

# Optimal Transport with Substructure

M S Nishanth

A Thesis Submitted to  
Indian Institute of Technology Hyderabad  
In Partial Fulfillment of the Requirements for  
The Degree of Master of Technology

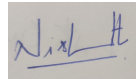


Department of Computer Science and Engineering

June 2020

## Declaration

I declare that this written submission represents my ideas in my own words, and where ideas or words of others have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources that have thus not been properly cited, or from whom proper permission has not been taken when needed.

A handwritten signature in blue ink, appearing to read 'Nishanth', with a horizontal line underneath.

M S Nishanth  
CS18MTECH11005

# Approval Sheet

This Thesis entitled Optimal Transport with Substructure by M S Nishanth is approved for the degree of Master of Technology from IIT Hyderabad



---

(Dr. C. Krishna Mohan) Examiner  
Dept. of Computer Science & Engineering  
IITH

---

(Dr. Vineeth N Balasubramanian) Examiner  
Dept. of Computer Science & Engineering  
IITH

---

(Dr. J. Saketha Nath) Adviser  
Dept. of Computer Science & Engineering  
IITH

---

(Dr. Sathya Peri) Chairman  
Dept. of Computer Science & Engineering  
IITH

## Acknowledgements

First and foremost, a sincere thanks to my guide Dr. J. Saketha Nath. You have been a continuous support for the past whole year without your guidance this would have been just another dream. There have been numerous occasions where I would have stupid silly doubts and still you handled them with your endless patience and guided me through these maze of equations. I also wanna thank you for giving me a second chance to present myself at various occasions. I wanna thank the IITH fraternity and the CSE department for delivering excellent lectures and helping us learn with a spark. I would also like to thank my friends and lab partners for their discussions. Finally, I thank my family for their everlasting support and encouragement.

## Dedication

*to love*

## Abstract

Optimal transport [OT] is a fun way of understanding the cost to reshaping/ moulding a distribution into another one. While doing a transport sometimes it is desired that the underlying substructure of the space that should be considered before doing the analysis. The structure entails the need to capture features other than the basic geometric properties such as, if the space is euclidean concepts such as Interpolation, Barycenter... which are captured. Our contribution here is to give more understanding to the solution of OT, by encapsulating the substructure of space into OT. Past work exists in this area where they had tried to use the concept of submodularity to preserve substructure/cohesiveness in the OT. This is an interesting application of the submodularity giving good results. However, The formulation of OT is that of a linear problem and by introducing the submodularity which is a nonlinear function the linear property of OT is destroyed. We try to cover the sub-structure aspect of OT using a Block Diagonal Representation. Where each block captures the substructure of the space in which the sample lies.

# Contents

Declaration . . . . .	ii
Approval Sheet . . . . .	iii
Acknowledgements . . . . .	iv
Abstract . . . . .	vi
<b>Nomenclature</b>	<b>viii</b>
<b>1 Introduction</b>	<b>2</b>
1.1 Overview . . . . .	2
1.2 Closer look . . . . .	4
<b>2 Related Work</b>	<b>5</b>
2.1 Optimal Transport . . . . .	5
2.2 Structured Optimal Transport . . . . .	6
2.3 Block Diagonal Representation . . . . .	6
<b>3 Methodology</b>	<b>7</b>
3.1 Block Diagonal Structure . . . . .	7
3.2 Enforcing Block Diagonal . . . . .	8
3.2.1 Solution as Block Diagonal . . . . .	8
3.2.2 Solution format . . . . .	8
3.2.3 Block Diagonal Regularizer . . . . .	9
3.3 $k$ -block regularizer to OT . . . . .	10
<b>4 Evaluation</b>	<b>12</b>
4.1 Dataset . . . . .	12
4.1.1 MNIST: . . . . .	12
4.1.2 USPS: . . . . .	12
4.2 Applications . . . . .	12
4.2.1 Color Transfer . . . . .	12
4.2.2 Domain Adaptation . . . . .	13
4.2.3 Analysis . . . . .	15
4.2.4 The good part: . . . . .	15
4.2.5 The bad part: . . . . .	15
4.2.6 The okay part: . . . . .	15

<b>5 Conclusion and future work</b>	<b>16</b>
<b>References</b>	<b>17</b>



# Nomenclature

- $\gamma$ : transport plan
- $C$ : optimal transport cost matrix
- Matrix is denoted by Upper case letters  $\mathbf{A}$ .
- vector is denoted by lower case letters  $\mathbf{a}$ .
- $\text{diag}(\mathbf{A})$  constructs a vector using all the diagonal elements from the matrix.
- $\text{Diag}(a)$  generates an Identity matrix where all the elements of the diagonal are the elements from the vector  $a$ .
- $\mathbf{1}$  denotes a all of all one's with appropriate dimension.
- $\mathbf{I}$  denotes the identity matrix.
- PSD matrix: If  $A$  is positive semi-definite it is represented as  $A \succeq 0$ .
- $\text{tr}(\mathbf{A})$ : represents the trace function of a square matrix.
- $A^T$ : transpose of the matrix  $A$ .

