Supplementary Materials for "Robust computation of optimal transport by β-potential regularization"

Shintaro Nakamura NAKAMURASHINTARO@G.ECC.U-TOKYO.AC.JP
The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa City, Chiba 277-8561

Han Bao BAO@I.KYOTO-U.AC.JP
Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501

Masashi Sugiyama SUGI@K.U-TOKYO.AC.JP
RIKEN AIP center, Nihonbashi 1-chome Mitsui Building, 15th floor, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027
The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa City, Chiba 277-8561

Editors: Emtiyaz Khan and Mehmet Gönen

Appendix A. Details of the Bregman projection

Here, we demonstrate the details of our algorithm inspired by the Non-negative alternate scaling algorithm (NASA) introduced by Dessein et al. (2018), which is an algorithm to obtain the solution for CROT. Although our outlier-robust CROT does not satisfy the assumptions required for CORT, we show our algorithm is constructed similarly to the NASA algorithm.

First, we introduce basics of convex analysis as preliminaries. Next, we explain the alternate scaling algorithm, which is the basis of the NASA algorithm, and the required assumptions for it. Finally, we show the NASA algorithm for the separable Bregman divergence, and how we borrowed their idea to construct our algorithm.

A.1. Convex analysis

Let $\mathcal{E}$ be a Euclidean space with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$. The boundary, interior, and relative interior of a subset $\mathcal{S} \subseteq \mathcal{E}$ are denoted by $\mathrm{bd}(\mathcal{S})$, $\mathrm{int}(\mathcal{S})$, and $\mathrm{ri}(\mathcal{S})$, respectively. Recall that for a convex set $\mathcal{C}$, we have

$$\mathrm{ri}(\mathcal{C}) = \{\boldsymbol{x} \in \mathcal{E} \mid \forall \boldsymbol{y} \in \mathcal{C},\ \exists \lambda > 1,\ \lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in \mathcal{C}\}. \tag{1}$$

In convex analysis, scalar functions are defined over the whole space $\mathcal{E}$ and take values in $\mathbb{R} \cup \{-\infty,\ \infty\}$. The effective domain, or simply domain, of a function f is defined as the set:

$$\mathrm{dom}\, f = \{\boldsymbol{x} \in \mathcal{E} \mid f(\boldsymbol{x}) < +\infty\}. \tag{2}$$

Definition 1 (Closed functions)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be closed if for each $\alpha \in \mathbb{R}$, the sublevel set $\{\boldsymbol{x} \in \mathrm{dom}\, f \mid f(\boldsymbol{x}) \leq \alpha\}$ is a closed set.*

If $\mathrm{dom}\, f$ is closed, then f is closed.

Definition 2 (Proper functions). *Suppose a convex function $f : \mathcal{E} \to \mathbb{R} \cup \{\pm\infty\}$ satisfies $f(\boldsymbol{x}) > -\infty$ for every $\boldsymbol{x} \in \operatorname{dom} f$ and there exists some point $\boldsymbol{x}_0$ in its domain such that $f(\boldsymbol{x}_0) < +\infty$. Then f is called a proper function.*

A proper convex function is closed if and only if it is lower semi-continuous. [1] A closed function f is continuous relative to any simplex, polytope of a polyhedral subset in $\operatorname{dom} f$. A convex function f is always continuous in the relative interior $\operatorname{ri}(\operatorname{dom} f)$.

Definition 3 (Essential smoothness Bauschke and Borwein (1997)). *Suppose f is a closed convex proper function on $\mathcal{E}$ with $\operatorname{int}(\operatorname{dom} f) \neq \emptyset$. Then f is essentially smooth, if f is differentiable on $\operatorname{int}(\operatorname{dom} f)$ and*

$$\left.\begin{array}{c} \forall n \in \mathbb{N},\ x_n \in \operatorname{int}(\operatorname{dom} f), \\ x_n \to x \in \operatorname{bd}(\operatorname{dom} f) \end{array}\right\} \Rightarrow \|\nabla f(\boldsymbol{x}_n)\| \to \infty.$$

Definition 4 (Essential strict convexity Bauschke and Borwein (1997)). *Let ∂f be the subgradient of f. Suppose f is closed convex proper on $\mathcal{E}$. Then, f is essentially strictly convex, if f is strictly convex on every convex subset of $\operatorname{dom}(\partial f)$.*

We define a set of functions called the Legendre type and Fenchel conjugate functions.

