MULTI-MARGINAL OPTIMAL TRANSPORT AND SCHRÖDINGER BRIDGES ON TREES

ISABEL HAASLER†, AXEL RINGH‡, YONGXIN CHEN§, AND JOHAN KARLSSON†

Abstract. The optimal transport problem has recently developed into a powerful framework for various applications in estimation and control. Many of the recent advances in the theory and application of optimal transport are based on regularizing the problem with an entropy term, which connects it to the Schrödinger bridge problem and thus to stochastic optimal control. Moreover, the entropy regularization makes the otherwise computationally demanding optimal transport problem feasible even for large scale settings. This has lead to an accelerated development of optimal transport based methods in a broad range of fields. Many of these applications have a underlying graph structure, for instance information fusion and tracking problems can be described by trees. In this work we consider multi-marginal optimal transport problems with a cost function that decouples according to a tree structure. The entropy regularized multi-marginal optimal transport problem can be viewed as a generalization of the Schrödinger bridge problem on the same tree, and by utilizing these connections we extend the computational methods for the classical optimal transport problem in order to solve structured multi-marginal optimal transport problems in an efficient manner. In particular, the algorithm requires only matrix-vector multiplications of relatively small dimensions. We show that the multi-marginal regularization introduces less diffusion, compared to the commonly used pairwise regularization, and is therefore more suitable for many applications. Numerical examples illustrate this, and we finally apply the proposed framework for tracking of an ensemble of indistinguishable agents.

Key words. Multi-marginal optimal transport, Schrödinger bridge, Hidden Markov chain, graph signal processing, ensemble estimation.

1. Introduction. An optimal transport problem is to find a transport plan that minimizes the cost of moving the mass of one distribution to another distribution [52]. Historically this problem has been important in economics and operations research, but as a result of recent progress in the area it has become a popular tool in a wide range of fields such as control theory [1, 6, 14, 32, 42, 54], signal processing [28, 38], computer vision [24, 48], and machine learning [2, 4, 43].

An extension to the standard optimal transport framework is multi-marginal optimal transport [44], which seeks a transport plan between not only two, but several distributions. In this work we consider multi-marginal optimal transport problems with cost functions that decouple according to a tree structure. We refer to such a problem as a multi-marginal optimal transport problem over a tree. This generalizes many structures that are commonly used in applications of optimal transport. For example, a path tree naturally appears in tracking and interpolation applications [8, 16, 48]. Similarly, star trees are used in barycenter problems, which occur for instance in information fusion applications [20, 27].

The optimal transport problem can be formulated as a linear program. However, in many practical applications the problem is too large to be solved directly. For the bi-marginal case, these computational limitations have recently been alleviated by regularizing the problem with an entropy term [19]. The optimal solution to the reg-
ularized optimal transport problem can then be expressed in terms of dual variables, which can be efficiently found by an iterative scheme, called Sinkhorn iterations [47]. For the multi-marginal case, although the Sinkhorn iterations can be generalized in a straightforward fashion [7], the computational complexity of the scheme increases dramatically with the number of marginals. Thus Sinkhorn iterations alone are not sufficient to address many multi-marginal problems. However, in specific settings, structures in the cost function can be exploited in order to derive computationally feasible methods, e.g., for computing Euler flows [7] and in tracking and information fusion applications [28].

Unregularized multi-marginal optimal transport problems with transport cost that decouples according to a tree structure can be equivalently formulated as a sum of coupled pairwise optimal transport problems, as for instance, barycenter problems [3, 7, 20, 48] and tracking and interpolation problems [8, 16, 26]. Typically these problems are solved using pairwise regularization (see, e.g., [7, Sec. 3.2]). However, we have empirically observed in some applications that the multi-marginal formulation yields favourable solutions compared to a corresponding pairwise optimal transport estimate [28]. In this work we develop a framework for solving tree-structured optimal transport problems with a multi-marginal regularization. For these problems, we show that the Sinkhorn algorithm can be performed in an efficient way, requiring only successive matrix-vector multiplications of relatively small size compared to that of the original multi-marginal problem. Thus we extend the computational results from [28] to general trees.

The entropy regularized formulation of optimal transport is connected to another classical topic, the Schrödinger bridge problem [12, 13, 39, 40]. Schrödinger was interested in determining the most likely evolution of a particle cloud observed at two time instances, where the particle dynamics have deviated from the expected Brownian motion [46]. Schrödinger showed that this particle evolution can be characterized as the one out of all the theoretically possible ones, that minimizes the relative entropy to the Wiener measure. Moreover, this optimal solution may be found by solving a so-called Schrödinger system, which turns out to be tightly connected to the Sinkhorn iterations for the entropy regularized optimal transport problem [12, 39]. This framework has been used in robust and stochastic control problems [14, 41, 53]. A version of the Schrödinger bridge problem, that is discrete in both time and space, can be formulated by modeling the evolutions of a number of particles as a Markov chain [31, 45]. In the infinite particle limit, the maximum likelihood solution of the Markov process can then be approximated by solving a relative entropy problem [15]. An analogous approach has recently been used to develop a framework for modelling ensemble flows on Hidden Markov chains [34]. Other optimal transport based state estimation problems for a continuum of agents and in continuous time have been considered in [10, 11, 16]. For further work on ensemble controllability and observability see, e.g., [9, 57].

Based on [34], we extend the discrete Schrödinger bridge in [45] to trees. In particular, we derive a maximum likelihood estimate for Markov processes defined on the edges of a rooted directed tree. Interestingly, it turns out that the solution to the entropy regularized multi-marginal optimal transport problem on a tree corresponds to the solution of the generalized Schrödinger bridge on the same tree. This generalizes the established equivalence of the classical bi-marginal entropy regularized optimal transport problem and the Schrödinger bridge problem [39]. Moreover, this gives an additional motivation for using multi-marginal optimal transport instead of the often used pairwise optimal transport for tree structured problems.
This paper is organized as follows: Section 2 is an introduction to the problems of optimal transport and Schrödinger bridges. The main results are presented in Section 3 and 4. In Section 3 we define the multi-marginal optimal transport problem on trees, and provide an algorithm for efficiently solving its entropy regularized version. In Section 4 we generalize the Schrödinger bridge problem to trees, and show that it is equivalent to an entropy regularized multi-marginal optimal transport problem on the same tree. Section 5 is on the discrepancy between the multi-marginal and pairwise optimal transport formulations on trees. A larger numerical simulation is detailed in Section 6, where the proposed framework is used to estimate ensemble flows from aggregate and incomplete measurements. The proofs are deferred to the Appendix.

2. Background. In this section we summarize some background material on optimal transport and Schrödinger bridges. This is also used to set up the notation. To this end, first note that throughout we let \( \exp(\cdot), \log(\cdot), \odot, \) and \( ./ \) denote the elementwise exponential, logarithm, multiplication, and division of vectors, matrices, and tensors, respectively. Moreover, \( \otimes \) denotes the outer product. By \( 1 \) we denote a vector of ones, the size of which will be clear from the context.

2.1. Optimal transport. Let \( \mu_1 \) and \( \mu_2 \) denote two nonnegative distributions with equal mass defined on the space \( X \subset \mathbb{R}^d \). The optimal transport problem is to find a mapping that moves the mass from \( \mu_1 \) to \( \mu_2 \), while minimizing the total transport cost. Here the transport cost is defined in terms of a continuous underlying cost function \( c : X \times X \to \mathbb{R}_+ \), where \( c(x_1, x_2) \) denotes the cost of moving a unit mass from \( x_1 \in X \) to \( x_2 \in X \). Analogously, a transport plan \( m \in \mathcal{M}_+(X \times X) \) is a nonnegative distribution on \( X \times X \), where \( m(x_1, x_2) \) represents the amount of mass moved from \( x_1 \in X \) to \( x_2 \in X \). The optimal transport plan from \( \mu_1 \) to \( \mu_2 \) is then a minimizing solution of

\[
\min_{m \in \mathcal{M}_+(X \times X)} \int_{X \times X} c(x_1, x_2) dm(x_1, x_2)
\]

subject to
\[
\int_{x_2 \in X} dm(x_1, x_2) = d\mu_1(x_1)
\]
\[
\int_{x_1 \in X} dm(x_1, x_2) = d\mu_2(x_2).
\]

Multi-marginal optimal transport extends the concept of the classical optimal transport problem (2.1) to the setting with a set of marginals \( \mu_1, \ldots, \mu_J \), where \( J \geq 2 \) \cite{7, 28, 44}. In this setting, the transport cost in (2.1) is described by the continuous function \( c : X^J \to \mathbb{R}_+ \) and a transport plan \( m \in \mathcal{M}_+(X^J) \) is a nonnegative measure on the product space \( X^J \). For \( x = (x_1, \ldots, x_j) \in X^J \), the value \( c(x) \) denotes the transport cost for a unit mass corresponding to the tuple \( x_1, \ldots, x_j \), and similarly \( m(x) \) represents the amount of transported mass associated with this tuple. The multi-marginal optimal transport problem reads

\[
\min_{m \in \mathcal{M}_+(X^J)} \int_{X^J} c(x) dm(x)
\]

subject to
\[
\int_{x_{\setminus j} \in X^{J-1}} dm(x) = d\mu_j(x_j), \text{ for } j \in \Gamma
\]

where \( x_{\setminus j} = (x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_J) \), and \( \Gamma \) is an index set. In the original multi-marginal optimal transport formulation, constraints are typically given on all
marginals, i.e., for the index set $\Gamma = \{1, 2, \ldots, J\}$. However, in this work we allow for the more general case, where constraints are only imposed on a subset of marginals, i.e., $\Gamma \subset \{1, 2, \ldots, J\}$. Note that the standard bi-marginal optimal transport problem (2.1) is a special case of the multi-marginal optimal transport problem (2.2), where $J = 2$ and $\Gamma = \{1, 2\}$.

2.2. Entropy regularized optimal transport. To solve the optimal transport problem (2.2) in practice, it is often discretized. For notational convenience we will use the same number of discretization points, $n$, for all marginals. The marginal distributions are then described by nonnegative vectors $\mu_j \in \mathbb{R}^n_+$, and the cost function $c(\cdot)$ and transport plan $m(\cdot)$ are represented by nonnegative $J$-mode tensors $C, M \in \mathbb{R}^{n \times \cdots \times n}_+$. This transforms the optimal transport problem (2.2) into the finite-dimensional linear program

$$\begin{align*}
\text{minimize} \quad & \langle C, M \rangle \\
\text{subject to} \quad & P_j(M) = \mu_j, \quad \text{for} \quad j \in \Gamma,
\end{align*}$$

where $\langle C, M \rangle = \sum_{i_1, \ldots, i_J} C_{i_1, \ldots, i_J} M_{i_1, \ldots, i_J}$, and the projection on the $j$-th marginal of $M$ is computed as

$$P_j(M) = \sum_{i_1, \ldots, i_{j-1}, i_{j+1}, \ldots, i_J} M_{i_1, \ldots, i_{j-1}, i_j, i_{j+1}, \ldots, i_J}.$$

Although linear, the number of variables in this multi-marginal problem (2.3) is often too large to be solved directly. A popular approach for the bi-marginal setting to bypass the size of the problem has been to add a regularizing entropy term to the objective. In theory the same approach can be used also for the multi-marginal case.

**Definition 2.1 ([33, Ch. 4]).** Let $p$ and $q$ be two nonnegative vectors, matrices or tensors of the same dimension. The normalized Kullback-Leibler (KL) divergence of $p$ from $q$ is defined as $H(p || q) := \sum_i (p_i \log (p_i/q_i) - p_i + q_i)$, where $0 \log 0$ is defined to be $0$. Similarly, the entropy of $p$ is then defined as $H(p) := H(p || 1) = \sum_i (p_i \log (p_i) - p_i + 1)$.

Note that $H(p || q)$ is jointly convex over $p, q$. For more properties and interpretation of the KL divergence see, e.g., [17, 18].

