On the complexity of the optimal transport problem with graph-structured cost

Jiaojiao Fan* Georgia Tech
Isabel Haasler* KTH
Johan Karlsson KTH
Yongxin Chen Georgia Tech

Abstract

Multi-marginal optimal transport (MOT) is a generalization of optimal transport to multiple marginals. Optimal transport has evolved into an important tool in many machine learning applications, and its multi-marginal extension opens up for addressing new challenges in the field of machine learning. However, the usage of MOT has been largely impeded by its computational complexity which scales exponentially in the number of marginals. Fortunately, in many applications, such as barycenter or interpolation problems, the cost function adheres to structures, which has recently been exploited for developing efficient computational methods. In this work we derive computational bounds for these methods. In particular, with m marginal distributions supported on n points, we provide a $\tilde{\mathcal{O}}(d(\mathcal{T})mn^{w(G)+1}\epsilon^{-2})$ bound for a ϵ-accuracy when the problem is associated with a graph that can be factored as a junction tree with diameter $d(\mathcal{T})$ and tree-width $w(G)$. For the special case of the Wasserstein barycenter problem, which corresponds to a star-shaped tree, our bound is in alignment with the existing complexity bound for it.

1 Introduction

The history of optimal transport can be traced back to the 18-th century when the French mathematician Monge introduced this tool for his engineering projects. In optimal transport problems one seeks an optimal strategy to move resources from an initial distribution to a target one. This theory has initially had a tremendous impact to fields such as economics and logistics. During the last decades, with new efficient computational methods (Villani, 2009; Cuturi, 2013) and more available computational power, optimal transport theory has also been used for addressing a broad class of problems both within the machine learning community (Peyré et al., 2019; Solomon et al., 2014, 2015; Arjovsky et al., 2017), but also in related fields such as imaging (Haker et al., 2004) and systems and control (Chen et al., 2016).

Multi-marginal optimal transport (MOT) is a natural extension of standard optimal transport to scenarios with more than two marginal distributions. In the discrete setting, the objective of MOT is to find an optimal coupling between m marginals $\mu_1, \dots, \mu_m \in \mathbb{R}_+^n$ over X, where X is a discrete space with support in n points. A m-mode tensor $\mathbf{B} \in \mathbb{R}_+^{n^m}$ is a feasible transport plan if it satisfies the assigned marginals, $P_k(\mathbf{B}) = \mu_k$, where

$$[P_k(\mathbf{B})](x_k) = \sum_{\mathbf{x}\setminus x_k} \mathbf{B}(\mathbf{x}), \quad \text{for all } x_k \in X, \tag{1}$$

where $\mathbf{x} = (x_1, \dots, x_m) \in X^m$. In this paper we consider a version of this problem where the marginals are typically only imposed on a subset of the transport tensors nodes, and we denote this subset of indices by $\Gamma \subset \{1, \dots, m\}$. The set of feasible transport plans consistent with these marginals $\{\mu_k\}_{k\in\Gamma}$ is then

$$\Pi_\Gamma^m((\mu_k)_{k\in\Gamma}) = \{\mathbf{B} \in \mathbb{R}^{n^m} : P_k(\mathbf{B}) = \mu_k, \forall k \in \Gamma\}.$$

Given a non-negative cost tensor $\mathbf{C} \in \mathbb{R}_+^{n^m}$, where $\mathbf{C}(\mathbf{x})$ denotes the cost associated with a unit mass on the tuple $\mathbf{x}$, the multi-marginal optimal transport problem reads

$$\min_{\mathbf{B}\in\Pi_\Gamma^m((\mu_k)_{k\in\Gamma})} \langle \mathbf{C}, \mathbf{B}\rangle. \tag{2}$$

The MOT problem is a linear program, thus, in principle, the simplex algorithm can be used to solve it exactly. The complexity however explodes quickly as the problem size increases. In practice, the MOT is

* Equal contribution. Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

solved approximately instead. The goal of these approximation algorithms is to find $\widehat{\mathbf{B}} \in \Pi_\Gamma^m((\mu_k)_{k\in\Gamma})$ such that $\langle \mathbf{C}, \widehat{\mathbf{B}}\rangle$ is an ϵ-approximation of the MOT problem (2). That is, $\widehat{\mathbf{B}}$ is an approximation of the transport tensor and satisfies

$$\langle \mathbf{C}, \widehat{\mathbf{B}}\rangle \leq \min_{\mathbf{B}\in\Pi_\Gamma^m((\mu_k)_{k\in\Gamma})} \langle \mathbf{C}, \mathbf{B}\rangle + \epsilon.$$

A popular method to approximately solve the MOT problem (2) is to solve an entropic regularized version of it where an entropy barrier term is added to the objective. This regularized problem can be solved by the renowned Sinkhorn iterations (Deming & Stephan, 1940; Cuturi, 2013).

Related work: A fundamental question in the study of MOT algorithms is understanding their complexities, and several complexity bounds have been derived over the last few years for various MOT algorithms (Lin et al., 2019; Altschuler & Boix-Adsera, 2020; Carlier, 2021). The best known complexity bound for the general multi-marginal Sinkhorn iterations is $\tilde{\mathcal{O}}(\frac{m^3 n^m}{\epsilon^2})$ (Lin et al., 2019) with greedy updates, which scales exponentially in the number of marginals m. This is not surprising as the size of the variable $\mathbf{B}$ grows exponentially. This complexity bound can be improved by exploiting the structure of the cost tensor $\mathbf{C}$. A well-known example is the Wasserstein barycenter problem where the cost can be decomposed into pairwise costs between the marginals and the barycenter. Kroshnin et al. (2019) shows that the iterative scaling algorithm finds an ϵ-approximate solution to the barycenter between L distributions in $\tilde{\mathcal{O}}(\frac{Ln^2}{\epsilon^2})$ operations. A more general class of costs where better computation complexity can be achieved is associated with the tree structure (see Section 2). Such structures appear in various applications, such as barycenter problems (Lin et al., 2020; Kroshnin et al., 2019), interpolation problems (Solomon et al., 2015), and estimation problems (Elvander et al., 2020). It was shown in Haasler et al. (2021c) that a complexity bound for MOT problems with tree-structured cost (including the barycenter problem as a special case) is $\tilde{\mathcal{O}}(\frac{m^4 n^2}{\epsilon^2})$, where m denotes the number of marginals. Many other MOT problems are structured according to graphs that contain cycles, e.g., in the generalized Euler flow problem (Benamou et al., 2015), control applications (Haasler et al., 2020), and multi-species problems (Haasler et al., 2021b). Tree-structured optimal transport problems are often formulated as a sum of bi-marginal optimal transport problems and in previous works the numerical scheme is often based on regularizing each of the bi-marginal problems locally. However, if the underlying graph structure contains cycles, there is no such representation of the problem. In Altschuler & Boix-Adsera (2020), it was shown that the complexity for MOT with general graph-structured cost scales polynomially as the number of marginals increases, as long as the tree-width of the graph is properly bounded, but they do not provide explicit dependencies on the parameters. Note that some other structures of the cost tensors such as the low rank property can be leveraged (Altschuler & Boix-Adsera, 2020), but these are very different to the graphical structure considered in this work.

Our contribution: The purpose of this work is to provide a tighter complexity bound for solving the MOT problem with general graph-structured costs. For the cases where the MOT problem is structured according to a tree, i.e., the graph does not contain any cycles, we show that an ϵ-approximation of the solution can be found within $\tilde{\mathcal{O}}(\bar{d}(G)mn^2\epsilon^{-2})$ operations, where $\bar{d}(G)$ denotes the average distance between two leaves of the tree. This improves on the previous result $\tilde{\mathcal{O}}(m^4n^2\epsilon^{-2})$ for tree-structured MOT in Haasler et al. (2021c). For the barycenter problem, which corresponds to the special case of a star-shaped graph, this matches the best known bound when no further acceleration of the method is applied. The framework in this paper also treats a class of MOT problems that is much larger than what can be described by bi-marginal OT problems. In the case of a general graph G, the complexity is $\tilde{\mathcal{O}}\left(\bar{d}(\mathcal{T})mn^{w(G)+1}\epsilon^{-2}\right)$, where $\mathcal{T}$ is a minimal junction tree over the graph G, and $w(G)$ is the tree-width of G. Our contribution can be summarized as follows:

i) By a novel analysis of the method, leveraging a random update scheme, we improve on the best known complexity result (cf. Table 1).

ii) By using a novel regularization term we simplify the complexity analysis (cf. Remark 1).

iii) We augment the Sinkhorn belief propagation method (Algorithm 1, see also Haasler et al. (2021c, Algorithm 2)) by a new rounding scheme for tree-structured MOT, given in Algorithm 2.

iv) We extend the analysis to MOT problems with general graph-structured costs.

We remark that our Algorithm 1 is very similar to those in Haasler et al. (2021c); Altschuler & Boix-Adsera (2020); the main results are on the explicit complexity bounds. There are accelerated versions of the Sinkhorn algorithm, see, e.g., Lin et al. (2019); Kroshnin et al. (2019) that can improve the dependence with respect to ϵ from ϵ^{-2} to ϵ^{-1}. Note that these accelerations cannot improve the dependence over m or n. Since the algorithm studied in this work is not accelerated, we compare the complexity bounds only to algorithms with no acceleration.

Table 1: Best-known complexity bounds for optimal transport without acceleration. Note that our bounds hold with high probability.

Problem	Complexity	Paper
Bi-marginal optimal transport	$\tilde{\mathcal{O}}(n^2\epsilon^{-2})$	Dvurechensky et al. (2018)
Barycenter optimal transport	$\tilde{\mathcal{O}}(mn^2\epsilon^{-2})$	Kroshnin et al. (2019)
General MOT	$\tilde{\mathcal{O}}(m^3n^m\epsilon^{-2})$	Lin et al. (2019)
Tree-structured MOT	$\tilde{\mathcal{O}}(m^4n^2\epsilon^{-2})$	Haasler et al. (2021c)
Tree-structured MOT	$\tilde{\mathcal{O}}(d(G)mn^2\epsilon^{-2})$	Ours
balanced Graph-structured MOT	$\tilde{\mathcal{O}}(d(\mathcal{T})mn^{w(G)+1}\epsilon^{-2})$	Ours

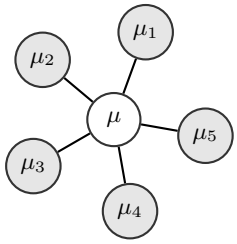


Figure 1: Graph associated with a barycenter problem (3), where $L = 5$.

Notation: For a matrix $C \in \mathbb{R}^{n\times n}$, we denote $\|C\|_\infty$ its largest element. We denote a graph as the tuple $G = (V, E)$, where V is the set of vertices, and E is the set of edges. For a vertex $k \in V$, we denote the set of neighbouring vertices by $N(k) \subset V$. Let $\mathbf{1}_d$ denote the all-ones vector/matrix/tensor in $\mathbb{R}^d$, and let $\exp(\cdot), \log(\cdot), \odot$, and $./$ denote the element-wise exponential, logarithm, multiplication, and division of tensors, respectively. The $p(m, n, \epsilon) = \tilde{\mathcal{O}}(q(m, n, \epsilon))$ notation absorbs polylogarithmic factors related to n, i.e., there exist positive constants c_2, c_3 such that $p(m, n, \epsilon) \leq c_2 q(m, n, \epsilon)(\log n)^{c_3}$.

2 Graph-structured MOT

In this paper we consider MOT problems with a cost that decouples according to a graph. Such structures appear in many applications, for instance in barycenter problems (Lin et al., 2020; Kroshnin et al., 2019), interpolation problems (Solomon et al., 2015), and estimation problems (Elvander et al., 2020; Singh et al., 2020). In fact, one of the very first studies of MOT, on the generalized Euler-flow problem, has a graph-structured cost (Brenier, 1989; Benamou et al., 2015).

