Handling Multiple Costs in Optimal Transport: 
Strong Duality and Efficient Computation

Meyer Scetbon\textsuperscript{1}  Laurent Meunier\textsuperscript{3,4}  Jamal Atif\textsuperscript{3}  Marco Cuturi\textsuperscript{1,2}

\textsuperscript{1}CREST, ENSAE \\ \textsuperscript{2}Google Brain \\ \textsuperscript{3}LAMSADE, Université Paris-Dauphine \\ \textsuperscript{4}Facebook AI Research

Abstract

We introduce an extension of the optimal transportation (OT) problem when multiple costs are involved. We consider a linear optimization problem which allows to choose locally among \( N \geq 1 \) cost functions in order to minimize the cost of transport, while making them contribute equally. When \( N = 1 \), we recover the classical OT problem; for \( N = 2 \) we are able to recover integral probability metrics defined by \( \alpha \)-Hölder functions, which includes the Dudley metric. We derive the dual formulation of the problem and show that strong duality holds under some mild assumptions. In the discrete case, as with regular OT, the problem can be solved with a linear program. We provide a faster, entropic regularized formulation of that problem. We validate our proposed approximation with experiments on real and synthetic datasets.

1 Introduction

Optimal Transport (OT) has gained interest last years in machine learning with diverse applications in neuroimaging \cite{22}, generative models \cite{3,28}, supervised learning \cite{9}, word embeddings \cite{2}, reconstruction cell trajectories \cite{30,40} or adversarial examples \cite{39}. The key to use OT in these applications lies in the gain of computation efficiency thanks to regularizations that smooths the OT problem. More specifically, when one uses an entropic penalty, one recovers the so called Sinkhorn divergences \cite{10}. Although OT has a rich theoretical history \cite{37}, the choice of the cost is often a difficult question. Choosing the “right” cost requires a strong knowledge about the problem and one may ask whether the used cost to compare distributions is relevant for the problem at hand. In this paper, we introduce a new family of variational problems extending the optimal transport problem when multiple costs are involved.

Given two distributions and \( N \geq 1 \) cost functions, we present a new transportation problem where the goal is to partition among the costs the task of transporting one distribution towards another, in order to minimize the global transporting cost while ensuring that every cost contributes equally to the transportation task. In that sense, we aim to get every cost involved equally in the transportation problem so that we can get the most out of it. For instance, consider a government who has \( N \) available shippers to transport its goods from one place to another and who tries to stimulate the economy after a pandemic outbreak. To do so, it may decide to pay its shippers equally for this task while still minimizing the global cost of transportation for the goods. We introduce MOT (\textbf{M}ultiple cost \textbf{O}ptimal \textbf{T}ransport), a formulation that solves this problem. We prove duality results and give interpretation for both primal and dual problems. MOT has strong links with notions of fairness and can be interpreted as a criterion of fair division. Every shipper feels that their payoff is at least as good as the payoff of any other shipper, and thus no shipper feels envy \cite{4}. As interesting properties, we recover some Integral Probability Metrics (IPMs) \cite{25} as Dudley metric \cite{13}, or standard Wasserstein metric \cite{37}.
Contributions. In this paper we introduce MOT an extension of Optimal Transport where multiple costs are considered under egalitarian constraints. We make the following contributions.

- In section 3, we introduce the problem, derive its dual and prove strong duality results. Moreover, we show that the problem we defined is closely related to some usual IPMs families, including Wasserstein distances and the Dudley metric, which is known to metrize weak convergence.
- In section 4, we propose an entropic regularized version of the problem, derive its dual formulation, obtain strong duality and derive an efficient algorithm to compute it. As a by-product, we obtain the differentiability of the regularized version and apply our approximation to compute barycenters between measures.

2 Related Work

Multiple Costs in Optimal Transport. Optimal transport aims to move a distribution towards another at lowest cost. More formally, if \( c \) is a cost function on the ground space \( \mathcal{X} \times \mathcal{Y} \), then the relaxed Kantorovich formulation of OT is defined for \( \mu \) and \( \nu \) two distributions as

\[
W_c(\mu, \nu) := \inf_{\gamma} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y)d\gamma(x, y)
\]

where the infimum is taken over all distributions \( \gamma \) with marginals \( \mu \) and \( \nu \). Kantorovich theorem states the following strong duality result under mild assumptions [37]

\[
W_c(\mu, \nu) = \sup_{f,g} \int_{\mathcal{X}} f(x)d\mu(x) + \int_{\mathcal{Y}} g(y)d\nu(y)
\]

where the supremum is taken over continuous bounded functions satisfying or all \( x, y, f(x) + g(y) \leq c(x, y) \). However, the problem of choosing the cost or handling multiple costs is a difficult question which, up to our knowledge, does not seem to be fully understood. So far, recent works try to handle multiple costs. For instance, [27] proposed a robust Wasserstein distance where the distributions are projected on a \( k \)-dimensional subspace that maximizes their transport cost. In that sense, they aim to choose the most expensive cost among Mahalanobis square distances with kernels of rank \( k \). Here, we do not aim to consider a worst case scenario among the available costs but rather consider that each cost has to perform the best it can under the constraint that it cannot do more than the others.

Regularized Optimal Transport. Computing exactly the optimal transport cost requires solving a linear program with a supercubic complexity \( (n^3 \log n) \) [35] that results in an output that is not differentiable with respect to the measures’ locations or weights [6]. Moreover, OT suffers from the curse of dimensionality [12, 17] and is therefore likely to be meaningless when used on samples from high-dimensional densities. Following the line of work introduced by [10], we propose an approximated computation of our problem by regularizing it with an entropic term. Such regularization in OT accelerates the computation, makes the problem differentiable with regards to the considered distributions [16] and reduces the curse of dimensionality [18]. Taking the dual of the approximation, we obtain a smooth and convex optimization problem under a simplicial constraint.

Integral Probability Metrics. In our work, we make links with some integral probability metrics. IPMs are (semi-)metrics on the space of probability measures. For a set of functions \( \mathcal{F} \) and two probability distributions \( \mu \) and \( \nu \), they are defined as \( \text{IPM}_\mathcal{F}(\mu, \nu) = \sup_{f \in \mathcal{F}} \int fd\mu - \int fd\nu \). For instance, when \( \mathcal{F} \) is chosen to be the set of bounded functions with uniform norm less or equal than 1, we recover the Total Variation distance [34] (TV). They recently regained interest in the Machine Learning community thanks to their application to Generative Adversarial Networks (GANs) [19] where IPMs are natural metrics for the
Figure 1: Comparison of the optimal couplings obtained from standard OT for three different costs and MOT. Blue dots and red squares represent the locations of two discrete uniform measures. *Left, middle left, middle right:* Kantorovich couplings between the two measures for Euclidean cost, square Euclidean cost and square L1 norm respectively. *Right:* transport couplings of MOT solving Eq. (1). Note that each cost contributes equally and its contribution is lower than the smallest OT cost.

discriminator $[3, 15, 21, 24]$. They also helped to build consistent two-sample tests $[20, 29]$. However when a closed form of the IPM is not available, exact computation of IPMs between discrete distributions may not be possible or can be costful. For instance, the Dudley metric can be written as a Linear Program $[33]$ which has at least the same complexity as standard OT. Here, we show that the Dudley metric is in fact a particular case of our problem and obtain a faster approximation thanks to the entropic regularization.

3 Multiple Cost Optimal Transport

3.1 Notations

Let $\mathcal{Z}$ be a Polish space, we denote $\mathcal{M}(\mathcal{Z})$ the set of Radon measures. We call $\mathcal{M}_+(\mathcal{Z})$ the sets of positive Radon measures, and $\mathcal{M}_1^+(\mathcal{Z})$ the set of probability measures. We denote $C_b(\mathcal{Z})$ the vector space of bounded continuous functions on $\mathcal{Z}$. Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces. We denote $\Pi_1 : (x, y) \in \mathcal{X} \times \mathcal{Y} \mapsto x$ and $\Pi_2 : (x, y) \in \mathcal{X} \times \mathcal{Y} \mapsto y$ respectively the projections on $\mathcal{X}$ and $\mathcal{Y}$, which are continuous applications. For an application $g$ and a measure $\mu$, we denote $g_#\mu$ the pushforward measure of $\mu$ by $g$. For $\mathcal{X}$ and $\mathcal{Y}$ two Polish spaces, we denote $\text{LSC}(\mathcal{X} \times \mathcal{Y})$ the space of lower semi-continuous functions on $\mathcal{X} \times \mathcal{Y}$. Let $N$ be a positive integer. We denote $c := (c_i)_{1 \leq i \leq N}$ a family of lower semi-continuous costs on $\mathcal{X} \times \mathcal{Y}$ with values in $\mathbb{R}_+ \cup \{+\infty\}$, for all $N \geq 1$, denote $\Delta_N^+: = \{\lambda \in \mathbb{R}_+^N \text{ s.t. } \sum_{i=1}^N \lambda_i = 1\}$, the probability simplex of $\mathbb{R}_+^N$ and $\Delta_N = \{\lambda \in \mathbb{R}^N \text{ s.t. } \sum_{i=1}^N \lambda_i = 1\}$.

3.2 Primal and Dual Formulations

We now introduce a more general version of the standard OT problem to handle multiple costs. Given $N \geq 1$ costs, the purpose here is to minimize the global cost of transportation of a distribution towards another under the constraint that the task of transportation has to be partitioned among the costs such that each cost cannot earn more than the others. One can imagine a government wishing equity among his shippers by paying them equally. Formally the problem studied here is defined as follows.

**Definition 1** (Primal problem). Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces. Let $c := (c_i)_{1 \leq i \leq N}$ be a family of nonnegative lower semi-continuous cost functions on $\mathcal{X} \times \mathcal{Y}$, and $\mu \in \mathcal{M}_1^+(\mathcal{X})$ and $\nu \in \mathcal{M}_1^+(\mathcal{Y})$. We define the multiple
Figure 2: **Left, middle left, middle right**: the size of dots and squares is proportional to the weight of their representing atom in the distributions $\mu_k^*$ and $\nu_k^*$ respectively.

The collection “price” $f_k^*$ for each point in $\mu_k^*$, and its delivery counterpart $g_k^*$ in $\nu_k^*$ are represented by the color of dots and squares according to the color scale on the right hand side. The gray dots and squares correspond to the points that are ignored by salesperson $k$ in the sense that there is no mass or almost no mass in distributions $\mu_k^*$ or $\nu_k^*$. **Right**: the size of dots and squares are uniform since they correspond to the weights of uniform distributions $\mu$ and $\nu$ respectively. The values of $f^*$ and $g^*$ are given also by the color at each point. Note that each salesperson earns exactly the same amount of money, corresponding exactly MOT cost. This value can be computed using dual formulation (2) or its reformulation (3) and for each figure it equals the sum of the values (encoded with colors) multiplied by the weight of each point (encoded with sizes).

The cost optimal transport primal problem:

$$
\text{MOT}_c(\mu, \nu) := \inf_{(t, \gamma) \in \mathbb{R} \times \Gamma_{\mu,\nu}^N} \left\{ t \text{ s.t. } \forall i = 1, \ldots, N \int_{X \times Y} c_i(x, y) d\gamma_i(x, y) = t \right\}
$$

where $\Gamma_{\mu,\nu}^N := \{ (\gamma_i)_{1 \leq i \leq N} \in \mathcal{M}_+ (X \times Y)^N \text{ s.t. } \Pi_1 \sum_i \gamma_i = \mu \text{ and } \Pi_2 \sum_i \gamma_i = \nu \}$.

We prove along with Theorem 1 that the problem is well defined and the infimum is attained. The well definition highly relies on the non negativity of the cost. For instance if two costs are considered, one always positive and the other always negative, then the constraints cannot be satisfied. Note that the problem defined here is a linear optimization problem and when $N = 1$ we recover standard optimal transport. Figure 1 illustrates the multiple cost optimal transport problem we consider. The definition above can be interpreted as in the following example.

**Government and goods.** Assume a government can work with $N$ shippers to transport its goods $\mu$ from its stocks to stores $\nu$. To transport one unit of good from $x$ the location to $y$, each shipper $i$ proposes a price $c_i(x, y)$. Assuming that the shipper $i$ move just a fraction of the goods to some stores, and by denoting this coupling $\gamma_i$, the shipper will bill $\int c_i d\gamma_i$. Therefore the government whose objective function is Problem 1 aims to find a partition of the work among the shippers that minimizes the total cost of transport while ensuring that shippers get equally paid. Note that from the definition, the government will pay a single shipper less than if this shipper was alone on the market: in mathematical terms, it means that $\text{MOT}_c(\mu, \nu) \leq \min_i W_{c_i}(\mu, \nu)$.

Let us now introduce the dual formulation of the problem and show that strong duality holds under some mild assumptions. See Appendix C.3 for the proof.

**Theorem 1** (Strong Duality). Let $X$ and $Y$ be Polish spaces. Let $c := (c_i)_{i=1}^N$ be nonnegative lower
semi-continuous costs. Then strong duality holds, i.e. for \((\mu, \nu) \in M^1_+(\mathcal{X}) \times M^1_+(\mathcal{Y})\):

\[
\text{MOT}_c(\mu, \nu) = \sup_{\lambda \in \Delta_N} \sup_{(f, g) \in \mathcal{F}_c^\lambda} \int_{x \in \mathcal{X}} f(x) d\mu(x) + \int_{y \in \mathcal{Y}} g(y) d\nu(y)
\]

where \(\mathcal{F}_c^\lambda := \{(f, g) \in C^b(\mathcal{X}) \times C^b(\mathcal{Y}) \text{ s.t. } \forall i \in \{1, \ldots, N\}, \ f \oplus g \leq \lambda_i c_i\}\).

**Remark 1.** It is worth noting that when we assume in addition that the costs \((c_i)_{1 \leq i \leq N}\) are continuous functions and the ground spaces \(\mathcal{X}\) and \(\mathcal{Y}\) are compact, the supremum in the dual formulation Eq. (2) is attained. See Appendix [C.1] for the proof.

This theorem holds under the same hypothesis and follows the same reasoning as the one in [37, Theorem 1.3]. This result is more difficult to interpret, but will be useful in Section 4 to compute an efficient algorithm. To give more interpretability to this problem, we reformulate the problem as follows. Let us introduce the set of functions:

\[\mathcal{G}_c^N := \{(f_k, g_k)_{k=1}^N \in (C^b(\mathcal{X}) \times C^b(\mathcal{Y}))^N \text{ s.t. } \forall k \in \{1, \ldots, N\}, \ f_k \oplus g_k \leq c_k\}\]

and the subset of \((M_+(\mathcal{X}) \times M_+(\mathcal{Y}))^N\), representing the marginals:

\[\mathcal{Y}_{\mu, \nu}^N := \{(\mu_i, \nu_i)_{i=1}^N \text{ s.t. } \sum_{i=1}^N \mu_i = \mu, \ \sum_{i=1}^N \nu_i = \nu \text{ and } \forall i, \ \mu_i(\mathcal{X}) = \nu_i(\mathcal{Y})\}\]

Let us now show the following reformulation of the problem. See Appendix [A.3] for the proof.

**Proposition 1.** Under the same assumptions of Theorem 1, we have

\[
\text{MOT}(\mu, \nu) = \sup_{(f_k, g_k)^N_{k=1} \in \mathcal{G}_c^N} \inf_{t \in \mathbb{R}} \left\{ t \text{ s.t. } \forall k, \ \int f_k d\mu_k + \int g_k d\nu_k = t \right\}
\]

**Remark 2.** As soon as Problem (1) admits a solution \((\lambda^*, f^*, g^*)\) such that for all \(k \in \{1, \ldots, N\}, \ \lambda_k^* \neq 0\), then \(f_k^* = \frac{\lambda_k^*}{N^*}\) and \(g_k^* = \frac{\mu_k^*}{N^*}\) is an optimal solution of the Problem (2).

Figure 2 illustrates this formulation of the problem with dual potentials. Let us now give a simple interpretation of what it means on the following example.

