Dynamic Screening Method on the Unbalanced Optimal Transport Problem

Anonymous Author
Anonymous Institution

Abstract

This paper applies dynamic screening framework on the L_2 penalized Unbalanced Optimal Transport (UOT) problem. Recently, researchers have linked the UOT problem to the Lasso problem, which encourages us to apply a technique that has been widely used in the Lasso problem, the Screening method, to the UOT problem. With the screening method, we can reasonably and safely freeze the unimportant elements of the sparse UOT solution, thus saving computational time. We demonstrate the effectiveness of the screening method for the UOT. Benefiting from the unique structure of the UOT problem, our proposed improved algorithm substantially improves the screening efficiency compared to the Lasso algorithm without significantly increasing the computational complexity. We demonstrate the advantages of the algorithm through constructed experiments.

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 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 created by modifying the restriction term to a penalty term. It 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. Entropy form can also be solved by the Sinkhorn algorithm. It has many advantages, for example its theoretical complexity is even better than OT problem [Pham et al., 2020], but it is not perfect, for example, its solution is always dense due to the KL penalty term and has a larger error compared with other regularizers [Blondel et al., 2018], which also brings some difficulties to many researches and applications. For this reason, many scholars have proposed other penalty terms for the UOT problem, such as L_2 and L_1 e.g. New problems have brought new solvers, such as FISTA, Majorization-Minimization method and Lagrange pairwise method. [Chapel et al., 2021] This is because the UOT problem has a similar structure to many other well-known problems such as non-negative matrix decomposition and Lasso problem, which encourages researchers to improve it by using the rich results in these fields.

Screening is a well-known technique promoted by [Ghaoui et al., 2010] in the field of lasso problems, where the penalty function of L_1 leads to a sparse solution of the problem, which can preselect solutions that must be zero by theory and freeze them before computation. The solutions of 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 before enabling the algorithm with linear complexity, thus saving optimization time. Screening method got attention in recent years

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

and have promoted a lot, such as Dynamic Screening [Bonnefoy et al., 2015], Gap screening method [Ndiaye et al., 2017] and Dynamic Sasvi [Yamada and Yamada, 2021]

Fortunately, the OT function in the UOT problem has the same effectiveness as L_1 in lasso, making the solution of the optimal problem always sparse. We believe that this means that the screening method that works in the Lasso problem can be applied to the UOT problem and, due to the unique structure of the UOT problem, will work better.

Contribution:

- We systematically provide the framework for Screening method on UOT problem. We give the correct projection method for UOT screening, which is better than the Lasso one.
- We promoted a two plane screening method for UOT problems, which benefits from its sparse constraints and outperforms the ordinary methods.

2 Background

2.1 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:

$$W(\mathbf{a},\mathbf{b}) := \min_{\mathbf{T}\in\mathbb{R}_+^{n\times n}} \langle \mathbf{C}, \mathbf{T}\rangle$$
$$\mathbf{T}\mathbb{1}_n = \mathbf{a}, \mathbf{T}^T\mathbb{1}_m = \mathbf{b}$$

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

$$W(\mathbf{a},\mathbf{b}) := \min_{t\in\mathbb{R}_+^{n^2}} c^\mathsf{T} t$$
$$\mathbf{N}\mathbf{t} = \alpha, \mathbf{M}\mathbf{t} = \beta$$

$\mathbf{N} \in \mathbb{R}^{m\times mn}, \mathbf{N} \in \mathbb{R}^{n\times mn}$ 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 $\mathbf{t}^*$ is not exist. We define $\mathbf{y} = [\mathbf{a},\mathbf{b}]^\mathsf{T}$, the UOT problem use a penalty function for the historgrams:

$$W(\mathbf{a},\mathbf{b}) := \min_{t\in\mathbb{R}_+^{mn}} \mathbf{c}^\mathsf{T} t + D_h(\mathbf{X}\mathbf{t},\mathbf{y}) \tag{1}$$

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

2.2 Relationship with Lasso

Lasso-like problem has a general formula as:

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

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.

