LOW-RANK MULTIMARGINAL OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimarginal Optimal Transport (MOT) suffers from exponential computational and memory costs in the number of marginals, as transport plans are represented by high-dimensional tensors. We introduce Low-Rank Multimarginal Optimal Transport (LRMOT), which restricts transport plans to tensors with bounded multilinear rank, providing a compact and structured representation of multimarginal couplings. This structure enables efficient algorithms with a low memory footprint. We establish approximation bounds for the resulting low-rank formulations and demonstrate experimentally that the proposed algorithms substantially reduce computational time while achieving solutions with low transport cost.

1 INTRODUCTION

Modelling strongly correlated particles systems in the presence of disorder (Byczuk et al., 2010) and multi-omic data integration (Hawinkel et al., 2020) are two instances where sets of distributions must be geometrically averaged. Multimarginal Optimal Transport (MOT) (Gangbo & Swiech, 1998; Pass, 2015) formalizes this problem, aligning a set of d probability distributions with respect to an efficiency cost. In the discrete case, MOT is a linear program between a cost tensor $\boldsymbol{\mathcal{C}}$ and a transport tensor $\boldsymbol{\mathcal{T}}$ coupling the distributions. As d increases, this problem becomes intractable. Although entropic regularization enables MOT to be solved via a sequence of scaling operations with the Sinkhorn algorithm (Benamou et al., 2015), the coupling $\boldsymbol{\mathcal{T}}$ itself remains a dense tensor of size n^d. This limitation is particularly problematic when MOT is used repeatedly, for example in machine learning problems such as GANs (see Cao et al. (2019)).

One way to alleviate the curse of dimensionality is to reduce the problem's density by imposing structure on the cost tensor $\boldsymbol{\mathcal{C}}$. When $\boldsymbol{\mathcal{C}}$ is tree-structured, Haasler et al. (2021a) exploit this structure to decouple the associated Gibbs kernel $\boldsymbol{\mathcal{K}}$. Strössner & Kressner (2023) combine this approach with low-rank approximations of $\boldsymbol{\mathcal{K}}$, further improving the problem's tractability. Haasler et al. (2021b) show that MOT is equivalent to a Bayesian marginal inference problem when $\boldsymbol{\mathcal{C}}$ has a graphical structure and certain conditions on the marginal are met. Ba & Quellmalz (2022) further exploit tree- or a circle- structured costs using FFT-based methods.

A second way to overcome this bottleneck is to impose structure on the transport tensor itself, as in the classic literature on low-rank (bipartite) optimal transport (Scetbon et al., 2021). While this line of work has been extended to yield linear-space algorithms for Wasserstein, Gromov-Wasserstein, and fused Gromov-Wasserstein objectives (Forrow et al., 2019; Lin et al., 2021; Scetbon et al., 2022; Halmos et al., 2024; Jawanpuria & Mishra, 2026), low-rank structure has not yet been imposed directly on the transport tensor in the multimarginal setting. To the best of our knowledge, the closest work is Tran et al. (2021), where the authors penalize the full tensor $\boldsymbol{\mathcal{T}}$ to be closer to a rank-$(1, \dots, 1)$ coupling.

Low-rank tensor structure encompasses a wide range of settings, most of which can be directly linked to tensor decompositions (Kolda & Bader, 2009). As shown in Figure 1, the barycenter view of MOT can be expressed directly through Candecomp/Parafac (CP) (Carroll & Chang, 1970; Harshman, 1970) decomposition (left), while Tucker decomposition (Tucker, 1966) offers a more general setting in which each distribution is synthesised into a smaller support (right), enabling a smaller MOT to be solved instead. Block-Term decomposition (De Lathauwer, 2008) provides an intermediary structure (center), where mass is transported through selected hubs. Other decompositions such as Tensor Trains (Oseledets, 2011) or Tensor Ring (Zhao et al., 2016) connect different tensors through their various unfoldings and lead to other factorization of the transport tensor.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

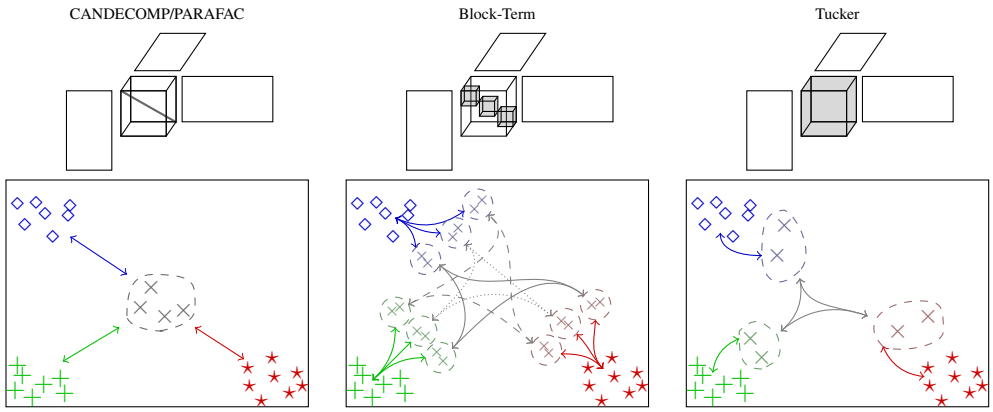


Figure 1: **Low-rank MOT through different tensor decompositions**. CP decomposition (left) aligns all distributions to one latent measure. Tucker decomposition (right) reduces the MOT problem through d latent measures. Block-Term decomposition provides an intermediary setting (center).

In this work, we define Low-Rank Multimarginal Optimal Transport (LRMOT) by restricting the admissible transport tensors to tensors having an upper bounded multilinear rank. We show that LRMOT is well defined and state a bound on the approximation error with respect to MOT. We instantiate different algorithms related to different decompositions and derive an algorithm taking advantage of a class of structured costs. Their performance is demonstrated through several experiments. The contributions of this paper can be summarized as follows:

- We define Low-Rank Multimarginal Optimal Transport (Section 3).
- We incorporate cost structure by formulating an "explicit" CPMOT, provide a bound relating on the optima of the original CPMOT and the explicit CPMOT (Section 4).
- We present three solvers for the instantiated low-rank problems CPMOT and TuckerMOT, based on tensor decompositions (Section 5).
- We show that our algorithms perform well on multi-marginal problems (Section 6).

2 Discrete Multimarginal Optimal Transport

2.1 Tensor operations and notations

For each integer $d \geq 1$, we denote by $[\![d]\!]$ the set $\{1, \dots, d\}$. Given an integer $d \geq 1$ and any positive integers $n_1, \dots, n_d$, we call $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ a $d-$mode tensor. For any two tensors $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ and $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{m_1 \times \cdots \times m_e}$, we define the algebraic operations *Frobenius Inner Product* (when $d = e$) and *Outer Product* $\circ$ by

$$\langle \boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{T}} \rangle \stackrel{\text{def}}{=} \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} \boldsymbol{\mathcal{S}}_{i_1 \dots i_d} \boldsymbol{\mathcal{T}}_{i_1 \dots i_d}, \quad (\boldsymbol{\mathcal{S}} \circ \boldsymbol{\mathcal{T}})_{i_1 \dots i_d j_1 \dots j_e} \stackrel{\text{def}}{=} \boldsymbol{\mathcal{S}}_{i_1 \dots i_d} \boldsymbol{\mathcal{T}}_{j_1 \dots j_e}.$$

We denote by $\otimes$ the Kronecker product between two matrices. Given a subset of indices $\mathcal{I} \subseteq [\![d]\!]$ and its complement $\bar{\mathcal{I}}$, any tensor can be unfolded into a matrix $\mathbf{T}^{(\mathcal{I})}$ with $n_{\mathcal{I}} = \prod_{i \in \mathcal{I}} n_i$ rows and $n_{\bar{\mathcal{I}}} = \prod_{j \in \bar{\mathcal{I}}} n_j$ columns whose elements are ordered according to a chosen lexicographical order of the modes. In particular, the matricization of $\boldsymbol{\mathcal{T}}$ with respect to mode k is a matrix denoted $\mathbf{T}^{(k)}$ with n_k rows and $\bar{n}_k = \prod_{\ell \in [\![d]\!] \setminus \{k\}} n_\ell$ columns. From these unfoldings, numerous tensor-matrix products can be defined. For instance, the product $\boldsymbol{\mathcal{T}} \times_{\mathcal{I}} \mathbf{M}$ refolds the result of the matrix product $\mathbf{M}\mathbf{T}^{(\mathcal{I})}$. In particular, the mode-k product, denoted by $\times_k$, corresponds to the matrix multiplication $\mathbf{M}\mathbf{T}^{(k)}$. The multilinear rank of a tensor $\boldsymbol{\mathcal{T}}$ is $(R_1, \dots, R_d)$ where R_k is the (column) rank of $\mathbf{T}^{(k)}$.

A tensor is a probability tensor if its entries are nonnegative and the sum of all entries is 1. In that case, we denote $\boldsymbol{\mathcal{T}} \in \Delta_{n_1 \times \cdots \times n_d}$. The Shannon entropy $\mathcal{H}(\boldsymbol{\mathcal{T}})$ is defined by $\mathcal{H}(\boldsymbol{\mathcal{T}}) = -\langle \boldsymbol{\mathcal{T}}, \log(\boldsymbol{\mathcal{T}}) - \mathbf{1} \rangle$, where the logarithm operator is element-wise, with the convention that $0 \log 0 =$

0. The marginal of $\boldsymbol{\mathcal{T}}$ according to the mode k denoted $\mathrm{Marg}_k(\boldsymbol{\mathcal{T}})$ is given by the vector of row sums of $\mathbf{T}^{(k)}$. More generally, the marginal of $\boldsymbol{\mathcal{T}}$ according to the subset of modes $\mathcal{I}$ denoted $\mathrm{Marg}_{\mathcal{I}}(\boldsymbol{\mathcal{T}})$ is given by the vector of row sums of $\mathbf{T}^{(\mathcal{I})}$.

2.2 Discrete Multimarginal Optimal Transport

Given a collection $\mathbf{a}$ of d discrete probability measures $a^{(1)}, \ldots, a^{(d)} \in \mathcal{P}(\mathcal{X})$ with $n_1, \ldots, n_d$ atoms respectively, a transport tensor $\boldsymbol{\mathcal{T}}$ associated to $\mathbf{a}$ is a probability tensor whose mode-k marginals satisfy $\mathrm{Marg}_k(\boldsymbol{\mathcal{T}}) = a^{(k)}$ for all $k \in [\![d]\!]$. The set of all such transport tensors is defined as:

$$\Pi(\mathbf{a}) \stackrel{\text{def}}{=} \{\boldsymbol{\mathcal{T}} \in \mathbb{R}_+^{n_1 \times \cdots \times n_d} \mid \forall k \in [\![d]\!],\ \mathrm{Marg}_k(\boldsymbol{\mathcal{T}}) = a^{(k)}\}.$$

Consequently, each mode-k matricization of $\boldsymbol{\mathcal{T}}$ implied by these marginal constraints can be interpreted as a coupling between $a^{(k)}$ and a joint measure $b^{(k)}$ supported on the product space $\prod_{\ell \in [\![d]\!] \setminus \{k\}} \mathrm{supp}(a^{(\ell)})$.

Given a cost tensor $\boldsymbol{\mathcal{C}} \in \mathbb{R}_+^{n_1 \times \cdots \times n_d}$, the discrete Multimarginal Optimal Transport problem (Gangbo & Swiech, 1998; Pass, 2015) is the multimarginal generalization of discrete optimal transport, and minimizes the total transport cost $\langle \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{T}} \rangle$ over the set of valid transport tensors associated to $\mathbf{a}$:

$$\mathrm{MOT}(\mathbf{a}, \boldsymbol{\mathcal{C}}) = \min_{\boldsymbol{\mathcal{T}} \in \Pi(\mathbf{a})} \langle \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{T}} \rangle. \tag{1}$$

Intuitively, MOT can be viewed as solving simultaneously d coupled, relaxed optimal transport problems. For each $k \in [\![d]\!]$, the k-th problem uses the mode-k matricization $\mathbf{C}^{(k)}$ as its cost matrix, with $a^{(k)}$ acting as the row (left) marginal. However, these d sub-problems are fundamentally interdependent: their respective transport plans are not free to vary independently, as they must form valid, structurally consistent matricizations of the same underlying tensor $\boldsymbol{\mathcal{T}}$. Finally, when $d = 2$, MOT exactly recovers the standard two-marginal (bipartite) optimal transport problem.

Since problem (1) consists of minimizing a continuous linear objective function over a compact convex polytope $\Pi(\mathbf{a})$, the existence of a global minimum is guaranteed. Despite this guarantee, the optimal solution is generally not unique due to the linear nature of the objective function over the polytope. To overcome this lack of uniqueness and mitigate computational challenges, a standard approach is to introduce entropic regularization (Benamou et al., 2015). The entropically regularized multimarginal optimal transport problem with regularization parameter $\varepsilon > 0$ is formulated as:

$$\mathrm{MOT}_\varepsilon(\mathbf{a}, \boldsymbol{\mathcal{C}}) = \min_{\boldsymbol{\mathcal{T}} \in \Pi(\mathbf{a})} \langle \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{T}} \rangle - \varepsilon \mathcal{H}(\boldsymbol{\mathcal{T}}). \tag{2}$$

Regularized MOT leads to algorithms based on Bregman projections, the Sinkhorn MOT algorithm (Benamou et al., 2015) (see Algorithm 2) and its greedy and accelerated versions (Lin et al., 2022). The first-order Karush-Kuhn-Tucker optimality conditions reveal that the unique solution $\boldsymbol{\mathcal{T}}^\varepsilon$ of Eq. (2) exhibits a factorized Gibbs-like structure. It can be expressed as a scaling of the kernel tensor $\boldsymbol{\mathcal{K}} = \exp(-\boldsymbol{\mathcal{C}}/\varepsilon)$ (Peyré, 2026):

$$\boldsymbol{\mathcal{T}}^\varepsilon_{i_1 \ldots i_d} = \boldsymbol{\mathcal{K}}_{i_1 \ldots i_d} \big(\textstyle\prod_{k=1}^d u^{(k)}_{i_k} \big),$$

where $u^{(k)} \in \mathbb{R}_+^{n_k}$ for $k \in [\![d]\!]$ are scaling vectors (or scaling factors), which are unique up to a positive multiplicative constant. The Sinkhorn algorithm needs the full tensor $\boldsymbol{\mathcal{K}}$, which cannot be stored as the size of the problem grows. This calls for the use of structured costs that reduce the memory footprint, as proposed in Haasler et al. (2021a), Haasler et al. (2021b), Ba & Quellmalz (2022) and Strössner & Kressner (2023).

3 Low-Rank Multimarginal Optimal Transport

Proofs of the stated properties are deferred to Appendix C.

3.1 Low-Rank Transport Tensors

Contrary to the aforementioned works relying on the structure of $\boldsymbol{\mathcal{C}}$, the low-rank formulations proposed by the literature in the standard two-marginal case enforce low-rank constraints on the transport plan and change the space of the problem, redefining the feasible set. In the multi-marginal

case, there exist numerous ways to enforce such a low-rank constraint as the rank of a tensor itself has several definitions (Kolda & Bader, 2009). Indeed, as the number of measures d increases, the number of available tensor unfoldings grows exponentially. Each of these unfolded tensors can then be constrained via a low-rank assumption.

To this end, take $\boldsymbol{\mathcal{I}} = (\mathcal{I}_m)_{m=1}^M$ a collection of M subsets of $[\![d]\!]$ and associated nonnegative ranks $\mathbf{r} = (r_m)_{m=1}^M$ (Cohen & Rothblum, 1993). The nonnegative rank satisfies $\operatorname{rank}(A) \leq \operatorname{rank}_+(A)$ for any nonnegative matrix A. Define $N_m = \prod_{j \in \mathcal{I}_m} n_j$. We define the set $\Pi_{\mathbf{r},\boldsymbol{\mathcal{I}}}(\mathbf{a}) \subseteq \Pi(\mathbf{a})$ of admissible couplings $\boldsymbol{\mathcal{T}}$ satisfying low-rank assumptions according to $\boldsymbol{\mathcal{I}}$ and $\mathbf{r}$ as

$$\Pi_{\mathbf{r},\boldsymbol{\mathcal{I}}}(\mathbf{a}) = \Pi(\mathbf{a}) \cap \{\boldsymbol{\mathcal{T}} \in \mathbb{R}_+^{n_1 \times \cdots \times n_d} \mid \operatorname{rank}_+(\mathbf{T}^{(\mathcal{I}_m)}) \leq r_m \text{ for } 1 \leq m \leq M\}.$$

Property 3.1. *The set $\Pi_{\mathbf{r},\boldsymbol{\mathcal{I}}}(\mathbf{a})$ is not empty.*

When $\boldsymbol{\mathcal{I}}$ is a general collections of indices, the rank constraints and the marginals constraints may not be aligned. To mitigate this issue, we impose that the elements of $\boldsymbol{\mathcal{I}}$ are *pairwise disjoint*. Without loss of generality, we can in fact impose that $\boldsymbol{\mathcal{I}}$ is a *partition* of $[\![d]\!]$ by adding every mode k not represented in $\boldsymbol{\mathcal{I}}$ as a singleton $\{k\}$ and setting the corresponding rank r_m to n_k, so that no rank restriction is imposed on these modes.

The following property shows that solving the MOT problem on a collection of product measures $\alpha^{(m)}$ induced by the $a^{(k)}$'s gives a valid transport tensor for $\Pi_{\mathbf{r},\boldsymbol{\mathcal{I}}}(\mathbf{a})$.

Property 3.2. *Let $\boldsymbol{\mathcal{I}}$ be a partition of $[\![d]\!]$. Let $\alpha^{(m)} = \otimes_{j \in \mathcal{I}_m} a^{(j)}$ be the product measure of the $a^{(j)}$'s associated to $\mathcal{I}_m$. If $\boldsymbol{\mathcal{T}} \in \mathbb{R}_+^{N_1 \times \cdots \times N_M}$ belongs to $\Pi(\boldsymbol{\alpha})$ and satisfies $\operatorname{rank}_+(\mathbf{T}^{(m)}) \leq r_m$ for any $1 \leq m \leq M$, then we can reshape $\boldsymbol{\mathcal{T}}$ into a tensor $\widetilde{\boldsymbol{\mathcal{T}}} \in \mathbb{R}_+^{n_1 \times \cdots \times n_d}$ belonging to $\Pi_{\mathbf{r},\boldsymbol{\mathcal{I}}}(\mathbf{a})$.*

Given a partition $\boldsymbol{\mathcal{I}}$, we can thus consider a new problem on the set of singletons $\boldsymbol{\mathcal{J}} = \{\{m\}\}_{m=1}^M$ by taking the product measures as marginals. In the following, we describe the case $\boldsymbol{\mathcal{I}} = \{\{k\}\}_{k=1}^d$, and write $\Pi_{\mathbf{r},\boldsymbol{\mathcal{I}}}(\mathbf{a}) = \Pi_{\mathbf{r}}(\mathbf{a})$ accordingly. In order to derive numerical methods solving low-rank problems, we enforce tensor decompositions structure to the transport tensor in the following subsections.