The entropy regularized multi-marginal optimal transport problem is the convex problem

$$\begin{align*}
\text{minimize} \quad & \langle C, M \rangle + \epsilon H(M) \\
\text{subject to} \quad & P_j(M) = \mu_j, \quad \text{for} \quad j \in \Gamma,
\end{align*}$$

where $\epsilon > 0$ is a regularization parameter. For the bi-marginal case (2.1), where the cost and mass transport tensors are matrices, the discretized and entropy regularized problem reads

$$T_\epsilon(\mu_1, \mu_2) := \text{minimize} \quad \text{trace}(C^T M) + \epsilon H(M)$$

$$\text{subject to} \quad M \mathbf{1} = \mu_1, \quad M^T \mathbf{1} = \mu_2.$$

It is well established that the entropy regularized bi-marginal optimal transport problem is connected to the Schrödinger bridge problem [12, 13, 39, 40], which is introduced
in Section 2.3. More importantly from a computational perspective, the introduction of the entropy term in problem (2.6) allows for expressing the optimal solution \( M \) in terms of the Lagrange dual variables, which may be efficiently computed by Sinkhorn iterations [19]. This procedure can be generalized to the setting of multi-marginal optimal transport [7].

In particular, for the multi-marginal entropy regularized optimal transport problem (2.5) it can be shown that the optimal solution is of the form [28]

\[
M = K \odot U,
\]

where \( K = \exp(-C/\epsilon) \) and \( U \) can be decomposed as

\[
U = u_1 \otimes u_2 \otimes \cdots \otimes u_J
\]

with \( u_j \in \mathbb{R}^n \) given by

\[
u_j = \begin{cases} \exp(\lambda_j/\epsilon), & \text{if } j \in \Gamma \\ 1, & \text{else.} \end{cases}
\]

Here \( \lambda_j \in \mathbb{R}^n \) for \( j \in \Gamma \) are optimal dual variables in the dual problem of (2.5), given by

\[
\begin{align*}
\max \quad & -\epsilon(\mathbf{K}, \mathbf{U}) + \sum_{j \in \Gamma} \lambda_j^T \mu_j, \\
\text{s.t.} \quad & \lambda_j \in \mathbb{R}^n, j \in \Gamma, \quad \langle K, U \rangle.
\end{align*}
\]

where \( U \) is specified by (2.8) and (2.9). Note that the dual variable \( \lambda_j \) corresponds to the constraint on the \( j \)-th marginal. For details the reader is referred to, e.g., [7,28].

The Sinkhorn scheme for finding \( U \) in (2.8), is to iteratively update \( u_j \) according to

\[
u_j \leftarrow u_j \odot \mu_j / P_j(M \odot U),
\]

for all \( j \in \Gamma \). This scheme may for instance be derived as Bregman projections [7] or a block coordinate ascend in the dual (2.10), [28,37,51]. As a result, global convergence of the Sinkhorn scheme (2.11) is guaranteed [5]. For the sake of completeness, we provide more details on the convergence in our presentation (see Corollary 3.4). We also note that (2.11) reduces to standard Sinkhorn iterations,

\[
u_1 \leftarrow \mu_1 / (K u_2), \quad u_2 \leftarrow \mu_1 / (K^T u_1),
\]

for the two-marginal case (2.6). The iterations (2.12) converge linearly to an optimal solution, which is unique up to multiplication with a constant [12,30].

The computational bottleneck of the Sinkhorn iterations (2.11) is computing the projections \( P_j(M) \), for \( j \in \Gamma \), which in general scales exponentially in \( J \). In fact, even storing the tensor \( M \) is a challenge as it consists of \( n^J \) elements. However, in many cases of interest, structures in the cost tensors can be exploited to make the computation of the projections feasible. In [28] this is shown on the example of cost functions that decouple sequentially, centrally, or a combination of both. Computing the projections requires then only repeated matrix vector multiplications. In this work, we show that these efficient methods can be generalized to the setting where the cost tensor decouples according to a tree structure, that is, when the marginals of the optimal transport problem are associated with the nodes of a tree, and cost matrices are defined on its edges (see Section 3).
2.3. Schrödinger bridge problem. The Schrödinger bridge problem is to determine the most likely evolution of a particle cloud observed at two time instances [46]. In the case that the second observation cannot be explained as a Brownian motion of the initially observed particle cloud, Schrödinger aimed to find the most likely particle evolution connecting, hence bridging, the two distributions.

A useful mathematical framework to solve this problem, however not yet developed in Schrödinger’s time, is the theory of large deviations, which studies so called rare events, meaning deviations from the law of large numbers [21, 22]. The probability of these rare events approaches zero, as the number of trials goes to infinity, and large deviation theory analyzes the rate of this decay, which can often be expressed as the exponential of a so called rate function [22, 25].

For a large deviation interpretation of the Schrödinger bridge [29, Sec. II.1.3], the particle evolutions are modelled as independent identically distributed random variables on path space, and the Schrödinger bridge is the probability measure $P$ on path space that is most likely to describe the rare event of observing the two particle distributions. Let $W$ be the Wiener measure, which corresponds to the probability law of the Brownian motion. Then the Schrödinger bridge $P$ is found by minimizing the corresponding rate function

$$\int \log \left( \frac{dP}{dW} \right) dP,$$

over all probability measures that are absolutely continuous with respect to $W$ and have the given particle cloud distributions as marginals. A space and time discrete Schrödinger bridge problem for Markov chains is treated in [31,45].

We follow the exposition in [34], where the space and time discrete Schrödinger bridge has been derived as a maximum likelihood approximation for Markov chains. Consider a cloud of $N$ particles and assume that each particle evolves according to a Markov chain. Denote the states of the Markov chain by $X = \{X_1, X_2, \ldots, X_n\}$ and let the transition probability matrix be $A^j \in \mathbb{R}^{n \times n}$, with elements $A^j_{kl} = P(q_{j+1} = X_l | q_j = X_k)$, where $q_j$ denotes the state at time $j$. Let the vector $\mu_j$ describe the particle distribution over the discrete state space $X$ at time $j$, for $j = 1, \ldots, J$. As in the optimal transport framework we define the mass transport matrix $M^j$ where element $M^j_{jt}$ describes the number of particles transitioning from state $k$ at time $j$ to state $t$ at time $j + 1$.

Note that given the distribution $\mu_j$ at time instance $j$, the expected transport matrix $M^j_{jt}$, according to the transition probability matrix $A^j$, is given by $\text{diag}(\mu_j)A^j$. In case the observed distribution $\mu_{j+1}$ differs from the expected distribution $(A^j)^T \mu_j$, one may pose a question similar to Schrödinger’s: what is the most likely mass transport plan that describes the true underlying state transitions? Note that in the language of large deviation theory, the event of observing the distribution $\mu_{j+1} \neq (A^j)^T \mu_j$ may be understood as a rare event. Indeed, the probability for the event that a given mass transfer matrix $M^j$ describes the underlying particle dynamics, satisfies a large deviation principle with rate function $H(\cdot | \text{diag}(\mu_j)A^j)$. That is, if the number of particles is large, then the log-likelihood of a transfer matrix $M$ to describe the particle dynamics can be approximated in terms of this rate function. Informally we write this as

$$P_{\mu_j,A^j}(M^j) \sim e^{-H(M^j | \text{diag}(\mu_j)A^j)},$$

where, given an initial distribution $\mu_j$ and a transition matrix $A^j$, the probability of
a mass transfer matrix $M^j$ to describe the underlying dynamics is

$$P_{\mu_j,A_j}(M^j) = \prod_{k=1}^{n} \left( M^j_{k1}, M^j_{k2}, \ldots, M^j_{kn} \right)^{A^j_{ki}}.$$ 

More precisely, the result proved in [34] is as follows$^1$.

**Proposition 2.2.** [34, Proposition 1] Given $A$, let $\mu_0^{(N)} \in \mathbb{N}^n$ be a sequence of distributions with $N$ particles, and $M^{(N)} \in \mathbb{N}^{n \times n}$ be a sequence of mass transfer matrices such that $M^{(N)}1 = \mu_0^{(N)}$ and $\text{supp}(M^{(N)}) \subseteq \text{supp}(\text{diag}(\mu_0^{(N)})A)$. Then there exists a constant $C > 0$, such that

$$| \log \left( P_{\mu_0^{(N)},A}(M^{(N)}) \right) + H \left( M^{(N)}|\text{diag}(\mu_0^{(N)})A \right) | \leq C \log(N).$$

Assuming that the initial and final marginal, $\mu_1$ and $\mu_J$, are given, these large deviation theoretic considerations motivate the formulation of an optimization problem to find the most likely mass transfer matrices $M^j$, for $j = 1, \ldots, J - 1$, and intermediate distributions $\mu_j$, for $j = 2, \ldots, J - 1$, as

$$\text{minimize } \sum_{j=1}^{J-1} H(M^j | \text{diag}(\mu_j)A^j)$$

subject to $M^j1 = \mu_j$, $(M^j)^T1 = \mu_{j+1}$, for $j = 1, \ldots, J - 1$.

Indeed, this problem is equivalent to the discrete time and space Schrödinger bridge problem in [45] (for details see [34] or Section 5.1).

An extension to a Hidden Markov Model formulation with aggregate indirect observations of the distributions was described in [34]. In this article the framework is extended to general tree structures, where each vertex is associated with a distribution, and each edge with a Markov process.

### 3. Multi-marginal optimal transport on trees.

In this section we introduce multi-marginal optimal transport problems on trees. For the entropy regularized version of these problems, the projections (2.4) can be computed by only matrix-vector multiplications. This yields an efficient Sinkhorn algorithm, which we present in this section. Moreover, some properties of multi-marginal optimal transport problems on trees are discussed.

**Definition 3.1.** A graph $T = (\mathcal{V}, \mathcal{E})$, with vertices $\mathcal{V}$ and edges $\mathcal{E}$, is a tree if it is acyclic and connected [23]. The vertices with degree 1 are called leaves, and we denote the set of leaves by $\mathcal{L}$. For a vertex $j \in \mathcal{V}$, the set of neighbours $\mathcal{N}_j$ is defined as the set of vertices, which have a common edge with $j$. The path between two vertices $j_1$ and $j_L$ is a sequence of edges that connect $j_1$ and $j_L$, such that all edges are distinct. We also denote a path by the set of intermediate vertices.

Let $T = (\mathcal{V}, \mathcal{E})$ be a tree. In this section we consider a multi-marginal optimal transport problem where the marginals correspond to the nodes of the tree, i.e., $\mathcal{V} = \{1,2,\ldots,J\}$. Assume that the cost function decouples as

$$c(x_1, \ldots, x_J) = \sum_{(j_1,j_2) \in \mathcal{E}} c^{(j_1,j_2)}(x_{j_1}, x_{j_2}),$$

$^1$Note that therein the result is proved for the classical KL divergence, i.e., $H(p|q) = \sum_i p_i \log(p_i/q_i)$. The result follows for the normalized KL divergence as in Definition 2.1, since $\sum_{k,j} (\mu_j)_k A^j_{kl} = \sum_{k,j} (\mu_j)_k = \sum_{k,l} M^j_{kl}$.
Hence, the projection on the first marginal of \( c \in \mathbb{R}^j \) is a cost function characterizing the cost of transportation between marginal \( j_1 \) and \( j_2 \), for all \( (j_1, j_2) \in \mathcal{E} \). Note that the cost on edge \( (j_1, j_2) \in \mathcal{E} \) can be interchangeably expressed by \( c_{j_1, j_2}(\cdot, \cdot) \) without changing the cost \( c(x_1, \ldots, x_J) \) by letting \( c_{j_1, j_2}(x_{j_1}, x_{j_2}) = c_{j_1, j_2}(x_{j_1}, x_{j_2}) \). We refer to problem (2.2) with a cost of the form (3.1) as a multi-marginal optimal transport problem on the tree \( T \). In the discretized form (2.3), the cost tensor \( C \in \mathbb{R}_+^{n_j} \) decouples as

\[
C_{i_1, \ldots, i_j} = \sum_{(j_1, j_2) \in \mathcal{E}} C_{i_1, j_1}^{(j_1, j_2)} C_{j_2, i_2}^{(j_1, j_2)} \]

where \( C_{i_1, j_1}^{(j_1, j_2)} \in \mathbb{R}_+^{n_j} \) is a cost matrix representing \( c_{j_1, j_2}(\cdot, \cdot) \), for all \( (j_1, j_2) \in \mathcal{E} \). Note that since \( c_{j_2, j_1}(x_{j_2}, x_{j_1}) = c_{j_1, j_2}(x_{j_1}, x_{j_2}) \), it holds that \( C_{j_2, j_1}^{(j_1, j_2)} = (C_{j_1, j_2}^{(j_1, j_2)})^T \).

