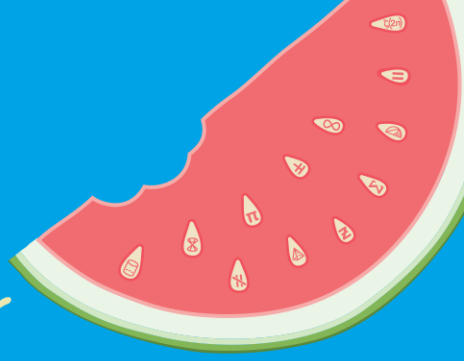
AMSI VACATIONRESEARCH SCHOLARSHIPS 2021–22

Get a taste for Research this Summer



Computational (Optimal) Transport for Domain Adaptation

Thanh Dat Tran

Supervised by Dr Tiangang Cui

Monash University

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Abstract

Because of the limited availability of labelled data and computing resources, it is often challenging to deploy machine learning algorithms to real-world scientific applications. Domain adaptation, which aims to transfer a well-trained model for a specific machine learning task to similar tasks within the same class, offers a viable route to solve scientific machine learning problems of this type. For example, how can the decision strategy for the COVID management plan of City M be adapted to City S? It is necessary to find the optimal plan without repeating the expensive experiments. By casting scientific machine learning tasks into a probabilistic framework, we want to investigate various avenues in applying (optimal) transport methods to address problems in domain adaptation.

Contents

1 Introduction

Modern machine learning tasks often requires the access to high volumes of labelled data to produce exceptional and reliable performance. But even when given a large amount of both data and computing resources, it can still be challenging to apply a well-trained model to new sets of data in the real-world applications. This can be seen as due to the differences in the distributions of the new data and the ones that were used for training such model. A simple solution is to obtain more and more observations from the new sets of data, but it can expensive and requires continuous acquisition, which can be burdensome or infeasible. Domain adaptation is a sub-field of machine learning that focuses on applying learnt information from the sets of labelled source data to some target data with unknown labels. In particular, it deals with the cases where the distributions of the source and target domains are different. Such differences (called *domain drift*) can be caused by multiple reasons with potential physical interpretations. In the case of computer vision, the drift can occur due to the changes in the angles, lightings, backgrounds, random noises or simply due to different acquisition devices. In the task of detecting epilepsy using data from Electroencephalogram (EEG) test, deploying predictive model developed with one patient's data to other patients can be obstructed because of the differences in patients' conditions.

In this work, we investigate the case of **unsupervised domain adaptation** with a single source domain associated with the outputs/labels and a single unlabelled target domain. This is to separate from semi-supervised domain adaptation with few known outputs/labels in the target domains, and multiple domain adaptation with multiple source and target domains. In more detail, we want to explore the **least effort principle** approaches [1] [2] to tackle the problem of domain adaptation with the assumption that the domain drift is in the form of some transformation from the source to the target domains, so that the transformation is minimal with respect to some cost metric. From this principle, the domain adaptation problem can be expressed as first finding a transformation making the distribution in the source domain to be similar to that in the target domain, and then making use of the learnt transformation to estimate the target outputs/labels. Our problem can then be formulated under the framework of Optimal Transportation (OT) theory, which has been well studied and applied in multiple fields due to the ability to compute distance between probability distribution with potentially non-overlapping support spaces. With two different formulations of domain adaptation problem under OT framework, we explore their properties in some examples to provide some deeper knowledge about such formulations.

Statement of Authorship

With the guidance from my supervisor, I produced analytical solutions to some simple optimal transport problems. I wrote the Python code to numerically compute the solution and visualize some examples as well as adapting the existing Python libraries to solve the problems. I presented some interpretations of the results inspired with the insights from my supervisor in the form of the report with included figures.

2 Domain Adaptation

2.1 Notation

Let $\Omega \subset \mathbb{R}^d$ be the input sample space and $\mathcal{C}$ be the set of all possible outputs/labels. Then we define $\mathcal{P}(\Omega)$ as the the set of all probability measures over the input sample space Ω.

In the normal setting of machine learning, we can assume the existences of a set of training data $\mathbf{X}_s = \{\mathbf{x}_i^s\}_{i=1}^{N_s}$, with $\mathbf{x}_i^s \in \Omega$, associated with a set of outputs/labels $\mathbf{Y}_s = \{y_i^s\}_{i=1}^{N_s}$, with $y_i^s \in \mathcal{C}$, and a set of test data $\mathbf{X}_t = \{\mathbf{x}_i^t\}_{i=1}^{N_t}$, where $\mathbf{x}_i^t \in \Omega$, with an existing but unknown set of outputs/labels $\mathbf{Y}_t = \{y_i^t\}_{i=1}^{N_t}$, with $y_i^t \in \mathcal{C}$.

To estimate the set of outputs/labels $\mathbf{Y}_t$ of the test data $\mathbf{X}_t$, we can relies on learning or estimating the joint probability distribution $\mathbf{P}(\mathbf{x}, y) \in \mathcal{P}(\Omega \times \mathcal{C})$ from the training data and outputs/labels $(\mathbf{X}_s, \mathbf{Y}_s)$, and then apply it to the test data $\mathbf{Y}_t$ under the assumption that $\mathbf{X}_s$ and $\mathbf{X}_t$ are drawn from the same distribution $\mathbf{P}(\mathbf{x}) \in \mathcal{P}(\Omega)$.

2.2 Domain Adaptation

In the case of domain adaptation, we assumes that the joint probability distributions $\mathbf{P}_s(\mathbf{x}^s, y)$ and $\mathbf{P}_t(\mathbf{x}^t, y)$ of the training and testing data are different. Here the two sets of data $\mathbf{X}_s$ and $\mathbf{X}_t$ will then be called as *source* and *target* data as they are corresponding to a *source* and a *target* domains, which are denoted as Ω_s and Ω_t. We also denote the marginal distributions over $\mathbf{X}$ of the source and target domains as μ_s and μ_t.

According to [1], two popular assumptions, that are frequently made in most domain adaptation approaches, relies on the similarity of conditional probability distributions while the marginal distributions are supposed to be different. One assumption related to the case of **Class Imbalance** states that the source and target output/labels distributions are different ($\mathbf{P}_s(y) \neq \mathbf{P}_t(y)$) while the conditional distributions of the input with respect to the outputs/labels are the same ($\mathbf{P}_s(\mathbf{x}^s|y) = \mathbf{P}_t(\mathbf{x}^t|y)$). Another popular assumption presents the case of **Covariate Shift** through the argument that the conditional distributions of the outputs/labels with respect to the inputs are the same ($\mathbf{P}_s(y|\mathbf{x}^s) = \mathbf{P}_t(y|\mathbf{x}^t)$) while the source and target input distribution are different ($\mathbf{P}_s(\mathbf{x}^s) \neq \mathbf{P}_t(\mathbf{x}^t)$).

2.3 Domain Adaptation as Optimal Transportation problem - An Initial Formulation

In the real world application, the difference in the joint distributions often occurs due to the changes in both the marginal and conditional distributions. Following the work in [1], it is proposed that such difference, called *domain drift*, is caused by a transformation solely in the input space $\mathbf{T} : \Omega_s \to \Omega_t$. This transformation can be explained as the differences in the physical process of obtaining the input data (e.g. different lighting, angles, backgrounds, random noises, etc in pictures) or the inherent but small differences in the input data (e.g. between pictures of a product from one manufacturer and pictures of the same product from another manufacturer).

Subsequently, it is supposed to preserve the conditional probabilities of the source and target domains under

the transformation, i.e.

$$\mathbf{P}_s(y|\mathbf{x}) = \mathbf{P}_t(y|\mathbf{T}(\mathbf{x})), \quad \forall \mathbf{x} \in \Omega_s,$$

that is, to maintain the outputs/labels of the source input data under the source true regression/labelling function $f_s : \Omega_s \to \mathcal{C}$ before the transformation, and under the target true regression/labelling function $f_t : \Omega_t \to \mathcal{C}$ after the transformation $\mathbf{T}$, i.e.

$$f_s(\mathbf{x}) = f_t(\mathbf{T}(\mathbf{x})), \quad \forall \mathbf{x} \in \Omega_s.$$

Furthermore, the transformation $\mathbf{T}$ can be understood as a **push-forward** from the source input measure μ_s to the target input measure μ_t, i.e.

$$\mathbf{T}_{\#}\mu_s = \mu_t$$

where for any measure $\alpha \in \mathcal{P}(\Omega_s)$ and any measurable set $\mathcal{B} \subset \Omega_t$, the push-forward measure $\mathbf{T}_{\#}\alpha \in \mathcal{P}(\Omega_t)$ is defined as

$$\mathbf{T}_{\#}\alpha(\mathcal{B}) = \alpha(\{\mathbf{x} \in \Omega_s : \mathbf{T}(\mathbf{x}) \in \mathcal{B}\}) = \alpha(\mathbf{T}^{-1}(\mathcal{B})).$$

Intuitively speaking, while the transformation $\mathbf{T}$ moves a single point between the input spaces, the operator $\mathbf{T}_{\#}$ moves the whole probability measure from the input space Ω_s towards the input space Ω_t.

