Stochastic Proximal Algorithms with SON Regularization: Towards Efficient Optimal Transport for Domain Adaptation

Ashkan Panahi
Arman Rahbar
Morteza Haghir Chehreghani
Devdatt Dubhashi
Department of Computer Science and Engineering
Chalmers University of Technology
Gothenburg, Sweden

Hamid Krim
Department of Electrical Engineering
North Carolina State University
Raleigh, NC 27606, USA

Abstract

We propose a new regularizer for optimal transport (OT) which is tailored to better preserve the class structure of the subjected process. Accordingly, we provide the first theoretical guarantees for an OT scheme that respects class structure. We derive an accelerated proximal algorithm with a closed form projection and proximal operator scheme thereby affording a highly 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 does not only result in a better preservation of the class structure but also in additional robustness relative to previous regularizers.

Keywords: Optimal Transport, Domain Adaptation, Regularization, Sum of Norm, Theoretical Guarantees, Geometric Analysis, Incremental Algorithms, Stochastic Averaging, Constrained Optimization

1. Introduction

Optimal Transport (OT), first proposed by Monge as an analysis problem Monge (1781), has become a classic topic in probability and statistics for transferring mass from one probability distribution to another Villani (2008); Santambrogio (2015). The OT problem seeks to find a transport map from a source distribution to a target distribution while minimizing the cost of the transport. As a richly adopted framework in many different disciplines, OT has also recently been very successfully used in many 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. This improvement, however, comes at a significant computational cost when solving the OT problem. Much attention has recently focused on
efficient computational and numerical algorithms for OT, and a monograph focusing on this topic recently appeared Peyré and Cuturi (2018).

Several advances in computational approaches to OT have been made in recent years, primarily focusing on applications in domain adaptation. In Courty et al. (2017), a generalized conditional gradient method is used to compute OT with the help of a couple of regularizers. 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). Another computational breakthrough was achieved by Genevay et al. (2016) who gave a stochastic incremental algorithm to solve the entropic regularized OT problem.

While the entropic regularization of OT has attracted a lot of attention on account of its many merits, it has some limitations, such as the blurring in the optimal transportation plan induced by the entropy term. An amelioration of this effect may be achieved by using a small regularization so long as it is carefully engineered. More importantly, the entropy term keeps the transportation plan strictly positive and therefore completely dense, unlike unregularized OT. This lack of sparsity can be problematic for applications where the optimal transportation plan is itself of interest as in domain adaptation Courty et al. (2017). For these applications, the principle of parsimony suggests that we should prefer to transfer one source class to as few target classes as possible. A case for exploring new regularizers was made in Courty et al. (2017) in the context of domain adaptation applications. In this paper, we accordingly propose a novel approach of class–based regularization of the OT problem, based on the recently proposed convex clustering framework of Sum of Norms (SON) Lindsten et al. (2011); Hocking et al. (2011), which presents an improvement on the state of the art on at least two grounds:

**SON Regularizer Benefits for OT:** The SON regularization allows one to exploit the class structure and to preserve the sparsity of the transport plan. While this approach may be superficially reminiscent of a Laplacian regularizer Courty et al. (2017), the latter only acts indirectly on the transported points and is quadratic in nature, in contrast to our transport plan. This difference is clearly illustrated in the experiments. We theoretically show and experimentally validate that this formulation ensures a transport plan adhering to the class structure. In the source domain, the class structure is given by the labels while it is latent (hidden) in the target domain. We further show that our formulation leads to the discovery of the underlying hidden class structure in the target domain, and provide for the first time, rigorous guarantees on the recovery of class structure. 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 in Courty et al. (2017).

**Computational Benefits of Stochastic Proximal Algorithm:** Our SON regularizer-based formulation also enjoys computational benefits – we propose a highly scalable stochastic incremental algorithm which operates in the primal formulation and explicitly produces the optimal coupling. In contrast to Courty et al. (2017) where full gradients are used, our algorithm operates in a stochastic incremental framework. We first construct an abstract framework for a proximal–projection scheme that is based on a combination of proximal and projection iterations – since projection is a special case of proximal operators, this is
a very natural combination. However, projection operators lack properties generally assumed for the convergence of stochastic proximal schemes and the combination of the two operators has not been sufficiently discussed in the existing literature. We propose a novel scheme for combining proximal and projection operators, which proves its stability in a number of experiments that we present. We subsequently specialize this general scheme by computing the proximal operator corresponding to our OT formulation in closed form. We derive sparse variants of our algorithms which exploit the fact that each term in the optimization only depends on very few coordinates. Together with using efficient projections to the simplex, this leads to an algorithm with computationally low-cost iterations. For gradient-based methods, there is recent evidence of convergence difficulties and instability for an inadequate choice of parameters Patrascu and Necoara (2018), which are avoided or ameliorated with our proximal scheme. We also use the optimal coupling to compute a transport map which can be used to map out-of-sample points as suggested in Perrot et al. (2016). The balance of this paper describes the main contributions which are developed as follows:

• We propose in section 2 a new regularized formulation of OT that promotes the sparsity of the transport plan, thereby ensuring a preservation of a class structure typically arising in domain adaptation problems.

• In section 2.1 we develop a new proof for the uniqueness of the solution optimum of our convex formulation in spite of its non-strong convexity. We believe that this proposed technique may have wider applicability.

• We develop in section 3 a general accelerated stochastic incremental proximal-projection optimization scheme that combines proximal and projection iterations. We derive a novel acceleration scheme for variance reduction in the stochastic projection steps. We specialize the general scheme with an explicit closed form of proximal operators and fast projections to simplices to yield a highly scalable stochastic incremental algorithm for computing our OT formulation, thus producing an optimal coupling explicitly.

• We derive the first rigorous results for an OT plan that respects class structure, section 4.