In the following, we consider discretized and entropy regularized multi-marginal optimal transport on trees. The transport plan can then be expressed as \( M = K \odot U \), where \( K = \exp(-C/\epsilon) \) (cf. (2.7)). Due to the structured cost (3.2), this tensor decouples as

\[
K_{i_1, \ldots, i_j} = \prod_{(j_1, j_2) \in \mathcal{E}} K_{i_1, j_1}^{(j_1, j_2)} K_{j_2, i_2}^{(j_1, j_2)}
\]

with the matrices defined as \( K_{i_1, j_1}^{(j_1, j_2)} = \exp(-C_{i_1, j_1}^{(j_1, j_2)} / \epsilon) \), for \( (j_1, j_2) \in \mathcal{E} \). We show that for these problems the projections (2.4) can therefore be computed by successive matrix vector multiplications. As the computation of the projections account for the bottleneck of the Sinkhorn iterations (2.11), this yields an efficient algorithm for solving entropy regularized multi-marginal optimal transport problem on trees.

We first illustrate the computation of the projections on a small example.

**Example 1.** Consider the tree \( T = (\mathcal{V}, \mathcal{E}) \) with vertices \( \mathcal{V} = \{1, 2, 3, 4\} \) and edges \( \mathcal{E} = \{(1, 2), (2, 3), (1, 4)\} \) as depicted in Figure 1. Assume that the cost in problem (2.5) decouples as

\[
C_{i_1, i_2, i_3, i_4} = C_{i_1, i_2}^{(1, 2)} + C_{i_2, i_3}^{(2, 3)} + C_{i_3, i_4}^{(3, 4)}
\]

where \( C_{i_1, i_2}^{(j_1, j_2)} \), for \( (j_1, j_2) \in \mathcal{E} \), are cost matrices defined on the respective edges. The transport tensor is then of the form \( M = K \odot U \), where \( K = \exp(-C/\epsilon) \) and \( U = u_1 \odot \cdots \odot u_4 \). Thus, denoting \( K_{i_1, i_2}^{(j_1, j_2)} = \exp(-C_{i_1, i_2}^{(j_1, j_2)} / \epsilon) \) it holds that

\[
K_{i_1, i_2, i_3, i_4}^{(j_1, j_2)} = K_{i_1, i_2}^{(1, 2)} K_{i_2, i_3}^{(2, 3)} K_{i_3, i_4}^{(3, 4)}
\]

Hence, the projection on the first marginal of \( K \odot U \) can be computed as

\[
P_{1}(K \odot U)_{i_1} = \sum_{i_2, i_3, i_4} K_{i_1, i_2, i_3, i_4} U_{i_1, i_2, i_3, i_4}
\]

\[
= (u_1)_{i_1} \sum_{i_2, i_3, i_4} K_{i_1, i_2}^{(1, 2)} K_{i_2, i_3}^{(2, 3)} K_{i_3, i_4}^{(3, 4)} u_2(i_2) u_3(i_3) u_4(i_4).
\]
The sum can be computed as successive matrix-vector multiplications, starting from the leaves of the tree. In particular, denote the products starting in the two leaves
\[ \sum_{i_3} K^{(2,3)}(u_3)_{i_3} = \left( K^{(2,3)} u_3 \right)_{i_2} =: (\alpha_{(2,3)})_{i_2}, \]
\[ \sum_{i_4} K^{(1,4)}(u_4)_{i_4} = \left( K^{(1,4)} u_4 \right)_{i_1} =: (\alpha_{(1,4)})_{i_1}. \]

On the lower branch of the tree, we have thereby brought the expression to a vector indexed by \( i_1 \). On the upper branch, another multiplication, corresponding to the edge \((1,2)\), is required. This leads to defining
\[ \sum_{i_2} K^{(1,2)}(u_2 \odot \alpha_{(2,3)})_{i_2} = \left( K^{(1,2)}(u_2 \odot \alpha_{(2,3)}) \right)_{i_1} =: (\alpha_{(1,2)})_{i_1}. \]

The full expression for the projection thus reads
\[ P_1(K \odot U) = u_1 \odot K^{(1,2)}(u_2 \odot K^{(2,3)} u_3) \odot K^{(1,4)} u_4 = u_1 \odot \alpha_{(1,2)} \odot \alpha_{(1,4)}. \]

The other marginals can be computed similarly by performing the corresponding matrix vector products, starting from the leaves of the tree.

Illustratively, the procedure in Example 1 for projecting the information of the full tensor down to one marginal can be understood as passing down the information from all branches of the tree to the desired vertex, where we interpret the vectors \( u_j \) as local information in each node. Starting from the leaves of the tree, this is done according to the following rules:

1. Information is passed down the edge \((j_1, j_2) \in E\) by multiplication with the matrix \( K^{(j_1,j_2)} \).
2. Information is collected in the node \( j \in V \) by elementwise multiplication of the vector \( u_j \), which contains local information, with information passed down from the connected branches.

The following theorem shows that these rules hold for any tensor of the form \( K \odot U \), with \( U = u_1 \otimes \cdots \otimes u_j \) and \( K \) decoupling according to an arbitrary tree.

**Theorem 3.2.** Let \( K = \exp(-C/\epsilon) \) with \( C \) as in (3.2) for the tree \( T = (V, E) \), and let \( U = u_1 \otimes u_2 \otimes \cdots \otimes u_J \). Define \( K^{(j_1,j_2)} = \exp(-C^{(j_1,j_2)}/\epsilon) \), for \((j_1,j_2) \in E\). Then the projection on the \( j \)-th marginal of \( K \odot U \) is of the form
\[ P_j(K \odot U) = u_j \odot \bigodot_{k \in N_j} \alpha_{(j,k)}. \]

The vectors \( \alpha_{(j,k)} \) can be computed recursively down the branches of the tree, according to
\[ \alpha_{(j,k)} = K^{(j,k)} \left( u_k \odot \bigodot_{\ell \in N_k \setminus \{j\}} \alpha_{(k,\ell)} \right). \]

Analogously to the marginal projections (2.4), we define the pairwise projections on marginal \( j_1 \) and \( j_2 \) as
\[ P_{j_1,j_2}(M) = \sum_{i_1, \ldots, i_J \setminus \{i_{j_1}, i_{j_2}\}} M_{i_1, \ldots, i_J}. \]
Algorithm 3.1 Sinkhorn method for the multi-marginal optimal transport problem on a tree.

Given: Initial guess $u_j$, for $j \in L$

for $(j_1, j_2) \in \mathcal{E}$ do
  Initialize $\alpha(j_1, j_2)$ according to (3.4)
end for

Initialize $j \in L$

while Sinkhorn not converged do
  for $(j_1, j_2) \in \mathcal{E}$ on the path from $j$ to $(j - 1 \mod |L|)$ do
    Update $\alpha(j_1, j_2)$ according to (3.4)
  end for

  $u_j \leftarrow u_j \odot \mu_j / \bigodot_{k \in N_j} \alpha(j, k)$

  $j \leftarrow j + 1 \mod |L|$
end while

return $u_j$ for $j \in L$

These pairwise projections can be expressed similarly to the marginal projections in Theorem 3.2.

Proposition 3.3. Let the assumptions in Theorem 3.2 hold, let $j_1, j_L \in \mathcal{V}$, and let $j_1, j_2, \ldots, j_L$ denote the path between $j_1$ and $j_L$. Then the pairwise projection of $K \odot U$ on the marginals $j_1$ and $j_L$, denoted by $P_{j_1, j_L}(K \odot U)$, is

$$
\left(\prod_{\ell=1}^{L-1} \text{diag} \left( u_{j_\ell} \odot \bigodot_{k \in \mathcal{N}_{j_\ell} \setminus \{j_1, \ldots, j_{\ell-1}\}} \alpha(j, k) \right) K_{(j_\ell, j_{\ell+1})} \right) \text{diag} \left( u_{j_L} \odot \bigodot_{k \in \mathcal{N}_{j_L} \setminus \{j_L-1\}} \alpha(j_L, k) \right),
$$

where $\alpha(j_1, j_2)$, for $(j_1, j_2) \in \mathcal{E}$, are defined as in Theorem 3.2.

Multi-marginal optimal transport problems on some specific trees have previously been introduced in [28]. It is worth noting that the expressions for the projections of the mass tensor $\hat{M}$ for the examples in [28] are special cases of the results in Theorem 3.2 and Proposition 3.3. For instance, in the tracking problem the tree $T$ is a path tree, and in the information fusion problem it is a star tree.

The expressions for the projections in Theorem 3.2 can be used to solve a multi-marginal optimal transport problem on a tree $T = (\mathcal{V}, \mathcal{E})$ by a Sinkhorn method as detailed in (2.11). Note that in such an iterative method, in each iteration step some of the factors $\alpha(j_1, j_2)$ do not change. Specifically, between the update of $u_{j_1}$ and $u_{j_2}$ only the factors on all edges that lie on the path between nodes $j_1$ and $j_2$ need to be updated. The full method for solving multi-marginal optimal transport problems on trees is summarized in Algorithm 3.1. In fact, this is a globally convergent algorithm, as detailed below.

Corollary 3.4. Let $\{u_j^k\}_{j \in \mathcal{L}}$ be the set of vectors after the $k$th iteration in the while-loop in Algorithm 3.1. Moreover, let $U^k = u_1^k \odot u_2^k \odot \cdots \odot u_J^k$ as in (2.8), let $\{\lambda_j^k\}_{j \in \mathcal{E}}$ be the corresponding vectors as in (2.9), and let $\hat{M}^k = K \odot U^k$ as in (2.7). Then, as $k \to \infty$,

- $\hat{M}^k$ converges to an optimal solution to (2.3),
- $-\epsilon(K, U^k) + \sum_{j \in \mathcal{V}} (\lambda_j^k)^T \mu_j$ converges to the optimal value of (2.10) and of (2.3),
the sequence $U^k$ is bounded, and all limit points of the sequence are optimal solutions to (2.10).

It may be noted that the updates can be scheduled depending on the tree structure in order to decrease the computational complexity of Algorithm 3.1. In particular they should be scheduled such that the paths between any two nodes that are updated successively are short. If the direct path between node $j-1$ and node $j$ is of length $p$, then the update of the vector $u_j$ requires $p$ matrix-vector multiplications. The complexity of one update is thus of the order $O(pn^2)$. In case there are additional structures in the cost matrices $C(j_1,j_2)$, for $(j_1,j_2) \in \mathcal{E}$, e.g., they describe the squared Euclidean distance, the cost of the matrix-vector multiplications can be decreased further, see, e.g., [28, Rem. 4], [37, Rem. 3.11].

Moreover, it should be noted that the set of given marginals in Algorithm 3.1 corresponds to the set of leaves, i.e., $\Gamma = \mathcal{L}$. The factors $u_j$, for $j \in \mathcal{V} \setminus \Gamma$, can therefore be neglected in Theorem 3.2 and Proposition 3.3. We can assume without loss of generality that $\Gamma = \mathcal{L}$, since otherwise one can solve a multi-marginal optimal transport problem on a subtree (or a set of subtrees), where the marginals are known exactly on the set of leaves of the subtree, in order to fully describes the solution to the original problem.

In the case where the marginal on one of the leaves is not given, no information is passed down on the corresponding branch. Specifically, assume that $\ell \in \mathcal{L}$, where $\mu_\ell$ is unknown. Recall from the definition of the tensor $U$ in (2.9) for the multi-marginal optimal transport problem, that $u_j = 1$ for all $j \in \mathcal{V} \setminus \Gamma$. Thus, for $k$ such that $(k,\ell) \in \mathcal{E}$, the corresponding vector (3.4) is $\alpha_{(k,\ell)} = K^{(k,\ell)}1$, which is constant and not updated in Algorithm 3.1. The same argument can be extended for adjoining edges as long as the degree is 2, thus no information is gained from other leaves.

In the case that an inner marginal is known, the tree $T$ may be decoupled into smaller subtrees, for which the regularized optimal transport problems can be solved separately. This is due to the fact that the marginal $P_j(K \odot U)$ on a vertex $j \in \mathcal{V}$, given the marginals on the neighboring vertices $P_\ell(K \odot U)$, for $\ell \in \mathcal{N}_j$, is independent of all other marginals. To see this, straightforward computations imply that

$$(P_{j,k}(K \odot U))_{i,j,k,l} = (P_{j,k}(K \odot U))_{i,j,k,l} (P_{k,l}(K \odot U))_{i,k,l},$$

whenever $k$ lies on the path between $j$ and $\ell$. In particular, for any fixed $i_k$, the matrix $(P_{j,k}(K \odot U))_{i,j,k,l}$ is of rank 1. Hence, given $P_k(K \odot U)$, the marginals $P_j(K \odot U)$ and $P_\ell(K \odot U)$ have no influence on each other. Thus, if the tensor $K$ is normalized to be a discrete probability distribution, it defines a Markov random field.

