Factored couplings in multi-marginal optimal transport via difference of convex programming

Huy Quang Tran, Hicham Janati, Ievgen Redko, Rémi Flamary, Nicolas Courty

Factored couplings in multi-marginal optimal transport via difference of convex programming

Quang Huy Tran
Univ. Bretagne-Sud, CNRS, IRISA
F-56000 Vannes
quang-huy.tran@univ-ubs.fr

Hicham Janati
École Polytechnique, CMAP, UMR 7641
F-91120 Palaiseau
hicham.janati@polytechnique.edu

Ievgen Redko
Univ Lyon, UJM-Saint-Etienne, CNRS, UMR 5516
F-42023 Saint-Etienne
ievgen.redko@univ-st-etienne.fr

Rémi Flamary
École Polytechnique, CMAP, UMR 7641
F-91120 Palaiseau
remi.flamary@polytechnique.edu

Nicolas Courty
Univ. Bretagne-Sud, CNRS, IRISA
F-56000 Vannes
nicolas.courty@irisa.fr

Abstract

Optimal transport (OT) theory underlies many emerging machine learning (ML) methods nowadays solving a wide range of tasks such as generative modeling, transfer learning and information retrieval. These latter works, however, usually build upon a traditional OT setup with two distributions, while leaving a more general multi-marginal OT formulation somewhat unexplored. In this paper, we study the multi-marginal OT (MMOT) problem and unify several popular OT methods under its umbrella by promoting structural information on the coupling. We show that incorporating such structural information into MMOT results in an instance of a different of convex (DC) programming problem allowing us to solve it numerically. Despite high computational cost of the latter procedure, the solutions provided by DC optimization are usually as qualitative as those obtained using currently employed optimization schemes.

1 Introduction

Broadly speaking, the classic OT problem provides a principled approach for transporting one probability distribution onto another following the principle of the least effort. Such a problem, and the distance on the space of probability distributions derived from it, arise in many areas of machine learning (ML) including generative modeling, transfer learning and information retrieval, where OT has been successfully applied. A natural extension of classic OT, in which the admissible transport plan (a.k.a coupling) can have more than two prescribed marginal distributions, is called the multi-marginal optimal transport (MMOT) [Gangbo and Swiech, 1998]. The latter has several attractive properties: it enjoys the duality theory [Kellerer, 1984] and finds connections with the Wasserstein barycenter problem [Agueh and Carlier, 2011] used for data averaging and probabilistic graphical models [Haasler et al., 2020]. While being far less popular than the classic OT with two marginals, MMOT is a very useful framework on its own with some notable recent applications in generative adversarial networks [Cao et al., 2019], clustering [Mi and Bento, 2020] and domain adaptation [Hui et al., 2018, He et al., 2019], to name a few.

Preprint. Under review.

The recent success of OT in ML is often attributed to the entropic regularization [Cuturi, 2013] where the authors imposed a constraint on the coupling matrix forcing it to be closer to the independent coupling given by the rank-one product of the marginals. Such a constraint leads to the appearance of the strongly convex entropy term in the objective function and allows the entropic OT problem to be solved efficiently using simple Sinkhorn-Knopp matrix balancing algorithm. In addition to this, it was also noticed that structural constraints allow to reduce the prohibitively high sample-complexity of the classic OT problem [Genevay et al., 2019, Forrow et al., 2019]. While these and several other recent works [Lin et al., 2021, Scetbon et al., 2021] addressed structural constraints in the classic OT problem with two marginals, none of them considered a much more challenging case of doing so in a multi-marginal setting. On the other hand, while the work of [Haasler et al., 2020] considers the MMOT problem in which the cost tensor induced by a graphical structure, it does not naturally promote the factorizability of transportation plans.

Contributions In this paper, we define and study a general MMOT problem with structural penalization on the coupling matrix. We start by showing that a such formulation includes several popular OT methods as special cases and allows to gain deeper insights into them. We further consider a relaxed problem where hard constraints are replaced by a regularization term and show that it leads to an instance of the difference of convex programming problem. A numerical study of the solutions obtained when solving the latter in cases of interest highlights their competitive performance when compared to solutions provided by the optimization strategies used previously.

2 Preliminary knowledge

Notations. For each integer $n \geq 1$, we write $[n] := \{1, ..., n\}$. For each discrete probability measure μ with finite support, its negative entropy is defined as $H(\mu) = \langle \mu, \log \mu \rangle$, where the logarithm operator is element-wise, with the convention that $0 \log 0 = 0$. Here, $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product. The Kullback-Leibler divergence between two discrete probability measures μ and ν with finite supports is defined as

$$\mathrm{KL}(\mu|\nu) = \begin{cases} \langle \mu, \log \frac{\mu}{\nu} \rangle, \text{ if } \mu \text{ is absolutely continuous with respect to } \nu \\ \infty, \text{ otherwise.} \end{cases}$$

where the division operator in the logarithm is element-wise.

In what follows, we write $\mathcal{T} = (1, ..., N)$, for some integer $N \geq 1$. For any positive integers $a_1, ..., a_N$, we call $P \in \mathbb{R}^{a_1 \times \dots \times a_N}$ a N-D tensor. In particular, a 1-D tensor is a vector and 2-D tensor is a matrix. A tensor is a probability tensor if its entries are nonnegative and the sum of all entries is 1. Given N probability vectors $\mu_1, ..., \mu_N$, we write $\mu = (\mu_n)_{n=1}^N$. We denote $\Sigma_{\mathcal{T}}$ the set of N-D probability tensors and $U_{\mathcal{T}} \subset \Sigma_{\mathcal{T}}$ the set of tensors whose N marginal distributions are $\mu_1, ..., \mu_N$. Any coupling in $U_{\mathcal{T}}$ is said to be *admissible*. We write $\mu_{\mathcal{T}} := \mu_1 \otimes ... \otimes \mu_N$ the tensor product (or measure product) of $\mu_1, ..., \mu_N$.

Multi-marginal OT problem. Given a collection of N probability vectors $\mu = (\mu_n \in \mathbb{R}^{a_n})_{n=1}^N$ and a N-D cost tensor $C \in \mathbb{R}^{a_1 \times \dots \times a_N}$, the MMOT problem reads

$$\mathrm{MMOT}(\mu) = \inf_{P \in U_{\mathcal{T}}} \langle C, P \rangle. \tag{1}$$

In practice, such a formulation is intractable to optimize in a discrete setting as it results in a linear program where the number of constraints grows exponentially in N. A more tractable strategy for solving MMOT is to consider the following entropic regularization problem

$$\inf_{P \in U_{\mathcal{T}}} \langle C, P \rangle + \varepsilon \mathrm{KL}(P|\mu_{\mathcal{T}}). \tag{2}$$

which can be solved using Sinkhorn's algorithm Benamou et al. [2014]. We refer the interested reader to Supplementary materials for algorithmic details.

3 Factored Multi-marginal Optimal Transport

In this section, we first define a factored MMOT (F-MMOT) problem where we seek to promote a structure on the optimal coupling given such as a factorization into a tensor product. Interestingly, such a formulation can be shown to include several other OT problems as special cases. Then, we introduce a relaxed version called MMOT-DC where the factorization constraint is smoothly promoted through a Kullback-Leibler penalty.

3.1 Motivation

Before a formal statement of our problem, we first give a couple of motivating examples showing why and when structural constraints on the coupling matrix can be beneficial. To this end, first note that a trivial example of the usefulness of such constraints in OT is the famous entropic regularization. Indeed, while most of the works define the latter by adding negative entropy of the coupling to the classic OT objective function directly, the original idea was to constraint the sought coupling to remain close (to some extent) to a rank-one product of the two marginal distributions. The appearance of negative entropy in the final objective function is then only a byproduct of such constraint due to the decomposition of the KL divergence into a sum of three terms with two of them being constant. Below we give two more examples of real-world applications related to MMOT problem where a certain decomposition imposed on the coupling tensor can be desirable.

Multi-source multi-target translation. A popular task in computer vision is to match images across different domains in order to perform the so-called image translation. Such tasks are often tackled within the GAN framework where one source domain from which the translation is performed, is matched with multiple target domains modeled using generators. While MMOT was applied in this context by Cao et al. [2019] when only one source was considered, its application in a multi-source setting may benefit from structural constraints on the coupling tensor incorporating the human prior on what target domains each source domain should be matched to.

Multi-task reinforcement learning. In this application, the goal is to learn individual policies for a set of agents while taking into account the similarities between them and hoping that the latter will improve the individual policies. A common approach is to consider an objective function consisting of two terms where the first term is concerned with learning individual policies, while the second forces a consensus between them. Similar to the example considered above, MMOT problem was used to promote the consensus across different agents' policies in [Cohen et al., 2021], even though such a consensus could have benefited from a prior regarding the semantic relationships between the learned tasks.

