Graph-structured tensor optimization for nonlinear density control and mean field games

Axel Ringh, Isabel Haasler, Yongxin Chen, and Johan Karlsson

***Abstract*—In this work we develop a numerical method for solving a type of convex graph-structured tensor optimization problems. This type of problems, which can be seen as a generalization of multi-marginal optimal transport problems with graph-structured costs, appear in many applications. In particular, we show that it can be used to model and solve nonlinear density control problems, including convex dynamic network flow problems and multi-species potential mean field games. The method is based on coordinate ascent in a Lagrangian dual, and under mild assumptions we prove that the algorithm converges globally. Moreover, under a set of stricter assumptions, the algorithm converges R-linearly. To perform the coordinate ascent steps one has to compute projections of the tensor, and doing so by brute force is in general not computationally feasible. Nevertheless, for certain graph structures we derive efficient methods for computing these projections. In particular, these graph structures are the ones that occur in convex dynamic network flow problems and multi-species potential mean field games. We also illustrate the methodology on numerical examples from these problem classes.**

***Index Terms*—Tensor optimization problems, Multi-marginal optimal transport, Density control, Mean field games.**

I. Introduction

A strong trend in many research fields is the study of large-scale systems consisting of components that are subsystems with specific characteristics. Examples of such technological systems that are currently emerging include smart electric grids [31], and road networks with self-driving cars [58]. There are also many such problems in biology, ecology, and social sciences, including, e.g., cell, animal, or human populations [78]. A major challenge is to understand and control the macroscopic behavior of such complex large-scale systems.

Since the number of agents is often too large to model each agent individually, the overall system is typically viewed as a flow or density control problem. In this setting, the aggregate state information of the agents is often described by a distribution or density function, and classical problems of this form include, e.g., network flow problems. More recently, there has been a large interest in control and estimation of densities, and one key result is that certain density control problems of first-order integrators can be seen as optimal transport problems [5]. This correspondence can be extended to general dynamics, and thus the optimal transport problem can be interpreted as a density control problem of agents (subsystems) with general dynamics [13], [17], [44].

Even though optimal transport problems are linear programs, the number of variables is often very large and thus the problems can be computationally challenging. A computational breakthrough is the use of Sinkhorn's method for computing an approximate solution. This builds on the insight that the entropy regularized problem can be efficiently solved using dual coordinate ascent [26], [65]. Interestingly, the entropy regularization term can also be interpreted in this setting as introducing stochasticity in the dynamics of the subsystems, and can in fact be shown to be equivalent to the Schrödinger bridge problem [19], [53], [54].

Many density flow problems can be viewed as a two-marginal problem. However, many problems involve using a time grid in order to model, e.g., congestion, instantaneous costs, or observations [39]. For such problems it is natural to use versions of the multi-marginal optimal transport problem, where marginals represent the distributions at different time points $j = 1, \dots, \mathcal{T}$. The multi-marginal optimal transport problem is an optimization problem where a nonnegative tensor is sought to minimize a linear cost subject to constraints on the marginals, where the marginals correspond to projections of the tensor on specific modes.

In this paper, we generalize the multi-marginal optimal transport problem and consider optimization problems on tensors with certain structures that appear in, e.g., density control problems and mean field games. More specifically, for such control problems with identical and indistinguishable agents, the Markov property may be used to separate the problem into $\mathcal{T} - 1$ parts, where each part represents the evolution during time interval $[j, j+1]$ for $j = 1, \dots, \mathcal{T} - 1$. The transition of the agents from time j to time $j + 1$ can thus be specified by the bimarginal projection of the tensor onto the joint two marginals j and $j + 1$, and thus this problem is a structured tensor problem with structure corresponding to a path graph (see, e.g., [23], [29], [40]). However, when the agents have heterogeneous dynamics or objectives, the distribution at a given time does not contain all necessary information about the past and thus the Markov property does not hold. Nevertheless, many problems of interest can instead be modeled by introducing additional dependencies between marginals. For example, traffic flow problems with origin destination constraints can be formulated by introducing dependence between the initial and final node [39], and Euler flow problems can be seen as a special case of this [6]. By introducing an additional marginal representing different types of agents we can also formulate and solve multi-species dynamic flow problems and large multi-commodity problems [41]. Interesting to note in this context is that the algorithms developed to solve this type of structured problems are closely related to the unified propagation and scaling algorithm for inference in graphical models [74].