Every Markov process is also a reciprocal process [35], which were first introduced in the context of studying Schrödinger bridges [36]. Recall that the maximum likelihood estimation problem for a Markov chain in Section 2.3 is equivalent to a discrete Schrödinger bridge. This suggests a connection between the multi-marginal optimal transport problem and the maximum likelihood estimation problem in Section 2.3. In fact, under certain conditions a generalization of the Markov chain problem (2.13) to trees, which is introduced in Section 4.1, yields the same solution as the regularized multi-marginal optimal transport problem (2.5) on the same tree, as we will see in Section 4.2.

4. Multi-marginal optimal transport vs. Schrödinger bridge on trees.

In this section, we extend the Schrödinger bridge problem which is defined on a path tree to a formulation valid on an arbitrary tree graph. We then show that under
certain conditions this problem yields a solution which is equivalent to the solution of the entropy regularized multi-marginal optimal transport problem (2.5) on the same tree, thus extending the well-known equivalence from the bi-marginal optimal transport problem and the Schrödinger bridge problem on a path tree [12,13,39,40].

4.1. Generalized Schrödinger bridges on trees. The notion of time of the Markov chain introduces a direction to the Schrödinger bridge problem, which can be seen directly in the optimization problem (2.13). This directionality needs to be taken into account when extending the problem to a general tree. To this end, consider a tree $T_r(V, E_r)$ that is rooted in a vertex $r \in L$, i.e., one of the leaves is defined to be the root of the tree. This defines a partial ordering on the tree, and we write $k < j$ (or $j > k$) if node $k$ lies on the path between $r$ and node $j$. To formulate the following results, we assume that all edges are directed according to this partial ordering, i.e., $j_1 < j_2$ for all $(j_1, j_2) \in E_r$. For a vertex $j \in V \setminus r$, its parent is then defined as the (unique) vertex $p(j)$ such that $(p(j), j) \in E_r$. In the following, without loss of generality we denote the root vertex by $r = 1$.

Let $A^{(j_1, j_2)}$ be the probability transition matrix on edge $(j_1, j_2) \in E_r$. Then, the Schrödinger bridge problem in (2.13) may be naturally extended to the tree structure as

\[
\begin{align*}
\text{minimize} & \quad \sum_{\mu_j, j \in V \setminus r} H \left( M^{(j_1, j_2)} \mid \text{diag}(\mu_j) A^{(j_1, j_2)} \right) \\
\text{subject to} & \quad M^{(j_1, j_2)} 1 = \mu_{j_1}, \quad (M^{(j_1, j_2)})^T 1 = \mu_{j_2}, \quad \text{for } (j_1, j_2) \in E_r.
\end{align*}
\] (4.1)

We can assume without loss of generality that $r = L$, by similar reasoning as in Section 3. The extension of (2.13) to (4.1) builds on the large deviation principle in Proposition 2.2, which requires that an initial distribution is known. This is why we have restricted our analysis to the case when the directed tree $T_r(V, E_r)$ is rooted in a leaf.

Similarly to the multi-marginal optimal transport problem, the optimal solution to (4.1) can be expressed in terms of its dual variables. In fact this is a natural generalization of the forward and backward propagation factors in the Schrödinger system [31,45] to tree graphs.

**Theorem 4.1.** Let $A^{(j_1, j_2)}$ be probability transition matrices $(j_1, j_2) \in E_r$ with strictly positive elements and assume that $1^T \mu_{j_1} = 1^T \mu_{j_2}$ for all leaves $j_1, j_2 \in L$. Then an optimal solution to (4.1) can be written as

\[
\begin{align*}
\mu_j &= \varphi_j \otimes \hat{\varphi}_j, \\
M^{(j_1, j_2)} &= \text{diag}(\hat{\varphi}_{j_1} \otimes \varphi_{j_1 \setminus j_2}) A^{(j_1, j_2)} \text{diag}(\varphi_{j_2}),
\end{align*}
\]

for a set of vectors $\varphi_j, \hat{\varphi}_j$, for $j \in V$, and $\varphi_{j_1 \setminus j_2}$ for $(j_1, j_2) \in E_r$.

Moreover, there exists a set of vectors $v_j$, for $j \in \Gamma$, such that the vectors $\varphi_j, \hat{\varphi}_j$ and $\varphi_{j_1 \setminus j_2}$ can be written as

\[
\begin{align*}
\varphi_j &= \begin{cases} v_j, & \text{if } j \in L \setminus \{1\} \\
A^{(j, k)} \varphi_k, & \text{else}, \end{cases} \\
\hat{\varphi}_j &= \begin{cases} 1^T v_1, & \text{if } j = 1 \\
(A^{(p(j), j)})^T (\hat{\varphi}_{p(j)} \otimes \varphi_{p(j) \setminus j}), & \text{else}, \end{cases}
\end{align*}
\] (4.2) (4.3)
and \( \varphi_{j_1 \setminus j_2} = \varphi_{j_1} \cdot (A^{(j_1,j_2)} \varphi_{j_2}) \).

Note that the vectors \( v_j \) for \( j \in L \) in Theorem 4.1 satisfy the nonlinear equations

\[
\begin{align*}
\varphi_1 \cdot v_1 &= \mu_1, \\
v_j \odot \hat{\varphi}_j &= \mu_j, \quad j \in L \setminus \{1\}.
\end{align*}
\]

When considering the Schrödinger bridge, which is the special case with two leaves, these equations correspond to the boundary conditions in the Schrödinger system [45, Thm. 4.1]. The factors \( \varphi_j \) and \( \hat{\varphi}_j \) can be seen as propagating information from the leaves to the other nodes in a similar manner as \( \alpha_{(j,k)} \) in Theorem 3.2.

Next, we show that the Schrödinger bridge problem on a tree, (4.1), is independent of the choice of root \( r \in L \). This follows as a corollary to the next proposition.

**Proposition 4.2.** Let \( \mathcal{T}_r = (\mathcal{V}, \mathcal{E}_r) \) be a rooted directed tree with root \( r \in L \). Then problem (4.1) is equivalent to the problem

\[
\begin{align*}
\min_{\mu_j, j \in \mathcal{V}\setminus L} & \sum_{j \in \mathcal{V}\setminus L} H \left( \frac{\mu_j}{M^{(j_1,j_2)}} A^{(j_1,j_2)} \right) - \sum_{j \in \mathcal{V}\setminus L} (\deg(j) - 1) H(\mu_j) \\
\text{subject to} & \quad M^{(j_1,j_2)} 1 = \mu_{j_1}, \quad (M^{(j_1,j_2)})^T 1 = \mu_{j_2}, \quad \text{for } (j_1,j_2) \in \mathcal{E}_r.
\end{align*}
\]

**Corollary 4.3.** Let \( \mathcal{T}_r = (\mathcal{V}, \mathcal{E}_r) \) be a directed tree with root \( r \in L \), and \( \mathcal{T}_T = (\mathcal{V}, \mathcal{E}_T) \) be a directed tree with root \( \hat{r} \in L \) with the same structure as \( \mathcal{T}_r \) and the edges on the direct path between \( r \) and \( \hat{r} \) reversed. Then, the solution to (4.1) on \( \mathcal{T}_r \) and on \( \mathcal{T}_T \) are equivalent in the sense that the marginal distributions \( \mu_j \), for \( j \in \mathcal{V} \), and transport plans \( M^{(j_1,j_2)} \), for \( (j_1,j_2) \in \mathcal{E}_r \), the same, and on the reversed edges it holds \( M^{(j_2,j_1)} = (M^{(j_1,j_2)})^T \).

### 4.2. Equivalence of problems (2.5) and (4.1) on a tree.

We will now verify that the generalized Schrödinger bridge (4.1) on a rooted tree \( \mathcal{T}_r \) is equivalent to an entropy regularized multi-marginal optimal transport problem (2.5) on an undirected tree \( \tilde{\mathcal{T}} \) with the same structure. In particular, given positive transition probability matrices \( A^{(j_1,j_2)} \) for \( (j_1,j_2) \in \mathcal{E}_r \) and a regularization parameter \( \epsilon \), there is a natural choice of cost matrices \( C^{(j_1,j_2)} \) for \( (j_1,j_2) \in \mathcal{E}_r \), so that the minimizers to both problems represent the same solution. For the original bi-marginal optimal transport problem (2.1) the equivalence between entropy regularized OMT and the Schrödinger bridge problem has been studied extensively [12, 13, 39, 40].

**Remark 1 (cf. [34, Rem. 1]).** In the discrete setting, it is easy to see that with a cost matrix defined as \( C = -\epsilon \log(A) \), the entropy regularized optimal transport problem (2.6) is equivalent to the bi-marginal Schrödinger bridge (2.13), by noting that the objective can then be written as

\[
\text{trace}(C^T M) + \epsilon H(M) = \sum_{i,j=1}^n \epsilon \left( M_{ij} \log \left( \frac{M_{ij}}{A_{ij}} \right) - M_{ij} + 1 \right) = \epsilon H(M \mid A).
\]

Both bi-marginal problems thus yield the same optimal solution \( M \).

We will show that this equivalence holds true even for the multi-marginal case defined on a tree.

**Theorem 4.4.** Let \( \mathcal{T}_r = (\mathcal{V}, \mathcal{E}_r) \) be a rooted directed tree with root \( r = 1 \in L \), and let \( \mathcal{T} = (\mathcal{V}, \mathcal{E}) \) be its undirected counterpart. Let \( \epsilon > 0 \) and \( C^{(j_1,j_2)} \) be such that

\[
K^{(j_1,j_2)} = A^{(j_1,j_2)}, \quad \text{for } (j_1,j_2) \in \mathcal{E}_r.
\]
Let $M = K \odot U$ be the solution to the entropy regularized multi-marginal optimal transport problem (2.5) on $T_r$, where $K$ and $U$ are defined as in Theorem 3.2. Moreover, let $\mu_j$ for $j \in V$, and $M^{(j_1,j_2)}$ for $(j_1,j_2) \in E_r$ be the solution to the generalized Schrödinger bridge (4.1) on $T_r$. Then,

$$P_j(M) = \mu_j, \quad \text{for } j \in V,$$

$$P_{(j_1,j_2)}(M) = M^{(j_1,j_2)}, \quad \text{for } (j_1,j_2) \in E_r.$$

Furthermore, if $v_j$ for $j \in L$ are the corresponding vectors in Theorem 4.1, then it holds that

$$u_j = \begin{cases} 1./v_1, & \text{if } j = 1 \\ v_j, & \text{otherwise.} \end{cases}$$

In particular, given a maximum likelihood problem (4.1) on a tree $T_r = (V,E_r)$, a corresponding multi-marginal optimal transport problem can be formulated by defining the cost matrices as $C^{(j_1,j_2)} = -\epsilon \log(A^{(j_1,j_2)})$, for all $(j_1,j_2) \in E_r$. Vice versa, given an entropy regularized multi-marginal optimal transport problem (2.5) on a tree $T = (V,E)$, a corresponding maximum likelihood problem can be formulated by transforming the undirected tree into a rooted tree $T_r = (V,E_r)$, where $r \in L$, and defining the matrices $A^{(j_1,j_2)} = \exp \left(-C^{(j_1,j_2)}/\epsilon \right)$, for all $(j_1,j_2) \in E_r$. However, if these matrices are not row-stochastic, they cannot be interpreted as transition probability matrices of a Markov process.