**Outsourcing logistics.** Assume that the government cannot solve the Linear Program (1) stated above (primal formulation), and decides instead to outsource that task to another organization which aims making everyone work equally for the cheapest price. Assume that this organization has at disposal \(N\) salespersons which may propose different prices to transport goods. Each salesperson \(k\) chooses a pricing scheme with the following structure: the salesperson splits the logistic task into that of collecting and then delivering the goods, and will apply a collection price \(f_k(x)\) for one unit of good located at \(x\) (no matter where that unit is sent to), and a delivery price \(g_k(y)\) for one unit to the location \(y\) (no matter from which place that unit comes from). Then the salesperson for transporting some goods \(\mu_k\) to some stores \(\nu_k\) will charge \(\int_{x \in \mathcal{X}} f_k(x) d\mu_k(x) + \int_{y \in \mathcal{Y}} g_k(y) d\nu_k(y)\).

**Checking prices.** The government must ensure the price of transport given by the outsourcing organization will be at least lower than if he has followed the primal problem (1). For each salesperson \(k\), the salesperson’s pricing scheme implies that transferring one unit of the resource from \(x\) to \(y\) costs exactly \(f_k(x) + g_k(y)\). Yet, the government also knows that the cost of shipping one unit from \(x\) to \(y\) as priced by the transportation company \(k\) is \(c_k(x, y)\). Therefore, if for any pair \((x, y)\) the aggregate price \(f_k(x) + g_k(y)\) is strictly larger that \(c_k(x, y)\), the salesperson is charging more than the fair price charged by the transportation company for that task, and the government should reject the \(k\)-th salesperson’s offer. It is therefore in the interest of the government
to check that for all pairs \((x, y)\) the prices offered by the salesperson verify \(f_k \oplus g_k(x, y) \leq c_k(x, y)\). Moreover the government wants all its goods have been transported by the salespersons at their destinations. Therefore the government needs to check that \(\sum_{i=1}^{N} \mu_i = \mu\) and that \(\sum_{i=1}^{N} \nu_i = \nu\). Finally recall the organization wants every salesperson to earn the same, which gives for all \(j, k, \int_{x \in X} f_j(x) d\mu_j(x) + \int_{y \in Y} g_j(y) d\nu_j(y) = \int_{x \in X} f_k(x) d\mu_k(x) + \int_{y \in Y} g_k(y) d\nu_k(y)\).

**Optimal Prices.** The salespersons must find a set of prices \((f_k, g_k)_{k=1}^{N}\) and a distribution of the masses \((\mu_k, \nu_k)_{k=1}^{N}\) that maximize their profits while minimizing the mass that they have to transport such that they earn exactly the same which is exactly the problem described in Eq. \((3)\).

**Remark 3 (Primal-Dual Optimality).** Note that the problem in Eq. \((3)\) admits a saddle point as soon as the cost functions \((c_i)_{1 \leq i \leq N}\) are continuous and the spaces \(X\) and \(Y\) are compact. Moreover at the optimum the new formulation of the problem implies that each salesperson receives the exact same amount of money from the government which was expected from the primal formulation. In fact at optimality we have for all \(k \in \{1, ..., N\}\) the following primal-dual relation

\[
 f_k^* \oplus g_k^*(x, y) = c_k(x, y) \text{ for all } (x, y) \in \text{Supp}(\gamma_k^*), \quad \Pi_2 \gamma_k^* = \mu_k^*, \quad \Pi_2 \gamma_k^* = \nu_k^*.
\]

**Discrete case.** When the distributions \(\mu\) and \(\nu\) are discretes, primal \((1)\) and dual \((2)\) formulations are Linear Programs which can be solved exactly using linear solvers with complexity at least super cubic. Note that the primal problem has less constraints than the dual problems, then using primal formulation will result in a faster algorithm to compute the solution. Details on the discrete case and its dual are left in Appendix \([3.1]\).

### 3.3 Link with other Probability Metrics

In this section, we provide some topological proprieties on the object defined by the MOT problem. In particular, we make links with other known probability metrics, such as Dudley and Wasserstein metrics and give a tight upper bound.

When \(N = 1\), recall from the definition \((1)\) that the problem considered is exactly the standard OT problem. Moreover any multiple cost problem with \(k \leq N\) costs can always be rewritten as a multiple cost problem with \(N\) costs. See Appendix \([C.2]\) for the proof. From this property, it is interesting to note that, for any \(N \geq 1\), MOT generalizes standard Optimal Transport.

**Remark 4 (MOT generalizes OT).** Given a cost function \(c\), if we consider the problem MOT with \(N\) costs such that, for all \(i, c_i = N \times c\) then, the problem MOT\(_c\) is exactly W\(_c\). See Appendix \([C.2]\) for the proof.

Now we have seen that all standard OT problems are sub-cases of the MOT problem, one may ask whether MOT can recover other family of metrics different from standard OT. Indeed we show that our multiple cost problem recovers an important family of IPMs that are those defined by the function space of \(\alpha\)-Hölder functions with \(\alpha \in (0, 1]\). See Appendix \([A.3]\) for the proof.

**Proposition 2.** Let \(X\) be a Polish space. Let \(d\) be a metric on \(X^2\) and \(\alpha \in (0, 1]\). Denote \(c_1 = 2 \times 1_{x \neq y}\), \(c_2 = d^\alpha\) and \(c := (c_1, (N-1) \times c_2, ..., (N-1) \times c_2) \in \text{LSC}(X \times Y)^N\) then for any \((\mu, \nu) \in \mathcal{M}_+^1(X) \times \mathcal{M}_+^1(X)\)

\[
\text{MOT}_c(\mu, \nu) = \text{MOT}_{(c_1, d^\alpha)}(\mu, \nu) = \sup_{f \in B_{d^\alpha}(X)} \int_X f(x) dP(x) - \int_Y f(y) dQ(y)
\]

where \(B_{d^\alpha}^b(X) := \{f \in C^b(X): \|f\|_\infty + \|f\|_\alpha \leq 1\}\) and \(\|f\|_\alpha := \sup_{x \neq y} \frac{|f(x) - f(y)|}{d^\alpha(x, y)}\).

**Example.** In particular, when \(\alpha = 1\), then for \((\mu, \nu) \in \mathcal{M}_+^1(X) \times \mathcal{M}_+^1(X)\), we have \(\text{MOT}_c(\mu, \nu) = \text{MOT}_{(c, d)}(\mu, \nu) = \beta_d(\mu, \nu)\) where \(\beta_d\) is the Dudley Metric \([13]\). In other words, the Dudley metric can be interpreted as an equally contributed optimal transport problem with the trivial cost and a metric \(d\).
Weak Convergence. When $d$ is a metric on $X$ and unbounded, it is well known that $W_{dp}$ with $p \in (0, +\infty)$ metrizes a convergence a bit stronger than weak convergence in general [37, Chapter 7]. A sufficient condition for Wasserstein distances to metrize weak convergence on the whole space of distributions is that the metric $d$ is bounded. In contrast, metrics defined by Eq. (4) do not require such assumptions and $\text{MOT}_{(c_i,d^p)}$ metrizes weak convergence of probability measures.

For an arbitrary choice of costs $(c_i)_{1 \leq i \leq N}$, we obtain a tight upper control of $\text{MOT}$ and show how it is related to the OT problem associated to each cost involved. See Appendix [A.4] for the proof.

**Proposition 3.** Let $X$ and $Y$ be Polish spaces. Let $c := (c_i)_{1 \leq i \leq N}$ be a family of nonnegative lower semi-continuous costs. For any $(\mu, \nu) \in \mathcal{M}_+^1(X) \times \mathcal{M}_+^1(Y)$

$$\text{MOT}_c(\mu, \nu) \leq \left( \sum_{i=1}^{N} \frac{1}{W_{c_i}(\mu, \nu)} \right)^{-1}. \quad (5)$$

Eq. (5) means that the solution given by $\text{MOT}_c$, which is the minimal cost to transport all goods under the constraint that all shippers get paid equally, is lower than the mean cost for transporting all goods from stocks to stores in the case where goods in the stocks are equally splitted among the shippers to transport them, which is, in other words, the harmonic sum as written above.

**Example.** Applying the above result in the case of the Dudley metric recovers the following inequality [35, Proposition 5.1]

$$\beta_d(\mu, \nu) \leq \frac{\text{TV}(\mu, \nu)W_d(\mu, \nu)}{\text{TV}(\mu, \nu) + W_d(\mu, \nu)}.$$ 

4 Entropic Relaxation

Linear Programs are standard in Optimal Transport but are generally slow. Considering the primal formulation instead of the dual one may speed up a bit the solving, but the problem remains slow to compute. Following the work of [10], we propose an entropic relaxation of the MOT problem to make the problem strongly convex and faster to compute. We obtain the dual formulation and derive an efficient algorithm from it to compute an approximation of MOT.

4.1 Primal-Dual Formulation

Let us first extend the notion of Kullback-Leibler divergence for positive Radon measures. Let $Z$ be a Polish space, for $\mu, \nu \in \mathcal{M}_+(Z)$, we define the generalized Kullback-Leibler divergence as $\text{KL}(\mu\|\nu) = \int \log \frac{d\mu}{d\nu} d\mu + \int d\nu - \int d\mu$ if $\mu \ll \nu$, and $+\infty$ otherwise. We introduce the following regularized version of MOT.

**Definition 2** (Entropic relaxed primal problem). Let $X$ and $Y$ be two Polish spaces, $c := (c_i)_{1 \leq i \leq N}$ a family of nonnegative lower semi-continuous costs on $X \times Y$ and $\varepsilon := (\varepsilon_i)_{1 \leq i \leq N}$ be non negative real numbers. For $(\mu, \nu) \in \mathcal{M}_+^1(X) \times \mathcal{M}_+^1(Y)$, we define the MOT regularized primal problem:

$$\text{MOT}_c^\varepsilon(\mu, \nu) := \inf_{(t, \gamma) \in \mathbb{R} \times \Gamma^N_{\mu, \nu}} \left\{ t + \sum_{i=1}^{N} \varepsilon_i \text{KL}(\gamma_i | \mu \otimes \nu) \text{ s.t. } \forall i, \int_{X \times Y} c_i d\gamma_i = t \right\} \quad (6)$$

This problem can be compared with the one from standard regularized OT. Note that here we sum the generalized Kullback-Leibler divergences since our objective is function of $N$ measures in $\mathcal{M}_+(X \times Y)$. In the case where $N = 1$, we recover the standard regularized OT. Moreover, thanks to the entropic terms, the underlying problem becomes $\sum_{i=1}^{N} \varepsilon_i$ strongly convex. In the following proposition, we prove the essential property that as $\varepsilon \to 0$, the regularized problem converges to the standard problem. As a consequence, entropic regularization is a consistent approximation of the original problem we introduced in Section [3.2]. See Appendix [A.6] for the proof.
Algorithm 1 Projected Alternating Minimization Algorithm

Input: \( C = (C_i)_{1 \leq i \leq N}, a, b, \varepsilon, L \)
Init: \( f^0 \leftarrow 1_n, \; g^0 \leftarrow 1_m, \; \lambda^0 \leftarrow (1/N, ..., 1/N) \in \mathbb{R}^N \)
for \( k = 1, 2, \ldots \) do
  \[
  K^k \leftarrow \sum_{i=1}^N K_i^{k-1}, \quad c_k \leftarrow (f^{k-1}, K^k g^{k-1}), \quad f^k \leftarrow \frac{c_k}{K^k g^{k-1}},
  \]
  \[
  d_k \leftarrow (f^k, K^k g^{k-1}), \quad g^k \leftarrow \frac{d_k b}{(K^k)^T}, \quad \lambda^k \leftarrow \text{Proj}_{\Delta_N} \left( \lambda^{k-1} + \frac{1}{N} \nabla \mathcal{F}^\varepsilon_{\varepsilon} (\lambda^{k-1}, f^k, g^k) \right).
  \]
end
Result: \( \lambda, f, g \)

Proposition 4. For \((\mu, \nu) \in \mathcal{M}^+_N(\mathcal{X}) \times \mathcal{M}^+_N(\mathcal{Y})\) we have \( \lim_{\varepsilon \to 0} \text{MOT}^\varepsilon_c(\mu, \nu) = \text{MOT}_c(\mu, \nu) \).

Next theorem shows that strong duality holds for lower semi-continuous costs and compact spaces. This is the basis of the algorithm we will propose in Section 4.2. See Appendix C.3 for the proof.

Theorem 2 (Duality for the regularized problem). Let \( \mathcal{X} \) and \( \mathcal{Y} \) be two compact Polish spaces, \( c := (c_i)_{1 \leq i \leq N} \) a family of nonnegative lower semi-continuous costs on \( \mathcal{X} \times \mathcal{Y} \) and \( \varepsilon := (\varepsilon_i)_{1 \leq i \leq N} \) be nonnegative real numbers. For \((\mu, \nu) \in \mathcal{M}^+_N(\mathcal{X}) \times \mathcal{M}^+_N(\mathcal{Y})\), strong duality holds:

\[
\text{MOT}^\varepsilon_c(\mu, \nu) = \sup_{\lambda \in \Delta_N} \sup_{(f, g) \in C_\varepsilon(\mathcal{X}) \times C_\varepsilon(\mathcal{Y})} \left( \int_{\mathcal{X}} f(x) d\mu(x) + \int_{\mathcal{Y}} g(y) d\nu(y) - \sum_{i=1}^N \varepsilon_i \left( \int_{\mathcal{X} \times \mathcal{Y}} e^{f(x) + g(y) - \lambda c_i(x, y)} d\mu(x) d\nu(y) - 1 \right) \right),
\]

and the infimum of the primal problem is attained.

Remark 5. It is worth noting that the constraint on \( \lambda \) to live in \( \Delta_N \) in the dual formulation Eq. 2 of the entropic-based version can be restricted to the simplex \( \Delta^+_N \). This may have practical interest when deriving an efficient algorithm to compute MOT. See Appendix C.3 for the proof.

4.2 Proposed Algorithms

We can now present algorithms obtained from entropic relaxation to approximately compute the solution to MOT. Let \( \mu = \sum_{i=1}^n a_i \delta_{x_i} \) and \( \nu = \sum_{j=1}^m b_j \delta_{y_j} \) be discrete probability measures where \( a \in \Delta^+_N \), \( b \in \Delta^+_m \), \( \{x_1, ..., x_n\} \subset \mathcal{X} \) and \( \{y_1, ..., y_m\} \subset \mathcal{Y} \). Then the objective \( \text{MOT}^\varepsilon_c(a, b) \) can be written as

\[
\text{MOT}^\varepsilon_c(a, b) := \inf_{P \in \Gamma^\varepsilon_{\varepsilon}} \left\{ t - \sum_{i=1}^N \varepsilon_i H(P_i) \text{ s.t. } \forall i, \; \langle P_i, C_i \rangle = t \right\}
\]

where \( H(P) = \sum_{i,j} P_{ij} \log P_{ij} - 1 \) the discrete Shannon entropy, \( C := (C_i)_{1 \leq i \leq N} \in (\mathbb{R}^{n \times m})^N \) are \( N \) nonnegative cost matrices with \( C_i := (c_i(x_k, y_l))_{k,l} \) and \( \Gamma^\varepsilon_{\varepsilon,a,b} := \left\{ (P_i)_{1 \leq i \leq N} \in (\mathbb{R}^{n \times m})^N \text{ s.t. } (\sum_i P_i)1_m = a \text{ and } (\sum_i P_i^T)1_n = b \right\} \).

We deferred to Appendix B.2 more details on the discrete regularized case. Note that, as in standard OT, applying the envelope theorem to Eq. 3 gives the gradients of \( \text{MOT}^\varepsilon_c(a, b) \) with respect to \( a \) and \( b \). This property will be useful to compute barycenters in the following. From now on we consider the case where \( \varepsilon_1 = \cdots = \varepsilon_N = \varepsilon \). In fact deriving the dual of the problem \( \text{Eq. 3} \) leads to the discretized version of Eq. 7.