3.2 Tucker decomposition

In this section, we restrict the set of admissible couplings $\Pi_{\mathbf{r}}(\mathbf{a})$ to its subset of tensors that admit a nonnegative Tucker decomposition (see Alexandrov et al. (2022) for characterizations). The tensor $\boldsymbol{\mathcal{T}}$ admits a nonnegative Tucker decomposition with multilinear ranks $\mathbf{r}$ if there exist $\boldsymbol{\mathcal{G}} \in \mathbb{R}_+^{r_1 \times \cdots \times r_d}$ and $P^{(k)} \in \mathbb{R}_+^{n_k \times r_k}$ such that

$$\boldsymbol{\mathcal{T}} = \boldsymbol{\mathcal{G}} \times_1 P^{(1)} \times_2 \cdots \times_d P^{(d)}. \tag{3}$$

The tensor $\boldsymbol{\mathcal{T}}$ can then be stored using $\prod_{k=1}^d r_k + \sum_{k=1}^d n_k r_k$ entries. Such representation can always be made probabilistic by normalizing each $P^{(k)}$ to be a column-stochastic matrix and absorbing the normalization constants into $\boldsymbol{\mathcal{G}}$. We thus consider the following subset of $\Pi_{\mathbf{r}}(\mathbf{a})$, with $\mathcal{S}_{\text{col}}^{n_k \times r_k}$ the set of column-stochastic matrices of size $n_k \times r_k$:

$$\Pi_{\mathbf{r}}^{\mathrm{T}}(\mathbf{a}) = \{\boldsymbol{\mathcal{T}} \in \Pi(\mathbf{a}) \mid \exists \boldsymbol{\mathcal{G}} \in \Delta_{r_1 \times \cdots \times r_d}, \exists P^{(k)} \in \mathcal{S}_{\text{col}}^{n_k \times r_k}, \boldsymbol{\mathcal{T}} = \boldsymbol{\mathcal{G}} \times_1 P^{(1)} \cdots \times_d P^{(d)}\}.$$

A particular case of this general decomposition structure appears in Tran et al. (2021), where the authors consider a partition $\boldsymbol{\mathcal{I}}$ of $[\![d]\!]$ and introduce the tensor $\widetilde{\boldsymbol{\mathcal{T}}} = \operatorname{Marg}_{\mathcal{I}_1}(\boldsymbol{\mathcal{T}}) \circ \cdots \circ \operatorname{Marg}_{\mathcal{I}_M}(\boldsymbol{\mathcal{T}})$. This outer product has multilinear rank $(1, \ldots, 1)$ with respect to the partition $\boldsymbol{\mathcal{I}}$, since each block factor $\operatorname{Marg}_{\mathcal{I}_m}(\boldsymbol{\mathcal{T}})$ spans a one-dimensional subspace.[1]

Property 3.3. *The set $\Pi_{\mathbf{r}}^{\mathrm{T}}(\mathbf{a})$ is compact and not empty.*

We can thus define the Tucker MOT problem as

$$\operatorname{TuckerMOT}_{\mathbf{r}}(\mathbf{a}, \boldsymbol{\mathcal{C}}) = \min_{\boldsymbol{\mathcal{T}} \in \Pi_{\mathbf{r}}^{\mathrm{T}}(\mathbf{a})} \langle \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{T}} \rangle. \tag{4}$$

We recover the original MOT problem by setting $r_k = n_k$ for every k so that $\Pi(\mathbf{a}) \subseteq \Pi_{\mathbf{r}}^{\mathrm{T}}(\mathbf{a})$. The following property states a bound on the approximation error of TuckerMOT.

[1] Property 3.3 ensures that the infimum in the definition of F-MMOT in Tran et al. (2021) is attained.

Property 3.4 (Approximation error)**.** *Let R be the minimum number of nonnegative rank one factors*[2] *necessary to construct the optimal transport tensor $\mathcal{T}^*$ of MOT. For $2 \leq r \leq R$, and for any choice of $\mathbf{r} = (r_k)_{k=1}^d$ such that $r_{\min} = r$, we have*

$$0 \leq \mathrm{TuckerMOT}_{\mathbf{r}}(\mathbf{a}, \mathcal{C}) - \mathrm{MOT}(\mathbf{a}, \mathcal{C}) \leq \|\mathcal{C}\|_\infty \ln\Big(\frac{R}{r-1}\Big).$$

To prove Property 3.4, we in fact establish the bound in the CP case (detailed below), which is more restrictive. Since the quantity R is difficult to compute, the following corollary provides a looser but more readily computable bound:

Corollary 3.5. *It is known that MOT always has a sparse solution with at most $\sum_{k=1}^d n_k - d + 1$ entries (Anderes et al. (2015), Lemma 3). Thus,*

$$0 \leq \mathrm{TuckerMOT}_{\mathbf{r}}(\mathbf{a}, \mathcal{C}) - \mathrm{MOT}(\mathbf{a}, \mathcal{C}) \leq \|\mathcal{C}\|_\infty \ln\Big(\frac{\sum_{k=1}^d n_k - d + 1}{r-1}\Big).$$

Introducing bipartite OT for mode k. Under the Tucker decomposition (3), each measure $a^{(k)}$ can be seen as being transported on a latent measure $g^{(k)} = \mathrm{Marg}_k(\mathcal{G})$. Indeed, any matricization $\mathbf{T}^{(k)}$ can be rewritten as a product of $P^{(k)}$ (coupling the measure $a^{(k)}$ with $g^{(k)}$), of the matricization $\mathbf{G}^{(k)}$ of the probability core tensor $\mathcal{G}$, and of the Kronecker product of the coupling matrices associated to the other modes (Kolda & Bader, 2009). The column stochasticity of the matrices $P^{(k)}$ implies that if $\mathcal{T} \in \Pi(\mathbf{a})$,

$$a^{(k)} = \mathrm{Marg}_k(\mathcal{T}) = P^{(k)}\mathbf{G}^{(k)}\big(\bigotimes_{\ell \neq k} P^{(\ell)}\big)^\mathsf{T}\mathbf{1}_{\bar{n}_k} = P^{(k)}\mathbf{G}^{(k)}\mathbf{1}_{\bar{r}_k} = P^{(k)}\mathrm{Marg}_k(\mathcal{G}) = P^{(k)}g^{(k)}.$$

As $P^{(k)}$ is not a transport plan, we define $Q^{(k)} = P^{(k)}\mathrm{diag}(g^{(k)})$. Remark that $Q^{(k)}\mathbf{1}_{r_k} = a^{(k)}$ and that $(Q^{(k)})^\mathsf{T}\mathbf{1}_{\bar{n}_k} = \mathrm{diag}(g^{(k)})(P^{(k)})^\mathsf{T}\mathbf{1}_{\bar{n}_k} = g^{(k)}$ as $P^{(k)}$ is column-stochastic. Thus, $Q^{(k)}$ can be interpreted as a transport plan between the measure $a^{(k)}$ and the latent measure $g^{(k)}$.

Conversely, the joint measure $b^{(k)}$ is coupled to $\mathcal{G}$ through the matrix $(\bigotimes_{\ell \neq k} P^{(\ell)})$. Denoting $h^{(k)} = (\mathbf{G}^{(k)})^\mathsf{T}\mathbf{1}_{\bar{r}_k}$ the latent joint measure, we can define $R^{(k)} = (\bigotimes_{\ell \neq k} P^{(\ell)})\mathrm{diag}(h^{(k)})$. The matrix $R^{(k)}$ is a transport plan from $h^{(k)}$ to $b^{(k)}$. This yields the hierarchical sequence of Figure 2. We now present another classical tensor decomposition of $\mathcal{T}$ more restrictive than the Tucker decomposition.

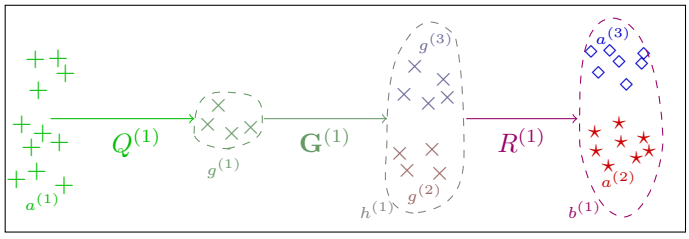


Figure 2: Transport from $a^{(1)}$ to $b^{(1)}$ through intermediary latent measures.

3.3 CP DECOMPOSITION

CP decomposition. CP decomposition imposes that the core tensor $\mathcal{G}$ in the Tucker decomposition is hyperdiagonal (see Figure 1, left panel). As a consequence, $r_1 = \cdots = r_d = r$, and the core tensor can be represented by a probability vector $g \in \Delta_r$. The tensor $\mathcal{T}$ can then be expressed as a sum of rank-1 tensors formed by columns of $P^{(k)}$, with $p_s^{(k)}$ denoting the s-th column of $P^{(k)}$:

$$\mathcal{T} = \sum_{s=1}^r g_s p_s^{(1)} \circ \cdots \circ p_s^{(d)}.$$

The tensor $\mathcal{T}$ can then be stored using $r + r\sum_{k=1}^d n_k$ entries.

[2] $R = \mathrm{rank}_{\mathrm{CP}+}(\mathcal{T}^\star)$ where $\mathrm{rank}_{\mathrm{CP}+}$ is defined in Lim & Comon (2009)

We define the CP formulation of the MOT problem as the minimization of the transport cost subject to the constraint that the nonnegative CP rank of the transport tensor $\boldsymbol{\mathcal{T}}$ is upper bounded. Define the feasible set of CP-rank-r admissible transport tensors as

$$\Pi_r^{\mathrm{CP}}(\mathbf{a}) \stackrel{\text{def}}{=} \Pi(\mathbf{a}) \cap \{\boldsymbol{\mathcal{T}} \in \mathbb{R}_+^{n_1 \times \cdots \times n_d} \mid \mathrm{rank}_{\mathrm{CP}+}(\boldsymbol{\mathcal{T}}) \leq r\}.$$

Property 3.6. *The set $\Pi_r^{\mathrm{CP}}(\mathbf{a})$ is compact and non empty.*

The CPMOT problem then reads

$$\mathrm{CPMOT}_r(\mathbf{a}, \boldsymbol{\mathcal{C}}) = \min_{\boldsymbol{\mathcal{T}} \in \Pi_r^{\mathrm{CP}}(\mathbf{a})} \langle \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{T}} \rangle. \tag{5}$$

Introducing bipartite OT for mode k. As in the Tucker formulation, each measure $a^{(k)}$ is mapped to the latent measure $g^{(k)}$ through a transport plan $Q^{(k)} = P^{(k)}\mathrm{diag}(g^{(k)})$. However as $\boldsymbol{\mathcal{G}}$ is hyperdiagonal, each marginal $\mathrm{Marg}_k(\boldsymbol{\mathcal{G}})$ is represented by the same vector $g \in \Delta_r$. Thus, there is only one common latent measure g. As a consequence, the transport plan associated to mode k is given by $Q^{(k)} = P^{(k)}\mathrm{diag}(g)$. In a similar way, the latent joint measure $h^{(k)}$ of the CP core over all latent modes except k reduces to a measure $\bar{g}^{(k)}$ whose only non-null entries are associated to the r elements of the hyperdiagonal of $\boldsymbol{\mathcal{G}}$. Consequently, $R^{(k)} = (\bigotimes_{\ell \neq k} P^{(\ell)})\mathrm{diag}(\bar{g}^{(k)})$. When computing $R^{(k)}$ only r rows of the Kronecker product are necessary. Finally, we can represent the transport from g to $b^{(k)}$ by a single $r \times \bar{n}_k$ matrix.

Regularization. We consider a regularized version of problem (5). Let $h(\mathbf{Q}) = \sum_k \mathcal{H}(Q^{(k)})$. Taking the infimum over $g \in \Delta_r^*$ and $Q^{(k)} \in \Pi(a^{(k)}, g)$, the regularized problem reads

$$\mathrm{CPMOT}_{r,\varepsilon}(\mathbf{a}, \boldsymbol{\mathcal{C}}) = \inf_{\mathbf{Q}, g} \langle \boldsymbol{\mathcal{C}}, \textstyle\sum_{s=1}^r g_s p_s^{(1)} \circ \cdots \circ p_s^{(d)} \rangle - \varepsilon h(\mathbf{Q}), \tag{6}$$

4 Incorporating cost structure

In this Section, we describe a class of cost tensors having a complete graph structure that allows us to define an *explicit* CPMOT (eCPMOT), where the support of the latent measure g is instantiated. This formulation yields an upper bound for CPMOT.

4.1 Explicit CPMOT

Conditionally of negative type costs. Following Houry et al. (2026), we define costs that are conditionally of negative type (CNT). Such a cost $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is CNT if it is continuous, symmetric, and verifies $c(x, x) = 0$ for all $x \in \mathcal{X}$, and for any $n \in \mathbb{N}$, any $x_1, \ldots, x_n \in \mathcal{X}$, any $\lambda_1, \ldots, \lambda_n$ with $\sum_{i=1}^n \lambda_i = 0$ we have $\sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j c(x_i, x_j) \leq 0$. Equivalently, a cost c is CNT if there exists a real Hilbert space $\mathcal{H}$ and a continuous mapping $\varphi : \mathcal{X} \to \mathcal{H}$ verifying $c(x, x') = \|\varphi(x) - \varphi(x')\|_{\mathcal{H}}^2$ for all $x, x' \in \mathcal{X}$.

Pairwise costs. Let $a^{(1)}, \ldots, a^{(d)}$ be d measures on $\mathcal{P}(\mathcal{X})$ with atoms $(x_{i_k}^{(k)})_{i_k=1}^{n_k}$ respectively. Let $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be a CNT cost, with its associated Hilbert space $\mathcal{H}$ and mapping φ. We consider pairwise costs of the form

$$\boldsymbol{\mathcal{C}}(x_{i_1}^{(1)}, \ldots, x_{i_d}^{(d)}) = \sum_{k<l} \|\varphi(x_{i_k}^{(k)}) - \varphi(x_{i_l}^{(l)})\|_{\mathcal{H}}^2.$$

Let us denote by $z = (z_s)_{s=1}^r$ the support of the latent measure g. Denote by $B^{(k)}(\varphi(z)) = [\|\varphi(x_{i_k}^{(k)}) - \varphi(z_s)\|_{\mathcal{H}}^2]_{i_k,s} \in \mathbb{R}_+^{n_k \times s}$ the cost matrix between $a^{(k)}$ and g. Introducing z in each cost pair yields an upper bound for CPMOT.

Theorem 4.1. *The following problem is an upper bound of CPMOT.*

$$\mathrm{eCPMOT}_r(\mathbf{a}, \boldsymbol{\mathcal{C}}) = d \inf_{\mathbf{Q}, g, \varphi(z)} \sum_{k=1}^d \langle B^{(k)}, Q^{(k)} \rangle, \tag{7}$$

where the minimization is taken over $g \in \Delta_r^\star$, $\varphi(z) \in \mathcal{H}^r$, and each $Q^{(k)} \in \Pi(a^{(k)}, g)$.

In Theorem 4.1, the vector g is required to have non zero entries. This does not pose a problem for numerical methods, since $\mathrm{CPMOT}_r(\mathbf{a}, \boldsymbol{\mathcal{C}}) \leq \mathrm{CPMOT}_{r-1}(\mathbf{a}, \boldsymbol{\mathcal{C}})$.

Adding regularization yields the following upper bound of $\mathrm{CPMOT}_{r,\varepsilon}$ (Eq. (6)):

$$\mathrm{eCPMOT}_{r,\varepsilon}(\mathbf{a}, \boldsymbol{\mathcal{C}}) = d \inf_{\mathbf{Q}, g, \varphi(z)} \sum_{k=1}^{d} \langle C^{(k)}, Q^{(k)} \rangle - \frac{\varepsilon}{d} \mathcal{H}(Q^{(k)}). \tag{8}$$

4.2 Optimality gap induced by eCPMOT

For a low-rank transport plan decomposition $(\mathbf{Q}, g)$, we denote by $F(\boldsymbol{\mathcal{T}})$ the objective from eCPMOT (Eq. (7)). Let $A_s^{(k)}$ be independent random variables of respective law $A_s^{(k)} \sim \sum_{i_k=1}^{n_k} p_{i_k s}^{(k)} \delta_{\varphi(x_{i_k}^{(k)})}$, where $p_{i_k s}^{(k)} = Q_{i_k s}^{(k)} / g_s$.

Theorem 4.2. *Denote by $\mu_s^{(k)} = \mathbb{E}[A_s^{(k)}]$. Let us denote*

$$V(\boldsymbol{\mathcal{T}}) = \sum_{s=1}^{r} g_s \sum_{k=1}^{d} \mathrm{Var}(A_s^{(k)}), \qquad B(\boldsymbol{\mathcal{T}}) = \frac{1}{2} \sum_{k \neq l} \|\mu_s^{(k)} - \mu_s^{(l)}\|_{\mathcal{H}}^2.$$

Then, the different objectives of Eq. (5) *and Eq.* (7) *can be expressed as*

$$\langle \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{T}} \rangle = (d-1) V(\boldsymbol{\mathcal{T}}) + B(\boldsymbol{\mathcal{T}}), \qquad F(\boldsymbol{\mathcal{T}}) = d V(\boldsymbol{\mathcal{T}}) + B(\boldsymbol{\mathcal{T}}).$$

Empirical interpretation. The transport from $a^{(k)}$ to g induces r measures $(a_s^{(k)})_{s=1}^r$ sharing the support of $a^{(k)}$ such that $a^{(k)} = \sum_{s=1}^r g_s a_s^{(k)}$. The quantity $V(\boldsymbol{\mathcal{T}})$ measures the within-component dispersion of each measure $a_s^{(k)}$ around its mean $\mu_s^{(k)}$. We refer to Figure 7 in Appendix B.2 for an example showing positions of the $\mu^{(k)}$'s.

