Dynamic Screening for Unbalanced Optimal Transport Problem

Anonymous Author
Anonymous Institution

Abstract

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 proposed a new projection method and a feasible area construction method for Screening on the UOT problem 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.

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 improves the computational speed of the OT problem from $\Theta(n^3)$ of Simplex's method to $\Theta(n^2)$. In order to extend the optimal transmission 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 regularized 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 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) 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.

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 Dy-

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

namic Sasvi (Yamada and Yamada, 2021)

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.

Contribution:

- 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

1 BACKGROUND

1.1 Optimal Transport and Unbalanced Optimal Transport

Given two histograms $\mathbf{a} \in \mathbb{R}^m, \mathbf{b} \in \mathbb{R}^n$, 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:

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

We can write it into a vector type, set $\mathbf{c}, \mathbf{t} \in \mathbb{R}^{mn}$:

$$\begin{aligned} \mathrm{OT}(\mathbf{a}, \mathbf{b}) := & \min_{t \in \mathbb{R}_+^{n^2}} c^\mathsf{T} \mathbf{t} \\ & \mathbf{N}\mathbf{t} = \alpha, \mathbf{M}\mathbf{t} = \beta \end{aligned} \tag{2}$$

$\mathbf{N} \in \mathbb{R}^{m \times mn}, \mathbf{N} \in \mathbb{R}^{n \times mn}$ are two matrix consisted with 0 and 1, listed in Appendix.A. When the $\|\mathbf{a}\|_2 = \|\mathbf{b}\|_2$, it is the OT problem. When $\|\mathbf{a}\|_2 \neq \|\mathbf{b}\|_2$, the solution $\hat{\mathbf{t}}$ is not exist. We define $\mathbf{y} = [\mathbf{a}, \mathbf{b}]^\mathsf{T}$, the UOT problem uses a penalty function for the historgrams:

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

D_h is the Bregman divergence derived from the norm h, $\mathbf{X} = [\mathbf{M}^\mathsf{T} \mathbf{N}^\mathsf{T}]^\mathsf{T}$.

1.2 Relationship with Lasso

The lasso-like problem has a general formula:

$$f(\mathbf{t}) = g(\mathbf{t}) + D_h(\mathbf{X}\mathbf{t}, \mathbf{y}), t \in \mathbb{R}^{mn}$$

When $g(\mathbf{t}) = \lambda \|\mathbf{t}\|_1$ and $D_h(\mathbf{X}\mathbf{t}, \mathbf{y}) = \|\mathbf{X}\mathbf{t} - \mathbf{y}\|_2^2$, this is the L_2 regression Lasso problem. It is important to note that $\mathbf{X}$ in UOT is a bit different from the $\mathbf{X}$ in the Lasso problem, the former $\mathbf{X}$ has a specific structure and has only two non-zero elements and is equal to 1, which is quite different to the irregular and dense $\mathbf{X}$ in Lasso problem.

1.3 Dynamic Screening Framework

We follow (Yamada and Yamada, 2021)'s framework to introduce the whole dynamic screening technique for the Lasso-like problem:

$$f(\mathbf{t}) = g(\mathbf{t}) + d(\mathbf{X}\mathbf{t}) \tag{4}$$

By Frenchel-Rockafellar Duality, we get the dual problem

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

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

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 $\tilde{\theta}$ by the strongly concave property:

Theorem 2. *(L-strongly concave) Considering problem 4, if function d and g are both convex, for $\forall \tilde{\theta} \in R^{m+n}$ and satisfied the constraints on the dual problem, we have the following area constructed by its L-strongly concave property:*

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

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

2 UNBALANCED OPTIMAL TRANSPORT SCREENING

2.1 Screening for UOT

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

following form:

$$d^*(-\theta) = \frac{1}{2}\|\theta\|_2^2 - \mathbf{y}^\mathsf{T}\theta \tag{5}$$

$$g^*(\mathbf{X}^\mathsf{T}\theta) = \begin{cases} 0 & (\forall \mathbf{t} \quad \theta^\mathsf{T}\mathbf{X}\mathbf{t} - g(\mathbf{t}) \leq 0) \\ \infty & (\exists t \quad \theta^\mathsf{T}\mathbf{X}\mathbf{t} - g(\mathbf{t}) \leq 0) \end{cases} \tag{6}$$

For UOT problem 3, we could get its dual form.