# Chapter 1

## Introduction

### 1.1 Overview

Optimal Transport [OT] [1] was first formulated by French mathematician Gaspard Monge where he describes it with an example: A worker has to move a huge pile of sand (Given a shovel and bucket). The job assigned to the worker is to move all the sand to target position and mould the new pile with a prescribed shape (example, sand castle). The experienced worker would do smart work to minimize the total effort put it in to complete the job. One quantity to minimize is the total distance or time spent carrying buckets of sand, And this inspired problem was formulated by Monge in mathematics as an Optimal Transport problem, with the ideology of comparing two probability distributions (here, two different piles of sand of the same volume.) The problem is still to find best transport among many of the possible ways to morph/ mould/ reshape/ transport the first pile into the second. This reconstruction would incur a cost for the worker, which is termed as the optimal transport cost.

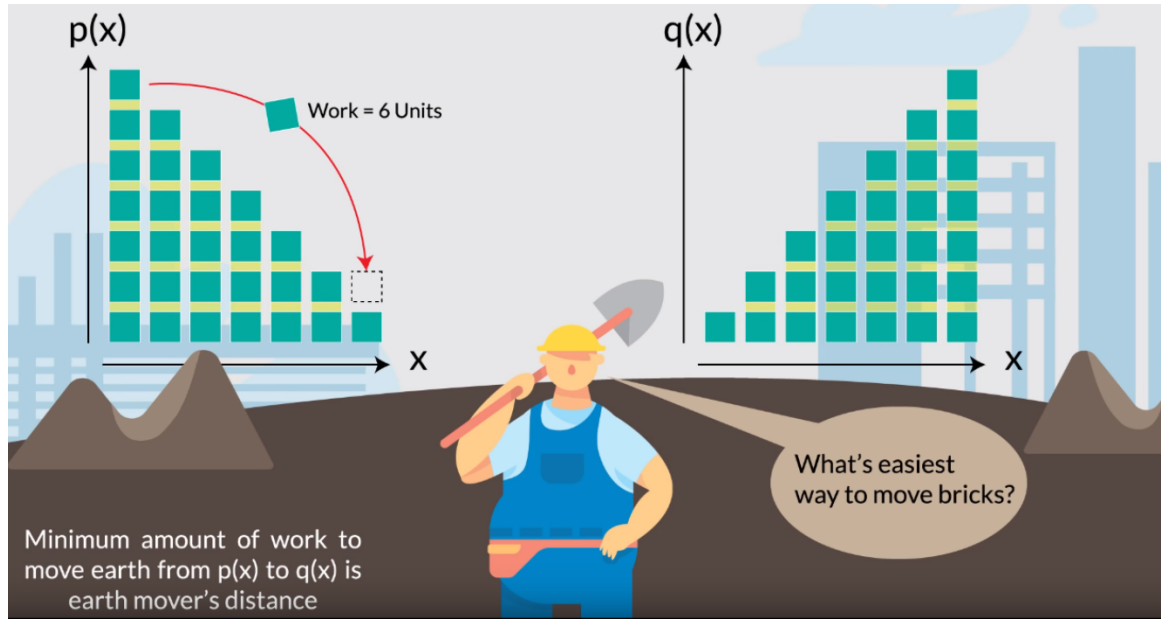


Figure 1.1: Representative image for optimal transport.

The OT metric is a metric of cost for transporting items from one space to another, which is also able to capture the the setting of each dimension such as the euclidean in the OT. This all seems so formal and mathematical yet we use this as a part of our real life in various situation without realizing it. To list a few examples, we often try to travel by the shortest path possible, we try to minimize the work done to move objects, The pizza service needs best plan to deliver multiple orders spread out in a region in a short time span with limited delivery people, using google translate to understand other languages, artist trying to find the best palette for his work.... each case above can be formalized as an OT problem, This realization would coming shocking as to how far we have been using OT in our lives.

Thou the formulation of OT itself is 200 years old, the ability of it to transform itself to wide range of problems is comprehended now. The Data Scientists and Machine Learning researchers were able to appreciate OT more, knowing that OT can provide an arsenal of weapons to study the probability distributions which would later on help in machine learning problems. It can be applied to various problem due to its simple formulation and the ability mode to fit in both discrete and continuous space. It basically generates a framework of cost for transporting from one space to another. Huge advantage of this being it is a linear minimization problem and convexity also, which cuts down the computation cost incurred by a large margin and decreases the variables in motion. It has also has found application in various models where it is used along side Deep Neural Networks. Some of its applications include: domain adaptation, sentence similarities, deep learning, shape analysis, color transfer, style transfer, image registration and interpolation, adversarial neural networks.

In plain OT the sub-structure of samples from the space, which is to be transported is not addressed. The structure can be anything from a intrinsic nature like distribution of colors or objects to extrinsic structure such as groupings based on label. Like as in Domain Adaptation

where certain structures of elements are captured to determine positive or negative, like in the case of Cancer. A nominal work came up with an response to this issue by using sub-modular function. The sub-modular function defined as a simple analogy to the structure of the space. This use of sub-modular function redefines the cost by assigning cost with partiality for different objects. In this framework of structural cost they analyze them self the cost function to be of non linear class contradicting to the linear property of OT. is a function. The final comments on the author by analysing the transport cost function is that the samples of a dimension from a similar pattern has more affinity to the similar sample pattern in the other dimension and hence causing the overall OT matrix to become a block diagonal. The block diagonal property is in alignment with the common understanding that the samples should follow the trend of similar pattern taking low cost and transport across different pattern should be have higher cost and hence always the transport to be followed between similar pattern, and proving the significance of the block diagonal in the OT problem.

