Dynamic Screening for Unbalanced Optimal Transport Problem

Anonymous Author
Anonymous Institution

Abstract

HK comment: We change the story: (OT and UOT → Lasso-type UOT → Propose Screening (No literature) → Evaluation) The Safe Screening technique saves computational time by freezing the zero elements in the sparse solution of the Lasso problem. Recently, researchers have linked the UOT problem to the Lasso problem. In this paper, we apply the newest Dynamic Screening framework to the L_2 penalized Unbalanced Optimal Transport (UOT) problem. We firstly apply the Screening method onto the UOT problem. We find out that the specific structure for the UOT problem allows it to get better screening results than the Lasso problem. We propose the new Dynamic Screening algorithm and demonstrate its extraordinary effectiveness and potential to benefit from the unique structure of the UOT problem, our algorithm substantially improves the screening efficiency compared to the ordinary Lasso algorithm without significantly increasing on the computational complexity. We demonstrate the advantages of the algorithm through some experiments on the Gaussian distributions and the MNIST dataset.

1 INTRODUCTION

Optimal Transfer (OT) has a long history in mathematics and has recently become prevalent due to its important role in the machine learning community for measuring distances between histograms. It has outperformed traditional methods in many different areas such as domain adaptation (Courty, 2017), generative models (Arjovsky et al., 2017),, graph machine learning (Petric Maretic et al., 2019) and natural language processing. (Chen et al., 2019) Its popularity is attributed to the introduction of Sinkhorn's algorithm for the entropy optimal transmission problem, (Cuturi, 2013) which performs faster in large scale OT problem than the Simplex's method. In order to extend the OT problem, which can only handle balanced samples, to a wider range of unbalanced samples. The unbalanced optimal transport (UOT) is proposed by modifying the restriction term to a penalty function term. UOT has been used in several applications like computational biology (Schiebinger et al., 2019) , machine learning (Janati et al., 2019) and deep learning (Yang and Uhler, 2019).

The UOT problem is a penalized version of Kantorovich formulation which replaced the equality constraints with penalty functions on the marginal distributions with a divergence. Many different divergences have been taken into consideration for UOT problems like KL divergence, L_1 norm, and L_2 norm. When it comes to the solving method, KL penalty with the entropy form can be solved by the Sinkhorn algorithm. It provides the UOT computation with scalability and differentiability but suffers from a larger error of KL divergence and lack of sparsity in solution compared with other regularizers (Blondel et al., 2018). However, L_2-norm reguralization could bring a sparse solution, which attracted the attention of researchers and many new algorithms are developed for it (Chapel et al., 2021; Nguyen et al., 2022) HK comment: As my comments below, the following descriptions should be located below. At the same time, The link between the UOT problem with many other well-known problems such as non-negative matrix decomposition and Lasso problem has been discovered, which encourages researchers to improve it by using the rich results in these fields.

HK comment: (We mentioned, so far, OT and UTO. Now, we raise a computational problem of OT/UOT). Then, focusing on the Lasso-type reformulation of OT/UOT proposed in the literature, we newly propose a screening technique designated to UOT. But,

Preliminary work. Under review by AISTATS 2022. Do not distribute.

this extension is NOT trivial because XXXXXX (difference and difficulty). For this we particularly propose a XXXX screening technique for UOT. For this, the detailed explanation of Screening are not needed. Some paper should be cited, and some basic concept should be provided briefly.)

The OT and UOT problems produce extremely sparse solutions due to the effectiveness of their optimal transport cost, which is a similar operator to the Lasso problem. We believe that it indicates the potential effectiveness of applying screening technical in the Lasso problem to the UOT problem. Furthermore, Different from the Lasso problem which has a dense constraints matrix, the UOT problem's constraint matrix is extremely sparse and has a unique transport matrix structure, which would benefit the design of screening and the outcome.

Contributions (should be strengthen).:

- We systematically provide the newest framework for the Screening method on the UOT problem. Considering the sparse and specific structure of the UOT problem, we design a new projection method for UOT screening, which hugely improves the screening performance over the general Lasso method.
- We propose a two-plane screening method for UOT problems, which benefits from UOT's sparse constraints and outperforms the ordinary methods adding only a negligible amount of computation

The paper is organized as follows. **Section 2** presents preliminary descriptions of optimal transport and unbalanced optimal transport. The screening methods in Lasso-like problems are explained by addressing a dynamic screening framework. In **Section 3**, our proposed screening method for the UOT problem is detailed. **Section 4** shows numerical experiments.