To reduce the search space of all possible transformations, [1] proposed to search for the transformation $\mathbf{T}$ that minimize a transportation cost

$$C(\mathbf{T}) = \int_{\Omega_s} c(\mathbf{x}, \mathbf{T}(\mathbf{x}))d\mu(\mathbf{x}), \quad \mathbf{T}_{\#}\mu_s = \mu_t \tag{1}$$

where $c : \Omega_s \times \Omega_t \to \mathbb{R}_+$ is a distance function, while satisfying .

The total cost $C(\mathbf{T})$ can be interpreted as the total energy cost to transport the every probability masses from the measure μ_s in the source domain to the measure μ_t in the target domain.

2.4 Joint Distribution Optimal Transport - Second Formulation

The initial formulation given as a minimization problem might be hindered by the two main factors:

- It assumes the preservation of the conditional distributions under the transformation, and subsequently the marginal distributions of the outputs/labels between the source and target domains. Such assumptions might not hold for many circumstances, for example, when the ranges of outputs in the two domains are clearly not aligned or simply when the set of labels of one domain has a much higher proportion of one class than that of the other domain.

- It also lacks a natural integration of the source outputs/labels in the searching for the optimal transformation. This has been partially resolved in the same work [1] proposing such formulation by adding additional regularization terms to accommodate the known information regarding the source outputs/labels but not at the fundamental level.

To handle the changes in both the marginal and conditional distributions between the source and target domains, instead of finding a transformation only in the input space Ω, [2] proposed to search for a transformation $\mathbf{T} : \Omega_s \times \mathcal{C} \to \Omega_t \times \mathcal{C}$ that would be able to directly align the joint distributions of the source and target domains, denoted as $\mu_{\mathbf{X}_s,\mathbf{Y}_s}$ and $\mu_{\mathbf{X}_t,\mathbf{Y}_t}$, by minimizing a similar transportation cost

$$C(\mathbf{T}) = \int_{\Omega_s \times \mathcal{C}} D(\mathbf{x}, y; \mathbf{T}(\mathbf{x}, y)) d\mu(\mathbf{x}, y), \quad \mathbf{T}_{\#}\mu_{\mathbf{X}_s,\mathbf{Y}_s} = \mu_{\mathbf{X}_t,\mathbf{Y}_t}$$

where $D : (\Omega_s \times \mathcal{C}) \times (\Omega_t \times \mathcal{C}) \to \mathbb{R}_+$ is a joint cost distance function.

In [2], they also adopted a separable type of cost functions $D(\mathbf{x}_s, y_s; \mathbf{x}_t, y_t) = \alpha c(\mathbf{x}_s, \mathbf{x}_t) + \mathcal{L}(y_s, y_t)$ instead of a more generic one. The distance function $c : \Omega_s \times \Omega_t \to \mathbb{R}_+$ in the first term is similar to the one in the initial formulation in Equation 1, while the second term $\mathcal{L} : \mathcal{C} \times \mathcal{C} \to \mathbb{R}_+$ is another distance function on the output/label space measuring the difference of the source and target outputs/labels. Finally, the coefficient $\alpha \in \mathbb{R}_+$ is a trade-off parameter to balance the cost function in the input and output/label spaces.

Here, given the problem of unsupervised domain adaptation, the information regarding the outputs/labels in the target domain is often unknown. Thus, the joint distribution $\mu_{\mathbf{X}_t,\mathbf{Y}_t}$ in the target domain can be replaced with an estimate $\mu_{\mathbf{X}_t,f(\mathbf{X}_t)}$ through a regression/classification function $f : \Omega_t \to \mathcal{C}$. At the same time, the function f should be chosen such that the estimated target joint distribution would be similar to that in the source domain. This leads to the problem of searching for a pair of function f and transformation $\mathbf{T}$ that minimizes the following transportation cost

$$C(\mathbf{T}) = \int_{\Omega_s \times \mathcal{C}} D(\mathbf{x}, y; \mathbf{T}(\mathbf{x}, y)) d\mu(\mathbf{x}, y), \quad \mathbf{T}_{\#}\mu_{\mathbf{X}_s,\mathbf{Y}_s} = \mu_{\mathbf{X}_t,f(\mathbf{X}_t)} \tag{2}$$

3 Optimal Transportation

3.1 Monge formulation

For two arbitrary probability measures α and β supported on two spaces $\mathcal{X}$ and $\mathcal{Y}$, and a distance function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, Monge formulation of the optimal transportation problem searches for a transformation $\mathbf{T} : \mathcal{X} \to \mathcal{Y}$ that satisfies the following minimization problem

$$\min_{\mathbf{T}} \left\{ \int_{\mathcal{X}} c(\mathbf{x}, \mathbf{T}(\mathbf{x})) d\alpha(\mathbf{x}) : \quad \mathbf{T}_{\#}\alpha = \beta \right\} \tag{3}$$

3.2 Kantorovich relaxation

To relax the deterministic nature of the transportation in Monge formulation, that is, each point $\mathbf{x}$ in the source domain $\mathcal{X}$ is assigned to another point $\mathbf{T}(\mathbf{x})$, Kantorovich's idea is to adopt a probabilistic transportation by allowing each source point $\mathbf{x}$ to be split towards multiple targets.

Let $\Pi(\alpha, \alpha) \subset \mathcal{P}(\mathcal{X} \times \mathcal{Y}$ be the set of all probabilistic couplings with marginals α and α. For two arbitrary probability measures α and β on two spaces $\mathcal{X}$ and $\mathcal{Y}$, and a distance function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, Kantorovich

formulation searches for a coupling $\boldsymbol{\gamma} \in \Pi(\alpha, \beta)$ that satisfies the minimization problem

$$\min_{\boldsymbol{\gamma} \in \Pi(\alpha,\beta)} \left\{ \int_{\mathcal{X} \times \mathcal{Y}} c(\mathbf{x}, \mathbf{y}) d\gamma(\mathbf{x}, \mathbf{y}) \right\} \tag{4}$$

3.3 Wasserstein distance

With the Kantorovich formulation, it allows the definition of the Wasserstein distance of order p between α and β as

$$W_p(\alpha, \beta) = \left(\inf_{\boldsymbol{\gamma} \in \Pi(\alpha,\beta)} \int_{\mathcal{X} \times \mathcal{Y}} d(\mathbf{x}, \mathbf{y})^p d\gamma(\mathbf{x}, \mathbf{y}) \right)^{\frac{1}{p}} \tag{5}$$

where $d : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ is a distance and $c(\mathbf{x}, \mathbf{y}) = d(\mathbf{x}, \mathbf{y})^p$.

3.4 Discrete Optimal Transport

In many practical problems, we often relies on empirically estimated distribution, which can be represented as discrete probability measures. A discrete probability measure α with weights $\mathbf{a} \in \mathbb{R}_+^n$ s.t. $\sum_{i=1}^n \mathbf{a}_i = 1$ and locations $\mathbf{x}_1, ...\mathbf{x}_n \in \mathcal{X}$ can be expressed as

$$\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{\mathbf{x}_i}$$

where δ_x is defined as the Dirac at point $\mathbf{x}$.