Corollary 4.3. *If $\boldsymbol{\mathcal{T}}^\star$ is an optimal transport plan for CPMOT (Eq.* (5)*) and $\widehat{\boldsymbol{\mathcal{T}}}$ is an optimal transport plan for eCPMOT (Eq.* (7)*), then*

$$0 \leq \langle \boldsymbol{\mathcal{C}}, \widehat{\boldsymbol{\mathcal{T}}} \rangle - \langle \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{T}}^\star \rangle \leq V(\boldsymbol{\mathcal{T}}^\star) - V(\widehat{\boldsymbol{\mathcal{T}}}).$$

Hence, the reduction in within-component dispersion resulting from the explicit barycentric optimization provides an upper bound on the excess transport cost of $\widehat{\boldsymbol{\mathcal{T}}}$. In particular, if $V(\boldsymbol{\mathcal{T}}^\star) = V(\widehat{\boldsymbol{\mathcal{T}}})$, then $\widehat{\boldsymbol{\mathcal{T}}}$ is also optimal for the original low-rank problem.

5 Algorithms

In this section, we derive two algorithms for solving Eq. (5) and one algorithm for solving Eq. (4). The first two algorithms presented below are designed for a general cost tensor $\boldsymbol{\mathcal{C}}$, and operations involving this tensor should be adapted to take advantage of any structure in $\boldsymbol{\mathcal{C}}$. In our experiments, we use the class of structured costs introduced in Section 4. When $d = 2$, we recover the methods proposed in Halmos et al. (2024), Scetbon et al. (2021) and Forrow et al. (2019); see Appendix B.2 for details. When using the low-rank solvers together with the pairwise cost structure, the per-iteration cost of the Sinkhorn MOT algorithm, $\mathcal{O}(d \prod_{k=1}^d n_k)$, is reduced to $\mathcal{O}(r \sum_{k=1}^d n_k)$ in the CP case and to $\mathcal{O}(r \sum_{k=1}^d n_k + dr^d)$ in the Tucker case.

5.1 Mirror Descent scheme optimizing the Tucker formulation

For each mode k, the matrix $Q^{(k)}$ is a transport plan from the marginal $a^{(k)}$ to the latent marginal $g^{(k)}$. At the latent level, the core tensor $\boldsymbol{\mathcal{G}}$ couples the latent marginals $g^{(1)}, \ldots, g^{(d)}$. Hence, the Tucker representation decomposes the MOT problem into a first level of d semi-relaxed transports $a^{(k)} \to g^{(k)}$ followed by a latent MOT problem. Costs associated to these sub-problems are derived from the following equalities

$$\langle \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{G}} \times_{k=1}^d P^{(k)} \rangle = \langle \boldsymbol{\mathcal{C}} \times_{k=1}^d (P^{(k)})^\mathsf{T}, \boldsymbol{\mathcal{G}} \rangle = \langle [\boldsymbol{\mathcal{C}} \times_{\ell \neq k} (P^{(\ell)})^\mathsf{T} \times_{\ell \neq k} \boldsymbol{\mathcal{G}}] \mathrm{diag}(g^{(k)})^{-1}, Q^{(k)} \rangle.$$

Thus, constructing the mode-k subcost directly from this expression requires access to the full cost tensor $\boldsymbol{\mathcal{C}}$. Denote by $\boldsymbol{\mathcal{C}}_{\boldsymbol{\mathcal{G}}}$ and $C_{Q^{(k)}}$ the cost tensors associated with the latent problem and the mode-k semi-relaxed transport problem, respectively. Since $\boldsymbol{\mathcal{C}}_{\boldsymbol{\mathcal{G}}} = C_{Q^{(k)}}\mathrm{diag}(g^{(k)}) \times_k (P^{(k)})^\mathsf{T}$, we can approximate $C_{Q^{(k)}}$ without explicitly forming $\boldsymbol{\mathcal{C}}$ by approximately inverting $(P^{(k)})^\mathsf{T}$. In particular, using the Moore-Penrose pseudoinverse A^+ of a matrix A (Barata & Hussein, 2011), we can consider $C_{Q^{(k)}} \approx [\boldsymbol{\mathcal{C}}_{\boldsymbol{\mathcal{G}}} \times_k ((P^{(k)})^\mathsf{T})^+]\mathrm{diag}(g^{(k)})^{-1}$.

Scetbon et al. (2021) explain how the classical regularized OT problem can be seen as a mirror descent step on the unregularized problem, justifying mirror descent schemes in low-rank approaches. After adding a regularization term $\mathcal{H}(\boldsymbol{\mathcal{G}})$ to TuckerMOT (Eq. (4)), we follow these steps an derive an algorithm (TUCKER) using a mirror descent scheme in Appendix E.

5.2 Mirror Descent scheme optimizing the CP formulation

We also perform a mirror descent in the CP formulation. In this case, we add entropic regularization on both $\mathbf{Q}$ and g. The algorithm CPMOT for Eq. (5) is derived in Appendix F.

5.3 Alternating minimization of the explicit CP formulation

To solve Eq. (8), we use an alternating minimization scheme, successively minimizing over $(\mathbf{Q}, g)$ and over $\varphi(z)$.

Property 5.1. *When the $Q^{(k)}$'s and g are fixed, the minimizing $\varphi(z)$ is defined by*

$$\varphi(z_s^\star) = \frac{1}{dg_s}\sum_k \sum_{i_k} Q^{(k)}_{i_k s}\varphi(x^{(k)}_{i_k}).$$

Thus, the arithmetic mean of the latent hubs minimizes $\varphi(z)$.

Minimization over $(\mathbf{Q}, g)$. When the support $\varphi(z)$ of the intermediate measure is fixed, Benamou et al. (2015) provide the SinkhornBary Algorithm 8 to find the minimizing weights of g and couplings $Q^{(k)}$.

This alternating minimization scheme results in Algorithm 1 (ECPMOT). In Appendix G.2, we present an extension of this method to the unbalanced setting (Beier et al., 2022).

Algorithm 1 Explicit CPMOT

function ECPMOT(r, $(x^{(k)})_{k=1}^d$, $\mathbf{a}$, ε)
  $\varphi(z) \leftarrow$ D-KMEANS-Z($(\varphi(x^{(k)}))_k, r$)
  $[C^{(k)}]_{k=1}^d \leftarrow \left[[\|\varphi(x^{(k)}_{i_k}) - \varphi(z_s)\|^2_{\mathcal{H}}]_{i_k s}\right]_{k=1}^d$
  for $t \geq 0$ **do** ▷ until convergence
    $\mathbf{Q}, g \leftarrow$ SinkhornBary($(C^{(k)})_{k=1}^d, \mathbf{a}, \varepsilon/d$)
    $\varphi(z) \leftarrow [\frac{1}{dg_s}\sum_{k=1}^d Q^{(k)}_{i_k s}\varphi(x^{(k)}_{i_k s})]_{s=1}^r$
    $[C^{(k)}]_{k=1}^d \leftarrow \left[[\|\varphi(x^{(k)}_{i_k}) - \varphi(z_s)\|^2_{\mathcal{H}}]_{i_k s}\right]_{k=1}^d$
  end for
  return $\mathbf{Q}, g$

6 Experiments

Time gain and optimality. In Table 1, we compare the methods derived in Section 5 with the standards and greedy Sinkhorn MOT algorithms in terms of transport cost and running time. We consider that convergence has been reached when the relative change in the objective is below 10^{-6}. We sample $(n_1, n_2, n_3, n_4) = (25, 50, 25, 50)$ points in $\mathbb{R}^2$ from four Gaussian mixtures (detailed in Appendix B.1). We set $\varepsilon = 0.005$ for the CP formulation, and $\varepsilon = 0.04$ for the Tucker formulation. For TUCKER, we report results using both Sinkhorn (2) (MSK) and Greedy Sinkhorn (3) (G-MSK) for the core subproblem. The cost tensors are normalized.

Table 1: Comparison of the transport costs and running times of the methods and Sinkhorn MOT. 2D Gaussian mixtures with $(n_1, n_2, n_3, n_4) = (25, 50, 25, 50)$. Values are averaged over 30 runs

r	Method	Transport cost	Marg. err	Iterations	Time (s)
	MSK	0.186 ± 0.022	10^{-8}	246 ± 62	20.3 ± 6.3
	G-MSK	0.186 ± 0.022	10^{-7}	580 ± 122	8.2 ± 1.9
2	ECPMOT	$\mathbf{0.229 \pm 0.022}$	10^{-8}	7 ± 3	$\mathbf{0.2 \pm 0.1}$
	CPMOT	0.258 ± 0.028	10^{-8}	2007 ± 690	2.8 ± 1.9
	TUCKER G-MSK	0.232 ± 0.022	10^{-17}	532 ± 415	60.2 ± 0.1
	TUCKER MSK	0.233 ± 0.022	10^{-17}	492 ± 377	60.2 ± 0.3
4	ECPMOT	$\mathbf{0.202 \pm 0.022}$	10^{-8}	20 ± 9	$\mathbf{0.4 \pm 0.2}$
	CPMOT	0.231 ± 0.028	10^{-8}	2379 ± 568	4.4 ± 2.2
	TUCKER G-MSK	0.231 ± 0.022	10^{-17}	1434 ± 786	59.6 ± 1.8
	TUCKER MSK	0.231 ± 0.022	10^{-17}	1677 ± 970	56.9 ± 6.6
10	ECPMOT	$\mathbf{0.191 \pm 0.022}$	10^{-8}	56 ± 36	$\mathbf{0.7 \pm 0.4}$
	CPMOT	0.219 ± 0.023	10^{-8}	2511 ± 432	5.1 ± 1.7
	TUCKER G-MSK	0.235 ± 0.022	10^{-17}	759 ± 219	60.3 ± 0.4
	TUCKER MSK	0.234 ± 0.022	10^{-17}	875 ± 295	60.1 ± 0.1

We observe that the running time is reduced with the CP formulation, particularly for the explicit method. Moreover, the low-rank methods achieve transport costs close to those obtained with Sinkhorn while maintaining low marginal errors, even for a very small rank such as $r = 4$. The cost decreases as r increases, except for TUCKER, for which there was insufficient time to perform as many iterations in the $r = 10$ case. Overall, Greedy Sinkhorn performs better than Sinkhorn. However, for a certain range of small r^d values in the Tucker core subproblem, the standard Sinkhorn routine performs better, allowing Tucker to perform more iterations within the allotted time.

Latent barycenters. In Figure 3, we solve TuckerMOT for a problem involving $d = 3$ measures, where the k-th measure $a^{(k)}$ is supported on r_k blobs. In the Tucker formulation, each measure $a^{(k)}$ is associated with a latent measure $g^{(k)}$, whose support points can be recovered as $\mu_s^{(k)} = \sum_{i_k=1}^{n_k} Q_{i_k s}^{(k)} x_{i_k}^{(k)} / g_s^{(k)}$. We observe that the resulting latent measures accurately represent the $a^{(k)}$'s.

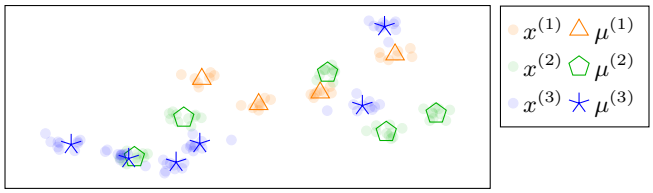


Figure 3: Supports of the $a^{(k)}$'s and $g^{(k)}$'s. $(n_1, n_2, n_3) = (30, 50, 60)$, $(r_1, r_2, r_3) = (4, 5, 6)$

Additionnal experiments are found in Appendix B.

7 Conclusion

We defined a low-rank multimarginal optimal transport problem and showed that it encompasses several formulations found in the literature. We instantiated this framework with several algorithms. In particular, the ECPMOT algorithm finds transport tensors with a low transport cost two orders of magnitude faster than the usual Sinkhorn algorithm. A natural extension of this work would be to generalize the algorithms to a broader class of structured costs. From a theoretical viewpoint, it would also be interesting to sharpen the approximation bound.

AI use statement In this work, we used generative AI tools to check language and grammar, and to give wording suggestions.

Reproducibility statement Source code for the methods and experiments is given in supplementary material.

REFERENCES

Boian Alexandrov, Derek DeSantis, Gianmarco Manzini, and Erik Skau. Nonnegative canonical tensor decomposition with linear constraints: nncandelinc, 2022. URL https://arxiv.org/abs/1909.07570.

Ethan Anderes, Steffen Borgwardt, and Jacob Miller. Discrete wasserstein barycenters: Optimal transport for discrete data, 2015. URL https://arxiv.org/abs/1507.07218.

David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. volume 8, pp. 1027–1035, 01 2007. doi: 10.1145/1283383.1283494.

Fatima Antarou Ba and Michael Quellmalz. Accelerating the sinkhorn algorithm for sparse multimarginal optimal transport via fast fourier transforms. Algorithms, 15(9):311, August 2022. ISSN 1999-4893. doi: 10.3390/a15090311. URL http://dx.doi.org/10.3390/a15090311.

João Carlos Alves Barata and Mahir Saleh Hussein. The moore–penrose pseudoinverse: A tutorial review of the theory. Brazilian Journal of Physics, 42(1-2):146–165, December 2011. ISSN 1678-4448. doi: 10.1007/s13538-011-0052-z. URL http://dx.doi.org/10.1007/s13538-011-0052-z.

Florian Beier, Johannes von Lindheim, Sebastian Neumayer, and Gabriele Steidl. Unbalanced multimarginal optimal transport. Journal of Mathematical Imaging and Vision, 65(3):394–413, October 2022. ISSN 1573-7683. doi: 10.1007/s10851-022-01126-7. URL http://dx.doi.org/10.1007/s10851-022-01126-7.

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(2):A1111–A1138, 2015. doi: 10.1137/141000439. URL https://doi.org/10.1137/141000439.

Nicolas Boumal. An introduction to optimization on smooth manifolds. Cambridge University Press, 2023. doi: 10.1017/9781009166164. URL https://www.nicolasboumal.net/book.

Krzysztof Byczuk, Walter Hofstetter, and Dieter Vollhardt. Anderson localization vs. mott–hubbard metal–insulator transition in disordered, interacting lattice fermion systems. International Journal of Modern Physics B, 24(12n13):1727–1755, May 2010. ISSN 1793-6578. doi: 10.1142/s0217979210064575. URL http://dx.doi.org/10.1142/S0217979210064575.

Jiezhang Cao, Langyuan Mo, Yifan Zhang, Kui Jia, Chunhua Shen, and Mingkui Tan. Multimarginal wasserstein gan. In Advances in Neural Information Processing Systems, 2019.

J. Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. Psychometrika, 35(3):283–319, Sep 1970. ISSN 1860-0980. doi: 10.1007/BF02310791. URL https://doi.org/10.1007/BF02310791.

Joel E. Cohen and Uriel G. Rothblum. Nonnegative ranks, decompositions, and factorizations of nonnegative matrices. Linear Algebra and its Applications, 190:149–168, 1993. ISSN 0024-3795. doi: https://doi.org/10.1016/0024-3795(93)90224-C. URL https://www.sciencedirect.com/science/article/pii/002437959390224C.

Lieven De Lathauwer. Decompositions of a higher-order tensor in block terms—part ii: Definitions and uniqueness. SIAM Journal on Matrix Analysis and Applications, 30(3):1033–1066, 2008. doi: 10.1137/070690729. URL https://doi.org/10.1137/070690729.

Aden Forrow, Jan-Christian Hütter, Mor Nitzan, Philippe Rigollet, Geoffrey Schiebinger, and Jonathan Weed. Statistical optimal transport via factored couplings. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, volume 89 of Proceedings of Machine Learning Research, pp. 2454–2465. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/forrow19a.html.

Wilfrid Gangbo and Andrzej Swiech. Optimal maps for the multidimensional monge-kantorovich problem. Communications on Pure and Applied Mathematics, 51(1):23–45, 1998. doi: https://doi.org/10.1002/(SICI)1097-0312(199801)51:1⟨23::AID-CPA2⟩3.0.CO;2-H. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0312%28199801%2951%3A1%3C23%3A%3AAID-CPA2%3E3.0.CO%3B2-H.

Isabel Haasler, Axel Ringh, Yongxin Chen, and Johan Karlsson. Multimarginal optimal transport with a tree-structured cost and the schrödinger bridge problem. SIAM Journal on Control and Optimization, 59(4):2428–2453, 2021a. doi: 10.1137/20M1320195. URL https://doi.org/10.1137/20M1320195.

Isabel Haasler, Rahul Singh, Qinsheng Zhang, Johan Karlsson, and Yongxin Chen. Multi-marginal optimal transport and probabilistic graphical models. IEEE Transactions on Information Theory, 67:4647–4668, 07 2021b. doi: 10.1109/TIT.2021.3077465.

Peter Halmos, Xinhao Liu, Julian Gold, and Benjamin J Raphael. Low-Rank Optimal Transport through Factor Relaxation with Latent Coupling. 2024.

R. A. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. UCLA Working Papers in Phonetics, 16:1–84, 1970.

Stijn Hawinkel, Luc Bijnens, Kim-Anh Lê Cao, and Olivier Thas. Model-based joint visualization of multiple compositional omics datasets. NAR Genomics and Bioinformatics, 2(3):lqaa050, 09 2020. ISSN 2631-9268. doi: 10.1093/nargab/lqaa050. URL https://doi.org/10.1093/nargab/lqaa050.

Guillaume Houry, Jean Feydy, and François-Xavier Vialard. Gromov-wasserstein at scale, beyond squared norms, 2026. URL https://arxiv.org/abs/2602.06658.

Pratik Jawanpuria and Bamdev Mishra. A Riemannian Approach to Low-Rank Optimal Transport, June 2026. URL http://arxiv.org/abs/2606.12120. arXiv:2606.12120 [cs.LG].

Tamara G. Kolda and Brett W. Bader. Tensor Decompositions and Applications. SIAM Review, 51 (3):455–500, January 2009. doi: 10.1137/07070111X.

Lek-Heng Lim and Pierre Comon. Nonnegative approximations of nonnegative tensors. Journal of Chemometrics, 23:432–441, 06 2009. doi: 10.1002/cem.1244.

Chi-Heng Lin, Mehdi Azabou, and Eva L. Dyer. Making transport more robust and interpretable by moving data through a small number of anchor points, July 2021. URL http://arxiv.org/abs/2012.11589. arXiv:2012.11589 [cs.LG].

Tianyi Lin, Nhat Ho, Marco Cuturi, and Michael I. Jordan. On the complexity of approximating multimarginal optimal transport, 2022. URL https://arxiv.org/abs/1910.00152.

I. V. Oseledets. Tensor-train decomposition. SIAM Journal on Scientific Computing, 33(5):2295–2317, 2011. doi: 10.1137/090752286. URL https://doi.org/10.1137/090752286.