2 PRELIMINARIES

$\mathbb{R}^n$ denotes n-dimensional Euclidean space, and $\mathbb{R}^n_+$ denotes the set of vectors in which all elements are non-negative. $\mathbb{R}^{m\times n}$ represents the set of $m\times n$ matrices. Also, $\mathbb{R}^{m\times n}_+$ stands for the set of $m\times n$ matrices in which all elements are non-negative. We present vectors as bold lower-case letters $\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\dots$ and matrices as bold-face upper-case letters $\mathbf{A},\mathbf{B},\mathbf{C},\dots$. The i-th element of $\boldsymbol{a}$ and the element at the (i,j) position of $\mathbf{A}$ are represented respectively as $\boldsymbol{a}_i$ and $\mathbf{A}_{i,j}$. In addition, $\mathbb{1}_n \in \mathbb{R}^n$ is the n-dimensional vector in which all the elements are one. For $\boldsymbol{x}$ and $\boldsymbol{y}$ of the same size, $\langle \boldsymbol{x},\boldsymbol{y}\rangle = \boldsymbol{x}^T\boldsymbol{y}$ is the Euclidean dot-product between vectors. For two matrices of the same size $\mathbf{A}$ and $\mathbf{B}$, $\langle \mathbf{A},\mathbf{B}\rangle = \mathrm{tr}(\mathbf{A}^T\mathbf{B})$ is the Frobenius dot-product. For a vector $\boldsymbol{x}$, the i-th element of $\exp(\boldsymbol{x})$ and $\log(\boldsymbol{x})$ respectively represent $\exp(\boldsymbol{x}_i)$ and $\log(\boldsymbol{x}_i)$. $\mathrm{KL}(\boldsymbol{x},\boldsymbol{y})$ stands for the KL divergence between $\boldsymbol{x}\in\mathbb{R}^n_+$ and $\boldsymbol{y}\in\mathbb{R}^n_+$, which is defined as $\sum_i \boldsymbol{x}_i \log(\boldsymbol{x}_i/\boldsymbol{y}_i) - \boldsymbol{x}_i + \boldsymbol{y}_i$. D_h is the Bregman divergence with the strictly convex and differentiable function h, i.e., $D_h(\boldsymbol{a},\boldsymbol{b}) = \sum_i d_h(a_i,b_i) = \sum_i [h(a_i) - h(b_i) - h'(a_i)(a_i - b_i)]$. In addition, a vectorization operator for $\mathbf{A}\in\mathbb{R}^{m\times n}$ into $\boldsymbol{a}\in\mathbb{R}^{mn}$ is defined as $\boldsymbol{a} = \mathrm{vec}(\mathbf{A}) = [\mathbf{A}_{1,1},\mathbf{A}_{1,2},\cdots,\mathbf{A}_{m,n-1},\mathbf{A}_{m,n}]$, which concatenates the row vectors of a matrix. (check it).

2.1 Optimal Transport and Unbalanced Optimal Transport

Optimal Transport (OT): Given two histograms $\boldsymbol{a}\in\mathbb{R}^m, \boldsymbol{b}\in\mathbb{R}^n$, For a cost matrix $\mathbf{C}\in{\mathbb{R}_+}^{m\times n}$, mordern Optimal transport problem is trying to get a corresponding transport matrix $\mathbf{T}\in\mathbb{R}^{m\times n}_+$ that minimize the whole transport cost, which could be formulated as:

$$\begin{aligned} \mathrm{OT}(\boldsymbol{a},\boldsymbol{b}) \quad &:= \min_{\mathbf{T}\in\mathbb{R}^{m\times n}_+} \langle \mathbf{C},\mathbf{T}\rangle \\ \text{subject to} \quad & \mathbf{T}\mathbb{1}_n = \boldsymbol{a}, \mathbf{T}^T\mathbb{1}_m = \boldsymbol{b}. \end{aligned} \tag{1}$$

Defining $\boldsymbol{t} = \mathrm{vec}(\mathbf{T})\in\mathbb{R}^{mn}$ and $\boldsymbol{c} = \mathrm{vec}(\mathbf{C})\in\mathbb{R}^{mn}$, we reformulate Eq.(1) in a vector format as

$$\begin{aligned} \mathrm{OT}(\boldsymbol{a},\boldsymbol{b}) \quad &:= \min_{\boldsymbol{t}\in\mathbb{R}^{mn}_+} \boldsymbol{c}^T\boldsymbol{t} \\ \text{subject to} \quad & \mathbf{N}\boldsymbol{t} = \boldsymbol{a}, \mathbf{M}\boldsymbol{t} = \boldsymbol{b}, \end{aligned} \tag{2}$$

where $\mathbf{N}\in\mathbb{R}^{m\times mn}$ and $\mathbf{M}\in\mathbb{R}^{n\times mn}$ are two matrices composed of "0" and "1". $\mathbf{N}$ and $\mathbf{M}$ in case of $m = n = 3$ are given, respectively, as

$$\mathbf{N} = \begin{pmatrix} 1&1&1&0&0&0&0&0&0\\ 0&0&0&1&1&1&0&0&0\\ 0&0&0&0&0&0&1&1&1 \end{pmatrix},$$
$$\mathbf{M} = \begin{pmatrix} 1&0&0&1&0&0&1&0&0\\ 0&1&0&0&1&0&0&1&0\\ 0&0&1&0&0&1&0&0&1 \end{pmatrix}.$$

Unbalanced Optimal Transport (UOT): The UOT problem is a penalized version of Kantorovich formulation which replaced the equality constraints with penalty functions on the marginal distributions with a divergence. Although $\|\boldsymbol{a}\|_2 = \|\boldsymbol{b}\|_2$ holds in the OT problem, the solution $\hat{\boldsymbol{t}}$ does not exist when $\|\boldsymbol{a}\|_2 \neq \|\boldsymbol{b}\|_2$. However, allowing $\|\boldsymbol{a}\|_2 \neq \|\boldsymbol{b}\|_2$, we derive the UOT problem as follows. Defining $\boldsymbol{y} = [\boldsymbol{a},\boldsymbol{b}]^T$

and $\mathbf{X} = [\mathbf{M}^T, \mathbf{N}^T]^T$, the UOT problem can be defined introducing a penalty function for the histograms:

$$\mathrm{UOT}(\boldsymbol{a}, \boldsymbol{b}) := \min_{\boldsymbol{t} \in \mathbb{R}_+^{mn}} \boldsymbol{c}^T \boldsymbol{t} + D_h(\mathbf{X}\boldsymbol{t}, \boldsymbol{y}), \tag{3}$$

where D_h is the Bregman divergence.