3.2 Factored MMOT and its relaxation

We start by giving several definitions used in the following parts of the paper.

Definition 3.1 (Tuple partition) *A sequence of tuples $(\mathcal{T}_m)_{m=1}^M$, with $M \leq N$, is called a partition of a N-tuple $\mathcal{T}$ if the tuples $\mathcal{T}_1, ..., \mathcal{T}_M$ are nonempty and disjoint, and their concatenation gives $\mathcal{T}$.*

Here, we implicitly take into account the order of the tuple, which is not the case for the partition of a set. If there exists a tuple in $(\mathcal{T}_m)_{m=1}^M$ which contains only one element, then we say $(\mathcal{T}_m)_{m=1}^M$ is *degenerate*.

Definition 3.2 (Marginal tensor) *Given a tensor $P \in \mathbb{R}^{a_1 \times \dots \times a_N}$ and a partition $(\mathcal{T}_m)_{m=1}^M$ of $\mathcal{T} = (1, ..., N)$, we call $P_{\#m}$ the $\mathcal{T}_m$-marginal tensor, by summing P over all dimensions not in $\mathcal{T}_m$. We write $P_{\#\mathcal{T}} = P_{\#1} \otimes ... \otimes P_{\#M} \in \mathbb{R}^{a_1 \times \dots \times a_N}$ the tensor product of its marginal tensors.*

For example, for $M = N = 2$ and $\mathcal{T} = (1, 2)$, we have $\mathcal{T}_1 = (1)$ and $\mathcal{T}_2 = (2)$. So, given a matrix $P \in \mathbb{R}^{a_1 \times a_2}$, its marginal tensors $P_{\#1}$ and $P_{\#2}$ are vectors in $\mathbb{R}^{a_1}$ and $\mathbb{R}^{a_2}$, respectively, defined by $(P_{\#1})_i = \sum_j P_{ij}$ and $(P_{\#2})_j = \sum_i P_{ij}$ for $(i, j) \in [a_1] \times [a_2]$. The tensor product $P_{\#\mathcal{T}} \in \mathbb{R}^{a_1 \times a_2}$ is defined by $(P_{\#\mathcal{T}})_{ij} = (P_{\#1})_i (P_{\#2})_j$.

Clearly, if P is a probability tensor, then so are its marginal tensors.

Suppose $\mathcal{T}_m = (p, ..., q)$ for $m = 1, ..., M$ and $1 \leq p \leq q \leq N$. We denote $\Sigma_{\mathcal{T}_m}$ the set of probability tensors in $\mathbb{R}^{a_p \times \dots \times a_q}$ and $U_{\mathcal{T}_m}$ the set of probability tensors in $\mathbb{R}^{a_p \times \dots \times a_q}$ whose (r)-marginal vector is μ_r, for every $r = p, ..., q$. We define $\mu_{\mathcal{T}_m} := \mu_p \otimes ... \otimes \mu_q$ by $(\mu_{\mathcal{T}_m})_{i_p \dots i_q} := \prod_r (\mu_r)_{i_r}$, for $i_r \in [a_r]$, with $r = p, ..., q$. Clearly, we have $\mu_{\mathcal{T}} = \mu_{\mathcal{T}_1} \otimes ... \otimes \mu_{\mathcal{T}_M}$.

Definition 3.3 (Factored MMOT) *Given a partition $(\mathcal{T}_m)_{m=1}^M$ of $\mathcal{T}$ and a collection of histograms μ, we consider the following OT problem*

$$\textit{F-MMOT}\big((\mathcal{T}_m)_{m=1}^M, \mu\big) = \inf_{P \in \mathcal{F}_M} \langle C, P \rangle, \tag{F-MMOT}$$

where $\mathcal{F}_M \subset U_\mathcal{T}$ is the set of admissible couplings which can be factorized as a tensor product of M component probability tensors in $\Sigma_{\mathcal{T}_1}, ..., \Sigma_{\mathcal{T}_M}$.

Several remarks are in order here. First, one should note that the partition considered above is in general not degenerate meaning that the decomposition can involve tensors of an arbitrary order $< N$. Second, the decomposition in this setting depicts the prior knowledge regarding the tuples of measures which should be independent: the couplings for the measures from different tuples will be degenerate and the optimal coupling tensor will be reconstructed from couplings of each tuple separately. Third, suppose the partition $(\mathcal{T}_m)_{m=1}^M$ is not degenerate and $M = 2$, i.e. the tensor is factorized as product of two tensors, the problem F-MMOT is equivalent to a variation of low nonnegative rank OT (see Appendix for a proof).

As for the existence of the solution to this problem, we have that $\mathcal{F}_M$ is compact because it is a close subset of the compact set $U_\mathcal{T}$, which implies that F-MMOT always admits a solution. Furthermore, observe that

$$\begin{aligned}\mathcal{F}_M &= \{P \in U_\mathcal{T} : P = P_1 \otimes ... \otimes P_M, \text{where } P_m \in \Sigma_{\mathcal{T}_m}, \forall m = 1, ..., M\} \\ &= \{P \in \Sigma_\mathcal{T} : P = P_1 \otimes ... \otimes P_M, \text{where } P_m \in U_{\mathcal{T}_m}, \forall m = 1, ..., M\}.\end{aligned}$$

Thus, the problem F-MMOT can be rewritten as

$$\text{F-MMOT}\big((\mathcal{T}_m)_{m=1}^M, \mu\big) = \inf_{\substack{P_m \in U_{\mathcal{T}_m} \\ \forall m=1,...,M}} \langle C, P_1 \otimes ... \otimes P_M \rangle.$$

So, if $\mathcal{T}_1, ..., \mathcal{T}_M$ are 2-tuples and two marginal distributions corresponding to each $U_{\mathcal{T}_m}$ are identical and uniform, then by Birkhoff's theorem [Birkhoff, 1946], F-MMOT admits an optimal solution in which each component tensor P_m is a permutation matrix.

COOT and GW as special cases. When $N = 4$ and $M = 2$ with $\mathcal{T}_1 = (a_1, a_2)$ and $\mathcal{T}_2 = (a_3, a_4)$, the problem F-MMOT becomes the CO-Optimal transport [Redko et al., 2020], where the two component tensors are known as sample and feature couplings. If furthermore, $a_1 = a_3, a_2 = a_4$, and $\mu_1 = \mu_3, \mu_2 = \mu_4$, it becomes a lower bound of the discrete Gromov-Wasserstein distance [Mémoli, 2011]. This means that our formulation can be seen as a generalization of several OT formulations.

Observe that if a probability tensor P can be factorized as a tensor product of probability tensors, i.e. $P = P_1 \otimes ... \otimes P_M$, then each P_m is also the $\mathcal{T}_m$-marginal tensor of P. This prompts us to consider the following relaxation of factored MMOT, where the hard constraint $\mathcal{F}_M$ is replaced by a regularization term.

Definition 3.4 (Relaxed Factored MMOT) *Given $\varepsilon \geq 0$, a partition $(\mathcal{T}_m)_{m=1}^M$ of $\mathcal{T}$ and a collection of measures μ, we define the following problem:*

$$\textit{MMOT-DC}_\varepsilon\big((\mathcal{T}_m)_{m=1}^M, \mu\big) = \inf_{P \in U_\mathcal{T}} \langle C, P \rangle + \varepsilon \textit{KL}(P | P_{\#\mathcal{T}}). \tag{MMOT-DC}$$

From the exposition above, one can guess that this relaxation is reminiscent of the entropic regularization in MMOT and coincides with it when $M = N$. As such it also recovers the classical entropic OT. One should note that the choice of the KL divergence is not arbitrary and its advantage will become clear when it comes to the algorithm. special case of MMOT-DC is when $M = N$, we recover the entropic MMOT.

After having defined the two optimization problems, we now set on exploring their theoretical properties.

3.3 Theoretical properties

Intuitively, the relaxed problem is expected to allow for solutions with a lower value of the final objective function. We formally prove the validity of this intuition below.

Proposition 3.1 *(Preliminary properties)*

1. $\forall \varepsilon \geq 0, \textit{MMOT}(\mu) \leq \textit{MMOT-DC}_\varepsilon\big((\mathcal{T}_m)_{m=1}^M, \mu\big) \leq \textit{F-MMOT}\big((\mathcal{T}_m)_{m=1}^M, \mu\big)$.
2. $\forall \varepsilon > 0, \textit{MMOT-DC}_\varepsilon\big((\mathcal{T}_m)_{m=1}^M, \mu\big) = 0$ *if and only if* $\textit{F-MMOT}\big((\mathcal{T}_m)_{m=1}^M, \mu\big) = 0$.