• Finally in section 5, we investigate the algorithm on several synthetic and benchmark data sets, and demonstrate the benefits of the new regularizer.

2. Optimal Transport with SON regularization

Consider two finite sets \( \{y_i^s\}_{i=1}^m, \{y_j^t\}_{j=1}^n \) of points, respectively sampled from the so-called source and target domains. Let \( \mathbf{D} = (D_{ij} = d(y_i^s, y_j^t)) \) be the \( m \times n \) distance matrix with \( D_{ij} \) representing a distance between the \( i \)th point in the source domain and the \( j \)th point in the target domain, being used as the transportation cost of a unit mass between them. 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 \( \mu_i, \nu_j \) be respectively assigned to the data points \( y_i^s \) and \( y_j^t \). In this discrete setup, the Monge problem amounts to finding a one-to-one assignment between the points in the two domains (assuming that \( m = n \)) with a minimal cost, that
transforms the source distribution \( \{ \mu_i \} \) to the target distribution \( \{ \nu_j \} \) (if feasible). This is generally considered to be a difficult and highly ill-posed problem to solve and hence its linear programming (LP) relaxation, known as the Kantarovich problem is more widely considered, which can be written as

\[
\min_{X \in B(\mu, \nu)} \langle D, X \rangle.
\] (1)

Here, the variable matrix \( X = (x_{i,j}) \) is known as the transport map and \( B(\mu, \nu) = \{ X \in R^{n_s \times n_t}, X1_{n_s} = \mu, X^T1_{n_t} = \nu \} \) is the set of all coupling distributions between \( \mu \) and \( \nu \), respectively denoting the vectors of elements \( \mu_i, \nu_j \). Moreover, \( \langle D, X \rangle = \text{Tr}(D^T X) = \sum_{i,j} X_{ij} D_{ij} \) is the Euclidean inner product of two matrices. In an ideal case, one hopes that the optimal solution for \( X \) become an assignment (permutation matrix) in which case it is seen to coincide with the solution of the Monge problem. On account of numerical difficulties and statistical instability, the Kantarovich problem is widely used by applying further regularization. In this respect, we introduce the following flexible convex optimization framework for optimal transport via the so-called SON regularizer:

\[
X^* = \arg \min_{X \in B(\mu, \nu)} \langle D, X \rangle + \lambda \left( \sum_{i,k} R_{i,k} \| x_i - x_k \|_2^2 + \sum_{i,k} S_{i,k} \| x_i^l - x_k^k \|_2^2 \right),
\] (2)

where \( x_i \) and \( x_k \) denote the (transpose of the) \( l \)th row and \( k \)th column of \( X \), respectively, and \( S_{i,k}, R_{i,k} \) are positive kernel coefficients on rows and columns of \( X \), respectively. \( \lambda \) is a tuning parameter. Comparing (1) to (2), we observe that the second line of (2) serves as the proposed regularizer.

The effect of the regularization in (2) is to enforce many columns and rows to be respectively identical to each other in order to achieve a large number of zero terms (sparsity) in the regularizer. Hence, the resulting map \( X^* \) after a suitable permutation of rows and columns is a block matrix with constant values in each block. Thus if the data in the source and target domains has a clear partitioning structure as in the well known stochastic block model, then the recovered blocks will reflect such a structure. Note that the Laplacian regularizer in Courty et al. (2017) acts only indirectly on the transported points and is quadratic, whereas ours acts directly on the transport plan and is of \( \ell_1 \) type. Compared to the entropy regularization, which accounts the mean energy of the distribution (per Boltzmann derivation of entropy), the SON formulation specifies a more refined characterization of the distributions. The blurry transport maps resulting from the entropy regularization is widely attributed to the dependence on the mean energy, which is avoided in the SON formulation. We show that under suitable conditions related to the well known stochastic block model, many blocks will be zero. Each row and column will contain exactly one non-zero block, and the solution reflects an assignment consistent of the classes, rather than individual samples. This is made precise and proved in Section 4.

The framework in (2) is useful, especially when the source samples \( y_s^i \) are readily assigned to different classes and the optimal transport is additionally required to map the points within each class to identical or similar points in the target domain. In this case, we
may set $R_{l,k} = 0$ if $y^s_l$ and $y^s_k$ are in different classes, otherwise set $R_{l,k} = k_s(y^s_l, y^s_k)$ for a suitable (differentiable) kernel $k_s$. On the target side where no class information is ordinarily provided, we may set $S_{l,k} = k_t(y^t_l, y^t_k)$ for a suitable kernel $k_t$ of choice. The framework also allows one to use different penalty hyperparameters $\lambda_1$ on the rows, and $\lambda_2$ on the columns by incorporating them into $R$ and $S$ respectively.

### 2.1 Uniqueness

An elementary question concerning any optimization formulation, including the Kantarovich problem and its regularization in (2), 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) 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 can also be used in many similar problems including a 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 1.** We call a (global) optimal point $X_0$ of a convex optimization problem

$$\min_{X \in \mathcal{S}} f(X),$$

where $f(\cdot)$ is a convex function and $\mathcal{S}$ is a convex set, a resistant optimal point if for any open neighborhood $\mathcal{N}$ of $X_0$ there exists an open neighborhood $\mathcal{M}$ of 0 such that

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

This means that a sufficiently small linear perturbation term in the objective is guaranteed to lead to an arbitrarily small perturbation in the solution.