Definition 5 (Legendre type Bauschke and Borwein (1997)). *Suppose f is a closed convex proper function on $\mathcal{E}$. Then, f is said to be of the Legendre type if f is both essentially smooth and essentially strictly convex.*

Definition 6 (Fenchel conjugate Dessein et al. (2018)). *The Fenchel conjugate f^* of a function f is defined for all $\boldsymbol{y} \in \mathcal{E}$ as follows:*

$$f^*(\boldsymbol{y}) = \sup_{\boldsymbol{x} \in \operatorname{int}(\operatorname{dom} f)} \langle \boldsymbol{x}, \boldsymbol{y} \rangle - f(\boldsymbol{x}). \tag{3}$$

The Fenchel conjugate f^* is always a closed convex function and if f is a closed convex function, then $(f^*)^* = f$, and f is of the Legendre type if and only if f^* is of the Legendre type. If f^* is of the Legendre type, the gradient mapping ∇f is a homeomorphism[2] between $\operatorname{int}(\operatorname{dom} f)$ and $\operatorname{int}(\operatorname{dom} f^*)$, with inverse mapping $(\nabla f)^{-1} = \nabla f^*$. This guarantees the existence of dual coordinate systems $\boldsymbol{x}(\boldsymbol{y}) = \nabla f^*(\boldsymbol{y})$ and $\boldsymbol{y}(\boldsymbol{x}) = \nabla f(\boldsymbol{x})$ on $\operatorname{int}(\operatorname{dom} f)$ and $\operatorname{int}(\operatorname{dom} f^*)$.

Finally, we say that a function f is a cofinite if it satisfies

$$\lim_{\lambda \to +\infty} f(\lambda \boldsymbol{x})/\lambda = +\infty, \tag{4}$$

for all nonzero $\boldsymbol{x} \in \mathcal{E}$. Intuitively, it means that f grows super-linearly in every direction. In particular, a closed convex proper function is cofinite if and only if $\operatorname{dom} f^* = \mathcal{E}$.

1. Let X be a topological space. A funtion $f : X \to \mathbb{R} \cup \{-\infty, \infty\}$ is called lower semi-continuous at a point $x_0 \in X$ if for every $y < f(x_0)$ there exists a neighborhood U of x_0 such that $f(x) > y$ for all $x \in U$.
2. A function $f : X \to Y$ between two topological spaces is a homeomorphism if it has the following three properties: (a) f is a bijection. (b) f is continuous. (c) The inverse function f^{-1} is continuous.

A.2. Alternate scaling algorithm

Here, we show the details of obtaining the Bregman projection onto a convex set. Let ϕ be a function of the Legendre type with Fenchel conjugate $\phi^* = \psi$. In general, computing Bregman projections onto an arbitrary closed convex set $\mathcal{C} \subseteq \mathcal{E}$ such that $\mathcal{C} \cap \mathrm{int}(\mathrm{dom}\,\phi) \neq \emptyset$ is nontrivial Dessein et al. (2018). Sometimes, it is possible to decompose $\mathcal{C}$ into an intersection of finitely many closed convex sets:

$$\mathcal{C} = \bigcap_{l=1}^{s} \mathcal{C}_l, \tag{5}$$

where the individual Bregman projections onto the respective sets $\mathcal{C}_1, \ldots, \mathcal{C}_s$ are easier to compute. It is then possible to obtain the Bregman projections onto $\mathcal{C}$ by alternate projections onto $\mathcal{C}_1, \ldots, \mathcal{C}_s$ according to Dykstra's algorithm (Boyle and Dykstra, 1986).