Example 1. *(Fixed-support Wasserstein Barycenter).*

A special case of a graph-structured optimal transport problem is the fixed support barycenter problem (Agueh & Carlier, 2011) with uniform weights

$$\min_{\mu\in\mathbb{R}^n} \sum_{\ell=1}^{L} \frac{1}{L} W(\mu, \mu_\ell), \quad \textit{with} \quad W(\mu,\nu) = \inf_{B\in\Pi(\mu,\nu)} \langle C, B\rangle, \tag{3}$$

where $\Pi(\mu, \nu) = \Pi^2_{\{1,2\}}(\mu, \nu)$ denotes the standard set of feasible transport plans for two marginals. The underlying structure can be described by a star-graph as illustrated in Figure 1. Problem (3) *can be written as the multi-marginal problem* (2)*, where the cost tensor*

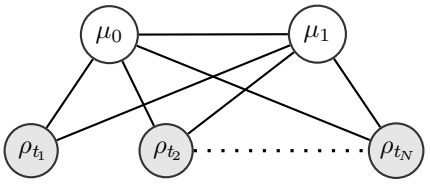


Figure 2: Graph associated with a Wasserstein least square problem (5).

$\mathbf{C} \in \mathbb{R}^{n^{L+1}}$ *is defined as*

$$\mathbf{C}(x_1, \ldots, x_L, x_{L+1}) = \sum_{\ell=1}^{L} \frac{1}{L} C(x_{L+1}, x_\ell), \tag{4}$$

and constraints are given on the set $\Gamma = \{1, \ldots, \ell\}$.

Example 2. *(Wasserstein least squares problem in Karimi et al. (2020)) Given a a set of probability measures $\rho_{t_1}, \rho_{t_2}, \ldots, \rho_{t_N}$ with $0 \leq t_1 < t_2 < \cdots < t_N \leq 1$, we aim to find two distributions μ_0 and μ_1 such that the set of measures is closest to the displacement interpolation between μ_0 and μ_1 in terms of the Wasserstein distance. More specifically, we seek the solution to*

$$\min_{\mu_0,\mu_1} \sum_{i=1}^{N} W(\rho_{t_i}, \mu_{t_i})^2 \tag{5}$$

where μ_t denotes the displacement interpolation connecting μ_0 and μ_1. The structure of this problem is described by the graph in Figure 2. By associating μ_0 and μ_1 with the first and last marginal, respectively, problem (5) *can be reformulated as a MOT problem* (2) *where constraints are given on the set $\Gamma = \{1, \ldots, N\}$ and the cost tensor is defined as*

$$\mathbf{C}(x_0, x_1, \ldots, x_{N+1}) = \sum_{i=1}^{N} \|x_i - (1-t_i)x_0 - t_i x_{N+1}\|^2 + \alpha\|x_0 - x_{N+1}\|^2,$$

where a regularization term $\alpha\|x_0 - x_{N+1}\|^2$ is added to promote short interpolation. Note that this cost can

be written in terms of pairwise interactions between marginals, which are represented by the edges in the graph in Figure 2.

Similar to Example 1 and 2, we can define a MOT problem that is structured according to any graph $G = (V, E)$. Therefore, we associate each vertex in V with a marginal of the transport plan $\mathbf{B}$, and each edge in E with a pair-wise cost. That is, for the interaction between vertices k_1 and k_2 we define a cost matrix $C^{(k_1,k_2)}$, and we let E be the set of all these pair-wise interactions. Then the graph-structured cost tensor is defined by

$$\mathbf{C}(x_1, \dots, x_m) = \sum_{(k_1,k_2)\in E} C^{(k_1,k_2)}(x_{k_1}, x_{k_2}). \tag{6}$$

Problem (2) with a cost tensor of the form (6) is called a graph-structured MOT problem (Haasler et al., 2021a,c).

Many graph-structured optimal transport problems, for instance interpolation and barycenter problems, are naturally described by tree graphs, i.e., graphs that do not contain any cycles. It is important to note that every tree-structured MOT problem can equivalently be expressed as a sum of coupled bi-marginal OT problems. However, general graph-structured MOT problems contain more information and *cannot* be represented as a sum of coupled OT problems, cf. Haasler et al. (2021a, Remark 3). Even in the case of tree-structured problems, the multi-marginal regularization is often favorable, because it yields less smoothed out solutions and is stable for smaller regularization parameters η (Haasler et al., 2021a, Section 5). In this work we utilize the fact that any graph can be converted into a tree using the junction tree technique (Koller & Friedman, 2009, Chapter 10), and we will use this representation to derive complexity bounds for general graph-structured MOT problems. It should be noted that in the case of a tree-structured MOT problem we can without loss of generality consider the case, where Γ is the set of leaves (Haasler et al., 2021a, Proposition 3.4).

3 Sinkhorn belief propagation algorithm

In practical applications the MOT problem is often prohibitively large for standard linear programming solvers, and one therefore has to resort to numerical methods to obtain an appropriate solution. A well-known approach, based on the seminal work by Cuturi (2013), is to regularize the objective in (2) with an entropic barrier term (Benamou et al., 2015). In particular, we introduce the barrier term

$$H(\mathbf{B} \mid \mathbf{M}) = \langle \mathbf{B}, \log(\mathbf{B}) - \log(\mathbf{M}) - \mathbf{1}_{n^m} \rangle,$$

where

$$\mathbf{M}(x_1, x_2, \dots, x_m) = \prod_{k\in\Gamma} \mu_k(x_k).$$

The regularized MOT problem reads then

$$\min_{\mathbf{B}\in\Pi_\Gamma^m((\mu_k)_{k\in\Gamma})} \langle \mathbf{C}, \mathbf{B} \rangle + \eta H(\mathbf{B} \mid \mathbf{M}), \tag{7}$$

where $\eta > 0$ is a small regularization parameter.

Remark 1. *Note that our choice of entropy regularizer is slightly different from the standard one $\langle \mathbf{B}, \log(\mathbf{B}) - \mathbf{1}_{n^m} \rangle$ often used for the Sinkhorn algorithm. The extra term $-\langle \mathbf{B}, \log(\mathbf{M}) \rangle$ turns out to simplify the approximation procedure (there is no need to alter the marginal distributions first to increase the minimum value of their elements as in Dvurechensky et al. (2018); Lin et al. (2019)) and the complexity analysis (see, e.g., Lemma 1).*

The optimal solution of the regularized multi-marginal optimal transport problem (7) can be compactly expressed in terms of the optimal variables of the dual problem. More precisely, the optimal transport tensor is of the form

$$[\mathbf{B}(\Lambda)](x_1, \dots, x_m) = \exp\left(-\mathbf{C}(x_1, \dots, x_m)/\eta\right) \cdot \prod_{k\in\Gamma} \Big(\exp\Big(\frac{\lambda_k(x_k)}{\eta}\Big) \mu_k(x_k) \Big), \tag{8}$$

where $\Lambda = \{\lambda_k\}_{k\in\Gamma}$ is the optimal solution of the dual of (7), which is given by (cf. Haasler et al. (2021a))

$$\min_{\Lambda} \psi(\Lambda) := \eta P(\mathbf{B}(\Lambda)) - \sum_{k\in\Gamma} \mu_k^{\mathrm{T}} \lambda_k. \tag{9}$$

Here, $P(\mathbf{B}) = \sum_{\mathbf{x}} \mathbf{B}(\mathbf{x}) \in \mathbb{R}$ is the projection over all marginals of $\mathbf{B}$, i.e., the sum over all elements.

The optimal solution to (9) can be efficiently found by the renowned Sinkhorn iterations (Benamou et al., 2015; Haasler et al., 2021c). In particular, the multi-marginal Sinkhorn algorithm is to find the scaled variables $u_k = \exp(\lambda_k/\eta)$, for $k \in \Gamma$, by iteratively updating them according to

$$u_k^{(t+1)} \leftarrow u_k^{(t)} \odot \mu_k ./ P_k(\mathbf{B}(\Lambda^{(t)})). \tag{10}$$

There are several approaches to perform these updates: At each iteration, the next marginal $k \in \Gamma$ to be updated can be picked in a random, cyclic, or greedy fashion (Benamou et al., 2015; Lin et al., 2019). In this paper we discuss the random updating rule. The greedy update requires more operations for each iteration as all the projections for $k \in \Gamma$ are needed for

an update. The traditional cyclic update introduces strong couplings between updates which makes the complexity analysis much more challenging.

For general MOT, computing the projections $P_k(\mathbf{B}(\Lambda^{(t)}))$ requires $\mathcal{O}(n^m)$ operations, which creates a large computational burden. However, in case the MOT problem has a tree-structure, the projections $P_k(\mathbf{B}(\Lambda^{(t)}))$ can be computed by a message-passing algorithm that utilizes the belief propagation algorithm (Yedidia et al., 2003), as described in Haasler et al. (2021c,a). This requires only matrix-vector multiplications of size n. In particular, the projections are of the form

$$[P_k(\mathbf{B}(\Lambda^{(t)}))](x_k) = \begin{cases} u_k^{(t)}(x_k)\mu_k(x_k)m_{\ell_k\to k}(x_k), & \text{if } k \in \Gamma \\ \prod_{\ell\in N(k)} m_{\ell\to k}(x_k), & \text{if } k \notin \Gamma, \end{cases} \quad (11)$$

where the messages are computed as

$$m_{\ell\to k}(x_k) = \begin{cases} \sum_{x_\ell} K^{(k,\ell)}(x_k,x_\ell) \prod_{j\in N(\ell)\setminus k} m_{j\to\ell}(x_\ell), & \text{if } \ell \notin \Gamma \\ \sum_{x_\ell} K^{(k,\ell)}(x_k,x_\ell) u_\ell^{(t)}(x_\ell)\mu_\ell(x_\ell), & \text{if } \ell \in \Gamma, \end{cases} \quad (12)$$

where $K^{(k,\ell)}(x_k,x_\ell) = \exp(-C^{(k,\ell)}(x_k,x_\ell)/\eta)$.

Since we can without loss of generality assume that Γ is the set of leaves of the tree, each vertex $k \in \Gamma$ has a unique neighbour $\ell_k \in N(k)$. The Sinkhorn iterations (10) with the projections (11) thus read

$$u_k^{(t+1)}(x_k) \leftarrow (m_{\ell_k\to k}(x_k))^{-1}.$$

Note that when we update the scaling vectors $u_{k^{(t)}}^{(t)}$ and in the previous iteration updated $u_{k^{(t-1)}}^{(t-1)}$ it is only required to recompute the messages between $k^{(t-1)}$ and $k^{(t)}$ (Haasler et al., 2021c; Singh et al., 2020). The Sinkhorn method is summarized in Algorithm 1. Here, we apply a random updating scheme, where the next scaling vector to be updated is picked from a uniform distribution of the remaining scaling vectors, except the previous one. Other common update rules for the Sinkhorn iterations, such as cyclic or greedy updates, can be obtained by simply changing the selection of $k^{(t)}$ in Algorithm 1.

From the scaling vectors $\{u_k\}_{k\in\Gamma}$ that are returned from Algorithm 1 we can construct the transport tensor $\widetilde{\mathbf{B}}$ as in (8). However, this tensor is not guaranteed to lie in the feasible set $\Pi_\Gamma^m((\mu_k)_{k\in\Gamma})$, and thus a rounding step is needed. Algorithm 2 describes a novel rounding scheme for tree-structured MOT that is based on the rounding for bi-marginal optimal transport in Altschuler et al. (2017, Algorithm 2).