Accordingly, we have the following result:

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

**Proof** Suppose that there exists a different optimal point $X'$. Take $D_0 = \frac{X_0 - X'}{\|X_0 - X'\|}$, $r = \|X_0 - X'\|$ and $\tilde{D} = \varepsilon D_0$ for arbitrary $\varepsilon > 0$. Further, define $\mathcal{N}$ as the ball of radius $\delta = r/2$ centered at $X_0$. Note that for each $Y \in \mathcal{N}$ we have

$$f(Y) + \langle \tilde{D}, Y \rangle \geq f(X_0) + \langle \tilde{D}, Y \rangle = f(X') + \langle \tilde{D}, X' \rangle + \langle \tilde{D}, (Y - X_0) + (X_0 - X') \rangle.$$  

Now, note that $\langle \tilde{D}, (Y - X_0) + (X_0 - X') \rangle \geq -\varepsilon \delta + r\varepsilon > 0$, which establishes

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

Hence, $\mathcal{N} \cap \arg \min_{X \in \mathcal{S}} f(X) + \langle \tilde{D}, X \rangle = \emptyset$ and since $\varepsilon = \|\tilde{D}\|$ is arbitrarily small, we conclude that $X_0$ is not a resistant optimal point. This contradicts the assumption and shows that the solution is unique.
Theorem 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 3.** If $F$ is continuous and strongly convex, then the global minimal point of $F$ over a convex set $S$ is resistant.

**Proof** Denote the optimal point by $X^*$. By strong convexity, there exists a $\gamma > 0$ such that for any feasible point $X \in S$, we have $F(X) - F(X^*) \geq \frac{\gamma}{2} \|X - X^*\|^2_F$. Take $G = F + \langle \hat{D}, X \rangle$ and note that $G(X) - G(X^*) \geq \frac{\gamma}{2} \|X - X^*\|^2_F + \langle \hat{D}, X - X^* \rangle \geq \frac{\gamma}{4} \|X - X^*\|^2_F - \frac{\gamma}{4} \|\hat{D}\|^2_F$. This shows that $G > G(X^*)$ and hence does not have any global optimal point outside the closed sphere $\{X | \|X - X^*\|_F \leq \sqrt{\frac{\gamma}{\gamma}} \|\hat{D}\|_F\}$. Since $G$ 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 $\|\hat{D}\| < \frac{\gamma \epsilon}{\sqrt{\gamma}}$ leads to an optimal solution inside a ball of radius $\epsilon$ centered at $x^*$. This shows that the solution is resistant.

**Uniqueness for (2):** One special case of resistant optimal points, that will be useful in our analysis, is when there exists a neighborhood $M$ of 0 such that $\forall \hat{D} \in M$, $X^* \in \arg \min_{X \in S} F(X) + \langle \hat{D}, X \rangle$.

We call such a resistant optimal point an extremal optimal point. Later, we consider an analysis where we give conditions on $D$ to ensure that a desired solution $X^*$ 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 (2), adding the term $\langle \hat{D}, X \rangle$ modifies the cost matrix $D$ to $D + \hat{D}$. Hence, being an extremal optimal point is in this case equivalent to the solution $X^*$ being maintained by perturbing the matrix $D$ in a sufficiently small open neighborhood. This is easy to achieve in our planted model analysis, because the optimality of $X^*$ is guaranteed by a set of inequalities on $D$, which remain valid under small perturbations, simply by requiring the inequalities to be strict. As seen, Theorem 2 and extremal optimality, in particular, can be powerful tools for establishing uniqueness beyond strong convexity.

3. Stochastic Incremental Algorithms

3.1 Accelerated Proximal-Projection Scheme

An important advantage of the framework in (2) 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":

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

and noting that the objective function in (2) can be written as

$$\sum_{l \neq k} \phi_{R[l,k], \frac{1}{2(m-1)} d[l,k]}(x_l, x_k) + \sum_{l \neq k} \phi_{S[l,k], \frac{1}{2(m-1)} d[l,k]}(x_l^l, x_k^k), \quad (4)$$
Stochastic Proximal Algorithms with SON Regularization

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) are not smooth, so gradient methods do not apply and second, (2) 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: We have a constrained optimization problem and calculating the proximal operators over the feasible set is numerically intractable. However, we observe an appealing structure in the constraint that can be utilized for a more efficient stochastic implementation: Recalling the definition of an \( n\)-dimensional standard simplex

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

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

In summary, the optimization problem in (2) can be written in the following abstract form:

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

where each term \( \phi_p \) denotes a template function term in the objective and each set \( S_q \) is a weighted cylinder-simplex. Bertsekas Bertsekas (2011) and Wang and Bertsekas 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. However, these do not 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 (2). Further, we shortly show in Lemma 4 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 in next iterations. We extend this scheme by introducing similar memory vectors to constraints. Each memory vector \( h_m \) for a constraint \( S_m \) stores the last observed normal (separating) vector to \( S_m \). At each iteration either an objective term \( \phi_p \) or a constraint component \( S_q \) is considered by random selection. Accordingly, we propose the following rule for updating the solution:

\[
x_{t+1} = \begin{cases} 
\text{prox}_{\mu \phi_p} \left( x_t + \mu g_p \right), & \phi_p \text{ is selected} \\
\text{proj}_{S_q} \left( x_t + \mu h_q \right), & S_q \text{ is selected}
\end{cases}
\]

(6)

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 \( \phi_p \) or \( S_q \), either \( g_p \leftarrow g_p + a_t \) or \( h_q \leftarrow h_q + a_t \), where

\[
a_t = \rho \frac{x_t - x_{t+1}}{\mu} - \alpha \left( \sum_n g_n + \sum_m h_m \right),
\]

(7)

where \( \rho \in (0, 1) \) and \( \alpha > 0 \) are design constants. The vector \( a_t \) consists of two parts: the first part \( \rho \frac{x_t - x_{t+1}}{\mu} \) calculates a sub-gradient or a normal vector at point \( x_{t+1} \) 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 rule of \( x_t \) to the update rule of \( g_t \). Also, the design parameters \( \rho \) and \( \alpha \) 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.