The main issue here is that the problem contains an additional variable \( \lambda \in \Delta^+_N \). Note that here we consider the simplex instead of the hyperplane for stability issues. When \( N = 1 \), one can use Sinkhorn algorithm. However when \( N \geq 2 \), we cannot apply directly an alternating minimization method to update \( \lambda \) by zeroing the gradient w.r.t. \( \lambda \). However, in order to enjoy from the strong convexity of the primal formulation, we
Figure 3: The two first row starting from the above are Wasserstein barycenters using respectively square $\ell_2$ cost and the cubic $\ell_3$ taken coordinate by coordinate. The last row represent the MOT barycenter with respect to these two costs. From left to right: Progressive barycentric transformation of the cross shape to the rectangle and circle shapes. Both shapes are normalized to probability distributions. One can notice that the separation of the cross into the rectangle and the circle induced by the MOT barycenter is more pronounced than for the ones induced by standard OT barycenters.

Consider instead the dual associated with the equivalent primal problem given when the additional trivial constraint $1^T_n (\sum_i P_i) 1_m = 1$ is considered. In that the dual obtained is

$$\text{MOT}^\varepsilon_C(a, b) = \sup_{\lambda \in \Delta^+_N} F^\varepsilon_C(\lambda, f, g) := \langle f, a \rangle + \langle g, b \rangle - \varepsilon \left[ \log \left( \sum_{i=1}^N \langle e^{f_i/\varepsilon}, K^\lambda_i e^{g_i/\varepsilon} \rangle \right) + 1 \right]$$

where $K^\lambda_i = \exp (-\lambda_i C_i/\varepsilon)$. We show that the new objective obtained above is smooth w.r.t $(\lambda, f, g)$. See Appendix C.4 for the proof. One can apply the accelerated projected gradient ascent [5, 36] which enjoys an optimal convergence rate for first order methods of $O(k^{-2})$. We denote by $\text{Proj}_{\Delta^+_N}$ the orthogonal projection on $\Delta^+_N$ [31], whose complexity is in $O(N \log N)$.

So far, it is also possible to adapt Sinkhorn algorithm to our problem. See Alg. 1. The smoothness constant in $\lambda$ in the algorithm is $L_\lambda = \max_i \|C_i\|_\infty^2/\varepsilon$. In practice Alg. 1 gives better results than the accelerated gradient descent. Further work will be devoted to study the complexity of this algorithm.

**Illustration with barycenters.** To illustrate MOT, we designed an experiment with barycenters between two distributions [1]. Whereas one need to solve costful linear programs to exactly compute such barycenters, we adapt the proposed method from [11] and use Alg. 1 to compute them more efficiently. Recall that the algorithm used in [11] is an accelerated mirror descent which needs the gradient of the MOT w.r.t. distributions to decrease the overall objective. Figure 3 displays the barycenters induced by the MOT problem compared to standard OT barycenters. We deferred in Appendix D additional barycenters experiments on MNIST dataset [23], and we also provide experiments on the Dudley Metric to show the accuracy of the proposed Alg. 1 w.r.t. the regularization $\varepsilon$.

**Conclusion and future work.** In this paper, we introduced a new variational problem to deal with multiple costs in OT by splitting the transportation problem among the costs such that all costs must contribute equally to the global transport problem. Following the idea of [10], we derived an entropic relaxation and an efficient algorithm to approximately compute solutions to our problem.
References

[1] M. Agueh and G. Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] D. Alvarez-Melis, S. Jegelka, and T. S. Jaakkola. Towards optimal transport with global invariances. *arXiv preprint 1806.09277*, 2018.

[3] M. Arjovsky, S. Chüntala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[4] M.-F. F. Balcan, T. Dick, R. Noothigattu, and A. D. Procaccia. Envy-free classification. In *Advances in Neural Information Processing Systems*, pages 1238–1248, 2019.

[5] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[6] D. Bertsimas and J. N. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1997.

[7] S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[8] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media, 2010.

[9] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

[10] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[11] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *Proceedings of ICML*, volume 32, pages 685–693, 2014.

[12] R. M. Dudley. The speed of mean Glivenko-Cantelli convergence. *Annals of Mathematical Statistics*, 40(1):40–50, 1969.

[13] R. M. Dudley et al. Weak convergence of probabilities on nonseparable metric spaces and empirical measures on euclidean spaces. *Illinois Journal of Mathematics*, 10(1):109–126, 1966.

[14] P. Dupuis and R. S. Ellis. *A weak convergence approach to the theory of large deviations*, volume 902. John Wiley & Sons, 2011.

[15] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015.

[16] J. Feydy, T. Séjourné, F.-X. Vialard, S.-I. Amari, A. Trouvé, and G. Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. *arXiv preprint arXiv:1810.08278*, 2018.

[17] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

[18] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of sinkhorn divergences. *arXiv preprint arXiv:1810.02733*, 2018.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[20] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.
[21] H. Husain, R. Nock, and R. C. Williamson. A primal-dual link between gans and autoencoders. In *Advances in Neural Information Processing Systems*, pages 413–422, 2019.

[22] H. Janati, T. Bazeille, B. Thirion, M. Cuturi, and A. Gramfort. Multi-subject meg/eeg source imaging with sparse multi-task regression. *NeuroImage*, page 116847, 2020.

[23] Y. LeCun, C. Cortes, and C. Burges. The mnist dataset of handwritten digits. 1998.

[24] Y. Mroueh and T. Sercu. Fisher gan. In *Advances in Neural Information Processing Systems*, pages 2513–2523, 2017.

[25] A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[26] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

[27] F.-P. Paty and M. Cuturi. Subspace robust wasserstein distances. *arXiv preprint arXiv:1901.08949*, 2019.

[28] T. Salimans, H. Zhang, A. Radford, and D. Metaxas. Improving GANs using optimal transport. In *International Conference on Learning Representations*, 2018.

[29] M. Scetbon and G. Varoquaux. Comparing distributions: $\ell_1$ geometry improves kernel two-sample testing, 2019.

[30] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

[31] S. Shalev-Shwartz and Y. Singer. Efficient learning of label ranking by soft projections onto polyhedra. *Journal of Machine Learning Research*, 7(Jul):1567–1599, 2006.

[32] M. Sion. On general minimax theorems. *Pacific J. Math.*, 8(1):171–176, 1958.

[33] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, G. R. Lanckriet, et al. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

[34] T. Steerneman. On the total variation and hellinger distance between signed measures; an application to product measures. *Proceedings of the American Mathematical Society*, 88(4):684–688, 1983.

[35] R. E. Tarjan. Dynamic trees as search trees via euler tours, applied to the network simplex algorithm. *Mathematical Programming*, 78(2):169–177, 1997.

[36] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 1, 2008.

[37] C. Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

[38] W. Wang and M. A. Carreira-Perpinan. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application, 2013.

[39] E. Wong, F. R. Schmidt, and J. Z. Kolter. Wasserstein adversarial examples via projected sinkhorn iterations, 2019.

[40] K. D. Yang, K. Damodaran, S. Venkatachalapathy, A. C. Soylemezoglu, G. Shivashankar, and C. Uhler. Predicting cell lineages using autoencoders and optimal transport. *PLoS computational biology*, 16(4):e1007828, 2020.
Supplementary material

Outline. In Sec. A we provide the proofs of the propositions and theorems given in the main paper. In Sec. B we consider the discretized problems and obtain similar results. In Sec. C we show additional properties of the problem considered. Finally in Sec. D we give more illustrations of barycenters and show the effect of the regularization on the accuracy of our proposed algorithm.

A Proofs

A.1 Notations

Let $Z$ be a Polish space, we denote $\mathcal{M}(Z)$ the set of Radon measures on $Z$ endowed with total variation norm: $\|\mu\|_{TV} = \mu_+(Z) + \mu_-(Z)$ with $\mu_+, \mu_-$ is the Dunford decomposition of the signed measure $\mu$. We call $\mathcal{M}_+(Z)$ the sets of positive Radon measures, and $\mathcal{M}^+_1(Z)$ the set of probability measures. We denote $C_b(Z)$ the vector space of bounded continuous functions on $Z$ endowed with $\|\cdot\|_\infty$ norm. We recall the Riesz-Markov theorem: if $Z$ is compact, $\mathcal{M}(Z)$ is the topological dual of $C_b(Z)$. Let $X$ and $Y$ be two Polish spaces. It is immediate that $X \times Y$ is a Polish space. We denote $\Pi_1 : (x,y) \in X \times Y \mapsto x$ and $\Pi_2 : (x,y) \in X \times Y \mapsto y$ respectively the projections on $X$ and $Y$, which are continuous applications. For an application $g$ and a measure $\mu$, we denote $\gamma_{\mu} g$ the pushforward measure of $\mu$ by $g$. For $f : X \to \mathbb{R}$ and $g : Y \to \mathbb{R}$, we denote $f \oplus g : (x,y) \in X \times Y \mapsto f(x) + g(y)$ the tensor sum of $f$ and $g$. For $X$ and $Y$ two Polish spaces, we denote $\text{LSC}(X \times Y)$ the space of lower semi-continuous functions on $X \times Y$.

A.2 Proof of Theorem 1

To prove this theorem, we use an equivalent form of the problem, presented in Section C.

Proof. In Proposition 11, we proved that

$$\text{MOT}_e(\mu, \nu) = \sup_{\lambda \in \Delta_+^N} \sup_{(f,g) \in \mathcal{F}^+} \int_{x \in X} f(x)d\mu(x) + \int_{y \in Y} g(y)d\nu(y).$$

Let show that we can extend the search space for $\lambda$ to $\Delta_N$. Denote by

$$\mathcal{F}_c := \{ (f,g) \in C_b(X) \times C_b(Y) \text{ s.t. } \exists \lambda \in \Delta_N, \forall i, f \oplus g \leq \lambda_i c_i \}$$

and

$$\mathcal{F}_c^+ := \{ (f,g) \in C_b(X) \times C_b(Y) \text{ s.t. } \exists \lambda \in \Delta_N^+, \forall i, f \oplus g \geq \lambda_i c_i \}.$$ 

Moreover by denoting

$$\mathcal{G}_c := \{ (f,g) \in \mathcal{F}_c \in C_b(X) \times C_b(Y) \text{ s.t. } \exists \lambda \in \Delta_N \setminus \Delta_N^+, \forall i, f \oplus g \leq \lambda_i c_i \} ,$$

we have therefore that $\mathcal{F}_c = \mathcal{F}_c^+ \cup \mathcal{G}_c$ and it is clear

$$\sup_{(f,g) \in \mathcal{F}_c} \int_{x \in X} f(x)d\mu(x) + \int_{y \in Y} g(y)d\nu(y) \leq \sup_{(f,g) \in \mathcal{F}_c} \int_{x \in X} f(x)d\mu(x) + \int_{y \in Y} g(y)d\nu(y)$$

Let us now consider $(f,g) \in \mathcal{G}_c$. Therefore there exists a $\lambda \in \Delta_N$ and $k$ such that $\lambda_k < 0$, so that $f \oplus g \leq \lambda_k c_k$, which by positivity of the cost functions leads that

$$\sup_{(f,g) \in \mathcal{G}_c} \int_{x \in X} f(x)d\mu(x) + \int_{y \in Y} g(y)d\nu(y) \leq 0$$

(8)
But thanks to Proposition \ref{prop:supremum}, we have that

$$\text{MOT}_c(\mu, \nu) = \sup_{(f,g) \in \mathcal{F}_c^+} \int_{x \in \mathcal{X}} f(x)d\mu(x) + \int_{y \in \mathcal{Y}} g(y)d\nu(y).$$

But by definition, \(\text{MOT}_c(\mu, \nu)\) is non-negative, therefore we have that

$$\sup_{(f,g) \in \mathcal{G}_c} \int_{x \in \mathcal{X}} f(x)d\mu(x) + \int_{y \in \mathcal{Y}} g(y)d\nu(y) \leq \sup_{(f,g) \in \mathcal{F}_c^+} \int_{x \in \mathcal{X}} f(x)d\mu(x) + \int_{y \in \mathcal{Y}} g(y)d\nu(y)$$

and the result follows.

### A.3 Proof of Proposition \ref{prop:supremum}

**Proof.** As shown in the proof of Proposition \ref{prop:supremum}, we have that for any \(\lambda \in \Delta_N\)

$$\sup_{(f,g) \in \mathcal{F}_c^\lambda} \int_{x \in \mathcal{X}} f(x)d\mu(x) + \int_{y \in \mathcal{Y}} g(y)d\nu(y) \leq \sup_{(f,g) \in \mathcal{G}_c^\lambda} \inf_{(\mu_k, \nu_k)_{k=1}^N \in \mathcal{T}_{\mu, \nu}^N} \sum_{k=1}^N \lambda_k \left[ \int_{x \in \mathcal{X}} f_k(x)d\mu_k(x) + \int_{y \in \mathcal{Y}} g_k(y)d\nu_k(y) \right] \leq \text{MOT}_c(\mu, \nu)$$

Then by taking the supremum over \(\lambda \in \Delta_N\), and by applying Theorem \ref{thm:sion} we obtain that

$$\text{MOT}_c(\mu, \nu) = \sup_{\lambda \in \Delta_N} \sup_{(f,g) \in \mathcal{F}_c^\lambda} \inf_{(\mu_k, \nu_k)_{k=1}^N \in \mathcal{T}_{\mu, \nu}^N} \sum_{k=1}^N \lambda_k \left[ \int_{x \in \mathcal{X}} f_k(x)d\mu_k(x) + \int_{y \in \mathcal{Y}} g_k(y)d\nu_k(y) \right]$$

Then applying Sion's theorem \ref{thm:sion} gives

$$\text{MOT}_c(\mu, \nu) = \sup_{(f,g) \in \mathcal{G}_c^N} \inf_{(\mu_k, \nu_k)_{k=1}^N \in \mathcal{T}_{\mu, \nu}^N} \sup_{\lambda \in \Delta_N} \sum_{k=1}^N \lambda_k \left[ \int_{x \in \mathcal{X}} f_k(x)d\mu_k(x) + \int_{y \in \mathcal{Y}} g_k(y)d\nu_k(y) \right]$$

Let us now fix \((f_k, g_k)_{k=1}^N \in \mathcal{G}_c^N\) and \((\mu_k, \nu_k)_{k=1}^N \in \mathcal{T}_{\mu, \nu}^N\), therefore we have:

$$\sup_{\lambda \in \Delta_N} \sum_{k=1}^N \lambda_k \left[ \int_{x \in \mathcal{X}} f_k(x)d\mu_k(x) + \int_{y \in \mathcal{Y}} g_k(y)d\nu_k(y) \right] = \sup_{\lambda} \inf_{t} t \times \left( 1 - \sum_{i=1}^N \lambda_i \right) + \sum_{k=1}^N \lambda_k \left[ \int_{x \in \mathcal{X}} f_k(x)d\mu_k(x) + \int_{y \in \mathcal{Y}} g_k(y)d\nu_k(y) \right]$$

$$= \inf_{t} \sum_{k=1}^N \lambda_k \left[ \int_{x \in \mathcal{X}} f_k(x)d\mu_k(x) + \int_{y \in \mathcal{Y}} g_k(y)d\nu_k(y) - t \right]$$

$$= \inf_{t} \left\{ t \text{ s.t. } \forall k, \int f_k d\mu_k + \int g_k d\nu_k = t \right\}$$

where the inversion is possible as the Slater’s conditions are satisfied and the result follows.