Lemma 3. *(Dual form of UOT problem)*

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

$\mathbf{x}_p$ is the p-th column of $\mathbf{X}$, It is clear that the strongly concave coefficient L for the dual function d is 1. These inequations 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 $\hat{\mathbf{t}}$:

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

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

8 indicates to us a potential method to screening the primal variable, as we do not know the information of $\hat{\mathbf{t}}$ directly, we construct an area $\mathcal{R}^S$ containing the $\hat{\mathbf{t}}$, if

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

then we have:

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

which means the corresponding $\hat{t}_p = 0$, and can be screened out. As for the UOT problem, $x_p = [..., 0, 1, 0, ..., 0, 1, 0, ...,]^\mathsf{T}$, which has only two elements p_1, p_2 equal to 1, we can set $\theta = [\mathbf{u}^\mathsf{T}, \mathbf{v}^\mathsf{T}]^\mathsf{T}$ and $\mathbf{u} \in \mathbb{R}^m, \mathbf{v} \in \mathbb{R}^n$, assuming $p = (I, J), I = p \mid m, J = p \mod m$. then we could rewrite 10 as

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

Before we start to construct the area containing $\hat{\theta}$, from 2 we know that we have to find a $\tilde{\theta}$ in the dual feasible area $\mathcal{R}^D$ firstly, there is a relationship between the primal variable and dual variable $\theta = \mathbf{y} - \mathbf{X}\mathbf{t}$, however, sometimes the outcome $\theta \notin \mathcal{R}^D$, which asks us to project. In the lasso problem, as the constraints limit the $\|\mathbf{x}_p\theta\|_1$, and every element of θ is multiplied by a dense x_i, researchers have to use a shrinking method to obtain a $\tilde{\theta} \in \mathcal{R}^D$ for further constructing the dual screening area:

$$\tilde{\theta} = \frac{(\mathbf{y} - \mathbf{X}\mathbf{t})}{\max(\lambda \mathbf{c}, \|\mathbf{X}^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{t})\|_\infty)} \tag{12}$$

This method pushes the θ far away from the optimum $\hat{\theta}$ and harms the screening effectiveness. As for the UOT problem, it only allows $\mathbf{t}_p \geq 0$, and the x_p only consists of two non-zero elements, which allows us to adapt a better projection method:

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

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

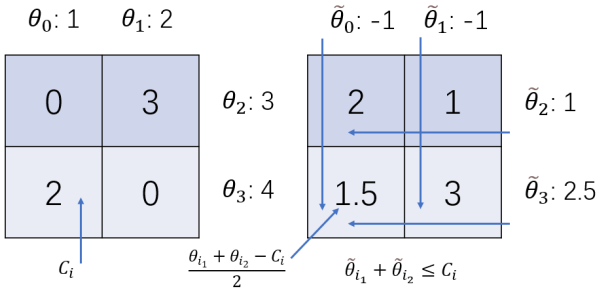

Figure 1: Shifting on a 2×2 matrix

As we have got the $\tilde{\theta}$ in the R^D and we also have another constraint area $\mathcal{R}^C$, we are sure that the $\hat{\mathbf{t}} \in \mathcal{R}^C \cap \mathcal{R}^D$. However, The intersection of a sphere and a polytope can not be computed in $O(knm)$, where k is a constant. We design a relaxation method. which divides the constraints into two parts, then we maximize the intersection of two hyperplanes and a hyperball.