3.2 Proximal Operator for the SON-Regularized Kantorovich Relaxation

We next show that we can explicitly compute the proximal operator for each term in (4):

**Lemma 4.** The proximal operator of the template function \( \phi_{\rho, \zeta, \eta} \) is given by \( \mathcal{T}_{\mu \rho}(p - \mu \zeta, q - \mu \eta) \), where

\[
\mathcal{T}_{\lambda}(a, b) = \left( \frac{a+b}{2} + \mathcal{T}_{\lambda} \left( \frac{a-b}{2} \right), \frac{a+b}{2} - \mathcal{T}_{\lambda} \left( \frac{a-b}{2} \right) \right),
\]

and

\[
\mathcal{T}_{\lambda}(c) = \begin{cases} 
\frac{\|c\| - \lambda}{\|c\|} c, & \|c\| \geq \lambda \\
0, & \text{otherwise}
\end{cases}
\]

**Proof** The proximal operator of \( \phi_{k, \zeta, \eta} \) is defined as

\[
\argmin_{(x,y) \in \mathbb{R}^m \times \mathbb{R}^m} \frac{1}{2\mu} \|x - p\|^2_2 + \frac{1}{2\mu} \|y - q\|^2_2 + \phi_{\rho, \zeta, \eta}(x, y).
\]

(9)
A change of variables by $u = (x+y) / 2$, $v = (x-y) / 2$ leads to a separable optimization over $u$ and $v$, which can be analytically solved and gives the result.

**Efficient Computation:** While the objective in (4) may appear complex as it involves $n^2$ terms, the associated algorithm is stochastic and incremental, thus only involving one term in (4) 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 Lemma 4 (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 (2), 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, which is deferred to the appendix A.

4. **Class Based Regularization: Guarantees**

We next show that our SON regularizer is able to provably compute transport plans that respect the class structure in the manner, explained in section 2. Our approach is to analyse it under a setting such as the well known stochastic block model (SBM) Holland et al. (1983); Snijders and Nowicki (1997), also known as the planted partition model Condon and Karp (2001) which has been used widely as a canonical generative model for the data with clear class structure. In this model, we already have a latent ground truth for the class structure which the algorithm is supposed to recover.

In the supervised version of the domain adaptation problem, the class structure is given explicitly in the source via the labels, but not in the target domain. In the unsupervised version, the class structure is unknown in both domains. In both cases, it is reasonable to assume that a latent (hidden) class structure exists. We show that our algorithm can discover this hidden class structure in both domains (unsupervised) or in the target (supervised) and computes a transport plan that respects the class structure in the two domains.

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

**Theorem 5.** Suppose in each of $K$ domains, that an equal number $m$ of random real vectors are drawn from each of $K$ individual Gaussian distributions, leading to a total number of $n = mK$ samples. The Gaussian distributions in the source and target domains are respectively centered at $\theta^s_\alpha, \theta^t_\alpha$ for $\alpha = 1, 2, \ldots, K$, and all have uncorrelated entries with equal variance $\omega^2$. Squared $\ell_2$ distance is used, $d(\mathbf{y}_1, \mathbf{y}_2) = \|\mathbf{y}_1 - \mathbf{y}_2\|^2_2$. With a probability
higher than $1 - \frac{1}{n^{10}}$ the solution of (2) with a suitable choice of $\lambda$ classifies the samples of each Gaussian distribution, and associate the $\alpha^{th}$ distributions of the two domains for each $\alpha \in [K]$ if

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

(10)

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

Fig. 1 clarifies in a simple example the geometric meaning of the concepts used in the above result. As seen, the left hand side of the condition in (10), requires the associated clusters to be substantially closer to each other than the other clusters. Moreover, the right hand side of (10) requires the distances to remain relatively bounded. An example of this situation is when $E, D, d \sim \sigma K \sqrt{K \log(nK)}$, i.e. all three grow proportionally with the number of samples, with a suitable proportion between them.

**Deterministic Guarantee:** Now, we present an extended deterministic result that is used to prove theorem 5. We use a setting inspired by the stochastic block model. For simplicity, we describe here a model in which the data points in the source and target domains are each partitioned into $K$ parts with equal size $m$. We respectively denote the partitions in the source and target domains by $\{\mathcal{S}_\alpha\}$, $\{\mathcal{T}_\beta\}$. The total number of points in each domain is $n = mK$ (these assumptions are relaxed in Appendix B). Further, $\mathcal{S}_\alpha$ is paired with $\Delta_\alpha$ for every $\alpha \in [K]$. We investigate that the plan obtained by solving (2) consists of blocks, recovering both the sets of clusters $\{\mathcal{S}_\alpha\}$, $\{\mathcal{T}_\beta\}$ 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 $i \in \mathcal{S}_\alpha$ and $j \in \mathcal{T}_\beta$, where $X_{\alpha,\beta}$ are constants satisfying $X_{\alpha,\beta} = 0$ for $\beta \neq \alpha$.  

![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.](image)
For simplicity, we take $S_{j,j'} = 1$ everywhere and study two cases where $R_{i,i'} = 1$ holds true either everywhere (no kernel) or for $i,i'$ belonging to the same cluster and $R_{i,i'} = 0$ otherwise (perfect kernels in the source domain). The general case is presented in Appendix B. Introducing an indicator variable $R$, the first case is referred to by $R = 0$ and the second one by $R = 1$. Note also that we assume the optimization in (2) to be feasible for our ideal solution, which requires for every $i,i' \in S_\alpha$ and $j,j' \in T_\alpha$ that $\mu_i = \mu_i' = \nu_j = \nu_j'$. In Appendix B, we treat the general infeasible cases by considering a relaxation of (2).