**Remark 2.** For a nonnegative matrix $A$ with at least one positive element in each row, define the vector $b = 1./A1$. Then the matrix $\bar{A} = \text{diag}(b)A$ is row stochastic. Note that we can write $H(M | \text{diag}(\mu)A) = H(M | \bar{A}) + H(\mu | b)$. For a given rooted tree $T_r = (V,E_r)$ assume that there is a set of matrices $A^{(j_1,j_2)}$, for $(j_1,j_2) \in E_r$, which are not necessarily row-stochastic, but are such that for every $j_2 \in E_r$, there is a vector $b_{j_2}$ such that it holds that $b_{j_1} = 1./A^{(j_1,j_2)}1$ for all $j_2$ such that $(j_1,j_2) \in E_r$. Then, according to Proposition 4.2, the objective function of the maximum likelihood estimation problem of a Markov process on $T_r$ in (4.1) may be written as

$$\begin{align*}
\text{minimize} & \quad \sum_{M^{(j_1,j_2)} \in E_r} \sum_{(j_1,j_2) \in E_r} H\left(M^{(j_1,j_2)} | \bar{A}^{(j_1,j_2)}\right) - \sum_{j \in V \setminus L} (\text{deg}(j) - 1)H(\mu_j | b_j).
\end{align*}$$

The Sinkhorn scheme (2.11) can equivalently be formulated in terms of the notation from the maximum likelihood estimation problem in Theorem 4.1. Given an initial set of positive vectors $v_j$ for $j \in L$, the updates are then expressed as

$$\begin{align*}
v_1 &= \varphi_1./\mu_1 \\
v_j &= \mu_j./\tilde{\varphi}_j, \quad \text{for } j \in L \setminus 1,
\end{align*}$$

where $\varphi_1$ is computed as in (4.2) before each update of $v_1$, and $\tilde{\varphi}_j$ is computed as in (4.3) before updating $v_j$ for each $j \in L \setminus 1$. This can be done efficiently, and by storing and reusing intermediate results appropriately, this algorithm is equivalent to Algorithm 3.1. It was shown that the classical Sinkhorn iterations are a block coordinate ascent in a Lagrange dual problem [37,51]. In this sense, the scheme (4.6) can be understood as Sinkhorn iterations even for the problem (4.1).

**Proposition 4.5.** Let $v_j$, for $j \in L$, be an initial set of positive vectors. Iteratively performing the updates (4.6) corresponds to a block coordinate ascent in a Lagrange dual problem to (4.1).
In the light of Proposition 4.5 it is not surprising, that the algorithm presented in [34], for the special case where the tree represents a hidden Markov chain, is of the form (4.6), where the vectors \(\varphi_j\) and \(\hat{\varphi}_j\) are constructed according to the special tree structure.

5. Multi-marginal vs. pairwise regularization. Another natural way to define an optimal transport problem on the tree \(\mathcal{T}\) is to minimize the sum of all bi-marginal transport costs on the edges of \(\mathcal{T}\). This problem has been considered for general networks in [49]. In particular, it is common to formulate optimal transport barycenter problems in this pairwise manner [20, 27]. Although, as argued in Remark 1, in the bi-marginal case the entropy regularized optimal transport problem is equivalent to the Schrödinger bridge, we show that this equivalence does not extend to the respective problems defined on trees.

Define the cost matrix 
\[
C(j_1, j_2) \in \mathbb{R}^{n \times n},
\]
which characterizes the cost of transportation between marginal \(j_1\) and \(j_2\), for all \((j_1, j_2) \in \mathcal{E}\). The discretized and entropy regularized pairwise optimal transport problem on \(\mathcal{T}\) is then
\[
\text{minimize}_{\mu_{j_1} \in \mathcal{V} \setminus \Gamma} \sum_{(j_1, j_2) \in \mathcal{E}} T_{\epsilon}^{(j_1, j_2)}(\mu_{j_1}, \mu_{j_2}),
\]
where \(T_{\epsilon}(\cdot, \cdot)^{(j_1, j_2)}\) is defined as the regularized bi-marginal optimal transport problem (2.6) with cost matrix \(C^{(j_1, j_2)}\).

Proposition 4.2 links the multi-marginal optimal transport problem and the pairwise optimal transport problem on a tree. Consider a rooted version of the tree, denoted \(\mathcal{T}_r = (\mathcal{V}, \mathcal{E}_r)\), where \(r \in \mathcal{L}\). In the case that \(A^{(j_1, j_2)} = K^{(j_1, j_2)}\), for all \((j_1, j_2) \in \mathcal{E}_r\), the multi-marginal optimal transport problem has the same solution as problem (4.4). Now note that this problem can be written as
\[
\text{minimize}_{\mu_{j_1} \in \mathcal{V} \setminus \Gamma} \sum_{(j_1, j_2) \in \mathcal{E}} T_{\epsilon}^{(j_1, j_2)}(\mu_{j_1}, \mu_{j_2}) - \sum_{j \in \mathcal{V} \setminus \mathcal{L}} (\deg(j) - 1)H(\mu_j).
\]
Thus, the multi-marginal optimal transport problem penalizes not only the transport cost between the marginals, but in addition favors marginal distributions with high entropy. One can thus expect less smoothed out distributions when solving the multi-marginal optimal transport problem, which is desirable in many applications, such as localization problems [28] and computer vision applications [48]. Note that this qualitative difference between the pairwise and multi-marginal formulation has been previously observed, yet not fully explained, for a barycenter optimal transport problem, i.e., a star graph [28]. Moreover, empirical study suggests that the multi-marginal problem is better conditioned compared to the pairwise problem, which allows for smaller values of the regularization parameter \(\epsilon\), while still yielding a numerically stable algorithm. We give empirical evidence for this behaviour in Section 5.2.

The solution of the pairwise optimal transport problem (5.1) is compared to the solution of the two equivalent problems (2.5) and (4.1) on the example of a path graph in Section 5.1, and on a more complex tree in Section 5.2.

5.1. Discrete time and discrete space Schrödinger bridge. Consider a path tree \(\mathcal{T}_r = (\mathcal{V}, \mathcal{E}_r)\), where \(\mathcal{V} = \{1, 2, \ldots, J\}\) and \(\mathcal{E}_r = \{(j, j+1) | j = 1, \ldots, J-1\}\), as sketched in Figure 2. Let \(A^j\) denote the probability transition matrix on \((j, j+1) \in \mathcal{E}_r\). This model corresponds to a Markov chain of length \(J\). Assume that the distributions on the leaves \(j = 1\) and \(j = J\) are known. The most likely particle evolutions between them are then found by solving (2.13). Following [34], we see that this problem is
\begin{align*}
    \text{Figure 2. Illustration of the linear path tree in Section 5.1.}
\end{align*}

\begin{align*}
\text{Figure 3. Solutions to the multi-marginal and pairwise entropy regularized optimal transport problem on a path graph for varying regularization parameter } \epsilon.
\end{align*}

\begin{align*}
   &\text{equivalent to the discrete time and discrete space Schrödinger bridge in [45]. Assume that the distributions } \mu_j, \text{ for } j = 1, \ldots, J, \text{ are strictly positive, and define the row stochastic matrices } \bar{A}^j = \text{diag}(\mu_j)^{-1}M^j, \text{ for } j = 1, \ldots, J. \text{ In terms of these matrices problem (2.13) reads}
   
   &\text{(5.2)}
   
   &\text{minimize } \sum_{j=1}^{J-1} \sum_{i=1}^{n} (\mu_j)^i H \left( \bar{A}^i, A^i \right)
   
   &\text{subject to } \bar{A}^j \mathbf{1} = \mathbf{1}, \mu_{j+1} = (\bar{A}^j)^T \mu_j, \text{ for } j = 1, \ldots, J - 1.
\end{align*}

Here \( A_i \) denotes the \( i \)-th row of \( A \). Problem (5.2) is exactly the formulation of a Schrödinger bridge over a Markov chain from [45, eq. (24)]. In [45] it is shown that a unique solution to a corresponding Schrödinger system exists if \( \mu_J \) is a strictly positive distribution and the matrix \( \prod_{j=1}^{J-1} A^j \) has only positive elements. The solution to the Schrödinger system may be obtained as a fixed point iteration [31], which is linked to the Sinkhorn iterations for entropy regularized optimal transport problems. We recall from [34] that the optimization problem (5.2) is non-convex, whereas the equivalent formulation (2.13) is convex.

We note that with cost matrices defined as \( C^j = -\epsilon \log(A^j) \), for \( j = 1, \ldots, J - 1 \), the time and space discrete Schrödinger bridge problem may be written as

\begin{align*}
   &\text{minimize } \sum_{j=1}^{J-1} T^{(j,j+1)}(\mu_j, \mu_{j+1}) - \sum_{j=2}^{J-1} H(\mu_j).
\end{align*}

In comparison to the pairwise optimal transport problem on the path graph \( T \), the Schrödinger bridge thus favors intermediate distributions with lower entropy, resulting in less smoothed out solutions. We illustrate this behaviour for a path tree \( T_r = (V, E_r) \), with vertices \( V = \{1, 2, \ldots, 6\} \), and where the initial and final distribution are given as \( \mu_1(x) = \exp\left(-\left(x - \frac{1}{2}\right)^2\right) \), and \( \mu_J(x) = \exp\left(-\left(x - \frac{0.8}{2}\right)^2\right) \), and the cost function is defined as the Euclidean distance. The results to problem (5.2) and problem (5.1) are compared for different values of the regularization parameter \( \epsilon \) in Figure 3. Both optimal transport solutions describe a smooth way of shifting the mass from distribution \( \mu_1 \) to \( \mu_J \). The larger the regularization parameter \( \epsilon \) is chosen,
5.2. Optimal transport on trees. In this section we compare the entropy regularized multi-marginal and pairwise optimal transport solutions on a more general tree. Consider the tree \( T = (V, E) \) illustrated in Figure 4(a) with 15 nodes, each representing a 50 \( \times \) 50 pixel image. The marginal images on the 8 leaves, coloured in gray, are known. Each edge on the tree \( T \) is associated with a cost function defined by the Euclidean distance between any two pixels. Using this choice of cost function in the corresponding optimal transport problems yields smooth translations in power for the intermediate marginals.

We solve the entropy regularized pairwise optimal transport problem (5.1) with regularization parameter \( \epsilon = 2 \cdot 10^{-3} \) on \( T \). The solution can be seen in Figure 4(b). Compared to the pairwise optimal transport estimate, the solution to the entropy regularized multi-marginal optimal transport problem (2.5) on the same tree \( T \) and with the same regularization parameter \( \epsilon \) is significantly sharper and less smoothed out, see Figure 4(c). For the pairwise optimal transport problem, the method diverges with a smaller regularization parameter, e.g., \( \epsilon = 10^{-3} \). In contrast, for the multi-marginal formulation the regularization parameter can be decreased further, still yielding a numerically stable algorithm. We have found that the method is still stable for a regularization parameter of \( \epsilon = 4 \cdot 10^{-4} \), which results in very clear estimates on the intermediate nodes.

6. Estimating ensemble flows on a hidden Markov chain. We consider the problem of tracking an ensemble of agents on a network based on aggregate measurements from sensors distributed around the network. This is similar to [34], where ensemble flows of indistinguishable agents have been estimated as the maximum likelihood solution on a hidden Markov chain. The present work generalizes this method and provides an algorithm for solving the problem introduced therein. In particular, the framework in [34] is a special case of the method in Theorem 4.1, and can therefore be solved with Algorithm 3.1. Herein, we study different observation models and robustness of the estimates with respect to the number of agents.

Consider a graph \( \mathcal{G} = (\mathcal{V}_G, \mathcal{E}_G) \) with vertices \( \mathcal{V}_G \) and edges \( \mathcal{E}_G \). Let the set of nodes \( \mathcal{V}_G \) be the set of states of a Markov model with transition probability matrix \( A^t \in \mathbb{R}^{n \times n} \), where \( n = |\mathcal{V}_G| \), for \( t = 1, \ldots, T-1 \). We simulate a finite number of agents to evolve according to this Markov model for a number of time steps \( t = 1, \ldots, T \).
Let $\mu_t \in \mathbb{R}^n$ denote the distribution of agents over the set of states for the times $t = 1, \ldots, T$. An observation model is represented by the detection probability matrices $B^s \in \mathbb{R}^{n \times m}$ where $m$ is the size of the observation space, and $s = 1, \ldots, S$ where $S$ denotes the number of uncoupled observations at each time point. Let $\Phi_{t,s} \in \mathbb{R}^m$ denote the aggregate measurements at time $t$ for observation $s$. We then estimate the ensemble evolution as the solution of

$$\begin{align*}
\text{minimize} & \quad M_{1:T-1}, D_{1:T,1:S} \mid \mu_{1:T}
\text{subject to} & \quad M^t \1 = \mu_{t-1}, \quad (M^t)^T \1 = \mu_t, \quad D^t \1 = \mu_t, \quad (D^t)^T \1 = \Phi_{t,s}, \\
& \quad \text{for } t = 1, \ldots, T, \text{ and } s = 1, \ldots, S.
\end{align*}$$

(6.1)

The tree corresponding to the Markov model (6.1) is sketched in Figure 5(a) This optimization problem is similar to the one in [34], but therein the initial distribution of agents, i.e., the marginal $\mu_1$, is assumed to be known.