Relationship with Lasso: The lasso-like problem has a general formula:

$$f(\boldsymbol{t}) \quad := \quad g(\boldsymbol{t}) + D_h(\mathbf{X}\boldsymbol{t}, \boldsymbol{y}). \tag{4}$$

When $g(\boldsymbol{t}) = \lambda\|\boldsymbol{t}\|_1$ with $\lambda > 0$ and $D_h(\mathbf{X}\boldsymbol{t}, \boldsymbol{y}) = \|\mathbf{X}\boldsymbol{t} - \boldsymbol{y}\|_2^2$, this problem is reduced to the L_2-regunarlized regression Lasso problem. It should be noted that $\mathbf{X}$ in UOT is different from the one in the Lasso problem. More concretely, the former $\mathbf{X}$ has a specific structure and has only two non-zero elements and is equal to 1 whereas the latter $\mathbf{X}$ in Lasso problem is irregular and dense.

2.2 Dynamic Screening Framework

Screening is a well-known technique proposed by (Ghaoui et al., 2010) in the field of lasso problems, where the L_1 regularizer leads to a sparse solution for the problem. It can pre-select solutions that must be zero theoretically and freeze them before computation. The solutions to many large-scale optimization problems are sparse, and a large amount of computation is wasted on updating the zero elements. With the Safe Screening method, we can identify and freeze the elements that are zero with linear complexity computation before starting the algorithm, thus saving optimization time. the Screening method get attention in recent years and has been improved a lot, New methods such as Dynamic Screening (Bonnefoy et al., 2015), Gap screening method (Ndiaye et al., 2017) and Dynamic Sasvi (Yamada and Yamada, 2021).

Hereinafter, we briefly detail the framework proposed in (Yamada and Yamada, 2021) to introduce the whole dynamic screening technique for the Lasso-like problem:

$$\min_{\boldsymbol{t}} \{f(\boldsymbol{t}) := g(\boldsymbol{t}) + d(\mathbf{X}\boldsymbol{t})\}. \tag{5}$$

The Fenchel-Rockafellar Duality yields the dual problem as presented below:

Theorem 1 (Fenchel-Rockafellar Duality)**.** *If d and g are proper convex functions on $\mathbb{R}^{m+n}$ and $\mathbb{R}^{mn}$. Then we have the following:*

$$\min_{\boldsymbol{t}} g(\boldsymbol{t}) + d(\mathbf{X}\boldsymbol{t}) = \max_{\boldsymbol{\theta}} -d^*(-\boldsymbol{\theta}) - g^*(\mathbf{X}^T\boldsymbol{\theta})$$

Because the primal function d is always convex, the dual function d^* is concave. Assuming d^* is an L-strongly concave problem, we can design an area for any feasible $\tilde{\boldsymbol{\theta}}$ by the strongly concave property:

Theorem 2 (L-strongly concave (Yamada and Yamada, 2021, Theorem 5))**.** *Considering problem in Eq.(5), if function d and g are both convex, for $\forall\ \tilde{\boldsymbol{\theta}} \in R^{m+n}$ and satisfied the constraints on the dual problem, we have the following area constructed by its L-strongly concave property:*

$$\mathcal{R}^C := \boldsymbol{\theta} \in \left\{ \frac{L}{2}\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2^2 + d^*(-\tilde{\boldsymbol{\theta}}) \leq d^*(-\boldsymbol{\theta}) \right\}.$$

We know that the optimal solution for the dual problem $\hat{\boldsymbol{\theta}}$ satisfied the inequality, so the set is not empty.

3 PROPOSED UOT SCREENING

This section proposes a dynamic screening method designated for the UOT problem. For this purpose, we first derive the dual formulation of the Lasso-like formulated UOT problem in Eq.(3). Concrete proofs of lemma and theorems are provided in the supplementary file.

3.1 Dual formulation of UOT

We can get the dual form of the UOT problem: For $d(\mathbf{X}\boldsymbol{t}) = \frac{1}{2}\|\mathbf{X}\boldsymbol{t} - \boldsymbol{y}\|_2^2$, the dual Lasso problem gives $d^*(-\boldsymbol{\theta})$ as

$$d^*(-\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}\|_2^2 - \boldsymbol{y}^T\boldsymbol{\theta}, \tag{6}$$

and $g^*(\mathbf{X}^T\boldsymbol{\theta})$ is given as

$$g^*(\mathbf{X}^T\boldsymbol{\theta}) = \begin{cases} 0 & (\forall \boldsymbol{t} \quad \boldsymbol{\theta}^T\mathbf{X}\boldsymbol{t} - g(\boldsymbol{t}) \leq 0) \\ \infty & (\exists t \quad \boldsymbol{\theta}^T\mathbf{X}\boldsymbol{t} - g(\boldsymbol{t}) > 0). \end{cases}$$

For UOT problem in Eq.(3), we obtain its dual form from **Theorem 1** as

Lemma 3 (Dual form of UOT problem)**.** *For UOT problem in Eq.(3), we have the following dual form:*

$$\begin{aligned} &-d^*(-\boldsymbol{\theta}) - g^*(\mathbf{X}^T\boldsymbol{\theta}) = -\frac{1}{2}\|\boldsymbol{\theta}\|_2^2 - \boldsymbol{y}^T\boldsymbol{\theta} \\ &s.t. \qquad \boldsymbol{x}_p^T\boldsymbol{\theta} - \lambda\boldsymbol{c}_p \leq 0, \forall p, \end{aligned} \tag{7}$$

where $\boldsymbol{x}_p$ corresponds to the p-th column of $\mathbf{X}$.

