \mathbf{X}')\rangle.$$

Now, note that $\langle\tilde{\mathbf{D}}, (\mathbf{Y} - \mathbf{X}_0) + (\mathbf{X}_0 - \mathbf{X}')\rangle \geq -\delta\epsilon + r\epsilon > 0$, which establishes

$$\mathscr{F}(\mathbf{Y}) + \langle\tilde{\mathbf{D}}, \mathbf{Y}\rangle > \mathscr{F}(\mathbf{X}') + \langle\tilde{\mathbf{D}}, \mathbf{X}'\rangle.$$

Hence, $\mathcal{N} \cap \arg\min_{\mathbf{X}\in\mathscr{S}} \mathscr{F}(\mathbf{X}) + \langle\tilde{\mathbf{D}}, \mathbf{X}\rangle = \emptyset$ and since $\epsilon = \|\tilde{\mathbf{D}}\|$ is arbitrarily small, we conclude that $\mathbf{X}_0$ is not a resistant optimal point. This contradicts the assumption and shows that the solution is unique. □

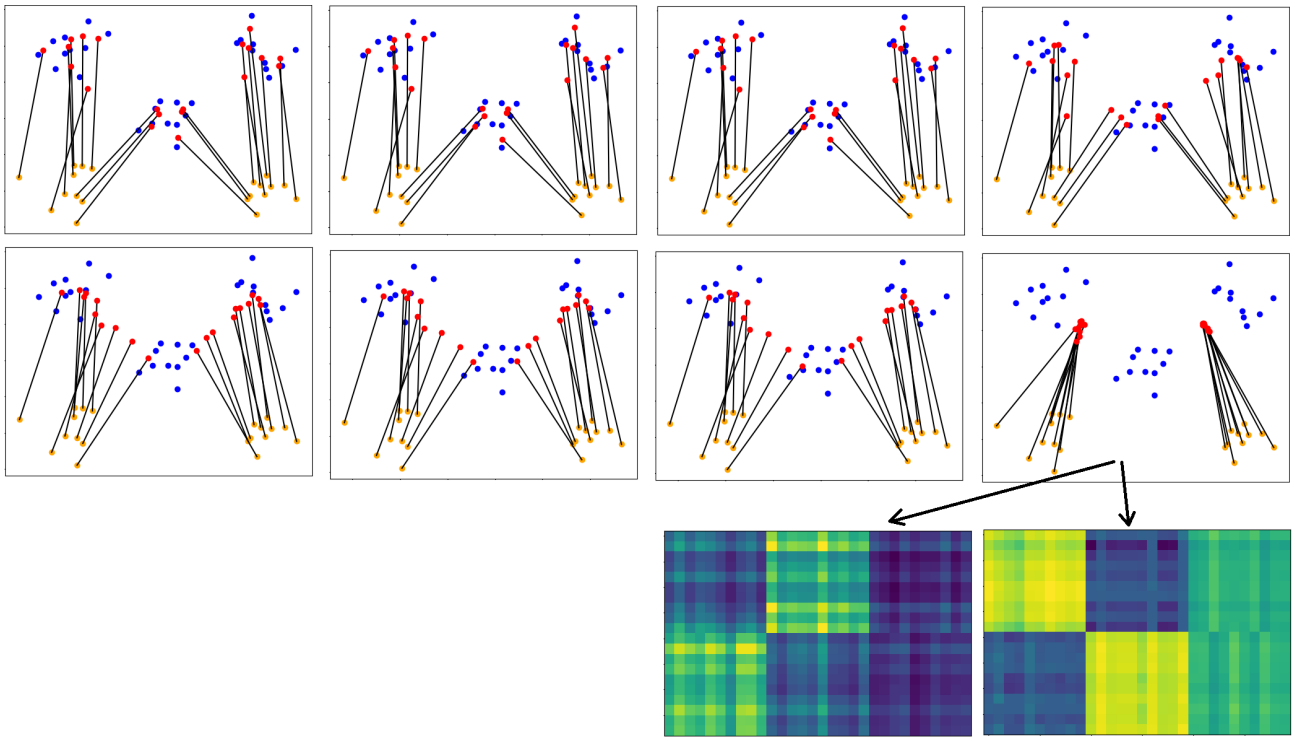


Figure 9. Performance of different methods when the source has two classes and the target consists of three classes. The columns in order represent OT-l1l2, OT-lpl1, OT-Sinkhorn and OT-SON. Among different methods, only OT-SON with high regularization prevents splitting the source data among all the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

Theorem D.2 is a general way to establish uniqueness. In fact, we can show that the strong convexity condition is a special case of this result:

Theorem D.3. *If $\mathscr{F}$ is continuous and strongly convex, then the global minimal point of $\mathscr{F}$ over a convex set $\mathscr{S}$ is resistant.*