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

**Definition 6.** 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 $\alpha_1 \rightarrow \alpha_2 \rightarrow \ldots \rightarrow \alpha_k \rightarrow \alpha_{k+1} = \alpha_1$ with length $k > 1$ we have

$$\sum_{l=1}^{k} D_{\alpha_l \alpha_{l+1}} > \sum_{l=1}^{k} D_{\alpha_l \alpha_1} + k\delta, \quad (11)$$

Compared to the standard notion of cyclic monotonicity, we introduce a constant $\delta \geq 0$ in the right hand side of (11), 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 $D_{\alpha,\beta} = \frac{1}{m^2} \sum_{i \in S_\alpha, j \in T_\beta} D_{i,j}$.

We denote by $\Delta$ the maximum of the values $\|d_i - d_i'\|/\sqrt{\pi}$ and $\|d_j - d_j'\|/\sqrt{\pi}$ where source points $i,i'$ and target points $j,j'$ belong to the same cluster and we remind that $d_i, d_j$ respectively refer to the rows and columns of $D$. We also define $\omega_\alpha := \sum_{i \in S_\alpha} \mu_i = \sum_{j \in D_\alpha} \nu_j$ and then take $T_{\alpha,\beta} = \sum_{\gamma \in [K]} \left( \frac{\omega_\gamma}{\sqrt{\omega_\alpha^2 + \omega_\gamma^2}} + \frac{\omega_\gamma}{\sqrt{\omega_\beta^2 + \omega_\gamma^2}} \right) - \sqrt{2}$. Finally, we define

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

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

**Theorem 7.** Suppose that $(D_{\alpha,\beta})$ is $\delta$–strongly cyclical monotone. Take $\lambda$ such that $\Delta \leq \lambda \sqrt{m/K}$. Then, the solution of (2) is given by $X_{ij} = X_{\alpha,\beta}$ for $i \in S_\alpha$ and $j \in T_\beta$ satisfying one of the following two conditions:

1. We have $X_{\alpha,\beta} = \omega_{i}/m^2 \delta_{\beta,\alpha}$ if $\Delta \sqrt{K} \leq \lambda \sqrt{m} \leq \Lambda \delta$

2. Otherwise, we have $\delta \sum_{\beta \neq \pi(\alpha)} X_{\alpha,\beta} \leq \lambda (1 + R) \sqrt{m} \sum_{\alpha \neq \alpha'} \sqrt{\omega_{\alpha}^2 + \omega_{\alpha'}^2}$.

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

**Proof** Proof can be found in the appendix.
The first part of theorem 7 establishes ideal recovery under the condition that the "effective cluster diameter" $\Delta$ is relatively smaller than $\Lambda \delta$. The second part gives an upper bound on the error $\sum_{\beta \neq \alpha} X_{\alpha,\beta}$. Note that $\Delta$ is always smaller with $R = 1$ compared to $R = 0$, making the conditions less restrictive. This reflects the intuitive fact that introducing kernels simplifies the estimation process.

Proof of Theorem 5: Based on theorem 7, we present a sketch of the proof for theorem 5. Under the assumptions of theorem 5, we directly verify that $\delta = D^2 - d^2$ is a valid choice. Moreover $\Lambda = \Lambda_{\alpha,\beta} = \sqrt{2} / K (1 + R)$. Finally, we may conclude by Chernoff bound that with a probability exceeding $1 - \frac{1}{n^{10}}$ (the power 10 is arbitrary) we have $\Delta = O(\sqrt{E^2 + \omega^2 \log(nK)}))$. Replacing these expression in the first part of theorem 7 gives us the result.

5. Experiments

In this section, we experimentally investigate the various aspects of different optimal transport domain adaptation models on several synthetic and real-world datasets. 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). We illustrate and evaluate the value of several other properties of our method, including several other properties of our method, such as early stopping, class diversity and unsupervised domain adaptation.

5.1 Impact of SON-Regularizer

We first 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: $\lambda_1 = 10, \lambda_2 = 5$). The source data, target data and transported source data are respectively shown as yellow, blue and red points. Each column of sub-figures in Fig. 2 corresponds to a particular model resulting performance of respectively OT-l1l2, OT-lpl1, OT-Sinkhorn and OT-SON (our proposed 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 class of the target domain.

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 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
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-lp1, 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.

Figure 3: Illustration of different methods where the source and target domains have different number of classes. 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.

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.

![Figure 4: Path-based data.](image1)
![Figure 4: Path-based transport.](image2)
![Figure 4: Accuracy results.](image3)

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.

5.2 Experiments on path-based data

In Fig. 4, we investigate the different OT-based domain adaptation models on a commonly-used synthetic dataset, wherein the three classes have diverse shapes and forms Chang and Yeung (2008). In particular, we consider the case where one of the source classes is absent in the target domain. With the same number of classes in the source and target domains, the different models perform equally well. Fig. 4(a) shows a case where the source data (yellow points) and the target data (blue points), differ in the fact that the target data is missing the upper left Gaussian cloud of points appearing in the source data. Fig. 4 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 of the target domain. This thus points 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.

5.3 Real-world experiments

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.
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 of the target data and then compute the respective accuracy. Table 1 shows the accuracy results for different OT-based models for different values of the regularization parameter $\lambda$ (i.e., $\lambda \in \{10^{-5}, ..., 10^3\}$). We observe, i) OT-SON yields the highest accuracy scores, and ii) it is significantly more robust to variation of the regularization parameter ($\lambda$), 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: The accuracy scores of different OT-based methods.

5.4 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 recent 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.](image)

5.5 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.

![Transport Maps for Different Epochs](image)

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

5.6 Diverse classes in the source

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

Figure 8: The impact of SON regularization when the class members are diverse. The plot in 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.

5.7 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
Panahi et. al.

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-lp11, 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.

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).