It is clear that the strongly concave coefficient L for the dual function d is 1. These in-equations in Eq.(7) make up a dual feasible area written as $\mathcal{R}^D$, and the optimal solution satisfied them.
From the KKT condition, we know that for the optimal primal solution $\hat{\boldsymbol{t}}$:

Theorem 4 (KKT condition)**.** *For the dual optimal solution $\hat{\boldsymbol{\theta}}$, we have the following relationship:*

$$\boldsymbol{x}_p^T\hat{\boldsymbol{\theta}} - \lambda\boldsymbol{c}_p \begin{cases} < 0 & \Rightarrow \hat{\boldsymbol{t}}_p = 0 \\ = 0 & \Rightarrow \hat{\boldsymbol{t}}_p \geq 0. \end{cases} \tag{8}$$

3.2 Detailed Screening Method

Eq.(8) indicates a potential method to screening the primal variable. Since we do not know the information of $\hat{\boldsymbol{t}}$ directly, we construct an area $\mathcal{R}^S$ containing the $\hat{\boldsymbol{t}}$, if

$$\max_{\boldsymbol{t}\in\mathcal{R}^S} \boldsymbol{x}_p^T\boldsymbol{\theta} - \lambda\boldsymbol{c}_p < 0. \tag{9}$$

Then we have

$$\boldsymbol{x}_p^T\hat{\boldsymbol{\theta}} - \lambda\boldsymbol{c}_p < 0, \tag{10}$$

which means the corresponding $\hat{\boldsymbol{t}}_p = 0$, and can be screened out. Noteworthy point is that, for the UOT problem, $\boldsymbol{x}_p = [..., 0, 1, 0, ..., 0, 1, 0, ...,]^T$ has only two nonzero elements, p_1 and p_2, that are equal to 1. Therefore, we can set $\boldsymbol{\theta} = [\boldsymbol{u}^T, \boldsymbol{v}^T]^T$ and $\boldsymbol{u} \in \mathbb{R}^m, \boldsymbol{v} \in \mathbb{R}^n$, assuming $p = (I, J)$ where $I = p \mid m, J = p \mod m$. Subsequently, we rewrite Eq.(10) as

$$\boldsymbol{u}_I + \boldsymbol{v}_J - \lambda\boldsymbol{c}_p < 0. \tag{11}$$

Before we start to construct the area containing $\hat{\boldsymbol{\theta}}$ from **Theorem 2**, we first need to find $\tilde{\boldsymbol{\theta}}$ in the dual feasible area $\mathcal{R}^D$. Although there exists a relationship between the primal variable and dual variable $\boldsymbol{\theta} = \boldsymbol{y} - \mathbf{X}\boldsymbol{t}$, we sometimes get $\boldsymbol{\theta} \notin \mathcal{R}^D$. This requires us to project $\boldsymbol{\theta}$ onto $\mathcal{R}^D$. In the Lasso problem, as the constraints limit the $\|\boldsymbol{x}_p\boldsymbol{\theta}\|_1$, and every element of $\boldsymbol{\theta}$ is multiplied by a dense x_i, one has to use a shrinking method to obtain a $\tilde{\boldsymbol{\theta}} \in \mathcal{R}^D$ for further constructing the dual screening area ?:

$$\tilde{\boldsymbol{\theta}} = \frac{\lambda\boldsymbol{c}^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{t})}{\max(\lambda\boldsymbol{c}, \|\mathbf{X}^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{t})\|_\infty)}$$

This projection pushes the $\boldsymbol{\theta}$ far away from the optimum $\hat{\boldsymbol{\theta}}$, and cannot work well when one of the costs is $\boldsymbol{c}_p = 0$. This situation fortunately never happens in the Lasso problem because xxxx, however, frequently happens in the UOT problem. Under this situation, the whole dual elements would degenerate to zero and disable the screening process. Therefore, noting that, fo the UOT problem, it only allows $\boldsymbol{t}_p \geq 0$, and the $\boldsymbol{x}_p$ only consists of two non-zero elements, we can adapt a better projection method. The following theorem states this:

Theorem 5 (UOT shifting projection)**.** *For any* $\boldsymbol{\theta} = [\boldsymbol{u}^T, \boldsymbol{v}^T]^T$*, we compute the projection* $\tilde{\boldsymbol{\theta}} = [\tilde{\boldsymbol{u}}^T, \tilde{\boldsymbol{v}}^T]^T \in \mathcal{R}^D$ *as*

$$\begin{aligned}\tilde{\boldsymbol{u}}_I &= \boldsymbol{u}_I - \max_{0\leq j\leq n} \frac{\boldsymbol{u}_I + \boldsymbol{v}_j - \lambda\boldsymbol{c}_p}{2} \\ &= \frac{\boldsymbol{u}_I + \lambda\boldsymbol{c}_p}{2} - \frac{1}{2}\max_{0\leq j\leq n} \boldsymbol{v}_j \end{aligned} \tag{12}$$

$$\begin{aligned}\tilde{\boldsymbol{v}}_J &= \boldsymbol{v}_J - \max_{0\leq i\leq m} \frac{\boldsymbol{u}_i + \boldsymbol{v}_J - \lambda\boldsymbol{c}_p}{2} \\ &= \frac{\boldsymbol{v}_J + \lambda\boldsymbol{c}_p}{2} - \frac{1}{2}\max_{0\leq i\leq m} \boldsymbol{u}_j \end{aligned} \tag{13}$$

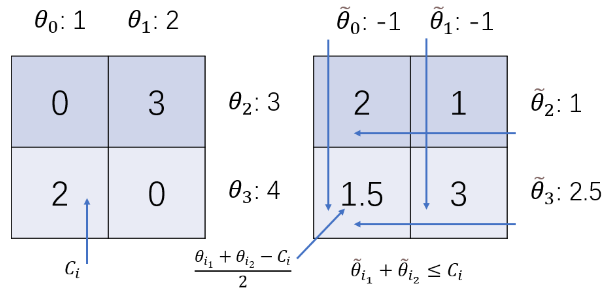

Figure 1: Shifting on a 2×2 matrix

Since we now obtain $\tilde{\boldsymbol{\theta}}$ in $\mathcal{R}^D$, we additionally consider $\mathcal{R}^C$ in **Theorem 2** to satisfy $\hat{\boldsymbol{t}} \in \mathcal{R}^C \cap \mathcal{R}^D$. However, the intersection of a sphere and a polytope cannot be computed (????) in $O(knm)$, where k is a constant. Therefore, we propose a relaxation method, denoted as *two plane screening*. This divides the constraints into two parts, then we maximize the intersection of two hyperplanes and a hyper-ball. The next theorem states this:

Theorem 6 (Two plane screening for UOT)**.** *For every single primal variable* t_p*, let* $A_p = \{i \| 0 \leq i < nm, i \mid m = I \vee i \mod m = J\}$*,* $B_p = \{i \| 0 \leq i < nm, i \notin A_p\}$*. Then, we construct a specific area* $\mathcal{R}_{IJ}^S$ *as presented below:*

$$\mathcal{R}_{IJ}^S = \left\{ \boldsymbol{\theta} \left\| \begin{array}{l} \sum\limits_{l\in A_p} (\boldsymbol{\theta}^T\boldsymbol{x}_l\boldsymbol{t}_l - \lambda\boldsymbol{c}_l\boldsymbol{t}) \leq 0, \\ \sum\limits_{l\in B_p} (\boldsymbol{\theta}^T\boldsymbol{x}_l\boldsymbol{t}_l - \lambda\boldsymbol{c}_l\boldsymbol{t}) \leq 0, \\ (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T(\boldsymbol{\theta} - \boldsymbol{y}) \leq 0 \ (\textit{where from??}) \end{array} \right. \right\}. \tag{14}$$

How does $\mathcal{R}_{IJ}^S$ play in the following??? Is in Algorithm 1 only?

We divide the constraints into two groups A_p and B_p for every single p, this problem can be solved easily by the Lagrangian method in constant time, the computational process is in Appendix. A

3.3 Screening Algorithms

(Insert equation numbers in appropriate lines in **Algorithm ??**

It should be emphasized that the proposed screening method is *independent* of the optimization solver that you choose. We give the specific algorithm for L_2 UOT problem to show the whole optimization process as in **Algorithm 1**. The update($\boldsymbol{t}$) operator in the algorithm indicates the updating process for $\boldsymbol{t}$ according to the adopted optimizer.

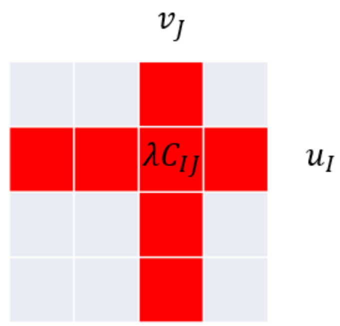

Figure 2: Selection of group A_{IJ}(red) and B_{IJ}(grey)

Algorithm 1 UOT Dynamic Screening Algorithm

Input: $\boldsymbol{t}_0, S \in R^{n\times m}, S_{ij} = 1, (i,j) = mi + j$
Output: S
  Choose a solver for the problem.
  for $k = 0$ to K **do**
    Projection $\tilde{\boldsymbol{\theta}} = \text{Proj}(\boldsymbol{t}^k)$
    for $i = 0$ to m **do**
      for $j = 0$ to n **do**
        $\mathcal{R}^S \leftarrow \mathcal{R}^S_{ij}(\tilde{\boldsymbol{\theta}}, \boldsymbol{t}^k)$
        $S \leftarrow S_{ij} = 0$ if $\max_{\boldsymbol{\theta}\in\mathcal{R}^S} x_{(i,j)}{}^T\boldsymbol{\theta} < \lambda c_{(i,j)}$
      end for
    end for
    for $(i,j) \in \{(i,j) \| S_{ij} = 0\}$ **do**
      $\boldsymbol{t}^k_{(i,j)} \leftarrow 0$
    end for
    $\boldsymbol{t}^{k+1} = \text{update}(\boldsymbol{t}^k)$
  end for
  return $\boldsymbol{t}^{K+1}, S$

3.4 Computational Cost Analysis

XXXXXX

4 EXPERIMENTS

In this section, we show the efficacy of the proposed methods using toy Gaussian models and the MNIST dataset.

4.1 Projection Method

To prove the effectiveness of our projection method compared with the traditional projection method in the Lasso problem, we compared the projection distance and screening ratio with randomly generated Gaussian measures by two projection methods. We set the $\lambda = \frac{\|\mathbf{X}^T y\|}{100}$ and test for 10 different pairs.We choose the FISTA for solving the L_2 penalized UOT problems. Our projection method has only moved the dual point by a very small order of magnitude. It ensures that the points are kept at a smaller distance from the optimal solution and cause better screening effects.