Theorem 6. *(Two plane Screening for UOT) For every single primal variable t_p, let $A_p = \{i \| 0 \leq i < nm, i \mid m = I \vee i \mod m = J\}$, $B_p = \{i \| 0 \leq i < nm, i \notin A_p\}$. we can construct the specific area $\mathcal{R}^S_{IJ}$*

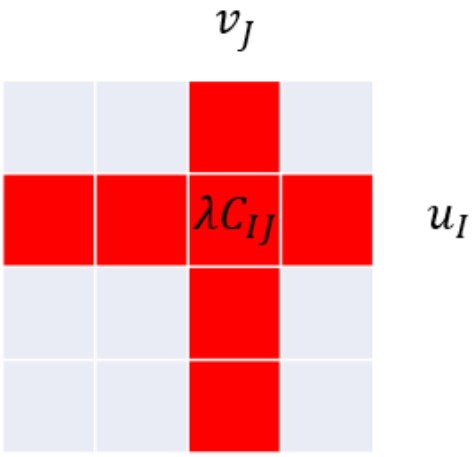

Figure 2: Selection of group A_{IJ}(red) and B_{IJ}(grey)

for it.

$$\mathcal{R}^S_{IJ} = \{\theta \| \begin{array}{c} \sum_{l \in A_p} (\theta^\mathsf{T} \mathbf{x}_l \mathbf{t}_l - \lambda \mathbf{c}_l \mathbf{t}) \leq 0 \\ \sum_{l \in B_p} (\theta^\mathsf{T} \mathbf{x}_l \mathbf{t}_l - \lambda \mathbf{c}_l \mathbf{t}) \leq 0 \} \\ (\theta - \tilde{\theta})^\mathsf{T} (\theta - \mathbf{y}) \leq 0 \end{array} \tag{14}$$

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

2.2 Screening Algorithms

The screening method is irrelevant to the optimization solver you choose. We give the specific algorithm for L_2 UOT problem to show the whole optimization process. The update indicates the updating process for $\mathbf{t}$ according to the optimizer you choose.

Algorithm 1 UOT Dynamic Screening Algorithm

Input: $\mathbf{t}_0, S \in R^{n \times m}, S_{ij} = 1, (i,j) = mi + j$
Output: S

```
Choose a solver for the problem.
for k = 0 to K do
  Projection θ̃ = Proj(t^k)
  for i = 0 to m do
    for j = 0 to n do
      R^S ⇐ R_)|^S(θ̃, t^k)
      S ⇐ S_ij = 0 if max_{θ∈R^S} x_(i,j)^T θ < λc_(i,j)
    end for
  end for
  for (i,j) ∈ {(i,j)‖S_ij = 0} do
    t^k_(i,j) ⇐ 0
  end for
  t^{k+1} = update(t^k)
end for
return t^{K+1}, S
```

3 EXPERIMENTS

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

3.1 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 $\lambda = \frac{\|\mathbf{X}^\mathsf{T} y\|}{100}$ 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.

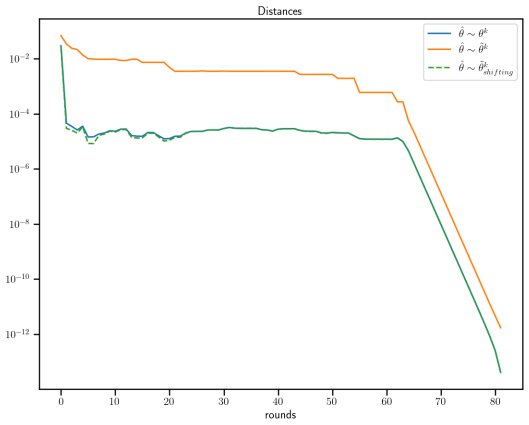

Figure 3: Distance of different projection method

3.2 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

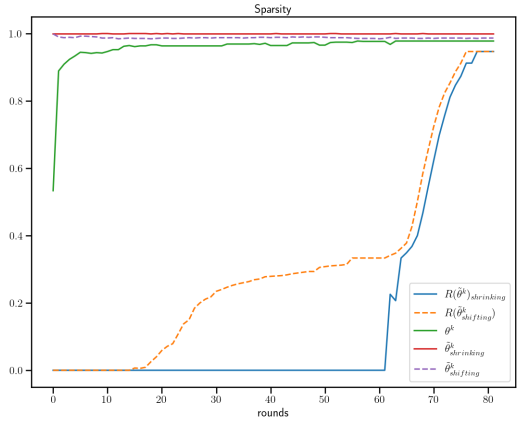

Figure 4: Screening ratio of different projection method

in the construction of the feasible domain.