In the following we consider a number of $N$ agents evolving on the network displayed in Figure 5(b) with $n = 100$ nodes and $180$ edges. Figure 5(b) also shows the location of the $N_S = 15$ sensors. In a first step, we compute the discrete Schrödinger bridge (2.13) between the distribution $\mu_1$, where all agents are in node 1, and the distribution $\mu_T$, where all agents are in node 100, with the random walk on $G$ as a prior. The resulting particle evolutions $M^t$, for $t = 1, \ldots, T - 1$, in (2.13) define the transition probability matrices $\bar{A}^t = \text{diag}(\mu_t)^{-1}M^t$, for $t = 1, \ldots, T - 1$, as in (5.2), used to simulate a number of $N$ agents with initial distribution $\mu_1$. By construction, the final distribution is then $\mu_T$.

We consider two observation models, one where the sensors make uncoupled measurements, similar to the example in [34], and one where the sensors are coupled and form a joint measurement. In the uncoupled setting, we consider an observation space of 2 states for each sensor, where one state denotes that an agent is detected, and the other one that the agent is undetected by the sensor. Hence, we define an observation probability matrix $B^s \in \mathbb{R}^{n \times 2}$ for each sensor $s = 1, \ldots, S = N_S$, where the probability for an agent on node $i$ to be detected by the sensor $s$ is defined as $B^s_{i1} = \min(0.99, 2e^{-d_{s,i}})$, where $d_{s,i}$ denotes the Euclidean distance between the location of sensor $s$ and the node $i$. Consequently the probability of not being detected is $B^s_{i2} = 1 - B^s_{i1}$. For the coupled observation model, the observation space consists of all possible sets of sensors that detect a given agent, i.e., all subsets of the set $\mathcal{S}$.
$N = 10$ agents. $N = 100$ agents. $N = 1000$ agents.

Figure 6. True ensemble flow and estimates with the two observation models for a varying number of agents.

$\{1, \ldots, N_S\}$. The size of the observation space is thus $2^{N_S}$, and there is only one observation at each time instance, i.e., $S = 1$. Hence, we define an observation probability matrix $B_{\text{joint}} \in \mathbb{R}^{n \times 2^2}$, where the probability for an agent in node $i$ to be detected by exactly the set $\mathcal{S}$ of sensors is given by $\prod_{s \in \mathcal{S}} B_{s,i}^1 \prod_{s \notin \mathcal{S}} B_{s,i}^2$.

We solve the multi-marginal optimal transport problem corresponding to (6.1) with assumed probability transition matrix $A$ describing a random walk on $\mathcal{G}$, and observation probability matrices $B^s$ for the two described observation models. The results for both observation models are compared for $N \in \{10, 100, 1000\}$ agents in Figure 6, where the estimated number of agents in each node is plotted as a circle with size corresponding to the log-scaled weighted number of agents in that node.

In Figure 6(a), one can see that for $N = 10$ agents the coupled estimate localizes the position of the agents slightly better than the uncoupled estimate. However, this effect decreases with an increasing number of agents, as the evolution of a single agent has less influence on the empirical distribution of agents. Already from a number of $N = 100$ agents the uncoupled estimate is competitive with the coupled estimate, albeit relying on significantly less information (cf. Figure 6(b)). For an ensemble of $N = 1000$ agents one can hardly see any difference between the two estimates, as shown in Figure 6(c).

Note that the number of observations at each time instance for the uncoupled estimate is only of the size $2^{N_S} = 30$, whereas the number of coupled observations is $2^{N_S} = 32768$. The improved observation model thus comes at an increased computational cost, which is exponential in the number of sensors, and becomes infeasible for larger numbers of sensors.

7. Conclusion. In this work we consider multi-marginal optimal transport problems with cost functions that decouple according to a tree, and we show that the entropy regularized formulation of this problem is equivalent to a generalization of a time and space discrete Schrödinger bridge defined on the same tree. Moreover, we derive an efficient algorithm for solving this problem. We also compare the multi-marginally regularized optimal transport problem to a commonly used pairwise regularized optimal transport problem and illustrate the benefits in theory and practice. Finally, we describe how to apply the framework to the problem of tracking an ensemble of indistinguishable agents.
Interestingly, the construction of the marginals in Theorem 3.2 is of the same form as the belief propagation algorithm [50, 55] for inference in graphical models. In this interpretation, the vectors \( u_j \) correspond to the local evidence in node \( j \in V \) and \( \alpha_{(j_1, j_2)} \) is the message passed from node \( j_2 \) to node \( j_1 \). Moreover, the objective function in (4.5) can be interpreted as the Bethe free energy [56], which is connected to belief propagation [55]. These similarities of our framework to belief propagation algorithms, could be a stepping stone to extending our framework to graphs with cycles, and will be analyzed in a future study. Another direction of interest is the extension to continuous state models.

**Appendix A. Proofs.**

For a index set \( \Gamma \), we denote the double sum \( \sum_{j \in \Gamma} \sum_{i_j} \) by \( \sum_{i_j,j \in \Gamma} \). We use the notation > and < for rooted trees as introduced in the beginning of Section 4.1. For simplicity of notation, we denote the set of matrices \( M^{(j_1, j_2)} \), for \((j_1, j_2) \in \mathcal{E} \), by \( \mathcal{M} \), and similarly write \( \mu \) and \( \lambda \) for the respective sets of optimization variables. The proof of Theorem 3.2 and Proposition 3.3 are based on the following lemma.

**Lemma A.1** ([28, Lemma 1 and 2]). Let \( \mathbf{U} = u_1 \otimes u_2 \otimes \cdots \otimes u_j \), for a set of vectors \( u_1, u_2, \ldots, u_j \), and \( \mathbf{K} \) be a tensor of the same size. If \( \langle \mathbf{K}, \mathbf{U} \rangle = \mathbf{w}^T \mathbf{u}_j \), for a vector \( \mathbf{w} \) that does not depend on \( u_j \), then it holds that

\[
P_j(\mathbf{K} \odot \mathbf{U}) = \mathbf{w} \otimes u_j.
\]

Similarly, if \( \langle \mathbf{K}, \mathbf{U} \rangle = \mathbf{w}^T \text{diag}(u_{j_1}) \mathbf{W} \text{diag}(u_{j_2}) \mathbf{w} \), for two vectors \( \mathbf{w}, \mathbf{w} \), and a matrix \( \mathbf{W} \) that does not depend on \( u_{j_1} \) and \( u_{j_2} \), then it holds that

\[
P_{j_1, j_2}(\mathbf{K} \odot \mathbf{U}) = \text{diag}(\mathbf{w} \otimes u_{j_1}) \mathbf{W} \text{diag}(u_{j_2} \otimes \mathbf{w}).
\]

**Proof of Theorem 3.2.** Due to the decoupling of the cost tensor \( \mathbf{C} \) as in (3.2), the tensor \( \mathbf{K} = \exp(-\mathbf{C}/\epsilon) \) decouples according to (3.3) with the matrices \( K^{(j_1, j_2)} = \exp(C^{(j_1, j_2)}/\epsilon) \), for \((j_1, j_2) \in \mathcal{E} \). We can therefore write

\[
\langle \mathbf{K}, \mathbf{U} \rangle = \sum_{i_j,j \in \mathcal{V}} \left( \prod_{(j_1, j_2) \in \mathcal{E}} K^{(j_1, j_2)}_{\ell_1, \ell_2} \right) \left( \prod_{k \in \mathcal{V}} (u_k)_{i_k} \right) = \sum_{i_j} (u_j)_{i_j} (w_j)_{i_j},
\]

where the vector \( w_j \) is of the form

\[
(w_j)_{i_j} = \sum_{i_k, k \in \mathcal{V} \setminus j} \left( \prod_{(j_1, j_2) \in \mathcal{E}} K^{(j_1, j_2)}_{\ell_1, \ell_2} \right) \left( \prod_{j \in \mathcal{V} \setminus j \neq \ell} (u_j)_{i_j} \right).
\]

Consider the underlying tree to be rooted in node \( j \). Then this can be written as \( (w_j)_{i_j} = \prod_{k \in \mathcal{L}} (\alpha_{(j, k)})_{i_j} \), where \( \alpha_{(p(k), k)} \in \mathbb{R}^n \) is defined by

\[
(\alpha_{(p(k), k)})_{i_{p(k)}} = \sum_{i_m, \ell \geq k, m > k} \prod_{i_{m}, \ell = k} K^{(p(m), i_m)}_{\ell_{p(m)}, i_{m}} (u_m)_{i_m}.
\]

If \( k \in \mathcal{L} \), then

\[
(\alpha_{(p(k), k)})_{i_{p(k)}} = \sum_{i_k} K^{(p(k), k)}_{\ell_{p(k)}, i_k} (u_k)_{i_k} = \left( K^{(p(k), k)} u_k \right)_{i_{p(k)}}.
\]
Otherwise, it holds
\[
\begin{align*}
(\alpha_{(p,k)})_{i_p(k)} &= \sum_{i_k} \left( K_{(p(k),k)}^{(p,k)}(u_k)_{i_k} \prod_{\ell \in N_k \setminus \ell \notin p(k)} \left( \sum_{i_m : m \ge \ell \ge j > \ell} K_{(p(j),j)}^{(p(k),k)}(u_j)_{i_j} \right) \right) \\
&= \sum_{i_k} \left( K_{(p(k),k)}^{(p(k),k)}(u_k)_{i_k} \prod_{\ell \in N_k \setminus \ell \notin p(k)} \left( \alpha_{(p(\ell),\ell)} \right)_{i_p(\ell)} \right) .
\end{align*}
\]

This inductively defines the vectors \( \alpha_{(j,k)} \) as in (3.4). The expression for the projection follows from Lemma A.1 with \( w_j = \bigcup_{k \in N_j} \alpha_{(j,k)} \).

**Proof of Proposition 3.3.** Let \( r \in L \) and assume that \( j_1 \) lies on the path from \( r \) to \( j_L \) (note that \( r = j_1 \) if \( j_1 \in L \)). For the pairwise marginals we would like to write the inner product as in Lemma A.1:
\[
\langle K, U \rangle = \sum_{i_j : j \in V} \left( \prod_{(k_1, k_2) \in E_r} K_{k_1, i_{k_2}}^{(k_1, k_2)} \right) \left( \prod_{j \in V} (u_j)_{i_j} \right) = \sum_{i_{j_1} : j_L} (w_{j_1})_{i_{j_1}} (u_{j_1})_{i_{j_1}} \prod_{i \in V} (\hat{w}_{j_L})_{i_{j_L}} .
\]

Note that the nodes \( \mathcal{V} \setminus \{j_1, j_L\} \) can be partitioned into three sets, which are separated by the nodes \( j_1 \) and \( j_L \): \( \{j \in \mathcal{V} : j \neq j_1, j \neq j_L\} \); \( \{j \in \mathcal{V} : j > j_1, j \neq j_L\} \); and \( \{j \in \mathcal{V} : j > j_L\} \). Then \( w_{j_1} \) corresponds to the contribution from the first set
\[
(w_{j_1})_{i_{j_1}} = \sum_{i_k : k \neq j_1, k \neq j_L} \left( \prod_{(k_1, k_2) \in E_r} K_{k_1, i_{k_2}}^{(k_1, k_2)} \right) \left( \prod_{k \in \mathcal{V}} \left( u_k \right)_{i_k} \right) = \prod_{k \in \mathcal{N}_{j_1} \setminus j_2} \left( \alpha_{(j_1, k)} \right)_{i_{j_1}},
\]
and similarly \( \hat{w}_{j_L} \) is the contribution from the third set
\[
(\hat{w}_{j_L})_{i_{j_L}} = \sum_{i_k : k > j_L} \left( \prod_{(k_1, k_2) \in E_r} K_{k_1, i_{k_2}}^{(k_1, k_2)} \right) \left( \prod_{k \in \mathcal{V}} \left( u_k \right)_{i_k} \right) = \prod_{k \in \mathcal{N}_{j_L} \setminus j_{L-1}} \left( \alpha_{(j_L, k)} \right)_{i_{j_L}} .
\]
The matrix \( W \) is then obtained by summing over all the indices corresponding to second index set \( \{i_k : k > j_1, k \neq j_L\} \)
\[
W_{i_{j_1}, i_{j_L}} = \sum_{i_k : k > j_1, k \neq j_L} \left( \prod_{(k_1, k_2) \in E_r} K_{k_1, i_{k_2}}^{(k_1, k_2)} \right) \left( \prod_{k \in \mathcal{V}} \left( u_k \right)_{i_k} \right) = \sum_{i_{j_2}, \ldots, i_{j_{L-1}}} \left( \prod_{\ell = 2}^{L-1} K_{i_{j_{\ell-1}}, i_{j_\ell}}^{(j_{\ell-1}, j_\ell)} \prod_{k \in \mathcal{N}_{j_\ell} \setminus j_{\ell+1}} \left( \alpha_{(j_\ell, k)} \right)_{i_{j_\ell}} K_{i_{j_{L-1}}, i_{j_L}}^{(j_{L-1}, j_L)} \right) .
\]

Applying Lemma A.1, we get the projection on the marginals \( j_1 \) and \( j_L \). \( \square \)
Proof of Corollary 3.4. The corollary follows from [51, Prop. 1 and 3] since (2.3) fulfills [51, Ass. A, B and C] (cf. [51, Sec. 6.3 and Prop. 11]) and the Sinkhorn iterations are block coordinate ascent in the dual (2.10), see [28,37] and Theorem 3.2.

