Appendix for "Outlier-Robust Optimal Transport"

Debarghya Mukherjee [1] **Aritra Guha** [1] **Justin Solomon** [2] **Yuekai Sun** [1] **Mikhail Yurochkin** [3]

A. Proof of Theorem 3.1

In the proofs, $a \wedge b$ denotes $\min\{a, b\}$ for any $a, b \in \mathbb{R}$.

A.1. Proof of discrete version

Proof. Define a matrix Π as:

$$\Pi(i,j) = \begin{cases} 0, & \text{if } C(i,j) > 2\lambda \\ \Pi_2^*(i,j), & \text{otherwise} \end{cases}$$

Also define $s \in \mathbb{R}^n$ and $t \in \mathbb{R}^m$ as:

$$s_1^*(i) = -\sum_{j=1}^m \Pi_2^*(i,j) \mathbb{1}_{C(i,j) > 2\lambda}$$

and similarly define:

$$t_1^*(j) = \sum_{i=1}^n \Pi_2^*(i,j) \mathbb{1}_{C(i,j) > 2\lambda}$$

These vectors corresponds to the row sums and the column sums of the elements of the optimal transport plan of Formulation 2, where the cost function exceeds 2λ. Note that, these co-ordinates of the optimal transport plan corresponding to those co-ordinates of cost matrix, where the cost is greater than 2λ and contribute to the objective value via their sum only, hence any different arrangement of these transition probabilities with same sum gives the same objective value.

Now based on this Π obtained we construct a feasible solution of Formulation 1 following Algorithm 1:

$$\Pi_1^* = \begin{bmatrix} \mathbf{0} & \Pi \\ \mathbf{0} & \mathsf{diag}(t_1^*) \end{bmatrix}$$

The row sums of Π_1^* is:

$$\Pi_1^* \mathbf{1} = \begin{bmatrix} \mu_n + s_1^* \\ t_1^* \end{bmatrix}$$

and it is immediate from the construction that the column sums of Π_1^* is ν_m. Also as:

$$\sum_{i=1}^n s_1^*(i) = \sum_{j=1}^m t_1^*(j) = \sum_{(i,j):C_{i,j} > 2\lambda} \Pi_2^*(i,j)$$

and $s_1^* \preceq 0, t_1^* \succeq 0$, we have:

$$\mathbf{1}^\top(\mu_n + s_1^* + t_1^*) = \mathbf{1}^\top p = 1\,.$$

Therefore, we have (Π_1^*, s_1^*, t_1^*) is a feasible solution of Formulation 1. Now suppose this is not an optimal solution. Pick an optimal solution $\tilde{\Pi}, \tilde{s}, \tilde{t}$ of Formulation 1 so that:

$$\langle C_{aug}, \tilde{\Pi}\rangle + \lambda\left[\|\tilde{s}\|_1 + \|\tilde{t}\|_1\right] < \langle C_{aug}, \Pi_1^*\rangle + \lambda\left[\|s_1^*\|_1 + \|t_1^*\|_1\right]$$

The following two lemmas provide some structural properties of any optimal solution of Formulation 1:

Lemma A.1. *Suppose Π_1^*, s_1^*, t_1^* are optimal solution for Formulation 1. Divide Π_1^* into four parts corresponding to augmentation as in algorithm 1:*

$$\Pi_1^* = \begin{bmatrix} \Pi_{1,11}^* & \Pi_{1,12}^* \\ \Pi_{1,21}^* & \Pi_{1,22}^* \end{bmatrix}$$

Then we have $\Pi_{1,11}^* = \Pi_{1,21}^* = \mathbf{0}$ and $\Pi_{1,22}^*$ is a diagonal matrix.

Lemma A.2. *If Π_1^*, s_1^*, t_1^* is an optimal solution of Formulation 1 then:*

1. *If $C_{i,j} > 2\lambda$ then $\Pi_1^*(i,j) = 0$.*
2. *If $C_{i,j} < 2\lambda$ for some i and for all $1 \le j \le n$, then $s_1^*(i) = 0$.*
3. *If $C_{i,j} < 2\lambda$ for some j and for all $1 \le i \le m$, then $t_1^*(j) = 0$.*
4. *If $C_{i,j} < 2\lambda$ then $s_1^*(i) t_1^*(j) = 0$.*

We provide the proofs in the next subsection. By Lemma A.1 we can assume without loss of generality:

$$\tilde{\Pi} = \begin{bmatrix} \mathbf{0} & \tilde{\Pi}_{12} \\ \mathbf{0} & \mathsf{diag}(\tilde{t}) \end{bmatrix}$$

[1] Department of Statistics, University of Michigan [2] MIT CSAIL, MIT-IBM Watson AI Lab [3] IBM Research, MIT-IBM Watson AI Lab. Correspondence to: Debarghya Mukherjee <mdeb@umich.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

Now based on $\left(\tilde{\Pi}, \tilde{s}, \tilde{t}\right)$ we create a feasible solution namely $\Pi^*_{2,new}$ of Formulation 2 as follows: Define the set of indices $\{i_1, \cdots, i_k\}$ and $\{j_1, \ldots, j_l\}$ as:

$$\tilde{s}_{i_1}, \tilde{s}_{i_2}, \ldots, \tilde{s}_{i_k} > 0 \quad \text{and} \quad \tilde{t}_{j_1}, \tilde{t}_{j_2}, \ldots, \tilde{t}_{j_l} > 0 \,.$$

Then by part (4) of Lemma A.2 we have $C_{i_\alpha, j_\beta} > 2\lambda$ for $\alpha \in \{1, \ldots, k\}$ and $\beta \in \{1, \ldots, l\}$. Also by part (2) of Lemma A.2 the value of transport plan at these co-ordinates is 0. Now distribute the mass of slack variables in these co-ordinates such that the marginals of new transport plan becomes exactly μ_n and ν_m. This new transport plan is our $\Pi^*_{2,new}$. Recall that, $\|\tilde{s}\|_1 = \|\tilde{t}\|_1$. Hence, here the regularizer value decreases by $2\lambda\|\tilde{s}\|_1$ and the cost value increased by exactly $2\lambda\|\tilde{s}\|_1$ as we are truncating the cost. Hence we have:

$$\begin{aligned} \langle C_\lambda, \Pi^*_{2,new} \rangle &= \langle C_{aug}, \tilde{\Pi} \rangle + \lambda \left[\|\tilde{s}\|_1 + \|\tilde{t}\|_1 \right] \\ &< \langle C_{aug}, \Pi^*_1 \rangle + \lambda \left[\|s^*_1\|_1 + \|t^*_1\|_1 \right] \\ &= \langle C_\lambda, \Pi^*_2 \rangle \end{aligned}$$

which is contradiction as Π^*_2 is the optimal solution of Formulation 2. This completes the proof for the discrete part.