Many of the problems in the previous paragraph can be formulated as linear optimization problems. Nevertheless, in many situations it is also natural to consider problems with convex costs, for example in potential mean field games [7], but standard convex optimization methods in general do not scale to this type of large-scale problems. The contribution of this paper is therefore to develop a theoretical framework for a type of convex structured tensor optimization problems, along with numerical solution methods and convergence results for these. We also illustrate how this type of problems can be used to model and solve, e.g., convex dynamic flow problems [63], density

This work was supported by the Swedish Research Council (VR) under grant 2020-03454, KTH Digital Futures, the NSF under grant 1942523 and 2008513, the Knut and Alice Wallenberg foundation under grant KAW 2018.0349, and by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

A. Ringh is with Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden. axelri@chalmers.se

I. Haasler, and J. Karlsson are with the Division of Optimization and Systems Theory, Department of Mathematics, KTH Royal Institute of Technology, Stockholm, Sweden. haasler@kth.se, johan.karlsson@math.kth.se

Y. Chen is with the School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, USA. yongchen@gatech.edu

control problems, and multi-species potential mean field games. An important observation is that the dual problem has a decomposable form, and can be efficiently solved using dual coordinate ascent [48] (cf. [65]). Moreover, the structure in these problems can be represented by a graph connecting the marginals, and by utilizing this graph we show how marginal and bimarginal projections of the tensor can be computed efficiently, thus alleviating the computational bottleneck of the algorithm (cf. [41]).

The outline of the paper is as follows: in Section II we introduce some background material on optimal transport and convex optimization. The main results are presented in Section III, where we formulate the graph-structured tensor optimization problem of interest and present a primal-dual framework for solving it, together with a Sinkhorn-type algorithm for iteratively solving the dual problem. Conditions for convergence and R-linear convergence are also presented. Based on this, in Section IV and V we develop algorithms for solving two important types of problems: convex dynamic network flow problems and multi-species potential mean field games, respectively. This is done by casting the corresponding problem as a graph-structured tensor optimization problem and then specializing the general algorithm to the particular instance. Moreover, in each of the two sections we also present numerical examples to illustrate the use and performance of the algorithms. Finally, Section VI contains some concluding remarks. Some proofs are deferred to the appendix for improved readability. This paper builds on [68], where we presented an algorithm, without proof of convergence, for the multi-species mean field game in a simplified setting (see also remark V.3).

II. Background

This section presents background material, in particular on graph-structured multi-marginal optimal transport. We also introduce some concepts from convex analysis and convex optimization that are needed in this work.

A. The graph-structured multi-marginal optimal transport problem

The optimal transport problem seeks a transport plan for how to move mass from an initial distribution to a target distribution with minimum cost. This topic has been extensively studied, see, e.g., the monograph [77] and references therein. An extension of this problem is the multi-marginal optimal transport problem, in which a minimum-cost transport plan between several distributions is sought [6], [29], [36], [59], [64], [70], [71]. In this work we consider the discrete case of the latter, where the marginal distributions are given by a finite set of $\mathcal{T}$ nonnegative vectors[1] $\mu_1, \dots, \mu_\mathcal{T} \in \mathbb{R}^N_+$. The transport plan and the corresponding cost of moving mass are both represented by $\mathcal{T}$-mode tensors $\mathbf{M} \in \mathbb{R}_+^{N^\mathcal{T}}$ and $\mathbf{C} \in \mathbb{R}^{N^\mathcal{T}}$, respectively. More precisely, $\mathbf{M}_{i_1 \dots i_\mathcal{T}}$ and $\mathbf{C}_{i_1 \dots i_\mathcal{T}}$ are the transported mass and the cost of moving mass associated with the tuple $(i_1, \dots, i_\mathcal{T})$, respectively. The total cost of transport is therefore given by