### A.4 Proof of Proposition \ref{prop:inversion}

**Proof.** Before proving the result let us first introduce the following lemma.
Lemma 1. Let \( \mathcal{X} \) and \( \mathcal{Y} \) be Polish spaces. Let \( \mathbf{c} := (c_i)_{1 \leq i \leq N} \) a family of nonnegative continuous costs. For \( (x, y) \in \mathcal{X} \times \mathcal{Y} \) and \( \lambda \in \Delta_N^+ \), we define

\[
c_\lambda(x, y) := \min_{i=1, \ldots, N} \left( \lambda_i c_i(x, y) \right)
\]

then for any \((\mu, \nu) \in \mathcal{M}_+^1(\mathcal{X}) \times \mathcal{M}_+^1(\mathcal{Y})\)

\[
\text{MOT}_\mathbf{c}(\mu, \nu) = \sup_{\lambda \in \Delta_N^+} W_{c_\lambda}(\mu, \nu)
\]

(9)

Proof. Let \((\mu, \nu) \in \mathcal{M}_+^1(\mathcal{X}) \times \mathcal{M}_+^1(\mathcal{Y})\) and \(\mathbf{c} := (c_i)_{1 \leq i \leq N}\) cost functions on \( \mathcal{X} \times \mathcal{Y} \). Let \( \lambda \in \Delta_N^+ \), then by Proposition [7]

\[
\text{MOT}_\mathbf{c}(\mu, \nu) = \sup_{\lambda \in \Delta_N^+} \sup_{(f, g) \in \mathcal{F}_\lambda} \int \mathcal{X} f(x)\,d\mu(x) + \int \mathcal{Y} g(y)\,d\nu(y)
\]

Therefore by denoting \( c_\lambda := \min_i (\lambda_i c_i) \) which is a continuous. The dual form of the classical Optimal Transport problem gives that:

\[
\sup_{(f, g) \in \mathcal{F}_\lambda} \int \mathcal{X} f(x)\,d\mu(x) + \int \mathcal{Y} g(y)\,d\nu(y) = W_{c_\lambda}(\mu, \nu)
\]

and the result follows.

Let us now prove the result. Let \( \mu \) and \( \nu \) be two probability measures. Let \( \gamma \in (0, 1) \). Note that if \( d \) is a metric then \( d^\gamma \) too. Therefore in the following we consider \( d \) a general metric on \( \mathcal{X} \times \mathcal{X} \). Let \( c_1 : (x, y) \to 2 \times 1_{x \neq y} \) and \( c_2 = d \). For all \( \alpha \in [0, 1) \):

\[
c_\alpha(x, y) := \min(\alpha c_1(x, y), (1 - \alpha)c_2(x, y)) = \min(2\alpha, (1 - \alpha)d(x, y))
\]

defines a distance on \( \mathcal{X} \times \mathcal{X} \). Then according to [37, Theorem 1.14]:

\[
W_{c_\alpha}(\mu, \nu) = \sup_{f \text{ s.t. } f \in L_{\alpha, \text{Lipschitz}}} \int f\,d\mu - \int f\,d\nu
\]

Then thanks to Lemma [7] we have

\[
\text{MOT}_{(c_1, c_2)}(\mu, \nu) = \alpha \in [0, 1], f \text{ s.t. } f \in L_{\alpha, \text{Lipschitz}} \int f\,d\mu - \int f\,d\nu
\]

Let now prove that in this case: \( \text{MOT}_{(c_1, c_2)}(\mu, \nu) = \beta_d(\mu, \nu) \). Let \( \alpha \in [0, 1) \) and \( f \) a \( c_\alpha \) Lipschitz function. \( f \) is lower bounded: let \( m = \inf f \) and \( (u_n)_n \) a sequence satisfying \( f(u_n) \to m \). Then for all \( x, y \), \( f(x) - f(y) \leq 2\alpha \) and \( f(x) - f(y) \leq (1 - \alpha)d(x, y) \). Let define \( g = f - m - \alpha \). For \( x \) fixed and for all \( n \), \( f(x) - f(u_n) \leq 2\alpha \), so taking the limit in \( n \) we get \( f(x) - m \leq 2\alpha \). So we get that for all \( x, y \), \( g(x) \in [\alpha, \alpha + \alpha] \) and \( g(x) - g(y) \in [-\alpha, \alpha + \alpha] \). Then \( ||g||_\infty \leq \alpha \) and \( ||g||_d \leq 1 - \alpha \). By construction, we also have \( \int f\,d\mu - \int f\,d\nu = \int g\,d\mu - \int g\,d\nu \). Then \( ||g||_\infty + ||g||_d \leq 1 \). So we get that \( \text{MOT}_{(c_1, c_2)}(\mu, \nu) \leq \beta_d(\mu, \nu) \). Reciprocally, let \( g \) be a function satisfying \( ||g||_\infty + ||g||_d \leq 1 \). Let define \( f = g + ||g||_\infty \) and \( \alpha = ||g||_\infty \). Then, for all \( x, y \), \( f(x) \in [0, 2\alpha] \) and so \( f(x) - f(y) \leq 2\alpha \). It is immediate that \( f(x) - f(y) \leq \min(\alpha, (1 - \alpha)d(x, y)) \). Then we get \( f(x) - f(y) \leq \min(\alpha, (1 - \alpha)d(x, y)) \). And by construction, we still have \( \int g\,d\mu - \int g\,d\nu = \int f\,d\mu - \int f\,d\nu \). So \( \text{MOT}_{(c_1, c_2)}(\mu, \nu) \geq \beta_d(\mu, \nu) \).

Finally we get \( \text{MOT}_{(c_1, c_2)}(\mu, \nu) = \beta_d(\mu, \nu) \) when \( c_1 : (x, y) \to 2 \times 1_{x \neq y} \) and \( c_2 = d \) a distance on \( \mathcal{X} \times \mathcal{X} \).
A.5 Proof of Proposition 3

Proof. Let \( \mu \) and \( \nu \) be two probability measures respectively on \( \mathcal{X} \) and \( \mathcal{Y} \). Let \( c := (c_i)_i \) be a family of cost functions. Let define for \( \lambda \in \Delta^+_N \), \( c_\lambda(x,y) := \min_i (\lambda_i c_i(x,y)) \). We have, by linearity \( W_{c_\lambda}(\mu,\nu) \leq \min_i (\lambda_i W_{c_i}(\mu,\nu)) \). So we deduce by Lemma 7:

\[
\text{MOT}_c(\mu,\nu) = \sup_{\lambda \in \Delta^+_N} W_{c_\lambda}(\mu,\nu) \\
\leq \sup_{\lambda \in \Delta^+_N} \min_i \lambda_i W_{c_i}(\mu,\nu) \\
= \frac{1}{\sum_i W_{c_i}(\mu,\nu)}
\]

which concludes the proof.

A.6 Proof of Proposition 4

Proof. Let \((\varepsilon_i)_i\) a sequence converging to 0. Let \( \gamma_l \) be the optimum of \( \text{MOT}_c^{\varepsilon_i}(\mu,\nu) \). By Lemma 2, up to an extraction, \( \gamma_l \to \gamma^* \in \Gamma^+_N \). Let now \( \gamma \) be the optimum of \( \text{MOT}_c(\mu,\nu) \). By optimality of \( \gamma \) and \( \gamma_l \), for all \( i \):

\[
0 \leq \int c_i d\gamma_{l,i} - \int c_i d\gamma_i \leq \sum_i \varepsilon_{l,i} (\text{KL}(\gamma_i || \mu \otimes \nu) - \text{KL}(\gamma_{l,i} || \mu \otimes \nu))
\]

By lower semi continuity of \( \text{KL}(\cdot || \mu \otimes \nu) \) and by taking the limit inferior as \( l \to \infty \), we get for all \( i \),

\[
\liminf_{l \to \infty} \int c_i d\gamma_{l,i} = \int c_i d\gamma_i.
\]

Moreover by continuity of \( \gamma \to \int c_i d\gamma_i \) we therefore obtain that for all \( i \),

\[
\int c_i d\gamma_i^* \leq \int c_i d\gamma_i.
\]

Then by optimality of \( \gamma \) the result follows.

A.7 Proof of Theorem 2

Proof. To show the strong duality of the regularized problem, we use the same sketch of proof as for the strong duality of the original problem. Let first assume that, for all \( i \), \( c_i \) is continuous on the compact set \( \mathcal{X} \times \mathcal{Y} \). Let fix \( \lambda \in \Delta_N \). We define, for all \( u \in C^b(\mathcal{X} \times \mathcal{Y}) \):

\[
V_i^\lambda(u) = \varepsilon_i \left( \int_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \exp \frac{-u(x,y) - \lambda_i c_i(x,y)}{\varepsilon_i} d\mu(x)d\nu(y) - 1 \right)
\]

and:

\[
E(u) = \left\{ \begin{array}{ll}
\int f d\mu + \int g d\nu & \text{if } \exists (f,g) \in C^b(\mathcal{X}) \times C^b(\mathcal{Y}), u = f + g \\
+\infty & \text{else}
\end{array} \right.
\]

Let compute the Fenchel-Legendre transform of these functions. Let \( \gamma \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}) \):

\[
V_i^{\lambda^*}(-\gamma) = \sup_{u \in C^b(\mathcal{X} \times \mathcal{Y})} -\int u d\gamma - \varepsilon_i \left( \int_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \exp \frac{-u(x,y) - \lambda_i c_i(x,y)}{\varepsilon_i} d\mu(x)d\nu(y) - 1 \right)
\]

However, by density of \( C^b(\mathcal{X} \times \mathcal{Y}) \) in \( L^1_{d\mu \otimes d\nu}(\mathcal{X} \times \mathcal{Y}) \) we deduce that

\[
V_i^{\lambda^*}(-\gamma) = \sup_{u \in L^1_{d\mu \otimes d\nu}(\mathcal{X} \times \mathcal{Y})} -\int u d\gamma - \varepsilon_i \left( \int_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \exp \frac{-u(x,y) - \lambda_i c_i(x,y)}{\varepsilon_i} d\mu(x)d\nu(y) - 1 \right)
\]
This supremum equals $+\infty$ if $\gamma$ is not positive and not absolutely continuous with regard to $\mu \otimes \nu$. Let us now denote $F_{\gamma,\lambda}(u) := -\int u \, d\gamma - \varepsilon_i \left( \int_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \exp \frac{u(x,y) - \lambda_i c_i (x,y)}{\varepsilon_i} \, d\mu(x) d\nu(y) - 1 \right)$. $F_{\gamma,\lambda}$ is Fréchet differentiable and its maximum is attained for $u^* = \varepsilon_i \log \left( \frac{d\gamma_i}{d\mu \otimes \nu} \right) + \lambda_i c_i$. Therefore we obtain that

$$V_i^{\lambda^*}(-\gamma) = \varepsilon_i \left( \int \log \left( \frac{d\gamma_i}{d\mu \otimes \nu} \right) \, d\gamma + 1 - \gamma(\mathcal{X} \times \mathcal{Y}) \right) + \lambda_i \int c_i d\gamma$$

Hence $\lambda^* \in \Gamma_{\mu,\nu}^N$. Moreover $\sup_{\lambda \in \Delta_N} \sum_{i=1}^N \lambda_i \int c_i d\gamma_i + \varepsilon_i \log (\gamma_i \| \mu \otimes \nu)$

Let $f : (\lambda,\gamma) \in \Delta_N \times \Gamma_{\mu,\nu}^N \rightarrow \sum_{i=1}^N \lambda_i \int c_i d\gamma_i + \varepsilon_i \log (\gamma_i \| \mu \otimes \nu)$. $f$ is clearly concave and continuous in $\lambda$. Moreover $\gamma \mapsto \log (\gamma_i \| \mu \otimes \nu)$ is convex and lower semi-continuous for weak topology [14, Lemma 1.4.3]. Hence $f$ is convex and lower-semi-continuous in $\gamma$. $\Delta_N$ is convex, and $\Gamma_{\mu,\nu}^N$ is compact for weak topology (see Lemma [4]). So by Sion’s theorem, we get the expected result:

$$\min_{\gamma \in \Gamma_{\mu,\nu}^N} \max_{\lambda \in \Delta_N} \sum_{i=1}^N \lambda_i \int c_i d\gamma_i + \sum_{i=1}^N \varepsilon_i \log (\gamma_i \| \mu \otimes \nu)$$

$$= \sup_{\lambda \in \Delta_N} \left( \int_{\mathcal{X} \times \mathcal{Y}} f(x) d\mu(x) + \int_{\mathcal{Y}} g(y) d\nu(y) \right) - \sum_{i=1}^N \frac{\varepsilon_i}{\gamma_i} \left( f(x) + g(y) \frac{\lambda_i c_i (x,y)}{\varepsilon_i} \right) \, d\mu(x) d\nu(y) - 1$$

16
Moreover by fixing $\gamma \in \Gamma^N_{\mu, \nu}$, we have

$$\sup_{\lambda \in \Delta_N} \sum_i \lambda_i \int c_i d\gamma_i + \sum_i \varepsilon_i \KL(\gamma_i || \mu \otimes \nu)$$

$$= \sup_{\lambda \in \Delta_N} \inf_i \left( 1 - \sum_{i=1}^{N} \lambda_i + \sum_i \lambda_i \int c_i d\gamma_i + \sum_i \varepsilon_i \KL(\gamma_i || \mu \otimes \nu) \right)$$

$$= \inf_t \sup_{\lambda \in \Delta_N} t \times \left( 1 - \sum_{i=1}^{N} \lambda_i + \sum_i \lambda_i \int c_i d\gamma_i + \sum_i \varepsilon_i \KL(\gamma_i || \mu \otimes \nu) \right)$$

$$= \inf_t \sup_{\lambda \in \Delta_N} \left\{ t + \sum_{i=1}^{N} \varepsilon_i \KL(\gamma_i || \mu \otimes \nu) \right\} \text{ s.t. } \forall i, \int_{\mathcal{X} \times \mathcal{Y}} c_i d\gamma_i = t$$

$$= \MOT^\varepsilon(\mu, \nu)$$

A similar proof as the one of the Theorem 3 allows to extend the results for lower semi-continuous cost functions.

B Discrete cases

B.1 Exact discrete case

Let $a \in \Delta_N^+$ and $b \in \Delta_m^+$ and $C := (C_i)_{1 \leq i \leq N} \in \left( \mathbb{R}^{n \times m}_+ \right)^N$ be $N$ cost matrices. Let also $X := \{x_1, ..., x_n\}$ and $Y := \{y_1, ..., y_m\}$ two subset of $\mathcal{X}$ and $\mathcal{Y}$ respectively. Moreover we define the two following discrete measure

$$\mu = \sum_i a_i \delta_{x_i} \quad \text{and} \quad \nu = \sum_i b_i \delta_{y_i} \quad \text{and for all } i, C_i = (c(x_k, y_l))_{1 \leq k \leq n, 1 \leq l \leq m} \text{ where } (c_i)_{i=1}^N \text{ a family of cost functions.}$$

The discretized multiple cost optimal transport primal problem can be written as follows:

$$\MOT_C(\mu, \nu) = \MOT^\varepsilon_C(a, b) := \inf_{P \in \Gamma^N_{a, b}} \{ t \text{ s.t. } \forall i, \langle P, C_i \rangle = t \}$$

where $\Gamma^N_{a, b} := \{(P_i)_{1 \leq i \leq N} \in \left( \mathbb{R}^{n \times m}_+ \right)^N \text{ s.t. } (\sum_i P_i)1_m = a \text{ and } (\sum_i P_i^T)1_n = b \}$. As in the continuous case, strong duality holds and we can rewrite the dual in the discrete case also.

**Proposition 5** (Duality for the discrete problem). Let $a \in \Delta_N^+$ and $b \in \Delta_m^+$ and $C := (C_i)_{1 \leq i \leq N} \in \left( \mathbb{R}^{n \times m}_+ \right)^N$ be $N$ cost matrices. Strong duality holds for the discrete problem and

$$\MOT_C(a, b) = \sup_{\lambda \in \Delta_N} \sup_{(f, g) \in \mathcal{F}_C} \langle f, a \rangle + \langle g, b \rangle.$$  

where $\mathcal{F}_C := \{(f, g) \in \mathbb{R}^n_+ \times \mathbb{R}^m_+ \text{ s.t. } \forall i \in \{1, ..., N\}, f1_m + g1_n \leq \lambda_i C_i \}$. 