□

A.2. Proof of equivalence for two sided formulation

Here we prove that our two sided formulation, i.e. Formulation 3 (equation 2.8) is equivalent to Formulation 1 (equation 2.6) for the discrete case. Towards that end, we introduce another auxiliary formulation and show that both Formulation 1 and Formulation 3 are equivalent to the following auxiliary formulation of the problem.

Formulation 4:

$$\begin{array}{ll} \min_{\Pi \in \mathbb{R}^{m \times n}, s_1 \in \mathbb{R}^m, s_2 \in \mathbb{R}^n} & \langle C, \Pi \rangle + \lambda \left[\|s_1\|_1 + \|s_2\|_1 \right] \\ \text{subject to} & \Pi 1_n = p + s_1 \\ & \Pi^T 1_m = q + s_2 \\ & \Pi \succeq 0 \end{array} . \tag{A.1}$$

First we show that Formulation 1 and Formulation 4 are equivalent in a sense that they have the same optimal objective value.

Theorem A.3. *Suppose C is a cost function such that $C(x, x) = 0$. Then Formulation 1 and Formulation 4 has same optimal objective value.*

Proof. Towards that end, we show that given one optimal variables of one formulation we can get optimal variables of other formulation with the same objective value. Before going into details we need the following lemma whose proof is provided in Appendix B:

Lemma A.4. *Suppose $\Pi^*_4, s^*_{4,1}, s^*_{4,2}$ are the optimal variables of Formulation 4. Then $s^*_{4,1} \preceq 0$ and $s^*_{4,2} \preceq 0$.*

Now we prove that optimal value of Formulation 1 and Formulation 4 are same. Let $(\Pi^*_1, s^*_{1,1}, t^*_{1,1})$ is an optimal solution of Formulation 1. Then we claim that $(\Pi^*_1, s^*_{1,1}, t^*_{1,1})$ is also an optimal solution of Formulation 4. Clearly it is feasible solution of Formulation 4. Suppose it is not optimal, i.e. there exists another optimal solution $(\tilde{\Pi}_4, \tilde{s}_{4,1}, \tilde{s}_{4,2})$ such that:

$$\langle C, \tilde{\Pi}_4 \rangle + \lambda(\|\tilde{s}_{4,1}\|_1 + \|\tilde{s}_{4,2}\|_2) < \langle C, \Pi^*_{1,12} \rangle + \lambda(\|s^*_{1,1}\|_1 + \|t^*_{1,1}\|_1)$$

Now based on $(\tilde{\Pi}_4, \tilde{s}_{4,1}, \tilde{s}_{4,2})$ we construct a feasible solution of Formulation 1 as follows:

$$\tilde{\Pi}_1 = \begin{bmatrix} \mathbf{0} & \tilde{\Pi}_4 \\ \mathbf{0} & -\mathsf{diag}(\tilde{s}_{4,2}) \end{bmatrix}$$

Note that we proved in Lemma A.4 $\tilde{s}_{4,2} \preceq 0$, hence we have $\tilde{\Pi}_1 \succeq 0$. Now as the column sums of $\tilde{\Pi}_4$ is $q + \tilde{s}_{4,2}$, we have column sums of $\tilde{\Pi}_1 = [\mathbf{0} \;\; q^\top]^\top$ and the row sums are $[(p + \tilde{s}_{4,1})^\top \quad \tilde{s}^\top_{4,2}]^\top$. Hence we take $\tilde{s}_{1,1} = \tilde{s}_{4,1}$ and $\tilde{s}_{1,2} = \tilde{s}_{4,2}$. Then it follows:

$$\begin{aligned} & \langle C_{aug}, \tilde{\Pi}_1 \rangle + \lambda \left[\|\tilde{s}_{1,1}\|_1 + \|\tilde{s}_{1,2}\|_1 \right] \\ & = \langle C, \tilde{\Pi}_4 \rangle + \lambda \left[\|\tilde{s}_{4,1}\|_1 + \|\tilde{s}_{4,2}\|_1 \right] \\ & < \langle C, \Pi^*_{1,12} \rangle + \lambda \left[\|s^*_{1,1}\|_1 + \|t^*_{1,1}\|_1 \right] \\ & = \langle C_{aug}, \Pi^*_1 \rangle + \lambda \left[\|s^*_{1,1}\|_1 + \|t^*_{1,1}\|_1 \right] \end{aligned}$$

This is contradiction as we assumed $(\Pi^*_1, s^*_{1,1}, t^*_{1,2})$ is an optimal solution of Formulation 1. Therefore we conclude $(\Pi^*_1, s^*_{1,1}, t^*_{1,1})$ is also an optimal solution of Formulation 4 which further concludes Formulation 1 and Formulation 4 have same optimal values. This completes the proof of the theorem. □

Theorem A.5. *The optimal objective value of Formulation 3 and Formulation 4 are same.*

Proof. Like in the proof of Theorem A.3 we also prove couple of lemmas.

Lemma A.6. *Any optimal transport plan Π^*_3 of Formulation 3 has the following structure: If we write,*

$$\Pi^*_3 = \begin{bmatrix} \Pi^*_{3,11} & \Pi^*_{3,12} \\ \Pi^*_{3,21} & \Pi^*_{3,22} \end{bmatrix}$$

then $\Pi^*_{3,11}$ and $\Pi^*_{3,22}$ are diagonal matrices and $\Pi^*_{3,21} = \mathbf{0}$.

Lemma A.7. *If $s^*_{3,1}, t^*_{3,1}, s^*_{3,2}, t^*_{3,2}$ are four optimal slack variables in Formulation 3, then $s^*_{3,1}, t^*_{3,1} \preceq 0$ and $s^*_{3,2}, t^*_{3,2} \succeq 0$.*

Proof. The line of argument is same as in proof of Lemma A.4. □

Next we establish equivalence. Suppose $(\Pi_3^*, s_{3,1}^*, t_{3,1}^*, s_{3,2}^*, t_{3,2}^*)$ are optimal values of Formulation 3. We claim that $(\Pi_{3,12}^*, s_{3,1}^* - s_{3,2}^*, t_{3,1}^* - t_{3,2}^*)$ forms an optimal solution of Formulation 4. The objective value will then also be same as $s_{3,1}^* \preceq 0, s_{3,2}^* \succeq 0$ (Lemma A.7) implies $\|s_{3,1}^* - s_{3,2}^*\|_1 = \|s_{3,1}^*\|_1 + \|s_{3,2}^*\|_1$ and similarly $t_{3,1}^* \preceq 0, t_{3,2}^* \succeq 0$ implies $\|t_{3,1}^* - t_{3,2}^*\|_1 = \|t_{3,1}^*\|_1 + \|t_{3,2}^*\|_1$. Feasibility is immediate. Now for optimality, we again prove by contradiction. Suppose they are not optimal. Then lets say $\tilde{\Pi}_4, \tilde{s}_{4,1}, \tilde{s}_{4,2}$ are an optimal triplet of Formulation 4. Now construct another feasible solution of Formulation 3 as follows: Set $\tilde{s}_{3,2} = \tilde{t}_{3,2} = 0, \tilde{s}_{3,1} = \tilde{s}_{4,1}$ and $\tilde{t}_{3,1} = \tilde{s}_{4,2}$. Set the matrix as:

$$\tilde{\Pi}_3 = \begin{bmatrix} \mathbf{0} & \tilde{\Pi}_4 \\ \mathbf{0} & -\mathsf{diag}(\tilde{s}_{4,2}) \end{bmatrix}$$