Proof of Theorem 4.1. Assume that the only edge with node $j = 1$ is denoted $(1,2)$. We add the trivial constraints $M^{(1,j_2)}1 = (M^{(p,j_1)})^T 1$ for $(j_1,j_2) \in \mathcal{E}_r \setminus (1,2)$. We relax this constraint, together with the constraint on $\mu_1$, and let $\lambda_{(j_1,j_2)}$, for $(j_1,j_2) \in \mathcal{E}_r$, denote the corresponding dual variables. Furthermore, we relax the constraints $(M^{(j,p-j)})^T 1 = \mu_j$ with dual variables $\lambda_j$, for the leaves $j \in \mathcal{L} \setminus 1$. A Lagrangian for (4.1) is then
\[
L(M, \mu, \lambda) = \sum_{(j_1,j_2) \in \mathcal{E}} H \left( M^{(j_1,j_2)} \text{diag}(\mu_{j_1}) A^{(j_1,j_2)} \right) + \lambda_{(1,2)}^T (M^{(1,2)} 1 - \mu_1) \\
+ \sum_{j_2 \in \mathcal{L}} \lambda_{j_2}^T (\mu_j - (M^{(j_1,j_2)})^T 1) + \sum_{(j_1,j_2) \in \mathcal{E}, j_1 \neq 1} \lambda_{(j_1,j_2)}^T (M^{(j_1,j_2)} 1 - (M^{(j_1,j_2)})^T 1).
\]

When $j_2$ is an inner node on the tree, i.e., $j_2 \notin \mathcal{L}$, the derivative with respect to the entries $M^{(j_1,j_2)}_{i_1,i_2}$, for all $i_1,i_2 = 1, \ldots, n$, and $(j_1,j_2) \in \mathcal{E}_r$ is
\[
(A.1) \quad \log \left( \frac{M^{(j_1,j_2)}_{i_1,i_2}}{(\mu_{j_1})_{i_1} A_{i_1,i_2}} \right) + (\lambda_{(j_1,j_2)})_{i_2} - \sum_{k: (j_2,k) \in \mathcal{E}} (\lambda_{(j_2,k)})_{i_2}.
\]

Since (4.1) is convex, a mass transport plan is optimal if this gradient vanishes, yielding the expression in terms of the other variables,
\[
(A.2) \quad M^{(j_1,j_2)} = \text{diag}(\mu_{j_1} / v_{(j_1,j_2)}) A^{(j_1,j_2)} \text{diag} \left( \bigodot_{k: (j_2,k) \in \mathcal{E}} v_{(j_2,k)} \right),
\]

where $v_{(j_1,j_2)} = \exp(\lambda_{(j_1,j_2)})$ for all $(j_1,j_2) \in \mathcal{E}_r$. In case $j_2 \in \mathcal{L}$, the last sum in the derivative (A.1) is replaced by $(\lambda_{j_2})$. Defining $v_{j_2} = \exp(\lambda_{j_2})$, the optimal mass transport plan is thus of the form
\[
(A.3) \quad M^{(j_1,j_2)} = \text{diag}(\mu_{j_1} / v_{(j_1,j_2)}) A^{(j_1,j_2)} \text{diag}(v_{j_2}).
\]

Next, note that the marginal of the optimal transport plan satisfies
\[
(A.4) \quad \mu_{j_2} = M_{(j_1,j_2)}^T 1 = \hat{\varphi}_{j_2} \odot \varphi_{j_2},
\]

with
\[
(A.5) \quad \varphi_{j_2} = \bigodot_{k: (j_2,k) \in \mathcal{E}} v_{(j_2,k)}, \quad \text{and} \quad \hat{\varphi}_{j_2} = (A^{(j_1,j_2)})^T (\mu_{j_1} / v_{(j_1,j_2)}).
\]

Since for all $k$ such that $(j_2,k) \in \mathcal{E}_r$ it holds that $M_{(j_2,k)} 1 = \mu_{j_2}$, we get
\[
v_{(j_2,k)} = A^{(j_2,k)} \left( \bigodot_{k: (k,\ell) \in \mathcal{E}} v_{(k,\ell)} \right) = A^{(j_2,k)} \varphi_k,
\]

which completes the definition of the vectors $\varphi_j$, for $j \in \mathcal{V}$. Similarly to (A.4) it holds that
\[
(A.6) \quad \mu_{j_1} = \left( \bigodot_{k: (j_1,k) \in \mathcal{E}} v_{(j_1,k)} \right) \odot A \left( \mu_{j_1} / v_{(j_1,j_1)} \right).
\]
Plugging (A.6) into \( \tilde{\varphi}_{j_2} \) in (A.5), yields the recursive definition of the vectors \( \tilde{\varphi}_j \), for \( j \in V \). The expression for the mass transport plans \( M^{(j_1,j_2)} \), for \( (j_1,j_2) \in E \), in terms of the vectors \( \tilde{\varphi}_{j_1} \), \( \varphi_{j_2} \) and \( \varphi_{j_1 \setminus j_2} \) follows by identifying them in expressions (A.2) and (A.3).

**Proof of Proposition 4.2.** For any matrices \( M, A \in \mathbb{R}^{n \times n} \) and \( \mu \in \mathbb{R}^n \) one can write the term \( H(M \mid \text{diag}(\mu)A) \) as

\[
(A.7) \quad \sum_{i,j=1}^n (M_{ij} \log \left( \frac{M_{ij}}{A_{ij}} \right) - M_{ij}) + \sum_{i,j=1}^n (\mu_i A_{ij} - M_{ij} \log(\mu_i))
\]

Since it holds that \( M1 = \mu \) and \( A1 = 1 \), the second sum can be simplified to \( \sum_{i,j=1}^n (\mu_i - \mu_i \log(\mu_i)) \). Furthermore, adding the term \( \sum_{i,j=1}^n A_{ij} - \sum_{i=1}^n 1 = 0 \) to (A.7), the expression can be written as \( H(M \mid A) - H(\mu) \). Due to the underlying tree structure of problem (4.1), the number of outgoing edges from the root node \( j_r \) is \( \text{deg}(j_r) \), and for all other vertices \( j \in V \setminus \{j_r\} \) the number of outgoing edges is \( \text{deg}(j) - 1 \). In the case that \( j_r \in L \), since the marginal \( \mu_{j_r} \) is known, the term \( H(\mu_{j_r}) \) is constant and can be removed from the objective without changing the optimal solution.

**Proof of Corollary 4.3.** Let the set of matrices \( M^{(j_1,j_2)} \), for \( (j_1,j_2) \in E_r \), and vectors \( \mu_j \), for \( j \in V \setminus L \), be the solution to (4.1) on the rooted directed tree \( T_r = (V,E_r) \). It is thus also the solution to (4.4) on the same tree. Changing the root to another vertex \( \hat{r} \in L \), requires switching the direction of all edges on the direct path between \( r \) and \( \hat{r} \). This is done by replacing the respective transition probability matrix \( A^{(j_1,j_2)} \) by the reverse transition probability matrix \( A^{(j_2,j_1)} \), which is given by \( A^{(j_2,j_1)} = \text{diag}(1/a_{j_2})(A^{(j_1,j_2)})^T \text{diag}(a_{j_1}) \), for two given vectors \( a_{j_1} \) and \( a_{j_2} \) [45]. Note that

\[
H \left( M^{(j_2,j_1)} \mid A^{(j_2,j_1)} \right) = H \left( M^{(j_2,j_1)} \mid \text{diag}(1/a_{j_2})(A^{(j_1,j_2)})^T \text{diag}(a_{j_1}) \right)
\]

\[
= H \left( (M^{(j_1,j_2)})^T \mid A^{(j_1,j_2)} \right) + H(\mu_{j_1} \mid \text{diag}(1/a_{j_1})) + H(\mu_{j_2} \mid \text{diag}(a_{j_2})).
\]

Summing over these terms for all edges on the path between \( r \) and \( \hat{r} \), the last two terms cancel for all inner nodes \( j_1,j_2 \in V \setminus \Gamma \), and are constants for \( j_1,j_2 \in \{r,\hat{r}\} \). From problem (4.4) we thus see that the optimal distributions \( \mu_j \), for \( j \in V \), are unchanged, and for the transport plans on the reversed edges it holds \( M^{(j_2,j_1)} = (M^{(j_1,j_2)})^T \).

**Proof of Theorem 4.4.** Note that for a rooted directed tree, in each node \( j \in V \setminus L \), there is one incoming edge and the rest of the connected edges are outgoing. Its neighbouring nodes are therefore given by the set \( N_j = p(j) \cup \{k : (j,k) \in E_r\} \). Thus,

\[
(A.8) \quad \bigoplus_{k \in N_j} \alpha_{(j,k)} = (K^{(p(j),j)})^T \alpha_{(j,p(j))} \bigoplus_{k : (j,k) \in E_r} K^{(j,k)} \alpha_{(j,k)}.
\]

For the edges \( (j,k) \in E \) it holds

\[
\alpha_{(j,k)} = \bigoplus_{\ell \in N_k \setminus \{j\}} K^{(k,\ell)} \alpha_{(k,\ell)} = \bigoplus_{\ell : (k,\ell) \in E} K^{(k,\ell)} \alpha_{k,\ell},
\]
For the reverse edge \((j, j_p)\) it holds
\[
\alpha_{(j, p(j))} = \bigodot_{\ell \in \mathcal{N}_{p(j)} \setminus \{j\}} K^{(p(j), \ell)} \alpha_{(j, \ell)}
\]
\[
= (K^{(p(p(j)), p(j))})^T \alpha_{(p(j), p(p(j)))} \bigodot_{\ell \in (p(j), \ell) \in \mathcal{E}},
\]
By associating the first term in (A.8) with \(\hat{\varphi}_{j}\), and the second term with \(\varphi_j\), we see that the tensor structure of \(K\) gives rise to the construction of \(\varphi_j\) and \(\hat{\varphi}_j\) as in Theorem 4.1. Hence, we can define a tensor in analogy to \(U\) of the form \(V = (1/v_1) \otimes v_2 \otimes \cdots \otimes v_J\), where \(v_j\) are the vectors from Theorem 4.1 if \(j \in \mathcal{L}\), and \(v_j = 1\) if \(j \in \mathcal{V} \setminus \mathcal{L}\). Then, the tensor \(K \odot V\) has the same marginals as the tensor \(K \odot U\). Due to an extension of Sinkhorn’s theorem to tensors [30], it follows \(U = V\) (up to scaling with a factor and its inverse in the vectors \(u_1, \ldots, u_J\)), and \(P_j(M) = \mu_j\) for all \(j \in \mathcal{V}\).