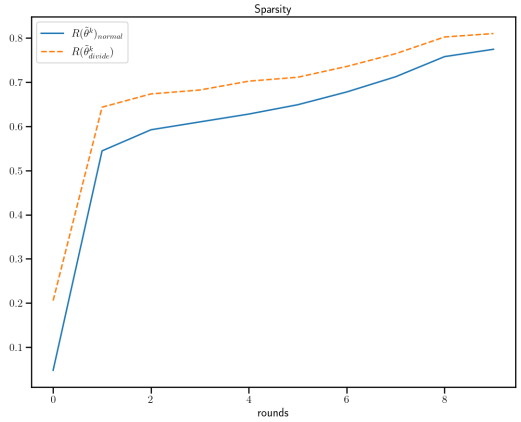

Figure 5: Screening ratio of dividing method

3.3 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.

3.4 Speed up Ratio

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

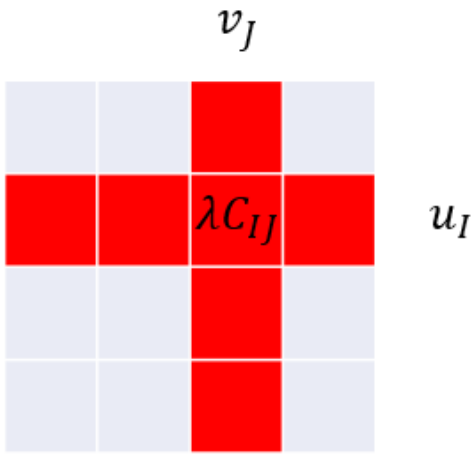

Figure 6: Comparing of our separation method with random separation method

4 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,

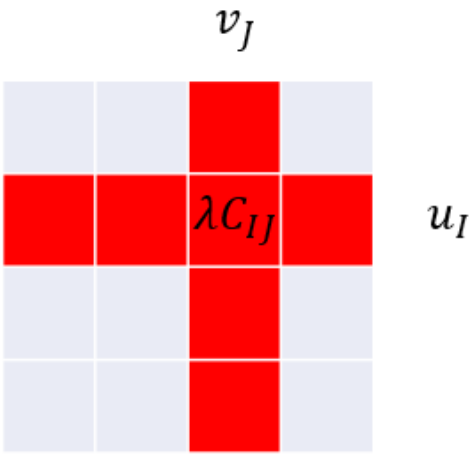

Figure 7: speed up ratio for different solver

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

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

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

B PROOFS

B.1 Proof of Theorem 6

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

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

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

B.2 Proof of Theorem 7

We Generalize the problem as

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

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

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

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

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

We can get the Lagrangian dual problem:

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

From the KKT optimum condition, we know that if

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

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:

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

we have

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

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:

$$\begin{aligned} 0 &= a\eta^2 + b\eta + c \\ a &= 4r^2 - s_1^2\|a\|^2 - u_1^2\|b\|^2 - 2s_1u_1 a^\mathsf{T} b \\ b &= 2(a_{I_1} + a_{I_2})s_1 + 2(b_{I_1} + b_{I_2})u_1 - 2s_1s_2\|a\|^2 - 2u_1u_2\|b\|^2 - 2(s_1u_2 + s_2u_1)a^\mathsf{T} b \\ c &= 2(a_{I_1} + a_{I_2})s_2 + 2(b_{I_1} + b_{I_2})u_2 - s_2^2\|a\|_1 - u_2^2\|b\|_1 - 2s_2u_2 a^\mathsf{T} b - 2 \end{aligned} \tag{28}$$

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, $\min_{\theta \in \mathcal{R}_I^S} -(\mathbf{q}_{I_1} + \mathbf{q}_{I_2}) = -\sqrt{2}r$

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.

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

To have a solution, the equations satisfied

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

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

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