6. Conclusion

We developed a regularized optimal transport algorithm which produces sparse maps which are suitable for problems with class specifications and geometric kernels. We provided theoretical guarantees for the sparsity of the resulting transform, 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.
References

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

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

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. Siam Reviews, 60(2):223–311, 2018. URL http://leon.bottou.org/papers/bottou-curtis-nocedal-2018.

Hong Chang and Dit-Yan Yeung. Robust path-based spectral clustering. Pattern Recognition, 41(1):191–203, 2008.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. Math. Comput., 87(314):2563–2609, 2018.

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

Laurent Condat. Fast projection onto the simplex and the l1 ball. Math. Program., 158(1-2):575–585, July 2016. ISSN 0025-5610.

A. Condon and R.M. Karp. Algorithms for graph partitioning on the planted partition model. Random Structures and Algorithms, 18(2):116–140, 2001.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. IEEE Trans. Pattern Anal. Mach. Intell., 39(9):1853–1865, 2017.

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

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

John C. Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l1-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, 2008. doi: 10.1145/1390156.1390191. URL http://doi.acm.org/10.1145/1390156.1390191.

A. Genevay, M. Cuturi, G. Peyre, and F. Bach. Stochastic optimization for large-scale optimal transport. In Adv. in Neural Information Processing Systems, 2016.
Toby Hocking, Jean-Philippe Vert, Francis R. Bach, and Armand Joulin. 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, 2011.

P.W. Holland, K.B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. Social Networks, 5(2):109–137, 1983.

Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. IEEE Signal Process. Mag., 34(4):43–59, 2017. doi: 10.1109/MSP.2017.2695801. URL https://doi.org/10.1109/MSP.2017.2695801.

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

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. Histoire de l’Académie Royale des Sciences de Paris, 1781.

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

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

Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. In Advances in Neural Information Processing Systems, pages 4197–4205, 2016.

G. Peyré and M. Cuturi. Computational Optimal Transport. Athena Scientific Belmont, 2018.

Filippo Santambrogio. Optimal transport for applied mathematicians. Birkäuser, NY, 2015.

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

T.A.B. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. Journal of Classification, 14(1):75–100, 1997.

Justin Solomon. Optimal transport on discrete domains. CoRR, abs/1801.07745, 2018. URL http://arxiv.org/abs/1801.07745.

Cédric Villani. Optimal transport: old and new, volume 338. Springer Science & Business Media, 2008.

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

A – Just-in-Time (JiT) Update

Here we explain more details about the JiT procedure, explained in section 3.2. 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 (6 in paper), (7 in paper) still operates on the full set of variables as the memory vectors become non-sparse by the updating rule in (7 in paper). We resolve this issue by following the Just-it-Time approach in [22] and modifying (7) to

\[
\mathbf{a}_t = \frac{x_{i} - x_{i+1}}{\mu} - \alpha \left( \sum_n g_n + \sum_m h_m \right)_{I_t}
\]

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

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

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

B – Extension and Proof of Theorem 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 \( C_{\alpha} \) and \( D_{\beta} \) by \( n_{\alpha} \) and \( m_{\beta} \), respectively. Further, we consider a permutation \( \pi \) on \([K]\) as the target of OT. Also, we address infeasibility by consider the following optimization:

\[
\min_{\mathbf{X} \in \mathbb{R}_{\geq 0}^{n \times n}} \langle \mathbf{D}, \mathbf{X} \rangle + \lambda \left( \sum_{\alpha,\alpha'} R_{\alpha,\alpha'} \|x_{\alpha} - x_{\alpha'}\|_2 + \sum_{j,j'} S_{j,j'} \|x^j - x^{j'}\|_2 \right) + \frac{\theta}{2} \left( \|\mathbf{X}1 - \mu\|_2^2 + \|\mathbf{X}^T1 - \nu\|_2^2 \right)
\]

(13)

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

\[
\min_{\mathbf{X}_{\alpha,\beta} \geq 0} \sum_{\alpha,\beta} n_{\alpha} m_{\beta} X_{\alpha,\beta} D_{\alpha,\beta} + \lambda \left( \sum_{\alpha,\alpha'} R_{\alpha,\alpha'} \|x_{\alpha} - x_{\alpha'}\|_M + \sum_{\beta,\beta'} S_{\beta,\beta'} \|x^\beta - x^{\beta'}\|_N \right) + \frac{\theta}{2} \left( \sum_{\alpha} n_{\alpha} (a_{\alpha}^T x_{\alpha} - \mu_{\alpha})^2 + \sum_{\beta} m_{\beta} (a_{\beta}^T x^\beta - \nu_{\beta})^2 \right)
\]

(14)
Theorem 8.

Then, we have the following more general result:

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

\[ D_{\alpha,\beta} = \frac{\sum_{i' \in C_{\alpha}, j \in D_{\beta}} D_{i,j} n_{\alpha m_{\beta}}}{\sum_{i' \in C_{\alpha}, j \in D_{\beta}} n_{\alpha m_{\beta}}}, \quad \mu_{\alpha} = \frac{\sum_{i' \in C_{\alpha}} \mu_{i}}{n_{\alpha}}, \quad \nu_{\beta} = \frac{\sum_{j \in D_{\beta}} \nu_{j}}{m_{\beta}}, \]

\[ \|x\|_{O} = \sqrt{x^{T}Ox}, \quad N, M \text{ are diagonal matrices with } n_{\alpha}, m_{\beta} \text{ as diagonals, respectively,} \]

\[ x_{\alpha} = (X_{\alpha,\beta})_{\beta} \text{ and } x^{\beta} = (X_{\alpha,\beta})_{\alpha}, \text{ and } a_{M} = (m_{\alpha})_{\alpha}, \quad a_{N} = (n_{\alpha})_{\alpha}. \]