Then it follows that $\left(\tilde{\Pi}_3, \tilde{s}_{3,1}, \tilde{s}_{3,2}, \tilde{t}_{3,1}, \tilde{t}_{3,2}\right)$ is a feasible solution of Formulation 3. Finally we have:

$$\begin{aligned}
&\langle C_{aug}, \tilde{\Pi}_3\rangle + \lambda\left[\|\tilde{s}_{3,1}\|_1 + \|\tilde{s}_{3,2}\|_1 + \|\tilde{t}_{3,1}\|_1 + \|\tilde{t}_{3,2}\|_1\right] \\
&= \langle C_{aug}, \tilde{\Pi}_3\rangle + \lambda\left[\|\tilde{s}_{4,1}\|_1 + \|\tilde{s}_{4,2}\|_1\right] \\
&= \langle C, \tilde{\Pi}_4\rangle + \lambda\left[\|\tilde{s}_{4,1}\|_1 + \|\tilde{s}_{4,2}\|_1\right] \\
&< \langle C, \Pi_{3,12}^*\rangle + \lambda\left[\|s_{3,1}^* - s_{3,2}^*\|_1 + \|t_{3,1}^* - t_{3,2}^*\|_1\right] \\
&= \langle C_{aug}, \Pi_3^*\rangle + \lambda\left[\|s_{3,1}^*\|_1 + \|s_{3,2}^*\|_1 + \|t_{3,1}^*\|_1 + \|t_{3,2}^*\|_1\right]
\end{aligned}$$

This contradicts the optimality of $(\Pi_3^*, s_{3,1}^*, s_{3,2}^*, t_{3,1}^*, t_{3,2}^*)$. This completes the proof. □

A.3. Proof of continuous version

Proof. In this proof we denote by F_1 the optimization problem of equation 2.3 and by F_2 the optimization problem equation 2.5. Let μ, ν be two absolutely continuous measures on $\mathbb{R}^d$. Moreover, we assume $c(x, y) = \|x - y\|$ for some norm $\|\cdot\|$ on $\mathbb{R}^d$. We assume that $\int \|x\|\nu(\mathrm{d}x), \int \|x\|\mu(\mathrm{d}x) < \infty$.

Step 1: Let K_ϵ be a compact set such that $\int_{K_\epsilon} \|x\|\mu(\mathrm{d}x), \int_{K_\epsilon} \|x\|\nu(\mathrm{d}x) > 1 - \epsilon$.

Also, let $\tilde{K}_\epsilon = \{x_1, \ldots, x_{n_\epsilon}\}$ be a maximal ϵ-packing set of K_ϵ. Starting from $\tilde{K}_\epsilon$, define $\{S_1, \ldots, S_{n_\epsilon}\}$ as a mutually disjoint covering of K_ϵ with internal points $x_1, \ldots, x_{n_\epsilon}$ respectively, so that $\mathrm{Diam}(S_i) \le 2\epsilon$. With $p_i = \int_{S_i} \mu(\mathrm{d}x)$, $q_i = \int_{S_i} \nu(\mathrm{d}x)$ for $i = 1, \ldots, n_\epsilon$, $p_0 = \int_{K_\epsilon^C} \mu(\mathrm{d}x)$, $q_0 = \int_{K_\epsilon^C} \nu(\mathrm{d}x)$ and $x_0 = 0 \in \mathbb{R}^d$, define

$$\begin{aligned}
\mu_\epsilon &= \sum_0^{n_\epsilon} p_i \delta_{x_i} \\
\nu_\epsilon &= \sum_0^{n_\epsilon} q_i \delta_{x_i}
\end{aligned}$$

A coupling Q between two probability distributions is a joint distribution with marginals as the given two distributions. The Wasserstein distance between two distributions P_1 and P_2 is defined as:

$$W_1(P_1, P_2) = \inf_{Q \in \mathscr{Q}(P_1, P_2)} \int Q(x, y)\|x - y\| \mathrm{d}x\mathrm{d}y, \tag{A.2}$$

where $\mathscr{Q}(P_1, P_2)$ is the collection of all couplings of P_1 and P_2.

Define $Q(x, y) = (\mathbb{1}_{x=x_0, y \in K_\epsilon^C} + \sum_{i=1}^{n_\epsilon} \mathbb{1}_{x=x_i, y \in S_i})\mu(\mathrm{d}y)$. Then Q is a coupling between μ and μ_ϵ. Therefore, clearly,

$$W_1(\mu, \mu_\epsilon) \le \int_{K_\epsilon^C} \|x\|\mu(\mathrm{d}x) + 2\epsilon\left(\sum_{i=1}^{n_\epsilon} p_i\right) \le 3\epsilon \tag{A.3}$$

Similarly, $W_1(\nu, \nu_\epsilon) \le 3\epsilon$. Therefore $\lim_{\epsilon \to 0} W_1(\nu, \nu_\epsilon) = 0$.

Moreover, $W_1(\mu, \nu) = \lim_{\epsilon \to 0} W_1(\mu_\epsilon, \nu_\epsilon)$, as $W_1(\mu_\epsilon, \nu_\epsilon) - 6\epsilon \le W_1(\mu, \nu) \le W_1(\mu_\epsilon, \nu_\epsilon) + 6\epsilon$ by triangle inequality.

Step 2: Let S be an arbitrary measure with $\|S\|_{\mathrm{TV}} = 2\gamma$, so that $\mu + S$ is a probability measure with $\int \|x\|(\mu + S)(\mathrm{d}x) < \infty$. Also, let us define $\epsilon_n = 2^{-(n+1)}$.

Let $S = S^+ - S^-$, where S^+ and S^- are positive measures on $\mathbb{R}^d$. Then, $\|S^-\|_{\mathrm{TV}} = \|S^+\|_{\mathrm{TV}} = \gamma$.

Clearly $(\mu - S^-)/(1 - \gamma), \mu, \nu, S^+/\gamma$ are tight probability measures. So we can construct compact sets $K_{\epsilon_n}^{(1)}$, similar to Step 1 to approximate all the four measures. Without loss of generality we assume that $0 \in K_{\epsilon_n}^{(1)}$ for all n. Moreover, we can also construct approximate measures $(\mu - S^-)_n = ((\mu - S^-)/(1 - \gamma))_{\epsilon_n}$ and $(S^+)_n = (S^+/\gamma)_{\epsilon_n}$ defined as in Step 1. $\mu_n = \mu_{\epsilon_n}, \nu_n = \nu_{\epsilon_n}$ are defined similarly. All four of the measures have support points in $K_{\epsilon_n}^{(1)}$.