Brendan Pass. Multi-marginal optimal transport: theory and applications. ESAIM: Mathematical Modelling and Numerical Analysis, 49(6):1771–1790, 2015.

Gabriel Peyré. Optimal Transport for Machine Learners, August 2026. URL http://arxiv.org/abs/2505.06589. arXiv:2505.06589 [stat.ML].

Meyer Scetbon and Marco Cuturi. Low-rank optimal transport: Approximation, statistics and debiasing. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), Advances in Neural Information Processing Systems, volume 35, pp. 6802–6814. Curran Associates, Inc., 2022. doi: 10.52202/068431-0493. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/2d69e771d9f274f7c624198ea74f5b98-Paper-Conference.pdf.

Meyer Scetbon, Marco Cuturi, and Gabriel Peyré. Low-Rank Sinkhorn Factorization. 2021.

Meyer Scetbon, Gabriel Peyre, and Marco Cuturi. Linear-Time Gromov Wasserstein Distances using Low Rank Couplings and Costs. 2022.

Henri Schmidt, Peter Halmos, and Benjamin Raphael. Transport clustering: Solving low-rank optimal transport via clustering, 2026. URL https://openreview.net/forum?id=YKTJJCNXF4.

Christoph Strössner and Daniel Kressner. Low-rank tensor approximations for solving multi-marginal optimal transport problems. SIAM Journal on Imaging Sciences, 16(1):169–191, 2023. doi: 10.1137/22M1478355. URL https://doi.org/10.1137/22M1478355.

Quang Huy Tran, Hicham Janati, Ievgen Redko, Rémi Flamary, and Nicolas Courty. Factored couplings in multi-marginal optimal transport via difference of convex programming, 2021. URL https://arxiv.org/abs/2110.00629.

Ledyard R. Tucker. Some mathematical notes on three-mode factor analysis. Psychometrika, 31(3): 279–311, Sep 1966. ISSN 1860-0980. doi: 10.1007/BF02289464. URL https://doi.org/10.1007/BF02289464.

Qibin Zhao, Guoxu Zhou, Shengli Xie, Liqing Zhang, and Andrzej Cichocki. Tensor ring decomposition, 2016. URL https://arxiv.org/abs/1606.05535.

APPENDICES

A Notations **13**

B Experiments **14**

B.1 More experiments and details 14

B.2 Experiments in the $d = 2$ case 16

C Proof details **19**

C.1 Proof of Property 3.1 19

C.2 Proof of Property 3.2 19

C.3 Proof of Property 3.3 19

C.4 Proof of Property 3.4 20

C.5 Proof of Property 3.6 21

C.6 Proof of Theorem 4.1 21

C.7 Proof of Theorem 4.2 22

C.8 Proof of Property 5.1 23

D Sinkhorn MOT **24**

D.1 Classical Sinkhorn MOT 24

D.2 Greedy and accelerated Sinkhorn MOT 24

E Mirror Descent for TuckerMOT **25**

F Mirror Descent for CPMOT **27**

G Alternating minimization for eCPMOT **29**

G.1 Scheme convergence 30

G.2 Explicit Unbalanced CP Multi-Marginal Optimal Transport 31

H Classical bipartite low-rank optimal transport **31**

A NOTATIONS

- d the number of measures in a MOT problem,
- $\boldsymbol{\mathcal{T}} \in \mathbb{R}_+^{n_1 \times \cdots \times n_d}$ a transport tensor,
- $a^{(k)} \in \mathcal{P}(\mathcal{X})$ the k-th measure in a MOT problem,
- n_k the size of the support of $a^{(k)}$,
- $b^{(k)}$ the product measure $(\mathbf{T}^{(k)})^\mathsf{T}\mathbf{1}_{n_k}$,
- $\bar{n}_k = \prod_{\ell \neq k} n_\ell$ the size of the support of $b^{(k)}$,
- $(x^{(k)})_{i_k=1}^{n_k} \in \mathcal{X}^{n_k}$ the support of $a^{(k)}$,
- r_k the rank constraint associated with the k-th measure,
- $\Pi(\mathbf{a})$ the set of admissible transport tensors with marginals $\mathbf{a} = (a^{(k)})_{k=1}^d$,

- $\boldsymbol{\mathcal{C}} \in \mathbb{R}_+^{n_1 \times \cdots \times n_d}$ a cost tensor,
- $\mathcal{H}$ the entropy function $\mathcal{H}(X) = -\langle X, \log(X) - \mathbf{1} \rangle$,
- $\varepsilon > 0$ a regularization parameter,
- $\boldsymbol{\mathcal{I}} = (\mathcal{I}_m)_{m=1}^M$ a collection of subsets of $[\![d]\!]$,
- $N_m = \prod_{j \in \mathcal{I}_m} n_j$,
- $\alpha^{(m)} = \otimes_{j \in \mathcal{I}_m} a^{(j)}$,
- $\boldsymbol{\mathcal{G}} \in \mathbb{R}_+^{r_1, \ldots, r_d}$ a Tucker core,
- $\Delta_{n_1, \ldots, n_d}$ the probability tensors of size $n_1 \times \cdots \times n_d$,
- $g^{(k)} = \mathrm{Marg}_k(\boldsymbol{\mathcal{G}}) \in \mathbb{R}_+^{r_k}$ a latent measure,
- $P^{(k)} \in \mathcal{S}_{\text{col}}^{n_k \times r_k}$ a column-stochastic matrix,
- $Q^{(k)} = P^{(k)} \mathrm{diag}(g^{(k)}) \in \mathbb{R}_+^{n_k \times r_k}$ a transport plan from $a^{(k)}$ to $g^{(k)}$, $\mathbf{Q} = (Q^{(k)})_{k=1}^d$,
- $\varphi : \mathcal{X} \to \mathcal{H}$ a mapping for the CNT cost c,
- $\mu^{(k)}$ hubs associated with $x^{(k)}$,
- $z \in \mathcal{X}$ the latent hubs positions.

B Experiments

B.1 More experiments and details

Time gain and optimality. The four densities considered in Table 3 are

$$f_1 = \frac{1}{3}\mathcal{N}([0,0],\Sigma) + \frac{1}{3}\mathcal{N}([0,1],\Sigma) + \frac{1}{3}\mathcal{N}([1,1],\Sigma)$$

$$f_2 = \frac{1}{2}\mathcal{N}([0.5,0.5],\Sigma) + \frac{1}{2}\mathcal{N}([-0.5,0.5],\Sigma)$$

$$f_3 = \frac{1}{4}\mathcal{N}([0.5,0.5],\Sigma) + \frac{1}{4}\mathcal{N}([-0.5,0.5],\Sigma) + \frac{1}{4}\mathcal{N}([0,0.5],\Sigma) + \frac{1}{4}\mathcal{N}([-0.5,0],\Sigma)$$

$$f_4 = \frac{1}{2}\mathcal{N}([0.5,-0.5],\Sigma) + \frac{1}{2}\mathcal{N}([-0.5,-0.5],\Sigma)$$

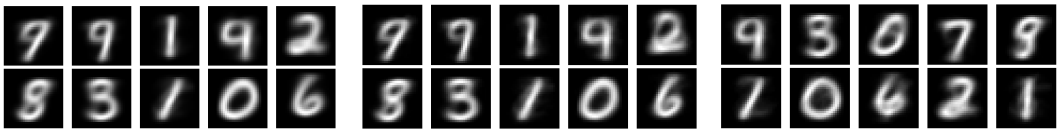
where $\Sigma = 0.05 \times I_2$.

We stopped each method after 3000 outer iterations or after 60 s in Table 1.

Table 2: More values of r in the same setting as Table 1.

r	Method	Transport cost	Marg. err	Iterations	Time (s)
	MSK	0.186 ± 0.022	10^{-8}	246 ± 62	20.3 ± 6.3
	G-MSK	0.186 ± 0.022	10^{-7}	580 ± 122	8.2 ± 1.9
2	ECPMOT	$\mathbf{0.229 \pm 0.022}$	10^{-8}	7 ± 3	$\mathbf{0.2 \pm 0.1}$
	CPMOT	0.258 ± 0.028	10^{-8}	2007 ± 690	2.8 ± 1.9
	TUCKER G-MSK	0.232 ± 0.022	10^{-17}	532 ± 415	60.2 ± 0.1
	TUCKER MSK	0.233 ± 0.022	10^{-17}	492 ± 377	60.2 ± 0.3
4	ECPMOT	$\mathbf{0.202 \pm 0.022}$	10^{-8}	20 ± 9	$\mathbf{0.4 \pm 0.2}$
	CPMOT	0.231 ± 0.028	10^{-8}	2379 ± 568	4.4 ± 2.2
	TUCKER G-MSK	0.231 ± 0.022	10^{-17}	1434 ± 786	59.6 ± 1.8
	TUCKER MSK	0.231 ± 0.022	10^{-17}	1677 ± 970	56.9 ± 6.6
7	ECPMOT	$\mathbf{0.193 \pm 0.023}$	10^{-8}	35 ± 18	$\mathbf{0.6 \pm 0.4}$
	CPMOT	0.222 ± 0.025	10^{-8}	2499 ± 439	5.1 ± 1.7
	TUCKER G-MSK	0.232 ± 0.023	10^{-17}	1127 ± 370	60.2 ± 0.3
	TUCKER MSK	0.231 ± 0.022	10^{-17}	1505 ± 680	59.8 ± 1.1
10	ECPMOT	$\mathbf{0.191 \pm 0.022}$	10^{-8}	56 ± 36	$\mathbf{0.7 \pm 0.4}$
	CPMOT	0.219 ± 0.023	10^{-8}	2511 ± 432	5.1 ± 1.7
	TUCKER G-MSK	0.235 ± 0.022	10^{-17}	759 ± 219	60.3 ± 0.4
	TUCKER MSK	0.234 ± 0.022	10^{-17}	875 ± 295	60.1 ± 0.1
15	ECPMOT	$\mathbf{0.189 \pm 0.022}$	10^{-8}	69 ± 34	$\mathbf{0.9 \pm 0.5}$
	CPMOT	0.218 ± 0.025	10^{-8}	2577 ± 438	5.4 ± 1.8
	TUCKER G-MSK	0.268 ± 0.031	10^{-17}	278 ± 74	62.1 ± 2.0
	TUCKER MSK	0.248 ± 0.022	10^{-17}	389 ± 110	60.2 ± 0.1
20	ECPMOT	$\mathbf{0.189 \pm 0.022}$	10^{-8}	76 ± 33	$\mathbf{0.8 \pm 0.4}$
	CPMOT	0.217 ± 0.025	10^{-8}	2508 ± 452	5.5 ± 1.9
	TUCKER G-MSK	0.295 ± 0.028	10^{-17}	189 ± 51	64.3 ± 4.7
	TUCKER MSK	0.282 ± 0.026	10^{-17}	253 ± 55	60.3 ± 0.2

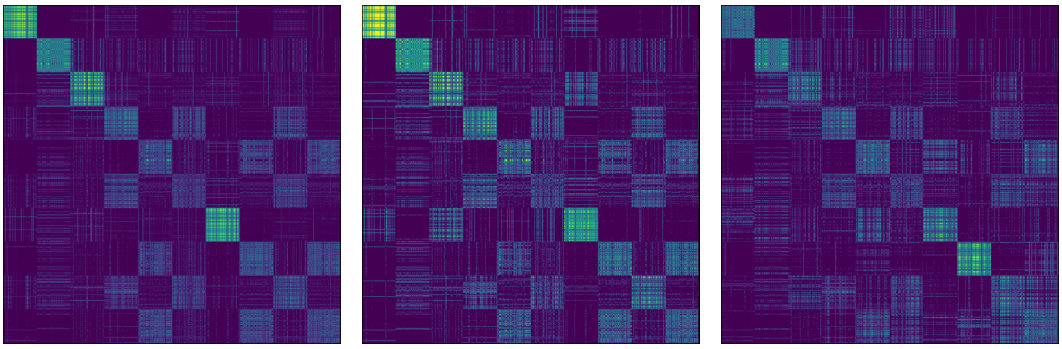
Matching MNIST. We create $d = 20$ measures $\mathbf{a} = (a^{(k)})_{k=1}^{20}$, each supported on 10×150 points which are vectors representing different digits in the MNIST dataset. We take our pairwise cost with the squared euclidean distance as cost tensor $\boldsymbol{\mathcal{C}}$, and solve $\mathrm{CPLRMOT}_{r,\varepsilon}(\mathbf{a}, \boldsymbol{\mathcal{C}})$ with $r = 10$ and $\varepsilon = 0.005$. Note that in this setup, the core $\boldsymbol{\mathcal{G}}$ in the Tucker method would have size 10^{20}, which is impractical. We then consider only the first two measures $a^{(1)}$ and $a^{(2)}$ and solve the $d = 2$ associated problem using RCG, the best competitor in the large scale experiment in Table 3 and ECPMOT taken only on these 2 measures. In Figure 4, we represent the latent barycenter positions z as 28×28 images for each method. In this experiment, we take $\varepsilon = 0.001$.



(a) ECPMOT (d) (b) ECPMOT (2) (c) RCG

Figure 4: Barycenters $z_s \in \mathbb{R}^{784}$ displayed as 28×28 images for the three methods.

We observe that the multimarginal problem obtains the cleanest barycenters. Some integers are merged together, but we are still able to distinguish all 10 of them. We compare the $d = 2$ transport costs with the marginal plan $\mathrm{Marg}_{\{1,2\}}(\boldsymbol{\mathcal{T}})$. Transport costs are ECPMOT (d): 0.3057, ECPMOT (2): 0.3043, RCG: 0.3312. In particular, taking the $(1, 2)$ marginal of $\boldsymbol{\mathcal{T}}$ provides a good transport plan from $a^{(1)}$ to $a^{(2)}$. Transport plans are displayed in Figure 5. Points are ordered according to their corresponding digit, and measures are uniform.



(a) ECPMOT (d) (b) ECPMOT (2) (c) RCG

Figure 5: Marginal plan $\mathrm{Marg}_{\{1,2\}}(\boldsymbol{\mathcal{T}})$ and plans from ECPMOT and RCG with $a^{(1)}$ and $a^{(2)}$ in the MNIST experiment.

Approximation error bound. We use the bound from Corollary 3.5. When the cost is normalized, and with $(n_1, n_2, n_3) = (15, 15, 15)$, $r = 5$ the bound is

$$\ln\left(\frac{43}{4}\right) \approx 2.37.$$

However, in this setting with Gaussian samples ECPMOT achieves a 0.1039 transport cost, when Sinkhorn gets 0.0830, so the error is of the order of 10^{-2}, which is far smaller than the bound. Similarly, in the $d = 2$ case, the bound $\ln(\min(n_1, n_2)/(r-1))$ provided by Scetbon & Cuturi (2022) gives $\ln(200/9) \approx 3.10$ in the setting of Figure 7, when the transport costs achieved by low-rank solvers are ≈ 0.1.

B.2 Experiments in the $d = 2$ case

Methods. In this section, we compare CPMOT, ECPMOT and TUCKER with other state-of-the-art methods, including LATOT (Lin et al., 2021) and RiemCG (RCG) (Jawanpuria & Mishra, 2026). In the $d = 2$ case, CPMOT, ECPMOT and TUCKER are variants of LOT, FC and FRLC from Scetbon et al. (2021), Forrow et al. (2019) and Halmos et al. (2024) respectively. For CPMOT and TUCKER, we use code from these works in this section. Note that the implementation of CPMOT from Scetbon et al. (2021) still takes advantage of the structure of the squared Euclidean cost. We also include TC (Schmidt et al., 2026) when applicable, namely, when $n = m$ and n is small enough for its initialization, which involves computing an optimal full-rank Monge map, to be completed within the prescribed time.

The LATOT formulation allows different inner marginals as in the Tucker formulation. Lin et al. (2021) optimize over the source and target hub locations μ^x and μ^y directly, using the same Sinkhorn barycenter algorithm as in ECPMOT to update couplings.

Setup. We adopt the experimental setup from Jawanpuria & Mishra (2026): each method terminates when $|\langle C, T_k\rangle - \langle C, T_{k-1}\rangle|/\langle C, T_{k-1}\rangle < 10^{-6}$ for five consecutive iterations, after 1000 iterations or after 180 seconds. Apart from this stopping criterion, we use the default parameters for each method. An important parameter of ECPMOT is the maximum number of iteration and tolerance of the inner Sinkhorn-like step, which we set to 1300 and 10^{-7} in order to achieve marginal error $< 10^{-7}$. In the tables below, the marginal error denotes the maximum violation among the marginal constraints, i.e. $\max(\|T\mathbf{1}_m - a\|_\infty, \|T^\mathsf{T}\mathbf{1}_n - b\|_\infty)$. If not specified, the regularization parameter ε is set to 0.005 and the cost is the squared euclidean cost. We use the normalized cost $C/C_{\max}$ in every experiment, except in the LATOT method where numerical errors occur otherwise.

In Table 3, we reproduce the Gaussian-mixture benchmark from Scetbon et al. (2021) (Appendix G). We sample $n = m = 10000$ points in $\mathbb{R}^2$ from two Gaussian mixtures and compare the transport costs obtained by the different low-rank solvers. We report the ratio $f/f^\star$, where $f^\star$ is the exact transport cost and $f = \langle C, T\rangle$ is the transport cost achieved by the concerned method. We also report

Table 3: Comparison of the low-rank transport cost ratio, marginal errors, and computation times of different solvers ($n = m = 10000$, $d = 2$, Gaussian mixtures). Values are averaged over 20 seeds. The TC method is excluded as its initialization takes more than 180 s.

Method	Ratio $f/f^\star$	Marg. err	Iterations	Time (s)	Success rate
Sinkhorn	1.087 ± 0.043	10^{-9}	135 ± 6	43.4 ± 1.8	1.00
Rank $r = 10$					
ECPMOT	1.181 ± 0.052	10^{-10}	86 ± 32	$\mathbf{3.7 \pm 1.2}$	1.00
CPMOT	1.180 ± 0.052	10^{-11}	979 ± 49	153.3 ± 10.3	1.00
TUCKER	1.240 ± 0.058	10^{-8}	101 ± 2	181.1 ± 0.6	1.00
RCG	$\mathbf{1.175 \pm 0.051}$	10^{-7}	31 ± 7	188.2 ± 4.0	1.00
LATOT	1.839 ± 0.015	10^{-9}	46 ± 21	26.2 ± 11.5	1.00
Rank $r = 50$					
ECPMOT	1.055 ± 0.045	10^{-10}	171 ± 76	$\mathbf{16.4 \pm 6.7}$	1.00
CPMOT	1.068 ± 0.043	10^{-11}	445 ± 71	180.3 ± 0.2	1.00
TUCKER	1.014 ± 0.045	10^{-8}	70 ± 2	182.1 ± 0.8	1.00
RCG	$\mathbf{1.002 \pm 0.045}$	10^{-8}	26 ± 3	199.7 ± 10.5	1.00
LATOT	1.832 ± 0.014	10^{-9}	44 ± 10	37.3 ± 8.0	1.00

the *success rate*, where a run fails if the obtained marginal error is above 10^{-4}. The experiment is realized 20 times over different sampling seeds, and values are averaged.