In more detail, let $\sigma : \mathbb{N} \to \{1, \ldots, s\}$ be a control mapping that determines the sequence of subsets onto which we project. For a given point $\boldsymbol{x}_0 \in \mathrm{int}(\mathrm{dom}\phi)$, the Bregman projection $T_{\mathcal{C}}(\boldsymbol{x}_0)$ of $\boldsymbol{x}_0$ onto $\mathcal{C}$ can be approximated with Dykstra's algorithm by iterating the following updates:

$$\boldsymbol{x}_{k+1} \leftarrow T_{\mathcal{C}_{\sigma(k)}}(\nabla\psi(\nabla\phi(\boldsymbol{x}_k + \boldsymbol{y}^{\sigma(k)})), \tag{6}$$

where the correction term $\boldsymbol{y}^1, \ldots, \boldsymbol{y}^s$ for the respective subsets are initialized with the null element of $\mathcal{E}$, and are updated after projection as follows:

$$\boldsymbol{y}^{\sigma(k)} \leftarrow \boldsymbol{y}^{\sigma(k)} + \nabla\phi(\boldsymbol{x}_k) - \nabla\phi(\boldsymbol{x}_{k+1}). \tag{7}$$

Under some technical assumptions, the sequence of updates $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ converges in terms of some norm to $P_{\mathcal{C}}(\boldsymbol{x_0})$ with a linear rate. Several sets of such conditions have been studied (Dhillon and Tropp, 2007; Bauschke and Lewis, 1993, 2000). Here, we use the following conditions proposed by Dhillon and Tropp [2007] for the CROT framework:

- The function ϕ is cofinite
- The constraint qualification $\mathrm{ri}(\mathcal{C}_1) \cap \cdots \cap \mathrm{ri}(\mathcal{C}_s) \cap \mathrm{int}(\mathrm{dom}\phi) \neq \emptyset$ holds
- The control mapping σ is essentially cyclic, that is , there exists a number $t \in \mathbb{N}$ such that σ takes each output value at least once during any t consecutive input values

Once these conditions are imposed, the convergence of Dykstra's algorithm is guaranteed.

A.3. Technical assumptions for CROT to hold

Some mild technical assumptions are required on the convex regularizer ϕ and its Fenchel conjugate $\psi = \phi*$ for the CROT framework to hold. The assumptions are as follows:

1. ϕ is of Legendre type.
2. $(0, 1)^{d\times d} \subseteq \mathrm{dom}\,\phi$
3. $\mathrm{dom}\,\psi = \mathbb{R}^{d\times d}$

Algorithm NASA algorithm

$\tilde{\boldsymbol{\theta}} \leftarrow -\boldsymbol{\gamma}/\lambda$
$\boldsymbol{\theta}^* \leftarrow \max\{\nabla\phi(\mathbf{0}_{m\times n}), \tilde{\boldsymbol{\theta}}\}$
repeat
  $\boldsymbol{\tau} \leftarrow \mathbf{0}_m$
  repeat
    $\boldsymbol{\tau} \leftarrow \boldsymbol{\tau} + \frac{\nabla\psi(\boldsymbol{\theta}^* - \boldsymbol{\tau}\mathbf{1}_n^\top)\mathbf{1}_n - \frac{\mathbf{1}_m}{m}}{\nabla^2\psi(\boldsymbol{\theta}^* - \boldsymbol{\tau}\mathbf{1}_n^\top)\mathbf{1}_n}$
  until convergence
  $\tilde{\boldsymbol{\theta}} \leftarrow \tilde{\boldsymbol{\theta}} - \boldsymbol{\tau}\mathbf{1}_n^\top$
  $\boldsymbol{\theta}^* \leftarrow \max\{\nabla\phi(\mathbf{0}_{m\times n}), \tilde{\boldsymbol{\theta}}\}$
  $\boldsymbol{\sigma} \leftarrow \mathbf{0}_n^\top$
  repeat
    $\boldsymbol{\sigma} \leftarrow \boldsymbol{\sigma} + \frac{\mathbf{1}_m^\top\nabla\psi(\boldsymbol{\theta}^* - \mathbf{1}_m\boldsymbol{\sigma}) - (\frac{\mathbf{1}_n}{n})^\top}{\mathbf{1}_m^\top\nabla^2\psi(\boldsymbol{\theta}^* - \mathbf{1}_m\boldsymbol{\sigma})}$
  until convergence
  $\tilde{\boldsymbol{\theta}} \leftarrow \tilde{\boldsymbol{\theta}} - \mathbf{1}_m\boldsymbol{\sigma}$
  $\boldsymbol{\theta}^* \leftarrow \max\{\nabla\phi(\mathbf{0}_{m\times n}), \tilde{\boldsymbol{\theta}}\}$
until convergence
$\boldsymbol{\pi}^* \leftarrow \boldsymbol{\nabla\psi}(\boldsymbol{\theta}^*)$