An interesting property of MMOT-DC$_\varepsilon$ is that it interpolates between MMOT and F-MMOT. Informally, for very large ε, the KL divergence term dominates, so the optimal transport plans tend to be factorizable. On the other hand, for very small ε, the KL divergence term becomes negligible and we approach MMOT. The result below formalizes this intuition.

Proposition 3.2 *(Interpolation between MMOT and F-MMOT) For any partition $(\mathcal{T}_m)_{m=1}^M$ of $\mathcal{T}$ and for $\varepsilon > 0$, let P_ε be a minimiser of the problem MMOT-DC$_\varepsilon\big((\mathcal{T}_m)_{m=1}^M, \mu\big)$.*

1. *When $\varepsilon \to \infty$, one has MMOT-DC$_\varepsilon\big((\mathcal{T}_m)_{m=1}^M, \mu\big) \to$ F-MMOT$\big((\mathcal{T}_m)_{m=1}^M, \mu\big)$. In this case, any cluster point of the sequence of minimisers $(P_\varepsilon)_\varepsilon$ is a minimiser of F-MMOT$\big((\mathcal{T}_m)_{m=1}^M, \mu\big)$.*
2. *When $\varepsilon \to 0$, then MMOT-DC$_\varepsilon\big((\mathcal{T}_m)_{m=1}^M, \mu\big) \to$ MMOT(μ). In this case, any cluster point of the sequence of minimisers $(P_\varepsilon)_\varepsilon$ is a minimiser of MMOT(μ).*

GW distance revisited. Somewhat surprisingly, the relaxation MMOT-DC also allows us to prove the equality between GW distance and COOT in the discrete setting. Let $\mathcal{X}$ be a finite subset (of size m) of a certain metric space. Denote $C_x \in \mathbb{R}^{m\times m}$ its similarity matrix (e.g. distance matrix). We define similarly the set $\mathcal{Y}$ of size n and the corresponding similarity matrix $C_y \in \mathbb{R}^{n\times n}$. We also assign two discrete probability measures $\mu_x \in \mathbb{R}^m$ and $\mu_y \in \mathbb{R}^n$ to $\mathcal{X}$ and $\mathcal{Y}$, respectively. The GW distance is then defined as

$$\text{GW}(C_x, C_y) = \inf_{P \in U(\mu_x, \mu_y)} \langle L_p(C_x, C_y), P \otimes P \rangle,$$

and the COOT reads

$$\text{COOT}(C_x, C_y) = \inf_{\substack{P \in U(\mu_x, \mu_y) \\ Q \in U(\mu_x, \mu_y)}} \langle L_p(C_x, C_y), P \otimes Q \rangle,$$

where the 4D tensor $L_p(X, Y) \in \mathbb{R}^{m\times n\times m\times n}$ is defined by $\big(L_p(X,Y)\big)_{i,j,k,l} = |X_{i,k} - Y_{j,l}|^p$, for some $p \geq 1$, and $U(\mu, \nu)$ is the set of couplings in $\mathbb{R}^{m\times n}$ whose two marginal distributions are μ and ν. When C_x and C_y are two squared Euclidean distance matrices, and $p = 2$, it can be shown that the GW distance is equal to the COOT [Redko et al., 2020]. This is also true when $L_p(C_x, C_y)$ is a negative definite kernel Séjourné et al. [2020]. Here, we establish another case where this equality still holds.

Corollary 3.3 *If $L_p(C_x, C_y)$ is a (symmetric) Lipschitz function with respect to both inputs, and induces a strictly positive definite kernel $\exp\big(-\frac{L_p(C_x, C_y)}{\varepsilon}\big)$ on $(\mathcal{X}\times\mathcal{Y})^2$, for every $\varepsilon > 0$, then there exists a solution (P, Q) of COOT such that $P = Q$. As a consequence, we have the equality between GW distance and COOT.*

The proof relies on the connection between MMOT-DC and COOT shown in the proposition 3.2, and that the two $\mathcal{T}_1, \mathcal{T}_2$-marginal matrices of the 4-D solution of MMOT-DC are in fact identical, under the assumption of the cost tensor. The proof of the second claim is deferred to the Appendix.

4 Numerical solution

We now turn to computational aspects of the problem MMOT-DC. First, note that the KL divergence term can be decomposed as

$$\text{KL}(P|P_{\#\mathcal{T}}) = H(P) - \sum_{m=1}^{m} H(P_{\#m}),$$

Algorithm 1 DC algorithm for the problem MMOT-DC

Input. Cost tensor C, partition $(\mathcal{T}_m)_{m=1}^M$ of $\mathcal{T}$, histograms $\mu_1, ..., \mu_N$, hyperparameter $\varepsilon > 0$, initialization $P^{(0)}$, tuple of initial dual vectors for the Sinkhorn step $(f_1^{(0)}, ..., f_N^{(0)})$.
Output. Tensor $P \in U_{\mathcal{T}}$.
While not converge

1. Compute $G^{(t)} = \sum_{m=1}^{M} \nabla_P H(P_{\#m}^{(t)})$ the gradient of teh concave term.
2. Solve
$$P^{(t+1)} \in \arg\min_{P \in U_{\mathcal{T}}} \langle L(X,Y) - \varepsilon G^{(t)}, P \rangle + \varepsilon H(P),$$
using the Sinkhorn algorithm 3, with the tuple of initial dual vectors $(f_1^{(0)}, ..., f_N^{(0)})$.

where the function H_m defined by $H_m(P) := H(P_{\#m})$ is continuous and convex with respect to P. Now, the problem MMOT-DC becomes

$$\text{MMOT-DC}_\varepsilon\big((\mathcal{T}_m)_{m=1}^M, \mu\big) = \inf_{P \in U_{\mathcal{T}}} \langle C, P \rangle + \varepsilon H(P) - \varepsilon \sum_{m=1}^{M} H_m(P). \tag{3}$$

This is nothing but a Difference of Convex (DC) programming problem (which explains the name MMOT-DC), thanks to the convexity of the set $U_{\mathcal{T}}$ and the entropy function H. Thus, it can be solved by the DC algorithm [Tao and Souad, 1986, Pham and Thi, 1997] as follows: at the iteration t,

1. Calculate $G^{(t)} \in \partial(\sum_{m=1}^M H_m)(P^{(t)})$.
2. Solve $P^{(t+1)} \in \arg\min_{P \in U_{\mathcal{T}}} \langle C - \varepsilon G^{(t)}, P \rangle + \varepsilon H(P)$.

This algorithm is very easy to implement. Indeed, the second step is a classical entropic-regularized MMOT problem and can be solved by the Sinkhorn algorithm 3. In the first step, the gradient can be calculated explicitly. For the sake of simplicity, we illustrate the calculation in a simple case, where $M = 2$ and $N = 4$ with $\mathcal{T}_1, \mathcal{T}_2$ are two 2-tuples. The function $H_1 + H_2$ is continuous, so $h^{(t)} = \nabla_P (H_1 + H_2)(P^{(t)})$. Given a 4-D probability tensor P, we have

$$H_1(P) + H_2(P) = \sum_{i,j,k,l} P_{i,j,k,l} \log\big(\sum_{i,j} P_{i,j,k,l}\big) + P_{i,j,k,l} \log\big(\sum_{k,l} P_{i,j,k,l}\big).$$

So,

$$\frac{\partial (H_1 + H_2)}{\partial P_{i,j,k,l}} = \log\left(\sum_{i,j} P_{i,j,k,l}\right) + \frac{P_{i,j,k,l}}{\sum_{i,j} P_{i,j,k,l}} + \log\left(\sum_{k,l} P_{i,j,k,l}\right) + \frac{P_{i,j,k,l}}{\sum_{k,l} P_{i,j,k,l}}.$$

The complete DC algorithm for the problem 3 can be found in the algorithm 1. We observed that initialization is crucial to the convergence of algorithm which is not surprising for a non-convex problem. To accelerate the algorithm for large ε, we propose to use the warm-start strategy, which is similar to the one used in the entropic OT problem with very small regularization parameter [Schmitzer, 2019]. Its idea is simple: we consider an increasing finite sequence $(\varepsilon_n)_{n=0}^N$ approaching ε such that the solution P_{ε_0} of the problem MMOT-DC$_{\varepsilon_0}(X,Y)$ can be estimated quickly and accurately using the initialization $P^{(0)}$. Then we solve each successive problem MMOT-DC$_{\varepsilon_n}(X,Y)$ using the previous solution $P_{\varepsilon_{n-1}}$ as initialization. Finally, the problem MMOT-DC$_\varepsilon(X,Y)$ is solved using the solution P_{ε_N} as initialization.