Further, define the ideal optimization:

\[
\begin{align*}
\min_{Y_{\alpha,\beta} \geq 0, \alpha, \beta} & \sum_{\alpha, \beta} Y_{\alpha,\beta} D_{\alpha,\beta} \\
\text{s.t.} & q_{\beta} : 1^{T} y^{\beta} = \sigma_{\beta}, p_{\alpha} : 1^{T} y_{\alpha} = \sigma_{\alpha}
\end{align*}
\]

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

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

Then, we have the following more general result:

**Theorem 8.**

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

\[
\sum_{l=1}^{k} \tilde{D}_{i_{l}i_{l+1}} \geq \sum_{l=1}^{k} \tilde{D}_{i_{l}i_{l}} + k \delta. \tag{16}
\]

The solution \( X_{\alpha,\beta} \) of the characteristic optimization in (14) satisfies the following condition:

\[
\delta \sum_{\beta \neq \pi(\alpha)} X_{\alpha,\beta} \leq \sum_{\alpha \neq \alpha'} \left( \frac{R_{\alpha,\alpha'}}{n_{\alpha} n_{\alpha'}} \left( \frac{n_{\alpha}^{2} \sigma_{\alpha}^{2}}{m_{\pi(\alpha)}} + \frac{n_{\alpha'}^{2} \sigma_{\alpha'}^{2}}{m_{\pi(\alpha')}} \right) \right)
\]

\[
+ \frac{\theta}{2} \left( \sum_{\alpha} \frac{\delta_{\alpha}^{2}}{n_{\alpha}} + \sum_{\alpha} \frac{\delta_{\alpha'}^{2}}{m_{\pi(\alpha')}} \right) + \frac{\Delta_{0}^{2} n}{\theta} + \Delta_{0} (\|\delta\|_{1} - \|\delta\|_{\infty})
\]

where

\[
\Delta_{0} = \max_{\alpha,\alpha'} \left| 2 \tilde{D}_{\alpha,\alpha'} - \tilde{D}_{\alpha,\alpha} - \tilde{D}_{\alpha',\alpha'} \right|, \quad \Delta_{1} = \frac{\Delta_{0} + \max_{\alpha} |\tilde{D}_{\alpha,\alpha}|}{2}
\]
2. The solution of (13) 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 $j, j' \in D_\beta$,

$$\sqrt{\sum_{j \in [m]} (D_{ij} - D_{i'j})^2} \leq 2am_\alpha \lambda_{R_{i'\alpha}}, \quad \sqrt{\sum_{j \in [n]} (D_{ij} - D_{i'j})^2} \leq 2am_\beta \lambda_{S_{j\beta}}$$

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

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

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

**Proof** Denote the optimal value of (15) and (14) by $C_0$ and $C_1$, respectively. Also, notice that since $D_{a,\alpha'}$ satisfies the strong cyclical monotonicity condition, $Y_{\alpha,\beta} = \delta_{\beta,\pi(\alpha)} \sigma_\alpha$ is the solution of (15) and there exist dual variables $p_\alpha, q_\beta$ such that

$$D_{\alpha,\beta} - p_\alpha - q_\beta \left\{ \begin{array}{ll} 0 & \beta = \pi(\alpha) \\ \geq & \beta \neq \pi(\alpha) \end{array} \right.$$  

Moreover,

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

Hence for the solution $X_{\alpha,\beta}$ of (14),

$$C_1 = F(\{X_{\alpha,\beta}\}) \geq \sum_{\alpha,\beta} n_\alpha m_\beta X_{\alpha,\beta} D_{\alpha,\beta} + \theta \left( \sum_{\alpha} n_\alpha (\bar{a}_M^T x_\alpha - \mu_\alpha)^2 + \sum_{\beta} m_\beta (\bar{a}_N^T x_\beta - \nu_\beta)^2 \right)$$

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

$$+ \sum_{\alpha} (\bar{a}_M^T x_\alpha - \mu_\alpha) p_\alpha n_\alpha + \sum_{\beta} (\bar{a}_N^T x_\beta - \nu_\beta) q_\beta m_\beta + \sum_{\alpha} (\mu_\alpha n_\alpha - \sigma_\alpha) p_\alpha + \sum_{\beta} (\nu_\beta m_\beta - \sigma^\beta q_\beta$$

$$+ \frac{\theta}{2} \left( \sum_{\alpha} n_\alpha (\bar{a}_M^T x_\alpha - \mu_\alpha)^2 + \sum_{\beta} m_\beta (\bar{a}_N^T x_\beta - \nu_\beta)^2 \right)$$

$$\geq \delta \sum_{\beta \neq \pi(\alpha)} X_{\alpha,\beta} + C_0 + \sum_{\alpha} p_\alpha \delta_\alpha + \sum_{\beta} \delta^\beta q_\beta - \frac{1}{2\theta} \left( \sum_{\alpha} p_\alpha^2 n_\alpha + \sum_{\beta} q_\beta^2 m_\beta \right),$$
where $F(.)$ denotes the objective function in (14). 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

$$C_1 \leq F(\{X_{\alpha,\beta}'\}) = C_0 + \lambda \sum_{\alpha \neq \alpha'} \left( \frac{R_{\alpha,\alpha'}}{n_\alpha n_{\alpha'}} \sqrt{\frac{n_{\alpha}^2 \sigma_{\alpha}^2}{m_{\pi(\alpha)}} + \frac{n_{\alpha'}^2 \sigma_{\alpha'}^2}{m_{\pi(\alpha')}}} + \frac{S_{\pi(\alpha),\pi(\alpha')}}{m_{\pi(\alpha)} m_{\pi(\alpha')}} \sqrt{\frac{m_{\pi(\alpha)}^2 \sigma_{\alpha}^2}{n_{\alpha}} + \frac{m_{\pi(\alpha')}^2 \sigma_{\alpha'}^2}{n_{\alpha'}}} \right)$$