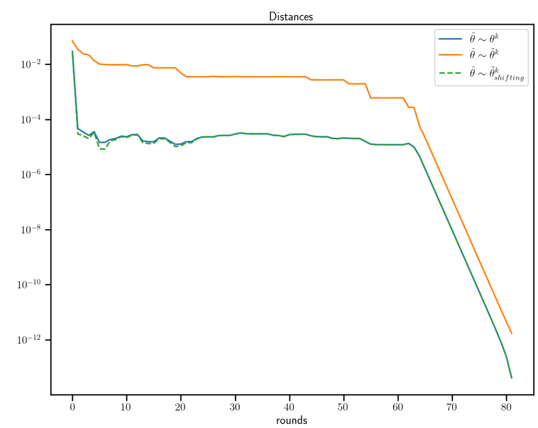

Figure 3: Distance of different projection method

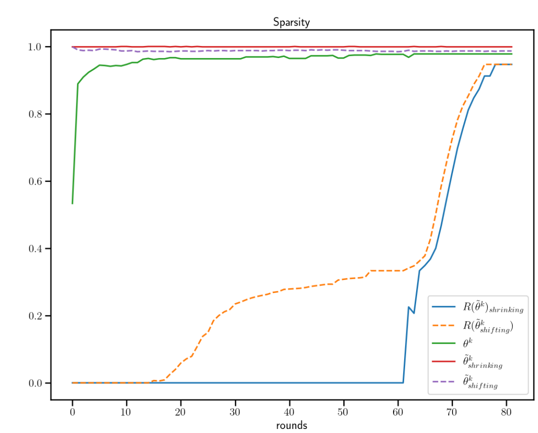

Figure 4: Screening ratio of different projection method

4.2 Divide Method

We compared the screening ratio with three different methods, including our Divide method, Dynamic Sasvi method, and Gap method. Every method would use our projection method to get a better outcome, which also makes sure the difference in performance is only in the construction of the feasible domain.

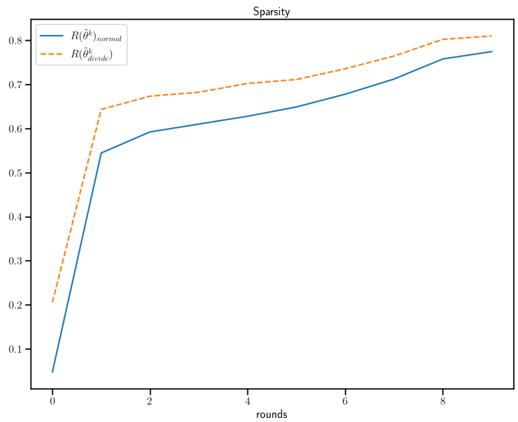

Figure 5: Screening ratio of dividing method

4.3 Best Divide Method

We compared the screening ratio with three different methods, including our Divide method, the Dynamic Sasvi method, and a random divide method.

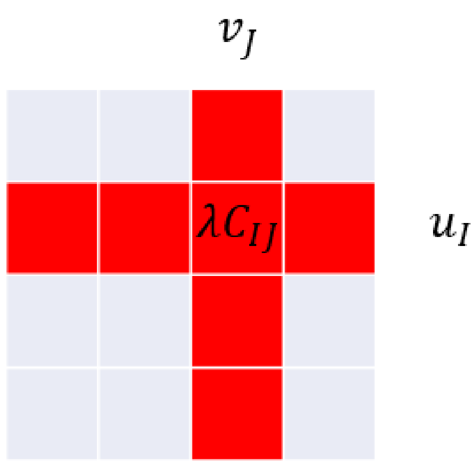

Figure 6: Comparing of our separation method with random separation method

4.4 Speed up Ratio

We choose the FISTA method, Newton method, and Language method to test the screening ratio.

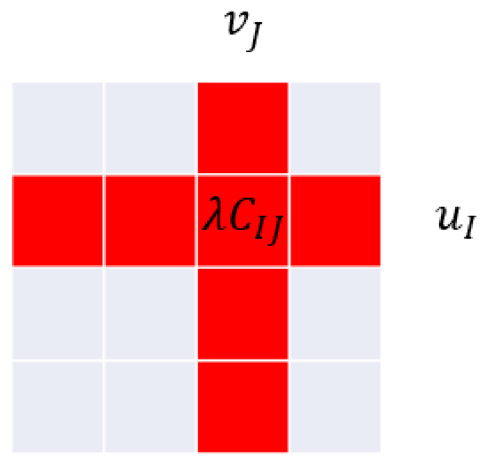

Figure 7: speed up ratio for different solver

5 CONCLUSION

Our algorithm is great, we are going to apply the method onto Sinkhorn

References

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR.

Blondel, M., Seguy, V., and Rolet, A. (2018). Smooth and sparse optimal transport. In Storkey, A. J. and Pérez-Cruz, F., editors, *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, volume 84 of *Proceedings of Machine Learning Research*, pages 880–889. PMLR.

Bonnefoy, A., Emiya, V., Ralaivola, L., and Gribonval, R. (2015). Dynamic screening: Accelerating first-order algorithms for the lasso and group-lasso. *IEEE Transactions on Signal Processing*, 63(19):5121–5132.

Chapel, L., Flamary, R., Wu, H., Févotte, C., and Gasso, G. (2021). Unbalanced optimal transport through non-negative penalized linear regression. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 23270–23282. Curran Associates, Inc.

Chen, L., Zhang, Y., Zhang, R., Tao, C., Gan, Z., Zhang, H., Li, B., Shen, D., Chen, C., and Carin, L. (2019). Improving sequence-to-sequence learning via optimal transport. In *7th International Conference on Learning Representations, ICLR 2019*. International Conference on Learning Representations, ICLR. Generated from Scopus record by KAUST IRTS on 2021-02-09.

Courty, N. (2017). Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Ghaoui, L. E., Viallon, V., and Rabbani, T. (2010). Safe feature elimination for the lasso and sparse supervised learning problems. *arXiv preprint arXiv:1009.4219*.

Janati, H., Cuturi, M., and Gramfort, A. (2019). Wasserstein regularization for sparse multi-task regression. In Chaudhuri, K. and Sugiyama, M., editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 1407–1416. PMLR.

Ndiaye, E., Fercoq, O., Alex, re Gramfort, and Salmon, J. (2017). Gap safe screening rules for sparsity enforcing penalties. *Journal of Machine Learning Research*, 18(128):1–33.

Nguyen, Q. M., Nguyen, H. H., Zhou, Y., and Nguyen, L. M. (2022). On unbalanced optimal transport: Gradient methods, sparsity and approximation error.