For two discrete probability measures

$$\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{\mathbf{x}_i} \quad \text{and} \quad \beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{\mathbf{y}_j},$$

and a distance function $c : \mathcal{X} \times \mathbf{Y} \to \mathbb{R}_+$, Monge formulation searches for a transportation map $\mathbf{T} : \{\mathbf{x}_1, ..., \mathbf{x}_n\} \to \{\mathbf{y}_1, ..., \mathbf{y}_2\}$ that assigns each point $\mathbf{x}_i$ to a single point $\mathbf{y}_1$ and satisfies the following discrete minimization problem

$$\min \left\{ \sum_{i=1}^{n} c(\mathbf{x}_i, \mathbf{T}(\mathbf{x}_i)) : \mathbf{T}_{\#}\alpha = \beta \right\} \tag{6}$$

where the condition $\mathbf{T}_{\#}\alpha = \beta$ for discrete optimal transportation problem can be expressed as

$$\forall j \in 1, 2, ..., m, \mathbf{b}_j = \sum_{i:\mathbf{T}(\mathbf{x}_i)=\mathbf{y}_j} \mathbf{a}_i.$$

For the discrete case of Kantorovich formulation, the set of all probabilistic couplings with marginals α and β can be defined as

$$\boldsymbol{\Gamma}(\alpha, \beta) = \{\boldsymbol{\gamma} \in \mathbb{R}_+^{n \times m} | \boldsymbol{\gamma} \mathbf{1}_n = \alpha, \boldsymbol{\gamma}^\top \mathbf{1}_m = \beta\} \tag{7}$$

where $\mathbf{1}_d$ is a d-dimensional vector of ones. Then Kantorovich formulation then searches for a coupling $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}(\alpha, \beta)$ that satisfies the following discrete minimization problem

$$\min_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}(\alpha,\beta)} \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F \tag{8}$$

where $\langle ., . \rangle_F$ is the Frobenius dot-product and $\mathbf{C} \in \mathbb{R}_+^{n \times m}$ is the cost matrix with $\mathbf{C}_i j = c(\mathbf{x}_i, \mathbf{y}_j)$.

4 Solving the discrete optimization problems of domain adaptation

4.1 Reformulate the problems with Kantorovich formulation

4.1.1 The inital formulation

In our first formulation of the domain adataption problem given in Equation 1, to apply it to practical source and target data, we relies on empirically estimating the distributions of the source and target input data $\mathbf{X}_s$ and $\mathbf{X}_t$, which can be represented as discrete measures

$$\mu_s = \sum_{i=1}^{N_s} \frac{1}{N_s} \delta_{\mathbf{x}_i^s} \quad \text{and} \quad \mu_s = \sum_{j=1}^{N_t} \frac{1}{N_t} \delta_{\mathbf{x}_j^t} \tag{9}$$

Following the discrete optimization problem under Kantorovich formulation given in Equation 8, the task is to find a coupling $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}(\mu_s, \mu_t)$ that satisfies the following discrete minimization problem

$$\min_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}(\mu_s, \mu_t)} \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F \tag{10}$$

with a cost matrix $\mathbf{C} \in \mathbb{R}_+^{N_s \times N_t}$.

4.1.2 The second formulation

Similarly, for the second formulation of the domain adaptation problem given in Equation 2, the empirically estimated joint distributions of the source and target, with the use of regression/classification function f, can be expressed as

$$\mu_{\mathbf{X}_s, \mathbf{Y}_s} = \sum_{i=1}^{N_s} \frac{1}{N_s} \delta_{(\mathbf{x}_i^s, y_i^s)} \quad \text{and} \quad \mu_{\mathbf{X}_t, f(\mathbf{X}_t)} = \sum_{j=1}^{N_t} \frac{1}{N_t} \delta_{(\mathbf{x}_i^s, f(\mathbf{x}_i^s))} \tag{11}$$

Then follow the discrete optimization problem under Kantorovich formulation given in Equation 8, the problem can be defined as

$$\min_{f, \boldsymbol{\gamma} \in \boldsymbol{\Gamma}(\mu_{\mathbf{X}_s, \mathbf{Y}_s}, \mu_{\mathbf{X}_t, f(\mathbf{X}_t)})} \langle \boldsymbol{\gamma}, \mathbf{D} \rangle_F \tag{12}$$

with a cost matrix $\mathbf{D} \in \mathbb{R}_+^{N_s \times N_t}$ between each data point in source and target domains.

4.2 Solving the optimal transportation problem in 1-dimensional space

For abitrary measures in high dimensional spaces, analytical solutions to Monge or Kantorovich formulations are often hard to find even if the minimizers exist especially for continuous measures. But in the simple cases where the source and target spaces Ω_s and Ω_t are in 1-dimensional space and the cost function $c(\mathbf{x}_s, \mathbf{x}_t) = |\mathbf{x}_s - \mathbf{x}_t|$ is a 1-norm function, the Monge map $\mathbf{T} : \Omega_s \to \Omega_t$ [3] is given by

$$\mathbf{T}(\mathbf{x}_s) = F_t^{-1}(F_s(\mathbf{x}_s)) \tag{13}$$

where $F_s : \Omega_s \to [0, 1]$ and $F_t : \Omega_s \to [0, 1]$ are the Cumulative Distribution Function (CDF) in the source and target domains.

4.3 Solving the optimal transportation problem in high dimensional space for Gaussian distributions

Let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the multivariate Gaussian distribution on $\mathbb{R}^d$ with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$. The 2-Wasserstein distance between two multivariate Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$ and $\mathcal{N}(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b)$ has a closed-form solution [4] [5] called the Wasserstein-Bures or Fréchet distance

$$W_2\left(\mathcal{N}(\boldsymbol{\mu_a}, \boldsymbol{\Sigma}_a), \mathcal{N}(\boldsymbol{\mu_b}, \boldsymbol{\Sigma}_b)\right) = \|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|^2 + \operatorname{tr}(\boldsymbol{\Sigma}_a) + \operatorname{tr}(\boldsymbol{\Sigma}_b) - 2\operatorname{tr}\left(\left(\boldsymbol{\Sigma}_a^{\frac{1}{2}}\boldsymbol{\Sigma}_b\boldsymbol{\Sigma}_a^{\frac{1}{2}}\right)^{\frac{1}{2}}\right) \tag{14}$$

5 One Dimensional Examples

5.1 Process description

5.1.1 Data generation

The set of source data $\mathbf{X}_s = \{x_i^s\}_{i=1}^{N_s}$ with $x_i^s \in \mathbb{R}$ is sampled from a Gaussian mixture distribution of n_g^s number of component Gaussian distributions. It corresponds to the set of known and available output $\mathbf{Y}_s = \{y_i^s\}_{i=1}^{N_s} = \{f_s(x_i^s)\}_{i=1}^{N_s}$ through the function $f_s : \mathbb{R} \to \mathbb{R}$

The set of target data $\mathbf{X}_t = \{x_i^t\}_{i=1}^{N_t}$ with $x_i^t \in \mathbb{R}$ is sampled from a Gaussian mixture distribution of n_g^t number of component Gaussian distributions with existing but unknown labels.

5.1.2 Target ouput generation

To ensure the distributions of the outputs in the source and target spaces are similar, a set of intermediate output for the target data is generated as $\bar{\mathbf{Y}}_t = \{\bar{y}_i^t\}_{i=1}^{N_t} = \{g_t(x_i^t)\}_{i=1}^{N_t}$ with $g_t : \mathbb{R} \to \mathbb{R}$, then it is transformed to the final set of unknown output $\mathbf{Y}_t = \{y_i^t\}_{i=1}^{N_t} = \{F_{\mathbf{Y}_s}^{-1}(F_{\bar{\mathbf{Y}}_t}(\bar{y}_i^t))\}_{i=1}^{N_t}$.

5.1.3 Data samples transformation

The target data samples $\mathbf{X}_t$ is transformed to follow the distribution of the source data samples as

$$\bar{\mathbf{X}}_t = \{\bar{x}_i^t\}_{i=1}^{N_t} = \{F_{\mathbf{X}_s}^{-1}(F_{\mathbf{X}_t}(x_i^t))\}_{i=1}^{N_t}$$

where $F_{\mathbf{X}_s}^{-1}$ is the approximated inverse CDF of the source distribution and $F_{\mathbf{X}_t}(x_i^t))$ is the approximated CDF of the target distribution.

The transformed data samples are then evaluated under the true regression function f_s to produce the predicted outputs of the target data samples.

5.2 Examples and Observations

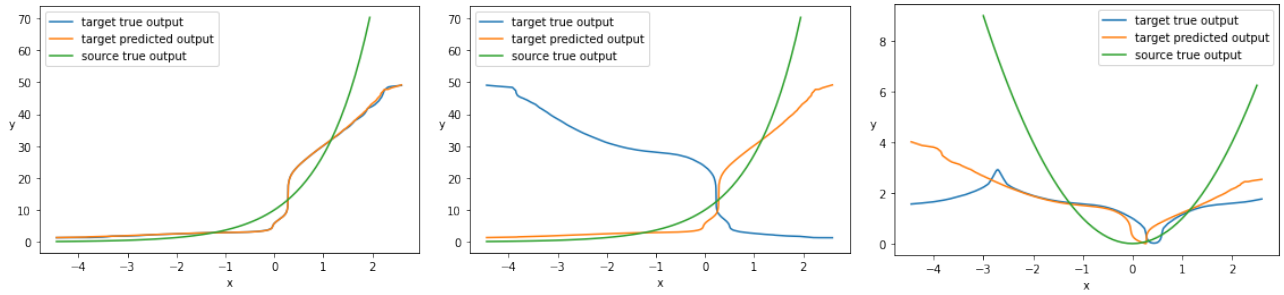


Figure 1: The three figures are of the cases where the function f_s and g_t are both monotonically increasing (left), f_s is monotonically increasing while g_t is monotonically decreasing (middle), and f_s and g_t are both non-monotonic(right).