$$\langle \mathbf{C}, \mathbf{M} \rangle := \sum_{i_1, \dots, i_\mathcal{T}} \mathbf{C}_{i_1 \dots i_\mathcal{T}} \mathbf{M}_{i_1 \dots i_\mathcal{T}}.$$

Moreover, for $\mathbf{M}$ to be a feasible transport plan, it must have the given distributions as its marginals. To this end, the marginal distributions of $\mathbf{M}$ are given by the projections $P_j(\mathbf{M}) \in \mathbb{R}^N_+$, where

$$(P_j(\mathbf{M}))_{i_j} := \sum_{i_1, \dots, i_{j-1}, i_{j+1}, i_\mathcal{T}} \mathbf{M}_{i_1 \dots i_\mathcal{T}},$$

and hence $\mathbf{M}$ is feasible if $P_j(\mathbf{M}) = \mu_j$ for $j = 1, \dots, \mathcal{T}$. A generalization of this optimization problem is to not necessarily impose marginal constraints on all projections $P_j(\mathbf{M})$, but only for an index set $\Gamma \subset \{1, \dots, \mathcal{T}\}$. The discrete multi-marginal optimal transport problem can thus be formulated as

$$\begin{aligned} \underset{\mathbf{M} \in \mathbb{R}_+^{N^\mathcal{T}}}{\text{minimize}} \quad & \langle \mathbf{C}, \mathbf{M} \rangle && \text{(1a)} \\ \text{subject to} \quad & P_j(\mathbf{M}) = \mu_j, \quad j \in \Gamma. && \text{(1b)} \end{aligned}$$

Problem (1) is a linear program, however solving it can be computationally challenging due to the large number of variables. An approach for obtaining approximate solutions in the bimarginal case is to add a small entropy term to the cost function and solve the corresponding perturbed problem [26] (see also [65]). This perturbed problem can be solved by using the so-called Sinkhorn iterations.[2] The approach has been extended to the multi-marginal setting [6], [29], [59], however in this case it only partly alleviates the computational difficulty. More precisely, in the multi-marginal setting the entropy term is defined[3] as

$$D(\mathbf{M}) := \sum_{i_1, \dots, i_\mathcal{T}} \big(\mathbf{M}_{i_1 \dots i_\mathcal{T}} \log(\mathbf{M}_{i_1 \dots i_\mathcal{T}}) - \mathbf{M}_{i_1 \dots i_\mathcal{T}} + 1 \big),$$

and the optimal solution to the perturbed problem can be shown to take the form [6], [29]

$$\mathbf{M} = \mathbf{K} \odot \mathbf{U},$$

where $\mathbf{K} = \exp(-\mathbf{C}/\epsilon)$, $\odot$ denotes the elementwise product, and $\mathbf{U}$ is the rank-one tensor

$$\mathbf{U}_{i_1 \dots i_\mathcal{T}} = \prod_{j \in \Gamma} u_j^{(i_j)},$$

i.e., $\mathbf{U} = (\otimes_{j \in \Gamma} u_j) \otimes (\otimes_{j \in \{1, \dots, \mathcal{T}\} \setminus \Gamma} \mathbf{1})$, where $\otimes$ denotes the tensor product and $\mathbf{1}$ denotes a vector of ones. In fact, the variables u_j are the logarithms of the Lagrangian dual variables in a relaxation of the entropy-regularized version of (1). Moreover, the (multi-marginal) Sinkhorn iterations iteratively update u_j to match the given marginals:

$$u_j \leftarrow u_j \odot \mu_j \oslash P_j(\mathbf{K} \odot \mathbf{U}), \qquad \text{for } j \in \Gamma,$$

where $\oslash$ denotes elementwise division. However, in the multi-marginal case, computing $P_j(\mathbf{K} \odot \mathbf{U})$ is challenging since the number of terms in the sum grows exponentially with the number of marginals, and the latter is also reflected in complexity bounds for the algorithm [55]. Nevertheless, in some cases when the underlying cost $\mathbf{C}$ is structured the projections can be computed efficiently. In particular, this is the case for certain graph-structured costs.