Algorithm 1 SINKHORN_BP($\epsilon', \{\mu_k\}_{k\in\Gamma}, \mathbf{C}, \eta$)

Initialization: $u_k^{(0)} = \mathbf{1} \in \mathbb{R}^n$, for $k \in \Gamma$; $t = 1$; $k^{(0)} \in \Gamma$

while $\sum_{k\in\Gamma} \|P_k(\mathbf{B}(\Lambda^{(t)})) - \mu_k\|_1 \geq \epsilon'$ **do**

1. Randomly pick $k^{(t)} \in \Gamma \setminus k^{(t-1)}$
2. Update messages $m_{k_1\to k_2}$ according to (12) on the path from $k^{(t-1)}$ to $k^{(t)}$
3. Update $u_k^{(t+1)}(x_k)$ to be

$$\begin{cases} (m_{\ell_k\to k}(x_k))^{-1}, & \text{for } k = k^{(t)}, \text{ and } \ell_k \in N(k), \\ u_k^{(t)}(x_k), & \text{for } k \in \Gamma \setminus k^{(t)}, \end{cases}$$

4. $t \leftarrow t+1$

end while

Output: $u_k^{(t+1)}$, $k \in \Gamma$

Algorithm 2 ROUND ($\mathbf{B}, \{\mu_k\}_{k\in\Gamma}$)

Initialization: $\mathbf{B}_{k,\ell_k} = P_{k,\ell_k}(\mathbf{B}) \in \mathbb{R}^{n\times n}$ for all $k \in \Gamma$ and each $\ell_k \in N(k)$

for $k \in \Gamma$ **do**

Input $(\mathbf{B}_{k,\ell_k}; P_{\ell_k}(\mathbf{B}), \mu_k)$ into (Altschuler et al., 2017, Algorithm 2) and get $\widehat{\mathbf{B}}_{k,\ell_k}$ such that $\widehat{\mathbf{B}}_{k,\ell_k} \in \Pi(P_{\ell_k}(\mathbf{B}), \mu_k)$

end for

Output: $\widehat{\mathbf{B}} = \{\widehat{\mathbf{B}}_{k,\ell_k}; k \in \Gamma\} \cup \{P_{k_1,k_2}(\mathbf{B}); (k_1,k_2) \in E, k_1, k_2 \notin \Gamma\}$

Note that a transport tensor that solves a graph-structured MOT problem (2) or (7) has the same tree-structure as $\mathbf{C}$ (or more precisely $\exp(-\mathbf{C}/\eta)$, see (8)) and is thus fully determined by the projections $P_{k_1,k_2}(\mathbf{B})$ on the edges $(k_1,k_2) \in E$ (Koller & Friedman, 2009), which are given by

$$[P_{k_1,k_2}(\mathbf{B})](x_{k_1},x_{k_2}) = \sum_{\mathbf{x}\setminus\{x_{k_1},x_{k_2}\}} \mathbf{B}(\mathbf{x}). \quad (13)$$

It is therefore not necessary to construct the full tensor $\mathbf{B}$, which would be computationally and memory-wise expensive. Instead it suffices to give the transport matrices for the edges as input and output to the rounding scheme Algorithm 2. By slight abuse of notation, we let $\mathbf{B}((B_{k_1,k_2})_{(k_1,k_2)\in E})$ denote this tensor that decouples according to the tree structure G and satisfies the projections $[P_{k_1,k_2}(\mathbf{B})] = B_{k_1,k_2}$ for $(k_1,k_2) \in E$ (Koller & Friedman, 2009). Note that the projections (13) can be cheaply computed from the scaling vectors $\{u_k\}_{k\in\Gamma}$ as described in Haasler et al. (2021c, Theorem 4). The full method for finding an ϵ-approximate solution to a tree-structured MOT problem is summarized in Algorithm 3.

Algorithm 3 ϵ-approximation of tree-structured MOT

$\eta \leftarrow \frac{\epsilon}{2m\log(n)}$; $\epsilon' \leftarrow \frac{\epsilon}{8R_C^\Gamma}$.
$\{u_k\}_{k\in\Gamma} \leftarrow$ SINKHORN_BP$(\epsilon', \{\mu_k\}_{k\in\Gamma}, \mathbf{C}, \eta)$. (Algorithm 1)
Construct $\widetilde{\mathbf{B}}((B_{k_1,k_2})_{(k_1,k_2)\in E})$ from $\{u_k\}_{k\in\Gamma}$.
$\widehat{\mathbf{B}} \leftarrow$ ROUND$(\widetilde{\mathbf{B}}, \{\mu_k\}_{k\in\Gamma})$. (Algorithm 2)
Output: $\widehat{\mathbf{B}}$

4 Tree structured MOT analysis

In this section, we present a complexity bound for the Sinkhorn belief propagation algorithm for solving MOT problems with tree-structured costs. We first provide a few technical lemmas that will be used in the proof. The proofs of all the supporting lemmas are given in the supplementary material. The first result provides bounds for the scaling vector iterates.

Lemma 1. *Let $\lambda_k = \eta\log(u_k)$, where u_k are generated by Algorithm 1. Let $\Lambda^* = \{\lambda_k^*\}_{k\in\Gamma}$ be a solution of (9). Then for each $k \in \Gamma$ it holds*

$$\max_{x_k}\lambda_k(x_k) - \min_{x_k}\lambda_k(x_k) \le R_C^k,$$
$$\max_{x_k}\lambda_k^*(x_k) - \min_{x_k}\lambda_k^*(x_k) \le R_C^k,$$

where

$$R_C^k := \|C^{(k,\ell_k)}\|_\infty,$$

and where $\ell_k \in N(k)$ is the (unique) neighbour of k.

Note that we use a novel regularization that was suggested in Marino & Gerolin (2020) and Carlier (2021), see Remark 1. This yields the improvement of the result in Lemma 1 compared to similar results in Dvurechensky et al. (2018); Lin et al. (2019) and slightly simplifies the following analysis.

The following Lemma relates the error in the dual objective value to the stopping criterion of Algorithm 1.

Lemma 2. *Let $\Lambda = \{\lambda_k\}_{k\in\Gamma}$, where $\lambda_k = \eta\log(u_k)$ and u_k are generated by Algorithm 1, and let $\Lambda^* = \{\lambda_k^*\}_{k\in\Gamma}$ be a solution to (9). Then it holds*

$$\psi(\Lambda) - \psi(\Lambda^*) \le R_C^\Gamma \sum_{k\in\Gamma}\|P_k(\mathbf{B}(\Lambda)) - \mu_k\|_1,$$

with $R_C^\Gamma = \max_{k\in\Gamma} R_C^k$, where R_C^k is defined as in Lemma 1.

The increment between two sequential Sinkhorn iterates is related to the stopping criterion of Algorithm 1 as described in the following.

Lemma 3. *For any $\Lambda^{(t)}$, let $\Lambda^{(t+1)}$ be the next iterate of the algorithm in (10). Then,*

$$\mathbb{E}\left[\psi(\Lambda^{(t)}) - \psi(\Lambda^{(t+1)})\right] \ge \frac{\eta}{2|\Gamma|^2}(e_t)^2$$

with

$$e_t := \sum_{k\in\Gamma}\|P_k(\mathbf{B}(\Lambda^{(t)})) - \mu_k\|_1.$$

The expectation is over the uniform distribution of $k^{(t+1)} \in \Gamma \setminus k^{(t)}$.

We are now ready to state our first main result, which gives two probabilistic bounds on the required number of iterations in Algorithm 1.

Theorem 1. *For sufficiently small η, Algorithm 1 generates a tensor $\mathbf{B}(\Lambda^{(t)})$ satisfying*

$$\sum_{k\in\Gamma}\|P_k(\mathbf{B}(\Lambda^{(t)})) - \mu_k\|_1 \le \epsilon',$$

within τ iterations, where

$$\mathbb{E}[\tau] \le \frac{8|\Gamma|^2 R_C^\Gamma}{\eta\epsilon'}.$$

Moreover, for any $\delta \in (0, 0.5)$, it holds that

$$\mathbb{P}\left(\tau \le \frac{48|\Gamma|^2 R_C^\Gamma}{\eta\epsilon'}\log\frac{1}{\delta}\right) \ge 1-\delta.$$

Proof sketch (see supplementary material for details). Define the stopping time $\tau := \min\{t : e_t \le \epsilon'\}$. Let $\{\mathcal{F}_t := \sigma\left(\Lambda^{(1)}, \dots, \Lambda^{(t)}\right)\}_t$ be the natural filtration. By Lemma 2 and Lemma 3,

$$\mathbb{E}\left[\psi(\Lambda^{(t)}) - \psi(\Lambda^{(t+1)})|\mathcal{F}_t, t<\tau\right] \ge \frac{\eta}{2|\Gamma|^2}\left(\max\left\{\frac{\psi(\Lambda^{(t)}) - \psi(\Lambda^*)}{R_C^\Gamma}, \epsilon'\right\}\right)^2,$$

Let τ_1 be the first iteration when $\psi(\Lambda^{(t)}) - \psi(\Lambda^*) \le R_C^\Gamma\epsilon'$ and $\tau_2 := \tau - \tau_1 \ge 0$. We can bound τ_1 and τ_2 as

$$\mathbb{E}[\tau_1] \le \frac{6|\Gamma|^2 R_C^\Gamma}{\eta\epsilon'} - 1, \text{ and } \mathbb{E}[\tau_2] \le \frac{2|\Gamma|^2 R_C^\Gamma}{\eta\epsilon'} + 1.$$

Summing up the two bounds results in the bound for $\mathbb{E}[\tau]$. The bound in probability follows similarly. □

Remark 2. *Many Sinkhorn methods for MOT use a greedy update rule. In principle, such an update rule could be applied to our method to get a deterministic complexity bound. However, this would increase the complexity of our method by the factor $|\Gamma|$. High probability bounds are often used in machine learning algorithms when randomness is involved. Due to the logarithmic dependence $\log(1/\delta)$ in terms of the probability $1-\delta$, the high probability bound can safely be used as a surrogate of the deterministic bound. We can claim that $\tau \le \mathcal{O}(|\Gamma|^2 R_C^\Gamma(\eta\epsilon')^{-1})$ with arbitrarily high probability.*

In order to provide the complexity on the full method in Algorithm 3 we need the following two lemmas, which deal with the rounding method in Algorithm 2.

Lemma 4. *Let $\mathbf{B} \in \mathbb{R}^{n^m}$, where $m \geq 3$, be a non-negative m-mode tensor and $\{\mu_k\}_{k\in\Gamma}$ be a sequence of probability vectors, Algorithm 2 returns $\widehat{\mathbf{B}}$ satisfying $P_k(\mathbf{B}) = P_k(\widehat{\mathbf{B}})$, for $k \in \Gamma$, and $P_k(\widehat{\mathbf{B}}) = \mu_k$, for $k \in \Gamma$. Moreover, it holds that*

$$\langle \mathbf{C}, \mathbf{B}\rangle - \langle \mathbf{C}, \widehat{\mathbf{B}}\rangle \leq 2\sum_{k\in\Gamma} \|C^{(k,\ell_k)}\|_\infty \|\mu_k - P_k(\mathbf{B})\|_1,$$

where ℓ_k is the unique neighbour of k, for each $k \in \Gamma$.

Lemma 5. *Let $\widetilde{\mathbf{B}}$ be the output of Algorithm 1, let $\widehat{\mathbf{B}}$ be the output of Algorithm 2 with input $(\widetilde{\mathbf{B}}, \{\mu_k\})$, and let $\mathbf{B}^*$ denote the optimal solution to the unregularized MOT problem (2). Then it holds that*

$$\langle \mathbf{C}, \widehat{\mathbf{B}}\rangle - \langle \mathbf{C}, \mathbf{B}^*\rangle \leq m\eta\log(n) + 4\sum_{k\in\Gamma} \|C^{(k,\ell_k)}\|_\infty \|\mu_k - P_k(\widetilde{\mathbf{B}})\|_1.$$

We now have the tools to state our new complexity bound for finding ϵ-approximate solutions to tree-structured MOT problems. Denote by $d(G)$ the maximum distance of two nodes in the graph G.