The block diagonal significance is covered in response to other problems such as clustering, where it is formally proved that the block diagonal is a significant feature which represents the affinity to items of same/similar features. We embark on the search for a better substructure analyzing function to make the cost between two dimensions compatible with the ability to capture the linearity in the existing space. The base motive here is to find the best substructure capturing function that can be utilized in OT. We try to embed the idea developed in this interesting work *Subspace Clustering by Block Diagonal Representation* [2] with ability to group data points which lie in approximately linear sub-spaces into clusters with each cluster corresponding to as subspace and achieving this in a block diagonal fashion. The use a block diagonal representation which generates the block diagonal similarity. Here they follow an assumption that the sampled data have the approximately linear subspace structure. Which is a valid assumption in our case also, since there the samples are all part of some pattern and they have a common intrinsic property in the fashion of a linear function.

Contributions: In short, we make the following contributions: (1) we propose a new framework for encoding sub-structural information into optimal transport problem; (2) we try to integrate a BDR [2] method to exploit the geometric structure. Subspace Clustering by Block Diagonal Representation

## 1.2 Closer look

The OT is a linear minimization problem and as there are a plethora of applications like: domain adaptation, sentence similarities, deep learning, shape analysis, color transfer, style transfer, image registration and interpolation, adversarial neural networks. The data set changes widely from corpus data, color, images to sound clips. But OT works with distribution as input. In general the data set is visualized as a distribution of the samples so as to allow the variety in the data set and all individual features of the data set can be represented in a format and where in all samples/items which are similar follow a distribution/ pattern. The distribution can be uniform or can be defined later according to the data set/ experiment. Note- all samples are assumed to follow a sub-structure property, as our goal is transfer the sub-structure property along with the plain OT solution, if there is no sub-structure in the sample that need to be addressed in the OT then plain OT can be straight applied.

# Chapter 2

## Related Work

The OT is a mathematical formulation of a transport metric cost between two space, to which we add the encapsulation module. This module is to capture the sub-structure property of the samples/items in the space. Since this work covers two broad aspects of research areas, we cover them in different sections.

### 2.1 Optimal Transport

The formulation of optimal transport [?] was given by Gaspar Monge: Let  $\mu, \nu$  be two probability measures over the metric spaces  $\mathcal{X}, \mathcal{Y}$ , and a measurable cost function  $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ , which represents the cost function for transporting a unit of mass from  $x \in \mathcal{X}$  to  $y \in \mathcal{Y}$ . The Monge's formulation of the optimal transportation problem is to find a transport map  $T : \mathcal{X} \rightarrow \mathcal{Y}$  that realizes the infimum

$$\inf_T \left\{ \int_{\mathcal{X}} c(x, T(x)) d\mu(x) \mid T_{\#}\mu = \nu \right\} \quad (2.1)$$

where  $T_{\#}$  denotes the push-forward of  $\mu$  by  $T$ . The solution to this might not exist. However, a convex relaxation of the Monge's problem which was proposed by Kantorovich is guaranteed to have a solution:

$$\inf_{\gamma} \left\{ \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\gamma(x, y) \mid \gamma \in \Gamma(\mu, \nu) \right\} \quad (2.2)$$

where  $\Gamma(\mu, \nu)$  is a set of transport plans, i.e., joint distributions with marginals  $\mu$  and  $\nu$ . The map  $T$  that gives this infimum is called an *optimal transport map*.

In the case of a discrete setting this cost needs to be specified for every pair of  $(x_i^s, x_j^t)$  which visualises to a matrix with each entry representing all possible pairs across  $\mu$  and  $\nu$ . The cost matrix is thus,  $C \in \mathbb{R}^{n \times m}$ , and the total transportation cost incurred by  $\gamma$  (the transport plan cost) is  $\sum_{i,j} \gamma_{i,j} \cdot c_{i,j}$ . So, in the case of discrete optimal transport (DOT) problem involves finding a transport plan that solves

$$\min_{\gamma \in \mathcal{M}_{\mu, \nu}} \langle \gamma, C \rangle \quad (2.3)$$

To this OT problem in order to capture the sub-structure property of the space, submodular functions are used. These are a class of non-linear functions that emphasizes on the same class/structure affinity as opposed to classes/structures across for the sample.

## 2.2 Structured Optimal Transport

Structured optimal transport [3] is a strategy to capture the information about the sub-space of the samples belonging to the space. The submodularity function is used. This function is able to generate biased cost so as to enable transport while preserving a structure. The methodology here is to use edmond's function to get the gradient of the cost function while implying a biased cost to the samples to ensure the sub-space property. After this a projection using sinkhorn [4] is done to map the solution back to the solution space. This is how structured optimal transport achieves the best transport plan.