Some assumptions relate to required conditions for the definition of Bregman projections and convergence of the algorithms, while others are more specific to CROT problems.

The first assumption (1) is required for the definition of the Bregman projection. In addition, it guarantees the existence of dual coordinate systems on $\mathrm{int}(\mathrm{dom}\,\phi)$ and $\mathrm{int}(\mathrm{dom}\,\psi)$ via the homeomorphism $\nabla\phi = \nabla\psi^{-1}$.

The second assumption (2) ensures the constraint qualification $\mathcal{G}(\frac{\mathbf{1}_m}{m}, \frac{\mathbf{1}_n}{n}) \cap \mathrm{int}(\mathrm{dom}\,\phi)$ for the Bregman projection onto the transport polytope.

The third assumption (3) equivalently requires ϕ to be cofinite for convergence.

A.4. NASA algorithm

In this subsection, we show the NASA algorithm based on Dykstra's algorithm constructed by projections on $\mathcal{C}_0$,$\mathcal{C}_1$, and $\mathcal{C}_2$.

A.4.1. Projcetion onto $\mathcal{C}_0$

Let us consider the projection of given matrix $\overline{\boldsymbol{\pi}}$ onto $\mathcal{C}_0$. We denote this projection $P_{\mathcal{C}_0}(\overline{\boldsymbol{\pi}})$ by $\boldsymbol{\pi}_0^*$. Then, the Karush-Kuhn-Tucker conditions (Kuhn and Tucker, 1951; Karush, 1939) for $\boldsymbol{\pi}_0^*$ are as follows:

$$\boldsymbol{\pi}_0^* \geq \mathbf{0}, \tag{8}$$
$$\nabla\phi(\boldsymbol{\pi}_0^*) - \nabla\phi(\overline{\boldsymbol{\pi}}) \geq \mathbf{0}, \tag{9}$$
$$(\nabla\phi(\boldsymbol{\pi}_0^*) - \nabla\phi(\overline{\boldsymbol{\pi}})) \odot \boldsymbol{\pi}_0^* = \mathbf{0}, \tag{10}$$

where (8) is the primal feasibility, (9) is the dual feasibility, and (10) is the complementary slackness.

Since we are thinking of the separable Bregman divergence, the projection onto $\mathcal{C}_0$ can be performed with a closed-form expression on primal parameters:

$$\pi^*_{0,ij} = \max\{0, \overline{\pi}_{ij}\}, \tag{11}$$

where, $\pi^*_{0,ij}$ is the (i,j)-element of matrix $\boldsymbol{\pi}^*_0$. Since ϕ' is increasing, this is equivalent on the dual parameters of $\boldsymbol{\pi}^*_0$, $\boldsymbol{\theta}^*_0$, to

$$\theta^*_{0,ij} = \max\{\phi'(0), \overline{\theta}_{ij}\}. \tag{12}$$

Here, the dual coordinate of the input matrix $\overline{\boldsymbol{\pi}}$ is denoted by $\overline{\boldsymbol{\theta}}$.