2.3 Dynamic Screening Framework

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

$$f(\mathbf{t}) = g(\mathbf{t}) + d(\mathbf{X}\mathbf{t}) \tag{2}$$

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 2, if d and g are both convex, for $\forall$ feasible $\tilde{\theta} \in R^{m+n}$, we have the following area:*

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

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

3 Dynamic Screening and UOT problem

3.1 Screening for UOT

We can get the dual form of UOT problem:

Lemma 3. *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:*

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

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

For UOT problem 1, we could get its dual form.

Lemma 4. *(Dual form of UOT problem)*

$$-d^*(-\theta) - g^*(\mathbf{X}^\mathsf{T}\theta) = -\frac{1}{2}\|\theta\|_2^2 - \mathbf{y}^\mathsf{T}\theta \quad \text{s.t.} \quad \forall i \quad \mathbf{x}_i^\mathsf{T}\theta - \lambda \mathbf{c}_i \leq 0 \tag{3}$$

$\mathbf{x}_i$ is the i-th column of $\mathbf{X}$, these inequations 3 make up a dual feasible area written as $\mathcal{R}^D$, and the optimal solution definitely satisfied them.
From the KKT condition, we konw that, for the optimal primal solution $\hat{\mathbf{t}}$:

Theorem 5. *(KKT condition) For the dual optimal solution $\hat{\theta}$, we have the following relationship:*

$$\mathbf{x}_i^\mathsf{T}\hat{\theta} - \lambda \mathbf{c}_i \begin{cases} < 0 & \Rightarrow \hat{\mathbf{t}}_i = 0 \\ = 0 & \Rightarrow \hat{\mathbf{t}}_i \geq 0 \end{cases} \tag{4}$$

As we do not know the information of $\hat{\mathbf{t}}$ directly, we can construct an area $\mathcal{R}^S$ containing the $\hat{\mathbf{t}}$, if

$$\max_{\mathbf{t} \in \mathcal{R}^S} \mathbf{x}_i^\mathsf{T}\theta - \lambda \mathbf{c}_i < 0 \tag{5}$$

then we have:

$$\mathbf{x}_i^\mathsf{T}\hat{\theta} - \lambda \mathbf{c}_i < 0 \tag{6}$$

which means the corresponding $\hat{t}_i = 0$, and can be screening out.

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 before we construct any area, there is a relationship between the primal variable and dual variable $\theta = \mathbf{y} - \mathbf{X}\mathbf{t}$, however, the outcome θ might not inside the dual feasible area, which encourage us to project. In lasso problem, as the constraints limit the$\|\mathbf{x}_i\theta\|_1$, researchers would use a shrinking method like:

$$\tilde{\theta} = \frac{(\mathbf{y} - \mathbf{X}\mathbf{t})}{\max(\lambda \mathbf{c}, \|\mathbf{X}^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{t})\|_\infty)} \tag{7}$$

Then $\tilde{\theta}$ would be in the dual feasible are. As for UOT problem, it only allow $\mathbf{t}_i \geq 0$, and its x_i only consists of two non-zero elements, which allows us to adapt a better projection method:

Theorem 6. *(UOT projection) For any any $\theta \in \mathbb{R}^{n+m}$, we can compute the projection $\tilde{\theta}$ onto the dual feasible area, j_1 and j_2 indicates the non-zero elements in each x_j*

$$\tilde{\theta}_i = \begin{cases} \theta_i - \max_{j|n=i}(\frac{\theta_{j_1} + \theta_{j_2} - \mathbf{c}_j}{2}) & 0 \leq i < m \\ \theta_i - \max_{j \bmod n = i-m}(\frac{\theta_{j_1} + \theta_{j_2} - \mathbf{c}_j}{2}) & m \leq i < n+m \end{cases} \tag{8}$$

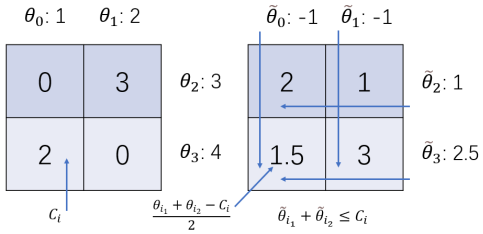

Figure 1: Shifting on a 2×2 matrix

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 compute in $O(knm)$, where k is a constant. We design a relaxation method. which divide the constrains into two parts, then we are maximizing on the intersection of two hyperplanes and a hyper-ball.