Theorem 2. *Algorithm 3 finds an ϵ-approximate solution to the tree-structured MOT problem (2) in T arithmetic operations, where*

$$\mathbb{E}[T] = \mathcal{O}\left(\frac{d(G)m|\Gamma|^2 n^2 (R_C^\Gamma)^2 \log(n)}{\epsilon^2}\right).$$

Moreover, for all $\delta \in (0, 0.5)$ it holds that

$$\mathbb{P}\left(T \leq \frac{cd(G)m|\Gamma|^2 n^2 (R_C^\Gamma)^2 \log(n)\log(1/\delta)}{\epsilon^2}\right) \geq 1 - \delta$$

where c is a universal constant.

Proof. With the specific choices $\eta = \frac{\epsilon}{2m\log(n)}$ and $\epsilon' = \frac{\epsilon}{8R_C^\Gamma}$ we get $\langle \mathbf{C}, \widehat{\mathbf{B}}\rangle - \langle \mathbf{C}, \mathbf{B}^*\rangle \leq \epsilon$. By Theorem 1, the stopping time τ satisfies

$$\mathbb{E}[\tau] = \frac{8|\Gamma|^2 R_C^\Gamma}{\eta\epsilon'} = \mathcal{O}\left(\frac{m|\Gamma|^2 (R_C^\Gamma)^2 \log(n)}{\epsilon^2}\right).$$

Since in each iteration of Algorithm 1 the messages between two leave nodes of the tree are updated, and each message update is of complexity $\mathcal{O}(n^2)$, one iteration takes at most $\mathcal{O}(d(G)n^2)$ operations. Thus, in expectation, a solution is achieved in

$$\mathcal{O}\left(\frac{d(G)n^2 m|\Gamma|^2 (R_C^\Gamma)^2 \log(n)}{\epsilon^2}\right)$$

operations. Algorithm 2 takes $\mathcal{O}(|\Gamma|n^2)$ (see Lemma 7 in Altschuler et al. (2017)). Hence, the bound on $\mathbb{E}[T]$ follows. The bound in probability follows similarly. □

5 Extension to general graphs

For a general graph, we cannot directly apply the belief propagation algorithm. One way to tackle this is to construct a tree factorization over the graph. A junction tree (also called tree decomposition) describes a partitioning of a graph, where several nodes are clustered together, such that the interactions between the clusters can be described by a tree (Koller & Friedman, 2009, Chapter 10). A cluster c is a collection of nodes, and we write $\mathbf{x}_c = \{x_k, k \in c\}$. Moreover, the matrices $K^{(k_1,k_2)} = \exp(-C^{(k_1,k_2)}/\eta)$, for $(k_1, k_2) \in E$, can be understood as pair-wise potentials. A junction tree is then defined as follows.

Definition 1. *A junction tree $\mathcal{T} = (\mathcal{C}, \mathcal{E})$ over a graph $G = (V, E)$ is a tree whose nodes $c \in \mathcal{C}$ are associated with subsets $\mathbf{x}_c \subset V$, and that satisfies the following properties:*

- *Family preservation: For each potential K there is a cluster c such that domain$(K) \subset \mathbf{x}_c$.*
- *Running intersection: For every pair of clusters $c_i, c_j \in \mathcal{C}$, every cluster on the path between c_i and c_j contains $\mathbf{x}_{c_i} \cap \mathbf{x}_{c_j}$.*

For two adjoining clusters c_i and c_j, we define the separation set $S_{ij} = \{v \in V : v \in c_i \cap c_j\}$.

A graph can be clustered into many different junction trees. It is often practical to find a junction tree that is as similar to a tree as possible. A measure of this is given by the following definition.

Definition 2. *For a junction tree $\mathcal{T} = (\mathcal{C}, \mathcal{E})$, we define its width as*

$$\textit{width}(\mathcal{T}) = \max_{c\in\mathcal{C}} |c| - 1.$$

For a graph G, we define its tree-width as

$$w(G) = \min\{\textit{width}(\mathcal{T}) \mid \mathcal{T} \textit{ is a junction tree for } G\}.$$

In order to extend Algorithm 1 to junction trees, the constraints need to be given on the leaf nodes of the tree. Thus, we need to define the junction tree such that all leaves are clusters containing only one vertex and correspond to the set Γ.

Example 3. *A junction tree with minimal tree-width for the Wasserstein least squares problem is illustrated in Figure 3. The graph in Figure 2 thus has tree-width 2.*

The problem of finding a minimal junction tree for a given graph is very challenging in itself (Koller & Friedman, 2009, Chapter 10). In this work we assume that a junction tree decomposition is known. Based on

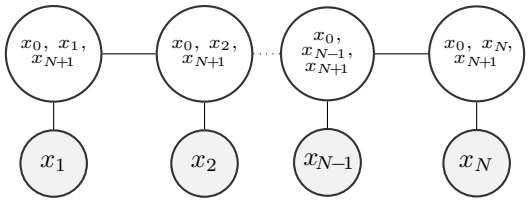


Figure 3: Junction tree for the graph in Figure 2 representing the Wasserstein least squares problem (5).

the junction tree partitioning, we achieve the following complexity bound for general graph-structured MOT problems. The derivation of the modified algorithm is deferred to Section C in the supplementary material.

Theorem 3. *Let* $R = \max_{k\in\Gamma} R^k_{\mathbf{C}}$, *where* $R^k_{\mathbf{C}} = \|\mathbf{C}_{c_{\ell_k}}(\mathbf{x}_{c_{\ell_k}})\|_\infty$, *and* c_{ℓ_k} *is the neighbouring clique to* c_k. *A generalization of Algorithm 3 finds an* ϵ*-approximate solution to the general graph structured MOT problem* (2) *in* T *arithmetic operations, where*

$$\mathbb{E}[T] = \mathcal{O}\left(\frac{d(\mathcal{T})m|\Gamma|^2 n^{w(G)+1}R^2\log(n)}{\epsilon^2}\right).$$

Moreover, for all $\delta \in (0, 0.5)$ *there exists a universal constant* c *such that*

$$\mathbb{P}\left(T \leq \frac{cd(\mathcal{T})m|\Gamma|^2 n^{w(G)+1}R^2\log(n)\log(1/\delta)}{\epsilon^2}\right) \geq 1-\delta.$$

6 Discussion of results

In Algorithm 1, the per iteration complexity is not independent of the random choice of the update, and thus not independent of the number of iterations. The results in Theorem 2 and 3 thus depend on the maximum iteration complexity, and can be improved by utilizing the expected (average) iteration complexity. Therefore, let $\bar{d}(G)$ denote the average distance between any two nodes in Γ.

Theorem 4. *A generalization of Algorithm 3 finds an* ϵ*-approximate solution to the graph-structured MOT problem* (2) *in* T *arithmetic operations, where*

$$\mathbb{E}[T] = \mathcal{O}\left(\frac{\bar{d}(\mathcal{T})m|\Gamma|^2 n^{w(G)+1}R^2\log(n)}{\epsilon^2}\right).$$

If the underlying graph is fully connected, the junction tree contains one "big" cluster that includes all the nodes. Then we have $\bar{d}(\mathcal{T}) = 2$, $w(G) = m-1$ and $R = \|\mathbf{C}\|_\infty$ in Theorem 4. In fact, in this case our algorithm does not exploit any graph structures, and thus the complexity is the same for general cost tensors that do not decouple into pairwise terms as in (6). Thus, the complexity of Algorithm 3 for general MOT problems matches the bound for general MOT problems in Lin et al. (2019).

We consider a class of tree-structured MOT problems, which contains many MOT applications of interest.

Definition 3. *Given a sequence of tree-structured MOT problems, where the number of nodes go to infinity, we call the sequence of such problems balanced if there is a constant* c *such that* $|\Gamma|R^\Gamma_C \leq c\|\mathbf{C}\|_\infty$.

Many MOT problems that arise in practice are balanced, see Section D.1 in the supplementary material for a number of examples. From Theorem 2 it follows that Algorithm 3 finds an ϵ-approximate solution to balanced MOT problem (2) in T operations, where

$$\mathbb{E}[T] = \mathcal{O}\left(\frac{\bar{d}(G)mn^2\|\mathbf{C}\|^2_\infty\log(n)}{\epsilon^2}\right).$$

This lets us compare our result with the bound for general MOT problems in Lin et al. (2019) without acceleration, which is given by $\mathcal{O}\left(m^3n^m\|\mathbf{C}\|^2_\infty\log(n)\epsilon^{-2}\right)$. Moreover, when the MOT problem on the junction tree is balanced, by a similar argumentation the expectation bound in Theorem 4 can be given by

$$\mathbb{E}[T] = \mathcal{O}\left(\frac{\bar{d}(\mathcal{T})mn^{w(G)+1}\|\mathbf{C}\|^2_\infty\log(n)}{\epsilon^2}\right).$$

Consider the barycenter problem introduced in Example 1. This problem is a MOT problem (2) with underlying graph as illustrated in Figure 1. Here, $d(G) = 2$, $|\Gamma| = L$, and $m = L+1$. Moreover, by (4), we have $R^\Gamma_C = \frac{1}{L}\|C\|_\infty$. Thus, Algorithm 3 is expected to return an $\epsilon-$approximate solution to problem 3 in $\mathcal{O}(Ln^2\|C\|^2_\infty\log(n)\epsilon^{-2})$. This coincides with the best known bound for the barycenter problem (Kroshnin et al., 2019; Lin et al., 2020) without acceleration. In fact, the argument can be extended to the case of non-uniform weights in the barycenter problem (3), see Section D.2 in the supplementary material. We also point out that the regularizer used in the Wasserstein barycenter literature is pairwise, whereas ours regularizes the full tensor $\mathbf{B}$. For more details on this comparison, see Haasler et al. (2021a, Section 5).

7 Experiments

We show numerical results for three types of MOT problems. We consider the barycenter problem in Example 1, which is structured according to the graph in Figure 1, and the Hidden Markov Model example in Haasler et al. (2021c, Section V.B), which is structured according to the graph in Figure 4. In particular, these two types of problems are tree-structured. The third ex-

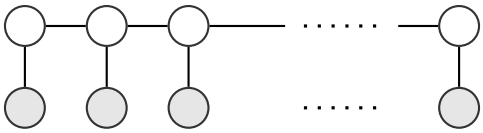
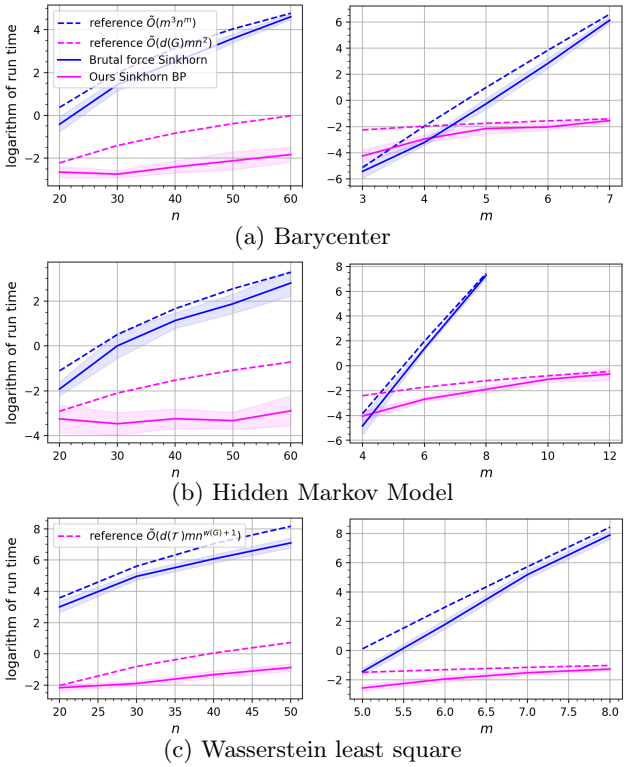


Figure 4: Graph associated with a Hidden Markov Model.

ample is the Wasserstein least square problem (Karimi et al., 2020), which is associated with the graph in Figure 2. Note that this is a graph with tree-width two. In all the above graphs, gray nodes correspond to fixed marginals $\{\mu_i\}_{k\in\Gamma}$, and white nodes are estimated in the problem.