The Figure 1 contains three prominent one-dimensional examples, more examples can be found in the Appendix A.

From these examples, our first formulation of domain adaptation problem given in Equation 1 can work well for the cases when the true regression functions f_s and f_t are both either monotonically increasing or decreasing. This is due to the monotonic nature of the Cumulative Distribution Functions (CDFs), which in turn making the Monge map in Equation 13 monotonic.

Differently, for the cases when the true regression functions f_s and f_t are both monotonic but one is increasing while the other is decreasing, or when either of them are non-monotonic, the estimated regression functions on the target domains are not able to produce the same results. This is also due to the monotonic nature of the Monge map, leading to the inability to deal with non-monotonic regression functions.

6 An example of Optimization Surface of Joint Distribution Optimal Transportation Approach

6.1 Analytical Solution to Simple Gaussian Distributions

Define the following random variable as follows:

$$\mathbf{X}_s \sim \mathcal{N}(\mu_s, \sigma_s)$$
$$\mathbf{X}_t \sim \mathcal{N}(\mu_t, \sigma_t)$$
$$\boldsymbol{\epsilon}_s \sim \mathcal{N}(0, 1)$$
$$\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, 1)$$
$$\mathbf{Y}_s = a\mathbf{X}_s + b\boldsymbol{\epsilon}_s$$
$$\mathbf{Y}_t = c\mathbf{X}_t + d\boldsymbol{\epsilon}_t$$

where $\mathbf{X}_s$, $\mathbf{X}_t$, $\boldsymbol{\epsilon}_s$ and $\boldsymbol{\epsilon}_t$ are independent random variables, $\mu_s, \mu_t, a, b, c, d \in \mathbb{R}$ and $\sigma_s, \sigma_t \in \mathbb{R}^+$.

Then the joint distribution of $\mathbf{X}_s$ and $\mathbf{Y}_s$ can be written as

$$\begin{pmatrix} \mathbf{X}_s \\ \mathbf{Y}_s \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ a & b \end{bmatrix} \begin{pmatrix} \mathbf{X}_s \\ \boldsymbol{\epsilon}_s \end{pmatrix} \tag{15}$$

Based on the independent of Gaussian random variables $\mathbf{X}_s$ and $\boldsymbol{\epsilon}_s$, their joint distribution is also normally distributed as

$$\begin{pmatrix} \mathbf{X}_s \\ \boldsymbol{\epsilon}_s \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_s \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_s & 0 \\ 0 & 1 \end{bmatrix} \right) \tag{16}$$

Thus, the joint distribution of $\mathbf{X}_s$ and $\mathbf{Y}_s$ is normally distributed as

$$\begin{aligned} \begin{pmatrix} \mathbf{X}_s \\ \mathbf{Y}_s \end{pmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} 1 & 0 \\ a & b \end{bmatrix} \begin{bmatrix} \mu_s \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ a & b \end{bmatrix} \begin{bmatrix} \sigma_s & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ a & b \end{bmatrix}^\top \right) \\ &= \mathcal{N}\left(\begin{bmatrix} \mu_s \\ a\mu_s \end{bmatrix}, \begin{bmatrix} \sigma_s & a\sigma_s \\ a\sigma_s & a^2\sigma_s + b^2 \end{bmatrix} \right) \end{aligned} \tag{17}$$

Similarly, the joint distribution of $\mathbf{X}_t$ and $\mathbf{Y}_s$ is normally distributed as

$$\begin{pmatrix} \mathbf{X}_t \\ \mathbf{Y}_t \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_t \\ c\mu_t \end{bmatrix}, \begin{bmatrix} \sigma_t & c\sigma_t \\ c\sigma_t & c^2\sigma_t + d^2 \end{bmatrix} \right) \tag{18}$$

Follow Equation 14, the 2-Wasserstein distance between the joint distributions of $(\mathbf{X}_s, \mathbf{Y}_s)$ and $(\mathbf{X}_t, \mathbf{Y}_t)$ is

$$\begin{aligned} &W_2\left(\mathcal{N}\left(\begin{bmatrix} \mu_s \\ a\mu_s \end{bmatrix}, \begin{bmatrix} \sigma_s & a\sigma_s \\ a\sigma_s & a^2\sigma_s + b^2 \end{bmatrix} \right), \mathcal{N}\left(\begin{bmatrix} \mu_t \\ c\mu_t \end{bmatrix}, \begin{bmatrix} \sigma_t & c\sigma_t \\ c\sigma_t & c^2\sigma_t + d^2 \end{bmatrix} \right) \right) \\ &= (\mu_s - \mu_t)^2 + (a\mu_s - c\mu_t)^2 + (\sigma_s + a^2\sigma_s + b^2) + (\sigma_t + c^2\sigma_t + d^2) \\ &- 2\mathrm{tr}\left(\begin{bmatrix} \sigma_s\sigma_t + ac\sigma_s\sigma_t & c\sigma_s\sigma_t + a\sigma_s(c^2\sigma_t + d^2) \\ a\sigma_s\sigma_t + (a^2\sigma_s + b^2)c\sigma_t & ac\sigma_s\sigma_t + (a^2\sigma_s + b^2)(c^2\sigma_t + d^2) \end{bmatrix}^{\frac{1}{2}} \right) \end{aligned} \tag{19}$$

6.2 Optimization Surfaces

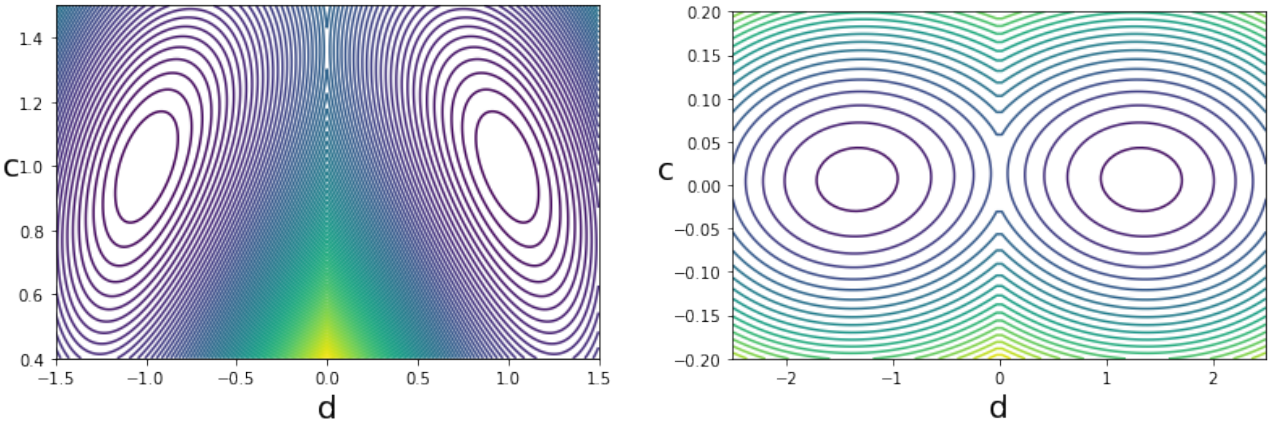


Figure 2: The optimization surfaces with respect to the coefficients c and d of target regression functions for the cases of $\mu_s = 0, \sigma_s = 1, a = 1, b = 1, \mu_t = 0, \sigma_t = 1$ (left) and $\mu_s = 0, \sigma_s = 1, a = 1, b = 1, \mu_t = 10, \sigma_t = 2$ (right)

For our continuous Gaussians distribution with linear regression functions, as shown in Figure 2, the optimization surfaces of the second formulation of domain adaptation problem are seemingly convex for the coefficient c of the inputs while being bimodal with respect to the coefficient d of the error term. It is expected since this is a simple problem and d only affects the size of the variance of the error term.