To this end, let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a connected graph with $\mathcal{T} = |\mathcal{V}|$ nodes, and consider the optimization problem

$$\begin{aligned} \underset{\mathbf{M} \in \mathbb{R}_+^{N^\mathcal{T}}}{\text{minimize}} \quad & \langle \mathbf{C}, \mathbf{M} \rangle + \epsilon D(\mathbf{M}) && \text{(2a)} \\ \text{subject to} \quad & P_t(\mathbf{M}) = \mu_t, \quad t \in \tilde{\mathcal{V}}, && \text{(2b)} \end{aligned}$$

where $\tilde{\mathcal{V}} \subset \mathcal{V}$ is a set of vertices. Moreover, consider cost tensor $\mathbf{C}$ with the structure

$$\mathbf{C}_{i_1 \dots i_\mathcal{T}} = \sum_{(t_1, t_2) \in \mathcal{E}} C^{(t_1, t_2)}_{i_{t_1}, i_{t_2}}, \tag{3}$$

where $C^{(t_1, t_2)} \in \mathbb{R}^{N \times N}$, which in particular means that the linear cost term takes the form

$$\langle \mathbf{C}, \mathbf{M} \rangle = \sum_{(t_1, t_2) \in \mathcal{E}} \langle C^{(t_1, t_2)}, P_{t_1, t_2}(\mathbf{M}) \rangle.$$

[1] To simplify the notation, we assume that all the marginals have the same number of elements, i.e., $\mu_j \in \mathbb{R}^N$. This can easily be relaxed.

[2] In fact, the iterations have been discovered in different settings and therefore also have many different names; see, e.g., [19], [51].

[3] In this work, we use the convention that $0 \cdot (\pm\infty) = (\pm\infty) \cdot 0 = 0$.

Here $P_{t_1,t_2}(\mathbf{M}) \in \mathbb{R}_+^{N\times N}$ denotes the joint projection of the tensor $\mathbf{M}$ on the two marginals t_1 and t_2, given by

$$(P_{t_1,t_2}(\mathbf{M}))_{i_{t_1},i_{t_2}} := \sum_{\{i_1,\dots,i_{\mathcal{T}}\}\setminus\{i_{t_1},i_{t_2}\}} \mathbf{M}_{i_1\dots i_{\mathcal{T}}}.$$

Problem (2) with a cost tensor structured according to (3) is called a (entropy-regularized) graph-structured multi-marginal optimal transport problem [30], [40], [41]. Moreover, for many graph structures, the projections $P_j(\mathbf{M})$ can be efficiently computed, see, e.g, [2], [6], [29], [30], [39]–[43], [72], and hence the Sinkhorn iterations can be used to efficiently solve such problems.