## 2.3 Block Diagonal Representation

Sub-space clustering is vastly studied due to its numerous applications in Machine learning fields. Existing clustering can be categorized into 4 methods broadly. (1) mixture of Gaussian, From a mixture of Gaussian distributions in dependant samples are drawn. Now the subspace clustering is converted to the model estimation problem and this can be performed by Expectation Maximization (EM [5]) algorithm, but they are also sensitive to errors and the initialization. (2) matrix factorization based methods, used to reveal the data segmentation based on the factorization of the given data matrix which are sensitive to noise and outliers. (3) Algebraic, Generalized Principal Component Analysis (GPCA[6]) fits the data points with a polynomial. However, this is generally difficult due to the data noise and its cost is high especially for high-dimensional data. (4) Due to the simplicity and outstanding performance, the spectral-type methods attract more attention in recent years. We give a more detailed review of this type of methods as follows. They first learn an affinity matrix to find the low-dimensional embedding of data and then a proximity based clustering algorithm is applied to achieve the final clustering result. The main difference among different spectral-type methods is the different ways the affinity matrix is constructed. The entries of the affinity matrix (or graph) measure the similarities of the data point pairs. Ideally, if the affinity matrix is block diagonal, i.e., the cross-cluster affinities are all zeros, one may achieve perfect data clustering since the experiment produces zero affinity for one cluster to be affiliated with another cluster. Typical kernel or locality based methods may not be a good choice for subspace clustering since the data points in a union of subspaces may be distributed arbitrarily but not necessarily locally. Hence, affinity matrix construction methods for subspace clustering by using global information have been proposed in recent years, each of which uses a variation in the used regularization for learning the representation coefficient matrix.

As in sub modular optimal transport, which uses submodularity to capture this substructure information. we replace this with block diagonal regularizer which captures the affinity and empowers the sub-structure of the sample in each space.

# Chapter 3

## Methodology

This chapter describes the integration of Block diagonal representation to the OT problem.

### 3.1 Block Diagonal Structure

Given any dataset  $X \in \mathbb{R}^{D \times n}$ , it will have  $k$  subspaces (each subspace following approximately similar substructure, here each subspace is column-matrix). The union of all these subspaces will give back the Data matrix  $X$ . Each subspace  $S_i$  has  $n_i$  samples with  $\sum_{i=1}^k n_i = n$ . Let  $X_i \in \mathbb{R}^{D \times n_i}$ , denote the sub-column-matrix in  $X$  that belongs to  $S_i$ . Let  $X = [X_1, X_2, \dots, X_k]$  be ordered according to their substructure membership. Now using the advantage of the subspace structure, the sampled data points obey the so called self expressiveness property, i.e., each data point in a union of subspaces can be well represented by a linear combination of other samples in the dataset. This is formulated as

$$X = XZ \quad (3.1)$$

Let  $Z^*$  be a feasible solution in to the above equation==3.1eq, where  $Z^* \in \mathbb{R}^{n \times n}$  is the representation coefficient matrix. So far, the  $Z^*$  is a set of feasible solution and there is no constraint to be a unique solution and to overdrive with the OT structure-biased cost reduction we need such a  $Z^*$  where in each sample is represented as a linear combination of samples belonging to the same subspace, i.e.,  $X_i = X_i Z_i^*$ , where  $Z_i^*$  is not the trivial solution, i.e., the Identity matrix. The structure of  $Z^*$  is said follow block *k-block diagonal* structure.

$$Z = \begin{bmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_k \end{bmatrix}, \quad Z_i \in \mathbb{R}^{n_i \times n_i}. \quad (3.2)$$

By structure the  $Z$  reveals the membership of data  $X$ . We can further derive the clustering/classification by defining the affinity matrix as  $(|Z| + |Z|)/2$ .

Issues to handle (1) May have many solutions, need a regularizer to find the block diagonal  $Z$ . (2) is sparse, but using the basic instinct to use  $\ell_0$  norm points may/ may not work, Need to study the

properties of block diagonal and various cases to achieve this.

## 3.2 Enforcing Block Diagonal

### 3.2.1 Solution as Block Diagonal

To understand the need of block diagonal affinity, Consider the previous dataset assumption where samples are drawn from a collection consisting of  $k$  independent subspaces  $\{\mathcal{S}_i\}_{i=1}^k$  of dimensions  $\{d_i\}_{i=1}^k$ . Let  $X = [X_1, \dots, X_k] \in \mathbb{R}^{D \times n}$ , where  $X_i \in \mathbb{R}^{D \times n_i}$  denotes the data point drawn from  $\mathcal{S}_i$ ,  $\text{rank}(X_i) = d_i$  and  $\sum_{i=1}^k n_i = n$ . For any feasible solution  $Z^* \in \mathbb{R}^{n \times n}$  to the following system

$$X = XZ \quad (3.3)$$

is decomposed into two parts  $Z^B$  and  $Z^C$ ,  $Z = Z^B + Z^C$  such that,

$$Z^B = \begin{bmatrix} Z_1^* & 0 & \cdots & 0 \\ 0 & Z_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_k^* \end{bmatrix}, \quad Z^C = \begin{bmatrix} 0 & * & \cdots & * \\ * & 0 & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & 0 \end{bmatrix}$$