Next, we define $(\mu + S)_n = \gamma(S^+)_n + (1 - \gamma)(\mu - S^-)_n$. Then by the construction, from (Villani, 2009), $\lim_{n \to \infty} W_1((\mu + S)_n, \mu + S) \to 0$ and thus $\lim_{n \to \infty} W_1((\mu + S)_n, \nu_n) \to W_1(\mu + S, \nu)$. Therefore we can define a signed measure $S_n = (\mu + S)_n - \mu_n$. Moreover,

$$\begin{aligned}
\|S_n\|_{\mathrm{TV}} &\le \gamma\|(S^+)_n\|_{\mathrm{TV}} + \|(1 - \gamma)(\mu - S^-)_n - \mu_n\|_{\mathrm{TV}} \qquad \text{(A.4)} \\
&= 2\gamma = \|S\|_{\mathrm{TV}} \qquad \text{(A.5)}
\end{aligned}$$

Note that $\mu_n, \nu_n, (\mu + S)_n$ put masses (sometimes zero masses) on a common set of support points given by $\tilde{K}_{\epsilon_n}^{(1)} \subset K_{\epsilon_n}^{(1)}$.

The $\tilde{K}^{(1)}_{\epsilon_n}$ is sequentially defined so that $\tilde{K}^{(1)}_{\epsilon_{n+1}}$ is a refinement of $\tilde{K}^{(1)}_{\epsilon_n}$. This can easily be achieved by the choice of ϵ_n defined.

Consider $\tilde{s}_n, \Pi_n$ such that

$$F_1(\mu_n, \nu_n) = \int \|x-y\|\Pi_n(\mathrm{d}x\mathrm{d}y) + \lambda\|\tilde{s}_n\|_{\mathrm{TV}} \quad \text{(A.6)}$$

By the discrete nature of μ_n, ν_n, using the proof of the discrete part $F_1(\mu_n, \nu_n) = F_2(\mu_n, \nu_n)$.

Since, $\min\{\|x-y\|, 2\lambda\}$ is a metric, whenever $\|x-y\|$ is, therefore, it is easy to check that $F_2(\mu,\nu) = \lim_n F_2(\mu_n,\nu_n) = \lim_n F_1(\mu_n,\nu_n)$.

Moreover, by construction, $F_1(\mu_n,\nu_n) \leq \int \|x-y\|\Pi(\mathrm{d}x\mathrm{d}y) + \lambda\|S_n\|_{\mathrm{TV}}$ for any arbitrary coupling Π of μ and $\mu+S$. Also $\lim_n W_1(\mu_n,\mu), W_1(\mu+S,(\mu+S)_n) \to 0$.

Thus, combining the above result with equation A.4, we get

$$\lim_n F_1(\mu_n,\nu_n) \leq \int \|x-y\|\tilde{\Pi}(\mathrm{d}x\mathrm{d}y) + \lambda\|S\|_{\mathrm{TV}}$$

for any coupling $\tilde{\Pi}$ of μ and $\mu+S$.

Therefore, $F_2(\mu,\nu) \leq F_1(\mu,\nu)$.

Step 3: Consider $\tilde{s}_n$ defined in equation A.6. As $\tilde{s}_n$ has support in the compact sets $K^{(1)}_{\epsilon_n}$ defined in Step 2, therefore, $\{\mu_n + \tilde{s}_n\}_{n\geq 1}$ are tight measures.

Therefore, by Prokhorov's Theorem for equivalence of sequential compactness and tightness for a collection of measures, there exists a probability measure $\mu \oplus s$ and a subsequence $\{n_k\}_{k\geq 1}$ such that $\mu_{n_k} + \tilde{s}_{n_k}$ converges weakly to $\mu \oplus s$. Moreover, by construction $\lim_{R\to\infty}\limsup_{n\to\infty}\int_{\|x\|>R}\|x\|(\mu_n+\nu_n)(\mathrm{d}x) = 0$ and so $\lim_{R\to\infty}\limsup_{n\to\infty}\int_{\|x\|>R}\|x\|(\mu_n+\tilde{s}_n)(\mathrm{d}x) = 0$.

Thus, by Definition 6.8 part (iii) and Theorem 6.9 of (Villani, 2009), $W_1(\mu_{n_k}+\tilde{s}_{n_k}, \mu\oplus s) \to 0$. Moreover, $W_1(\mu_{n_k},\mu) \to 0$. Therefore $\|\tilde{s}_{n_k}\|_{\mathrm{TV}} \to \|\mu\oplus s - \mu\|_{\mathrm{TV}}$. Thus, $W_1(\mu_{n_k}+\tilde{s}_{n_k},\nu_{n_k}) + \lambda\|\tilde{s}_{n_k}\|_{\mathrm{TV}} \to W_1(\mu\oplus s,\nu) + \lambda\|\mu\oplus s-\mu\|_{\mathrm{TV}}$. But by the proof of the discrete part, $W_1(\mu_{n_k}+\tilde{s}_{n_k},\nu_{n_k}) + \lambda\|\tilde{s}_{n_k}\|_{\mathrm{TV}} = F_1(\mu_{n_k},\nu_{n_k}) = F_2(\mu_{n_k},\nu_{n_k}) \to F_2(\mu,\nu)$. Therefore, with $s = \mu\oplus s - \mu$, $W_1(\mu+s,\nu) + \lambda\|s\|_{\mathrm{TV}} = F_2(\mu,\nu)$.

Therefore, $F_2(\mu,\nu) = \limsup_{n\to\infty} F_1(\mu_n,\nu_n) \geq F_1(\mu,\nu)$. Thus the equality holds.