B.2 Entropic regularized discrete case

We now extend the regularization in the discrete case. Let $a \in \Delta_N^+$ and $b \in \Delta_m^+$ and $C := (C_i)_{1 \leq i \leq N} \in \left( \mathbb{R}^{n \times m}_+ \right)^N$ be $N$ cost matrices and $\varepsilon = (\varepsilon_i)_{1 \leq i \leq N}$ be nonnegative real numbers. The discretized regularized primal problem is:

$$\MOT^\varepsilon_C(a, b) = \inf_{P \in \Gamma^N_{a, b}} \left\{ t - \sum_{i=1}^{N} \varepsilon_i H(P_i) \text{ s.t. } \forall i, \langle P_i, C_i \rangle = t \right\}$$

where $H(P) = \sum_{i,j} P_{i,j} (\log P_{i,j} - 1)$ for $P = (P_{i,j})_{i,j} \in \mathbb{R}^{n \times m}$ is the discrete entropy. In the discrete case, strong duality holds thanks to Lagrangian duality and Slater sufficient conditions.
Proposition 6 (Duality for the discrete regularized problem). Let \( a \in \Delta_N^+ \) and \( b \in \Delta_m \) and \( C := (C_i)_{1 \leq i \leq N} \in (\mathbb{R}^{n \times m})^N \) be \( N \) cost matrices and \( \varepsilon := (\varepsilon_i)_{1 \leq i \leq N} \) be non negative reals. Strong duality holds and by denoting \( K_i^\lambda = \exp (-\lambda_i C_i / \varepsilon_i) \), we have

\[
\hat{\text{MOT}}_C^\varepsilon (a, b) = \sup_{\lambda \in \Delta_N} \sup_{f \in \mathbb{R}^N, \ g \in \mathbb{R}^m} \langle f, a \rangle + \langle g, b \rangle - \sum_{i=1}^N \varepsilon_i \langle \varepsilon_i, K_i^\lambda \rangle.
\]

The objective function for the dual problem is strictly concave in \((\lambda, f, g)\) but is neither smooth or strongly convex.

Proof. The proofs in the discrete case are simpler and only involves Lagrangian duality [7, Chapter 5]. Let do the proof in the regularized case, the one for the standard problem follows exactly the same path.

Let \( a \in \Delta_N^+ \) and \( b \in \Delta_m \) and \( C := (C_i)_{1 \leq i \leq N} \in (\mathbb{R}^{n \times m})^N \) be \( N \) cost matrices.

\[
\hat{\text{MOT}}_C^\varepsilon (a, b) = \inf_{P \in \mathbb{T}_a, 1 \leq i \leq N} \sup_{\{P_i, C_i\} \in (\mathbb{R}^{n \times m})^N} \langle P_i, C_i \rangle - \sum_{i=1}^N \varepsilon_i H(P_i)
\]

\[
= \inf_{(t, P) \in \mathbb{R} \times \mathbb{R}^{n \times m}} \sup_{\sum_{i=1}^N P_i 1_m = a} \sup_{\sum_{i=1}^N P_i 1_n = b} \left[ t + \sum_{j=1}^N \lambda_j (\langle P_j, C_j \rangle - t) - \sum_{i=1}^N \varepsilon_i H(P_i) \right] + f^T \left( a - \sum_{i=1}^N P_i 1_m \right) + g^T \left( b - \sum_{i=1}^N P_i 1_n \right)
\]

The constraints are qualified for this convex problem, hence by Slater’s sufficient condition [7, Section 5.2.3], strong duality holds and:

\[
\hat{\text{MOT}}_C^\varepsilon (a, b) = \sup_{f \in \mathbb{R}^n, \ g \in \mathbb{R}^m, \ \lambda \in \Delta_N} \inf_{(t, P) \in \mathbb{R} \times \mathbb{R}^{n \times m}} \left[ t + \sum_{j=1}^N \lambda_j (\langle P_j, C_j \rangle - t) - \sum_{i=1}^N \varepsilon_i H(P_i) \right] + f^T \left( a - \sum_{j=1}^N P_j 1_m \right) + g^T \left( b - \sum_{j=1}^N P_j 1_n \right)
\]

\[
= \sup_{f \in \mathbb{R}^n, \ g \in \mathbb{R}^m} \inf_{\lambda \in \Delta_N} \left[ \langle P_j, \lambda_j C_j - f 1_n^T - 1_m g^T \rangle - \varepsilon_j H(P_j) \right]
\]

But for every \( i = 1, \ldots, N \) the solution of

\[
\inf_{P_j \in \mathbb{R}^{n \times m}} \left( \langle P_j, \lambda_j C_j - f 1_n^T - 1_m g^T \rangle - \varepsilon_j H(P_j) \right)
\]

is

\[
P_j = \exp \left( \frac{f 1_n^T + 1_m g^T - \lambda_j C_j}{\varepsilon_i} \right)
\]
Finally we obtain that

$$\text{MOT}_c^\varepsilon(a, b) = \sup_{f \in \mathbb{R}^m, g \in \mathbb{R}^m, \lambda \in \Delta_N^x} (f, a) + (g, b) - \sum_{k=1}^N \varepsilon_k \sum_{i,j} \exp \left( \frac{f_i + g_j - \lambda_k C_k^{(i,j)}}{\varepsilon_k} \right)$$

C Other Useful Properties

C.1 The dual supremum is attained for continuous cost on compact sets

Proposition 7. Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish compact spaces. Let $c := (c_i)_{i=1}^N$ be nonnegative lower continuous costs. Then for $(\mu, \nu) \in \mathcal{M}_+^1(\mathcal{X}) \times \mathcal{M}_+^1(\mathcal{Y})$

$$\sup_{\lambda \in \Delta_N} \sup_{(f,g) \in \mathcal{F}^a} \int_{x \in \mathcal{X}} f(x) d\mu(x) + \int_{y \in \mathcal{Y}} g(y) d\nu(y)$$

is attained.

Proof 1. Let recall that, from standard optimal transport results:

$$\text{MOT}_c(\mu, \nu) = \sup_{u \in \Phi_c} \int u d\mu d\nu$$

with $\Phi_c := \{u \in C^b(\mathcal{X} \times \mathcal{Y}) \text{s.t. } \exists \lambda \in \Delta_N^x, \exists \phi \in C^b(\mathcal{X}), \text{ u} = \phi^c \oplus \phi^c \text{ with } c = \min_i \lambda_i c_i \}$ where $\phi^c$ is the c-transform of $\phi$, i.e. for $y \in \mathcal{Y}$, $\phi^c(y) = \inf_{x \in \mathcal{X}} c(x, y) - \phi(x)$.

Let denote $\omega_1, \ldots, \omega_N$ the continuity moduli of $c_1, \ldots, c_N$. Then a modulus of continuity for $\min_i \lambda_i c_i$ is $\sum_i \lambda_i \omega_i$. As $\phi^c$ and $\phi^c$ share the same modulus of continuity than $c = \min_i \lambda_i c_i$, for $u$ is $\Phi_c$, a common modulus of continuity is $2 \times \sum_i \omega_i$. More over, it is clear that for all $x, y$, $\{u(x, y) \text{ s.t. } u \in \Phi_c\}$ is compact. Then, applying Ascoli's theorem, we get, that $\Phi_c$ is compact for $\|\|_\infty$ norm. By continuity of $u \rightarrow \int u d\mu d\nu$, the supremum is attained, and we get the existence of the optimum $u^*$. The existence of optimas $(\lambda^*, f^*, g^*)$ immediately follows.

C.2 MOT generalizes OT

Proposition 8. Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces. Let $N \geq 0$, $c = (c_i)_{1 \leq i \leq N}$ be a family of nonnegative lower semi-continuous costs and let us denote for all $k \in \{1, \ldots, N\}$, $c_k = (c_i)_{1 \leq i \leq k}$. Then for all $k \in \{1, \ldots, N\}$, there exists a family of costs $d_k \in LSC(\mathcal{X} \times \mathcal{Y})^N$ such that

$$\text{MOT}_{d_k}(\mu, \nu) = \text{MOT}_c(\mu, \nu)$$

Proof. For all $k \in \{1, \ldots, N\}$, we define $d_k := (c_1, \ldots, (N-k+1) \times c_k, \ldots, (N-k+1) \times c_k)$. Therefore, thanks to proposition 1 we have

$$\text{MOT}_{d_k}(\mu, \nu) = \sup_{\lambda \in \Delta_N^x} W_{c_k}(\mu, \nu)$$

$$= \sup_{(\lambda, \gamma) \in \Delta_N^x} \inf_{\gamma \in \Gamma_{\mu, \nu}} \int_{X \times Y} \min(\lambda_1 c_1, \ldots, \lambda_k c_k, \ldots, \gamma c_k) d\gamma$$

where $\Delta_N^x := \{ (\lambda, \gamma) \in \Delta_N^x \times \mathbb{R}_+: \gamma = (N-k+1) \times \min(\lambda_k, \ldots, \lambda_N) \}$. First remarks that

$$\gamma = 1 - \sum_{i=1}^{k-1} \lambda_i \iff (N-k+1) \times \min(\lambda_k, \ldots, \lambda_N) = \sum_{i=k}^N \lambda_i$$

$$\iff \lambda_k = \ldots = \lambda_N$$
But in that case \((\lambda_1, \ldots, \lambda_{k-1}, \gamma) \in \Delta_k\) and therefore we obtain that
\[
\text{MOT}_{d_k}(\mu, \nu) \geq \sup_{\lambda \in \Delta_k} \inf_{\gamma \in \Gamma_{\mu, \nu}} \int_{X \times Y} \min(\lambda_1 c_1, \ldots, \lambda_{k-1} c_{k-1}, \gamma c_k) \, d\gamma = \text{MOT}_{c_k}(\mu, \nu)
\]
Finally by definition we have \(\gamma \leq \sum_{i=1}^N \lambda_i = 1 - \sum_{i=1}^{k-1} \lambda_i\) and therefore
\[
\int_{X \times Y} \min(\lambda_1 c_1, \ldots, \lambda_{k-1} c_{k-1}, \gamma c_k) \, d\gamma \leq \int_{X \times Y} \min(\lambda_1 c_1, \ldots, \lambda_{k-1} c_{k-1}, \left(1 - \sum_{i=1}^{k-1} \lambda_i\right) c_k) \, d\gamma
\]
Then we obtain that
\[
\text{MOT}_{d_k}(\mu, \nu) \leq \text{MOT}_{c_k}(\mu, \nu)
\]
and the result follows.

**Proposition 9.** Let \(X\) and \(Y\) be Polish spaces and \(c := (c_i)_{1 \leq i \leq N}\) a family of nonnegative lower semi-continuous costs on \(X \times Y\). We suppose that, for all \(i\), \(c_i = N \times c_1\). Then for any \((\mu, \nu) \in \mathcal{M}_+^1(X) \times \mathcal{M}_+^1(Y)\)
\[
\text{MOT}_c(\mu, \nu) = \text{MOT}_{c_1}(\mu, \nu) = W_{c_1}(\mu, \nu).
\]
**Proof.** Let \(c := (c_i)_{1 \leq i \leq N}\) such that for all \(i\), \(c_i = c_1\). For all \((x, y) \in X \times Y\) and \(\lambda \in \Delta_N^\lambda\), we have:
\[
c_\lambda(x, y) := \min_i (\lambda_i c_i(x, y)) = \min_i (\lambda_i c_1(x, y))
\]
Therefore we obtain from Lemma 7 that
\[
\text{MOT}_c(\mu, \nu) = \sup_{\lambda \in \Delta_N^\lambda} W_{c_\lambda}(\mu, \nu)
\]
But we also have that:
\[
W_{\min(a, 1-a)c}(\mu, \nu) \leq \inf_{\gamma \in \Gamma_{\mu, \nu}} \int_{X \times Y} \min(\lambda_i, c_i(x, y)) \, d\gamma(x, y) = \min(\lambda_i) \inf_{\gamma \in \Gamma_{\mu, \nu}} \int_{X \times Y} c_i(x, y) \, d\gamma(x, y) = \min(\lambda_i) W_{c_1}(\mu, \nu)
\]
Finally by taking the supremum over \(\lambda \in \Delta_N\) we conclude the proof.

**C.3 An equivalent problem**

**Proposition 10.** Let \(X\) and \(Y\) be Polish spaces and \(c := (c_i)_{1 \leq i \leq N}\) a family of nonnegative lower semi-continuous costs on \(X \times Y\). We suppose that, for all \(i\), \(c_i = N \times c_1\). Then for any \((\mu, \nu) \in \mathcal{M}_+^1(X) \times \mathcal{M}_+^1(Y)\)
\[
\text{MOT}_c(\mu, \nu) = \inf_{\gamma \in \Gamma_{\mu, \nu}} \max_i \int c_i d\gamma_i
\]
and the infimum is attained
**Proof.** First, it is clear that \(\text{MOT}_c(\mu, \nu) \geq \inf_{\gamma \in \Gamma_{\mu, \nu}} \max_i \int c_i d\gamma_i\). Let now show that in fact it is an equality. The infimum is attained for \(\inf_{\gamma \in \Gamma_{\mu, \nu}} \max_i \int c_i d\gamma_i\) by applying Weierstrass theorem. Indeed recall that \(\Gamma_{\mu, \nu}\) is compact and that the objective is l.s.c. Let \(\gamma^*\) be such a minimizer. Assume that there exists \(i, j\) such that, \(\int c_i d\gamma_i^* > \int c_j d\gamma_j^*\). By non negativity of the costs, there exists \((x_0, y_0)\) such that \(c_i(x_0, y_0) > 0\). Let \(\tilde{\gamma}\) defined as for all \(k \neq i, j\), \(\tilde{\gamma}_k = \gamma_k^*\) and \(\tilde{\gamma}_i = \gamma_i^* - \frac{\epsilon}{c_i(x_0, y_0)} \delta_{(x_0, y_0)}\) and \(\tilde{\gamma}_j = \gamma_j^* + \frac{\epsilon}{c_j(x_0, y_0)} \delta_{(x_0, y_0)}\) for \(\epsilon\) sufficiently small so that \(\tilde{\gamma} \in \Gamma_{\mu, \nu}\). Now, \(\max_k \int c_k d\gamma_k^* > \max_k \int c_k d\tilde{\gamma}_k\), which contradicts that \(\gamma^*\) is a minimizer. Then for \(i, j\), \(\int c_i d\gamma_i^* = \int c_j d\gamma_j^*\). And then: \(\text{MOT}_c(\mu, \nu) = \inf_{\gamma \in \Gamma_{\mu, \nu}} \max_i \int c_i d\gamma_i\).
Proposition 11. Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces and $c := (c_i)_{1 \leq i \leq N}$ a family of nonnegative lower semi-continuous costs on $\mathcal{X} \times \mathcal{Y}$. We suppose that, for all $i$, $c_i = N \times c_1$. Then for any $(\mu, \nu) \in \mathcal{M}_1^N(\mathcal{X}) \times \mathcal{M}_1^N(\mathcal{Y})$

$$\text{MOT}_c(\mu, \nu) = \sup_{\lambda \in \Delta_\mathcal{X}} \sup_{(f, g) \in F_c} \int fd\mu + \int g d\nu$$  

(17)

To prove this theorem, one need to prove the three following technical lemmas. The first one shows the weak compactness of $\Gamma_{\mu, \nu}^N$.

Lemma 2. Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces, and $\mu$ and $\nu$ two probability measures respectively on $\mathcal{X}$ and $\mathcal{Y}$. Then $\Gamma_{\mu, \nu}^N$ is sequentially compact for the weak topology induced by $\|\gamma\|_{TV} = \max_{i=1, \ldots, N} \|\gamma_i\|_{TV}$.