7 Examples of Joint Distribution Optimal Transport (JDOT)

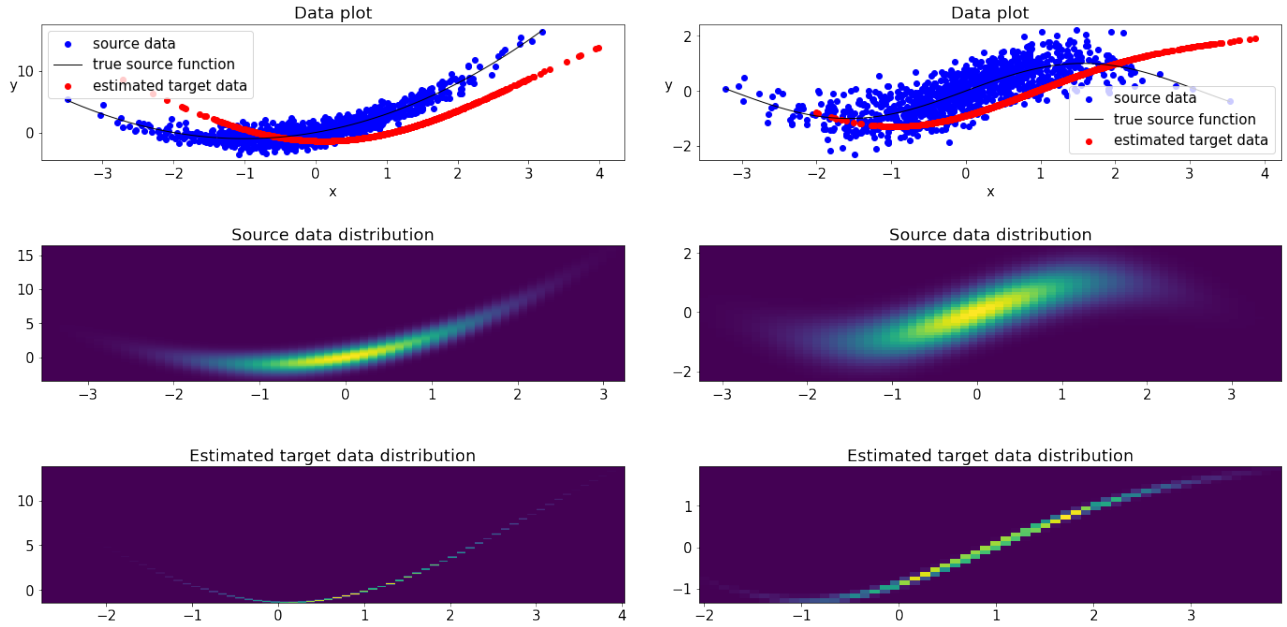


Figure 3: Two examples of applying JDOT formulation, with the cost functions chosen to use Support Vector Machines (SVMs) for estimated regression functions, to some toy examples with data generated from Gaussians distributions with different regressions functions in the source domains (For estimated target data distributions, small noises are added to visualize degenerate joint distribution).

As shown in Figure 3, the second formulation given in Equation 2 can provides estimates for the true regression functions in the target domains using the information from the joint distribution of source data. Note that it is still unclear about the properties of such estimates.

8 Discussion

Under the first formulation of domain adaptation problem given in Equation 1, it provides a great performance for monotonic regressions equations in 1 dimensional spaces. In higher dimensional space, the concept of monotonicity is less of a problem, thus further research can be carried out on the performance and properties of this formulation for high dimensional spaces. Additionally, there exists different ways to use the optimal transportation plan to estimate the target outputs/labels under different motivations, which can greatly affect the performance and properties of the estimated functions.

For the second formulation of domain adaptation in Equation 2, without any further assumptions on the distributions of the unknown target outputs/labels, it can only provide an initial estimate on the target regression/classification functions without clear properties and guaranteed performance. Future work can be used on detecting its properties and limitation, as well as how to incorporate some assumptions on the distribution of the target outputs/labels.

At the same time, our formulation relies heavily on the ability to solve the optimal transportation problem or equivalently computing the Wasserstein distance. Thus, although solving the optimal transportation problems for empirical discrete measures numerically is feasible despite high computational complexity, it is important to consider the cases between continuous and discrete distributions and how to compute the Wasserstein distance effectively.

A One Dimensional Example Results

A.1 Data descriptions

The source data $\mathbf{X}_s$ consist of $N_s = 100000$ samples from a Gaussian mixture distribution of $n_g^s = 3$ component Gaussian distribution as shown in Figure (4).

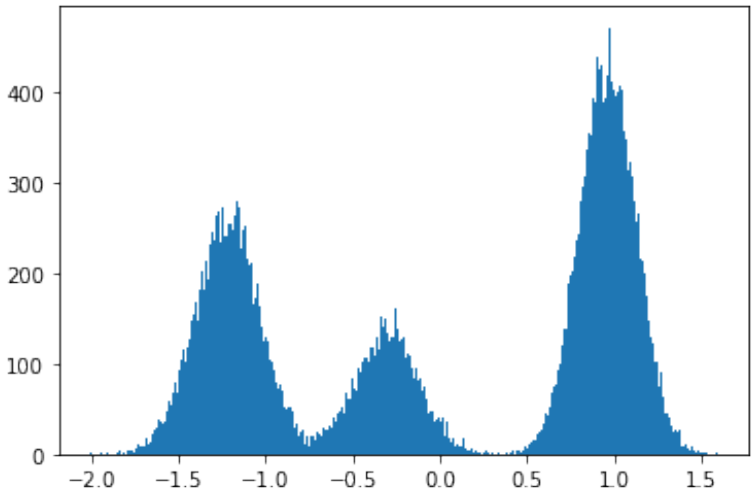


Figure 4: The histogram of the source data samples with bin size of 1000.

The target data $\mathbf{X}_t$ consist of $N_t = 90000$ samples from a Gaussian mixture distribution of $n_g^t = 2$ component Gaussian distribution as shown in Figure (5).

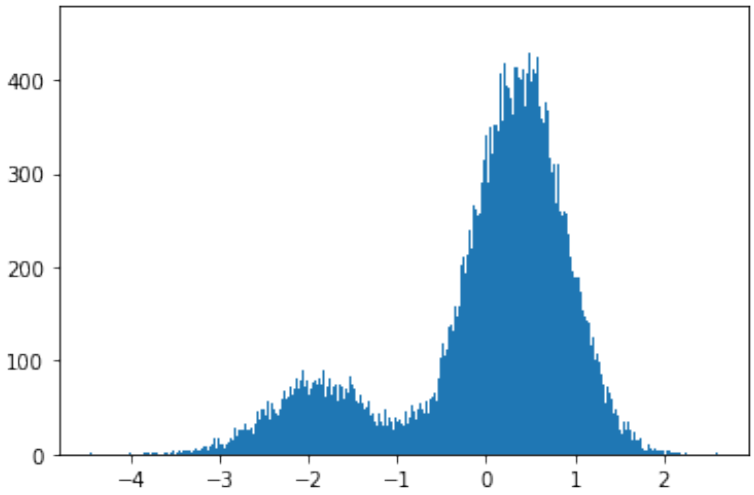


Figure 5: The histogram of the target data samples with bin size of 1000.

A.2 Example results

A.2.1 Example 1.1: f_s and g_t are both linearly increasing function

The functions f_s and g_t are of the forms

$$f_s(x) = x + 2$$

and

$$g_t(x) = 9x + 2$$

The histograms of the source true regression outputs, the target intermediate regression outputs and the target true regression outputs are shown in Figure (6).

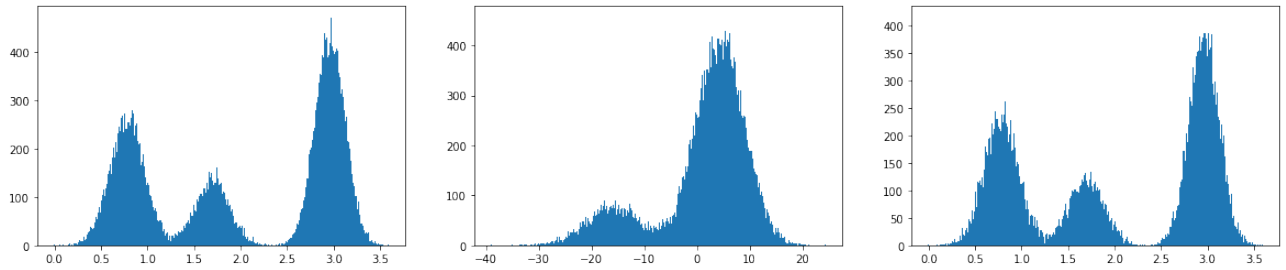


Figure 6: (Example 1.1) The histograms of the source true outputs (left), the target intermediate outputs (middle), the target true outputs (right) with bin size of 1000.