□

B. Proof of Theorem 2.1

Proof. The proof is immediate from the Formulation 1. Recall that the Formulation 1 can restructured as:

$$\mathrm{ROBOT}(\tilde{\mu},\nu) = \inf_P \left\{\mathrm{OT}(P,\nu) + \lambda\|P-\tilde{\mu}\|_{\mathrm{TV}}\right\} .$$

where the infimum is taking over all measure dominated by some common measure σ (with respect to which μ, μ_c, ν are dominated). Hence,

$$\mathrm{ROBOT}(\tilde{\mu},\nu) \leq \mathrm{OT}(P,\nu) + \lambda\|P-\tilde{\mu}\|_{\mathrm{TV}}$$

for any particular choice of P. Taking $P=\mu$ we get that

$$\begin{aligned}\mathrm{ROBOT}(\tilde{\mu},\nu) &\leq \mathrm{OT}(\mu,\nu) + \lambda\|\mu-\tilde{\mu}\|_{\mathrm{TV}} \\ &= \mathrm{OT}(\mu,\nu) + \lambda\epsilon\|\mu-\mu_c\|_{\mathrm{TV}}\end{aligned}$$

Taking $P=\nu$ we get $\mathrm{ROBOT}(\tilde{\mu},\nu) \leq \lambda\|\nu-\tilde{\mu}\|_{\mathrm{TV}}$ and finally taking $P=\tilde{\mu}$ we get $\mathrm{ROBOT}(\tilde{\mu},\nu) \leq \mathrm{OT}(\tilde{\mu},\nu)$. This completes the proof. □

C. Proof of Lemma 3.2

As defined in the main text, let Π_2^* be the optimal solution of equation 2.7 and $\Pi_{2,\alpha}^*$ be the optimal solution of equation 3.1. Then by Proposition 4.1 from Peyré & Cuturi (2018) we conclude:

$$\Pi_{2,\alpha}^* \overset{\alpha\to 0}{\longrightarrow} \Pi_2^* . \quad \text{(C.1)}$$

Now we have defined $\left(\Pi_{1,\alpha}^*, \mathbf{s}_{1,\alpha}^*\right)$ as the *approximate* solution of equation 2.6 obtained via Algorithm 1 from $\Pi_{2,\alpha}^*$. Note that we can think of Algorithm 1 as a map from $\mathbb{R}^{m\times n}$ to $\mathbb{R}^{(m+n)\times(m+n)} \times \mathbb{R}^m$. Define this map as F.

$$F(\Pi_2) \mapsto (\Pi_1, \mathbf{s}_1)$$

Hence, by our notation, $\left(\Pi_{1,\alpha}^*, \mathbf{s}_{1,\alpha}^*\right) = F(\Pi_{2,\alpha}^*)$ and $(\Pi_1^*, \mathbf{s}_1^*) = F(\Pi_2^*)$. Now if we show that F is a continuous map, then by continuous mapping theorem, it is also immediate from equation C.1 that:

$$F(\Pi_{2,\alpha}^*) \overset{\alpha\to 0}{\longrightarrow} F(\Pi_2^*) .$$

which implies:

$$\begin{aligned}\Pi_{1,\alpha}^* &\overset{\alpha\to 0}{\longrightarrow} \Pi_1^* \\ \mathbf{s}_{1,\alpha}^* &\overset{\alpha\to 0}{\longrightarrow} \mathbf{s}_1^* .\end{aligned}$$

which will complete the proof. Therefore all we need to show is that F is a continuous map. Towards that direction, first fix a sequence of matrices $\{\bar{\Pi}_{2,i}\}_{i\in\mathbb{N}} \to \bar{\Pi}_2$. Define $F(\bar{\Pi}_{2,i}) = \left(\bar{\Pi}_{1,i}, \bar{\mathbf{s}}_{1,i}\right)$ and $F(\bar{\Pi}_2) = \left(\bar{\Pi}_1, \bar{\mathbf{s}}_1\right)$. By Step 3 - Step 5 of Algorithm 1, we obtain $\bar{\Pi}_{1,i}$ by first setting $\bar{\Pi}_{1,i,12} = \bar{\Pi}_{2,i}$ and for each of the columns of $\bar{\Pi}_{1,i,12}$, dumping the sum of its entries for which the cost is $> 2\lambda$ to the diagonals of $\bar{\Pi}_{1,i,22}$. Also, we have all the entries of the first n columns of $\bar{\Pi}_{1,i}$ to be 0. In step 6 of Algorithm 1, we obtain $\mathbf{s}_{1,i}$ by taking the negative of the sum of the elements of each rows of $\bar{\Pi}_{1,i,12}$ for which the cost is $> 2\lambda$. Note that these operations (Step 3 - Step 6 of Algorithm 1) are continuous. Therefore we conclude:

1. $0 = \bar{\Pi}_{1,i,11} \to \bar{\Pi}_{1,11} = 0\,.$
2. $0 = \bar{\Pi}_{1,i,21} \to \bar{\Pi}_{1,21} = 0\,.$
3. $\bar{\Pi}_{1,i,12} = \bar{\Pi}_{2,i} \odot \mathbb{1}_{\mathcal{I}^c} \to \bar{\Pi}_2 \odot \mathbb{1}_{\mathcal{I}^c} = \bar{\Pi}_{1,12}\,.$
4.
$$\begin{aligned}\bar{\Pi}_{1,i,22} &= \mathsf{diag}\left(\mathbf{1}^\top\left(\bar{\Pi}_{2,i} \odot \mathbb{1}_{\mathcal{I}}\right)\right)\\ &\to \mathsf{diag}\left(\mathbf{1}^\top\left(\bar{\Pi}_{2} \odot \mathbb{1}_{\mathcal{I}}\right)\right)\\ &= \bar{\Pi}_{1,22}\,.\end{aligned}$$
5.
$$\begin{aligned}\mathbf{s}_{1,i} &= -\left(\bar{\Pi}_{i,n} \odot \mathbb{1}_{\mathcal{I}}\right)\mathbf{1}\\ &\to -\left(\bar{\Pi}_{2} \odot \mathbb{1}_{\mathcal{I}}\right)\mathbf{1} = \mathbf{s}_1\,.\end{aligned}$$

where $A \odot B$ denotes the Hadamard product (element-wise multiplication) between two matrices. Hence we have established:

$$\begin{aligned}F(\bar{\Pi}_{2,i}) &= \left(\bar{\Pi}_{1,i}, \bar{\mathbf{s}}_{1,i}\right)\\ &\overset{n\to\infty}{\longrightarrow} \left(\bar{\Pi}_1, \bar{\mathbf{s}}_1\right)\\ &= F(\bar{\Pi}_2)\,.\end{aligned}$$

This completes the proof of continuity of F.

D. Proof of auxiliary lemmas

D.1. Proof of Lemma A.1

Proof. The fact that $\Pi^*_{1,11} = \Pi^*_{1,21} = \mathbf{0}$ follows from the fact that $\Pi^*_1 \succeq 0$ and $\Pi^*_1\mathbf{1} = \mathbf{Q}$. To prove that $\Pi^*_{1,22}$ is diagonal, we use the fact that the any diagonal entry the cost matrix is 0. Now suppose $\Pi^*_{1,22}$ is not diagonal. Then define a matrix $\hat{\Pi}$ as following: set $\hat{\Pi}_{11} = \hat{\Pi}_{21} = \mathbf{0}$, $\hat{\Pi}_{12} = \Pi^*_{1,12}$ and:

$$\hat{\Pi}_{22}(i,j) = \begin{cases}\sum_{k=1}^m \Pi^*_{1,22}(k,i), & \text{if } j = i\\ 0, & \text{if } j \neq i\end{cases}$$