Proof. Let $(\gamma^n)_{n \geq 0}$ a sequence in $\Gamma_{\mu, \nu}^N$, and let us denote for all $n \geq 0$, $\gamma^n = (\gamma^n_i)_{i=1}^N$. We first remarks that for all $i \in \{1, \ldots, N\}$ and $n \geq 0$, $\|\gamma^n_i\|_{TV} \leq 1$ therefore for all $i \in \{1, \ldots, N\}$, $(\gamma^n_i)_{n \geq 0}$ is uniformly bounded. Moreover as $\{\mu\}$ and $\{\nu\}$ are tight, for any $\delta > 0$, there exists $K \subset \mathcal{X}$ and $L \subset \mathcal{Y}$ compact such that

$$\mu(K^c) \leq \frac{\delta}{2} \quad \text{and} \quad \nu(L^c) \leq \frac{\delta}{2}.$$  

Therefore, we obtain that for any $i \in \{1, \ldots, N\}$,

$$\gamma^n_i(K^c \times L^c) \leq \sum_{k=1}^N \gamma^n_k(K^c \times L^c)$$  

(19)

$$\leq \sum_{k=1}^N \gamma^n_k(K^c \times \mathcal{Y}) + \gamma^n_i(K \times L^c)$$  

(20)

$$\leq \mu(K^c) + \nu(L^c) = \delta.$$  

(21)

Therefore, for all $i \in \{1, \ldots, N\}$, $(\gamma^n_i)_{n \geq 0}$ is tight and uniformly bounded and Prokhorov’s theorem [14, Theorem A.3.15] guarantees for all $i \in \{1, \ldots, N\}$, $(\gamma^n_i)_{n \geq 0}$ admits a weakly convergent subsequence. By extracting a common convergent subsequence, we obtain that $(\gamma^n_i)_{n \geq 0}$ admits a weakly convergent subsequence. By continuity of the projection, the limit also lives in $\Gamma_{\mu, \nu}^N$ and the result follows.

Next lemma generalizes Rockafellar-Fenchel duality to our case.

Lemma 3. Let $V$ be a normed vector space and $V^*$ its topological dual. Let $V_1, \ldots, V_N$ be convex functions and lower semi-continuous on $V$ and $E$ a convex function on $V$. Let $V_1^*, \ldots, V_N^*, E^*$ be the Fenchel-Legendre transforms of $V_1, \ldots, V_N, E$. Assume there exists $z_0 \in V$ such that for all $i$, $V_i(z_0) < \infty$, $E(z_0) < \infty$, and for all $i$, $V_i$ is continuous at $z_0$. Then:

$$\inf_{u \in V} \sum_i V_i(u) + E(u) = \sup_{\gamma_1, \ldots, \gamma_N \in V^*} \sum_i V_i^*(-\gamma_i) - E^*(\gamma)$$

Proof. This Lemma is an immediate application of Rockafellar-Fenchel duality theorem [8, Theorem 1.12] and of Fenchel-Moreau theorem [8, Theorem 1.11]. Indeed, $V = \sum_{i=1}^N V_i(u)$ is a convex function, lower semi-continuous and its Legendre-Fenchel transform is given by

$$V^*(\gamma^*) = \inf_{\sum_{i=1}^N \gamma_i^* = \gamma^*} \sum_{i=1}^N V_i^*(\gamma_i^*).$$  

(22)

Last lemma is an application of Sion’s Theorem to this problem.
Lemma 4. Let \( \mathcal{X} \) and \( \mathcal{Y} \) be Polish spaces. Let \( c = (c_i)_{1 \leq i \leq N} \) be a family of nonnegative lower semi-continuous costs on \( \mathcal{X} \times \mathcal{Y} \), then for \( \mu \in \mathcal{M}^+_+ (\mathcal{X}) \) and \( \nu \in \mathcal{M}^+_+ (\mathcal{Y}) \), we have

\[
\text{MOT}_c (\mu, \nu) = \sup_{\lambda \in \Delta^+_N} \inf_{\gamma \in \Gamma^+_N} \sum_{i=1}^N \lambda_i \int_{\mathcal{X} \times \mathcal{Y}} c_i (x, y) d\gamma_i (x, y)
\]

and the infimum is attained.

Proof. Taking for granted that a minmax principle can be invoked, we have

\[
\sup_{\lambda \in \Delta^+} \inf_{\gamma \in \Gamma^+_N} \sum_{i=1}^N \lambda_i \int_{\mathcal{X} \times \mathcal{Y}} c_i (x, y) d\gamma_i (x, y) = \inf_{\gamma \in \Gamma^+_N} \sup_{\lambda \in \Delta^+_N} \sum_{i=1}^N \lambda_i \int_{\mathcal{X} \times \mathcal{Y}} c_i (x, y) d\gamma_i (x, y)
\]

But thanks to Lemma 4, we have that \( \Gamma^+_N \) is compact for the weak topology. And \( \Delta^+_N \) is convex. Moreover the objective function \( f : (\lambda, \gamma) \in \Delta^+_N \times \Gamma^+_N \rightarrow \sum_{i=1}^N \lambda_i \int_{\mathcal{X} \times \mathcal{Y}} c_i^0 d\gamma_i \) is bilinear, hence convex and concave in its variables, and continuous with respect to \( \lambda \). Moreover, let \( (c^0_i)_n \) be non-decreasing sequences of non-negative bounded cost functions such that \( c_i = \sup_n c^0_i \). By monotone convergence, we get \( f(\lambda, \gamma) = \sup_n \sum_i \lambda_i \int_{\mathcal{X} \times \mathcal{Y}} c^0_i d\gamma_i \). So \( f \) the supremum of continuous functions, then \( f \) is lower semi-continuous with respect to \( \gamma \), therefore Sion’s minmax theorem [32] holds.

We are now able to prove Proposition 11.

Proof. Let \( \mathcal{X} \) and \( \mathcal{Y} \) be two Polish spaces. For all \( i \in \{1, \ldots, N\} \), we define \( c_i : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^+ \) a cost function. The proof follows the exact same steps as those in the proof of [37, Theorem 1.3]. First we suppose that \( \mathcal{X} \) and \( \mathcal{Y} \) are compact and that for all \( i \), \( c_i \) is continuous, then we show that it can be extended to \( \mathcal{X} \) and \( \mathcal{Y} \) non compact and finally to \( c_i \) only lower semi continuous.

First, let assume \( \mathcal{X} \) and \( \mathcal{Y} \) are compact and that for all \( i \), \( c_i \) is continuous. Let fix \( \lambda \in \Delta^+_N \). We recall the topological dual of the space of bounded continuous functions \( \mathcal{C}^b (\mathcal{X} \times \mathcal{Y}) \) endowed with \( \| \cdot \|_{\infty} \) norm, is the space of Radon measures \( \mathcal{M} (\mathcal{X} \times \mathcal{Y}) \) endowed with total variation norm. We define, for \( u \in \mathcal{C}^b (\mathcal{X} \times \mathcal{Y}) \):

\[
V^\lambda (u) = \left\{ \begin{array}{ll} 0 & \text{if } u \geq -\lambda_i c_i \\ +\infty & \text{else} \end{array} \right.
\]

and:

\[
E(u) = \left\{ \begin{array}{ll} \int f d\mu + \int g d\nu & \text{if } \exists (f, g) \in \mathcal{C}^b (\mathcal{X}) \times \mathcal{C}^b (\mathcal{Y}), \ u = f + g \\ +\infty & \text{else} \end{array} \right.
\]

One can show that for all \( i \), \( V^\lambda_i \) is convex and lower semi-continuous (as the sublevel sets are closed) and \( E^\lambda \) is convex. Moreover for all \( i \), these functions continuous in \( u_0 \equiv 1 \) the hypothesis of Lemma 4 are satisfied.

Let now compute the Fenchel-Legendre transform of these function. Let \( \gamma \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}) \):

\[
V^\lambda \ast (-\gamma) = \sup_{u \in \mathcal{C}^b (\mathcal{X} \times \mathcal{Y})} \left\{ -\int u d\gamma; \ u \geq -\lambda_i c_i \right\}
\]

\[
= \left\{ \begin{array}{ll} \int \lambda_i c_i d\gamma & \text{if } \gamma \in \mathcal{M}_+ (\mathcal{X} \times \mathcal{Y}) \\ +\infty & \text{otherwise} \end{array} \right.
\]

On the other hand:

\[
E^\lambda \ast (\gamma) = \left\{ \begin{array}{ll} 0 & \forall (f, g) \in \mathcal{C}^b (\mathcal{X}) \times \mathcal{C}^b (\mathcal{Y}), \ \int f d\mu + \int g d\nu = \int (f + g) d\gamma \\ +\infty & \text{else} \end{array} \right.
\]
This dual function is finite and equals 0 if and only if that the marginals of the dual variable $\gamma$ are $\mu$ and $\nu$. Applying Lemma 3, we get:

$$\inf_{u \in C^b(X \times Y)} \sum_i V_i^\lambda(u) + E(u) = \sup_{\sum \gamma_i = \gamma} \sum_i -V_i^{\lambda^*}(\gamma_i) - E^\lambda(-\gamma)$$

Hence, we have shown that, when $X$ and $Y$ are compact sets, and the costs $(c_i)_i$ are continuous:

$$\sup_{(f,g) \in F^\lambda_0} \int f \, d\mu + \int g \, d\nu = \inf_{\gamma \in \Gamma^\lambda_{N}} \sum_i \lambda_i \int c_i d\gamma_i$$

Let now prove the result holds when the spaces $X$ and $Y$ are not compact. We still suppose that for all $i$, $c_i$ is uniformly continuous and bounded. We denote $\|c\|_{\infty} := \sup_{s \in X \times Y} c(s,x,y)$. Let define $I^\lambda(\gamma) := \sum_i \lambda_i \int_{X \times Y} c_i d\gamma_i$

Let $\gamma^* \in \Gamma^N_{\mu,\nu}$ such that $I^\lambda(\gamma^*) = \min_{\gamma \in \Gamma^N_{\mu,\nu}} I^\lambda(\gamma)$. The existence of the minimum comes from the lower-semi continuity of $I^\lambda$ and the compacity of $\Gamma^N_{\mu,\nu}$ for weak topology.

Let fix $\delta \in (0,1)$. $X$ and $Y$ are Polish spaces then $\exists X_0 \subset X, Y_0 \subset Y$ compact such that $\mu(X_0^c) \leq \delta$ and $\mu(Y_0^c) \leq \delta$. It follows that for all $i$, $\gamma_i^0((X_0 \times Y_0)^c) \leq 2\delta$. Let define $\gamma^*_i$ such that for all $i$, $\gamma_i^0 = \frac{1}{\sum_i \gamma_i((X_0 \times Y_0)^c)} \gamma_i^*$. We define $\mu_0 = \Pi_{11} \sum \gamma_i^0$ and $\nu_0 = \Pi_{22} \sum \gamma_i^0$. We then naturally define $\Gamma^N_{\mu_0,\nu_0} := \{ (\gamma_i)_{1 \leq i \leq N} \in \mathcal{M}_+(X_0 \times Y_0)^N s.t. \Pi_{11} \sum \gamma_i = \mu_0 \text{ and } \Pi_{22} \sum \gamma_i = \nu_0 \}$ and $I^\lambda_0(\gamma_0) := \sum_i \lambda_i \int_{X_0 \times Y_0} c_i d\gamma_0, i$ for $\gamma_0 \in \Gamma^N_{\mu_0,\nu_0}$.

Let $\tilde{\gamma}_0$ verifying $I_0^\lambda(\tilde{\gamma}_0) = \min_{\gamma \in \Gamma^N_{\mu_0,\nu_0}} I_0^\lambda(\gamma_0)$. Let $\tilde{\gamma} = (\sum_i \gamma_i^0((X_0 \times Y_0)) \tilde{\gamma}_0 + 1_{(X_0 \times Y_0)^c})^* \gamma^* \in \Gamma^N_{\mu,\nu}$. Then we get

$$I^\lambda(\tilde{\gamma}) \leq \min_{\gamma \in \Gamma^N_{\mu_0,\nu_0}} I_0^\lambda(\gamma_0) + 2 \sum_i |\lambda_i||c|_{\infty}\delta$$

We have already proved that:

$$\sup_{(f,g) \in F^\lambda_{0,c}} J^\lambda_0(f,g) = \inf_{\gamma \in \Gamma^N_{\mu_0,\nu_0}} I_0^\lambda(\gamma_0)$$

with $J^\lambda_0(f,g) = \int f \, d\mu_0 + \int g \, d\nu_0$ and $F^\lambda_0, c$ is the set of $(f,g) \in C^b(X_0) \times C^b(Y_0)$ satisfying, for every $i$, $f \oplus g \leq \min, \lambda_i c_i$. Let $(\tilde{f}_0, \tilde{g}_0) \in F^\lambda_{0,c}$ such that:

$$J^\lambda_0(\tilde{f}_0, \tilde{g}_0) = \sup_{(f,g) \in F^\lambda_{0,c}} J^\lambda_0(f,g) - \delta$$

Since $J^\lambda_0(0,0) = 0$, we get $\sup J^\lambda_0 \geq 0$ and then, $J^\lambda_0(\tilde{f}_0, \tilde{g}_0) \geq \delta \geq -1$. For every $\gamma_0 \in \Gamma^N_{0,0}$:

$$J^\lambda_0(\tilde{f}_0, \tilde{g}_0) = \int (\tilde{f}_0(x) + \tilde{g}_0(y)) d\gamma_0(x,y)$$

then we have the existence of $(x_0, y_0) \in X_0 \times Y_0$ such that : $\tilde{f}_0(x_0) + \tilde{g}_0(y_0) \geq -1$. If we replace $(\tilde{f}_0, \tilde{g}_0)$ by $(\tilde{f}_0 - s, \tilde{g}_0 + s)$ for an accurate $s$, we get that: $\tilde{f}_0(x_0) \geq \frac{1}{2}$ and $\tilde{g}_0(y_0) \geq \frac{1}{2}$, and then $\forall(x,y) \in X_0 \times Y_0$:

$$\tilde{f}_0(x) \leq c^*(x,y) - \tilde{g}_0(y_0) \leq c^*(x,y) + \frac{1}{2}$$

$$\tilde{g}_0(y) \leq c^*(x_0,y) - \tilde{f}_0(x_0) \leq c^*(x_0,y) + \frac{1}{2}$$

23
where \( c' := \min_i \lambda_i c_i \). Let define \( \tilde{f}_0(x) = \inf_{y \in \mathcal{Y}_0} c'(x, y) - \tilde{g}_0(y) \) for \( x \in \mathcal{X} \). Then \( \tilde{f}_0 \leq \bar{f}_0 \) on \( \mathcal{X}_0 \). We then get \( J^0_\lambda(\tilde{f}_0, \tilde{g}_0) \geq J^0_\lambda(\bar{f}_0, \bar{g}_0) \) and \( \tilde{f}_0 \leq c'(\cdot, y_0) + \frac{1}{2} \) on \( \mathcal{X} \). Let define \( \bar{g}_0(y) = \inf_{x \in \mathcal{X}} c'(x, y) - \bar{f}_0(y) \). By construction \( (\tilde{f}_0, \tilde{g}_0) \in \mathcal{F}_c^\lambda \) since the costs are uniformly continuous and bounded and \( J^0_\lambda(\tilde{f}_0, \tilde{g}_0) \geq J^0_\lambda(\bar{f}_0, \bar{g}_0) \). We also have \( \bar{g}_0 \geq c'(x_0, \cdot) + \frac{1}{2} \) on \( \mathcal{Y} \). Then we have in particular: \( \bar{g}_0 \geq -\|c\|_\infty - \frac{1}{2} \) on \( \mathcal{X} \) and \( \tilde{f}_0 \geq -\|c\|_\infty - \frac{1}{2} \) on \( \mathcal{Y} \). Finally:

\[
J^\lambda(\tilde{f}_0, \tilde{g}_0) := \int_{\mathcal{X}_0} \tilde{f}_0 d\mu_0 + \int_{\mathcal{X}_0} \tilde{g}_0 d\nu_0 \\
= \sum_i \gamma_i^*(\mathcal{X}_0 \times \mathcal{Y}_0) \int_{\mathcal{X}_0 \times \mathcal{Y}_0} (\tilde{f}_0(x) + \tilde{g}_0(y)) d\left( \sum_i \gamma_i^0(x, y) \right) \\
+ \int_{(\mathcal{X}_0 \times \mathcal{Y}_0)^c} \tilde{f}_0(x) + \tilde{g}_0(y) d\left( \sum_i \gamma_i^*(x, y) \right) \\
\geq (1 - 2\delta) \left( \int_{\mathcal{X}_0} \tilde{f}_0 d\mu_0 + \int_{\mathcal{Y}_0} \tilde{g}_0 d\nu_0 \right) - (2\|c\|_\infty + 1) \sum_i \gamma^*((\mathcal{X}_0 \times \mathcal{Y}_0)^c) \\
\geq (1 - 2\delta) J^\lambda_0(\tilde{f}_0, \tilde{g}_0) - 2 \sum_i |\lambda_i| (2\|c\|_\infty + 1)\delta \\
\geq (1 - 2\delta) J^\lambda_0(\tilde{f}_0, \tilde{g}_0) - 2 \sum_i |\lambda_i| (2\|c\|_\infty + 1)\delta \\
\geq (1 - 2\delta) (\inf I^\lambda - 2 \sum_i |\lambda_i| (2\|c\|_\infty + 1)\delta ) - 2 \sum_i |\lambda_i| (2\|c\|_\infty + 1)\delta \\
\geq (1 - 2\delta) (\inf I^\lambda - (2\sum_i |\lambda_i| (2\|c\|_\infty + 1)\delta ) - 2 \sum_i |\lambda_i| (2\|c\|_\infty + 1)\delta )
\]

This being true for arbitrary small \( \delta \), we get \( \sup J^\lambda \geq \inf I^\lambda \). The other sense is always true then:

\[
\sup_{(f,g) \in \mathcal{F}^\lambda_c} \int f d\mu + \int g d\nu = \inf_{\gamma \in \Gamma^N_{\mu,\nu}} \sum_i \lambda_i \int c_i d\gamma_i
\]

for \( c_i \) uniformly continuous and \( \mathcal{X} \) and \( \mathcal{Y} \) non necessarily compact.