The cost matrices $C^{(k_1,k_2)}$ in (6) are set to be the squared Euclidean distance. The constrained marginal distributions $\{\mu_k\}_{k\in\Gamma}$ are supported on a uniform grid with n points between 0 and 1, where the values are generated from the log-normal distribution and normalized to sum to one. We choose the accuracy $\epsilon = 0.2$ in Algorithm 3. As a comparison, we implemented a brute force Sinkhorn method, which computes the projections $P_k(\mathbf{B}(\Lambda^{(t)}))$ in the Sinkhorn iterates (10) by directly summing over the elements of the tensor $\mathbf{B}(\Lambda^{(t)})$ as in (1). We use a random update rule for both methods. The number of iterations of both brute force Sinkhorn and Sinkhorn belief propagation are nearly the same. For brute force Sinkhorn and Sinkhorn BP, we use the code given by `https://github.com/qshzh/cbp` and make necessary modifications, such as random update rules. We repeat every experiment 5 times with different random seeds and report the total run time in Figure 5. The theoretical complexity bound is also presented as dashed lines. The run time of brute force Sinkhorn grows in a higher polynomial of n and grows exponentially with respect to m. This coincides with the general MOT bound and our bounds in Table 1. We can also tell our bound is a bit pessimistic about the dependence over n.



(a) Barycenter

(b) Hidden Markov Model

(c) Wasserstein least square

Figure 5: Logarithm of total run time in seconds for brute force Sinkhorn and Sinkhorn belief propagation. The left column shows the run time as a function of n when m is fixed ($m = 4$ in (a) and (b); $m = 5$ in (c)), and the right column vice versa with $n = 10$. We choose $\alpha = 10$ in the Wasserstein least square example.

8 Conclusion

In this work we considered a class of multi-marginal optimal transport problems where the cost functions can be decomposed according to a graph. It turns out that the computational complexity of MOT can be significantly reduced by exploiting graphical structures. More specifically, without any structure, the complexity grows exponentially as the number of marginals increases. With graphical structure, the dependence becomes polynomial. We provide a complexity bound $\tilde{\mathcal{O}}(d(\mathcal{T})mn^{w(G)+1}\epsilon^{-2})$ for solving graph-structured MOT problems based on the Sinkhorn belief propagation algorithm (Haasler et al., 2021c; Singh et al., 2020) with the random updating rule. One limitation of the present work is that the proof techniques do not seem to be applicable to Sinkhorn iterations with cyclic updating rule, which is the most popular strategy used in practice. This will be a future research direction. We also plan to accelerate the Sinkhorn belief propagation algorithm using ideas from Lin et al. (2019); Kroshnin et al. (2019).

Acknowledgements

The authors would like to thank the anonymous reviewers for useful comments. JF and YC are supported in part by grants NSF CAREER ECCS-1942523 and NSF CCF-2008513. IH and JK are supported in part by Swedish Research Council (VR) under grant 2020-03454 and Digital Futures. The authors also thank Qinsheng Zhang for fruitful discussions.

References

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

Altschuler, J., Niles-Weed, J., and Rigollet, P. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in neural information processing systems*, pp. 1964–1974, 2017.

Altschuler, J. M. and Boix-Adsera, E. Polynomial-time algorithms for multimarginal optimal transport problems with structure. *arXiv preprint arXiv:2008.03006*, 2020.

Altschuler, J. M. and Parrilo, P. A. Random Osborne: a simple, practical algorithm for matrix balancing in near-linear time. *arXiv preprint arXiv:2004.02837*, 2020.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Arnborg, S., Corneil, D. G., and Proskurowski, A. Complexity of finding embeddings in a k-tree. *SIAM Journal on Algebraic Discrete Methods*, 8(2):277–284, 1987.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Brenier, Y. The least action principle and the related concept of generalized flows for incompressible perfect fluids. *Journal of the American Mathematical Society*, 2(2):225–255, 1989.

Carlier, G. On the linear convergence of the multimarginal sinkhorn algorithm. 2021.

Chen, Y., Georgiou, T. T., and Pavon, M. On the relation between optimal transport and Schrödinger bridges: A stochastic control viewpoint. *Journal of Optimization Theory and Applications*, 169(2): 671–691, 2016.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

Deming, W. E. and Stephan, F. F. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444, 1940.

Dvurechensky, P., Gasnikov, A., and Kroshnin, A. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In *International conference on machine learning*, pp. 1367–1376. PMLR, 2018.

Elvander, F., Haasler, I., Jakobsson, A., and Karlsson, J. Multi-marginal optimal transport using partial information with applications in robust localization and sensor fusion. *Signal Processing*, 171:107474, 2020.

Haasler, I., Chen, Y., and Karlsson, J. Optimal steering of ensembles with origin-destination constraints. *IEEE Control Systems Letters*, 5(3):881–886, 2020.

Haasler, I., Ringh, A., Chen, Y., and Karlsson, J. Multimarginal optimal transport with a tree-structured cost and the Schrödinger bridge problem. *SIAM Journal on Control and Optimization*, 59(4):2428–2453, 2021a.

Haasler, I., Ringh, A., Chen, Y., and Karlsson, J. Scalable computation of dynamic flow problems via multimarginal graph-structured optimal transport. *arXiv preprint arXiv:2106.14485*, 2021b.

Haasler, I., Singh, R., Zhang, Q., Karlsson, J., and Chen, Y. Multi-marginal optimal transport and probabilistic graphical models. *IEEE Transactions on Information Theory*, 2021c.

Haker, S., Zhu, L., Tannenbaum, A., and Angenent, S. Optimal mass transport for registration and warping. *International Journal of computer vision*, 60(3):225–240, 2004.

Karimi, A., Ripani, L., and Georgiou, T. T. Statistical learning in Wasserstein space. *IEEE Control Systems Letters*, 5(3):899–904, 2020.

Koller, D. and Friedman, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Kroshnin, A., Tupitsa, N., Dvinskikh, D., Dvurechensky, P., Gasnikov, A., and Uribe, C. On the complexity of approximating Wasserstein barycenters. In *International conference on machine learning*, pp. 3530–3540. PMLR, 2019.

Lin, T., Ho, N., Cuturi, M., and Jordan, M. I. On the complexity of approximating multimarginal optimal transport. *arXiv preprint arXiv:1910.00152*, 2019.

Lin, T., Ho, N., Chen, X., Cuturi, M., and Jordan, M. Fixed-support Wasserstein barycenters: Computational hardness and fast algorithm. *Advances in Neural Information Processing Systems*, 33:5368–5380, 2020.

Marino, S. D. and Gerolin, A. An optimal transport approach for the schrödinger bridge problem and convergence of sinkhorn algorithm. *Journal of Scientific Computing*, 85(2):1–28, 2020.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6): 355–607, 2019.

Singh, R., Haasler, I., Zhang, Q., Karlsson, J., and Chen, Y. Inference with aggregate data: An optimal transport approach. *arXiv preprint arXiv:2003.13933*, 2020.

Solomon, J., Rustamov, R., Guibas, L., and Butscher, A. Wasserstein propagation for semi-supervised learning. In *International Conference on Machine Learning*, pp. 306–314. PMLR, 2014.

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):1–11, 2015.

Villani, C. *Optimal transport: old and new*, volume 338. Springer, 2009.

Wald, A. Some generalizations of the theory of cumulative sums of random variables. *The Annals of Mathematical Statistics*, 16(3):287–293, 1945.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8:236–239, 2003.

A Tensor operations

In this section we briefly introduce notations that are used for tensors. The inner product between two m-mode tensors $\mathbf{B}$ and $\mathbf{C}$ is defined as

$$\langle \mathbf{C}, \mathbf{B} \rangle = \sum_{x_1,\dots,x_m} \mathbf{C}(x_1,\dots,x_m)\mathbf{B}(x_1,\dots,x_m).$$

It is worth noting that this can be seen as a generalization of the Frobenius inner product for matrices. The projection on the k-th marginal of the tensor $\mathbf{B}$ is defined as

$$[P_k(\mathbf{B})](x_k) = \sum_{x_1,\dots,x_{k-1},x_{k+1},\dots,x_m} \mathbf{B}(x_1,\dots,x_m),$$

see also (1). By summing out all indexes except for x_k, we get a vector with elements in x_k. This can be interpreted as a marginalization of a probability measure. More precisely, if the tensor $\mathbf{B}$ describes a m-dimensional probability measure of the variables $x_1,\dots,x_m$, then $P_k(\mathbf{B})$ describes the probability distribution for x_k. Similarly, the bi-marginal projection $P_{k_1,k_2}(\mathbf{B})$ defined in (13) describes the joint distribution of x_{k_1} and x_{k_2}.

B The dual of the regularized MOT problem and the Sinkhorn iterations

In this section we provide details to Section 3. In particular, we derive the dual of the regularized MOT problem and the Sinkhorn belief propagation algorithm.

The Lagrangian function of problem (7) is

$$L(\mathbf{B}, \Lambda) = \langle \mathbf{C}, \mathbf{B} \rangle + \eta H(\mathbf{B} \mid \mathbf{M}) - \sum_{k\in\Gamma} \lambda_k^{\mathrm{T}} \left(P_k(\mathbf{B}) - \mu_k \right), \tag{14}$$

where $\Lambda = (\lambda_k)_{k\in\Gamma}$ and $\lambda_k \in \mathbb{R}^n$ for $k \in \Gamma$. Minimizing the Lagrangian with respect to $\mathbf{B}$ gives the optimum

$$[\mathbf{B}(\Lambda)](x_1,\dots,x_m) = \exp\left(-\mathbf{C}(x_1,\dots,x_m)/\eta\right) \prod_{k\in\Gamma} \Big(\exp\left(\lambda_k(x_k)/\eta\right) \mu_k(x_k) \Big),$$

and plugging this into (14) yields

$$\inf_{\mathbf{B}} L(\mathbf{B}, \Lambda) = L(\mathbf{B}(\Lambda), \Lambda) = -\eta P(\mathbf{B}(\Lambda)) + \sum_{k\in\Gamma} \mu_k^{\mathrm{T}} \lambda_k.$$

Therefore, the dual problem (formulated as a minimization problem) is given by

$$\min_{\Lambda} \psi(\Lambda) := \eta P(\mathbf{B}(\Lambda)) - \sum_{k\in\Gamma} \mu_k^{\mathrm{T}} \lambda_k.$$

In each iteration the block coordinate descent algorithm picks some $k \in \Gamma$ and minimizes $\psi(\Lambda)$ over λ_k, while keeping the other variables fixed. The minimum is achieved when the gradient of ψ with respect to λ_k vanishes, i.e., when

$$e^{\lambda_k(x_k)/\eta} \mu_k(x_k) \left(\sum_{\mathbf{x}\setminus x_k} e^{-\mathbf{C}(x_1,\dots,x_m)/\eta} \prod_{\ell\in\Gamma\setminus k} \Big(e^{\lambda_\ell(x_\ell)/\eta} \mu_\ell(x_\ell) \Big) \right) - \mu_k(x_k) = 0.$$

In the scaled variables $u_k = \exp(\lambda_k/\eta)$ this can be expressed as

$$u_k^{(t+1)} \odot \mu_k \odot \Big(P_k(\mathbf{B}(\Lambda^{(t)})) ./ \Big(u_k^{(t)} \odot \mu_k \Big) \Big) - \mu_k = 0.$$

This yields the Sinkhorn updates (10).