B. Convex analysis and optimization

We need the following definitions and results from convex analysis and optimization. For extensive treatments of the topic, see, e.g., the monographs [4], [56], [69]. To this end, let $f : \mathbb{R}^n \to \bar{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ be an extended real-valued function. The *epigraph* of f is defined as $\text{epi}(f) := \{(x,\eta) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) \leq \eta\}$, and f is called convex if $\text{epi}(f) \subset \mathbb{R}^{n+1}$ is a convex set. A function f is *lower-semicontinuous* if and only if $\text{epi}(f)$ is closed [69, Thm. 7.1]. The *effective domain* of f is defined as $\text{dom}(f) := \{x \in \mathbb{R}^n \mid f(x) < \infty\}$, and f is called *proper* if $f(x) > -\infty$ for all $x \in \mathbb{R}^n$ and $\text{dom}(f) \neq \emptyset$. A convex set C is called *polyhedral* if it can be written as the intersection of a finite number of closed half spaces. A convex function f is called polyhedral if $\text{epi}(f)$ is polyhedral. The *Fenchel conjugate* of a function f is defined as $f^*(x^*) := \sup_x \langle x^*, x\rangle - f(x)$. A convex, proper, lower-semicontinuous function f is called *co-finite* if $\text{epi}(f)$ contains no non-vertical half-lines, which is equivalent to that f^* is finite everywhere, i.e., that $\text{dom}(f^*) = \mathbb{R}^n$, [69, Cor. 13.3.1].

The *subdifferential* of a function f in a point x is the set $\partial f(x) := \{u \in \mathbb{R}^n \mid \langle y - x, u\rangle + f(x) \leq f(y)\ \forall\, y \in \mathbb{R}^n\}$, and if f is proper, convex, and differentiable in x with gradient $\nabla f(x)$, then $\partial f(x) = \{\nabla f(x)\}$ [4, Prop. 17.31]. A convex, proper, lower-semicontinuous function f is called *essentially smooth* if i) it is differentiable on $\text{int}(\text{dom}(f))$, i.e., on the interior of the effective domain, and ii) $\lim_{\ell\to\infty} \|\nabla f(x_\ell)\| \to \infty$ for any sequence $\{x_\ell\}_\ell \subset \text{int}(\text{dom}(f))$ that either converges to the boundary of $\text{int}(\text{dom}(f))$ or is such that $\|x_\ell\| \to \infty$. Finally, an operator $A : \mathbb{R}^n \to \mathbb{R}^n$ is called strongly monotone if there exists a $\gamma > 0$ such that $\langle Ax - Ay, x - y\rangle \geq \gamma\|x - y\|^2$ for all $x, y \in \mathbb{R}^n$.

III. Convex graph-structured tensor optimization

In this work, we consider a family of optimization problems that generalizes problems of the form (2). To this end, let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a connected graph with $\mathcal{T} = |\mathcal{V}|$ nodes, and let $\mathbf{C} \in \mathbb{R}^{N^{\mathcal{T}}}$ be a cost tensor that takes the form (3). The convex graph-structured tensor optimization problems of interest are problems of the form

$$\begin{aligned}\underset{\mathbf{M}\in\mathbb{R}_+^{N^{\mathcal{T}}}}{\text{minimize}} \quad & \langle \mathbf{C}, \mathbf{M}\rangle + \epsilon D(\mathbf{M}) + \sum_{t\in\mathcal{V}} g^{(t)}(P_t(\mathbf{M})) \\ & + \sum_{(t_1,t_2)\in\mathcal{E}} f^{(t_1,t_2)}(P_{t_1,t_2}(\mathbf{M})),\end{aligned} \tag{4}$$

where $g^{(t)}$ and $f^{(t_1,t_2)}$ are proper, convex, and lower-semicontinuous functionals; further assumptions on these functionals will be imposed where needed. The reason for our interest in problems of the form (4) is that a number of different applications can be modeled as such problems. In particular, this is true for convex dynamic network flow problems, and potential multi-species mean field games. These two applications are studied in detail in Sections IV and V, respectively.

Remark III.1. *To see that problems of the form* (4) *is a generalization of the graph-structured optimal transport problem* (2)*, let $I_A(\cdot)$ denote the indicator function on the set A, i.e., the function*

$$I_A(x) := \begin{cases} 0, & \text{if } x \in A \\ \infty, & \text{else,}\end{cases}$$

and note that this function is proper, convex, and lower-semicontinuous if and only if A is a nonempty, closed, convex set. Now, (2) *is recovered from* (4) *by taking $g^{(t)}(P_t(\mathbf{M})) = I_{\{\mu_t\}}(P_t(\mathbf{M}))$ for $t \in \tilde{\mathcal{V}}$ and $g^{(t)}(P_t(\mathbf{M})) \equiv 0$ otherwise, and $f^{(t_1,t_2)}(P_{t_1,t_2}(\mathbf{M})) \equiv 0$ for all $(t_1,t_2) \in \mathcal{E}$. Another particular case of interest is a version of* (2)*, but where some of the equality constraints are replaced by inequality constraints, cf. [41].*