Let now prove that the result holds for lower semi-continuous costs. Let \( c := (c_i)_i \) be a collection of lower semi-continuous costs. Let \( (c^n_i)_n \) be non-decreasing sequences of non-negative bounded cost functions such that \( c_i = \sup_n c^n_i \). Let fix \( \lambda \in \Delta_+^N \). From last step, we have shown that for all \( n \):

\[
\inf_{\gamma \in \Gamma^N_{\mu,\nu}} I^\lambda_n(\gamma) = \sup_{(f,g) \in \mathcal{F}^\lambda_n} \int f d\mu + \int g d\nu
\]

where \( I^\lambda_n(\gamma) = \sum_i \lambda_i \int c^n_i d\gamma_i \). First it is clear that:

\[
\sup_{(f,g) \in \mathcal{F}^\lambda_n} \int f d\mu + \int g d\nu \leq \sup_{(f,g) \in \mathcal{F}^\lambda_n} \int f d\mu + \int g d\nu
\]

Let show that:

\[
\inf_{\gamma \in \Gamma^N_{\mu,\nu}} I^\lambda(\gamma) = \sup_n \inf_{\gamma \in \Gamma^N_{\mu,\nu}} I^\lambda_n(\gamma) = \lim_n \inf_{\gamma \in \Gamma^N_{\mu,\nu}} I^\lambda_n(\gamma)
\]

where \( I^\lambda(\gamma) = \sum_i \lambda_i \int c_i d\gamma_i \).

Let \( (\gamma^{n,k})_k \) a minimizing sequence of \( \Gamma^N_{\mu,\nu} \) for the problem \( \inf_{\gamma \in \Gamma^N_{\mu,\nu}} \sum_i \lambda_i \int c_i d\gamma_i \). By Lemma 2, up to an extraction, there exists \( \gamma^n \in \Gamma^N_{\mu,\nu} \) such that \( (\gamma^{n,k})_k \) converges weakly to \( \gamma^n \). Then:

\[
\inf_{\gamma \in \Gamma^N_{\mu,\nu}} I^\lambda(\gamma) = I^\lambda_n(\gamma_n)
\]

24
Up to an extraction, there also exists $\gamma^* \in \Gamma^N_{\mu,\nu}$ such that $\gamma^n$ converges weakly to $\gamma^*$. For $n \geq m$, $I^\lambda_m(\gamma_n) \geq I^\lambda_m(\gamma_m)$, so by continuity of $I^\lambda_m$:

$$\lim_{n} I^\lambda_n(\gamma_n) \geq \limsup_{n} I^\lambda_m(\gamma^n) \geq I^\lambda_m(\gamma^*)$$

By monotone convergence, $I^\lambda_m(\gamma^*) \to I^\lambda(\gamma^*)$ and $\lim_{n} I^\lambda_m(\gamma_n) \geq I^\lambda(\gamma^*) \geq \inf_{\gamma \in \Gamma^N_{\mu,\nu}} I^\lambda(\gamma)$.

Along with Eqs. [24] and [25] we get that:

$$\inf_{\gamma \in \Gamma^N_{\mu,\nu}} I^\lambda(\gamma) \leq \sup_{(f,g) \in F^2_N} \int f d\mu + \int g d\nu$$

The other sense being always true, we have then shown that, in the general case we still have:

$$\inf_{\gamma \in \Gamma^N_{\mu,\nu}} I^\lambda(\gamma) = \sup_{(f,g) \in F^2_N} \int f d\mu + \int g d\nu$$

To conclude, we apply Lemma 4 and we get:

$$\sup_{\lambda \in \Delta^+_N} \sup_{(f,g) \in F^2_N} \int f d\mu + \int g d\nu = \sup_{\lambda \in \Delta^+_N} \inf_{\gamma \in \Gamma^N_{\mu,\nu}} I^\lambda(\gamma) = \text{MOT}_c(\mu, \nu)$$

**Proposition 12.** Let $X$ and $Y$ be Polish spaces and $c := (c_i)_{1 \leq i \leq N}$ a family of nonnegative lower semi-continuous costs on $X \times Y$. Then for any $\lambda, \nu \in M^1_+(X) \times M^1_+(Y)$:

$$\text{MOT}_c(\mu, \nu) = \sup_{(f,g) \in G^N_c} \inf_{\lambda \in \Gamma^N_{\mu,\nu}} \max_k \left[ \int_{x \in X} f_k(x) d\mu_k(x) + \int_{y \in Y} g_k(y) d\nu_k(y) \right]$$

*Proof.* Let $\lambda \in \Delta^+_N$, $\gamma \in \Gamma^N_{\mu,\nu}$, and $(f_k, g_k)_{k=1}^N \in G^N_c$, then we have:

$$\sum_{k=1}^N \lambda_k \left[ \int_{x \in X} f_k(x) d\mu_k(x) + \int_{y \in Y} g_k(y) d\nu_k(y) \right] \leq \sum_{k=1}^N \lambda_k \int_{x \in X} c_k(x, y) d\gamma_k(x, y) \leq \max_k \left( \int_{x \in X} c_k(x, y) d\gamma_k(x, y) \right)$$

where $\mu_k$ and $\nu_k$ are defined as $\mu_k = \Pi_{1\gamma_k}$ and $\nu_k = \Pi_{2\gamma_k}$. Note that it holds

$$\sum_{k=1}^N \lambda_k \left[ \int_{x \in X} f_k(x) d\mu_k(x) + \int_{y \in Y} g_k(y) d\nu_k(y) \right] \leq \max_k \left( \int_{x \in X} c_k(x, y) d\gamma_k(x, y) \right)$$

even if $\lambda \in \Delta^+_N$ thanks to the positivity of the cost functions. Then $(\mu_k, \nu_k)_{k=1}^N \in \Upsilon^N_{\mu,\nu}$ and we obtain that

$$\text{MOT}_c(\mu, \nu) \geq \sup_{\lambda \in \Delta^+_N} \sup_{(f,g) \in G^N_c} \inf_{\lambda \in \Gamma^N_{\mu,\nu}} \sum_{k=1}^N \lambda_k \left[ \int_{x \in X} f_k(x) d\mu_k(x) + \int_{y \in Y} g_k(y) d\nu_k(y) \right]$$

Let us now consider $(f, g) \in F^2_N$ and $\lambda \in \Delta^+_N$. Moreover assume that for each $i \in \{1, \ldots, N\}$, $\lambda_i > 0$. We can now define for all $i \in \{1, \ldots, N\}$, $f_i := f/\lambda_i$ and $g_i := g/\lambda_i$. In the case where $\lambda_i = 0$, $f_i := 0$ and $g_i := 0$. Then $(f_i, g_i) \in G^N_c$. We remark that for any $(\mu_k, \nu_k)_{k=1}^N \in \Upsilon^N_{\mu,\nu}$ we have:

$$\int_{x \in X} f(x) d\mu(x) + \int_{y \in Y} g(y) d\nu(y) = \sum_{k=1}^N \lambda_k \left[ \int_{x \in X} f_k(x) d\mu_k(x) + \int_{y \in Y} g_k(y) d\nu_k(y) \right]$$

25
Therefore we obtain that:

\[ \int_{x \in X} f(x) d\mu(x) + \int_{y \in Y} g(y) d\nu(y) = \inf_{(\mu_k, \nu_k)} \sum_{k=1}^{N} \lambda_k \left[ \int_{x \in X} f_k(x) d\mu_k(x) + \int_{y \in Y} g_k(y) d\nu_k(y) \right] \]

Finally we obtain that

\[ \sup_{(f,g) \in F^+} \int_{x \in X} f(x) d\mu(x) + \int_{y \in Y} g(y) d\nu(y) \]

\[ \leq \sup_{(f_k,g_k) \in \mathcal{G}^N} \inf_{(\mu_k,\nu_k)} \sum_{k=1}^{N} \lambda_k \left[ \int_{x \in X} f_k(x) d\mu_k(x) + \int_{y \in Y} g_k(y) d\nu_k(y) \right] \]

and the result follows by strong duality from Theorem 1. Then we deduce:

\[ \text{MOT}_{c}(\mu,\nu) = \sup_{\lambda \in \Delta_{X}^H} \sup_{(f,g) \in \mathcal{G}_{\lambda}^N} \inf_{(\mu_k,\nu_k)} \max_{k} \left[ \int_{x \in X} f_k(x) d\mu_k(x) + \int_{y \in Y} g_k(y) d\nu_k(y) \right] \]

Applying Sion’s theorem [32]:

\[ \text{MOT}_{c}(\mu,\nu) = \sup_{(f,g) \in \mathcal{G}_{\lambda}^N} \inf_{(\mu_k,\nu_k)} \max_{k} \left[ \int_{x \in X} f_k(x) d\mu_k(x) + \int_{y \in Y} g_k(y) d\nu_k(y) \right] \]

Proposition 13. Let \( X \) and \( Y \) be two compact Polish spaces, \( c := (c_i)_{1 \leq i \leq N} \) a family of nonnegative lower semi-continuous costs on \( X \times Y \) and \( \varepsilon := (\varepsilon_i)_{1 \leq i \leq N} \) be non negative real numbers. For \( (\mu,\nu) \in \mathcal{M}^1(X) \times \mathcal{M}^1(Y) \), strong duality holds:

\[ \text{MOT}_{c}^\varepsilon(\mu,\nu) = \sup_{\lambda \in \Delta_{X}^H} \sup_{(f,g) \in \mathcal{C}_{\lambda}^N \times \mathcal{C}_{\lambda}^N} \int_{X} f(x) d\mu(x) + \int_{Y} g(y) d\nu(y) \]

\[ - \sum_{i=1}^{N} \varepsilon_i \left( \int_{X \times Y} e^{\frac{f(x) + g(y) - \lambda_i c_i(x,y)}{\varepsilon_i}} d\mu(x)d\nu(y) - 1 \right) \]

and the infimum of the primal problem is attained.

Proof. To show the strong duality of the regularized problem, we use the same sketch of proof as for the strong duality of the original problem. Let first assume that, for all \( i \), \( c_i \) is continuous on the compact set \( X \times Y \). Let fix \( \lambda \in \Delta_{X}^H \). We define, for all \( u \in \mathcal{C}^0(X \times Y) \):

\[ V^\lambda_i(u) = \varepsilon_i \left( \int_{(x,y) \in X \times Y} \exp \frac{-u(x,y) - \lambda_i c_i(x,y)}{\varepsilon_i} d\mu(x)d\nu(y) - 1 \right) \]

and:

\[ E(u) = \begin{cases} \int f d\mu + \int g d\nu & \text{if } \exists (f,g) \in \mathcal{C}^0(X) \times \mathcal{C}^0(Y), \ u = f + g \\ +\infty & \text{else} \end{cases} \]

Let compute the Fenchel-Legendre transform of these functions. Let \( \gamma \in \mathcal{M}(X \times Y) \):

\[ V^\lambda_i(\gamma) = \sup_{u \in \mathcal{C}^0(X \times Y)} - \int u d\gamma - \varepsilon_i \left( \int_{(x,y) \in X \times Y} \exp \frac{-u(x,y) - \lambda_i c_i(x,y)}{\varepsilon_i} d\mu(x)d\nu(y) - 1 \right) \]
However, by density of $C_b(X \times Y)$ in $L^1_{d\mu \otimes \nu}(X \times Y)$ we deduce that

$$V_i^{\lambda^*}(-\gamma) = \sup_{u \in L^1_{d\mu \otimes \nu}(X \times Y)} - \int u d\gamma - \frac{1}{\varepsilon_i} \left( \int_{(x,y) \in X \times Y} \exp \frac{-u(x,y) - \lambda_i c_i(x,y)}{\varepsilon_i} d\mu(x) d\nu(y) - 1 \right)$$

This supremum equals $+\infty$ if $\gamma$ is not positive and not absolutely continuous with regard to $\mu \otimes \nu$. Let us now denote $F_{\gamma, \lambda^*, \lambda}(u) := - \int u d\gamma - \frac{1}{\varepsilon_i} \left( \int_{(x,y) \in X \times Y} \exp \frac{-u(x,y) - \lambda_i c_i(x,y)}{\varepsilon_i} d\mu(x) d\nu(y) - 1 \right)$. $F_{\gamma, \lambda^*, \lambda}$ is Fréchet differentiable and its maximum is attained for $u^* = \varepsilon_i \log \left( \frac{d\gamma}{d\mu \otimes \nu} \right) + \lambda_i c_i$. Therefore we obtain that

$$V_i^{\lambda^*}(-\gamma) = \varepsilon_i \left( \int \log \left( \frac{d\gamma}{d\mu \otimes \nu} \right) d\gamma + 1 - \gamma(X \times Y) \right) + \lambda_i \int c_i d\gamma$$

Thanks to the compactness of $X \times Y$, all the $V_i^\lambda$ for $i \in \{1, \ldots, N\}$ are continuous on $C_b(X \times Y)$. Therefore by applying Lemma 3 we obtain that:

$$\inf_{u \in C_b(X \times Y)} \sum_{i=1}^N V_i^\lambda(u) + E(u) = \sup_{\gamma_1 \ldots \gamma_N \in \Lambda(\gamma \in M(X \times Y))} - \sum_{i=1}^N V_i^{\lambda^*}(\gamma_i) - E^*(-\gamma)$$

$$\sup_{f \in C_b(X), g \in C_b(Y)} \int f d\mu + \int g d\nu$$

$$- \sum_{i=1}^N \varepsilon_i \left( \int_{(x,y) \in X \times Y} \exp \frac{f(x) + g(y) - \lambda_i c_i(x,y)}{\varepsilon_i} d\mu(x) d\nu(y) - 1 \right)$$

$$= \inf_{\gamma \in \Gamma \lambda^*, \varepsilon_i \lambda_i \int c_i d\gamma_i + \varepsilon_i \text{KL}(\gamma_i \| \mu \otimes \nu)$$

Therefore by considering the supremum over the $\lambda \in \bigtriangleup_N^+$, we obtain that

$$\text{MOT}_c = \sup_{\lambda \in \bigtriangleup_N^+} \inf_{\gamma \in \Gamma \lambda^*, \varepsilon_i \lambda_i \int c_i d\gamma_i + \varepsilon_i \text{KL}(\gamma_i \| \mu \otimes \nu)$$