C Algorithm for MOT with general graph structure

In order to apply Sinkhorn belief propagation, we first decompose the underlying graph into a tree with minimal tree-width. Finding such a tree decomposition is a NP hard problem (Arnborg et al., 1987). Luckily, in many applications of graph-structured MOT a tree decomposition is known. Thus, in the following we assume that the cost tensor $\mathbf{C}$ decouples according to $\mathcal{T} = (\mathcal{C}, \mathcal{E})$ into tensors $\mathbf{C}_c$, for $c \in \mathcal{C}$, such that

$$\mathbf{C} = \sum_{c \in \mathcal{C}} \mathbf{C}_c(\mathbf{x}_c).$$

The potential tensor $\mathbf{K} = \exp(-\mathbf{C}/\eta)$ is then factorized, into tensors $\mathbf{K}_c = \exp(-\mathbf{C}_c/\eta)$, for $c \in \mathcal{C}$, and can be written as

$$\mathbf{K}(\mathbf{x}) = \prod_{c \in \mathcal{C}} \mathbf{K}_c(\mathbf{x}_c).$$

To apply Algorithm 1, the constraints have to be given on the leaf nodes of the tree. Thus, we define the junction tree such that the all leaves are clusters containing only one vertex and correspond to the set Γ. We denote this set of leaf cliques by $\Gamma_{\mathcal{C}}$. In particular, note that then $S_{\ell_k k} = x_k$, if $c_k = \{x_k\} \in \Gamma_{\mathcal{C}}$, and c_{ℓ_k} is its unique neighbour clique. The Sinkhorn iterations are then of the form (10), where the projections on the marginals $c_k \in \Gamma_{\mathcal{C}}$, with neighbour clique $c_{\ell_k} \in \mathcal{C}$, are computed as

$$[P_k(\mathbf{B}(\Lambda^{(t)}))](x_k) = u_k^{(t)}(x_k)\mu_k(x_k)m_{\ell_k \to k}(x_k). \tag{15}$$

Here, the messages between clusters of the junction tree are given by

$$m_{\ell \to k}(S_{\ell k}) = \sum_{\mathbf{x}_{c_\ell} \setminus S_{\ell k}} \mathbf{K}_{c_\ell}(\mathbf{x}_{c_\ell}) \prod_{j \in N(\ell) \setminus k} m_{j \to \ell}(S_{j\ell}), \quad \text{if } c_\ell \notin \Gamma_{\mathcal{C}} \tag{16a}$$

$$m_{\ell \to k}(x_\ell) = u_\ell^{(t)}(x_\ell)\mu_\ell(x_\ell), \qquad \text{if } c_\ell \in \Gamma_{\mathcal{C}}. \tag{16b}$$

It follows that the Sinkhorn iterations (10) with the projections (15) read, as before,

$$u_k^{(t+1)}(x_k) \leftarrow (m_{\ell_k \to k}(x_k))^{-1}.$$

Algorithm 1 can thus simply be modified to general graphs by replacing the messages (12) by the messages (16). This lets us formulate the result in Theorem 3.

D Details on the discussion of results in Section 6

This Section provides details on the discussion of the results.

D.1 Balanced MOT problems

There are many structured MOT problems of interest that are balanced. In the following we check the condition in Definition 3 for a few special cases.

Example 4. *The Wasserstein barycenter problem discussed in Example 1 is balanced. With the barycenter cost tensor* $\mathbf{C}$ *defined in* (4) *it holds* $\|\mathbf{C}\|_\infty = \|C\|_\infty$*, and thus* $|\Gamma| R_C^\Gamma = L\frac{1}{L}\|C\|_\infty = \|\mathbf{C}\|_\infty$.

Example 5. *A tree-structured MOT problem where the costs on all edges are equal and symmetric is balanced. Note that if* $C^{(k_1,k_2)}$*, for all* $(k_1, k_2) \in E$*, are equal and symmetric, then* $\|\mathbf{C}\|_\infty = |E| R_C^\Gamma$*. Thus, it holds*

$$|\Gamma| R_C^\Gamma = \frac{|\Gamma|}{|E|}\|\mathbf{C}\|_\infty \leq \|\mathbf{C}\|_\infty.$$

The barycenter case in Example 1 is a special case of this.

Example 6. *Consider a tree-structured MOT problem, where the shortest distance between any two leaf nodes is* 3*, and the maximum cost entries on the edges connecting to the leaf nodes are of the same order. Such a problem*

is balanced. Let $\|C^{(k,\ell_k)}\|_\infty$ *be of the same order for all* $k \in \Gamma$*, where* ℓ_k *is the neighbour of* k*. Then there is a constant* c *such that*

$$R_C^\Gamma = \max_{k\in\Gamma} \|C^{(k,\ell_k)}\|_\infty \leq \frac{c}{|\Gamma|} \sum_{k\in\Gamma} \|C^{(k,\ell_k)}\|_\infty.$$

If the shortest distance between any two leaf nodes is 3*, there is no node that has two leaf nodes as neighbours. Thus, it holds*

$$\sum_{k\in\Gamma} \|C^{(k,\ell_k)}\|_\infty \leq \|\mathbf{C}\|_\infty.$$

Hence, it follows $|\Gamma| R_C^\Gamma \leq c\|\mathbf{C}\|_\infty$.

Example 7. *Consider a tree-structured MOT problem with cost tensor* $\mathbf{C}$*. Let* $\tilde{\mathbf{x}} = (\tilde{x}_1, \ldots, \tilde{x}_m)$ *be a maximizer of* $\mathbf{C}(\mathbf{x})$*, that is* $\mathbf{C}(\tilde{\mathbf{x}}) = \|\mathbf{C}\|_\infty$*, and assume that*

$$\left| \frac{\max_{k\in\Gamma} \|C^{(k,\ell_k)}\|_\infty}{\min_{k\in\Gamma} C^{(k,\ell_k)}(\tilde{x}_{k_1}, \tilde{x}_{k_2})} \right| \leq c$$

for some constant c*. Then the MOT problem is balanced. To see this note that*

$$\|\mathbf{C}\|_\infty = \sum_{(k_1,k_2)\in E} C^{(k_1,k_2)}(\tilde{x}_{k_1}, \tilde{x}_{k_2}) \geq |E| \min_{(k_1,k_2)\in E} C^{(k_1,k_2)}(\tilde{x}_{k_1}, \tilde{x}_{k_2}).$$

Thus, it follows

$$|\Gamma| R_C^\Gamma = |\Gamma| \max_{k\in\Gamma} \|C^{(k,\ell_k)}\|_\infty \leq c|\Gamma| \min_{k\in\Gamma} C^{(k,\ell_k)}(\tilde{x}_{k_1}, \tilde{x}_{k_2}) \leq c\frac{|\Gamma|}{|E|}\|\mathbf{C}\|_\infty \leq c\|\mathbf{C}\|_\infty.$$

D.2 Barycenter problem with nonuniform weights

We provide a complexity bound for the barycenter problem with nonuniform weights. Let w_ℓ be the weight and C_ℓ be the cost matrix for the ℓ-th term in the barycenter problem (3). Then the cost tensor in the corresponding MOT problem is given by

$$\mathbf{C}(x_1, \ldots, x_L, x_{L+1}) = \sum_{\ell=1}^{L} w_\ell C_\ell(x_{L+1}, x_\ell).$$

With this cost, the bound in Lemma 2 becomes

$$\psi(\Lambda) - \psi(\Lambda^*) \leq R \sum_k w_k \|P_k(\mathbf{B}(\Lambda)) - \mu_k\|_1, \quad \text{where } R = \max_\ell \|C_\ell\|_\infty.$$

Now, if in Algorithm 1 we pick the next update according to the weight $w_1, w_2, \ldots, w_L$ instead of a uniform distribution, then the bound in Lemma 3 becomes

$$\mathbb{E}\left[\psi(\Lambda^{(t)}) - \psi(\Lambda^{(t+1)})\right] \geq \frac{\eta}{2}(e_t)^2, \quad \text{with } e_t = \sum_{k\in\Gamma} w_k \|P_k(\mathbf{B}(\Lambda^{(t)})) - \mu_k\|_1.$$

The bound in Theorem 1 then becomes $\mathcal{O}(\frac{R}{\eta\epsilon'})$. Putting everything together, the iteration complexity becomes $\tilde{\mathcal{O}}(\frac{mR^2}{\epsilon^2})$ and the arithmetic complexity becomes $\tilde{\mathcal{O}}(\frac{mn^2R^2}{\epsilon^2})$, which matches the result in Kroshnin et al. (2019).

E Deferred proofs

In this section we provide the proofs that are omitted in the main paper.

E.1 Proof of Lemma 1

Proof. Denote $v_k(x_\ell) = \prod_{j\in N(\ell)\setminus k} m_{j\to\ell}(x_\ell)$, where $\ell \in N(k)$ is the unique neighbour of k, since k is a leaf of the tree. Assume variable u_k was updated in the previous step of the algorithm. Then it holds

$$u_k(x_k) = 1/m_{\ell\to k}(x_k) = 1/\left([K^{(k,\ell)}v_k](x_k)\right).$$

Thus,

$$\max_{x_k} \lambda_k(x_k) \leq -\eta \log\left(e^{-\|C^{(k,\ell)}\|_\infty/\eta} v_k^T \mathbf{1}\right) = \|C^{(k,\ell)}\|_\infty - \eta \log\left(v_k^T \mathbf{1}\right). \tag{17}$$

Moreover,

$$\min_{x_k} \lambda_k(x_k) \geq -\eta \log\left(v_k^T \mathbf{1}\right). \tag{18}$$

Combining (17) and (18) it follows

$$\max_{x_k} \lambda_k(x_k) - \min_{x_k} \lambda_k(x_k) \leq \|C^{(k,\ell)}\|_\infty.$$

Note that the gradient of $\psi(\cdot)$ vanishes in Λ^*, since it is optimal to (9). Thus, it holds $P_k(\mathbf{B}(\Lambda^*)) = \mu_k$ for $k = 1, \dots, m$ and the bound for λ_k^* follows in the same way as before.