Also define $\hat{s} = s^*_1$ and $\hat{t}$ as $\hat{t}(i) = \hat{\Pi}_{22}(i,i)$. Then clearly $(\hat{\Pi}, \hat{s}, \hat{t})$ is a feasible solution of Formulation 1. Note that:

$$\|\hat{t}\|_1 = 1^\top \hat{\Pi}_{22} 1 = 1^\top \Pi^*_{1,22} 1 = \|t^*_1\|_1$$

and by our construction $\langle C_{aug}, \hat{\Pi}\rangle < \langle C_{aug}, \Pi^*_1\rangle$. Hence $(\hat{\Pi}, \hat{s}, \hat{t})$ reduces the value of the objective function of Formulation 1 which is a contradiction. This completes the proof. □

D.2. Proof of Lemma A.2

Proof. 1. Suppose $\Pi^*_1(i,j) > 0$. Then dump this mass to $s^*_1(j)$ and make it 0. In this way $\langle C_{aug}, \Pi^*_1\rangle$ will decrease by $> 2\lambda\Pi^*_1(i,j)$ and the regularizer value will increase by atmost $2\lambda\Pi^*_1(i,j)$, resulting in overall reduction in the objective value, which leads to a contradiction.

2. Suppose each entry of i^{th} row of C is $< 2\lambda$. Then if $s^*_1(i) > 0$, we can distribute this mass in the i^{th} row such that, $s^*_1(i) = a_1 + a_2 + \cdots + a_m$ with the condition that $t^*_1(j) \geq a_j$. Now we reduce t^*_1 as:
$$t^*_1(j) \leftarrow t^*_1(j) - a_j$$
Hence the value $\langle C_{aug}, \Pi^*_1(i,j)\rangle$ will increase by a value $< 2\lambda s^*_1(i)$ but the value of regularizer will decrease by the value of $2\lambda s^*_1(i)$, resulting in overall decrease in the value of objective function.
3. Same as proof of part (2) by interchanging row and column in the argument.
4. Suppose not. Then choose $\epsilon < s^*_1(i) \wedge t^*_1(j)$, Add ϵ to $\Pi^*_1(i,j)$. Hence the cost function value $\langle C_{aug}, \Pi^*_1\rangle$ will increase by $< 2\lambda\epsilon$ but the regularizer value will decrease by $2\lambda\epsilon$, resulting in overall decrease in the objective function.

□

D.3. Proof of Lemma A.4

Proof. For the notational simplicity, we drop the subscript 4 now as we will only deal with the solution of Formulation 4 and there will be no ambiguity. We prove the Lemma by contradiction. Suppose $s^*_{1,i} > 0$. Then we show one can come up with another solution $(\tilde{\Pi}, \tilde{s}_1, \tilde{s}_2)$ of Formulation 4 such that it has lower objective value. To construct this new solution, make:

$$\tilde{s}_{1,j} = \begin{cases} s^*_{1,j}, & \text{if } j \neq i\\ 0, & \text{if } j = i\end{cases}$$

Now to change the optimal transport plan, we will only change i^{th} row of Π^*. We subtract $a_1, a_2, \dots, a_n \geq 0$ from i^{th} column of Π^* in such a way, such that none of the elements are negative. Hence the column sum will be change, i.e. the value of $\tilde{s}_2$ will be:

$$\tilde{s}_{2,j} = s^*_{2,j} - a_j \quad \forall 1 \leq j \leq n\,.$$

Now clearly from our construction:

$$\langle C, \tilde{\Pi}\rangle \leq \langle C, \Pi^*\rangle$$

For the regularization part, note that, as we only reduced i^{th} element of s^*_1, we have $\|\tilde{s}_1\|_1 = \|s^*_1\|_1 - s^*_{1,i}$. And by simple triangle inequality,

$$\|\tilde{s}_2\|_1 \leq \|s^*_2\|_1 + \|a_1\|_1 = \|s^*_2\|_1 + s^*_{1,i}$$

by construction a_i's, as $a_i \geq 0$ and $\sum_i a_i = s^*_{1,i}$. Hence we have:

$$\|\tilde{s}_1\|_1 + \|\tilde{s}_2\|_1 \leq \|s^*_1\|_1 - s^*_{1,i} + \|s^*_2\|_1 + s^*_{1,i} = \|s^*_1\|_1 + \|s^*_2\|_1 \, .$$

Hence the value corresponding to regularizer will also decrease. This completes the proof. □

D.4. Proof of Lemma A.6

Proof. We prove this lemma by contradiction. Suppose Π^*_3 does not have the structure mentioned in the statement of Lemma. Construct another transport plan for Formulation 3 $\tilde{\Pi}_3$ as follows: Keep $\tilde{\Pi}_{3,12} = \Pi^*_{3,12}$ and set $\tilde{\Pi}_{3,12} = \mathbf{0}$. Construct the other parts as:

$$\tilde{\Pi}_{3,11}(i,j) = \\ \begin{cases} \sum_{k=1}^m \Pi^*_{3,11}(i,k) + \sum_{k=1}^n \Pi^*_{3,21}(k,i), & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}$$

and

$$\tilde{\Pi}_{3,22}(i,j) = \begin{cases} \sum_{k=1}^n \Pi^*_{3,22}(k,i), & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}$$

It is immediate from the construction that:

$$\langle C_{aug}, \tilde{\Pi}_3 \rangle \leq \langle C_{aug}, \Pi^*_3 \rangle$$

As for the regularization term: Note the by our construction $\tilde{s}_4$ will be same as s^*_4 as column sum of $\tilde{\Pi}_{3,22}$ is same as $\Pi^*_{3,22}$. For the other three:

$$\tilde{s}_3(i) = \tilde{\Pi}_{3,11}(i,i) = \sum_{k=1}^m \Pi^*_{3,11}(i,k) + \sum_{k=1}^n \Pi^*_{3,21}(k,i)$$

$$\tilde{s}_2(i) = \tilde{\Pi}_{3,22}(i,i) = \sum_{k=1}^n \Pi^*_{3,22}(k,i)$$

and hence by construction:

$$\|\tilde{s}_2\|_1 = \mathbf{1}^\top \Pi^*_{3,22} \mathbf{1} = \|s^*_2\|_1 - \mathbf{1}^\top \Pi^*_{3,21} \mathbf{1} \, .$$

$$\|\tilde{s}_3\|_1 = \mathbf{1}^\top \Pi^*_{3,11} \mathbf{1} + \mathbf{1}^\top \Pi^*_{3,21} \mathbf{1} = \|s^*_3\|_1$$