Petric Maretic, H., El Gheche, M., Chierchia, G., and Frossard, P. (2019). Got: An optimal transport framework for graph comparison. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., Lee, L., Chen, J., Brumbaugh, J., Rigollet, P., Hochedlinger, K., Jaenisch, R., Regev, A., and Lander, E. S. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943.e22.

Yamada, H. and Yamada, M. (2021). Dynamic sasvi: Strong safe screening for norm-regularized least squares. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 14645–14655. Curran Associates, Inc.

Yang, K. D. and Uhler, C. (2019). Scalable unbalanced optimal transport using generative adversarial networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Supplementary Material: Dynamic Screening for Unbalanced Optimal Transport Problem

A NOTATIONS

$$M = \begin{pmatrix} 1 & & & & & \cdots\cdots & 1 & & & & \\ & 1 & & & & \cdots\cdots & & 1 & & & \\ & & \ddots & & & \ddots\ddots & & & \ddots & & \\ & & & 1 & & & & & & 1 & \\ & & & & 1 & \cdots\cdots & & & & & 1 \end{pmatrix} \tag{15}$$

$$N = \begin{pmatrix} 1 & 1 & \dots & 1 & 1 & \cdots\cdots & & & & & \\ & & & & & \ddots\ddots & & & & & \\ & & & & & \cdots\cdots & 1 & 1 & \dots & 1 & 1 \end{pmatrix} \tag{16}$$

B PROOFS

B.1 Proof of Theorem 5

For any $p \in 0, 1, ..., nm-1$ we assume that $p = (I, J)$, then we can compute that:

$$\begin{aligned} \boldsymbol{x}_p^T \tilde{\boldsymbol{\theta}} &= \tilde{\boldsymbol{u}}_I + \tilde{\boldsymbol{v}}_J \\ &= \boldsymbol{u}_I + \boldsymbol{v}_J - \max_{0 \geq j \geq n} \frac{\boldsymbol{u}_I + \boldsymbol{v}_j - \lambda \boldsymbol{c}_p}{2} - \max_{0 \geq i \geq m} \frac{\boldsymbol{u}_i + \boldsymbol{v}_J - \lambda \boldsymbol{c}_p}{2} \\ &= \frac{\boldsymbol{u}_I + \boldsymbol{v}_J}{2} - \max_{0 \geq j \geq n} \frac{\boldsymbol{v}_j}{2} - \max_{0 \geq i \geq m} \frac{\boldsymbol{u}_i}{2} + \lambda \boldsymbol{c}_p \\ &= \frac{1}{2} \boldsymbol{x}_p^T \boldsymbol{\theta} - \max_{0 \geq j \geq n} \frac{\boldsymbol{v}_j}{2} - \max_{0 \geq i \geq m} \frac{\boldsymbol{u}_i}{2} + \lambda \boldsymbol{c}_p \\ &\leq \lambda \boldsymbol{c}_p \end{aligned} \tag{17}$$

For $\forall p$, we have $\tilde{\boldsymbol{\theta}} \in \mathcal{R}^D$

B.2 Proof of Theorem 6

We Generalize the problem as

$$\max_{\boldsymbol{\theta} \in \mathcal{R}_I^S} \boldsymbol{\theta}_{I_1} + \boldsymbol{\theta}_{I_2} \tag{18}$$

Considering the center of the circle as $\boldsymbol{\theta}^o$, we define $\boldsymbol{\theta} = \boldsymbol{\theta}^o + q$, as $\boldsymbol{\theta}^o_{I_1} + \boldsymbol{\theta}^o_{I_2}$ is a constant, the problem is equal to $\min_{\boldsymbol{\theta} \in \mathcal{R}_I^S} -(\boldsymbol{q}_{I_1} + \boldsymbol{q}_{I_2})$, we compute the Lagrangrian function of later:

$$\min_{\boldsymbol{q}} \max_{\eta, \mu, \nu \geq 0} L(\boldsymbol{q}, \eta, \mu, \nu) = \min_{\boldsymbol{q}} \max_{\eta, \mu, \nu \geq 0} -\boldsymbol{q}_{I_1} - \boldsymbol{q}_{I_2} + \eta(\boldsymbol{q}^T \boldsymbol{q} - r^2) + \mu(a^T \boldsymbol{q} - e_a) + \nu(b^T \boldsymbol{q} - e_b) \tag{19}$$

$$\frac{\partial L}{\partial \boldsymbol{q}_i} = \begin{cases} -1 + 2\eta \boldsymbol{q}_i + \mu a_i + \nu b_i & i = I_1, I_2 \\ 2\eta \boldsymbol{q}_i + \mu a_i + \nu b_i & i \neq I_1, I_2 \end{cases} \tag{20}$$

$$\boldsymbol{q}_i{}^* = \begin{cases} \dfrac{1-\mu a_i - \nu b_i}{2\eta} & i = I_1, I_2 \\ -\dfrac{\mu a_i + \nu b_i}{2\eta} & i \neq I_1, I_2 \end{cases} \tag{21}$$

We can get the Lagrangian dual problem:

$$\max_{\eta,\mu,\nu\geq 0} L(\eta,\mu,\nu) = \max_{\eta,\mu,\nu\geq 0} \frac{\mu a_{I_1} + \nu b_{I_1} - 1}{2\eta} + \frac{\mu a_{I_2} + \nu b_{I_2} - 1}{2\eta} + \eta(\boldsymbol{q}^{*T}\boldsymbol{q}^* - r^2) + \mu(a^T\boldsymbol{q}^* - e_a) + \nu(b^T\boldsymbol{q}^* - e_b) \tag{22}$$