This specific assumption is to show that the solution can indeed be a block diagonal by proving  $XZ^C = \text{Zero}$  and  $Z^B$  a diagonal matrix is enough to represent the affinity of the subspaces. From above  $Z_i^* \in \mathbb{R}^{n_i \times n_i}$  corresponds to  $X_i$ . This can be proved easily with the assumption that the subspaces are independent and the fact that the samples form one subspace don't have unique to that subspace. This characterizes the underlying representation contributions of all data points. However, such contributions are not explicitly reflected by the representation matrix  $Z$  since the decomposition  $Z^* = Z^B + Z^C$  is unknown when  $Z^C \neq 0$ . In such cases, the solution does not necessarily obey the block diagonal property, and thus it does not align with the idea of a sample being a linear combination of other data point. To handle cases like we introduce a regularizer to the feasible solution such that  $Z^C = 0$ .  $Z^* = Z^B$  obeys the block diagonal property.

A wide variety of subspace clustering methods exist and revolve around the generalized equation [7],

$$\begin{aligned} \min_Z ||Z||_*, \\ \text{st. } X = XZ \end{aligned} \quad (3.4)$$

### 3.2.2 Solution format

For a given problem the dataset will not be the ideal one as in where we assumed a perfectly ordered dataset like,  $X = [X_1, X_2, \dots, X_k]$ . So as to satisfy the block diagonal property in various setting the following general cases are considered.

The k-block diagonal matrix and block diagonal property are not the same. The block diagonal property is widely used in clustering algorithms but k-block diagonal matrix is not. A matrix obeying the block diagonal property is k-block.diagonal, but not vice versa.



Case (i): The matrix does not follow ideal formatting, like where all subspaces are arranged in order. Since the data is collected in random the initial assumption of  $\mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_k]$  might not be valid. We need to make sure so that the block diagonal property is satisfied to invariant to the permutation of the columns of the data input matrix  $\mathbf{X}$ .

Case (ii): We need a condition to ascertain the solution matrix to be block diagonal under certain subspace assumption. From 3.5, we have  $\mathbf{X} = \mathbf{X}\mathbf{Z} = \mathbf{X}\mathbf{Z}^B$ . This guarantees that  $\mathbf{Z} = \mathbf{Z}^B$  when minimizing the objective.

Case (iii): To see the connection between the structure of each block of the block diagonal solutions and the used objective  $f$  we need a disassociate property from the solution. This is not a necessary property but through we through the lens of this condition, we can verify the linear combination of samples forming the dataset.

To follow up with the above cases the following conditions are to be imposed on the solution affinity matrix.

$$(1) f(Z, X) = f(P^T Z P, X P) \text{ for any permutation } P^T Z P \in \Omega$$

$$(2) f(Z, X) \geq f(Z^B, X), \text{ where the equality holds if } Z = Z^B \text{ (or } Z_3 = Z_4 = 0)$$

$$(3) f(Z^B, X) = f(Z_1, X_1) + f(Z_2, X_2)$$

These conditions are termed as enforcing block diagonal conditions (EBD). One best way to go for enforcing conditions use these conditions in the part of regularizer. Norms [8] have been best choice of regularizers and checking for a match between the EBD conditions and norms. The  $\ell_0$  and  $\ell_1$  norms follow only the first EBD condition. Interestingly the nuclear norm satisfies the conditions 1 and 3. Now to fit in condition 2 we use the formulation for case 1 and case 3 and the condition 2 can be proved.

### 3.2.3 Block Diagonal Regularizer

In the definition of block diagonal we say it is block diagonal by a factor  $k$  or simply as  $k$ -block diagonal. This  $K$  for our purposes defines the number of sub-spaces observed in the dataset. These  $k$  subspace in the dataset should produce a  $k$ -block diagonal, i.e. number of block observed should be  $k$ . But this  $k$  at this point is more ambiguous. Consider a matrix

$$A = \begin{bmatrix} A_0 & 0 & 0 \\ 0 & A_1 & 0 \\ 0 & 0 & A_2 \end{bmatrix}$$

Where,  $A_0 = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$  here  $A_0$ ,  $A_1$  and  $A_2$  could all be a matrix by them self of any order, with the only constraint that the dimensions are in consistent to the expected dimensions of  $\mathbf{Z}$ . We are building the  $k$ -block on the  $\mathbf{Z}$  matrix (cost coefficient matrix). In the above example we can say that it is 3-block diagonal but by construction and definition we can say that it is 1-block or 2-block diagonal too, by grouping the  $A_0$  and  $A_1$ .. and so on... to remove this ambiguity a more formal method of defining connectedness can be used such as the Laplacian which in most graphical

applications are used to determine the number of connected components. Laplacian of a matrix [9] is denoted by

$$L_B = \text{Diag}(\mathbf{B}\mathbf{1}) - \mathbf{B} \quad (3.5)$$

The regularizer can be defined more structurally now. For any affinity matrix  $\mathbf{B} \in \mathbb{R}^{n \times n}$ , the *k-block diagonal regularizer* is defined as the sum of the  $k$  smallest eigenvalues of  $L_B$ , i.e.,

$$\|B_K\| = \sum_{i=n-k+1}^n \lambda_i(L_B). \quad (3.6)$$

It can be seen that  $\|B\|_K = 0$  is equivalent to the fact that the affinity matrix  $\mathbf{B}$  is  $k$ -block diagonal. So  $\|B\|_K$  can be used as the block diagonal matrix structure induced regularizer.