Remark III.2. *In problem* (4) *the functions $f^{(t_1,t_2)}$ and the tensor $\mathbf{C}$ are defined on the same set of edges $\mathcal{E}$. This is done for convenience of notation, and is not a restrictive assumption. To see this, note that it is possible to define certain functions $f^{(t_1,t_2)}$ to be the zero function, or to take certain matrices $C^{(t_1,t_2)}$ in the decomposition* (3) *to be the zero-matrix.*

Note that (4) is typically a large-scale problem, where the full set of variables may neither be stored nor manipulated directly. Therefore one must utilize the problem structure in order to compute the solution. In this section, we develop a method for such problems, based on generalized Sinkhorn iterations. This methodology for handling the problem builds on deriving the Lagrangian dual of an optimization problem that is equivalent to (4), and solving this dual using coordinate ascent. As we will see, the method exploits the graph structure and the algorithm is efficient when the graph is simple, i.e., the tree-width is low (cf. [30], [41], [43]), and when the functionals $g^{(t)}$ and $f^{(t_1,t_2)}$ are in some sense simple.

A. An equivalent problem and existence of solution

We first introduce and analyze a problem that is equivalent to (4), and give conditions under which the latter has an optimal solution. To this end, introducing the variables μ_t, $t \in \mathcal{V}$, and R_{t_1,t_2}, $(t_1,t_2) \in \mathcal{E}$, problem (4) can be rewritten as

$$\begin{aligned}\underset{\substack{\mathbf{M}\in\mathbb{R}_+^{N^{\mathcal{T}}},\ \mu_t\in\mathbb{R}^N,\ t\in\mathcal{V} \\ R_{t_1,t_2}\in\mathbb{R}^{N\times N},\ (t_1,t_2)\in\mathcal{E}}}{\text{minimize}} \quad & \langle \mathbf{C}, \mathbf{M}\rangle + \epsilon D(\mathbf{M}) + \sum_{t\in\mathcal{V}} g^{(t)}(\mu_t) \\ & + \sum_{(t_1,t_2)\in\mathcal{E}} f^{(t_1,t_2)}(R_{t_1,t_2}) && \text{(5a)} \\ \text{subject to} \quad & P_t(\mathbf{M}) = \mu_t,\ t \in \mathcal{V} && \text{(5b)} \\ & P_{t_1,t_2}(\mathbf{M}) = R_{t_1,t_2},\ (t_1,t_2) \in \mathcal{E}. && \text{(5c)}\end{aligned}$$

In order for this to be a well-posed problem, we impose the following assumptions on the functionals involved.

Assumption A. *Assume that all elements of $\mathbf{C}$ are strictly larger than $-\infty$, and that $g^{(t)}$, $t \in \mathcal{V}$, and $f^{(t_1,t_2)}$, $(t_1,t_2) \in \mathcal{E}$, are all proper, convex, and lower-semicontinuous, Moreover, assume that there exists a feasible point to* (5) *with finite objective function value, i.e., a nonnegative tensor $\mathbf{M}$ such that $\langle \mathbf{C}, \mathbf{M}\rangle < \infty$, and with marginals and bimarginals as in* (5b)-(5c)*, respectively, such that*

$$\begin{aligned} & g^{(t)}(\mu_t) < \infty, \quad \text{for all } t \in \mathcal{V}, \\ & f^{(t_1,t_2)}(R_{t_1,t_2}) < \infty, \quad \text{for all } (t_1,t_2) \in \mathcal{E}.\end{aligned}$$

In fact, this assumption ensures that (5) has an optimal solution, as stated in the following lemma.