Theorem 7. *(Screening Area for UOT) With the help of $\tilde{\theta}$, we can construct following area $\mathcal{R}^S$, and the optimal dual solution $\hat{\theta}$ must be inside the area.*

$$\mathcal{R}^S = \{\theta \| \begin{matrix} \theta^\mathsf{T}\mathbf{X}^A\beta - \lambda g^A\beta \leq 0 \\ \theta^\mathsf{T}\mathbf{X}^B\beta - \lambda g^B\beta \leq 0 \\ (\theta - \tilde{\theta})^\mathsf{T}(\theta - \mathbf{y}) \leq 0 \end{matrix}\} \tag{9}$$

We devide the constraints into two group A and B, we have $X^A + X^B = X$ and $g^A + g^B = g$ the computational process is in Appendix.A

3.2 Screening Algorithms

Algorithm 1 UOT Dynamic Screening Algorithm

Input: $t_0, S \in R^{n \times m}, S_{ij} = 1$
Output: S

Choose a solver for the problem.
for $t = 0$ to K **do**
  Projection $\tilde{\theta} = \text{Proj}(t^k)$
  if $(i \neq 0)$ **then**
    break
  end if
  $\mathcal{R} \Leftarrow \mathcal{R}^S(\tilde{\theta}, t^k)$
  $S \Leftarrow S_{ij} = 0$ if $\max_{\theta \in \mathcal{R}^S} {x_{k(i,j)}}^\mathsf{T}\theta < \lambda c_{k(i,j)}$
  for $a \in A_{ij} \| A_{ij} = 0$ **do**
    $t^k(i,j) \Leftarrow 0$
  end for
  $t^{k+1} = \text{update}(t^k)$
end for
return t^{K+1}, S

screening method is irrelevent to the optimization solver you choose. We give the specific algorithm for

L_2 UOT problem to show the whole optimization process.

4 Experiments

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

4.1 Screening Ratio

5 Conclusion

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

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport. In Amos J. Storkey and Fernando Pérez-Cruz, 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, 2018. URL http://proceedings.mlr.press/v84/blondel18a.html.

Antoine Bonnefoy, Valentin Emiya, Liva Ralaivola, and Rémi Gribonval. Dynamic screening: Accelerating first-order algorithms for the lasso and group-lasso. *IEEE Transactions on Signal Processing*, 63(19):5121–5132, 2015. doi: 10.1109/TSP.2015.2447503.

Laetitia Chapel, Rémi Flamary, Haoran Wu, Cédric Févotte, and Gilles Gasso. Unbalanced optimal transport through non-negative penalized linear regression. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 23270–23282. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/c3c617a9b80b3ae1ebd868b0017cd34b-Paper.pdf.

Liqun Chen, Yizhe Zhang, Ruiyi Zhang, Chenyang Tao, Zhe Gan, Haichao Zhang, Bai Li, Dinghan Shen, Changyou Chen, and Lawrence Carin. Improving sequence-to-sequence learning via optimal transport. In *7th International Conference on Learning Representations, ICLR 2019*. International Conference on Learning Representations, ICLR, January 2019. Generated from Scopus record by KAUST IRTS on 2021-02-09.

Nicolas Courty. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017. doi: 10.1109/TPAMI.2016.2615921.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/af21d0c9

Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. Safe feature elimination for the lasso and sparse supervised learning problems. *arXiv preprint arXiv:1009.4219*, 2010.

Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Wasserstein regularization for sparse multi-task regression. In Kamalika Chaudhuri and Masashi Sugiyama, 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, 2019. URL http://proceedings.mlr.press/v89/janati19a.html.