And also by our construction, $\tilde{s}_1 = s^*_1 + c$ where $c = (\Pi^*_{3,21})^\top \mathbf{1}$. As a consequence we have $\|c\|_1 = \mathbf{1}^\top \Pi^*_{3,21} \mathbf{1}$. Then it follows:

$$\begin{aligned} \sum_{i=1}^4 \|\tilde{s}_i\|_1 &= \|s^*_1 + c\| + \|s^*_2\|_1 - \mathbf{1}^\top \Pi^*_{3,21} \mathbf{1} + \|s^*_3\|_1 + \|s^*_4\|_1 \\ &\leq \sum_{i=1}^4 \|s^*_i\|_1 + \|c\|_1 - \mathbf{1}^\top \Pi^*_{3,21} \mathbf{1} \\ &= \sum_{i=1}^4 \|s^*_i\|_1 \end{aligned}$$

So the objective value is overall reduced. This contradicts the optimality of Π^*_3 which completes the proof. □

E. Change of support of outliers with respect to λ

For any λ, define the set $\mathcal{I}_\lambda = \{(i,j) : C_{i,j} > 2\lambda\}$, i.e. $\mathcal{I}_\lambda$ denotes the costs which exceeds the threshold 2λ. As before we define by C_λ to be truncated cost $C \wedge 2\lambda$. Denote by π_λ to be the optimal transport plan with respect to C_λ and the marginal measures μ, ν. Borrowing our notations from previous theorems, we define a "slack vector" $\mathbf{s}_\lambda$ as:

$$\mathbf{s}_\lambda(i) = \sum_{j=1}^n \pi_\lambda(i,j) \mathbb{1}_{C(i,j) > 2\lambda} = \sum_{j:(i,j) \in \mathcal{I}_\lambda} \pi_\lambda(i,j) \, .$$

And we define the observation i_0 to be an outlier if $\mathbf{s}_\lambda(i_0) > 0$. It is immediate that for any $\lambda_1 < \lambda_2$, $\mathcal{I}_{\lambda_1} \supseteq \mathcal{I}_{\lambda_2}$. We goal is to establish the following theorem:

Theorem E.1. *For any $\lambda_1 < \lambda_2$, if $\mathbf{s}_{\lambda_2}(i_0) > 0$, then $\mathbf{s}_{\lambda_1}(i_0) > 0$, i.e. if a point is selected as outlier for larger λ, then it is also selected as outlier for smaller λ.*

Proof. Fix $\lambda_1 < \lambda_2$. Note that for any $\pi \in \Pi(\mu, \nu)$ we have:

$$\begin{aligned} \langle C_{\lambda_2} - C_{\lambda_1}, \pi \rangle = & \sum_{(i,j) \in I_{\lambda_1} \cap \mathcal{I}^c_{\lambda_2}} (C(i,j) - 2\lambda_1) \pi(i,j) \\ & + 2(\lambda_2 - \lambda_1) \sum_{(i,j) \in \mathcal{I}_{\lambda_2}} \pi(i,j) \\ := & \, T_1(\pi) + T_2(\pi) \, . \end{aligned}$$

Now as π_{λ_2} is optimal with respect to C_{λ_2} and π_{λ_1} is optimal with respect to C_{λ_1} we have:

$$\begin{aligned} & \langle C_{\lambda_1}, \pi_{\lambda_2} \rangle + T_1(\pi_{\lambda_2}) + T_2(\pi_{\lambda_2}) \\ & = \langle C_{\lambda_2}, \pi_{\lambda_2} \rangle \\ & \leq \langle C_{\lambda_2}, \pi_{\lambda_1} \rangle \\ & = \langle C_{\lambda_1}, \pi_{\lambda_1} \rangle + T_1(\pi_{\lambda_1}) + T_2(\pi_{\lambda_1}) \\ & \leq \langle C_{\lambda_1}, \pi_{\lambda_2} \rangle + T_1(\pi_{\lambda_1}) + T_2(\pi_{\lambda_1}) \end{aligned}$$

Therefore we have:

$$T_1(\pi_{\lambda_2}) + T_2(\pi_{\lambda_2}) \leq T_1(\pi_{\lambda_1}) + T_2(\pi_{\lambda_1}) \, . \tag{E.1}$$

But this is not enough. Note that we can further decompose T_1 (and similarly T_2) as:

$$\begin{aligned} T_{1,i}(\pi) &= \sum_{j:(i,j) \in I_{\lambda_1} \cap \mathcal{I}^c_{\lambda_2}} (C_{i,j} - 2\lambda_1) \, \pi(i,j) \\ T_{2,i}(\pi) &= 2(\lambda_2 - \lambda_1) \sum_{j:(i,j) \in \mathcal{I}_{\lambda_2}} \pi(i,j) \, . \end{aligned}$$

Hence we have:

$$T_1(\pi) = \sum_{i=1}^n T_{1,i}(\pi), \quad T_2(\pi) = \sum_{i=1}^n T_{2,i}(\pi) \, .$$

In equation E.1 we have established that $T_1(\pi_{\lambda_2}) + T_2(\pi_{\lambda_2}) \le T_1(\pi_{\lambda_1}) + T_2(\pi_{\lambda_1})$. In addition if we can show that:

$$T_{1,i_0}(\pi_{\lambda_1}) + T_{2,i_0}(\pi_{\lambda_1}) = 0 \implies T_{2,i_0}(\pi_{\lambda_1}) = 0\,. \tag{E.2}$$

holds for all $1 \le i \le n$ then we are done. This is because, suppose $\mathbf{s}_{\lambda_1}(i_0) = 0$, Then

$$T_{1,i_0}(\pi_{\lambda_1}) + T_{2,i_0}(\pi_{\lambda_1}) = 0\,.$$

This in turn by equation E.2 implies

$$T_{2,i_0}(\pi_{\lambda_2}) = 0\,,$$

i.e. $\mathbf{s}_{\lambda_2}(i_0) = 0$.

Lemma E.2. *Suppose C is a 2×2 cost matrix with all unequal cost:*

$$C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$$

If $C_{22} > 2\lambda_2$ and $C_{21} < 2\lambda_1$, then the following two inequalities won't occur simultaneously:

$$C_{11}^{\lambda_2} + C_{22}^{\lambda_2} \le C_{12}^{\lambda_2} + C_{21}^{\lambda_2}\,,$$
$$C_{12}^{\lambda_1} + C_{21}^{\lambda_1} \le C_{11}^{\lambda_1} + C_{22}^{\lambda_1}\,.$$

Now suppose i_2 is an outlier with respect to λ_2 but not with respect to λ_1. Then there exists j_1 and j_2 ($j_1 \ne j_2$) such that:

$$C_{i_2,j_1} < 2\lambda_1, C_{i_2,j_2} > 2\lambda_2$$

such that $\pi^{\lambda_2}_{i_2,j_2} > 0$ and $\pi^{\lambda_1}_{i_2,j_1} > 0$.

Case 1: Now assume that we can find $i_1 \ne i_2$ such that $\pi^{\lambda_2}_{i_1,j_1} > 0$ and $\pi^{\lambda_1}_{i_1,j_2} > 0$.