Lemma III.3. *If Assumption A holds, then there exists a unique optimal solution to problem* (5)*.*

Proof: See Appendix A. ∎

Remark III.4. *A necessary condition for Assumption A to hold is that there exist vectors $\mu_t \in \mathbb{R}^N_+ \cap dom(g^{(t)})$, for all $t \in \mathcal{V}$, matrices $R_{t_1,t_2} \in \mathbb{R}^{N\times N}_+ \cap dom(f^{(t_1,t_2)})$, for all $(t_1,t_2) \in \mathcal{E}$, and a constant $\gamma \geq 0$ such that*

$$\begin{aligned}&\mu_t^T \mathbf{1} = \gamma, \quad \text{for all } t \in \mathcal{V}\\ &R_{t_1,t_2}\mathbf{1} = \mu_{t_1},\ R_{t_1,t_2}^T\mathbf{1} = \mu_{t_2}, \quad \text{for all } (t_1,t_2) \in \mathcal{E},\\ &\langle C^{(t_1,t_2)}, R_{t_1,t_2}\rangle < \infty, \quad \text{for all } (t_1,t_2) \in \mathcal{E}.\end{aligned}$$

However, unless the graph $(\mathcal{V},\mathcal{E})$ is a tree, this is not a sufficient condition for the existence of a tensor that fulfills Assumption A. More precisely, the existence of marginals and bimarginals that are consistent with each other does, in general, not guarantee that there exists a tensor that matches the marginals and bimarginals. A counterexample can be found in [40, Rem. 3].

B. Form of the optimal solution and Lagrangian dual

Next, we derive the Lagrangian dual of (5) and show that there is no duality gap between the primal and the dual problem.

Theorem III.5. *A Lagrangian dual of* (5) *is, up to a constant, given by*

$$\begin{aligned}\sup_{\substack{\lambda_t\in\mathbb{R}^N,\ t\in\mathcal{V}\\ \Lambda_{t_1,t_2}\in\mathbb{R}^{N\times N},\ (t_1,t_2)\in\mathcal{E}}} \quad &-\epsilon\langle \mathbf{K},\mathbf{U}\rangle - \sum_{t\in\mathcal{V}} (g^{(t)})^*(-\lambda_t)\\ &- \sum_{(t_1,t_2)\in\mathcal{E}} (f^{(t_1,t_2)})^*(-\Lambda_{t_1,t_2}),\end{aligned} \tag{6}$$

where $\mathbf{K}$ and $\mathbf{U}$ are given by

$$\mathbf{K}_{i_1\dots i_\mathcal{T}} = \exp(-\mathbf{C}_{i_1\dots i_\mathcal{T}}/\epsilon), \tag{7a}$$

$$\begin{aligned}\mathbf{U}_{i_1\dots i_\mathcal{T}} &= \prod_{t\in\mathcal{V}} u_t^{(i_t)} \prod_{(t_1,t_2)\in\mathcal{E}} U_{t_1,t_2}^{(i_{t_1},i_{t_2})}\\ &= \prod_{t\in\mathcal{V}} \exp\left(\lambda_t^{(i_t)}/\epsilon\right) \prod_{(t_1,t_2)\in\mathcal{E}} \exp\left(\Lambda_{t_1,t_2}^{(i_{t_1},i_{t_2})}/\epsilon\right).\end{aligned} \tag{7b}$$

Moreover, under Assumption A, the minimum in (5) *equals the supremum in* (6) *(up to the discarded constant). Finally, if the dual* (6) *has an optimal solution, then the optimal solution to the primal problem takes the form $\mathbf{M}^\star = \mathbf{K}\odot\mathbf{U}^\star$, where $\mathbf{U}^\star$ is obtained via* (7b) *from an optimal solution to* (6)*.*