Eugene Ndiaye, Olivier Fercoq, Alex, re Gramfort, and Joseph Salmon. Gap safe screening rules for sparsity enforcing penalties. *Journal of Machine Learning Research*, 18(128):1–33, 2017. URL http://jmlr.org/papers/v18/16-577.html.

Hermina Petric Maretic, Mireille El Gheche, Giovanni Chierchia, and Pascal Frossard. Got: An optimal transport framework for graph comparison. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/fdd5b16f

Khiem Pham, Khang Le, Nhat Ho, Tung Pham, and Hung Bui. On unbalanced optimal transport: An analysis of sinkhorn algorithm. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 7673–7682. PMLR, 2020. URL http://proceedings.mlr.press/v119/pham20a.html.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, Jenny Chen, Justin Brumbaugh, Philippe Rigollet, Konrad Hochedlinger, Rudolf Jaenisch, Aviv Regev, and Eric S. Lander. Optimal-transport

analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943.e22, 2019. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2019.01.006. URL https://www.sciencedirect.com/science/article/pii/S009286741930039X.

Hiroaki Yamada and Makoto Yamada. Dynamic sasvi: Strong safe screening for norm-regularized least squares. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 14645–14655. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/7b5b23f4aadf9513306bcd59afb6e4c9-Paper.pdf.

Karren D. Yang and Caroline Uhler. 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, 2019. URL https://openreview.net/forum?id=HyexAiA5Fm.

Supplementary Material: Dynamic Screening Method on the Unbalanced Optimal Transport Problem

A Entropic UOT dual

We know that

$$P(t) := \min_{t \in \mathbb{R}_+^{mn}} f(\mathbf{X}t) + g(t) \tag{10}$$

the dual problem is

$$D(\theta) := \max_{\theta} -f^*(-\theta) - g^*(X^\mathsf{T}\theta) \tag{11}$$

The Entropic UOT problem is:

$$W(\alpha, \beta) := \min_{t \in \mathbb{R}_+^{mn}, t_i > \epsilon > 0} \lambda c^\mathsf{T} t + D_h(\mathbf{X}t, y) + \varepsilon H(t) \tag{12}$$

$H(x) = x^\mathsf{T}(\ln x - 1)$, In order to screening, we introduce a smoothing threshold $t \geq \epsilon > 0$

$$\begin{aligned} f(\mathbf{X}t) &= D_h(\mathbf{X}t, y) \\ g(t) &= \lambda c^\mathsf{T} t + \varepsilon H(t) \end{aligned}$$

We have

$$\begin{aligned} f^*(\theta) &= \max_t \theta^\mathsf{T} t - t \ln \frac{t}{y} + t^\mathsf{T}\mathbb{1} - y^\mathsf{T}\mathbb{1} \quad (t^* = y \odot e^\theta) \\ &= y^\mathsf{T} e^\theta - y^\mathsf{T}\mathbb{1} \\ g^*(\theta) &= \max_t \theta^\mathsf{T} t - \varepsilon t^\mathsf{T}(\ln t - 1) - \lambda c^\mathsf{T} t \quad (t^* = \exp \frac{\theta - \lambda c}{\varepsilon}) \\ &= \sum_{i, \theta_i \geq \lambda c_i + \varepsilon \ln \epsilon} \varepsilon \exp(\frac{\theta_i - \lambda c_i}{\varepsilon}) + \sum_{i, \theta_i < \lambda c_i + \varepsilon \ln \epsilon} \epsilon(\theta_i - \varepsilon(\ln \epsilon - 1) - \lambda c_i) \end{aligned}$$

The dual problem is:

$$D(\theta) := \max_{\theta} -y^\mathsf{T} e^{-\theta} + y^\mathsf{T}\mathbb{1} - \sum_{i, x_i^\mathsf{T}\theta \geq \lambda c_i + \varepsilon \ln \epsilon} \varepsilon \exp(\frac{x_i^\mathsf{T}\theta - \lambda c_i}{\varepsilon}) - \sum_{i, x_i^\mathsf{T}\theta < \lambda c_i + \varepsilon \ln \epsilon} \epsilon(x_i^\mathsf{T}\theta - \varepsilon(\ln \epsilon - 1) - \lambda c_i) \tag{13}$$