Then $(i_1,j_2),(i_2,j_1) \in \mathsf{supp}(\pi^{\lambda_1})$ and $(i_1,j_1),(i_2,j_2) \in \mathsf{supp}(\pi^{\lambda_2})$. Hence from c-cyclical monotonicity properties of the support of the optimal transport plan we have for π^{λ_1}:

$$C^{\lambda_1}_{i_1,j_2} + C^{\lambda_1}_{i_2,j_1} \le C^{\lambda_1}_{i_1,j_1} + C^{\lambda_1}_{i_2,j_2}\,,$$

and for π^{λ_2}:

$$C^{\lambda_2}_{i_1,j_1} + C^{\lambda_2}_{i_2,j_2} \le C^{\lambda_2}_{i_1,j_2} + C^{\lambda_2}_{i_2,j_1}\,.$$

which is a contradiction from Lemma E.2. This completes the proof.

Case 2: Need to be proved Now we need to consider the other case, there does not exist any row $i_1 \ne i_2$ such that both $\pi^{\lambda_2}_{i_1,j_1} > 0$ and $\pi^{\lambda_1}_{i_1,j_2} > 0$ occur simultaneously. This means that the columns j_1, j_2 are orthogonal, i.e. $\langle \pi^{\lambda_2}_{:,j_1}, \pi^{\lambda_1}_{:,j_2} \rangle = 0$. □

E.1. Proof of Lemma E.2

As $C_{21} < 2\lambda_1$ and $C_{22} > 2\lambda_2$ we can modify the inequalities in Lemma E.2 as:

$$C_{11}^{\lambda_2} + 2\lambda_2 \le C_{12}^{\lambda_2} + C_{21}\,, \tag{E.3}$$
$$C_{12}^{\lambda_1} + C_{21} \le C_{11}^{\lambda_1} + 2\lambda_1\,. \tag{E.4}$$

Now as we have assume $C_{21} < 2\lambda_1$, from equation E.3 we obtain:

$$2\lambda_1 > C_{11}^{\lambda_2} - C_{12}^{\lambda_2} + 2\lambda_2$$
$$\iff 2(\lambda_2 - \lambda_1) < C_{12}^{\lambda_2} - C_{11}^{\lambda_2}\,. \tag{E.5}$$

Hence $C_{12}^{\lambda_2} - C_{11}^{\lambda_2} > 0$, which implies $C_{11} < C_{12}, C_{11} < 2\lambda_2$ and also both C_{11} and C_{12} can not lie within $(2\lambda_1, 2\lambda_2)$. We divide the rest of the proofs into four small cases:

Case 1: Assume $2\lambda_1 < C_{11} < 2\lambda_2, C_{12} > 2\lambda_2$. In this case from equation E.3 we have:

$$C_{11} + 2\lambda_2 \le 2\lambda_2 + C_{21}$$

i.e. $C_{21} \ge C_{11}$ which is not possible as $C_{11} > 2\lambda_1$ and $C_{21} < 2\lambda_1$.

Case 2: Assume $C_{11} < 2\lambda_1, C_{12} > 2\lambda_2$. Then from equation E.4 we have $C_{21} \le C_{11}$ and from equation E.3 we have: $C_{11} \le C_{21}$ which cannot occur simultaneously.

Case 3: Assume $C_{11} < 2\lambda_1$ and $2\lambda_1 < C_{12} < 2\lambda_2$. Then from equation E.3 and equation E.4 we have respectively:

$$C_{11} + 2\lambda_2 \le C_{12} + C_{21}\,,$$
$$2\lambda_1 + C_{21} \le C_{11} + 2\lambda_1$$

From the second inequality we have $C_{21} \le C_{11}$, which putting back in the first inequality yields:

$$C_{11} + 2\lambda_2 \le C_{12} + C_{11} \implies C_{12} \ge 2\lambda_2$$

which is a contradiction.

Case 4: Assume $C_{11} < 2\lambda_1$ and $C_{12} < 2\lambda_1$. This form equation E.3 yields:

$$C_{11} + 2\lambda_2 \le C_{12} + C_{21}$$
$$\implies C_{21} \ge C_{11} - C_{12} + 2\lambda_2\,. \tag{E.6}$$

Also from equation E.4 we have:

$$C_{12} + C_{21} \le C_{11} + 2\lambda_1$$
$$\implies C_{21} \le C_{11} - C_{12} + 2\lambda_1\,. \tag{E.7}$$

From equation E.6 and equation E.7 we have:

$$C_{11} - C_{12} + 2\lambda_2 \le C_{11} - C_{12} + 2\lambda_1$$

i.e. $\lambda_2 \le \lambda_1$ which is a contradiction. This completes the proof.

Table 1. Robust mean estimation with GANs using different distribution divergences. True mean is $\eta_0 = \mathbf{0}_5$; sample size $n = 1000$; contamination proportion $\epsilon = 0.2$. We report results over 30 experiment restarts.

Contamination	JS Loss	SH Loss	ROBOT
Cauchy$(0.1 \cdot \mathbf{1_5}, I_5)$	0.2 ± 0.06	**0.17** ± 0.04	**0.17** ± 0.05
Cauchy$(0.5 \cdot \mathbf{1_5}, I_5)$	0.3 ± 0.07	0.26 ± 0.05	**0.25** ± 0.05
Cauchy$(1 \cdot \mathbf{1_5}, I_5)$	0.45 ± 0.14	0.37 ± 0.06	**0.36** ± 0.07
Cauchy$(2 \cdot \mathbf{1_5}, I_5)$	0.39 ± 0.3	0.26 ± 0.06	**0.2** ± 0.07

F. Robust mean experiment with Cauchy distribution

In this section we present our results corresponding to the robust mean estimation with the generative distribution $g_\theta(x) = x + \theta$ where $x \sim$ Cauchy$(0, 1)$. As in Subsection 4.1, we assume that we have observation $\{x_1, \ldots, x_n\}$ from a contaminated distribution $(1 - \epsilon)$ Cauchy$(\eta_0, 1) + \epsilon$ Cauchy$(\eta_1, 1)$. For our experiments we take $\eta_0 = \mathbf{0}_5$ and vary $\eta_1 \in \{0.1 \cdot \mathbf{1_5}, 0.5 \cdot \mathbf{1_5}, 1 \cdot \mathbf{1_5}, 2 \cdot \mathbf{1_5}\}$ along wth $\epsilon = 0.2$. We compare our method with Wu et al. (2020) and results are presented in Table 1.

References

Gabriel Peyré and Marco Cuturi. Computational Optimal Transport. *arXiv:1803.00567 [stat]*, March 2018.

C. Villani. *Optimal Transport: Old and New. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathemtical Sciences]*. Springer, Berlin, 2009.

Kaiwen Wu, Gavin Weiguang Ding, Ruitong Huang, and Yaoliang Yu. On Minimax Optimality of GANs for Robust Mean Estimation. In *International Conference on Artificial Intelligence and Statistics*, pp. 4541–4551, June 2020.