A.4.2. Projection onto $\mathcal{C}_1$ and $\mathcal{C}_1$

Next, we consider the Bregman projections of a given matrix $\overline{\boldsymbol{\pi}} \in \mathrm{int}(\mathrm{dom}\phi)$ onto $\mathcal{C}_1$ and $\mathcal{C}_2$. For the projection onto $\mathcal{C}_1$ and $\mathcal{C}_2$, we employ the method of Lagrange multipliers. The Lagrangians with Lagrange multipliers $\boldsymbol{\mu} \in \mathbb{R}^m$ and $\boldsymbol{\nu} \in \mathbb{R}^n$ for the Bregman projections $\boldsymbol{\pi}^*_1$ and $\boldsymbol{\pi}^*_2$ of a given matrix $\overline{\boldsymbol{\pi}} \in \mathrm{int}(\mathrm{dom}\phi)$ onto $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively write as follows:

$$\mathcal{L}_1(\boldsymbol{\pi}, \boldsymbol{\mu}) = \phi(\boldsymbol{\pi}) - \langle \boldsymbol{\pi}, \nabla\phi(\overline{\boldsymbol{\pi}})\rangle + \mu^\top(\boldsymbol{\pi}\mathbf{1} - \frac{\mathbf{1}}{m}), \tag{13}$$

$$\mathcal{L}_2(\boldsymbol{\pi}, \boldsymbol{\nu}) = \phi(\boldsymbol{\pi}) - \langle \boldsymbol{\pi}, \nabla\phi(\overline{\boldsymbol{\pi}})\rangle + \nu^\top(\boldsymbol{\pi}^\top\mathbf{1} - \frac{\mathbf{1}}{n}). \tag{14}$$

Their gradients are given on $\mathrm{int}(\mathrm{dom}\phi)$ by

$$\nabla\mathcal{L}_1(\boldsymbol{\pi}, \boldsymbol{\mu}) = \nabla\phi(\boldsymbol{\pi}) - \nabla\phi(\overline{\boldsymbol{\pi}}) + \boldsymbol{\mu}\mathbf{1}^\top, \tag{15}$$

$$\nabla\mathcal{L}_2(\boldsymbol{\pi}, \boldsymbol{\nu}) = \nabla\phi(\boldsymbol{\pi}) - \nabla\phi(\overline{\boldsymbol{\pi}}) + \mathbf{1}\boldsymbol{\nu}^\top, \tag{16}$$

and vanish at $\boldsymbol{\pi}^*_1, \boldsymbol{\pi}^*_2 \in \mathrm{int}(\mathrm{dom}\phi)$ if and only if

$$\boldsymbol{\pi}^*_1 = \nabla\psi(\nabla\phi(\overline{\boldsymbol{\pi}}) - \boldsymbol{\mu}\mathbf{1}^\top), \tag{17}$$

$$\boldsymbol{\pi}^*_2 = \nabla\psi(\nabla\phi(\overline{\boldsymbol{\pi}}) - \mathbf{1}\boldsymbol{\nu}^\top). \tag{18}$$

By duality, the Bregman projections onto $\mathcal{C}_1$, $\mathcal{C}_2$ are thus equivalent to finding the unique vectors $\boldsymbol{\mu}$, $\boldsymbol{\nu}$, such that the rows of $\boldsymbol{\pi}^*_1$ sum up to $\frac{\mathbf{1}}{m}$, respectively the columns of $\boldsymbol{\pi}^*_2$ sum up to $\frac{\mathbf{1}}{n}$:

$$\nabla\psi(\nabla\phi(\overline{\boldsymbol{\pi}}) - \boldsymbol{\mu}\mathbf{1}^\top)\mathbf{1} = \frac{\mathbf{1}}{m}, \tag{19}$$

$$\nabla\psi(\nabla\phi(\overline{\boldsymbol{\pi}}) - \mathbf{1}\boldsymbol{\nu}^\top)^\top\mathbf{1} = \frac{\mathbf{1}}{n}. \tag{20}$$

Again, since we are resticting ourselves to the separable Bregman divergence, we can compute the projection step more efficiently. Due to the separability, the projections onto $\mathcal{C}_1$ and $\mathcal{C}_2$ can be divided into m and n parallel subproblems in the search space of 1-dimension as follows:

$$\sum_{j=1}^{n} \psi'(\overline{\theta}_{ij} - \mu_i) = \frac{1}{m}, \tag{21}$$

$$\sum_{i=1}^{m} \psi'(\overline{\theta}_{ij} - \nu_j) = \frac{1}{n}. \tag{22}$$

Here, we denote the dual coordinate of $\overline{\boldsymbol{\pi}}$ by $\overline{\boldsymbol{\theta}}$.