Algorithm 2 DC algorithm with warm start for the problem MMOT-DC

Input. Cost tensor C, partition $(\mathcal{T}_m)_{m=1}^M$ of $\mathcal{T}$, histograms $\mu_1, ..., \mu_N$, hyperparameter $\varepsilon > 0$, initialization $P^{(0)}$, initial $\varepsilon_0 > 0$, step size $s > 1$, tuple of initial dual vectors $(f_1^{(0)}, ..., f_N^{(0)})$.
Output. Tensor $P \in U_{\mathcal{T}}$.

1. While $\varepsilon_0 < \varepsilon$:
   (a) Using algorithm 1, solve the problem MMOT-DC$_{\varepsilon_0}(X, Y)$ with initialization $P^{(0)}$ and $(f_1^{(0)}, ..., f_N^{(0)})$ to find the solution P_{ε_0} and its associated tuple of dual vectors $(f_1^{(\varepsilon_0)}, ..., f_N^{(\varepsilon_0)})$.
   (b) Set $P^{(0)} = P_{\varepsilon_0}, f_i^{(0)} = f_i^{(\varepsilon_0)}$, for $i = 1, ..., N$.
   (c) Increase regularization: $\varepsilon_0 := s\varepsilon_0$.
2. Using algorithm 1, solve the problem MMOT-DC$_\varepsilon(X, Y)$ using the initialization $P^{(0)}$ and $(f_1^{(0)}, ..., f_N^{(0)})$.

5 Experimental evaluation

In this section, we illustrate the use of MMOT-DC on simulated data. Rather than performing experiments in full generality, we choose the setting where $N = 4$ and $M = 2$ with $\mathcal{T}_1 = (1, 2)$ and $\mathcal{T}_2 = (3, 4)$, so that we can compare MMOT-DC with other popular solvers of COOT and GW distance. Given two matrices X and Y, we always consider the 4-D cost tensor C, where $C_{i,j,k,l} = |X_{i,k} - Y_{j,l}|^2$. On the other hand, we are not interested in the 4-D minimiser of MMOT-DC, but only in its two $\mathcal{T}_1, \mathcal{T}_2$-marginal matrices.

Solving COOT on a toy example. We generate a random matrix $X \in \mathbb{R}^{30\times 25}$, whose entries are drawn independently from the uniform distribution on the interval $[0, 1)$. We equip the rows and columns of X with two discrete uniform distributions on $[30]$ and $[25]$. We fix two permutation matrices $P \in \mathbb{R}^{30\times 30}$ (called sample permutation) and $Q \in \mathbb{R}^{25\times 25}$ (called feature permutation), then calculate $Y = PXQ$. We also equip the rows and columns of Y with two discrete uniform distributions on $[30]$ and $[25]$.

It is not difficult to see that COOT$(X, Y) = 0$ because (P, Q) is a solution. As COOT is a special case of F-MMOT, we see that MMOT-DC$_\varepsilon\big((\mathcal{T}_m)_{m=1}^2, \mu\big) = 0$, for every $\varepsilon > 0$, by proposition 3.1. In this experiment, we will check if marginalizing the minimizer of MMOT-DC allows us to recover the permutation matrices P and Q. As can be seen from the figure 1, MMOT-DC can recover the permutation positions, for various values of ε. On the other hand, it can not recover the true sparse permutation matrices because the Sinkhorn algorithm applied to the MMOT problem implicitly results in a dense tensor, thus having dense marginal matrices. For this reason, the loss only remains very close to zero, but never exactly.

We also plot, with some abuse of notation, the histograms of the error between the $(1, 3), (1, 4), (2, 3), (2, 4)$-marginal matrices of MMOT-DC and their independent counterpart from F-MMOT. In this example, in theory, as the F-MMOT optimal tensor P can be factorized as $P = P_{12} \otimes P_{34}$, it is immediate to see that $P_{12} = P_{14} = P_{23} = P_{24} \in \mathbb{R}^{30\times 25}$ are uniform matrices whose entries are $\frac{1}{750}$.

Quality of the MMOT-DC solutions. Now, we consider the situation where the true matching between two matrices is not known in advance and investigate the quality of the solutions returned by MMOT-DC to solve the COOT and GW problems. This means that we will look at the COOT loss $\langle C, P \otimes Q\rangle$, where the smaller the loss, the better when using both exact COOT and GW solvers and our relaxation.

We generate two random matrices $X \in \mathbb{R}^{20\times 3}$ and $Y \in \mathbb{R}^{30\times 2}$, whose entries are drawn independently from the uniform distribution on the interval $[0, 1)$. Then we calculate two corresponding squared Euclidean distance matrices of size 20 and 30. Their rows and columns are equipped with the discrete uniform distributions. In this case, the COOT loss coincides with the GW distance [Redko et al., 2020].

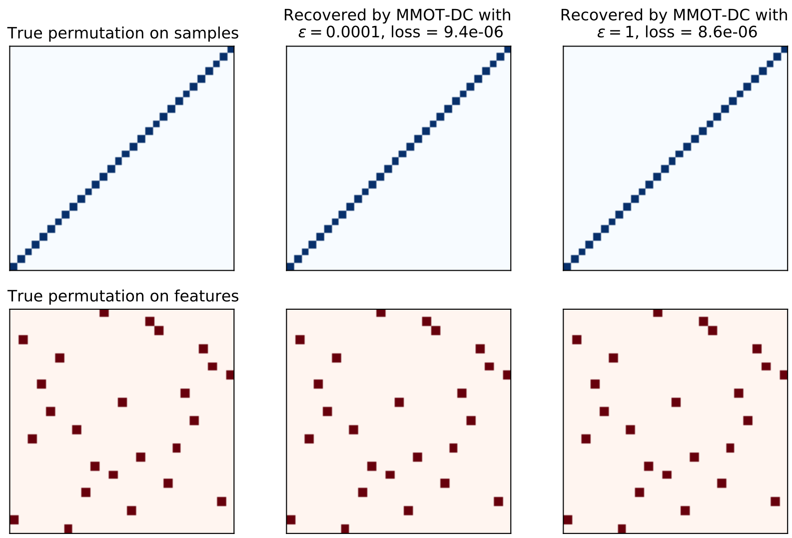

Figure 1: Couplings generated by COOT and MMOT-DC on the matrix recovering task.

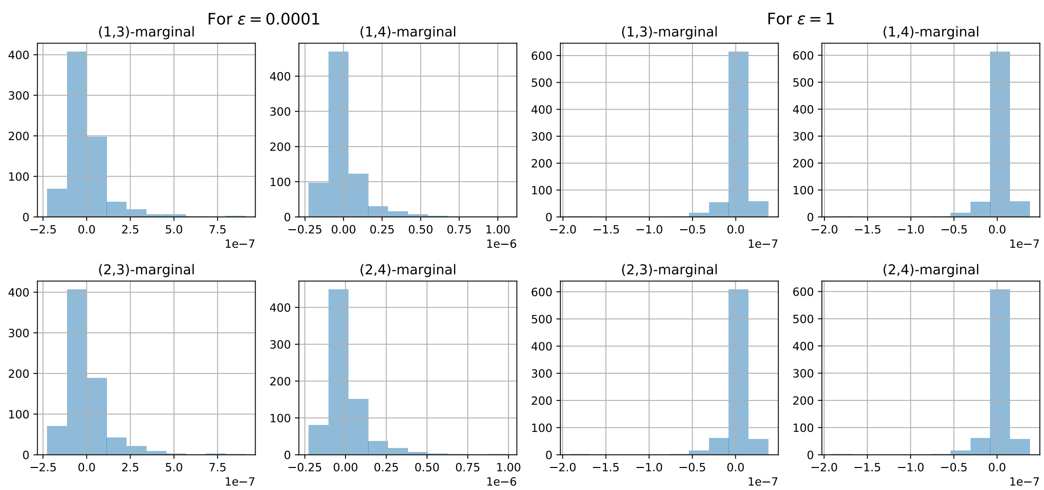

Figure 2: Histograms of difference between true independent marginal matrices and their approximations. We see that the marginal matrices obtained by the algorithm 1 approximate well the theoretical uniform matrices.

We compare four solvers:

1. The Frank-Wolfe algorithm [Frank and Wolfe, 1956] to solve the GW (GW-FW).
2. The projected gradient algorithm to solve the entropic GW distance Peyré et al. [2016] (EGW-PG). We choose the regularization parameter from $\{0.0008, 0.0016, 0.0032, 0.0064, 0.0128, 0.0256\}$ and pick the one which corresponds to smallest COOT loss.
3. The Block Coordinate Descent algorithm to approximate COOT and its entropic approximation Redko et al. [2020] (GW-BCD and EGW-BCD, respectively), where two additional KL divergences corresponding to two couplings are introduced. Both regularization parameters are tuned from $\{0, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1\}$, where 0 means that there is no regularization term for the corresponding coupling and we pick the pair whose COOT loss is the smallest.