It stands out that ECPMOT converges substantially faster than all other method while achieving good f/f^* ratios. We observe that RCG requires fewer but more costly iterations and achieves the best f/f^* values. Such a low f/f^* ratio for RCG can be explained by the fact that low-rank methods can achieve a lower cost than the exact OT algorithm while maintaining a valid transport plan with marginal violation below 10^{-7}. ECPMOT has the fastest iterations and requires fewer iterations than CPMOT and TUCKER to converge. In Figure 6, we study the evolution of the f/f^* ratio as a function of the elapsed optimization time.

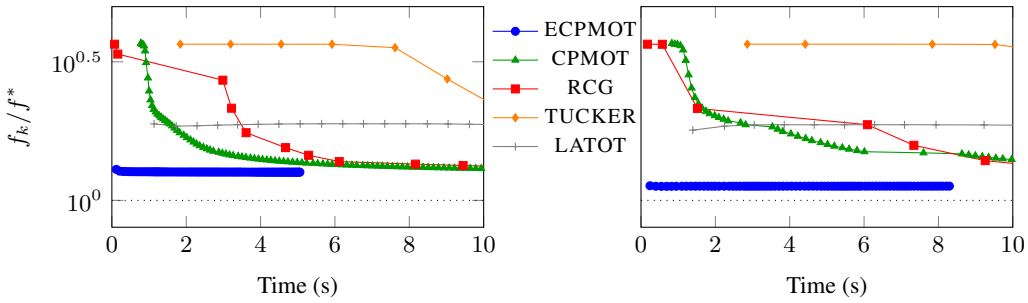


Figure 6: Evolution of the f/f^* ratio during the optimization. Setting is the same as in Table 3. Left $r = 10$, right $r = 50$.

Again, we see that ECPMOT achieves a good time-accuracy tradeoff. In this setting, we see that the optimum is in fact reached in the first iterations. We also observe that LATOT fails to optimize further than its initialization in this setting with default ultra-parameter values, one could try to select a suitable value of ε using a bisection method (as advised by Lin et al. (2021)) at the cost of increased runtime.

In Table 4, we reproduce the study of Jawanpuria & Mishra (2026). We sample $n = 50$, $m = 60$ points from the standard Gaussian distribution in 2D and report the evolution of the low-rank transport cost and the marginal error when the rank r varies. The experiment is realized 50 times over different sampling seeds, and values are averaged.

Table 4: Transport cost, marginal error, and time for each method as a function of the rank r. ($n = 50$, $m = 60$, $d = 2$, standard Gaussians). Values are averaged over 50 runs.

r	Transport cost ($\cdot 10^2$)				Time (s)			
	CPMOT	ECPMOT	TUCKER	RCG	CPMOT	ECPMOT	TUCKER	RCG
2	**8.60 ± 1.19**	8.70 ± 1.15	9.38 ± 1.39	8.68 ± 1.21	2.6	**0.1**	62.9	0.8
3	5.91 ± 0.81	**5.84 ± 0.82**	6.82 ± 1.34	5.97 ± 0.80	3.3	**0.2**	93.9	1.7
5	**4.04 ± 0.55**	4.08 ± 0.59	5.19 ± 0.96	4.17 ± 0.57	9.4	**0.3**	103.8	1.4
8	**2.85 ± 0.44**	2.86 ± 0.46	3.70 ± 0.58	2.91 ± 0.48	16.3	**0.4**	110.7	1.7
10	**2.34 ± 0.40**	2.35 ± 0.37	3.05 ± 0.51	2.40 ± 0.45	17.5	**0.4**	120.6	2.8
15	1.95 ± 0.37	**1.92 ± 0.39**	2.46 ± 0.39	2.02 ± 0.39	14.9	**0.4**	120.0	1.8
20	1.72 ± 0.36	**1.70 ± 0.37**	2.10 ± 0.35	1.77 ± 0.39	13.9	**0.4**	123.3	2.1
30	**1.55 ± 0.35**	**1.55 ± 0.35**	1.77 ± 0.36	**1.55 ± 0.35**	13.6	**0.5**	134.5	1.8

As expected, we observe that the transport cost decreases as the rank increases. In this small scale setting, we see that CPMOT and ECPMOT dominate in terms of transport cost optimality, while ECPMOT remains fastest to converge.

In Table 5, we study the behavior of the algorithms in a setting where the optimal transport plan has rank r. To construct such a setting, we create two sets of r hubs, $(\mu_i^x)_i$ and $(\mu_j^Y)_j$, and place n points exactly at the locations μ_i^x and $m = n$ points at the locations μ_j^y. Measures are taken uniform. We report the average f/f^* value of each method and the rate at witch a method is close ($|f/f^* - 1| < 10^{-3}$) to the exact optimal transport cost over 50 sampling seeds.

Table 5: Transport cost ratio and optimality rate for each method when the optimal transport plan has rank r. $(n, m, r) = (20, 20, 5)$. 50 iterations.

Method	Sinkhorn	ECPMOT	CPMOT	TUCKER	RCG	LATOT	TC
Mean f/f^*	1.003	**1.001**	1.032	1.245	1.035	1.802	2.440
Optimality rate	0.56	**0.80**	0.42	0.20	0.60	0.14	0.00

We observed that low-rank methods tended to merge some latent hubs z_s together, resulting in a sub-optimal transport plan. Except for TUCKER, LATOT and TC, each method finds the optimal transport plan at good rate. In particular, ECPMOT achieves the highest success rate.

In Figure 7, we study the position of the hubs z_s and latent source and target hubs μ_s^x and μ_s^y for the 4 best performing methods in Table 5 in the following setting from Halmos et al. (2024): source points are sampled from Gaussians centered at the 5th roots of unity, while target points are sampled from Gaussians centered at the 10th roots of unity. Low-rank transport plan are computed with $r = 10$. Note that each "blob" does not have the same mass (number of points).

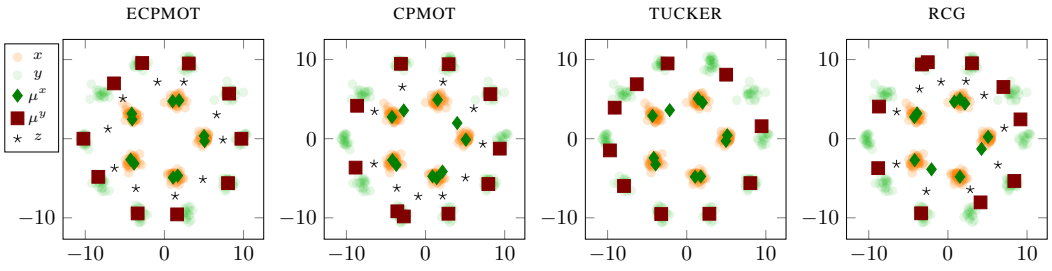


Figure 7: Position of the hubs z_s, the source and target hubs μ_s^x and μ_s^y when the x_i are sample from Gaussians centered at the 5th roots of unity ($\times 5$), the y_j are sampled from Gaussians centered at the 10th roots of unity ($\times 10$), and $(n, m, r) = (200, 200, 10)$. Running times are ECPMOT: 0.29 s, CPMOT: 25.1 s, TUCKER: 73.3 s, RCG: 3.60 s. Transport cost are ECPMOT: 0.1059, CPMOT: 0.1030, TUCKER: 0.1079, RCG: 0.1039.

C PROOF DETAILS

C.1 PROOF OF PROPERTY 3.1

Proof. Consider $\boldsymbol{\mathcal{T}} = a^{(1)} \circ \cdots \circ a^{(d)}$ so that an entry of $\boldsymbol{\mathcal{T}}$ is

$$\boldsymbol{\mathcal{T}}_{i_1 \dots i_d} = \prod_{k=1}^{d} a^{(k)}_{i_k}.$$

Let $\mathcal{I}_m$ be a set of modes. In order to show that $\boldsymbol{\mathcal{T}} \in \Pi_{\mathbf{r},\boldsymbol{\mathcal{I}}}(\mathbf{a})$, we need to show that $\operatorname{rank}_+(\mathbf{T}^{(\mathcal{I}_m)}) \leq r_m$. If k correspond to the indices $(i_j)_{j\in\mathcal{I}_m}$ and l to the indices $(i_j)_{j\in\bar{\mathcal{I}}_m}$, then

$$\mathbf{T}^{(\mathcal{I}_m)}_{kl} = \prod_{j=1}^{d} a^{(j)}_{i_j} = \prod_{j\in\mathcal{I}_m} a^{(j)}_{i_j} \times \prod_{j\in\bar{\mathcal{I}}_m} a^{(j)}_{i_j}.$$

So, defining $u = \otimes_{j\in\mathcal{I}_m} a^{(j)}$ and $v = \otimes_{j\in\bar{\mathcal{I}}_m} a^{(j)}$, we have $\mathbf{T}^{(\mathcal{I}_m)} = uv^{\mathsf{T}}$.

Thus, $\operatorname{rank}_+(\mathbf{T}^{(\mathcal{I}_m)}) = 1 \leq r_m$. □

C.2 PROOF OF PROPERTY 3.2

Proof. For each block $\mathcal{I}_m = \{j_1 < \cdots < j_J\}$ define a bijection

$$\phi_m : \prod_{j\in\mathcal{I}_m} [\![n_j]\!] \to [\![N_m]\!],$$

for example $\phi_m(i_1, \dots, i_J) = 1 + \sum_{\ell=1}^{J}(i_{j_\ell} - 1)\prod_{q=\ell+1}^{J} n_{j_q}$. We define the ungrouped tensor $\widetilde{\boldsymbol{\mathcal{T}}} \in \mathbb{R}_+^{n_1\times\cdots\times n_d}$ by

$$\widetilde{\boldsymbol{\mathcal{T}}}_{i_1,\dots,i_d} = \boldsymbol{\mathcal{T}}_{\phi_1((i_j)_{j\in\mathcal{I}_1}),\dots,\phi_M((i_j)_{j\in\mathcal{I}_M})}.$$

Let $j \in \mathcal{I}_m$, then as $\boldsymbol{\mathcal{I}}$ is a partition of $[\![d]\!]$, taking the j-th marginal of the ungrouped tensor is equivalent to first taking the m-th marginal of the grouped tensor $\boldsymbol{\mathcal{T}}$, and then the marginal of the result on the mode corresponding to j (that we call j with an abuse of notation):

$$\operatorname{Marg}_j(\widetilde{\boldsymbol{\mathcal{T}}}) = \operatorname{Marg}_j(\operatorname{Marg}_m(\boldsymbol{\mathcal{T}})) = \operatorname{Marg}_j(\otimes_{\ell\in\mathcal{I}_m} a^{(\ell)}) = a^{(j)},$$

so $\boldsymbol{\mathcal{T}} \in \Pi(\mathbf{a})$. Now, the two matrices $\widetilde{\mathbf{T}}^{(\mathcal{I}_m)}$ and $\mathbf{T}^{(m)}$ differ only by a permutation of the rows and columns. Nonnegative rank is invariant under such permutations, and thus

$$\operatorname{rank}_+(\widetilde{\mathbf{T}}^{(\mathcal{I}_m)}) = \operatorname{rank}_+(\mathbf{T}^{(m)}) \leq r_m.$$

Finally, $\widetilde{T} \in \Pi_{\mathbf{r},\boldsymbol{\mathcal{I}}}(\mathbf{a})$. □

C.3 PROOF OF PROPERTY 3.3

Proof. • **Not empty** : The independent coupling $\boldsymbol{\mathcal{T}} = a^{(1)} \circ \cdots \circ a^{(d)}$ belongs to $\Pi^{\mathrm{T}}_{\mathbf{r}}(\mathbf{a})$. Indeed, setting $P^{(k)} = \begin{bmatrix} p^{(k)} & \cdots & p^{(k)} \end{bmatrix} \in \mathbb{R}_+^{n_k\times r_m}$, each $P^{(k)}$ is column-stochastic. For any $\boldsymbol{\mathcal{G}} \in \Delta_{r_1\times\cdots\times r_d}$,

$$\boldsymbol{\mathcal{T}} = \boldsymbol{\mathcal{G}} \times_1 P^{(1)} \cdots \times_d P^{(d)}.$$

Consequently, $\Pi^{\mathrm{T}}_{\mathbf{r}}(\mathbf{a})$ is nonempty.

• **Compact** : The parameter space

$$\Delta_{r_1\times\cdots\times r_d} \times \prod_{k=1}^{d} \mathcal{S}^{n_k\times r_m}_{\mathrm{col}}$$

is compact and the map

$$(\boldsymbol{\mathcal{G}}, P^{(1)}, \dots, P^{(d)}) \mapsto \boldsymbol{\mathcal{G}} \times_1 P^{(1)} \cdots \times_d P^{(d)}$$

is continuous. Therefore, its image is compact and intersecting it with the compact set $\Pi(\mathbf{a})$ gives that $\Pi^{\mathrm{T}}_{\mathbf{r}}(\mathbf{a})$ is compact. □

C.4 Proof of Property 3.4

Proof. We follow ideas from the proof of Proposition 1 in Scetbon & Cuturi (2022). Let $\boldsymbol{\mathcal{T}}^\star$ be an optimal solution of $\mathrm{MOT}(\mathbf{a}, \boldsymbol{\mathcal{C}})$. Since $\boldsymbol{\mathcal{T}}^\star \geq 0$, it admits a nonnegative rank-one decomposition. Since $\boldsymbol{\mathcal{T}}^\star$ has total mass one, this decomposition can be written as

$$\boldsymbol{\mathcal{T}}^\star = \sum_{s=1}^{R} \lambda_s r_s^{(1)} \circ \cdots \circ r_s^{(d)}$$

with $\lambda \in \Delta_R^*$ and $r_s^{(k)} \in \Delta_{n_k}$. Without loss of generality, assume $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_R$. Since $\lambda \in \Delta_R$, we have $k\lambda_k \leq \sum_{s=1}^{k} \lambda_s \leq 1$, and hence $\lambda_k \leq \frac{1}{k}$. Define the mass

$$\Lambda_r = \sum_{s=r}^{R} \lambda_s$$

and, for every $k \in [\![d]\!]$, the normalized tail marginal

$$\bar{r}^{(k)} = \frac{1}{\Lambda_r} \sum_{s=r}^{R} \lambda_s r_s^{(k)}.$$

Since every $r_s^{(k)}$ belongs to Δ_{n_k}, we have $\bar{r}^{(k)} \in \Delta_{n_k}$. Now, we construct a transport tensor $\widetilde{\boldsymbol{\mathcal{T}}}_r \in \Pi_r(\mathbf{a})$:

$$\widetilde{\boldsymbol{\mathcal{T}}}_r = \sum_{s=1}^{r-1} \lambda_s r_s^{(1)} \circ \cdots \circ r_s^{(d)} + \Lambda_r \bar{r}^{(1)} \circ \cdots \circ \bar{r}^{(d)}.$$

Indeed, the k-th marginal of $\widetilde{\boldsymbol{\mathcal{T}}}_r$ is

$$\mathrm{Marg}_k(\widetilde{\boldsymbol{\mathcal{T}}}_r) = \sum_{s=1}^{r-1} \lambda_s r_s^{(k)} + \Lambda_r \bar{r}^{(k)} = a^{(k)}.$$

It remains to verify that $\widetilde{\boldsymbol{\mathcal{T}}}_r$ belongs to $\Pi_{\mathbf{r}}^{\mathrm{T}}(\mathbf{a})$. The tensor $\widetilde{\boldsymbol{\mathcal{T}}}_r$ has a CP decomposition with r nonnegative components. Therefore, for every mode k, the unfolding $\widetilde{\mathbf{T}}_r^{(k)}$ admits a nonnegative rank-r decomposition, and hence $\mathrm{rank}_+(\widetilde{\mathbf{T}}_r^{(k)}) \leq r$.

Write

$$\widetilde{\boldsymbol{\mathcal{T}}}_r = \sum_{s=1}^{r} \mu_s u_s^{(1)} \circ \cdots \circ u_s^{(d)},$$

where the last component corresponds to the tail marginal. For each mode k, define

$$p_s^{(k)} = u_s^{(k)} \in \Delta_{n_k}.$$

Complete these r columns with arbitrary probability vectors $p_{r+1}^{(k)}, \ldots, p_{r_k}^{(k)}$ and set $P^{(k)} = \left[p_1^{(k)} \cdots p_{r_k}^{(k)}\right]$. Thus $P^{(k)} \in \mathcal{S}_{\mathrm{col}}^{n_k \times r_m}$. Then,

$$\widetilde{\boldsymbol{\mathcal{T}}}_r = \boldsymbol{\mathcal{G}} \times_1 P^{(1)} \cdots \times_d P^{(d)},$$

where $\boldsymbol{\mathcal{G}}_{i_1 \ldots i_d} = \mu_s$ if $i_1 = \cdots = i_d = s$ and $s \leq r$, and 0 otherwise. Therefore $\widetilde{\boldsymbol{\mathcal{T}}}_r \in \Pi_{\mathbf{r}}^{\mathrm{T}}(\mathbf{a})$. Since $\mathrm{TuckerMOT}_{\mathbf{r}}$ minimizes over a subset of $\Pi(\mathbf{a})$,

$$0 \leq \mathrm{TuckerMOT}_{\mathbf{r}}(\mathbf{a}, \boldsymbol{\mathcal{C}}) - \mathrm{MOT}(\mathbf{a}, \boldsymbol{\mathcal{C}}).$$

Furthermore, using $\widetilde{\boldsymbol{\mathcal{T}}}_r$ as a feasible candidate,

$$\begin{aligned}
\mathrm{TuckerMOT}_{\mathbf{r}}(\mathbf{a}, \boldsymbol{\mathcal{C}}) - \mathrm{MOT}(\mathbf{a}, \boldsymbol{\mathcal{C}}) &\leq \langle \boldsymbol{\mathcal{C}}, \widetilde{\boldsymbol{\mathcal{T}}}_r \rangle - \langle \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{T}}^\star \rangle \\
&\leq \langle \boldsymbol{\mathcal{C}}, \Lambda_r \bar{r}^{(1)} \circ \cdots \circ \bar{r}^{(d)} \rangle - \langle \boldsymbol{\mathcal{C}}, \sum_{s=r}^{R} \lambda_s r_s^{(1)} \circ \cdots \circ r_s^{(d)} \rangle \\
&\leq \|\boldsymbol{\mathcal{C}}\|_\infty \Lambda_r.
\end{aligned}$$

Using $\lambda_k \leq 1/k$ and $\sum_{s=r}^{R} \frac{1}{s} \leq \int_{r-1}^{R} \frac{1}{x}\,\mathrm{d}x = \ln(R/(r-1))$, we get

$$\mathrm{TuckerMOT}_{\mathbf{r},\boldsymbol{\mathcal{I}}}(\mathbf{a},\boldsymbol{\mathcal{C}}) - \mathrm{MOT}(\mathbf{a},\boldsymbol{\mathcal{C}}) \leq \|\boldsymbol{\mathcal{C}}\|_\infty \ln\Big(\frac{R}{r-1}\Big),$$

which proves the result. □

C.5 Proof of Property 3.6

Proof. • **Non empty** : The independent coupling $\boldsymbol{\mathcal{T}} = a^{(1)} \circ \cdots \circ a^{(d)}$ is in $\Pi_r^{\mathrm{CP}}(\mathbf{a})$.