To embed the EBD conditions on to the regularizer we use the nuclear norm of the Laplacian and mark it as the regularizer.

$$\begin{aligned} \|L_B\|_* &= \text{Tr}(L_B) = \text{Tr}(\text{Diag}(\mathbf{B}\mathbf{1}) - \mathbf{B}) \\ &= \|\mathbf{B}\|_1 - \|\text{diag}(\mathbf{B})\|_1 \end{aligned} \quad (3.7)$$

where the following facts are used:

$$B = B^T, \quad B \geq 0 \quad \text{and} \quad L_B \succeq 0. \quad (3.8)$$

So, the model is equivalent to

$$\begin{aligned} \min_Z \|L_B\|_* \\ \text{s.t. } X = XZ, \quad \text{diag}(Z) = 0, \quad B = (|Z| + |Z|^T)/2. \end{aligned} \quad (3.9)$$

### 3.3 *k-block* regularizer to OT

To the original OT problem the regularizer is added to get the equation,

$$\min_{\gamma \in \mathcal{M}_{\mu, \nu}} \langle \gamma, C \rangle + \gamma \|B\|_k \quad (3.10)$$

The existing result by Ky Fan can be used to reformulate  $\|B\|_k$  (our regularizer). Ky Fan states that for  $\mathbf{L} \in \mathbb{R}^{n \times n}$  and  $\mathbf{L} \geq 0$  then,

$$\begin{aligned} \sum_{i=n-k+1}^n \lambda_i(\mathbf{L}) &= \min_w \langle \mathbf{L}, \mathbf{W} \rangle, \\ \text{s.t. } 0 &\preceq \mathbf{W} \preceq \mathbf{I}, \text{Tr}(\mathbf{W}) = k. \end{aligned} \quad (3.11)$$

Using Ky Fan's result, we can rewrite the block diagonal regularizer as a convex program

$$\begin{aligned} \|\mathbf{B}\|_k &= \sum_{i=n-k+1}^n \lambda_i \mathbf{L}_B = \min_w \langle \mathbf{L}, \mathbf{W} \rangle, \\ \text{s.t. } 0 &\preceq \mathbf{W} \preceq \mathbf{I}, \text{Tr}(\mathbf{W}) = k. \end{aligned} \quad (3.12)$$

Now, the overall OT is:

$$\min_{\gamma \in \mathbf{M}_{\mu, \nu}} tr\langle \gamma, \mathbf{C} \rangle + \lambda tr(\mathbf{W}^T (Diag(\mathbf{B}1) - \mathbf{B})) \quad (3.13)$$

note, in our case  $\mathbf{W} = \mathbf{B} = (\gamma + \gamma^T)/2$

On substituting for  $\mathbf{B}$ . we get,

$$\min_{\gamma \in \mathbf{M}_{\mu, \nu}} tr\langle \gamma, \mathbf{C} \rangle + \lambda tr(\mathbf{W}^T (Diag((\gamma + \gamma^T).1)/2 - \mathbf{B})) \quad (3.14)$$

We can solve the above equation by alternate optimization. In this we consider one block to be constant while the other is updated. The two blocks are:

$$\min_{\gamma \in \mathbf{M}_{\mu, \nu}} tr(\mathbf{C} - \lambda(\mathbf{W} + \mathbf{W}^T - diag(\mathbf{W})1^T - 1diag(\mathbf{W}))^T \gamma_{st} : 0 \preceq \mathbf{W} \preceq \mathbf{I}, tr(\mathbf{W}) = k. \quad (3.15)$$

$$\min_{\gamma \in \mathbf{M}_{\mu, \nu}} M_{\mu, \nu} \lambda tr(\mathbf{W}^T (Diag((\gamma + \gamma^T).1)/2 - \mathbf{B})) \quad (3.16)$$

Here both equations are convex and hence they have closed form solutions.

The closed form solution to the above equation would be:

$$\mathbf{W}^{k+1} = \mathbf{U}\mathbf{U}^T \quad (3.17)$$

Where  $\mathbf{U} \in \mathbb{R}^{n \times k}$  is build up of  $K$  eigen vectors corresponding to the  $k$  smallest eigen vector of the Laplacian of B; which is  $k$  *smallest* eigen vectors of  $Diag(\mathbf{B}1) - \mathbf{B}$ .

$$\mathbf{C} = \mathbf{C} - \lambda(\mathbf{W} + \mathbf{W}^T - diag(\mathbf{W})1^T - 1diag(\mathbf{W}))^T \quad (3.18)$$

$\mathbf{W}^{k+1} = \mathbf{U}\mathbf{U}^T$  where  $\mathbf{U} \in \mathbb{R}^{n \times k}$  consist of  $k$  *eigenvectors* associated with the  $k$  *smallest eigenvalues* of Laplacian of B. Which is the laplacian of  $Diag(\mathbf{B}1) - \mathbf{B}$ .