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

We conclude that

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

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

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

Lemma 1 gives the result in part 1. For part 2, notice that the optimality condition of $X_{\alpha,\beta}$ yields

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

$$+ \theta n_\alpha m_{\beta}(a^T M x_\alpha - \mu_\alpha) + \theta m_{\beta} n_\alpha(a^T N x_\beta - \nu_\beta) = 0$$

where

$$z_{\alpha,\alpha'} = \frac{x_{\alpha} - x_{\alpha'}}{\|x_{\alpha} - x_{\alpha'}\|_M}, \quad z_{\beta,\beta'} = \frac{x_{\beta} - x_{\beta'}}{\|x_{\beta} - x_{\beta'}\|_N}$$

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

$$(z_{i,i'})_j = \frac{1}{2 \lambda n_\alpha R_{i,i'}} \left( -D_{ij} + D_{i'j} - \frac{\sum_{j'' \in D_{\beta}} D_{ij''}}{m_{\beta}} + \frac{\sum_{j'' \in D_{\beta}} D_{ij''}}{m_{\beta}} - 2\theta \mu_i + 2\theta \mu_{i'} \right)$$

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

$$(z_{j,j'})_i = \frac{1}{2 \lambda m_{\beta} S_{j,j'}} \left( -D_{ij} + D_{i'j} - \frac{\sum_{j'' \in C_{\alpha}} D_{ij''}}{n_\alpha} + \frac{\sum_{j'' \in C_{\alpha}} D_{ij''}}{n_\alpha} - 2\theta \nu_j + 2\theta \nu_{j'} \right)$$

$$- \frac{1}{m_{\beta} S_{j,j'}} \sum_{\beta' \neq \beta} \left( S_{j,\beta'} - S_{j',\beta'} \right) (z_{\beta,\beta'})_\alpha$$
Also for \( i \in C, i' \in C' \) and \( j \in D, j' \in D' \), where \( \alpha \neq \alpha' \) and \( \beta \neq \beta' \), take \((z_{ii'})_j = (z_{\alpha, \alpha'})_\beta, (z_{jj'})_i = (z_{\beta, \beta'})_\alpha\). Then, it’s simple to check that \( X_{ij} = X_{\alpha, \beta} \) satisfies the optimality conditions of (13) under conditions of the theorem and noticing that by the root-means-square and arithmetic mean (RMS-AM) inequality, we also have

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

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

\( \Box \)

**Lemma 9.** Suppose that the ideal optimization in (15) has a solution where \( X_{\alpha, \pi(\alpha)} > 0 \) holds for every \( \alpha \). For every \( \delta = (\delta_\alpha)_\alpha \) satisfying \( 1^T \delta = 0 \) and any choice of the optimal dual parameters \( \{p_\alpha, q_\beta\} \) we have that

\[
\sum_\alpha p_\alpha \delta_\alpha + \sum_\beta q_\beta \delta^\beta \leq \Delta_0 (\|\delta\|_1 - \|\delta\|_\infty)
\]

where \( \delta^\beta = -\delta_{\pi^{-1}(\beta)} \). As a result in this case, (15) has optimal dual parameters \( \{p_\alpha, q_\beta\} \) satisfying

\[
|p_\alpha| \leq \Delta_1, \quad |q_\beta| \leq \Delta_1
\]

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

Denote the optimal value of (15) 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

\[
C_1 = \min_{Y_{\alpha, \beta} \geq 0} \sum_{\alpha, \beta} Y_{\alpha, \beta} D_{\alpha, \beta}
\]

s.t

\[
1^T y^\beta = \sigma^\beta + \delta^\beta, \quad 1^T y_\alpha = \sigma_\alpha + \delta_\alpha
\]  

(17)

We notice that \( \{p_\alpha, q_\beta\} \) are feasible dual vectors for (17). Hence, from the weak duality theorem we have

\[
C_1 \geq \sum_\alpha p_\alpha (\sigma_\alpha + \delta_\alpha) + \sum_\beta q_\beta (\sigma^\beta + \delta^\beta)
\]

\[
= C_0 + \sum_\alpha p_\alpha \delta_\alpha + \sum_\beta q_\beta \delta^\beta
\]
Now take the solution

\[
Y_{\alpha,\beta}' = Y_{\alpha,\beta},
\]

\[
\begin{cases}
-|\delta_\alpha| & \alpha \neq \alpha_0, \beta = \pi(\alpha) \\
- \sum_{\alpha \neq \alpha_0} |\delta_\alpha| & \alpha = \alpha_0, \beta = \pi(\alpha_0) \\
+(\delta^3)_+ & \alpha = \alpha_0, \beta \neq \pi(\alpha_0) \\
+(\delta_\alpha)_+ & \alpha \neq \alpha_0, \beta = \pi(\alpha_0) \\
0 & \text{Otherwise}
\end{cases}
\]

It is simple to check that \(Y_{\alpha,\beta}'\) is feasible in (17). Moreover, we have

\[
C_1 \leq \sum_{\alpha,\beta} Y_{\alpha,\beta}' D_{\alpha,\beta} = C_0 + \\
\sum_{\alpha \neq \alpha_0} (2D_{\alpha,\pi(\alpha)}(\delta_\alpha)_+ + 2D_{\alpha_0}(\delta_\alpha)_-) - \\
(D_{\alpha,\alpha} + D_{\alpha_0,\alpha_0})|\delta_\alpha|) \\
\leq C_0 + \Delta_0 \sum_{\alpha \neq \alpha_0} |\delta_\alpha|
\]

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 \((\alpha_1, \alpha_2)\) 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 \(\alpha_1, \alpha_2\) yield

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

Now, notice that from the optimality of (15) 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 \(\lambda\) 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.