• **Compact** : Proposition 6.2 from Lim & Comon (2009) gives us that $\{\boldsymbol{\mathcal{T}} \in \mathbb{R}_+^{n_1 \times \cdots \times n_d} \mid \mathrm{rank}_{\mathrm{CP}+}(\boldsymbol{\mathcal{T}}) \leq r\}$ is closed. This closure comes from the choice of the *nonnegative* rank and the *inequality* constraint. Therefore, intersecting it with the compact set $\Pi(\mathbf{a})$ gives that $\Pi_r^{\mathrm{CP}}(\mathbf{a})$ is comapct. □

C.6 Proof of Theorem 4.1

For any s, let $(X_s^{(k)})_{k=1}^d$ be independent random variables with respective laws $X_s^{(k)} \sim \sum_{i_k=1}^{n_k} p_{i_k s}^{(k)} \delta_{\varphi(x_{i_k}^{(k)})}$ which are denoted by $A_s^{(k)}$ in the main article. Therefore, using the joint law of the $(X_s^{(k)})_{k=1}^d$, the total cost in Eq. (7) rewrites

$$\begin{aligned}
\sum_{i_1,\ldots,i_d} \mathcal{C}_{i_1,\ldots,i_d} \sum_{s=1}^{r} \frac{Q_{i_1 s}^{(1)} \times \cdots \times Q_{i_d s}^{(d)}}{g_s^{d-1}} &= \sum_{i_1,\ldots,i_d} \sum_{s=1}^{r} g_s \sum_{k<l} \|\varphi(x_{i_k}^{(k)}) - \varphi(x_{i_l}^{(l)})\|_{\mathcal{H}}^2 p_{i_1 s}^{(1)} \times \cdots \times p_{i_d s}^{(d)} \\
&= \sum_{s=1}^{r} g_s \sum_{k<l} \sum_{i_1,\ldots,i_d} f_{kl}(x_{i_1}^{(1)},\ldots,x_{i_d}^{(d)}) P((X_s^{(1)},\ldots,X_s^{(d)}) = (x_{i_1}^{(1)},\ldots,x_{i_d}^{(d)})) \\
&= \sum_{s=1}^{r} g_s \sum_{k<l} \mathbb{E}\big[f_{kl}(X_s^{(1)},\ldots,X_s^{(d)})\big] \\
&= \sum_{s=1}^{r} g_s \mathbb{E}\big[\sum_{k<l} \|X_s^{(k)} - X_s^{(l)}\|_{\mathcal{H}}^2\big] \qquad (9)
\end{aligned}$$

where $f_{kl}(x_{i_1}^{(1)},\ldots,x_{i_d}^{(d)}) = \|\varphi(x_{i_k}^{(k)}) - \varphi(x_{i_l}^{(l)})\|_{\mathcal{H}}^2$.

In order to introduce the hubs positions, we denote $\bar{X} = \frac{1}{d}\sum_k X_s^{(k)}$ and state the following lemma:

Lemma C.1. *For all $z \in \mathcal{X}$, the following equalities holds:*

$$\sum_{k<l} \|X_s^{(k)} - X_s^{(l)}\|_{\mathcal{H}}^2 = d\sum_{k=1}^{d} \|X_s^{(k)} - \varphi(z)\|_{\mathcal{H}}^2 - d^2\|\bar{X} - \varphi(z)\|_{\mathcal{H}}^2,$$

$$\mathbb{E}\big[\|X_s^{(k)} - \varphi(z)\|_{\mathcal{H}}^2\big] = \sum_{i_k=1}^{n_k} \|X_s^{(k)} - \varphi(z)\|_{\mathcal{H}}^2 p_{i_k s}^{(k)}$$

In particular, $\sum_{k<l} \|X_s^{(k)} - X_s^{(l)}\|_{\mathcal{H}}^2 \leq d\sum_{k=1}^{d} \|X_s^{(k)} - \varphi(z)\|_{\mathcal{H}}^2$. Plugging this inequality in Eq. (9) and using the second equality yields

$$(9) \leq d\sum_{s=1}^{r} g_s \mathbb{E}\big[\sum_{k=1}^{d} \|X_s^{(k)} - \varphi(z_s)\|_{\mathcal{H}}^2\big] = d\sum_{s=1}^{r}\sum_{k=1}^{d}\sum_{i_k=1}^{n_k} Q_{i_k s}^{(k)} \|\varphi(x_{i_k}^{(k)}) - \varphi(z_s)\|_{\mathcal{H}}^2, \qquad (10)$$

so that the explicit CPMOT is

$$\inf_{\mathbf{Q},g} d\sum_{s=1}^{r}\sum_{k=1}^{d}\sum_{i_k=1}^{n_k} Q_{i_k s}^{(k)} \|\varphi(x_{i_k}^{(k)}) - \varphi(z_s)\|_{\mathcal{H}}^2.$$

Proof of Lemma C.1. First,

$$\sum_{k<l} \|X_s^{(k)} - X_s^{(l)}\|_{\mathcal{H}}^2 = (d-1)\sum_{k=1}^{d} \|X_s^{(k)}\|_{\mathcal{H}}^2 - 2\sum_{k<l}\langle X_s^{(k)}, X_s^{(l)}\rangle_{\mathcal{H}}$$
$$= d\sum_{k=1}^{d} \|X_s^{(k)}\|_{\mathcal{H}}^2 - d^2\|\bar{X}\|_{\mathcal{H}}^2.$$

The last equality comes from the fact that $\|\sum_k X_s^{(k)}\|_{\mathcal{H}}^2 = \sum_k \|X_s^{(k)}\|_{\mathcal{H}}^2 + 2\sum_{k<l}\langle X_s^{(k)}, X_s^{(l)}\rangle_{\mathcal{H}}$. Then,

$$\sum_{k=1}^{d} \|X_s^{(k)} - \bar{X}\|_{\mathcal{H}}^2 = \sum_{k=1}^{d} \|X_s^{(k)}\|_{\mathcal{H}} - 2\langle\sum_{k=1}^{d} X_s^{(k)}, \bar{X}\rangle_{\mathcal{H}} + d\|\bar{X}\|_{\mathcal{H}}^2$$
$$= \sum_{k=1}^{d} \|X_s^{(k)}\|_{\mathcal{H}}^2 - d\|\bar{X}\|_{\mathcal{H}}^2. \tag{11}$$

So that

$$\sum_{k<l} \|X_s^{(k)} - X_s^{(l)}\|_{\mathcal{H}}^2 = d\sum_{k=1}^{d} \|X_s^{(k)} - \bar{X}\|_{\mathcal{H}}^2. \tag{12}$$

For any $z \in \mathcal{X}$,

$$\sum_{k=1}^{d} \|X_s^{(k)} - \bar{X}\|_{\mathcal{H}}^2 = \sum_{k=1}^{d} \|X_s^{(k)} - \varphi(z) - (\bar{X} - \varphi(z))\|_{\mathcal{H}}^2$$
$$= \sum_{k=1}^{d} \|X_s^{(k)} - \varphi(z)\|_{\mathcal{H}}^2 - d\|\bar{X} - \varphi(z)\|_{\mathcal{H}}^2 \quad \text{as in Eq. (11).}$$

Finally,

$$\forall z \in \mathcal{X}, \quad \sum_{k<l} \|X_s^{(k)} - X_s^{(l)}\|_{\mathcal{H}}^2 = d\sum_{k=1}^{d} \|X_s^{(k)} - \varphi(z)\|_{\mathcal{H}}^2 - d^2\|\bar{X} - \varphi(z)\|_{\mathcal{H}}^2.$$

Since $\|X_s^{(k)} - \varphi(z_s)\|_{\mathcal{H}}^2$ is a function of $X_s^{(k)}$ only, its expectation depends only on the marginal distribution of $X_s^{(k)}$. Hence, the second equality:

$$\mathbb{E}\big[\|X_s^{(k)} - \varphi(z_s)\|_{\mathcal{H}}^2\big] = \sum_{i_k=1}^{n_k} \|\varphi(x_{i_k}^{(k)}) - \varphi(z_s)\|_{\mathcal{H}}^2 p_{i_k s}^{(k)}.$$

C.7 Proof of Theorem 4.2

We denote by $X_s^{(k)}$ the quantity $A_s^{(k)}$ from the main article. From $\varphi(z_s^\star) = \frac{1}{d}\sum_{k=1}^{d} \mathbb{E}[X_s^{(k)}]$, we have for any k,

$$\mathbb{E}[X_s^{(k)}] - \varphi(z_s^\star) = \frac{1}{d}\sum_{l=1}^{d} \mathbb{E}[X_s^{(k)}] - \mathbb{E}[X_s^{(l)}].$$

Thus, denoting $\mu_s^{(k)} = \mathbb{E}[X_s^{(k)}]$, the expectation from Eq. (10), it rewrites

$$\mathbb{E}\big[\|X_s^{(k)} - \varphi(z_s^\star)\|_{\mathcal{H}}^2\big] = \|\mu_s^{(k)} - \varphi(z_s^\star)\|_{\mathcal{H}}^2 + \mathrm{Var}(X_s^{(k)})$$
$$= \frac{1}{d^2}\|\sum_{l\neq k} \mu_s^{(k)} - \mu_s^{(l)}\|_{\mathcal{H}}^2 + \mathrm{Var}(X_s^{(k)}).$$

So that the objective of explicit CPMOT (10) is

$$F(\boldsymbol{\mathcal{T}}) = \sum_s g_s \big(d \sum_k \mathrm{Var}(X_s^{(k)}) + \frac{1}{d} \sum_k \| \sum_{l \neq k} \mu_s^{(k)} - \mu_s^{(l)} \|_{\mathcal{H}}^2 \big).$$

On the other hand, expectation from Eq. (9) is

$$\mathbb{E}\big[\|X_s^{(k)} - X_s^{(l)}\|_{\mathcal{H}}^2\big] = \mathrm{Var}(X_s^{(k)}) + \mathrm{Var}(X_s^{(l)}) + \|\mu_s^{(k)} - \mu_s^{(l)}\|_{\mathcal{H}}^2.$$

and the total transport cost (9) is

$$\langle \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{T}} \rangle = \sum_s g_s \big((d-1) \sum_{k=1}^{d} \mathrm{Var}(X_s^{(k)}) + \frac{1}{2} \sum_{k \neq l} \|\mu_s^{(k)} - \mu_s^{(l)}\|_{\mathcal{H}}^2 \big).$$

Lemma C.2. *The following equality holds,*

$$\frac{1}{d} \sum_k \| \sum_{l \neq k} \mu_s^{(k)} - \mu_s^{(l)}] \|_{\mathcal{H}}^2 = \frac{1}{2} \sum_{k \neq l} \|\mu_s^{(k)} - \mu_s^{(l)}\|_{\mathcal{H}}^2.$$

Proof. Let $\bar{\mu}_s = \frac{1}{d} \sum_{k=1}^{d} \mu_s^{(k)}$ For fixed s and k, we have

$$\sum_{l \neq k} (\mu_s^{(k)} - \mu_s^{(l)}) = d(\mu_s^{(k)} - \bar{\mu}_s).$$

Therefore,

$$\frac{1}{d} \sum_{k=1}^{d} \| \sum_{l \neq k} \mu_s^{(k)} - \mu_s^{(l)}] \|_{\mathcal{H}}^2 = d \sum_{k=1}^{d} \|\mu_s^{(k)} - \bar{\mu}_s\|_{\mathcal{H}}^2.$$

Using that $\sum_{k=1}^{d} \|\mu_s^{(k)} - \bar{\mu}_s\|_{\mathcal{H}}^2 = \frac{1}{d} \sum_{k<l} \|\mu_s^{(k)} - \mu_s^{(l)}]\|_{\mathcal{H}}^2$ (Eq. (11)), we get

$$\frac{1}{d} \sum_{k=1}^{d} \| \sum_{l \neq k} \mu_s^{(k)} - \mu_s^{(l)}] \|_{\mathcal{H}}^2 = \sum_{k<l} \|\mu_s^{(k)} - \mu_s^{(l)}]\|_{\mathcal{H}}^2 = \frac{1}{2} \sum_{k \neq l} \|\mu_s^{(k)} - \mu_s^{(l)}]\|_{\mathcal{H}}^2.$$

□

Combining the expressions for $F(\boldsymbol{\mathcal{T}})$ and $\langle \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{T}} \rangle$, and the Lemma C.2, we get Theorem 4.2.

C.8 Proof of Property 5.1

Proof. For a fixed index s, the terms depending on $\varphi(z_s) = u$ are

$$F_s(u) = \sum_{k=1}^{d} \sum_{i_k=1}^{n_k} Q_{i_k s}^{(s)} \|\varphi(x_{i_k}^{(k)}) - u\|_{\mathcal{H}}^2.$$

Expanding the squared norm gives

$$F_s(u) = \sum_{k,i_k} Q_{i_k s}^{(k)} \|\varphi(x_{i_k}^{(k)})\|_{\mathcal{H}}^2 - 2\langle \sum_{k,i_k} Q_{i_k s}^{(k)} \varphi(x_{i_k}^{(k)}), u \rangle_{\mathcal{H}} + \big(\sum_{k,i_k} Q_{i_k s}^{(k)} \big) \|u\|_{\mathcal{H}}^2.$$

We have $\sum_{k,i_k} Q_{i_k s}^{(k)} = d g_s$, so

$$F_s(u) = \mathrm{const} - 2\langle \sum_k g_s \mathbb{E}[X_s^{(k)}], u \rangle_{\mathcal{H}} + d g_s \|u\|_{\mathcal{H}}^2.$$

Completing the square,

$$F_s(\varphi(z_s)) = \mathrm{const} + d g_s \Big\| u - \frac{\sum_k g_s \mathbb{E}[X_s^{(k)}]}{d g_s} \Big\|_{\mathcal{H}}^2 - \mathrm{const}_2.$$

Therefore,

$$\varphi(z_s^\star) = \frac{1}{d} \sum_{k=1}^{d} \mathbb{E}[X_s^{(k)}],$$

which is the result. □

D SINKHORN MOT

D.1 CLASSICAL SINKHORN MOT

Benamou et al. (2015) introduce an entropic regularization of Eq. (1), yielding the entropic MOT problem:

$$\mathrm{MOT}_\varepsilon(\mathbf{a}, \boldsymbol{\mathcal{C}}) = \min_{\boldsymbol{\mathcal{T}} \in \Pi(\mathbf{a})} \langle \mathcal{C}, \mathcal{T} \rangle - \varepsilon \mathcal{H}(\boldsymbol{\mathcal{T}}). \tag{13}$$

Authors show that Eq. (13) has a unique minimizer $\boldsymbol{\mathcal{T}}_\varepsilon^\star$ that converges to a solution of Eq. (1) as $\varepsilon \to 0$. They also derive the multimarginal Sinkhorn algorithm, which we present in the log domain in Algorithm 2, using the broadcasting function $\mathrm{broadcast}_i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_1 \times \cdots \times n_d}$ defined by $[\mathrm{broadcast}_i(\mathrm{f}^{(i)})]_{j_1,\ldots,j_d} = \mathrm{f}^{(i)}_{j_i}$. This work was further extended in Lin et al. (2022), introducing a greedy step and acceleration.

Algorithm 2 Multimarginal Sinkhorn

function MSK($\boldsymbol{\mathcal{C}}$, $\mathbf{a}$, ε)
  for $1 \le k \le d$ **do**
    $\mathrm{f}_0^{(k)} \leftarrow \mathbf{0} \in \mathbb{R}^{n_k}$
  end for
  $\log \boldsymbol{\mathcal{K}} \leftarrow -\boldsymbol{\mathcal{C}}/\varepsilon$
  $\log \boldsymbol{\mathcal{T}} \leftarrow \log \boldsymbol{\mathcal{K}}$
  for $t \ge 0$ **do** ▷ until convergence
    for $1 \le i \le d$ **do**
      $\mathrm{f}_{t+1}^{(i)} \leftarrow \mathrm{f}_t^{(i)} + \log(a^{(i)}) - \log(\mathrm{Marg}_i(\boldsymbol{\mathcal{T}}))$
      $\log \boldsymbol{\mathcal{T}} \leftarrow \log \boldsymbol{\mathcal{K}} + \sum_{i=1}^d \mathrm{broadcast}_i(\mathrm{f}_t^{(i)})$
    end for
  end for
  return $\boldsymbol{\mathcal{T}} = \exp(\log \boldsymbol{\mathcal{T}})$
end function=0

D.2 GREEDY AND ACCELERATED SINKHORN MOT

In the greedy version (Algorithm 3), we choose an index K on which we perform the Bregman projection with

$$K \leftarrow \arg\max_{i \in [\![d]\!]} \mathbf{1}_{n_i}^\mathsf{T} (\mathrm{Marg}_i(\boldsymbol{\mathcal{T}}) - a^{(i)}) + \mathrm{KL}(a^{(i)} || \mathrm{Marg}_i(\boldsymbol{\mathcal{T}})). \tag{14}$$

The accelerated version (Algorithm 4) consider a dual entropic regularized problem:

$$\min_{(\mathrm{f}^{(k)})_k} \varphi(\mathrm{f}^{(1)}, \ldots, \mathrm{f}^{(d)})$$

where $\varphi(\mathrm{f}^{(1)}, \ldots, \mathrm{f}^{(d)}) = \log(\|\boldsymbol{\mathcal{T}}(\mathrm{f})\|_1) - \sum_{k=1}^d (\mathrm{f}^{(k)})^\mathsf{T} a^{(k)}$.