□

E.2 Proof of Lemma 2

Proof. Note that

$$\begin{aligned}\psi(\Lambda) - \psi(\Lambda^*) &= \eta P(\mathbf{B}(\Lambda)) - \sum_{k\in\Gamma} \mu_k^{\mathrm{T}} \lambda_k - \eta P(\mathbf{B}(\Lambda^*)) + \sum_{k\in\Gamma} \mu_k^{\mathrm{T}} \lambda_k^* \\ &= \eta P(\mathbf{B}(\Lambda)) - \sum_{k\in\Gamma} \lambda_k^{\mathrm{T}} P_k(\mathbf{B}(\Lambda)) - \eta P(\mathbf{B}(\Lambda^*)) + \sum_{k\in\Gamma} (\lambda_k^*)^{\mathrm{T}} P_k(\mathbf{B}(\Lambda)) + \sum_{k\in\Gamma} (\lambda_k - \lambda_k^*)^{\mathrm{T}} \left(P_k(\mathbf{B}(\Lambda)) - \mu_k\right).\end{aligned} \tag{19}$$

Consider the convex function of $\widehat{\Lambda} = \{\widehat{\lambda}_k\}_{k\in\Gamma}$ given by

$$h(\widehat{\Lambda}) = \eta P(\mathbf{B}(\widehat{\Lambda})) - \sum_{k\in\Gamma} \widehat{\lambda}_k^{\mathrm{T}} P_k(\mathbf{B}(\Lambda)).$$

Note that its gradient vanishes if and only if $\widehat{\Lambda} = \Lambda$, since $\nabla_{\widehat{\lambda}_k} h = P_k(\mathbf{B}(\widehat{\Lambda})) - P_k(\mathbf{B}(\Lambda)) = 0$. Thus, Λ is the minimizer of h, and it follows with (19) that

$$\psi(\Lambda) - \psi(\Lambda^*) \leq \sum_{k\in\Gamma} (\lambda_k - \lambda_k^*)^{\mathrm{T}} \left(P_k(\mathbf{B}(\Lambda)) - \mu_k\right). \tag{20}$$

Define $\bar{\lambda}_k = \frac{1}{2}(\max_{x_k} \lambda_k(x_k) + \min_{x_k} \lambda_k(x_k))$, and note that $\bar{\lambda}_k^{\mathrm{T}} \left(P_k(\mathbf{B}(\Lambda)) - \mu_k\right) = 0$. By Hölder's inequality and Lemma 1 , it holds

$$\begin{aligned}\lambda_k^{\mathrm{T}} \left(P_k(\mathbf{B}(\Lambda)) - \mu_k\right) =& (\lambda_k - \bar{\lambda}_k)^{\mathrm{T}} \left(P_k(\mathbf{B}(\Lambda)) - \mu_k\right) \\ \leq& \|\lambda_k - \bar{\lambda}_k)\|_\infty \left\|P_k(\mathbf{B}(\Lambda)) - \mu_k\right\|_1 \\ =& \frac{1}{2}\left(\max_{x_k} \lambda_k(x_k) - \min_{x_k} \lambda_k(x_k)\right) \left\|P_k(\mathbf{B}(\Lambda)) - \mu_k\right\|_1 \\ \leq& \frac{R_C^k}{2} \left\|P_k(\mathbf{B}(\Lambda)) - \mu_k\right\|_1.\end{aligned} \tag{21}$$

Similarly, defining $\bar{\lambda}_k^* = \frac{1}{2}(\max_{x_k} \lambda_k^*(x_k) + \min_{x_k} \lambda_k^*(x_k))$, we derive the bound

$$-\lambda_k^{*\,\mathrm{T}} \left(P_k(\mathbf{B}(\Lambda)) - \mu_k\right) = (\bar{\lambda}_k^* - \lambda_k^*)^{\mathrm{T}} \left(P_k(\mathbf{B}(\Lambda)) - \mu_k\right) \leq \frac{R_C^k}{2} \left\|P_k(\mathbf{B}(\Lambda)) - \mu_k\right\|_1. \tag{22}$$

Summing (21) and (22) over $k \in \Gamma$ yields

$$\sum_{k\in\Gamma} (\lambda_k - \lambda_k^*)^{\mathrm{T}} \left(P_k(\mathbf{B}(\Lambda)) - \mu_k\right) \leq \sum_{k\in\Gamma} R_C^k \left\|P_k(\mathbf{B}(\Lambda)) - \mu_k\right\|_1 \leq R_C^\Gamma \sum_{k\in\Gamma} \left\|P_k(\mathbf{B}(\Lambda)) - \mu_k\right\|_1.$$

Together with (20) this completes the proof. □

E.3 Proof of Lemma 3

Proof. Since $P(\mathbf{B}(\Lambda^t)) = 1$ for all t and $u^{(t+1)}_{k^{(t+1)}} ./ u^{(t)}_{k^{(t+1)}} = \mu_{k^{(t+1)}} ./ P_{k^{(t+1)}}(\mathbf{B}(\Lambda^{(t)}))$,

$$\begin{aligned}\psi(\Lambda^{(t)}) - \psi(\Lambda^{(t+1)}) =& \mu^{\mathrm{T}}_{k^{(t+1)}} \left(-\lambda^t_{k^{(t+1)}} + \lambda^{t+1}_{k^{(t+1)}}\right) \\ =& \eta \mu^{\mathrm{T}}_{k^{(t+1)}} \log \frac{\mu_{k^{(t+1)}}}{P_\ell(\mathbf{B}(\Lambda^{(t)}))} \\ =& \eta \mathrm{KL}(\mu_{k^{(t+1)}} \mid P_{k^{(t+1)}}(\mathbf{B}(\Lambda^{(t)}))).\end{aligned}$$

where KL is the Kullback–Leibler divergence. By Pinsker's inequality, we get

$$\psi(\Lambda^{(t)}) - \psi(\Lambda^{(t+1)}) \geq \frac{\eta}{2} \|\mu_{k^{(t+1)}} - P_{k^{(t+1)}}(\mathbf{B}(\Lambda^{(t)}))\|_1^2. \tag{23}$$

Since $k^{(t+1)}$ is randomly picked from a uniform distribution over $\Gamma \setminus k^{(t)}$ the expected value of (23) is

$$\psi(\Lambda^{(t)}) - \mathbb{E}_{k^{(t+1)}} \left[\psi(\Lambda^{(t+1)})\right] \geq \frac{\eta}{2(|\Gamma|-1)} \sum_{k \in \Gamma} \|\mu_k - P_k(\mathbf{B}(\Lambda^{(t)}))\|_1^2.$$

By Cauchy–Schwarz inequality, it holds

$$\mathbb{E}_{k^{(t+1)}} \left[\psi(\Lambda^{(t)}) - \psi(\Lambda^{(t+1)})\right] \geq \frac{\eta}{2(|\Gamma|-1)^2} \left(\sum_{k \in \Gamma} \|\mu_k - P_k(\mathbf{B}(\Lambda^{(t)}))\|_1\right)^2.$$

□

E.4 Proof of Theorem 1

We need the following lemma from Altschuler & Parrilo (2020) to connect the per-iteration expected improvement and the number of iterations.

Lemma 6. *(Altschuler & Parrilo, 2020, Lemma 5.3) Assume $A > a, h > 0$. Let $(Y_t)_{t=1}^\infty$ be a sequence of random variables adapted to a filtration $(\mathcal{F}_t)_{t=0}^\infty$ such that (i) $Y_0 \leq A$ almost surely, (ii) $0 \leq Y_{t-1} - Y_t \leq 2(A-a)$ almost surely, and*

$$\textit{(iii)}\ \mathbb{E}\left[Y_t - Y_{t+1} | \mathcal{F}_t, Y_t \geq a\right] \geq h \quad \forall t = 0, 1, 2, \ldots.$$

Then the stopping time $s = \min\{t : Y_t \leq a\}$ satisfies 1) the expectation bound $\mathbb{E}[s] \leq \frac{A-a}{h} + 1$; and 2) $\forall \delta \in (0, 1/e)$, the probability bound $\mathbb{P}(s \leq \frac{6(A-a)}{h} \log \frac{1}{\delta}) \geq 1 - \delta$ holds.

Proof of Theorem 1. Define the stopping time $\tau := \min\{t : e_t \leq \epsilon'\}$. Let $\{\mathcal{F}_t := \sigma\left(\Lambda^{(1)}, \ldots, \Lambda^{(t)}\right)\}_t$ be the natural filtration. By Lemma 2 and Lemma 3,

$$\mathbb{E}\left[\psi(\Lambda^{(t)}) - \psi(\Lambda^{(t+1)}) | \mathcal{F}_t, t < \tau\right] \geq \frac{\eta}{2|\Gamma|^2} \left(\max\left\{\frac{\psi(\Lambda^{(t)}) - \psi(\Lambda^*)}{R^\Gamma_C}, \epsilon'\right\}\right)^2,$$

For shorthand, denote $\widetilde{\psi}(\Lambda^{(t)}) = \psi(\Lambda^{(t)}) - \psi(\Lambda^*)$, and let τ_1 be the first iteration when $\widetilde{\psi}(\Lambda^{(t)}) \leq R^\Gamma_C \epsilon'$ and $\tau_2 := \tau - \tau_1 \geq 0$. Define

$$Z_t = \begin{cases} \widetilde{\psi}(\Lambda^{(t)}) & \text{if } t \leq \tau, \\ \widetilde{\psi}(\Lambda^{(t)}) - (t-\tau)\frac{\eta(\epsilon')^2}{2|\Gamma|^2} & \text{if } t > \tau. \end{cases}$$

A direct observation is that Z_t is monotonically decreasing. For $t \in [\tau_1, \tau]$, let $Y_{t-\tau_1} = Z_t$. Then the expected improvement of Y_t per iteration is at least $\frac{\eta(\epsilon')^2}{2|\Gamma|^2}$, that is

$$\mathbb{E}\left[Y_t - Y_{t+1} | \mathcal{F}_t, Y_t \geq 0\right] \geq \frac{\eta(\epsilon')^2}{2|\Gamma|^2}.$$

With choices $A = R_C^\Gamma \epsilon'$, $a = 0$, and $h = \frac{\eta(\epsilon')^2}{2|\Gamma|^2}$, clearly $Y_t \le A$ and $0 \le Y_t - Y_{t+1} \le 2(A-a)$. Thus, Lemma 6 implies

$$\mathbb{E}[\tau_2'] \le \frac{2|\Gamma|^2 R_C^\Gamma}{\eta\epsilon'} + 1 \quad \text{where } \tau_2' = \min\{t : Y_t \le 0\}.$$

Whenever $t \le \tau$, we have $\widetilde{\psi}(\Lambda^{(t)}) \ge 0$ and as such $Z_t \ge 0$. So $\tau := \min\{t : e_t \le \epsilon'\}$ is achieved earlier than $\min\{t : Z_t \le 0\}$ and this implies

$$\tau - \tau_1 = \tau_2 \le \tau_2' = \min\{t : Z_t \le 0\} - \tau_1 \quad \Rightarrow \quad \mathbb{E}[\tau_2] \le \mathbb{E}[\tau_2'] \le \frac{2|\Gamma|^2 R_C^\Gamma}{\eta\epsilon'} + 1. \tag{24}$$

To bound τ_1, we define $D_0 = R_C^\Gamma e_0$ and $D_i := D_{i-1}/2$ for $i = 1, 2, \dots$ until $D_N \le R_C^\Gamma \epsilon'$. Let $\tau_{1,i}$ be the number of iterations when $D_i \le \widetilde{\psi}(\Lambda^{(t)}) \le D_{i-1}$. Let $t_{1,i} = \min\{t : \widetilde{\psi}(\Lambda^{(t)}) \le D_{i-1}\}$. Consider $A = D_{i-1}$, $a = D_i$, $h = \frac{\eta}{2|\Gamma|^2 {R_C^\Gamma}^2} D_i^2$, and $Y_t = Z_{t+t_{1,i}}$. It holds

$$\mathbb{E}\left[Y_t - Y_{t+1} | \mathcal{F}_t, Y_t \ge D_i\right] \ge \frac{\eta}{2|\Gamma|^2 {R_C^\Gamma}^2} \widetilde{\psi}(\Lambda^{(t)})^2 \ge \frac{\eta}{2|\Gamma|^2 {R_C^\Gamma}^2} D_i^2.$$

In addition $Y_t \le A$ and $0 \le Y_t - Y_{t+1} \le D_{i-1} \le 2(A-a)$ by the nonnegativity and monotonicity of Y_t. From Lemma 6 and the definition of the sequence D_i it follows that

$$\mathbb{E}[\tau_{1,i}] \le \frac{D_{i-1} - D_i}{\eta D_i^2} 2|\Gamma|^2 {R_C^\Gamma}^2 + 1 \le \frac{2|\Gamma|^2 {R_C^\Gamma}^2}{\eta D_i} + 1. \tag{25}$$