Proof. Denote the optimal point by $\mathbf{X}^*$. By strong convexity, there exists a $\gamma > 0$ such that for any feasible point $\mathbf{X} \in \mathscr{S}$, we have $\mathscr{F}(\mathbf{X}) - \mathscr{F}(\mathbf{X}^*) \geq \frac{\gamma}{2}\|\mathbf{X} - \mathbf{X}^*\|_{\mathrm{F}}^2$. Take $\mathscr{G} = \mathscr{F} + \langle \tilde{\mathbf{D}}, \mathbf{X} \rangle$ and note that $\mathscr{G}(\mathbf{X}) - \mathscr{G}(\mathbf{X}^*) \geq \frac{\gamma}{2}\|\mathbf{X} - \mathbf{X}^*\|_{\mathrm{F}}^2 + \langle \tilde{\mathbf{D}}, \mathbf{X} - \mathbf{X}^* \rangle \geq \frac{\gamma}{4}\|\mathbf{X} - \mathbf{X}^*\|_{\mathrm{F}}^2 - \frac{2}{\gamma}\|\tilde{\mathbf{D}}\|_{\mathrm{F}}^2$. This shows that $\mathscr{G} > \mathscr{G}(\mathbf{X}^*)$ and hence does not have any global optimal point outside the closed sphere $\{\mathbf{X} \mid \|\mathbf{X} - \mathbf{X}^*\|_{\mathrm{F}} \leq \frac{\sqrt{8}}{\gamma}\|\tilde{\mathbf{D}}\|_{\mathrm{F}}\}$. Since $\mathscr{G}$ is continuous, it also attains a minimum inside the sphere, which then becomes the global optimal point. We conclude that for any $\epsilon > 0$, taking $\|\tilde{\mathbf{D}}\| < \frac{\gamma\epsilon}{\sqrt{8}}$ leads to an optimal solution inside a ball of radius ϵ centered at $\mathbf{X}^*$. This shows that the solution is resistant. □

Uniqueness for equation 2.3: One special case of resistant optimal points, that will be useful in our analysis, is when there exists a neighborhood $\mathcal{M}$ of $\mathbf{0}$ such that

$$\forall \tilde{\mathbf{D}} \in \mathcal{M}, \quad \mathbf{X}^* \in \arg\min_{\mathbf{X} \in \mathscr{S}} \mathscr{F}(\mathbf{X}) + \langle \tilde{\mathbf{D}}, \mathbf{X} \rangle.$$

We call such a resistant optimal point an **extremal optimal point**. Later, we consider an analysis where we give conditions on $\mathbf{D}$ to ensure that a desired solution $\mathbf{X}^*$ is achieved. Our strategy for uniqueness in this analysis is to show that under the same conditions, the desired optimal point is also extremal and hence unique, according to Theorem 1. In the case of the problem in equation 2.3, adding the term $\langle \tilde{\mathbf{D}}, \mathbf{X} \rangle$ modifies the cost matrix $\mathbf{D}$ to $\mathbf{D} + \tilde{\mathbf{D}}$. Hence, being an extremal optimal point is in this case equivalent to the solution $\mathbf{X}^*$ being maintained following a perturbation of the matrix $\mathbf{D}$ in a sufficiently small open neighborhood. This is easy to achieve in our planted model analysis, because the optimality of $\mathbf{X}^*$ is guaranteed by a set of inequalities on $\mathbf{D}$, which remain valid under small perturbations, simply by requiring the inequalities to be strict.

As seen, Theorem D.2 and extremal optimality, in particular, can be powerful tools for establishing uniqueness beyond strong convexity.

E. Remarks on the Optimization Algorithm

Efficient Computation: While the objective in (4.2) may appear complex as it involves n^2 terms, the associated algorithm is stochastic and incremental, thus only involving one term in (4.2) for each iteration, thus greatly reducing the complexity as a result. The simplification of the algorithm is also due to the proximal update detailed in Theorem 4.2 (and subsequent projection) used in each iteration update of a pair of rows or columns. We further note that an early stopping typical of stochastic schemes is likely, making a full-run to convergence unnecessary (see Section 4.4 in (Bottou et al., 2018)), and in practice avoiding the impact of the n^2 terms on the performance. When the underlying data satisfies the structure of the stochastic block model, the problem size is essentially $B^2 \ll n^2$, as the number of required iterations is determined by an adequate sampling of all blocks.

Just-in-Time Update: In our problem of interest in equation 2.3, the number of variables quadratically grows with the problem size. For such problems, incremental algorithms may become infeasible in large-scale. Note that each iteration of our algorithm includes proximal and projection operators, that update only a small group of variables. This allows us to apply the Just-in-Time approach in (Schmidt et al., 2017) to resolve the problem with the number of variables. In our problem, each term $\phi_n(x)$ and constraint S_m only involves a small subset $x_{I_n} := (x_i, i \in I_n)$ of the variables, where $I_n \subseteq [D]$. Hence, the projection and proximal operators alter only a small subset of variables, dramatically reducing the amount of computation. We exploit this to give an algorithm that has much cheaper per-iteration cost. Note that the vanilla algorithm explained in equation 4.4 and equation 4.5 still operates on the full set of variables as the memory vectors become non-sparse by the updating rule in equation 4.5. We resolve this issue by following the Just-in-Time approach in (Schmidt et al., 2017) and modifying equation 4.5 to

$$\mathbf{a}_t = \rho \frac{\mathbf{x}_t - \mathbf{x}_{t+1}}{\mu} - \alpha \left(\sum_n \mathbf{g}_n + \sum_m \mathbf{h}_m \right)_{I_t} \tag{E.1}$$

where I_t denotes the set of variables involved in the t^{th} iteration and we define $(\mathbf{y})_I$ for a vector $\mathbf{y} = (y_1, y_2, \ldots, y_d)$ as a vector $\mathbf{y}' = (y'_1, y'_2, \ldots, y'_d)$ such that

$$y'_i = \begin{cases} \frac{K y_i}{K_i} & i \in I \\ 0 & i \notin I \end{cases}$$

where $K = M + N$ and K_i is the number of objective terms ϕ_n and constraint sets S_m including the i^{th} variable x_i.