Algorithm 3 Greedy Multimarginal Sinkhorn

function G-MSK($\boldsymbol{\mathcal{C}}$, $\mathbf{a}$, ε)
  for $1 \leq k \leq d$ **do**
    $\mathrm{f}_0^{(k)} \leftarrow \mathbf{0} \in \mathbb{R}^{n_k}$
  end for
  $\log \boldsymbol{\mathcal{K}} \leftarrow -\boldsymbol{\mathcal{C}}/\varepsilon$
  $\log \boldsymbol{\mathcal{T}} \leftarrow \log \boldsymbol{\mathcal{K}} + \sum_{k=1}^{d} \mathrm{broadcast}_k(\mathrm{f}_0^{(k)})$
  for $t \geq 0$ **do** ▷ until convergence
    Choose K according to (14)
    $\mathrm{f}_{t+1}^{(K)} \leftarrow \mathrm{f}_t^{(K)} + \log(a^{(K)}) - \log(\mathrm{Marg}_K(\boldsymbol{\mathcal{T}}))$
    $\log \boldsymbol{\mathcal{T}} \leftarrow \log \boldsymbol{\mathcal{K}} + \sum_{i=1}^{d} \mathrm{broadcast}_i(\mathrm{f}_t^{(i)})$
  end for
  return $\boldsymbol{\mathcal{T}} = \exp(\log \boldsymbol{\mathcal{T}})$
end function

Algorithm 4 Accelerated Multimarginal Sinkhorn

function A-MSK($\boldsymbol{\mathcal{C}}$, $\mathbf{a}$, ε)
  for $1 \leq k \leq d$ **do**
    $\check{\mathrm{f}}_0^{(k)} \leftarrow \mathbf{0} \in \mathbb{R}^{n_k}$, $\tilde{\mathrm{f}}_0^{(k)} \leftarrow \mathbf{0} \in \mathbb{R}^{n_k}$
  end for
  $\theta_0 \leftarrow 1$, $K \leftarrow 1$
  $\log \boldsymbol{\mathcal{K}} \leftarrow -\boldsymbol{\mathcal{C}}/\varepsilon$ ▷ used to compute $\boldsymbol{\mathcal{T}}(\mathrm{f})$
  for $t \geq 0$ **do** ▷ until convergence
    $\bar{\mathrm{f}}_t \leftarrow (1-\theta_t)\check{\mathrm{f}}_t + \theta_t \tilde{\mathrm{f}}_t$
    for $1 \leq k \leq d$ **do**
      $\tilde{\mathrm{f}}_{t+1}^{(k)} \leftarrow \tilde{\mathrm{f}}_t^{(k)} - \frac{1}{d\theta_t}(\mathrm{Marg}_k(\boldsymbol{\mathcal{T}}(\bar{\mathrm{f}}))/\|(\boldsymbol{\mathcal{T}}(\bar{\mathrm{f}})\|_1) - a^{(k)})$
    end for
    $\grave{\mathrm{f}}_t \leftarrow \bar{\mathrm{f}}_t + \theta_t(\tilde{\mathrm{f}}_{t+1} - \tilde{\mathrm{f}}_t)$
    for $1 \leq k \leq d$ **do**
      if $k = K$ **then**
        $\hat{\mathrm{f}}_t^{(k)} \leftarrow \grave{\mathrm{f}}_t^{(k)} + \log(a^{(k)}) - \log(\mathrm{Marg_k}(\boldsymbol{\mathcal{T}}(\grave{\mathrm{f}}_t)))$
      else
        $\hat{\mathrm{f}}_t^{(k)} \leftarrow \grave{\mathrm{f}}_t^{(k)}$
      end if
    end for
    $\mathrm{f}_t \leftarrow \arg\min\{\varphi(\mathrm{f}) \mid \mathrm{f} \in \{\check{\mathrm{f}}_t, \hat{\mathrm{f}}_t\}\}$
    Choose K according to (14)
    for $1 \leq k \leq d$ **do**
      if $k = K$ **then**
        $\check{\mathrm{f}}_{t+1}^{(k)} \leftarrow \mathrm{f}_t^{(k)} + \log(a^{(k)}) - \log(\mathrm{Marg_k}(\boldsymbol{\mathcal{T}}(\mathrm{f_t})))$
      else
        $\hat{\mathrm{f}}_{t+1}^{(k)} \leftarrow \mathrm{f}_t^{(k)}$
      end if
    end for
    $\theta_{t+1} \leftarrow \theta_t(\sqrt{\theta_t^2 + 4} - \theta_t)/2$
  end for
  return $\boldsymbol{\mathcal{T}}(\mathrm{f}_t)$
end function

E MIRROR DESCENT FOR TUCKERMOT

We consider the regularized problem

$$\min_{\boldsymbol{\mathcal{G}}, \mathbf{Q}} \langle \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{G}} \times_1 P^{(1)} \cdots \times_d P^{(d)} \rangle - \varepsilon \mathcal{H}(\boldsymbol{\mathcal{G}}) \tag{15}$$

with the constraints $Q^{(k)} \in \Pi(a^{(k)}, g^{(k)})$ and $g^{(k)} = \mathrm{Marg}_k(\boldsymbol{\mathcal{G}})$ where $g^{(k)}$ is an independent variable and $Q^{(k)} = P^{(k)}\mathrm{diag}(g^{(k)})$. We perform alternating KL mirror-descent/proximal updates on the factors and an entropic Sinkhorn update on the core.

Step on Q. Fix an index k, an keep $\boldsymbol{\mathcal{G}}$ and the other factors $P^{(\ell)}$ fixed. The entropy term $-\varepsilon\mathcal{H}(\boldsymbol{\mathcal{G}})$ is then constant. Define

$$A^{(k)} = \boldsymbol{\mathcal{C}} \times_{\ell \neq k} (P^{(\ell)})^{\mathsf{T}} \times_{\ell \neq k} \boldsymbol{\mathcal{G}}.$$

The cost $\langle \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{T}} \rangle$ is then equal to

$$\langle A^{(k)}, P^{(k)} \rangle = \langle \underbrace{A^{(k)}\mathrm{diag}(g^{(k)})^{-1}}_{C_{Q^{(k)}}}, Q^{(k)} \rangle.$$

We want to minimize

$$F(Q^{(k)}, g^{(k)}) = \langle C_{Q^{(k)}}, Q^{(k)} \rangle$$

under the constraints

$$g^{(k)} - \mathrm{Marg}_k(\boldsymbol{\mathcal{G}}) = 0, \qquad (Q^{(k)})^{\mathsf{T}}\mathbf{1}_{n_k} - g^{(k)} = 0.$$

Denoting $\alpha^{(k)}$ and $\beta^{(k)}$ the Lagrange multipliers associated with these constraints, the Lagrangian reads

$$\mathcal{L} = F(Q^{(k)}, g^{(k)}) + \langle \alpha^{(k)}, g^{(k)} - \mathrm{Marg}_k(\boldsymbol{\mathcal{G}}) \rangle + \langle \beta^{(k)}, (Q^{(k)})^{\mathsf{T}}\mathbf{1}_{n_k} - g^{(k)} \rangle.$$

Consequently,

$$\nabla_{Q^{(k)}}\mathcal{L} = \nabla_{Q^{(k)}}F + \mathbf{1}_{n_k}(\beta^{(k)})^{\mathsf{T}}.$$

We now compute the dependence of F on $g^{(k)}$. Since $C_{Q^{(k)}} = A^{(k)}/g^{(k)}$, we have $\nabla_{g^{(k)}} C_{Q^{(k)}} = -A^{(k)}/(g^{(k)})^2$. We define

$$\bar{A}^{(k)}_s = \sum_i \frac{Q^{(k)}_{is}}{g^{(k)}_s} A^{(k)}_{is}$$

so that

$$\nabla_{g^{(k)}}F = -\frac{\bar{A}^{(k)}}{g^{(k)}}.$$

The first order condition with respect to $g^{(k)}$ gives

$$\nabla_{g^{(k)}}\mathcal{L} = -\frac{\bar{A}^{(k)}}{g^{(k)}} + \alpha^{(k)} - \beta^{(k)} = 0.$$

In particular, $\beta^{(k)} = \alpha^{(k)} - \frac{\bar{A}^{(k)}}{g^{(k)}}$.

The gradient with respect to $Q^{(k)}$ becomes

$$\nabla_{Q^{(k)}}\mathcal{L} = C_{Q^{(k)}} - \mathbf{1}_{n_k}(\bar{C}^{(k)})^{\mathsf{T}} + \mathbf{1}_{n_k}(\alpha^{(k)})^{\mathsf{T}}$$

with $\bar{C}^{(k)} = \frac{\bar{A}^{(k)}}{g^{(k)}}$.

We then perform a KL mirror-descent step based on this Lagrangian gradient. Given the current factor $Q^{(k)}_t$, define

$$K^{(k)} = Q^{(k)}_t \odot \exp(-\eta \nabla_{Q^{(k)}}\mathcal{L}).$$

The marginal constraint $Q^{(k)}\mathbf{1}_{r_k} = a^{(k)}$ is imposed exactly through a KL projection, while the latent marginal is only semi-relaxed during this factor step. Writing $Q^{(k)}_{t+1} = \mathrm{diag}(u^{(k)})K^{(k)}\mathrm{diag}(v^{(k)})$, the row scaling is given by

$$u^{(k)} = \frac{a^{(k)}}{K^{(k)}v^{(k)}}.$$

The latent scaling is updated according to the semi-relaxed KL projection with parameter τ

$$v^{(k)} = \Big(\frac{g^{(k)}_t}{(K^{(k)})^{\mathsf{T}}u^{(k)}}\Big)^{\frac{\tau}{\tau+\eta^{-1}}}.$$

The resulting latent marginal is then allowed to change:

$$g_{t+1}^{(k)} = (Q_{t+1}^{(k)})^\mathsf{T} \mathbf{1}_{n_k}.$$

The factors are updated for all modes using the same current feasible pair $(\mathbf{Q}, \boldsymbol{\mathcal{G}})$. After all factor updates have been performed, the core is recomputed using the resulting latent marginals.

Step on $\boldsymbol{\mathcal{G}}$. Given the updated factors $\mathbf{P}$ and latent marginals $\mathbf{g}$, we update the core by solving the entropic MOT problem with cost

$$\boldsymbol{\mathcal{C}}_{\boldsymbol{\mathcal{G}}} = \boldsymbol{\mathcal{C}} \times_{k=1}^{d} P^{(k)\mathsf{T}}$$

and prescribed marginals $\mathbf{g}$. This is performed using the Sinkhorn MOT algorithm.

Algorithm 5 Tucker MOT - Mirror Descent

function TUCKER($\boldsymbol{\mathcal{C}}, \mathbf{a}, \varepsilon, \mathbf{r}, \tau$)
  Initialize $\mathbf{Q}, \mathbf{P}, \boldsymbol{\mathcal{G}}$
  for $t \geq 0$ **do** ▷ until convergence
    for $1 \leq k \leq d$ **do**
      $D^{(k)} \leftarrow C_{Q^{(k)}} - \mathbf{1}(\bar{C}^{(k)})^\mathsf{T} + \mathbf{1}(\alpha^{(k)})^\mathsf{T}$
      $Q^{(k)} \leftarrow$ SEMIRELAXED($Q^{(k)}$, $D^{(k)}$, $a^{(k)}$, τ)
      $g^{(k)} \leftarrow (Q^{(k)})^\mathsf{T} \mathbf{1}_{n_k}$
    end for
    Update $\mathbf{P}$
    $\boldsymbol{\mathcal{C}}_{\boldsymbol{\mathcal{G}}} \leftarrow \mathcal{C} \times_{k=1}^{d} (P^{(k)})^\mathsf{T}$
    $\boldsymbol{\mathcal{G}} \leftarrow$ SINKHORN-MOT($\boldsymbol{\mathcal{C}}_{\boldsymbol{\mathcal{G}}}$, $\mathbf{g}$, ε) ▷ or Greedy Sinkhorn MOT
  end for
  return $\mathbf{Q}, \mathbf{P}, \boldsymbol{\mathcal{G}}$
end function

F MIRROR DESCENT FOR CPMOT

In this Section, we write $\xi = (\mathbf{Q}, g)$ and $f(\xi)$ the objective from Eq. (5). Following Scetbon et al. (2021), we define the three sets

$$\mathcal{C}_1(\mathbf{a}, r, \alpha) = \{(\mathbf{Q}, g) \in \mathbb{R}_+^{n_1 r} \times \cdots \times \mathbb{R}_+^{n_d r} \times \mathbb{R}_+^r \mid \forall k,\ Q^{(k)} \mathbf{1}_r = a^{(k)}, g \geq \alpha\},$$

and

$$\mathcal{C}_2(r) = \{(\mathbf{Q}, g) \in \mathbb{R}_+^{n_1 r} \times \cdots \times \mathbb{R}_+^{n_d r} \times \mathbb{R}_+^r \mid \forall k,\ Q^{(k)\mathsf{T}} \mathbf{1}_{n_k} = g\},$$

and $\Xi = \mathcal{C}_1(\mathbf{a}, r, \alpha) \cap \mathcal{C}_2(r)$ so that adding a regularization term on g in Eq. (6) is equivalent to solving

$$\inf_{\xi \in \Xi} f(\xi) - \varepsilon h(\xi),$$

for α sufficiently small. We perform a KL mirror descent with step size η:

$$\begin{aligned} \xi_{t+1} &= \mathrm{Proj}_{\Xi}^{\mathrm{KL}}(\xi_t \odot \exp(-\eta[\nabla f(\xi_t) + \varepsilon \log(\xi_t)])) \\ &= \mathrm{Proj}_{\Xi}^{\mathrm{KL}}(\xi_t^{1-\eta\varepsilon} \odot \exp(-\eta \nabla f(\xi_t)) \end{aligned}$$

In order to perform the KL projection onto Ξ, we use the Dykstra algorithm with projection onto $\mathcal{C}_1(\mathbf{a}, r, \alpha)$ and $\mathcal{C}_2(r)$.

Gradient of f. We write $f(\xi) = \sum_{i_1,\ldots,i_d} \boldsymbol{\mathcal{C}}_{i_1,\ldots,i_d} \sum_{s=1}^{r} \frac{\prod_k Q_{i_k s}^{(k)}}{g_s^{d-1}}$.

For any k, j, s, we have

$$\frac{\partial f}{\partial Q_{js}^{(k)}} = \sum_{\substack{i_1,\ldots,i_d \\ i_k = j}} \boldsymbol{\mathcal{C}}_{i_1,\ldots,i_d} \frac{\prod_{\ell \neq k} Q_{i_\ell s}^{(\ell)}}{g_s^{d-1}} = \sum_{\substack{i_1,\ldots,i_d \\ i_k = j}} \boldsymbol{\mathcal{C}}_{i_1,\ldots,i_d} \prod_{\ell \neq k} p_{i_\ell s}^{(\ell)}.$$

Using $\frac{\partial f}{\partial Q_{js}^{(k)}} = u_{js}^{(k)}$ with $u_{.s}^{(k)} = \boldsymbol{\mathcal{C}} \times_{\ell \neq k} p_s^{(\ell)} \in \mathbb{R}^{n_k}$, we have

$$\nabla_{Q^{(k)}} f = \left(u_{.1}^{(k)}, \dots, u_{.r}^{(k)} \right) \in \mathbb{R}^{n_k r}.$$

Considering g,

$$\frac{\partial f}{\partial g_s} = -(d-1) \sum_{i_1,\dots,i_d} \boldsymbol{\mathcal{C}}_{i_1,\dots,i_d} \big(\prod_k Q_{i_k s}^{(k)} \big) / g_s^d = -(d-1)(\boldsymbol{\mathcal{C}} \times_1 p_s^{(1)} \times_2 \cdots \times_d p_s^{(d)})$$

Denoting $\mathbf{c} = \begin{pmatrix} \boldsymbol{\mathcal{C}} \times_1 p_1^{(1)} \cdots \times_d p_1^{(d)} \\ \vdots \\ \boldsymbol{\mathcal{C}} \times_1 p_r^{(1)} \cdots \times_d p_r^{(d)} \end{pmatrix}$, we have $\nabla_g f = -(d-1)\mathbf{c} \in \mathbb{R}^r$.

Projection onto $\mathcal{C}_1$. Each constraint in $\mathcal{C}_1$ is separable between the $Q^{(k)}$'s and g. For each k, we seek

$$\min_{Q \in \mathcal{C}_1^{(k)}} \mathrm{KL}(Q \| Q^{(k)}) = \min_{Q \in \mathcal{C}_1^{(k)}} \sum_{is} Q_{is}(\log Q_{is} / Q_{is}^{(k)} - 1).$$

The Lagrangian reads

$$\mathcal{L}(Q, \lambda) = \sum_{is} Q_{is}(\log Q_{is} / Q_{is}^{(k)} - 1) + \sum_i \lambda_i (\sum_s Q_{is} - a_i^{(k)}).$$

First order optimality conditions gives

$$\nabla_Q \mathcal{L} = \log Q / Q^{(k)} + \lambda \mathbf{1}_r^\mathsf{T} = \mathbf{0}_{n_k r} \implies Q^\star = \mathrm{diag}(e^{-\lambda}) Q^{(k)}$$

The $Q^\star \mathbf{1}_r = a^{(k)}$ condition enforces $e^{-\lambda} = a^{(k)} / Q^{(k)} \mathbf{1}_r$. Finally,

$$\mathrm{Proj}_{\mathcal{C}_1}^{\mathrm{KL}}(Q^{(k)}) = \mathrm{diag}(a^{(k)} / Q^{(k)} \mathbf{1}_r) Q^{(k)}$$

We also have $\mathrm{Proj}_{\mathcal{C}_1}^{\mathrm{KL}}(g) = \max(g, \alpha)$, as in Scetbon et al. (2021).