GW-FW	EGW-PG	GW-BCD	EGW-BCD	MMOT-DC
0.083 ($\pm$ 0.035)	**0.079 ($\pm$ 0.035)**	0.083 ($\pm$ 0.035)	0.080 ($\pm$ 0.035)	0.082 ($\pm$ 0.036)

Table 1: Average and standard deviation of COOT loss of the solvers. MMOT-DC is competitive to other solvers, except for EGW-PG and EGW-BCD.

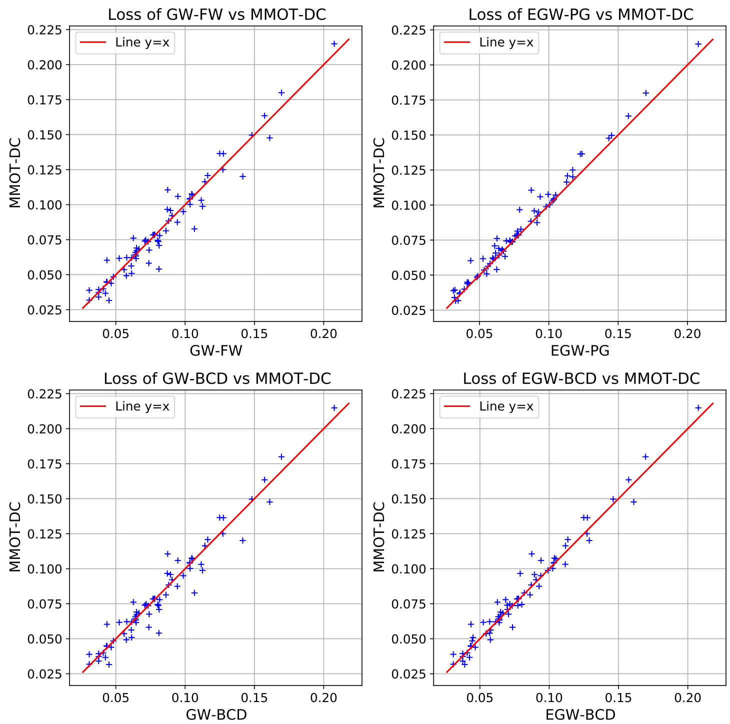

Figure 3: Scatter plots of MMOT-DC versus other solvers. In all four plots, the points tend to concentrate around the line $y = x$, which indicates the comparable performance of MMOT-DC. On the other hand, the top-right plot shows the clear superiority of EGW-PG.

4. The algorithm 1 to solve the MMOT-DC. We tune $\varepsilon \in \{1, 1.4, 1.8, 2.2, 2.6\}$ and we pick the one which corresponds to smallest COOT loss.

For GW-FW and EGW-PG, we use the implementation from the library `PythonOT` [Flamary et al., 2021].

Given two random matrices, we record the COOT loss corresponding to the solution generated by these three methods. We simulate this process 70 times and compare the overall performance of these methods. We can see in Table 1 the average value and standard deviation and the comparison for the values of the loss between the different algorithms in Figure 3. The performance is quite similar across methods with a slight advantage for EGW-PG. This is in itself a very interesting result that has never been noted to the best of our knowledge: the entropic version of GW can provide better solution than solving the exact problem maybe because of the "convexification "of the problem due to the entropic regularization. Our approach is also interestingly better than the exact GW-FW which illustrates that the relaxation might help in finding better solutions despite the non-convexity of the problem.

6 Discussion and conclusion

In this paper, we present a novel relaxation of the factorized MMOT problem called *MMOT-DC*. More precisely, we replace the hard constraint on factorization constraint by a smooth regularization term. The resulting problem not only enjoys an interpolation property between MMOT and factorized

MMOT, but also is a DC problem, which can be solved easily by the DC algorithm. We illustrate the use of MMOT-DC the via some simulated experiments and show that it is competitive with the existing popular solvers of COOT and GW distance. One limitation of the current DC algorithm is that, it is not scalable because it requires storing a full-size tensor in the gradient step computation. Thus, future work may focus on more efficiently designed algorithms, in terms of both time and memory footprint. Moreover, incorporating additional structure on the cost tensor may also be computationally and practically beneficial. From a theoretical viewpoint, it is also interesting to study the extension of MMOT-DC to the continuous setting, which can potentially allow us to further understand the connection between GW distance and COOT.

Acknowledgements. The authors thank to Thibault Séjourné for the fruitful discussion on the connection between GW distance and COOT. The authors thank to Titouan Vayer for pointing out the error in the definition of the lower bound of GW distance.

References

Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein Space. *SIAM Journal on Mathematical Analysis*, 43:904–924, 2011.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman Projections for Regularized Transportation Problems. *SIAM Journal on Scientific Computing*, 37:1111–1138, 2014.

George David Birkhoff. Tres observaciones sobre el algebra lineal. *Universidad Nacional de Tucuman, Revista*, 5:147–150, 1946.

Karsten M. Borgwardt, Arthur Gretton, Malte J. Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J. Smola. Integrating structured biological data by Kernel Maximum Mean Discrepancy. *Bioinformatics*, 22(14):49–57, 7 2006.

Jiezhang Cao, Langyuan Mo, Yifan Zhang, Kui Jia, Chunhua Shen, and Mingkui Tan. Multi-marginal Wasserstein GAN. *Advances in Neural Information Processing Systems*, pages 1774–1784, 2019.

Joel E. Cohen and Uriel G. Rothblum. Nonnegative ranks, decompositions, and factorizations of nonnegative matrices. *Linear Algebra and its Applications*, 190:149–168, 1993.

Samuel Cohen, K. S. Sesh Kumar, and Marc Peter Deisenroth. Sliced multi-marginal optimal transport. *ICML*, 2021.

Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *NeurIPS*, pages 2292–2300, 2013.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL `http://jmlr.org/papers/v22/20-451.html`.

Aden Forrow, Jan-Christian Hütter, Mor Nitzan, Philippe Rigollet, Geoffrey Schiebinger, and Jonathan Weed. Statistical Optimal Transport via Factored Couplings. *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2454–2465, 2019.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

Wilfrid Gangbo and Andrzej Swiech. Optimal maps for the multidimensional Monge-Kantorovich problem. *Communications on Pure and Applied Mathematics*, 51:23–45, 1998.

Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 89:1574–1583, 2019.

Isabel Haasler, Rahul Singh, Qinsheng Zhang, Johan Karlsson, and Yongxin Chen. Multi-marginal optimal transport and probabilistic graphical models. *arXiv preprint arXiv:2006.14113*, 2020.

Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Trans. Image Process.*, 28(11):5464–5478, 2019.

Le Hui, Xiang Li, Jiaxin Chen, Hongliang He, and Jian Yang. Unsupervised multi-domain image translation with domain-specific encoders/decoders. In *24th International Conference on Pattern Recognition, ICPR 2018, Beijing, China, August 20-24, 2018*, pages 2044–2049. IEEE Computer Society, 2018.

Hans G Kellerer. Duality theorems for marginal problems. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 67:399–432, 1984.

Chi-Heng Lin, Mehdi Azabou, and Eva Dyer. Making transport more robust and interpretable by moving data through a small number of anchor points. *Proceedings of the 38th International Conference on Machine Learning*, 139:6631–6641, 2021.

Liang Mi and José Bento. Multi-marginal optimal transport defines a generalized metric. *CoRR*, abs/2001.11114, 2020.

Facundo Mémoli. Gromov-Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, pages 1–71, 2011.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein Averaging of Kernel and Distance Matrices. *International Conference on Machine Learning*, 48, 2016.

Tao Dinh Pham and Hoai An Le Thi. Convex analysis approach to D.C. programming: Theory, Algorithm and Applications. *Acta Mathematica Vietnamica*, 22:289–355, 1997.

Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein Two-Sample Testing and Related Families of Nonparametric Tests. *Entropy*, 19, 2017.

Ievgen Redko, Titouan Vayer, Rémi Flamary, and Nicolas Courty. CO-Optimal Transport. *Advances in Neural Information Processing Systems*, 2020.

Meyer Scetbon, Marco Cuturi, and Gabriel Peyré. Low-Rank Sinkhorn Factorization. *Proceedings of the 38th International Conference on Machine Learning*, 139:9344–9354, 2021.

Bernhard Schmitzer. Stabilized Sparse Scaling Algorithms for Entropy Regularized Transport Problems. *SIAM Journal on Scientific Computing*, 41:1443–1481, 2019.

Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. The Unbalanced Gromov Wasserstein Distance: Conic Formulation and Relaxation. *arXiv preprint arXiv:2009.04266*, 2020.

Pham Dinh Tao and El Bernoussi Souad. Algorithms for Solving a Class of Nonconvex Optimization Problems. Methods of Subgradients. *North-Holland Mathematics Studies*, 129:249–271, 1986.