The target true regression outputs and predicted outputs are shown in Figure (7)

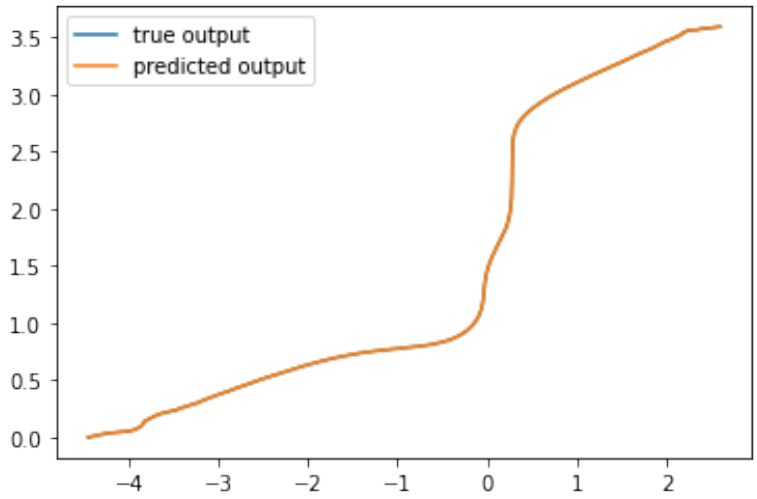


Figure 7: (Example 1.1) The target true outputs and predicted outputs.

A.2.2 Example 1.2: f_s and g_t are both linearly decreasing

The functions f_s and g_t are of the forms

$$f_s(x) = -x + 2$$

and

$$g_t(x) = -9x + 2$$

The histograms of the source true regression outputs, the target intermediate regression outputs and the target true regression outputs are shown in Figure (8).

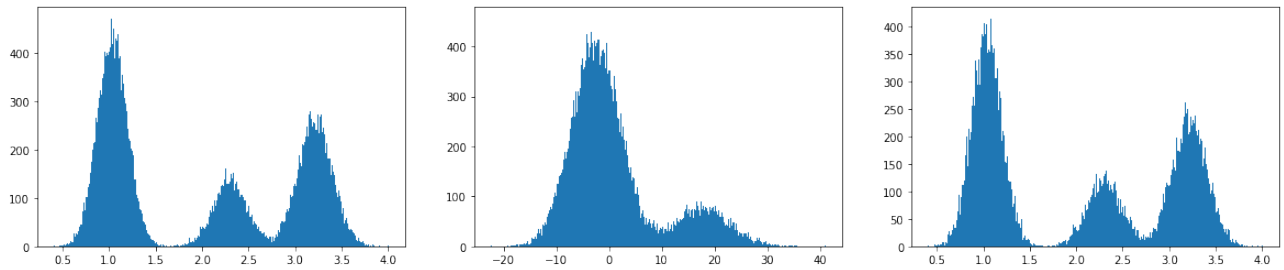


Figure 8: (Example 1.2) The histograms of the source true outputs (left), the target intermediate outputs (middle), the target true outputs (right) with bin size of 1000.

The target true regression outputs and predicted outputs are shown in Figure (9)

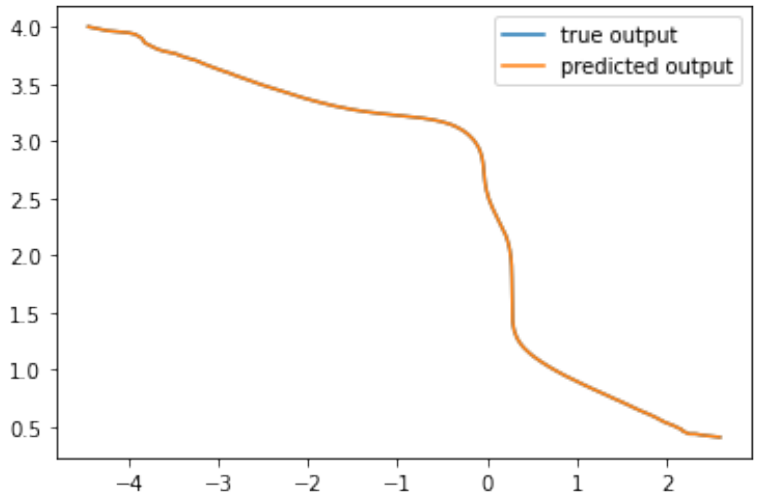


Figure 9: (Example 1.2) The target true outputs and predicted outputs.

A.2.3 Example 1.3: f_s is linearly increasing and g_t is linearly decreasing

The functions f_s and g_t are of the forms

$$f_s(x) = x + 2$$

and

$$g_t(x) = -9x + 2$$

The histograms of the source true regression outputs, the target intermediate regression outputs and the target true regression outputs are shown in Figure (10).

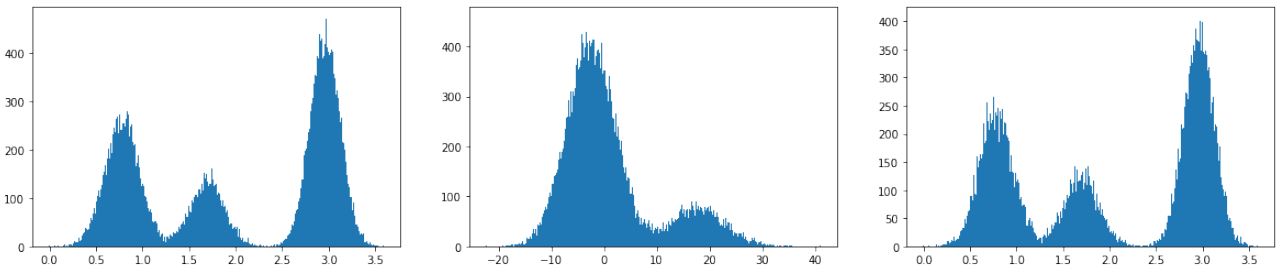


Figure 10: (Example 1.3) The histograms of the source true outputs (left), the target intermediate outputs (middle), the target true outputs (right) with bin size of 1000.

The target true regression outputs and predicted outputs are shown in Figure (11)

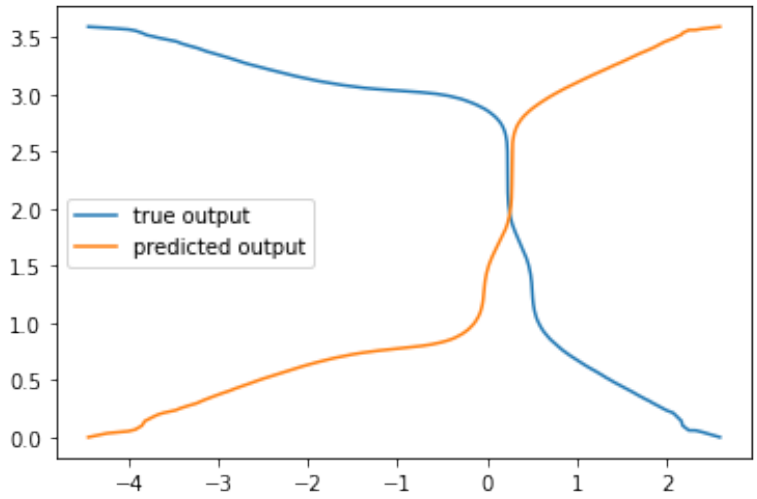


Figure 11: (Example 1.3) The target true outputs and predicted outputs.

A.2.4 Example 1.4: f_s is linearly decreasing and g_t is linearly increasing

The functions f_s and g_t are of the forms

$$f_s(x) = -x + 2$$

and

$$g_t(x) = 9x + 2$$

The histograms of the source true regression outputs, the target intermediate regression outputs and the target true regression outputs are shown in Figure (12).

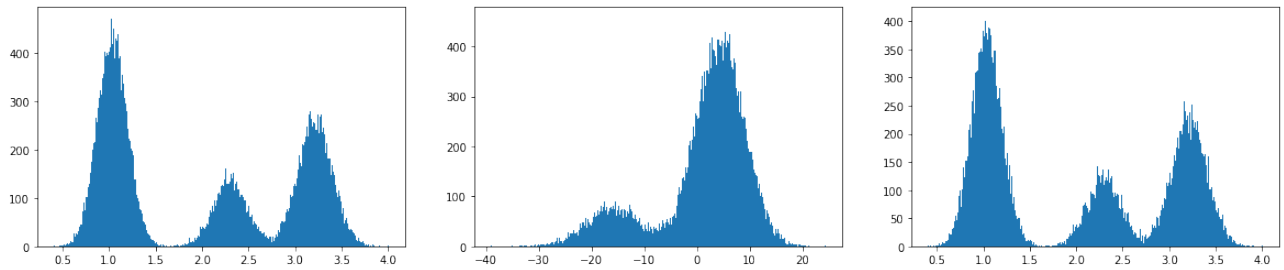


Figure 12: (Example 1.4) The histograms of the source true outputs (left), the target intermediate outputs (middle), the target true outputs (right) with bin size of 1000.