Projection onto $\mathcal{C}_2$. Denoting $\lambda_k \in \mathbb{R}^r$ associated with the constraint $Q^{(k)\mathsf{T}} \mathbf{1}_{n_k} = g$ in the Lagrangian, we obtain

$$\begin{cases} Q^{(k)\star} = Q^{(k)} \mathrm{diag}(e^{-\lambda_k}), \\ g^\star = g \odot e^{\sum_k \lambda_k}. \end{cases}$$

As $Q^{(k)\star\mathsf{T}} \mathbf{1}_{n_k} = g^\star$, we have $e^{-\lambda_k} = g^\star / Q^{(k)\mathsf{T}} \mathbf{1}_{n_k}$. Thus, $(g^\star)^{(d+1)} = g \bigodot_k Q^{(k)\mathsf{T}} \mathbf{1}_{n_k}$. Finally,

$$\mathrm{Proj}_{\mathcal{C}_2}^{\mathrm{KL}}(g) = \big(g \bigodot_{k=1}^{d} Q^{(k)\mathsf{T}} \mathbf{1}_{n_k} \big)^{\frac{1}{d+1}}, \qquad \mathrm{Proj}_{\mathcal{C}_2}^{\mathrm{KL}}(Q^{(k)}) = Q^{(k)} \mathrm{diag}(\mathrm{Proj}_{\mathcal{C}_2}^{\mathrm{KL}}(g) / Q^{(k)\mathsf{T}} \mathbf{1}_{n_k}).$$

We use the Dykstra algorithm to project onto $\Xi = \mathcal{C}_1 \cap \mathcal{C}_2$, as stated in Algorithm 6.

Algorithm 6 CPMOT - Mirror Descent

function CPMOT($\mathcal{C}$, $\mathbf{a}$,ε,r,α, η_t,δ)
  Initialize $(Q_0^{(k)})_k, g_0$
  for $t \geq 0$ **do** ▷ until convergence
    for $1 \leq k \leq d$ **do**
      $K_t^{(k)} \leftarrow (Q_t^{(k)})^{1-\eta\varepsilon} \odot \exp(-\eta \nabla_{Q_t^{(k)}} f(\xi_t))$ ▷ $\xi_t = (Q_t^{(1)}, \dots, Q_t^{(d)}, g_t)$
    end for
    $K_t^g \leftarrow g_t^{1-\eta\varepsilon} \odot \exp(-\eta \nabla_g(\xi_t))$
    $(Q_{t+1}^{(k)})_k, g_{t+1} \leftarrow$ DYKSTRA($(K_t^{(k)}, K_t^g), \alpha, \delta$) ▷ using projections onto $\mathcal{C}_1$ and $\mathcal{C}_2$
  end for
  return $(Q_{t+1}^{(k)})_k, g_{t+1}$
end function

Computational cost. If no assumption is made on the cost $\boldsymbol{C}$, computing all $u_{\cdot s}^{(k)}$'s has cost $\mathcal{O}(dr\prod_{k=1}^{d} n_k)$. This calls for the use of low-rank cost factorization, such as in Strössner & Kressner (2023), to reduce the computation of $u_{\cdot s}^{(k)}$ to $\mathcal{O}(dn_k r)$. In particular, for the class of pairwise costs we consider, we can use $u_{js}^{(k)} = (d-1)\|x_{i_k}^{(k)}\|_{\mathcal{H}}^2 + \sum_{\ell\neq k}\sum_{i_\ell} p_{i_\ell}^{(\ell)}\|x_{i_\ell}^{(\ell)}\|_{\mathcal{H}}^2 - 2\langle\varphi(x_{i_k}^{(k)}), \sum_{\ell\neq k}\sum_{i_\ell} p_{i_\ell}^{(\ell)}\varphi(x_{i_\ell}^{(\ell)})\rangle_{\mathcal{H}}$ reducing the overall cost to $\mathcal{O}(TDNr)$, where D is the average number of Dykstra iterations, and T the number of mirror descent steps, and $N = \sum_{k=1}^{d} n_k$.

G ALTERNATING MINIMIZATION FOR ECPMOT

Initialization As the optimal $\varphi(z_s^\star)$ is an arithmetic mean of the source hubs $\mu_s^{(k)}$, we can initialize the hub positions $\varphi(z)$ using the k-means++ initialization method (Arthur & Vassilvitskii, 2007) on the sets $\{\varphi(x_{i_k}^{(k)})\}_{i_k=1}^{n_k}$ for each k, and then take the mean of the resulting sets of centers. To compute this mean, we need to assign a point from each set to each index s. At the cost of breaking the symmetry, we use the following procedure, which requires $d-1$ calls to LINEARSUMASSIGNEMENT. For each $k > 1$, we get τ_k by solving LINEARSUMASSIGNEMENT$([c(\tilde{x}_s^{(1)}, \tilde{x}_{s'}^{(k)})]_{s,s'})$ and set

$$\varphi(z_s) = \frac{1}{d}\big(\varphi(\tilde{x}_s^{(1)} + \sum_{k=2}^{d}\varphi(\tilde{x}_{\tau_k(s)}^{(k)})\big).$$

Initialization sensitivity. To assess the effect of the initialization of $\varphi(z)$ on the results, Figure 8 show the distribution of the resulting transport cost for random and k-means-z initialization across 500 random seeds on the same problem. We consider 4 measures, 30 points each sampled from a standard 2-D Gaussian distribution, and take $r = 7$. We observe that k-means-z- initialization substantially reduces the overall runtimes while finding the best local minimum more frequently.

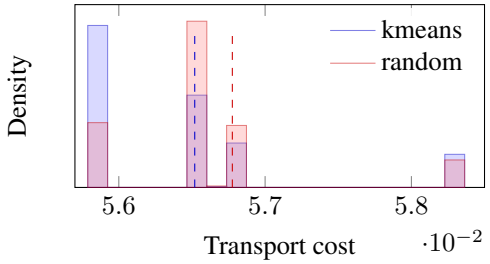


Figure 8: Distribution of the transport-cost (lower is better) across random seeds for different initialization methods. Average runtimes are 0.37 s and 0.47 s for k-means-z and random initialization, respectively, while the corresponding average values of the transport cost are 0.0565 and 0.0568.

Algorithm 7 k-means z initializer for d sets

function D-KMEANS-Z($(x^{(k)})_k, r$)
  for $1 \le k \le d$ **do**
    $\varphi(\tilde{x}^{(k)}) \leftarrow$ KMEANS++$((\varphi(x_{i_k}^{(k)}))_{i_k}, r)$
  end for
  for $2 \le k \le d$ **do**
    $C^{(k)} \leftarrow [\|\varphi(\tilde{x}_s^{(1)}) - \varphi(\tilde{x}_{s'}^{(k)})\|_{\mathcal{H}}^2]_{s,s'}$
    $\tau_k \leftarrow$ LINEARSUMASSIGNMENT$(C^{(k)})$
  end for
  $\varphi(z) \leftarrow \frac{1}{d}\left[\varphi(\tilde{x}_s^{(1)}) + \sum_{k=2}^{d}\varphi(\tilde{x}_{\tau_k(s)}^{(k)})\right]_s$
  return $\varphi(z)$
end function

Algorithm 8 Sinkhorn barycenter couplings Benamou et al. (2015)

function SINKHORNBARY($(C^{(k)})_k$, $\mathbf{a}$, ε, r)
  $d \leftarrow \mathrm{Card}(\{C^{(k)}\}_k)$
  for $1 \leq k \leq d$ **do**
    $K^{(k)} \leftarrow \exp\left(-C^{(k)}/\varepsilon\right)$
    $v^{(k)} \leftarrow \mathbf{1}_r$
  end for
  $g \leftarrow \frac{1}{r}\mathbf{1}_r$
  for $t \geq 0$ **do**
    for $1 \leq k \leq d$ **do**
      $u^{(k)} \leftarrow \dfrac{a^{(k)}}{K^{(k)}v^{(k)}}$
    end for
    $g \leftarrow \exp\left(\frac{1}{d}\sum_{k=1}^{d}\log(K^{(k)\mathsf{T}}u^{(k)})\right)$
    $g \leftarrow g/\mathbf{1}_r^{\mathsf{T}}g$
    for $1 \leq k \leq d$ **do**
      $v^{(k)} \leftarrow \dfrac{g}{K^{(k)\mathsf{T}}u^{(k)}}$
    end for
  end for
  for $t \geq 0$ **do**
    $Q^{(k)} \leftarrow \mathrm{diag}(u^{(k)})K^{(k)}\mathrm{diag}(v^{(k)})$
  end for
  return $\left(\mathbf{Q}, g\right)$
end function

Algorithm 9 Explicit CPMOT

function EXPCPMOT(r, $(x^{(k)})_k$, $\mathbf{a}$, ε)
  $\varphi(z) \leftarrow$ D-KMEANS-Z($(\varphi(x^{(k)}))_k, r$)
  for $1 \leq k \leq d$ **do**
    $C^{(k)} \leftarrow [\|\varphi(x_{i_k}^{(k)}) - \varphi(z_s)\|_{\mathcal{H}}^2]_{i_k s}$
  end for
  for $t \geq 0$ **do** ▷ until convergence
    $\mathbf{Q}, g \leftarrow$ SINKHORNBARY($(C^{(k)})_k, \mathbf{a}, \varepsilon/d$)
    $\varphi(z) \leftarrow [\frac{1}{dg_s}\sum_{k=1}^{d} Q_{i_k s}^{(k)}\varphi(x_{i_k s}^{(k)})]_{s=1}^{r}$
    for $1 \leq k \leq d$ **do**
      $C^{(k)} \leftarrow [\|\varphi(x_{i_k}^{(k)}) - \varphi(z_s)\|_{\mathcal{H}}^2]_{i_k s}$
    end for
  end for
  return $\mathbf{Q}, g$
end function

G.1 SCHEME CONVERGENCE

As stated in (Peyré, 2026) (Section 12.1), the regularization gives strict convexity for the objective in $(\mathbf{Q}, g)$, but it remains non-convex in $\varphi(z)$. Consequently, the joint problem is non-convex. The sequence of values produced by Algorithm 1 is non-increasing because the $(\mathbf{Q}, g)$ step, as well as the $\varphi(z)$ step, diminish the objective. By definition, the objective is below-bounded, and thus Algorithm 1 converges to a local minimum.

Computational cost. Denote $N = \sum_{k=1}^{d} n_k$. Each iteration of the inner SINKHORNBARY loop has a computational cost of $\mathcal{O}(Nr)$. Initialization, construction of the cost matrices $C^{(k)}$ and updating $\varphi(z)$ cost $\mathcal{O}(Nr)$ as well. Therefore, denoting T the average number of iterations of SINKHORNBARY, and S the number of iterations of the outer loop, the total computational cost is

$\mathcal{O}(TSNr)$. By comparison, the Sinkhorn algorithm 2 has a computational cost of $\mathcal{O}(Td\prod_{k=1}^{d} n_k)$, dominated by the computation of $\mathrm{Marg}_k(\boldsymbol{\mathcal{T}})$.

G.2 Explicit Unbalanced CP Multi-Marginal Optimal Transport

In the unbalanced setting (Beier et al., 2022), we allow the transport plan $\boldsymbol{\mathcal{T}}$ to not have its k-th marginal equal to $a^{(k)}$, but we add a penalization term to the objective to enforce $\mathrm{Marg}_k(\boldsymbol{\mathcal{T}}) \approx a^{(k)}$. Denoting the set of tensors of CP-rank at most r having positive entries $J_r = \{\boldsymbol{\mathcal{T}} \in \mathbb{R}_+^{n_1\times\cdots\times n_d} \mid \mathrm{rank}_{\mathrm{CP}}(\boldsymbol{\mathcal{T}}) \leq r\}$, the problem reads

$$\mathrm{UCPMOT}_r(\mathbf{a}) = \inf_{\boldsymbol{\mathcal{T}}\in J_r} \langle \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{T}}\rangle + \tau \sum_{k=1}^{d} \mathrm{KL}(\mathrm{Marg}_k(\boldsymbol{\mathcal{T}}) \| a^{(k)}). \tag{16}$$

Using the low-rank factorization of $\boldsymbol{\mathcal{T}}$, with $Q^{(k)} \in \mathbb{R}_+^{n_k r}$ satisfying $Q^{(k)\mathsf{T}}\mathbf{1}_{n_k} = g$, the problem (16) is equivalent to minimizing

$$\mathrm{UCPMOT}_r(\mathbf{a}) = \inf_{\mathbf{Q},g} \langle \mathcal{C}, \sum_{s=1}^{r} g_s \bigcirc_{k=1}^{d} p_s^{(k)}\rangle + \tau \sum_{k=1}^{d} \mathrm{KL}(Q^{(k)}\mathbf{1}_r \| a^{(k)}).$$

The proof of Theorem 4.1 relying on the $Q^{(k)}$'s having g as right marginal, it applies in the unbalanced case. Adding entropic regularization, the upper bound is

$$\inf_{\mathbf{Q},g} \langle \mathcal{C}, \sum_{s=1}^{r} g_s \bigcirc_{k=1}^{d} p_s^{(k)}\rangle + \tau \sum_{k=1}^{d} \mathrm{KL}(Q^{(k)}\mathbf{1}_r \| a^{(k)}) + \varepsilon h(\mathbf{Q}) \leq \inf_{\mathbf{Q},g,\varphi(z)} d\Big(\sum_{k=1}^{d} \langle C^{(k)}, Q^{(k)}\rangle - \frac{\varepsilon}{d}\mathcal{H}(Q^{(k)}) + \tau \mathrm{KL}(Q^{(k)}\mathbf{1}_r \| a^{(k)})\Big). \tag{17}$$

In order to solve Eq. (17), we can apply Algorithm 1 replacing SinkhornBary in line 7 by its unbalanced version. Theorem 4.2 also holds in the case.

H Classical bipartite low-rank optimal transport

In the bipartite classical case, the optimal transport setup involves two probability measures $a = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $b = \sum_{j=1}^{m} b_j \delta_{y_j} \in \mathcal{P}(\mathcal{X})$, and a cost matrix $C \in \mathbb{R}^{nm}$, such that $C_{ij} = c(x_i, y_j)$ where $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a cost function over pairs of points in $\mathcal{X}$. Solving optimal transport is finding a transport plan $T \in \mathbb{R}_+^{nm}$, where T_{ij} represents the amout of mass going from x_i to y_j, such that T satisfies marginal constraints, i.e.

$$T \in \Pi(a,b) = \{T \in \mathbb{R}_+^{nm} \mid T\mathbf{1}_m = a \text{ and } T^{\mathsf{T}}\mathbf{1}_n = b\},$$

and such that T minimizes the total transport cost. The problem reads

$$\mathrm{OT}(a,b,C) = \min_{T\in\Pi(a,b)} \langle C, T\rangle.$$

The low-rank optimal transport (LOT) problem is defined as

$$\mathrm{LOT}_r(a,b,C) = \min_{T\in\Pi_r(a,b)} \langle C, T\rangle \tag{18}$$

where $\Pi_r(a,b) = \{T \in \Pi(a,b) \mid \mathrm{rank}_+(T) \leq r\}$. Scetbon et al. (2021) introduce a factorization of the low-rank coupling T as a triple (Q,R,g) such that $g \in \Delta_r^*$, $Q \in \Pi(a,g)$, $R \in \Pi(b,g)$, and

$$T = Q\mathrm{diag}(1/g)R^{\mathsf{T}}, \qquad T_{ij} = \sum_{s=1}^{r} \frac{Q_{is}R_{js}}{g_s}.$$

The vector g represents weights of an implicit intermediate measure $\sum_{s=1}^{r} g_s\delta_{z_s}$ where the locations of the hubs z_s do not take part in the formulation of the problem. The transport from a to b can be seen as the transport from a to g and from g to b, with transport plans Q and R^{T}.

Thus, Scetbon et al. (2021) show that solving Eq. (18) amounts to solving

$$\min_{Q,R,g} \langle C, Q\mathrm{diag}(1/g)R^\mathsf{T}\rangle$$

where $(Q, R, g) \in \mathbb{R}_+^{nr} \times \mathbb{R}_+^{mr} \times (\mathbb{R}_+^*)^r$ such that $Q\mathbf{1}_r = a$, $R\mathbf{1}_r = b$ and $Q^\mathsf{T}\mathbf{1}_n = R^\mathsf{T}\mathbf{1}_m = g$. This problem can be regularized by solving

$$\min_{Q,R,g} \langle C, Q\mathrm{diag}(1/g)R^\mathsf{T}\rangle - \varepsilon h(Q, R, g) \tag{19}$$

where $h(Q, R, g) = \mathcal{H}(Q) + \mathcal{H}(R) + \mathcal{H}(g)$. Eq. (19) is solved using a Mirror Descent scheme.

Riemannian formulation. Jawanpuria & Mishra (2026) show that the space of low-rank factorizations $\mathcal{M} = \{(Q, R) \in (\mathbb{R}_+^*)^{nr} \times (\mathbb{R}_+^*)^{mr} \mid Q\mathbf{1}_r = a, R\mathbf{1}_r = b, Q^\mathsf{T}\mathbf{1}_n = R^\mathsf{T}\mathbf{1}_m\}$ is a non-empty embedded smooth manifold of $(\mathbb{R}_+^*)^{nr} \times (\mathbb{R}_+^*)^{mr}$ with dimension $(n+m-1)(r-1)$. Authors derive Riemannian ingredients to use second-order Riemannian algorithms such as Riemannian conjugate gradients and Riemannian trust regions (Boumal, 2023). This method does not require a regularization parameter ε.

Approximation Error. Scetbon & Cuturi (2022) provide an upper bound on the error made by restricting the transport plan to have rank r. We have for any $2 \leq r \leq \min(n, m)$:

$$|\mathrm{LOT}_r(a, b, C) - \mathrm{OT}(a, b, C)| \leq \|C\|_\infty \ln(\min(n, m)/(r-1)).$$

LOT as clustering. Scetbon & Cuturi (2022) state that LOT can be seen as a generalization of the k-means methods. Indeed, if we take $a = b$, then

$$\mathrm{LOT}_r(a, a, C) = \min_Q \langle C, Q\mathrm{diag}(1/Q^\mathsf{T}\mathbf{1}_n)Q^\mathsf{T}\rangle$$

where $Q\mathbf{1}_r = a$. We recover the k-means algorithm by taking $r = k$ and c the squared Euclidean distance. A co-clustering algorithm using this formulation is derived in Schmidt et al. (2026): authors also propose an approximation of the low-rank optimal transport by a generalized k-means and provide the transport clustering (TC) algorithm, which is specific to the $n = m$ case, and require solving a full-rank optimal transport problem as an initialization.

Computational cost. Algorithms considered in Appendix B.2 have computational cost ECPMOT: $\mathcal{O}(NT(n + m)r)$, CPMOT: $\mathcal{O}(NT(n + m)r)$, TUCKER: $\mathcal{O}(NLr^2(n + m))$, RCG: $\mathcal{O}(N(k_{\mathrm{CG}} + L)r(n + m))$, LATOT: $\mathcal{O}(N(8r^3 + T(n + m)r))$ where N is the number of iterations of the outer loop, and other constants are related to the number of iterations in inner loops. We refer to related articles for further details.