Let $f : (\lambda, \gamma) \in \bigtriangleup_N^+ \times \Gamma \lambda^*, \nu \rightarrow \sum_{i=1}^N \lambda_i \int c_i d\gamma_i + \varepsilon_i \text{KL}(\gamma_i \| \mu \otimes \nu)$. $f$ is clearly concave and continuous in $\lambda$. Moreover $\gamma \mapsto \text{KL}(\gamma_i \| \mu \otimes \nu)$ is convex and lower semi-continuous for weak topology [3]. Hence $f$ is convex and lower-semi continuous in $\gamma$. $\bigtriangleup_N^+$ is convex, and $\Gamma \lambda^*, \mu, \nu$ is compact for weak topology (see Lemma 3). By Sion’s theorem, we get the expected result:

$$\text{MOT}_c(\mu, \nu) = \min_{\gamma \in \Gamma \lambda^*, \varepsilon_i \lambda_i \int c_i d\gamma_i + \varepsilon_i \text{KL}(\gamma_i \| \mu \otimes \nu)$$

$$= \sup_{\lambda \in \bigtriangleup_N^+} \sup_{(f,g) \in C_b(X) \times C_b(Y)} \int_X f(x) d\mu(x) + \int_Y g(y) d\nu(y)$$

$$- \sum_{i=1}^N \varepsilon_i \left( \int_{X \times Y} e^{f(x) + g(y) - \lambda_i c_i(x,y)} d\mu(x) d\nu(y) - 1 \right)$$

A similar proof as the one of the Theorem 4 allows to extend the results for lower semi-continuous cost functions.
C.4 Projected Accelerated Gradient Descent

Proposition 14. Let \( a \in \Delta_N \) and \( b \in \Delta_m \) and \( C := (C_i)_{1 \leq i \leq N} \in (\mathbb{R}^{n \times m})_N \) be \( N \) cost matrices and \( \varepsilon := (\varepsilon, ..., \varepsilon) \) where \( \varepsilon > 0 \). Then by denoting \( K_{i} = \exp(-\lambda_i C_i / \varepsilon) \), we have

\[
\overline{\text{MOT}}_C^\varepsilon(a, b) = \sup_{\lambda \in \Delta_N} \sup_{f \in \mathbb{R}^n, g \in \mathbb{R}^m} F_C^\varepsilon(\lambda, f, g) := \langle f, a \rangle + \langle g, b \rangle - \varepsilon \left[ \log \left( \sum_{i=1}^{N} \langle e^f, K_i e^{b/\varepsilon} \rangle \right) + 1 \right].
\]

Moreover, \( F_C^\varepsilon \) is concave, differentiable and \( \nabla F \) is Lipschitz-continuous on \( \mathbb{R}^N \times \mathbb{R}^m \).

Proof. Let \( \mathcal{Q} := \left\{ P := (P_1, ..., P_N) \in (\mathbb{R}^{n \times m})^N : \sum_{i=1}^{N} \sum_{i,j} P_{i,j} = 1 \right\} \). Note that \( \Gamma_{a,b}^N \subset \mathcal{Q} \), therefore from the primal formulation of the problem we have that

\[
\overline{\text{MOT}}_C^\varepsilon(a, b) = \sup_{\lambda \in \Delta_N} \inf_{P \in \Gamma_{a,b}^N} \sum_{i=1}^{N} \lambda_i \langle P_i, C_i \rangle - \varepsilon H(P_i)
\]

\[
= \sup_{\lambda \in \Delta_N} \inf_{P \in \mathcal{Q}} \sup_{f \in \mathbb{R}^n, g \in \mathbb{R}^m} \sum_{i=1}^{N} \lambda_i \langle P_i, C_i \rangle - \varepsilon H(P_i)
\]

\[
+ f^T \left( a - \sum_i P_i 1_m \right) + g^T \left( b - \sum_i P_i^T 1_n \right)
\]

The constraints are qualified for this convex problem, hence by Slater’s sufficient condition [3, Section 5.2.3], strong duality holds. Therefore we have

\[
\overline{\text{MOT}}_C^\varepsilon(a, b) = \sup_{\lambda \in \Delta_N} \sup_{f \in \mathbb{R}^n, g \in \mathbb{R}^m} \inf_{P \in \mathcal{Q}} \sum_{i=1}^{N} \lambda_i \langle P_i, C_i \rangle - \varepsilon H(P_i)
\]

\[
+ f^T \left( a - \sum_i P_i 1_m \right) + g^T \left( b - \sum_i P_i^T 1_n \right)
\]

\[
= \sup_{\lambda \in \Delta_N} \sup_{f \in \mathbb{R}^n, g \in \mathbb{R}^m} \inf_{P \in \mathcal{Q}} \left( \sum_{k=1}^{N} \sum_{i,j} P_{i,j}^k \left( \lambda_k C_{i,j}^k + \varepsilon \left( \log(P_{i,j}^k) - 1 \right) - f_i - g_j \right) \right)
\]

Let us now focus on the following problem:

\[
\inf_{P \in \mathcal{Q}} \sum_{k=1}^{N} \sum_{i,j} P_{i,j}^k \left( \lambda_k C_{i,j}^k + \varepsilon \left( \log(P_{i,j}^k) - 1 \right) - f_i - g_j \right)
\]

Note that for all \( i, j, k \) and some small \( \delta \),

\[
P_{i,j}^k \left( \lambda_k C_{i,j}^k - \varepsilon \left( \log(P_{i,j}^k) - 1 \right) - f_i - g_j \right) < 0
\]

if \( P_{i,j}^k \in (0, \delta) \) and this quantity goes to 0 as \( P_{i,j}^k \) goes to 0. Therefore \( P_{i,j}^k > 0 \) and the problem becomes

\[
\inf_{P > 0} \sup_{\nu \in \mathbb{R}} \sum_{k=1}^{N} \sum_{i,j} P_{i,j}^k \left( \lambda_k C_{i,j}^k + \varepsilon \left( \log(P_{i,j}^k) - 1 \right) - f_i - g_j \right) + \nu \left( \sum_{k=1}^{N} \sum_{i,j} P_{i,j}^k - 1 \right).
\]
The solution to this problem is for all \( k \in \{1,..,N\}, \)

\[
P_k = \frac{\exp\left(\frac{f_1^T + 1_m g^T - \lambda_k C_k}{\epsilon}\right)}{\sum_{k=1}^N \sum_{i,j} \exp\left(\frac{f_i + g_j - \lambda_k C_{i,j}^k}{\epsilon}\right)}
\]

Therefore we obtain that

\[
\hat{\text{MOT}}^\epsilon_C(a,b) = \sup_{\lambda \in \Delta_N^+} \sup_{f \in \mathbb{R}^n, \ g \in \mathbb{R}^m} \langle f, a \rangle + \langle g, b \rangle - \epsilon \sum_{k=1}^N \sum_{i,j} P_{i,j}^k \left[ \log \left( \sum_{k=1}^N \sum_{i,j} \exp\left(\frac{f_i + g_j - \lambda_k C_{i,j}^k}{\epsilon}\right) \right) + 1 \right]
\]

From now on, we denote for all \( \lambda \in \Delta_N^+ \)

\[
\hat{\text{MOT}}^{\epsilon,\lambda}_C(a,b) := \inf_{P \in \Gamma_{a,b}^N} \lambda_i(P_i, C_i) - \epsilon H(P_i)
\]

\[
\hat{\text{MOT}}^{\epsilon,\lambda}_C(a,b) := \sup_{f \in \mathbb{R}^n, \ g \in \mathbb{R}^m} \langle f, a \rangle + \langle g, b \rangle - \epsilon \left[ \log \left( \sum_{k=1}^N \sum_{i,j} \exp\left(\frac{f_i + g_j - \lambda_k C_{i,j}^k}{\epsilon}\right) \right) + 1 \right]
\]

which has just been shown to be dual and equal. Thanks to [26, Theorem 1], as for all \( \lambda \in \mathbb{R}^N, P \in \Gamma_{a,b}^N \rightarrow \sum_{i=1}^N \lambda_i(P_i, C_i) - \epsilon H(P_i) \) is \( \epsilon \)-strongly convex, then for all \( \lambda \in \mathbb{R}^N, (f,g) \rightarrow \nabla_{(f,g)} F(\lambda, f, g) \) is \( \|A\|_{1 \rightarrow 2} = \sqrt{2N} \) Lipschitz-continuous where \( A \) is the linear operator of the equality constraints of the primal problem. Moreover this norm is equal to the maximum Euclidean norm of a column of \( A \). By definition, each column of \( A \) contains only \( 2N \) non-zero elements, which are equal to one. Hence, \( \|A\|_{1 \rightarrow 2} = \sqrt{2N} \). Let us now show that for all \( (f,g) \in \mathbb{R}^n \times \mathbb{R}^m \lambda \in \mathbb{R}^N \rightarrow \nabla_{\lambda} F(\lambda, f, g) \) is also Lipschitz-continuous. Indeed we remarks that

\[
\frac{\partial^2 F}{\partial \lambda_q \partial \lambda_k} = \frac{1}{\epsilon^2} \left[ \sigma_{q,1}(\lambda) \sigma_{k,1}(\lambda) - \nu(\sigma_{k,2}(\lambda) \mathbb{1}_{k=q}) \right]
\]

where \( \mathbb{1}_{k=q} = 1 \) iff \( k = q \) and 0 otherwise, for all \( k \in \{1,...,N\} \) and \( p \geq 1 \)

\[
\sigma_{k,p}(\lambda) = \sum_{i,j} (C_{i,j}^k)^p \exp\left(\frac{f_i + g_j - \lambda_k C_{i,j}^k}{\epsilon}\right)
\]

\[
\nu = \sum_{k=1}^N \sum_{i,j} \exp\left(\frac{f_i + g_j - \lambda_k C_{i,j}^k}{\epsilon}\right).
\]
Let $v \in \mathbb{R}^N$, and by denoting $\nabla^2_{(f,g)} F$ the Hessian of $F$ with respect to $\lambda$ for fixed $f, g$ we obtain first that

$$
\begin{align*}
    v^T \nabla^2_{(f,g)} F v &= \frac{1}{\varepsilon v^2} \left[ \left( \sum_{k=1}^N v_k \sigma_{q_1}(\lambda) \right)^2 - \nu \sum_{k=1}^N v_k^2 \sigma_{q_2} \right] \\
    &\leq \frac{1}{\varepsilon v^2} \left( \sum_{k=1}^N v_k \sigma_{q_1}(\lambda) \right)^2 \\
    &\leq \frac{1}{\varepsilon v^2} \left( \sum_{k=1}^N |v_k| \sum_{i,j} \exp \left( \frac{f_i + g_j - \lambda_k C_{k}^{i,j}}{\varepsilon} \right) \right)^2 \\
    &\leq \frac{1}{\varepsilon v^2} \left[ \left( \sum_{k=1}^N v_k \sigma_{q_1}(\lambda) \right)^2 - \left( \sum_{k=1}^N |v_k| \sum_{i,j} |C_{k}^{i,j}| \exp \left( \frac{f_i + g_j - \lambda_k C_{k}^{i,j}}{\varepsilon} \right) \right)^2 \right] \\
    &\leq 0
\end{align*}
$$

Indeed the last two inequalities come from Cauchy Schwartz. Moreover we have

$$
\begin{align*}
    \frac{1}{\varepsilon v^2} \left[ \left( \sum_{k=1}^N v_k \sigma_{q_1}(\lambda) \right)^2 - \nu \sum_{k=1}^N v_k^2 \sigma_{q_2} \right] &= v^T \nabla^2_{(f,g)} F v \leq 0 \\
    - \sum_{k=1}^N v_k^2 \sigma_{q_2} &\leq \sum_{k=1}^N \max_{1 \leq i \leq N} (||C_i||_{\infty}^2) \\
    \varepsilon &\leq \frac{\varepsilon}{\varepsilon v^2}
\end{align*}
$$

Therefore we deduce that $\lambda \in \mathbb{R}^N \rightarrow \nabla \lambda F(\lambda, f, g)$ is \( \frac{\max (\max_{1 \leq i \leq N} (||C_i||_{\infty}^2))}{\varepsilon} \) Lipschitz-continuous, hence $\nabla F(\lambda, f, g)$ is \( \frac{\max (\max_{1 \leq i \leq N} (||C_i||_{\infty}^2))}{\varepsilon} \) Lipschitz-continuous on $\mathbb{R}^N \times \mathbb{R}^n \times \mathbb{R}^m$.

Denote $L := \frac{\max (\max_{1 \leq i \leq N} (||C_i||_{\infty}^2))}{\varepsilon}$ the Lipschitz constant of $F^\varepsilon_{\lambda}$. Moreover for all $\lambda \in \mathbb{R}^N$, let $\text{Proj}_{\Delta_N^\varepsilon}(\lambda)$ the unique solution of the following optimization problem

$$
\min_{x \in \Delta_N^\varepsilon} ||x - \lambda||^2_2.
$$

(27)

Let us now introduce the following algorithm.

Algorithm 2 Accelerated Projected Gradient Ascent Algorithm

Input: $C = (C_i)_{1 \leq i \leq N}$, $a$, $b$, $\varepsilon$, $L$

Init: $f^{-1} = f^0 \leftarrow 0_n$; $g^{-1} = g^0 \leftarrow 0_m$; $\lambda^{-1} = \lambda^0 \leftarrow (1/N, ..., 1/N) \in \mathbb{R}^N$

for $k = 1, 2, \ldots$

\begin{itemize}
    \item \( (v, w, z)^T \leftarrow (\lambda^{k-1}, f^{k-1}, g^{k-1} - \lambda^{k-1})^T + \frac{k-2}{k+2} ((\lambda^{k-1}, f^{k-1}, g^{k-1})^T - (\lambda^{k-2}, f^{k-2}, g^{k-2})^T) \);
    \item $\lambda^k \leftarrow \text{Proj}_{\Delta_N^\varepsilon}(v + \frac{1}{L} \nabla \lambda F^\varepsilon_{\lambda}(v, w, z))$;
    \item $(g^k, f^k)^T \leftarrow (w, z)^T + \frac{1}{L} \nabla_{(f,g)} F^\varepsilon_{\lambda}(v, w, z)$.
\end{itemize}

end

Result: $\lambda, f, g$
give us that the accelerated projected gradient ascent algorithm achieves the optimal rate for first order methods of $O(1/k^2)$ for smooth functions. To perform the projection we use the algorithm proposed in [31] which finds the solution of (27) after $O(N \log(N))$ algebraic operations [38].

D Additional Experiments

Barycenter experiments. To illustrate MOT, we provide two more barycenter transformations in Figures 4 and 5

Figure 4: The two first row starting from the above are Wasserstein barycenters using respectively square $\ell_2$ cost and the cubic $\ell_3$ taken coordinate by coordinate. The last row represent the MOT barycenter with respect to these two costs. From left to right: Progressive barycentric transformation of “9” to “0”. Both shapes are normalized to probability distributions.

Figure 5: The two first row starting from the above are Wasserstein barycenters using respectively square $\ell_2$ cost and the cubic $\ell_3$ taken coordinate by coordinate. The last row represent the MOT barycenter with respect to these two costs. From left to right: Progressive barycentric transformation of “2” to “8”. Both shapes are normalized to probability distributions.

Approximation for the Dudley Metric. Figure 6 illustrates the convergence of the entropic regularization approximation when $\epsilon \to 0$. To do so we plot the relative error from the ground truth defined as $RE := \frac{\text{MOT}_\epsilon - \beta_d}{\beta_d}$ for different regularizations where $\beta_d$ is obtained by solving the exact linear program and $\text{MOT}_\epsilon$ is obtained by our proposed Alg. [1]
Figure 6: In this experiment, we draw 100 samples from two normal distributions and we plot the relative error from ground truth for different regularizations. We consider the case where two costs are involved: $c_1 = 2 \times 1_{x \neq y}$, and $c_2 = d$ where $d$ is the Euclidean distance. This case corresponds exactly to the Dudley metric (see Proposition 2). We remark that as $\varepsilon \to 0$, the approximation error goes also to 0.