In order to obtain the Lagrange multipliers μ_i and ν_j, we use the Newton-Raphson method. More specifically, we eploit the following functions:

$$f(\mu_i) = -\sum_{j=1}^{n} \psi'(\overline{\theta}_{ij} - \mu_i), \tag{23}$$

$$g(\nu_j) = -\sum_{i=1}^{m} \psi'(\overline{\theta}_{ij} - \nu_j). \tag{24}$$

These functions are defined on the open intervals $(\hat{\theta}_i - \theta_{\text{limit}}, +\infty)$ and $(\check{\theta}_j - \theta_{\text{limit}}, +\infty)$, where $0 < \theta_{\text{limit}} < +\infty$ is such that $\operatorname{dom}\psi = (-\infty, \theta_{\text{limit}})$, and $\hat{\theta}_i = \max\{\overline{\theta}_{ij}\}_{1\leq j\leq n}$, $\check{\theta}_j = \max\{\overline{\theta}_{ij}\}_{1\leq i\leq m}$. We can now obtain the unique solution to $f(\mu_i) = -\frac{1}{m}$ and $g(\nu_j) = -\frac{1}{n}$. Starting with $\mu_i = 0$ and $\nu_j = 0$, the Newton-Raphson updates:

$$\mu_i \leftarrow \mu_i + \frac{\sum_{j=1}^{n} \psi'(\overline{\theta}_{ij} - \mu_i) - \frac{1}{m}}{\sum_{j=1}^{n} \psi''(\overline{\theta}_{ij} - \mu_i)}, \tag{25}$$

$$\nu_j \leftarrow \nu_j + \frac{\sum_{i=1}^{m} \psi'(\overline{\theta}_{ij} - \nu_i) - \frac{1}{n}}{\sum_{i=1}^{m} \psi''(\overline{\theta}_{ij} - \nu_i)}, \tag{26}$$

converge to the optimal solution with a quadraitic rate. To avoid storing the intermediate Lagrange multipliers, the updates can be directly written in terms of the dual parameters:

$$\theta^*_{1,ij} \leftarrow \theta^*_{1,ij} - \frac{\sum_{j=1}^{n} \psi'(\theta^*_{1,ij}) - \frac{1}{m}}{\sum_{j=1}^{n} \psi''(\theta^*_{1,ij})}, \tag{27}$$

$$\theta^*_{2,ij} \leftarrow \theta^*_{2,ij} - \frac{\sum_{i=1}^{m} \psi'(\theta^*_{2,ij}) - \frac{1}{n}}{\sum_{i=1}^{m} \psi''(\theta^*_{2,ij})}, \tag{28}$$

after initilalization by $\theta^*_{1,ij} \leftarrow \overline{\theta}_{ij}$, $\theta^*_{2,ij} \leftarrow \overline{\theta}_{ij}$. Here, $\theta^*_{1,ij}$ and $\theta^*_{2,ij}$ are the ith row and jth column of $\boldsymbol{\theta}^*_1$ and $\boldsymbol{\theta}^*_2$ respectively. $\boldsymbol{\theta}^*_1$ and $\boldsymbol{\theta}^*_2$ are the dual coordinates of $\boldsymbol{\pi}^*_1$ and $\boldsymbol{\pi}^*_2$ respectively.

Table 1: Domain of Euclidean norm and β-potential ($\beta > 1$).