The target true regression outputs and predicted outputs are shown in Figure (13)

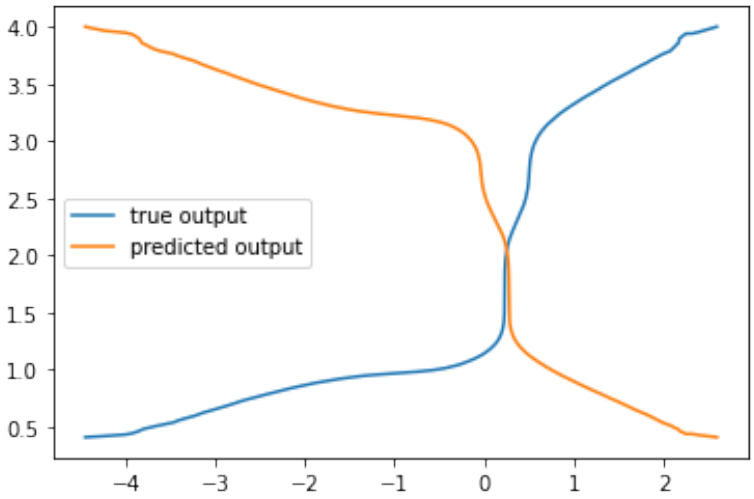


Figure 13: (Example 1.4) The target true outputs and predicted outputs.

A.2.5 Example 1.5: f_s and g_t are monotonically increasing

The functions f_s and g_t are of the forms

$$f_s(x) = 10e^x$$

and

$$g_t(x) = (x-5)^3$$

The histograms of the source true regression outputs, the target intermediate regression outputs and the target true regression outputs are shown in Figure (14).

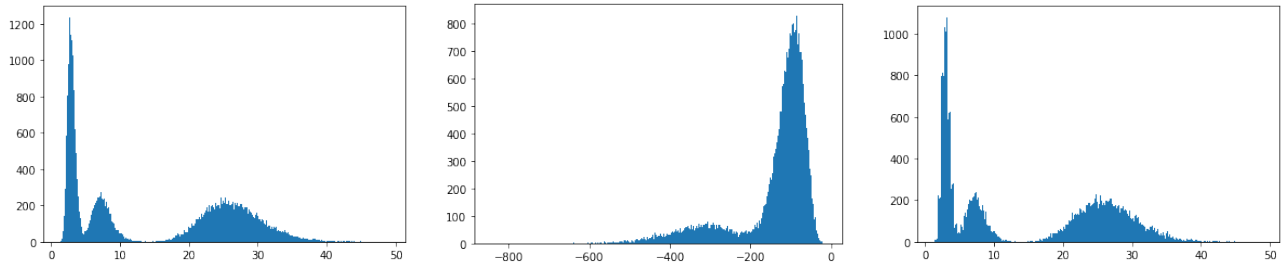


Figure 14: (Example 1.5) The histograms of the source true outputs (left), the target intermediate outputs (middle), the target true outputs (right) with bin size of 1000.

The target true regression outputs and predicted outputs are shown in Figure (15)

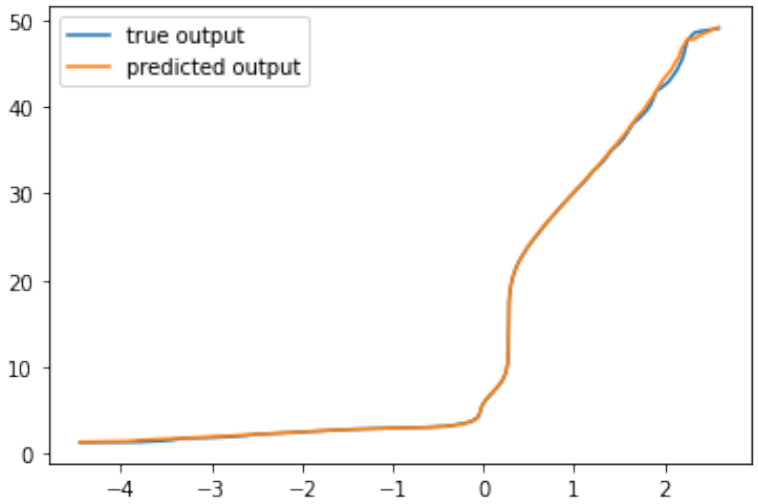


Figure 15: (Example 1.5) The target true outputs and predicted outputs.

A.2.6 Example 1.6: f_s and g_t are monotonically decreasing

The functions f_s and g_t are of the forms

$$f_s(x) = -10e^x$$

and

$$g_t(x) = -(x-5)^3$$

The histograms of the source true regression outputs, the target intermediate regression outputs and the target true regression outputs are shown in Figure (16).

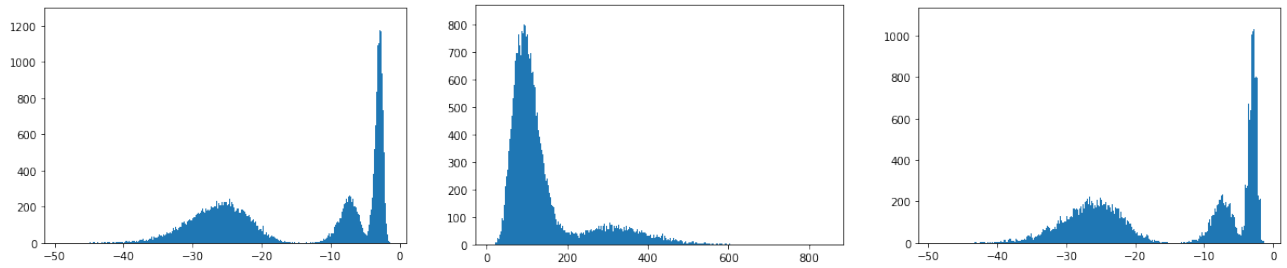


Figure 16: (Example 1.6) The histograms of the source true outputs (left), the target intermediate outputs (middle), the target true outputs (right) with bin size of 1000.

The target true regression outputs and predicted outputs are shown in Figure (17)

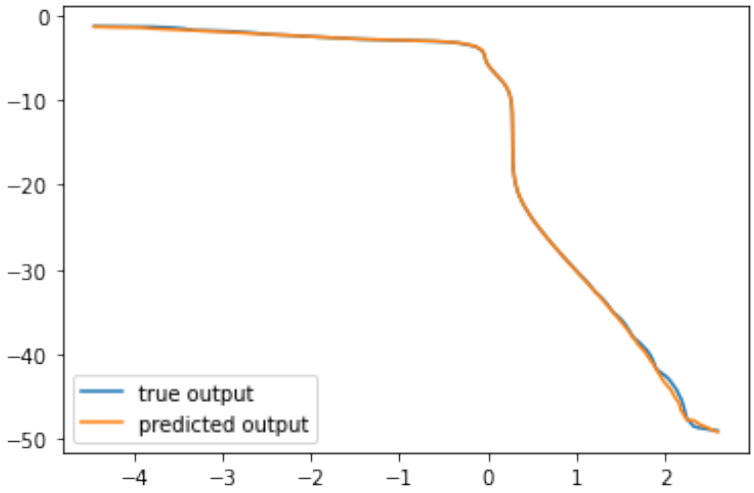


Figure 17: (Example 1.6) The target true outputs and predicted outputs.

A.2.7 Example 1.7: f_s is monotonically increasing and g_t is monotonically decreasing

The functions f_s and g_t are of the forms

$$f_s(x) = 10e^x$$

and

$$g_t(x) = -(x-5)^3$$

The histograms of the source true regression outputs, the target intermediate regression outputs and the target true regression outputs are shown in Figure (18).

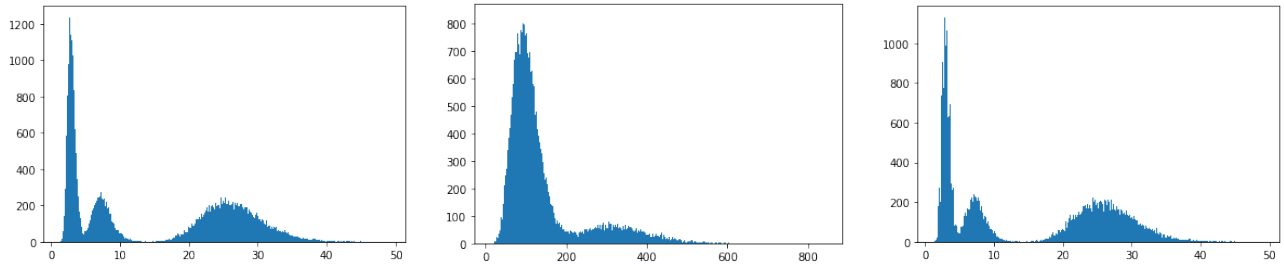


Figure 18: (Example 1.7) The histograms of the source true outputs (left), the target intermediate outputs (middle), the target true outputs (right) with bin size of 1000.