**Proof of Proposition 4.5.** Based on the proof to Theorem 4.1, a Lagrange dual to problem (4.1) can be formulated as to maximize
\[
- \sum_{(j_1, j_2) \in \mathcal{E}} \left( \mu_{j_1} \odot \exp(-\lambda_{(j_1, j_2)}) \right) A^{(j_1, j_2)} \left( \bigodot_{k:(j_2, k) \in \mathcal{E}} \exp(\lambda_{(j_2, k)}) \right) - \sum_{j \in \mathcal{L}} \left( \mu_{j_p} \odot \exp(-\lambda_{j_p}) \right) A^{(j_1, j_2)} (\exp(\lambda_j)) - \lambda_{1,2}^T \mu_1 + \sum_{j \in \mathcal{L}} \lambda_j^T \mu_j
\]
with respect to \(\mu_j\), for all inner nodes \(j\), and the dual variables \(\lambda_{(j_1, j_2)}\), for \((j_1, j_2) \in \mathcal{E}\), and \(\lambda_j\), for \(j \in \mathcal{L}\). A block coordinate ascent in the dual is then to iteratively maximize (A.9) with respect to one of the dual variable vectors, while keeping the others fixed. Denote \(v_{(j_1, j_2)} = \exp(\lambda_{(j_1, j_2)})\), for \((j_1, j_2) \in \mathcal{E}\), and \(v_j = \exp(\lambda_j)\), for \(j \in \mathcal{L}\). The gradient of (A.9) with respect to \(\lambda_{(1,2)}\) vanishes if it holds that
\[
v_{(1,2)} = A^{(j_1, j_2)} \left( \bigodot_{k:(2, k) \in \mathcal{E}} v_{(2,k)} \right),
\]
and the gradient with respect to \(\lambda_j\), where \(j \in \mathcal{L}\), vanishes if
\[
v_j = \mu_{j_p} / \left( (A^{(j_p,j)})^T (\mu_{j_p} / v_{(j_p,j)}) \right),
\]
where \(j_p \in \mathcal{V}\) is the parent of node \(j\). Finally, the gradient of (A.9) with respect to \(\lambda_{(j_1, j_2)}\), where \(j_1 \neq 1\), vanishes if \(M_{(j_1, j_2)} 1 = M_{(j_p, j)} 1\), where \(j_p\) is the parent of node \(j_1\), which leads to the recursive definition of \(\varphi_j\) and \(\hat{\varphi}_j\), for \(j \in \mathcal{V}\), as in Theorem 4.1. Hence, given an initial set of positive vectors \(v_{(1,2)}\) and \(v_j\), for \(j \in \mathcal{L}\), the scheme (4.6) is a block coordinate ascend in a Lagrange dual of problem (4.1). □

**REFERENCES**

[1] B. Acciaio, J. Backhoff-Veraguas, and R. Carmona, Extended mean field control problems: stochastic maximum principle and transport perspective, SIAM Journal on Control and Optimization, 57 (2019), pp. 3666–3693.

[2] J. Adler, A. Ringh, O. Öktem, and J. Karlsson, Learning to solve inverse problems using Wasserstein loss, arXiv preprint arXiv:1710.10898, (2017).

[3] M. Ager and G. Carlier, Barycenters in the Wasserstein space, SIAM Journal on Mathematical Analysis, 43 (2011), pp. 904–924.
[4] M. Arjovsky, S. Chintala, and L. Bottou, Wasserstein Generative Adversarial Networks, in Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 214–223.

[5] H. Bauschke and A. Lewis, Dykstra’s algorithm with Bregman projections: A convergence proof, Optimization, 48 (2000), pp. 409–427.

[6] E. Bayraktar, A. M. Cox, and Y. Stoev, Martingale optimal transport with stopping, SIAM Journal on Control and Optimization, 56 (2018), pp. 417–433.

[7] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré, Iterative Bregman Projections for Regularized Transportation Problems, SIAM Journal on Scientific Computing, 37 (2015), p. A1111–A1138.

[8] N. Bonneel, M. van de Panne, S. Paris, and W. Heidrich, Displacement interpolation using lagrangian mass transport, ACM Transactions on Graphics (TOG), 30 (2011), pp. 158:1–158:12.

[9] X. Chen, Structure theory for ensemble controllability, observability, and duality, Mathematics of Control, Signals, and Systems, 31 (2019), p. 7.

[10] Y. Chen, G. Conforti, and T. Georgiou, Measure-valued spline curves: An optimal transport viewpoint, SIAM Journal on Mathematical Analysis, 50 (2018), pp. 5947–5968.

[11] Y. Chen, G. Conforti, T. Georgiou, and L. Ripani, Multi-marginal Schrödinger bridges, arXiv preprint arXiv:1902.08319, (2019).

[12] Y. Chen, T. Georgiou, and M. Pavon, Entropic and Displacement Interpolation: A Computational Approach Using the Hilbert Metric, SIAM Journal on Applied Mathematics, 76 (2016), pp. 2375–2396.

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

[14] Y. Chen, T. Georgiou, and M. Pavon, Optimal Steering of a Linear Stochastic System to a Final Probability Distribution, Part I, IEEE Transactions on Automatic Control, 61 (2016), pp. 1158–1169.

[15] Y. Chen, T. Georgiou, M. Pavon, and A. Tannenbaum, Robust transport over networks, IEEE Transactions on Automatic Control, 62 (2016), pp. 4675–4682.

[16] Y. Chen and J. Karlsson, State tracking of linear ensembles via optimal mass transport, IEEE Control Systems Letters, 2 (2018), pp. 260–265.

[17] T. Cover and J. Thomas, Elements of information theory, John Wiley & Sons, Hoboken, NJ, 2006.

[18] I. Csiszar, Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems, The Annals of Statistics, 19 (1991), pp. 2032–2066.

[19] M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transport, in Advances in Neural Information Processing Systems (NIPS), Neural Information Processing Systems Foundation, 2013, pp. 2292–2300.

[20] M. Cuturi and A. Doucet, Fast computation of Wasserstein barycenters, in International Conference on Machine Learning, 2014, pp. 685–693.

[21] D. Dawson, L. Gorostiza, and A. Wakolbinger, Schrödinger processes and large deviations, Journal of mathematical physics, 31 (1990), pp. 2385–2388.

[22] A. Dembo and O. Zeitouni, Large deviations techniques and applications, Springer-Verlag Berlin Heidelberg, 2009.

[23] R. Diestel, Graph Theory, Springer, Berlin, Heidelberg, 2010.

[24] A. Dominitz and A. Tannenbaum, Texture Mapping via Optimal Mass Transport, IEEE Transactions on Visualization and Computer Graphics, 16 (2010), pp. 419–433.

[25] R. Ellis, Entropy, Large Deviations, and Statistical Mechanics, Springer-Verlag Berlin Heidelberg, 2006.

[26] F. Elvander, I. Haasley, A. Jakobsson, and J. Karlsson, Tracking and Sensor Fusion in Direction of Arrival Estimation Using Optimal Mass Transport, in 2018 26th European Signal Processing Conference (EUSIPCO), IEEE, 2018, pp. 1617–1621.

[27] F. Elvander, I. Haasley, A. Jakobsson, and J. Karlsson, Non-Coherent Sensor Fusion via Entropy Regularized Optimal Mass Transport, ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2019), pp. 4415–4419.

[28] F. Elvander, I. Haasley, A. Jakobsson, and J. Karlsson, Multi-Marginal Optimal Mass Transport using Partial Information with Applications in Robust Localization and Sensor Fusion, Signal Processing, (2020).

[29] H. Föllmer, Random fields and diffusion processes, in Ecole d’Été de Probabilités de Saint-Flour XV–XVII, 1985–87, P.-L. Hennequin, ed., vol. 1362 of Lecture Notes in Mathematics, Springer, Berlin, Heidelberg, 1988, pp. 101–203.
[30] J. Franklin and J. Lorenz, *On the scaling of multidimensional matrices*, Linear Algebra and its Applications, 114-115 (1989), pp. 717 – 735. Special Issue Dedicated to Alan J. Hoffman.

[31] T. Georgiou and M. Pavon, *Positive contraction mappings for classical and quantum Schrödinger systems*, Journal of Mathematical Physics, 56 (2015), p. 033301.

[32] N. Ghouchoub, Y.-H. Kim, and A. Z. Palmer, *Optimal transport with controlled dynamics and free end times*, SIAM Journal on Control and Optimization, 56 (2018), pp. 3239-3259.

[33] H. Gezuy, *The Method of Maximum Entropy*, World Scientific, 1995.

[34] I. Haasler, A. Ringh, Y. Chen, and J. Karlsson, *Estimating ensemble flows on a hidden Markov chain*, in 2019 IEEE 58th Conference on Decision and Control (CDC), IEEE, 2019, pp. 1331–1338.

[35] B. Jamison, *Reciprocal processes*, Probability Theory and Related Fields, 30 (1974), pp. 65–86.

[36] B. Jamison, *The markov processes of schrödinger*, Probability Theory and Related Fields, 32 (1975), pp. 323–331.

[37] J. Karlsson and A. Ringh, *Generalized Sinkhorn iterations for regularizing inverse problems using optimal mass transport*, SIAM Journal on Imaging Sciences 10(4), 1935–1962, (2017).

[38] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde, *Optimal Mass Transport: Signal processing and machine-learning applications*, IEEE Signal Process. Mag., 34 (2017), pp. 43–59.

[39] C. Léonard, *A survey of the Schrödinger problem and some of its connections with optimal transport*, Discrete & Continuous Dynamical Systems - A, 34 (2014), pp. 1533–1574.

[40] T. Mikami, *Monge’s problem with a quadratic cost by the zero-noise limit of h-path processes*, Probability Theory and Related Fields, 129 (2004), pp. 245–260.

[41] T. Mikami, *Two end points marginal problem by stochastic optimal transportation*, SIAM Journal on Optimization, 53 (2015), pp. 2449–2461.

[42] T. Mikami and M. Thieullen, *Optimal transportation problem by stochastic optimal control*, SIAM Journal on Control and Optimization, 47 (2008), pp. 1127–1139.

[43] G. Montavon, K.-R. Müller, and M. Cuturr, *Wasserstein training of restricted boltzmann machines*, in Advances in Neural Information Processing Systems, 2016, pp. 3718–3726.

[44] B. Pass, *Multi-marginal optimal transport: theory and applications*, ESAIM: Mathematical Modelling and Numerical Analysis, 49 (2015), pp. 1771–1790.

[45] M. Pavon and F. Ticozzi, *Discrete-time classical and quantum Markovian evolutions: Maximum entropy problems on path space*, Journal of Mathematical Physics, 51 (2010), p. 042104.

[46] E. Schrödinger, *Über die Umkehrung der Naturgesetze*, Sitzungsberichte der Preussischen Akademie der Wissenschaften, Physikalisch-mathematische Klasse, (1931), pp. 144–153.

[47] R. Sinkhorn, *Diagonal Equivalence to Matrices with Prescribed Row and Column Sums*, The American Mathematical Monthly, 74 (1967), pp. 402–405.

[48] J. Solomon, F. de Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas, *Convolutional Wasserstein distances: efficient optimal transportation on geometric domains*, ACM Transactions on Graphics (TOG), 34 (2015), pp. 66:1–66:11.

[49] J. Solomon, R. Rustamov, L. Guibas, and A. Butscher, *Wasserstein Propagation for Semi-Supervised Learning*, Proceedings of the 31st International Conference on Machine Learning, 32 (2014), pp. 306–314.

[50] Y. Teh and M. Welling, *The unified propagation and scaling algorithm*, Advances in neural information processing systems, (2002), pp. 953–960.

[51] P. Tseng, *Dual ascent methods for problems with strictly convex costs and linear constraints: A unified approach*, SIAM Journal on Control and Optimization, 28 (1990), pp. 214–242.

[52] C. Villani, *Optimal transport: Old and new*, Springer, Berlin Heidelberg, 2008.

[53] I. G. Vladimirov and P. I. R., *State distributions and minimum relative entropy noise sequences in uncertain stochastic systems: The discrete-time case*, SIAM Journal on Control and Optimization, 53(3) (2015), p. 1107–1153.

[54] I. Yang, *A Convex Optimization Approach to Distributionally Robust Markov Decision Processes With Wasserstein Distance*, IEEE Control Systems Letters, 1 (2017), pp. 2475–1456.

[55] J. S. Yedidia, W. T. Freeman, and Y. Weiss, *Understanding belief propagation and its generalizations*, in Exploring Artificial Intelligence in the New Millennium, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003, pp. 239–269.

[56] J. S. Yedidia, W. T. Freeman, and Y. Weiss, *Constructing free-energy approximations and generalized belief propagation algorithms*, IEEE Transactions on Information Theory, 51 (2005), pp. 2282–2312.

[57] S. Zeng, *Sample-based population observers*, Automatica, 101 (2019), pp. 166–174.