Thanh Xuan Vo. *Learning with sparsity and uncertainty by Difference of Convex functions optimization*. PhD thesis, Université de Lorraine, 2015.

A Appendix

Derivation of the Sinkhorn algorithm in MMOT. The corresponding entropic dual problem of the primal problem 2 reads

$$\sup_{f_n \in \mathbb{R}^{a_n}} \sum_{n=1}^{N} \langle f_n, \mu_n \rangle - \varepsilon \sum_{i_1,...,i_N} \exp\Big(\frac{\sum_n (f_n)_{i_n} - C_{i_1,...,i_N}}{\varepsilon}\Big) + \varepsilon. \tag{4}$$

For each $n = 1, ..., N$ and $i_n \in [a_n]$, the first order optimality condition reads

$$0 = (\mu_n)_{i_n} - \exp\big(\frac{(f_n)_{i_n}}{\varepsilon}\big) \sum_{i_{-n}} \exp\Big(\frac{\sum_{j \neq n} (f_j)_{i_j} - C_{i_1,...,i_N}}{\varepsilon}\Big),$$

where, with some abuse of notation, we write $i_{-n} = (i_1, ..., i_{n-1}, i_{n+1}, ..., i_N)$. Or, equivalently

$$(f_n)_{i_n} = \varepsilon \log(\mu_n)_{i_n} - \varepsilon \log \sum_{i_{-n}} \exp\Big(\frac{\sum_{j \neq n} (f_j)_{i_j} - C_{i_1,...,i_N}}{\varepsilon}\Big),$$

or even more compact form

$$f_n = \varepsilon \log \mu_n - \varepsilon \log \sum_{i_{-n}} \exp\Big(\frac{\sum_{j \neq n} (f_j)_{i_j} - C_{\cdot, i_{-n}}}{\varepsilon}\Big).$$

Using the primal-dual relation, we obtain the minimiser of the primal problem 2 by

$$P_{i_1,...,i_N} = \exp\Big(\frac{\sum_n (f_n)_{i_n} - C_{i_1,...,i_N}}{\varepsilon}\Big),$$

for $i_n \in [a_n]$, with $n = 1, ..., N$. Similar to the entropic OT, the Sinkhorn algorithm 3 is also usually implemented in log-domain to avoid numerical instability.

Algorithm 3 Sinkhorn algorithm for the entropic MMOT problem 2 from Benamou et al. [2014]

Input. Histograms $\mu_1, ..., \mu_N$, hyperparameter $\varepsilon > 0$, cost tensor C and tuple of initial dual vectors $(f_1^{(0)}, ...f_N^{(0)})$.
Output. Optimal transport plan P and tuple of dual vectors $(f_1, ...f_N)$ (optional).

1. While not converge: for $n = 1, ..., N$,

$$f_n^{(t+1)} = \varepsilon \log \mu_n - \varepsilon \log \sum_{i_{-n}} \Big[\exp\Big(\frac{\sum_{j<n} (f_j^{(t+1)})_{i_j} + \sum_{j>n} (f_j^{(t)})_{i_j} - C_{\cdot, i_{-n}}}{\varepsilon}\Big)\Big].$$

2. Return tensor P, where for $i_n \in [a_n]$, with $n = 1, ..., N$,

$$P_{i_1,...,i_N} = \exp\Big(\frac{\sum_n (f_n)_{i_n} - C_{i_1,...,i_N}}{\varepsilon}\Big). \tag{5}$$

F-MMOT of two components (i.e. $M = 2$) as a variation of low nonnegative rank OT. For the sake of notational ease, we only consider the simplest case, but the same argument holds in the general case. Consider $\mathcal{T}_1 = (1, 2)$ and $\mathcal{T}_2 = (3, 4)$, then the factored MMOT problem reads

$$\text{F-MMOT}\big((\mathcal{T}_m)_{m=1}^2, \mu\big) = \min_{P \in \mathcal{F}_2} \langle C, P \rangle. \tag{6}$$

First, we define three reshaping operations.

- vectorization: concatenates rows of a matrix into a vector.

$$\text{vec} : \mathbb{R}^{m \times n} \to \mathbb{R}^{mn},$$

  where each element $A_{i,j}$ of the matrix $A \in \mathbb{R}^{m \times n}$ is mapped to a unique element $b_{(i-1)n+j}$ of the vector $b \in \mathbb{R}^{mn}$, with $A_{i,j} = b_{(i-1)n+j}$, for $i = 1, ..., m$ and $j = 1, ..., n$. Conversely, each element b_k is mapped to a unique element $A_{k//n, n-k\%n}$, for every $k = 1, ..., mn$. Here, $k//n$ is the quotient of the division of k by n and $k\%n$ is the remainder of this division, i.e. if $k = qn + r$, with $0 \leq r < n$, then $k//n = q$ and $k\%n = r$.

- Matrization: transforms a 4D tensor to a 2D tensor (matrix) by vectorizing the first two and the last two dimensions of the tensor.

$$\text{mat} : \mathbb{R}^{n_1 \times n_2 \times n_3 \times n_4} \to \mathbb{R}^{(n_1 n_2) \times (n_3 n_4)},$$

where, similar to the vectorization, each element $P_{i,j,k,l}$ of the tensor $P \in \mathbb{R}^{n_1 \times n_2 \times n_3 \times n_4}$ is mapped to the unique element $A_{(i-1)n_2+j,(k-1)n_4+l}$ of the matrix $A \in \mathbb{R}^{(n_1 n_2) \times (n_3 n_4)}$, with $P_{i,j,k,l} = A_{(i-1)n_2+j,(k-1)n_4+l}$.
- Concatenation: stacks vertically two equal-column matrices.

$$\begin{aligned} \text{con}_v : &\mathbb{R}^{m \times d} \times \mathbb{R}^{n \times d} \to \mathbb{R}^{(m+n) \times d} \\ &\big((u_1, ..., u_m), (v_1, ..., v_n)\big) \to (u_1, ..., u_m, v_1, ..., v_n)^T. \end{aligned}$$

Or, stacks horizontally two equal-row matrices

$$\begin{aligned} \text{con}_h : &\mathbb{R}^{n \times p} \times \mathbb{R}^{n \times q} \to \mathbb{R}^{n \times (p+q)} \\ &\big((u_1, ..., u_p), (v_1, ..., v_q)\big) \to (u_1, ..., u_p, v_1, ..., v_q). \end{aligned}$$

Lemma A.1 *For any 4D tensor $\pi \in \mathbb{R}^{n_1 \times n_2 \times n_3 \times n_4}$, denote P its matrisation. We have,*

$$vec\Big(\sum_{k,l} \pi_{\cdot,\cdot,k,l}\Big) = \sum_{n=1}^{n_3 n_4} P_{\cdot,n} = P 1_{n_3 n_4},$$

where 1_n is the vector of ones in $\mathbb{R}^n$.

Proof of lemma A.1. For $(i,j) \in [n_1] \times [n_2]$, we have

$$\text{vec}\Big(\sum_{k,l} \pi_{\cdot,\cdot,k,l}\Big)_{(i-1)n_2+j} = \sum_{k,l} \pi_{i,j,k,l} = \sum_{k,l} P_{(i-1)n_2+j,(k-1)n_4+l} = \sum_{n=1}^{n_3 n_4} P_{(i-1)n_2+j,n}.$$

Now, let $(e_i)_{i=1}^{n_1 n_2}$ be the standard basis vectors of $\mathbb{R}^{(n_1 n_2)}$, i.e. $(e_i)_k = 1_{\{i=k\}}$. For each $\pi \in U_{\mathcal{T}}$, denote P its matrix form, then by lemma A.1, we have, for $i \in [n_1]$,

$$(\mu_1)_i = \sum_j \sum_{k,l} \pi_{i,j,k,l} = \sum_{j=1}^{n_2} \sum_{n=1}^{n_3 n_4} P_{(i-1)n_2+j,n},$$

which can be written in matrix form as

$$A_1^T P 1_{n_3 n_4} = \mu_1$$

where the matrix $A_1 = \text{con}_h(v_1, ..., v_{n_1}) \in \mathbb{R}^{(n_1 n_2) \times n_1}$, with $v_i \in \mathbb{R}^{(n_1 n_2)}$, where $v_i = \sum_{j=(i-1)n_2+1}^{in_2} e_j$, with $i \in [n_1]$. Similarly, $A_2 P 1_{n_3 n_4} = \mu_2$, where the matrix $A_2 = \text{con}_h(I_{n_2}, ..., I_{n_2}) \in \mathbb{R}^{n_2 \times (n_1 n_2)}$, where $I_n \in \mathbb{R}^{n \times n}$ is the identity matrix. Both conditions can be compactly written as

$$A_{12}^T P 1_{n_3 n_4} = \mu_{12},$$

where the matrix $A_{12} = \text{con}_h(A_1, A_2^T) \in \mathbb{R}^{(n_1 n_2) \times (n_1 + n_2)}$ and $\mu_{12} = \text{con}_v(\mu_1, \mu_2) \in \mathbb{R}^{(n_1 + n_2)}$. Note that μ_{12} is not a probability because its mass is 2. The matrix A_{12} has exactly $2 n_1 n_2$ ones and the rest are zeros. Similarly, for A_{34} and μ_{34} defined in the same way as A_{12} and μ_{12}, respectively, we establish the equality $A_{34}^T P^T 1_{n_1 n_2} = \mu_{34}$. As a side remark, both matrices A_{12}^T and A_{34}^T are *totally unimodular*, i.e. every square submatrix has determinant $-1, 0$, or 1.