Regularization term	dom ϕ	dom ψ
β-potential ($\beta > 1$)	$\mathbb{R}_+$	$(\frac{1}{1-\beta}, \infty)$
Euclidean norm	$\mathbb{R}$	$\mathbb{R}$

From the above, starting from $\boldsymbol{\xi}$ and writing the successive vectors $\boldsymbol{\mu}^{(k)}$, $\boldsymbol{\nu}^{(k)}$ along iterations, we have:

$$
\begin{aligned}
\psi'(-\boldsymbol{\gamma}/\lambda) &\to \psi'\Big(\max\{\phi'(\mathbf{0}),\ -\boldsymbol{\gamma}/\lambda\}\Big) \\
&\to \psi'\Big(\max\{\phi'(\mathbf{0}),\ -\boldsymbol{\gamma}/\lambda\} - \boldsymbol{\mu}^{(1)}\mathbf{1}^\top\Big) \\
&\to \psi'\Big(\max\{\phi'(\mathbf{0}),\ -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(1)}\mathbf{1}^\top\ )\}\Big) \\
&\to \psi'\Big(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(1)}\mathbf{1}^\top)\} - \mathbf{1}\boldsymbol{\nu}^{(1)\top}\Big) \\
&\to \psi'\Big(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(1)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\nu}^{(1)\top}\}\Big) \\
&\to \psi'\Big(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(1)}\mathbf{1}^\top) - \mathbf{1}\boldsymbol{\nu}^{(1)\top}\} + \boldsymbol{\mu}^{(1)}\mathbf{1}^\top - \boldsymbol{\mu}^{(2)}\mathbf{1}^\top\Big) \\
&\to \psi'\Big(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(2)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\nu}^{(1)\top}\}\Big) \\
&\to \psi'\Big(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(2)}\mathbf{1}^\top) - \mathbf{1}\boldsymbol{\nu}^{(1)\top}\} + \mathbf{1}\boldsymbol{\nu}^{(1)\top} - \mathbf{1}\boldsymbol{\nu}^{(2)\top}\Big) \\
&\to \psi'\Big(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(2)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\nu}^{(2)\top}\}\Big) \\
&\to \cdots \\
&\to \psi'\Big(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(k)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\nu}^{(k)\top}\}\Big) \\
&\to \psi'\Big(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(k)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\nu}^{(k)\top}\} + \boldsymbol{\mu}^{(k)}\mathbf{1}^\top - \boldsymbol{\mu}^{(k+1)}\mathbf{1}^\top\Big) \\
&\to \psi'\Big(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(k+1)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\nu}^{(k)\top}\}\Big) \\
&\to \psi'\Big(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(k+1)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\nu}^{(k)\top}\} + \mathbf{1}\boldsymbol{\nu}^{(k)\top} - \mathbf{1}\boldsymbol{\nu}^{(k+1)\top}\Big) \\
&\to \psi'\Big(\max\{\phi'(\mathbf{0}), -\boldsymbol{\gamma}/\lambda - \boldsymbol{\mu}^{(k+1)}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\nu}^{(k+1)\top}\}\Big) \\
&\to \cdots \\
&\to \boldsymbol{\pi}^*.
\end{aligned}
$$

An efficent algorithm then exploits the differences $\boldsymbol{\tau}^{(k)} = \boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}^{(k-1)}$ and $\boldsymbol{\sigma}^{(k)} = \boldsymbol{\nu}^{(k)} - \boldsymbol{\nu}^{(k-1)}$ to scale the rows and columns (Algorithm).

A.5. The different point of our algorithm from NASA

An example of applying NASA algorithm is when the regulariler is the Euclidean norm $\phi(\pi) = \frac{1}{2}(\pi - 1)^2$. As it is shown in Table 1, we can easily confirm the Euclidean norm satisfies the three assumptions introduced in A.3. However, for the outlier-robust CROT, we use β-potential ($\beta > 1$) as the regularizer, which violates the third assumption, dom $\psi = \mathbb{R}$

(Table 1). Therefore, we cannot naively apply the NASA algorithm for the outlier-robust CROT. For instance, lines 2, 7, and 11 in our algorithm are not mathematically correct as projections onto $\mathcal{C}_0$. Similarly, lines 4–6 and 8–10 are not mathematically correct for projections onto $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively.