Let's have a look whether it is strongly conconcave...

$$\frac{\partial D}{\partial \theta} = y^\mathsf{T} e^{-\theta} - \sum_{i, x_i^\mathsf{T}\theta \geq \lambda c_i + \varepsilon \ln \epsilon} x_i \exp(\frac{x_i^\mathsf{T}\theta - \lambda c_i}{\varepsilon}) - \sum_{i, x_i^\mathsf{T}\theta < \lambda c_i + \varepsilon \ln \epsilon} \epsilon x_i$$

$$\frac{\partial^2 D}{\partial \theta^2} = -\operatorname{Diag}(y^\mathsf{T} e^{-\theta}) - \sum_{i, x_i^\mathsf{T}\theta \geq \lambda c_i + \varepsilon \ln \epsilon} \frac{x_i x_i^\mathsf{T}}{\varepsilon} \exp(\frac{x_i^\mathsf{T}\theta - \lambda c_i}{\varepsilon})$$

L is

$$\begin{aligned} L(t, v, \eta, u, m) &= \min_{t,v} \max_{\eta > 0, u, m} \lambda c^\mathsf{T} t + D_h(y, v) + \varepsilon H(t) + \eta^\mathsf{T}(-t) + u^\mathsf{T}(v - \mathbf{X}t) + m(t^\mathsf{T}\mathbb{1} - a) \\ &= \max_{\eta > 0, u, m} -ma + \min_{t,v} \lambda c^\mathsf{T} t + D_h(y, v) + \varepsilon H(t) + \eta^\mathsf{T}(-t) + u^\mathsf{T}(v - \mathbf{X}t) + mt^\mathsf{T}\mathbb{1} \end{aligned}$$

$$\frac{\partial L}{\partial v} = \tag{14}$$

$$W(\alpha, \beta) := \min_{t \in \mathbb{R}_+^{mn}} f(\mathbf{X}t) + g(t) \tag{15}$$

We have

$$f(Xt) = D_h(y, \mathbf{X}t)$$

$$g(t) = \lambda c^\mathsf{T} t + \varepsilon(t + \eta_2)\ln(t + \eta_2)$$

$$\begin{aligned} g^*(\theta) &= \max_t \theta^\mathsf{T} t - g(x) \\ \frac{\partial g^*}{\partial t} &= \theta - \lambda c - \varepsilon(\ln(t + \eta_2) + 1) \\ t^* &= \exp(\frac{\theta - \lambda c}{\varepsilon} - 1) - \eta_2 \\ g^*(\theta) &= (\lambda c - \theta)\eta_2 + \varepsilon \exp(\frac{\lambda c - \theta}{\varepsilon} - 1) \mathbf{s.t.} \end{aligned}$$

B FORMATTING INSTRUCTIONS FOR THE SUPPLEMENTARY MATERIAL

Your supplementary material should go here. It may be in one-column or two-column format. To display the supplementary material in two-column format, comment out the line

`\onecolumn \makesupplementtitle`

and uncomment the following line:

`\twocolumn[ \makesupplementtitle ]`

Please submit your paper (including the supplementary material) as a single PDF file. Besides the PDF file, you may submit a single file of additional non-textual supplementary material, which should be a ZIP file containing, e.g., code.

If you require to upload any video as part of the supplementary material of your camera-ready submission, do not submit it in the ZIP file. Instead, please send us via email the URL containing the video location.

Note that reviewers are under no obligation to examine your supplementary material.

C MISSING PROOFS

The supplementary materials may contain detailed proofs of the results that are missing in the main paper.

C.1 Proof of Lemma 3

In this section, we present the detailed proof of Lemma 3 and then [...]

D ADDITIONAL EXPERIMENTS

If you have additional experimental results, you may include them in the supplementary materials.

D.1 The Effect of Regularization Parameter

Our algorithm depends on the regularization parameter λ. Here we illustrate the effect of this parameter on the performance of our algorithm [...]