To handle the factorization constraint, first we recall the following concept.

Definition A.1 *Given a nonnegative matrix A, we define its nonnegative rank by*

$$rank_+(A) := \min \Big\{ r \geq 1 : A = \sum_{i=1}^r M_i, \text{ where } rank(M_i) = 1, M_i \geq 0, \forall i \Big\}.$$

By convention, zero matrix has zero (thus nonnegative) rank.

$$\begin{array}{cc|ccc} \multicolumn{2}{c|}{\overleftrightarrow{\ A_1\ }} & \multicolumn{3}{c}{\overleftrightarrow{\quad A_2^T \quad}} \\ \hline 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{array}$$

Figure 4: An example of the matrix A_{12} when $n_1 = 2$ and $n_2 = 3$.

So, the constraint $P = P_1 \otimes P_2$ is equivalent to $\text{mat}(P) = \text{vec}(P_1)\text{vec}(P_2)^T$. By lemma 2.1 in [Cohen and Rothblum, 1993], $\text{rank}_+(A) = 1$ if and only if there exist two nonnegative vectors u, v such that $A = uv^T$. Thus, the factorization constraint can be rewritten as $\text{rank}_+\big(\text{mat}(P)\big) = 1$.

Denote $L = \text{mat}(C)$ and $M = n_1 n_2, N = n_3 n_4$. The problem 6 can be rewritten as

$$\begin{aligned} \min_{Q \in \mathbb{R}_{\geq 0}^{M \times N}} & \langle L, Q \rangle \\ \text{such that } & A_{12}^T Q 1_N = \mu_{12} \\ & A_{34}^T Q^T 1_M = \mu_{34} \\ & \text{rank}_+(Q) = 1. \end{aligned}$$

Proof of proposition 3.1. The inequality $\text{MMOT}(\mu) \leq \text{MMOT-DC}_\varepsilon\big((\mathcal{T}_m)_{m=1}^M, \mu\big)$ follows from the positivity of the KL divergence. On the other hand,

$$\text{F-MMOT}\big((\mathcal{T}_m)_{m=1}^M, \mu\big) = \inf_{P \in \mathcal{F}_M} \langle C, P \rangle + \varepsilon \text{KL}(P | P_\#),$$

because $\text{KL}(P|P_\#) = 0$, for every $P \in \mathcal{F}_M$. As $\mathcal{F}_M \subset U_\mathcal{T}$, we have $\text{MMOT-DC}_\varepsilon\big((\mathcal{T}_m)_{m=1}^M, \mu\big) \leq \text{F-MMOT}\big((\mathcal{T}_m)_{m=1}^M, \mu\big)$.

Now, if $\text{F-MMOT}\big((\mathcal{T}_m)_{m=1}^M, \mu\big) = 0$, then $\text{MMOT-DC}_\varepsilon\big((\mathcal{T}_m)_{m=1}^M, \mu\big) = 0$. Conversely, if $\text{MMOT-DC}_\varepsilon\big((\mathcal{T}_m)_{m=1}^M, \mu\big) = 0$, for $\varepsilon > 0$, then there exists $P^* \in U_\mathcal{T}$ such that $\langle C, P^* \rangle = 0$ and $P^* = P^*_\#$. Thus $\langle C, P^*_\# \rangle = 0$, which means $\text{F-MMOT}\big((\mathcal{T}_m)_{m=1}^M, \mu\big) = 0$.

Proof of proposition 3.2. The function $\varepsilon \to \text{MMOT-DC}_\varepsilon\big((\mathcal{T}_m)_{m=1}^M, \mu\big)$ is increasing on $\mathbb{R}_{\geq 0}$ and bounded, thus admits a finite limit $L \leq \text{F-MMOT}\big((\mathcal{T}_m)_{m=1}^M, \mu\big)$, when $\varepsilon \to \infty$, and a finite limit $l \geq \text{MMOT}(\mu)$, when $\varepsilon \to 0$.

Let P_ε be a solution of the problem $\text{MMOT-DC}_\varepsilon\big((\mathcal{T}_m)_{m=1}^M, \mu\big)$. As $U_\mathcal{T}$ is compact, when either $\varepsilon \to 0$ or $\varepsilon \to \infty$, one can extract a converging subsequence (after reindexing) $(P_{\varepsilon_k})_k \to \widetilde{P} \in U_\mathcal{T}$, when either $\varepsilon_k \to 0$ or $\varepsilon_k \to \infty$. Thus, the convergence of the marginal distributions is also guaranteed, i.e $(P_{\varepsilon_k})_{\#m} \to \widetilde{P}_{\#m} \in U_{\mathcal{T}_m}$, for every $m = 1, ..., M$, which implies that $P_{\varepsilon_k} - (P_{\varepsilon_k})_{\#\mathcal{T}} \to \widetilde{P} - \widetilde{P}_{\#\mathcal{T}}$.

When $\varepsilon \to 0$, let P^* be a solution of the problem $\text{MMOT}(\mu)$. Then,

$$\langle C, P^* \rangle \leq \langle C, P_\varepsilon \rangle + \varepsilon \text{KL}(P_\varepsilon | (P_\varepsilon)_{\#\mathcal{T}}) \leq \langle C, P^* \rangle + \varepsilon \text{KL}(P^* | P^*_{\#\mathcal{T}}).$$

By the sandwich theorem, when $\varepsilon \to 0$, we have $\text{MMOT-DC}_\varepsilon\big((\mathcal{T}_m)_{m=1}^M, \mu\big) \to \langle C, P^* \rangle = \text{MMOT}(\mu)$. Furthermore, as

$$0 \leq \langle C, P_{\varepsilon_k} \rangle - \langle C, P^* \rangle \leq \varepsilon_k \text{KL}(P^* | P^*_{\#\mathcal{T}}),$$

when $\varepsilon_k \to 0$, it follows that $\langle C, \widetilde{P} \rangle = \langle C, P^* \rangle$. So $\widetilde{P}$ is a solution of the problem $\text{MMOT}(\mu)$. We conclude that any cluster point of the sequence of minimisers of $\text{MMOT-DC}_\varepsilon\big((\mathcal{T}_m)_{m=1}^M, \mu\big)$ when

$\varepsilon \to 0$ is a minimiser of MMOT(μ). As a byproduct, since

$$\text{KL}(P^*|P^*_{\#\mathcal{T}}) - \text{KL}(P_{\varepsilon_k}|(P_{\varepsilon_k})_{\#\mathcal{T}}) \geq \frac{\langle C, P_{\varepsilon_k}\rangle - \langle C, P^*\rangle}{\varepsilon_k} \geq 0,$$

we deduce that $\text{KL}(\widetilde{P}|\widetilde{P}_{\#\mathcal{T}}) \leq \text{KL}(P^*|P^*_{\#\mathcal{T}})$ (so the cluster point $\widetilde{P}$ has minimal mutual information).

On the other hand, when $\varepsilon \to \infty$, for $\mu = \mu_1 \otimes ... \otimes \mu_N$, one has

$$\langle C, \mu\rangle + \varepsilon \times 0 \geq \langle C, P_\varepsilon\rangle + \varepsilon\text{KL}(P_\varepsilon|(P_\varepsilon)_{\#\mathcal{T}}) \geq \varepsilon\text{KL}(P_\varepsilon|(P_\varepsilon)_{\#\mathcal{T}}).$$

Thus,

$$0 \leq \text{KL}(P_\varepsilon|(P_\varepsilon)_{\#\mathcal{T}}) \leq \frac{1}{\varepsilon}\langle C, \mu\rangle \to 0, \text{ when } \varepsilon \to \infty,$$

which means $\text{KL}(P_\varepsilon|(P_\varepsilon)_{\#\mathcal{T}}) \to 0$, when $\varepsilon \to \infty$. In particular, when $\varepsilon_k \to \infty$, we have $\text{KL}(P_{\varepsilon_k}|(P_{\varepsilon_k})_{\#\mathcal{T}}) \to 0$. Thus, $\text{KL}(\widetilde{P}|\widetilde{P}_{\#\mathcal{T}}) = 0$, which implies $\widetilde{P} = \widetilde{P}_{\#\mathcal{T}}$.