Summing up Equation (25) for $i = 1, 2, \dots, N$ and Equation (24) yields

$$\mathbb{E}[\tau] \le \frac{2|\Gamma|^2 R_C^\Gamma}{\eta\epsilon'} + 1 + \sum_{i=1}^{N} \frac{2|\Gamma|^2 {R_C^\Gamma}^2}{\eta D_i} + N \le \frac{2|\Gamma|^2 R_C^\Gamma}{\eta\epsilon'} + 1 + \frac{4|\Gamma|^2 R_C^\Gamma}{\eta\epsilon'} + \log_2\left\lceil \frac{e_0}{\epsilon'} \right\rceil.$$

Since

$$e_0 := \sum_{k\in\Gamma} \|P_k(\mathbf{B}(\Lambda^{(t)})) - \mu_k\|_1 \le \sum_{k\in\Gamma} \|P_k(\mathbf{B}(\Lambda^{(t)}))\|_1 + \|\mu_k\|_1 = 2|\Gamma|,$$

there is $\log_2\left(\frac{e_0}{\epsilon'}\right) \le \frac{e_0}{\epsilon'} \le \frac{2|\Gamma|}{\epsilon'}$. And the mild assumption $\eta \le 0.5|\Gamma|R_C^\Gamma$ implies that

$$1 + \log_2\lceil a \rceil \le 1 + \log_2\lceil b \rceil \le \frac{b|\Gamma|R_C^\Gamma}{\eta}, \quad \forall\, b \ge a > 0,$$

resulting in $1 + \log_2\left\lceil \frac{e_0}{\epsilon'} \right\rceil \le \frac{2|\Gamma|^2 R_C^\Gamma}{\eta\epsilon'}$. It further follows

$$\mathbb{E}[\tau] \le \frac{8|\Gamma|^2 R_C^\Gamma}{\eta\epsilon'}.$$

Next we prove the high probability bound. By Lemma 6, $\forall \delta \in (0, 0.5)$,

$$\mathbb{P}\left(\tau_2 > \frac{12|\Gamma|^2 R_C^\Gamma}{\eta\epsilon'} \log\frac{2}{\delta}\right) < \frac{\delta}{2} \tag{26}$$

and with $\delta_i := \delta/2^{N-i+2}$ for each $i = 1, \dots, N$,

$$\mathbb{P}\left(\tau_{1,i} > \frac{12|\Gamma|^2 {R_C^\Gamma}^2}{\eta D_i} \log\frac{1}{\delta_i}\right) < \delta_i.$$

Given the series summation $\sum_{i=0}^{\infty} 2^{-i} = \sum_{i=0}^{\infty} i \cdot 2^{-i} = 2$ and the definition of δ_i and D_N, we have

$$\sum_{i=1}^{N} \frac{\log\frac{1}{\delta_i}}{D_i} = \frac{1}{D_N} \sum_{i=0}^{N-1} 2^{-i} \left(\log\frac{4}{\delta} + i\log 2\right) \le \frac{2}{D_N}\left(\log\frac{4}{\delta} + \log 2\right) \le \frac{3}{R_C^\Gamma \epsilon'} \log\frac{4}{\delta}.$$

By taking the union over $\tau_{1,i}$ it follows

$$\mathbb{P}\left(\tau_1 > \frac{36|\Gamma|^2 R_C^\Gamma}{\eta\epsilon'}\log\frac{4}{\delta}\right) \leq \sum_{i=1}^{N}\mathbb{P}\left(\tau_{1,i} > \frac{12|\Gamma|^2 {R_C^\Gamma}^2}{\eta D_i}\log\frac{1}{\delta}\right) < \frac{\delta}{2}. \tag{27}$$

Taking a union bound over Equation (26) and Equation (27), we conclude that

$$\mathbb{P}\left(\tau > \frac{48|\Gamma|^2 R_C^\Gamma}{\eta\epsilon'}\log\frac{4}{\delta}\right) < \delta.$$

□

E.5 Proof of Lemma 4

Proof. Due to the underlying tree structure of the problem it holds

$$\langle \mathbf{C}, \mathbf{B}\rangle - \langle \mathbf{C}, \widehat{\mathbf{B}}\rangle = \sum_{(k_1,k_2)\in E}\langle C^{(k_1,k_2)}, P_{k_1,k_2}(\mathbf{B}) - P_{k_1,k_2}(\widehat{\mathbf{B}})\rangle.$$

By Hölder's inequality and (Altschuler et al., 2017, Lemma 7),

$$\begin{aligned}\langle \mathbf{C}, \mathbf{B}\rangle - \langle \mathbf{C}, \widehat{\mathbf{B}}\rangle &\leq \sum_{(k_1,k_2)\in E}\|C^{(k_1,k_2)}\|_\infty \|P_{k_1,k_2}(\mathbf{B}) - P_{k_1,k_2}(\widehat{\mathbf{B}})\|_1 \\ &\leq 2\sum_{k\in\Gamma}\|C^{(k,\ell_k)}\|_\infty\|\mu_k - P_k(\mathbf{B})\|_1.\end{aligned}$$

In the second step note that $P_{k_1,k_2}(\mathbf{B}) = P_{k_1,k_2}(\widehat{\mathbf{B}})$ by construction whenever $k_1, k_2 \notin \Gamma$. Also, note that for $k \in \Gamma$ we can use the bound in (Altschuler et al., 2017, Lemma 7) to get

$$\begin{aligned}\|P_{k,\ell_k}(\widehat{\mathbf{B}}) - P_{k,\ell_k}(\mathbf{B})\|_1 \leq& 2\left(\|P_k(\widehat{\mathbf{B}}) - P_k(\mathbf{B})\|_1 + \|P_{\ell_k}(\widehat{\mathbf{B}}) - P_\ell(\mathbf{B})\|_1\right) \\ =& 2\|P_k(\widehat{\mathbf{B}}) - P_k(\mathbf{B})\|_1 \\ =& 2\|\mu_k - P_k(\mathbf{B})\|_1.\end{aligned}$$

□

E.6 Proof of Lemma 5

Proof. Let $\widetilde{\mathbf{Y}}$ denote the tensor that is returned from Algorithm 2 with inputs $\mathbf{B}^*$ and $\{P_k(\widetilde{\mathbf{B}})\}_{k\in\Gamma}$. Note that $\widetilde{\mathbf{B}}$ is the optimal solution to

$$\min_{\mathbf{B}\in\Pi_\Gamma^m((P_k(\widetilde{\mathbf{B}}))_{k\in\Gamma})}\langle \mathbf{C}, \mathbf{B}\rangle + \eta H(\mathbf{B}|\mathbf{M}),$$

which can easily be verified by checking the KKT conditions. Thus, it holds

$$\langle \mathbf{C}, \widetilde{\mathbf{B}}\rangle + \eta H(\widetilde{\mathbf{B}}|\mathbf{M}) \leq \langle \mathbf{C}, \widetilde{\mathbf{Y}}\rangle + \eta H(\widetilde{\mathbf{Y}}|\mathbf{M}).$$

Since $\langle \widetilde{\mathbf{B}}, \log(\widetilde{\mathbf{B}})\rangle \geq -m\log(n)$ and $\langle \widetilde{\mathbf{Y}}, \log(\widetilde{\mathbf{Y}})\rangle \leq 0$ it follows that

$$\begin{aligned}\langle \mathbf{C}, \widetilde{\mathbf{B}}\rangle - \langle \mathbf{C}, \widetilde{\mathbf{Y}}\rangle &\leq \eta H(\widetilde{\mathbf{Y}}|\mathbf{M}) - \eta H(\widetilde{\mathbf{B}}|\mathbf{M}) \\ &\leq -\langle \widetilde{\mathbf{B}}, \log(\widetilde{\mathbf{B}})\rangle + \langle \widetilde{\mathbf{B}} - \widetilde{\mathbf{Y}}, \log\mathbf{M}\rangle \\ &\leq \eta m\log(n) + \eta\sum_{k\in\Gamma}\langle P_k(\tilde{\mathbf{B}}) - P_k(\tilde{\mathbf{Y}}), \log\mu_k\rangle \\ &= \eta m\log(n).\end{aligned} \tag{28}$$

Lemma 4 gives

$$\langle \mathbf{C}, \widetilde{\mathbf{Y}} \rangle - \langle \mathbf{C}, \mathbf{B}^* \rangle \leq 2 \sum_{k \in \Gamma} \|C^{(k,\ell_k)}\|_\infty \|P_k(\widetilde{\mathbf{Y}}) - \mu_k\|_1, \tag{29}$$

$$\langle \mathbf{C}, \widehat{\mathbf{B}} \rangle - \langle \mathbf{C}, \widetilde{\mathbf{B}} \rangle \leq 2 \sum_{k \in \Gamma} \|C^{(k,\ell_k)}\|_\infty \|P_k(\widetilde{\mathbf{B}}) - \mu_k\|_1. \tag{30}$$

Since $P_k(\widetilde{\mathbf{B}}) = P_k(\widetilde{\mathbf{Y}})$, summing up (28), (29), and (30) concludes the proof.

□

E.7 Proof of Theorem 3

Proof. In the case of a general graph, we factorize it according to a junction tree with minimal tree-width and modify the messages in Algorithm 1 to the message passing scheme in (16). Note that each message update requires at most $\mathcal{O}(n^{w(G)+1})$ operations. In order to perform one iteration of Algorithm 1 on a junction tree, at most $d(\mathcal{T})$ messages have to be updated. Thus, each iteration of Algorithm 1 on a junction tree requires $\mathcal{O}(d(\mathcal{T})n^{w(G)+1})$ operations. The results in Lemma 1-5 and and Theorem 1 can be applied to the junction tree version of the presented methods. In particular, the constant in Lemma 1 is modified to $R^k_{\mathbf{C}} = \|\mathbf{C}_{c_{\ell_k}}(\mathbf{x}_{c_{\ell_k}})\|_\infty$, where c_{ℓ_k} is the neighbouring clique to c_k. Letting $R = \max_{k\in\Gamma} R^k_{\mathbf{C}}$, the proof follows as the proof of Theorem 2, where the per-iteration complexity is now $\mathcal{O}(d(\mathcal{T})n^{w(G)+1})$. □

E.8 Proof of Theorem 4

Proof. First, note that the expected time of one iteration is $\mathbb{E}[T_t] = \mathcal{O}(\bar{d}(G)n^{w(G)+1})$, for all t. The expectation of the random variables T_t is thus bounded and equal for all t. We also note that

$$\mathbb{E}[T_t \mathbf{1}_{\tau \geq t}] = \mathbb{E}[T_t | \tau \geq t]\mathbb{P}(\tau \geq t) = \mathbb{E}[T_t]\mathbb{P}(\tau \geq t).$$

Moreover,

$$\sum_{t=1}^{\infty} \mathbb{E}[T_t \mathbf{1}_{\tau \geq t}] = \sum_{t=1}^{\infty} \mathbb{E}[T_t]\mathbb{P}(\tau \geq t) = \mathbb{E}[T_t] \sum_{t=1}^{\infty} \mathbb{P}(\tau \geq t) = \mathbb{E}[T_t]\mathbb{E}[\tau] < \infty.$$

Thus, by the general Wald's equation (Wald, 1945)(Altschuler & Parrilo, 2020, Lemma 5.6) it follows

$$\mathbb{E}[T] = \mathbb{E}\left[\sum_{t=1}^{\tau} T_t\right] = \mathbb{E}\left[\tau\right]\mathbb{E}[T_1] = \mathcal{O}\left(\frac{\bar{d}(G)mn^{w(G)+1}|\Gamma|^2 R^2 \log(n)}{\epsilon^2}\right).$$

□