The target true regression outputs and predicted outputs are shown in Figure (19)

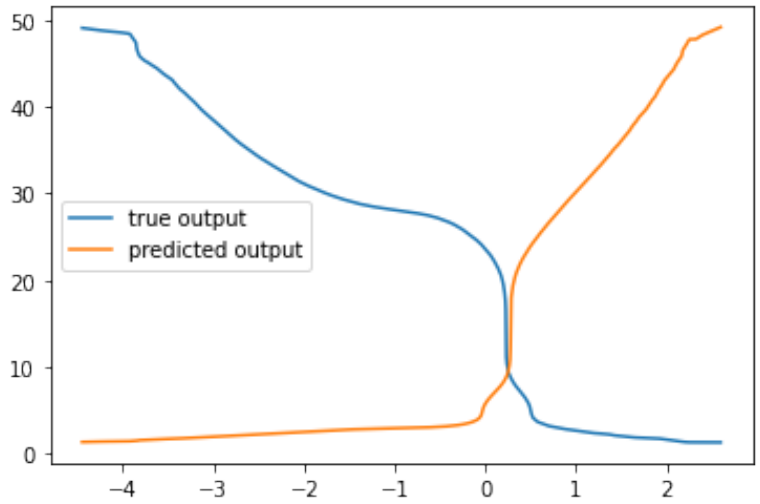


Figure 19: (Example 1.7) The target true outputs and predicted outputs.

A.2.8 Example 1.8: f_s is monotonically decreasing and g_t is monotonically increasing

The functions f_s and g_t are of the forms

$$f_s(x) = -10e^x$$

and

$$g_t(x) = (x-5)^3$$

The histograms of the source true regression outputs, the target intermediate regression outputs and the target true regression outputs are shown in Figure (20).

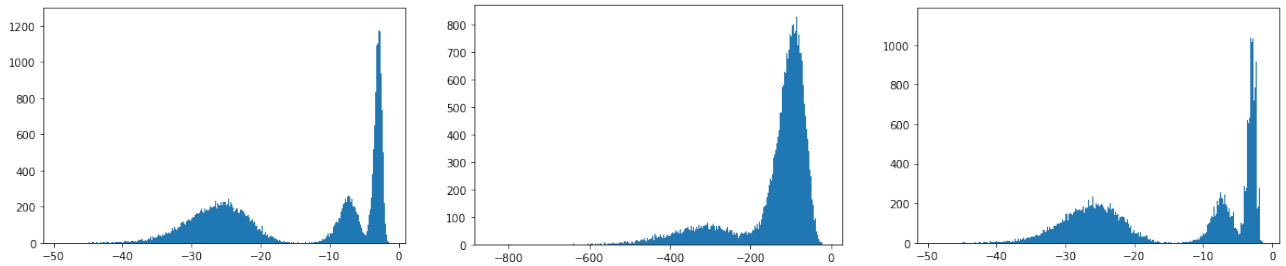


Figure 20: (Example 1.8) The histograms of the source true outputs (left), the target intermediate outputs (middle), the target true outputs (right) with bin size of 1000.

The target true regression outputs and predicted outputs are shown in Figure (21)

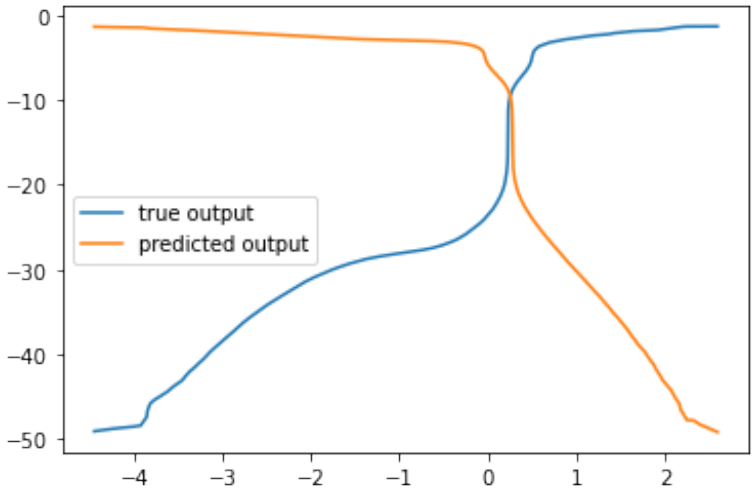


Figure 21: (Example 1.8) The target true outputs and predicted outputs.

A.2.9 Example 1.9: f_s and g_t are non-monotonic

The functions f_s and g_t are of the forms

$$f_s(x) = x^2$$

and

$$g_t(x) = 2 * sin(x - 2) + 2$$

which are shown in Figure (22).

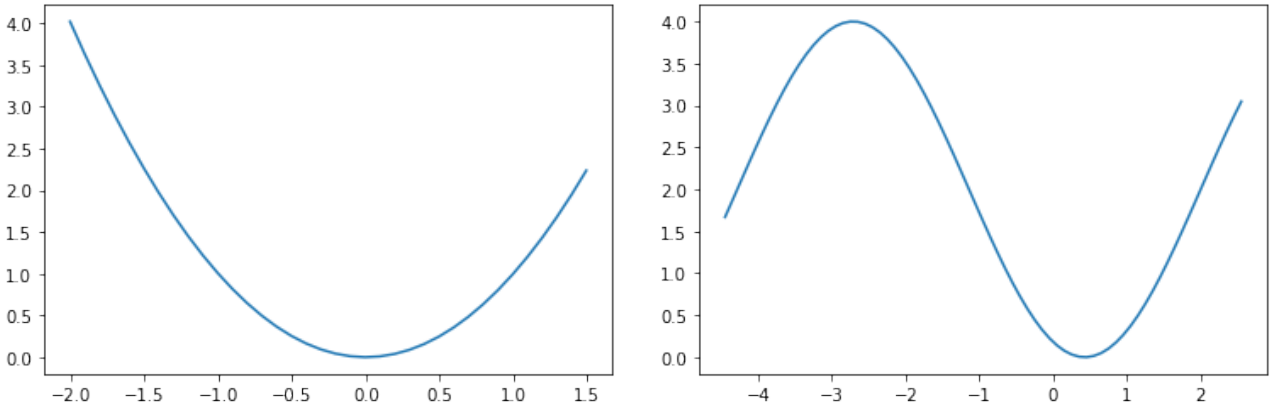


Figure 22: (Example 1.9) The plots of the functions f_s (left) and g_t (right) on the range of data samples.

The target true regression outputs and predicted outputs are shown in Figure (23)

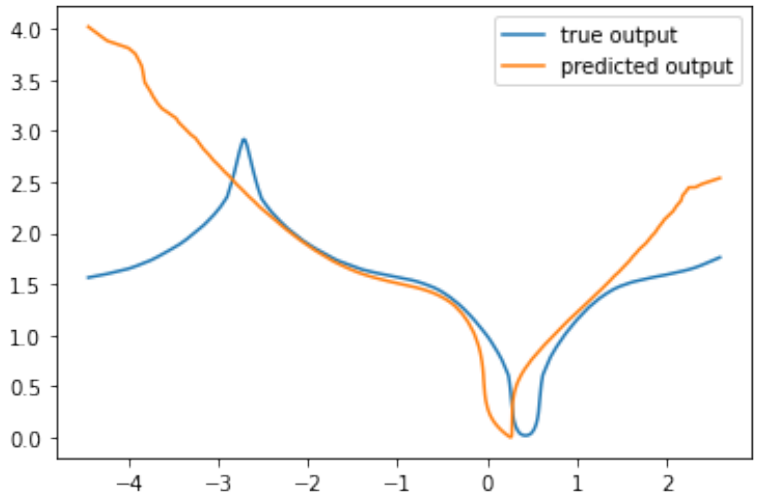


Figure 23: (Example 1.9) The target true outputs and predicted outputs.

References

[1] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation, 2016.

[2] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation, 2017.

[3] Gabriel Peyré and Marco Cuturi. Computational optimal transport, 2020.

[4] D.C Dowson and B.V Landau. The fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982.

[5] Asuka Takatsu. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005 – 1026, 2011.