Now, as $\text{MMOT-DC}_\varepsilon\big((\mathcal{T}_m)_{m=1}^M, \mu\big) \geq \langle C, P_\varepsilon\rangle$, taking limit when $\varepsilon \to \infty$ gives us $L \geq \langle C, \widetilde{P}\rangle = \langle C, \widetilde{P}_{\#\mathcal{T}}\rangle \geq \text{F-MMOT}\big((\mathcal{T}_m)_{m=1}^M, \mu\big)$. Thus $L = \langle C, \widetilde{P}\rangle = \text{F-MMOT}\big((\mathcal{T}_m)_{m=1}^M, \mu\big)$, i.e. $\text{MMOT-DC}_\varepsilon\big((\mathcal{T}_m)_{m=1}^M, \mu\big) \to \text{F-MMOT}\big((\mathcal{T}_m)_{m=1}^M, \mu\big)$ when $\varepsilon \to \infty$. We also have that any cluster point of the sequence of minimisers of $\text{MMOT-DC}_\varepsilon\big((\mathcal{T}_m)_{m=1}^M, \mu\big)$ when $\varepsilon \to \infty$ is a minimiser of $\text{F-MMOT}\big((\mathcal{T}_m)_{m=1}^M, \mu\big)$.

Proof of corollary 3.3. In this proof, we write $C := L_p(C_x, C_y)$, for notational ease. In the setting of GW distance, we have $N = 4, M = 2$ and $\mathcal{T}_1 = (1, 2), \mathcal{T}_2 = (3, 4)$. Given a solution P_ε of the problem MMOT-DC, let $Q_i \in U\big((P_\varepsilon)_{\#i}, (P_\varepsilon)_{\#i}\big) \subset U_{\mathcal{T}}$, for $i = 1, 2$. The optimality of P_ε implies that

$$\langle C, P_\varepsilon\rangle + \varepsilon\Big[H(P_\varepsilon) - H\big((P_\varepsilon)_{\#1}\big) - H\big((P_\varepsilon)_{\#2}\big)\Big] \leq \langle C, Q_i\rangle + \varepsilon\Big[H(Q_i) - 2H\big((P_\varepsilon)_{\#i}\big)\Big].$$

Thus,

$$2\big(\langle C, P_\varepsilon\rangle + \varepsilon H(P_\varepsilon)\big) \leq \sum_{i=1}^2 \langle C, Q_i\rangle + \varepsilon H(Q_i).$$

As this is true for every $Q_i \in U\big((P_\varepsilon)_{\#i}, (P_\varepsilon)_{\#i}\big)$, we have

$$\begin{aligned}
\frac{1}{2}\sum_{i=1}^2 \text{OT}_\varepsilon\big((P_\varepsilon)_{\#i}, (P_\varepsilon)_{\#i}\big) &= \frac{1}{2}\sum_{i=1}^2 \inf_{Q_i \in U\big((P_\varepsilon)_{\#i}, (P_\varepsilon)_{\#i}\big)} \langle C, Q_i\rangle + \varepsilon H(Q_i) \\
&\geq \langle C, P_\varepsilon\rangle + \varepsilon H(P_\varepsilon) \\
&\geq \inf_{P \in U\big((P_\varepsilon)_{\#1}, (P_\varepsilon)_{\#2}\big)} \langle C, P\rangle + \varepsilon H(P) \\
&= \text{OT}_\varepsilon\big((P_\varepsilon)_{\#1}, (P_\varepsilon)_{\#2}\big).
\end{aligned}$$

The second inequality holds because $P_\varepsilon \in U\big((P_\varepsilon)_{\#1}, (P_\varepsilon)_{\#2}\big)$. Thus,

$$\text{OT}_\varepsilon\big((P_\varepsilon)_{\#1}, (P_\varepsilon)_{\#2}\big) - \frac{1}{2}\sum_{i=1}^2 \text{OT}_\varepsilon\big((P_\varepsilon)_{\#i}, (P_\varepsilon)_{\#i}\big) \leq 0. \tag{7}$$

The left-hand side of 7 is nothing but the Sinkhorn divergence SD_ε between $(P_\varepsilon)_{\#1}$ and $(P_\varepsilon)_{\#2}$ [Ramdas et al., 2017]. As a strictly positive definite kernel is necessarily universal in the finite setting (see for example section 2.3 in [Borgwardt et al., 2006]), by theorem 1 in [Feydy et al., 2019], the inequality 7 is in fact an equality and we must have $(P_\varepsilon)_{\#1} = (P_\varepsilon)_{\#2}$.

Now, by proposition 3.2, when $\varepsilon \to \infty$, a cluster point $P_{\#1} \otimes P_{\#2}$ of the sequence of minimisers $(P_\varepsilon)_\varepsilon$ induces a solution $(P^{(1)}, P^{(2)})$ of the COOT. The above result implies that $P_{\#1} = P_{\#2}$ and the equality between GW and COOT then follows.

An empirical variation. Intuitively, for sufficiently large ε, the minimisation of the KL divergence is prioritised over the linear term in the objective function of the MMOT-DC problem, which implies that the optimal tensor P^* is "close" to its corresponding tensor product $P^*_{\#\mathcal{T}}$. So, instead of calculating the gradient at P, one may calculate at $P_{\#\mathcal{T}}$. In this case, the gradient reads

$$\sum_{m=1}^{M} \nabla_P H_m(P_{\#\mathcal{T}}) = \big[\log P_{\#1} + P_{\#1}\big] \oplus ... \oplus \big[\log P_{\#M} + P_{\#M}\big],$$

where $\oplus$ represents the tensor sum operator between two arbitrary-size tensors: $(P \oplus Q)_{i,j} := P_i + Q_j$, where with some abuse of notation, i or j can be understood as a tuple of indices. Thus, we avoid storing the N-D gradient tensor (as in the algorithm 1) and only need to store M smaller-size tensors. Not only saving the memory, this variation also seems to be empirically competitive with the original algorithm 1, if not sometimes better, in terms of COOT loss. The underlying reason might be related to the approximate DCA scheme [Vo, 2015], where one replaces both steps in each DC iteration by their approximation. We leave the formal theoretical justification of this variation to the future work. We call this variation *MMOT-DC-v1* and use the same setup as in the experiment 5.

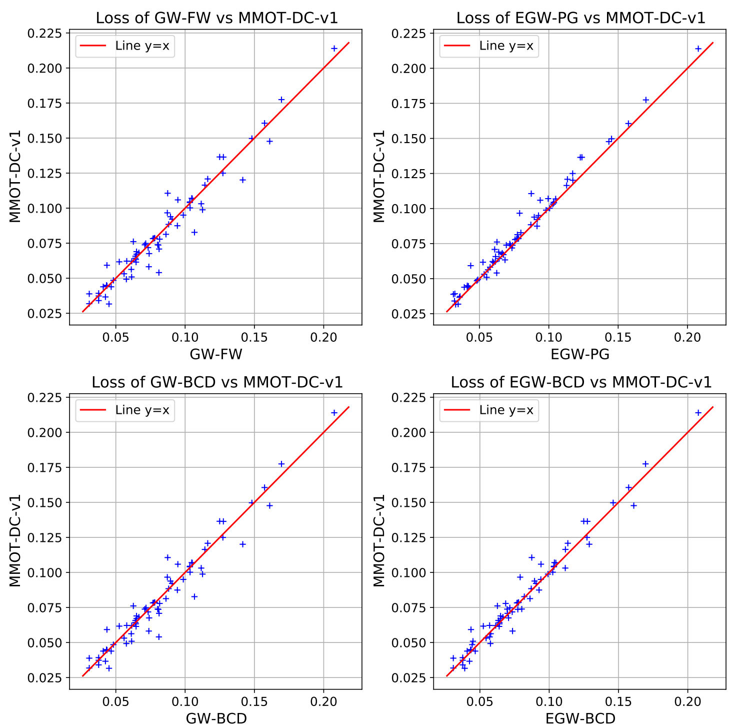

Figure 5: Scatter plots of MMOT-DC-v1 versus other solvers. In all four plots, the points tend to concentrate around the line $y = x$, which indicates the comparable performance of MMOT-DC-v1. On the other hand, the top-right plot shows the clear superiority of EGW-PG.

MMOT-DC	MMOT-DC-v1
0.0822 ($\pm$ 0.0364)	0.0820 ($\pm$ 0.0361)

Table 2: Average and standard deviation of COOT loss of MMOT-DC and MMOT-DC-v1. The performance of the two algorithms is very similar.