In spite of these mathematical issues, we still see lines 2, 7, and 11 in our algorithm as projections onto $\mathcal{C}_0$. In addition, since we cannot update the Newton-Raphson more than twice for projections onto $\mathcal{C}_1$ and $\mathcal{C}_2$ because $\boldsymbol{\theta}^* \in \operatorname{dom} \nabla\psi (= \operatorname{dom} \nabla\psi)$ is no longer guaranteed, we overcome this issue by only updating it once.

Appendix B. The proof of Proposition 1

Proposition 1. *For a given z ($> \frac{\lambda}{\beta-1}$), let $J \subseteq \{1, \dots, n\}$ be a subset of indices which satisfies the condition shown in Definition 2. Suppose we obtained a transport matrix $\boldsymbol{\pi}^{\text{output}}$ by running the alogrithm T times satisfying the following condition:*

$$T < \frac{\frac{z}{\lambda}(\beta-1)-1}{(\frac{1}{m})^{\beta-1} + (\frac{1}{n})^{\beta-1}}. \tag{29}$$

Then, $\boldsymbol{\pi}^{\text{output}}$ transports no mass to J.

Proof Before the algorithm starts,

$$-\frac{z}{\lambda} < \frac{1}{1-\beta} \tag{30}$$

holds. Since every element in $\boldsymbol{\theta}^*$ is greater than or equal to $\phi'(0) = \frac{1}{1-\beta}$, the following inequality holds for every i in the algorithm:

$$\begin{aligned} \tau_i &\geq \frac{1}{1-\beta} - \phi'\Big(\frac{1}{m}\Big) && (31)\\ &= \frac{1}{1-\beta} - \left(\frac{1}{\beta-1}\left(\left(\frac{1}{m}\right)^{\beta-1} - 1\right)\right) \\ &= -\frac{1}{\beta-1}\left(\frac{1}{m}\right)^{\beta-1}. && (32) \end{aligned}$$

Therefore,

$$-\tau_i \leq \frac{1}{\beta-1}\left(\frac{1}{m}\right)^{\beta-1}. \tag{33}$$

Similarly, for every j, the following inequality holds:

$$-\sigma_j \leq \frac{1}{\beta-1}\left(\frac{1}{n}\right)^{\beta-1}. \tag{34}$$

Therefore, if the algorithm finished running T times and the following inequality holds,

$$-\frac{z}{\lambda} + T \times \frac{1}{\beta-1}\left(\frac{1}{m}\right)^{\beta-1} + T \times \frac{1}{\beta-1}\left(\frac{1}{n}\right)^{\beta-1} < \frac{1}{1-\beta}, \tag{35}$$

then,

$$\forall i, \tilde{\theta}_{ij} < \frac{1}{1-\beta} \quad \text{if} \quad j \in J \tag{36}$$

$$\tag{37}$$

holds. Therefore,

$$\forall i, \boldsymbol{\pi}_{ij}^{\text{output}} = 0 \quad \text{if} \quad j \in J. \tag{38}$$

References

H. H. Bauschke and A. S. Lewis. Dual coordinate ascent methods for non-strictly convex minimization. *Mathematical Programming*, 59(1-3):231–247, 1993.

H. H. Bauschke and A. S. Lewis. Dykstra's algorithm with bregman projections: A convergence proof. *Optimization*, 48(4):409–427, 2000.

H.H. Bauschke and J.M. Borwein. Legendre functions and the method of random bregman projections. *Jounal of Convex Analysis*, 1997.

J.P. Boyle and R.L. Dykstra. A method for finding projections onto the intersection of convex sets in hilbert spaces. *Lecture Notes in Statistics.*, pages 28–47, 1986.

Arnaud Dessein, Nicolas Papadakis, and Jean-Luc Rouas. Regularized optimal transport and the rot mover's distance. *Journal of Machine Learning*, 25(10):2734–2775, 2018.

I.S. Dhillon and J.A. Tropp. Matrix nearness problems with bregman divergences. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1120–1146, 2007.

William Karush. Minima of functions of several variables with inequalities as side conditions. Master's thesis, Department of Mathematics, University of Chicago, Chicago, IL, USA, 1939.

H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, pages 481–492, Berkeley and Los Angeles, 1951. University of California Press.