From the KKT optimum condition, we know that if

$$\begin{aligned} \eta(\boldsymbol{q}^{*T}\boldsymbol{q}^* - r^2) &= 0 \\ \mu(a^T\boldsymbol{q}^* - e_a) &= 0 \\ \nu(b^T\boldsymbol{q}^* - e_b) &= 0 \end{aligned} \tag{23}$$

We set η^*, μ^*, ν^* as the solution of the equations, which is also the solution of the dual problem. Firstly, we assume that $\eta^*, \mu^*, \nu^* \neq 0$, then the solution is equal to compute the following equations:

$$\begin{aligned} &(1-\mu a_{I_1} - \nu b_{I_1})^2 + (1-\mu a_{I_2} - \nu b_{I_2})^2 + \sum_{i\neq I_1,I_2}^{m+n} (a_i\mu + b_i\nu)^2 - 4\eta^2 r^2 = 0 \\ &a_{I_1} - \mu a_{I_1}^2 - \nu b_{I_1} a_{I_1} + a_{I_2} - \mu a_{I_2}^2 - \nu b_{I_2} a_{I_2} - \sum_{i\neq I_1,I_2}^{m} (a_i^2\mu + b_i a_i \nu) - 2\eta e_a = 0 \\ &b_{I_1} - \nu b_{I_1}^2 - \mu b_{I_1} a_{I_1} + b_{I_2} - \nu b_{I_2}^2 - \mu b_{I_2} a_{I_2} - \sum_{i\neq I_1,I_2}^{m} (b_i^2\nu + b_i a_i \mu) - 2\eta e_b = 0 \end{aligned} \tag{24}$$

Rearranged as:

$$\begin{aligned} &2 - 2\mu(a_{I_1} + a_{I_2}) - 2\nu(b_{I_1} + b_{I_2}) + \|a\|^2\mu^2 + \|b\|^2\nu^2 + 2\mu\nu a^T b - 4\eta^2 r^2 = 0 \\ &(a_{I_1} + a_{I_1}) - \|a\|^2\mu - a^T b\nu - 2\eta e_a = 0 \\ &(b_{I_1} + b_{I_2}) - \|b\|^2\nu - a^T b\mu - 2\eta e_b = 0 \end{aligned} \tag{25}$$

we have

$$\begin{aligned} \mu &= \frac{2(e_b a^T b - e_a\|b\|^2)\eta + (a_{I_1} + a_{I_2})\|b\|^2 - (b_{I_1} + b_{I_2})(a^T b)}{\|a\|^2\|b\|^2 - a^T b} \\ \nu &= \frac{2(e_a a^T b - e_b\|a\|^2)\eta + (b_{I_1} + b_{I_2})\|a\|^2 - (a_{I_1} + a_{I_2})(a^T b)}{\|a\|^2\|b\|^2 - a^T b} \end{aligned} \tag{26}$$

set it as:

$$\begin{aligned} \mu &= s_1\eta + s_2 \\ \nu &= u_1\eta + u_2 \end{aligned} \tag{27}$$

Then we can solve the η as a quadratic equation:

$$\begin{aligned} &0 = a\eta^2 + b\eta + c \\ &a = 4r^2 - s_1^2\|a\|^2 - u_1^2\|b\|^2 - 2s_1u_1a^T b \\ &b = 2(a_{I_1} + a_{I_2})s_1 + 2(b_{I_1} + b_{I_2})u_1 - 2s_1s_2\|a\|^2 - 2u_1u_2\|b\|^2 - 2(s_1u_2 + s_2u_1)a^T b \\ &c = 2(a_{I_1} + a_{I_2})s_2 + 2(b_{I_1} + b_{I_2})u_2 - s_2^2\|a\|_1 - u_2^2\|b\|_1 - 2s_2u_2a^T b - 2 \end{aligned} \tag{28}$$

Then we can put it back into 27 and get μ, ν.

If the solution satisfied the constraints $\eta^*, \mu^*, \nu^* > 0$, then it is the solution. However, if one of the dual variables is less than 0, the problem would degenerate into a simpler question.

If only η^* is larger than 0, $\min_{\boldsymbol{\theta} \in \mathcal{R}_I^S} -(\boldsymbol{q}_{I_1} + \boldsymbol{q}_{I_2}) = -\sqrt{2}r$

If only μ^* or ν^* is less than 0, we are optimizing on a sphere cap, the solution can be found in (Yamada and Yamada, 2021, Appendix B)

if only $\eta^* \leq 0$: As the sphere is inactivated, the problem gets maximum at every point of the intersection of two planes.

$$\min_{\boldsymbol{q}} \max_{\mu,\nu \geq 0} L(\boldsymbol{q}, \mu, \nu) = \min_{\boldsymbol{q}} \max_{\mu,\nu \geq 0} -\boldsymbol{q}_{I_1} - \boldsymbol{q}_{I_2} + \mu(a^T \boldsymbol{q} - e_a) + \nu(b^T \boldsymbol{q} - e_b) \tag{29}$$

To have a solution, the equations satisfied

$$\frac{\partial L}{\partial q} = \begin{cases} -1 + \mu a_i + \nu b_i = 0 \quad i = I_1, I_2 \\ \quad -\mu a_i - \nu b_i \quad = 0 i \neq I_1, I_2 \end{cases} \tag{30}$$

As the equation satisfied, we can just set $\boldsymbol{q}_i^* = 0, i \neq I_1, I_2$, then we compute the

$$\min_{\boldsymbol{\theta} \in \mathcal{R}_I^S} -(\boldsymbol{q}_{I_1} + \boldsymbol{q}_{I_2}) = \frac{a_{I_2} e_b - b_{I_2} e_a - a_{I_1} e_b + b_{I_1} e_a}{a_{I_1} b_{I_2} - a_{I_2} b_{I_1}} \tag{31}$$