# Chapter 4

## Evaluation

Our implementation of the algorithm uses Python Optimal Transport library [10] Experiments were run on server without using GPU. (OT library can be optimised to use GPU). Note, that our equation on  $B$  is an affinity matrix that stores cross affinity information. This being affinity matrix should always be symmetric, square and positive semi-definite. For this another important point that should hold true is that the number of source and target subspaces should be equal.

Parameters: there are two parameters that can be fine tuned according to the experiment. One is the  $k$  value which denotes the number of subspaces. This can be viewed in the block diagonal. The other is the  $\lambda$  parameter which denotes the level of regularization effect caused by the block diagonal. this can be customised according to the dataset size and subspaces.

### 4.1 Dataset

#### 4.1.1 MNIST:

A database of handwritten digits, has a 10 classes with total sample size of 60,000 images and a test set of 10,000 examples. The digits have been size-normalized and centered in a fixed-size image. Here we use fixed number of random samples from each class of training set. The images are resized and down-scaled to 16x16 pixels so as to be compatible with USPS dataset.

#### 4.1.2 USPS:

A set of handwritten digits of varying orientation and scaling with 7291 train and 2007 test images. The images are in 16x16 grayscale pixels. A random number of samples from each class of training set is taken arbitrarily.

### 4.2 Applications

#### 4.2.1 Color Transfer

This is an very interesting use case of Optimal transport. Here the goal is to use the color from source image as the palette to color the target image. The subspace for this experiment is the

color and their corresponding occurrences. We build the base cost for this optimal transport by calculating the difference in RGB space between the images. Note there is a strict condition for the source and target subspaces to be equal to make this possible we add proxy colors. To denote that this transport is very costly and should be avoided in the Optimal transport we define a fixed high cost for transporting the color from source to target, Since optimal transport always selects the least work/ cost for transport the proxy colors won't affect the experiment. With this as base the transport of colors is done by the OT+BDR. Using this optimal transport plan the color mapping from source to target image is defined.

We observe that the OT+BDR method is more sharper than the plain OT method.

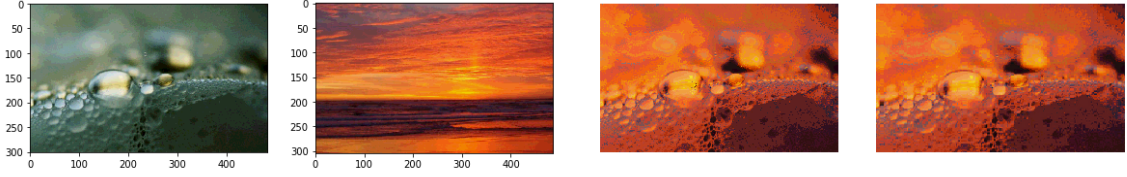


Figure 4.1: Images from left: Target image, Source image, OT+BDR solution, EMD Solution

We can observe the OT+BDR method to have sharpened color transfer results. tip: when dealing with high resolution images the pixel binning can be done to decrease the resolution to get faster results for comparison/ studying.

### 4.2.2 Domain Adaptation

Domain adaptation is about learning the source distribution very well then using that information to do analysis on the target distribution. Here we try to do a map between the MNIST and the USPS dataset. We do not use the target distribution's label information, by this we are also able to achieve unsupervised learning mechanism. The subspaces are constructed for both source and target images such that each subspace contains samples from its own class. To make sure the data is compatible among the source and target images, i.e. MNIST and USPS dataset, The MNIST is down scaled to 16x16 pixel size since it has higher resolution. We show experiment in two sets where we consider random 3 class and 10 image set in each class and second setting where we use 5 classes with each class having 50 samples in both source and target dataset.

We set the baseline cost for transport and then perform the OT+BDR transport. Using this we try to understand the nature of the block diagonal features expressed in the transport plan.

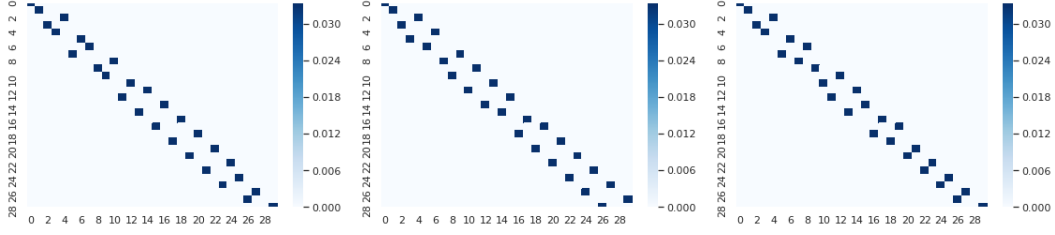


Figure 4.2: Images from left: Target image, Source image, OT+BDR solution, EMD Solution. MNIST to USPS domain adaptation with classes = 3, samples = 10 each.

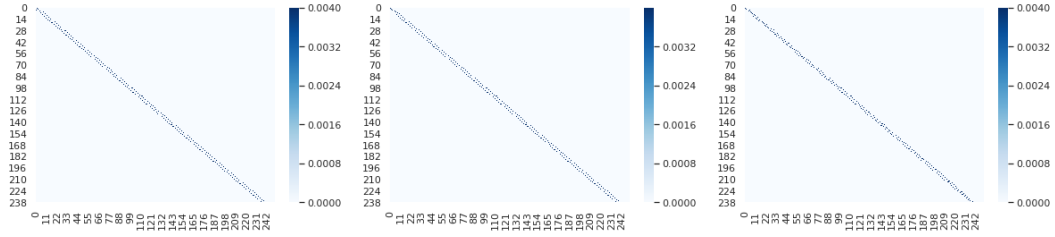
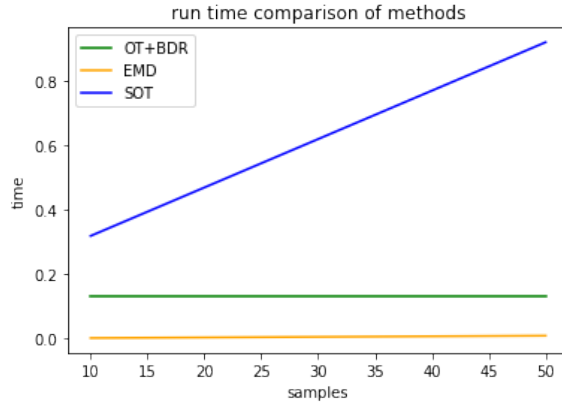


Figure 4.3: Images from left: Target image, Source image, OT+BDR solution, EMD Solution. MNIST to USPS domain adaptation with classes = 5, samples = 50 each.

We do this DA experiment under various settings to find the  $\lambda$  value which controls the level of regularization of the block diagonal. For these settings the  $\lambda$  is set to be 10, with  $k$  val respect to the number of classes in experiment. We can observe a thicker block alignment from the OT+BDR solution. The blocks seems to be spread out a more in the cases of plain OT and Structured Optimal Transport.



Comparing the run time for EMD, OT enforced EBD and SOT for this domain adaptation dataset. We observe that the SOT has the highest run time which is expected with its pixel wise search into each of the sub-spaces. EMD has the least and OT enforced BDR takes a little more time comparatively. Since EMD is plain OT without much computation it is able to achieve low run time. Note, with the subspace computation also we are almost able to match the run time of the EMD solution.

### 4.2.3 Analysis

The regularizer induced by the Block diagonal perform well in various settings, by considering the set of applications it can work in with only changes as to how the data should be fed.

### 4.2.4 The good part:

It is able to generalize and adapt to various settings. Like in the case of color transfer it is hard to define the bubble lining and the different lighting was a challenge in itself, where this method proved to work okay. In structured Optimal transport there is always the overhead cost of finding the membership of the sample to its subspace in each iteration of optimization. However, there is not computation time involved to find the source sample's membership as that information is captured in the optimization function via block diagonal. This finding membership is a time consuming process when you consider an image with  $720 \times 1080$  which makes the cost of finding 7 lac samples in that subspace.

### 4.2.5 The bad part:

It adds to computation overload since it requires the computation of eigen vectors and eigen values. Since the matrix being a symmetric matrix the cost is not as high as computing the eigen values for a full non-symmetric matrix however the run time could be reduced.

### 4.2.6 The okay part:

The visual appearance and the block diagonal visualization is in close call with other existing methods but numerically the method performs slightly better than other methods, with better convergence rate.

## Chapter 5

# Conclusion and future work

We had a goal to define a better framework for Optimal Transport which is able to capture the sub-structure property of the samples in the subspaces, and we were able to achieve this. The framework created is a generalized framework that can be used in various applications with only change required being the adjustment needed for different datasets. Integrating a technique from clustering was a good turning point. There can be other classification/ Clustering strategies that can help one capture sub-structure property in a more optimized manner, this can be a different perspective to look for where the block diagonal can be captured.

Certain aspects for extension of this work, as this being a cost metric it can be applied in almost all machine learning problems like adversarial networks and other areas of domain adaptation where the less information about the target domain is known/ there is some structure in the sample that needs to be studied more.



# References

- [1] L. Ambrosio. Lecture notes on optimal transport problems. In Mathematical aspects of evolving interfaces, 1–52. Springer, 2003.
- [2] C. Lu, J. Feng, Z. Lin, T. Mei, and S. Yan. Subspace clustering by block diagonal representation. *IEEE transactions on pattern analysis and machine intelligence* 41, (2018) 487–501.
- [3] D. Alvarez-Melis, T. Jaakkola, and S. Jegelka. Structured optimal transport. In International Conference on Artificial Intelligence and Statistics. 2018 1771–1780.
- [4] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In Advances in neural information processing systems. 2013 2292–2300.
- [5] T. K. Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine* 13, (1996) 47–60.
- [6] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (GPCA). *IEEE transactions on pattern analysis and machine intelligence* 27, (2005) 1945–1959.
- [7] J. Feng, Z. Lin, H. Xu, and S. Yan. Robust subspace segmentation with block-diagonal prior. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2014 3818–3825.
- [8] F. R. Bach. Structured sparsity-inducing norms through submodular functions. In Advances in Neural Information Processing Systems. 2010 118–126.
- [9] A. Marsden. Eigenvalues of the laplacian and their relationship to the connectedness of a graph. *University of Chicago, REU* .
- [10] R. Flamary and N. Courty. POT Python Optimal Transport library 2017.