Proof: Relaxing each constraint (5b) and (5c) with a multiplier $\lambda_t \in \mathbb{R}^N$ and $\Lambda_{t_1,t_2} \in \mathbb{R}^{N\times N}$, respectively, we get the Lagrangian

$$\begin{aligned}L(\mathbf{M},\mu,R,\lambda,\Lambda) &:= \langle \mathbf{C},\mathbf{M}\rangle + \epsilon D(\mathbf{M}) + \sum_{t\in\mathcal{V}} g^{(t)}(\mu_t)\\ &+ \sum_{(t_1,t_2)\in\mathcal{E}} f^{(t_1,t_2)}(R_{t_1,t_2}) + \sum_{t\in\mathcal{V}} \lambda_t^T(\mu_t - P_t(\mathbf{M}))\\ &+ \sum_{(t_1,t_2)\in\mathcal{E}} \mathrm{tr}[\Lambda_{t_1,t_2}^T(R_{t_1,t_2} - P_{t_1,t_2}(\mathbf{M}))],\end{aligned} \tag{8}$$

where μ denote $(\mu_t)_{t\in\mathcal{V}}$, and similar for all other variables. The dual function is given by $\inf L$ over $\mathbf{M}$, μ, and R, but the Lagrangian decouples over $\mathbf{M}$, μ_t, and R_{t_1,t_2}. For the inf over μ_t we have that

$$\inf_{\mu_t} \lambda_t^T\mu_t + g^{(t)}(\mu_t) = -\sup_{\mu_t}(-\lambda_t)^T\mu_t - g^{(t)}(\mu_t) = -(g^{(t)})^*(-\lambda_t)$$

where * denotes the Fenchel conjugate; and analogous result follows for $f^{(t_1,t_2)}$ and the inf over R_{t_1,t_2}. This means that

$$\begin{aligned}\inf_{\mathbf{M}\geq 0,\mu,R} L(\mathbf{M},\mu,R,\lambda,\Lambda) &= \inf_{\mathbf{M}\geq 0} \mathcal{L}(\mathbf{M},\lambda,\Lambda)\\ &- \sum_{t\in\mathcal{V}} (g^{(t)})^*(-\lambda_t) - \sum_{(t_1,t_2)\in\mathcal{E}} (f^{(t_1,t_2)})^*(-\Lambda_{t_1,t_2})\end{aligned} \tag{9}$$

where $\mathcal{L}(\mathbf{M},\lambda,\Lambda)$ is defined as

$$\langle \mathbf{C},\mathbf{M}\rangle + \epsilon D(\mathbf{M}) - \sum_{t\in\mathcal{V}} \lambda_t^T P_t(\mathbf{M}) - \sum_{(t_1,t_2)\in\mathcal{E}} \mathrm{tr}[\Lambda_{t_1,t_2}^T P_{t_1,t_2}(\mathbf{M})].$$

Now, noticing that

$$\begin{aligned}\lambda_t^T P_t(\mathbf{M}) &= \sum_{i_t=1}^N \lambda_t^{(i_t)} \sum_{i_1,\dots,i_\mathcal{T}\setminus\{i_t\}} \mathbf{M}_{i_1\dots i_\mathcal{T}} = \sum_{i_1,\dots,i_\mathcal{T}} \lambda_t^{(i_t)}\mathbf{M}_{i_1\dots i_\mathcal{T}},\\ \mathrm{tr}[\Lambda_{t_1,t_2}^T P_{t_1,t_2}(\mathbf{M})] &= \sum_{i_1,\dots,i_\mathcal{T}} \Lambda_{t_1,t_2}^{(i_{t_1},i_{t_2})}\mathbf{M}_{i_1\dots i_\mathcal{T}},\end{aligned}$$

$\mathcal{L}(\mathbf{M},\lambda,\Lambda)$ decouples over the elements of the tensor. Therefore, the inf in each element is either attained in 0, or found by setting the first variation to 0. If $\mathbf{C}_{i_1\dots i_\mathcal{T}} = \infty$, then the trivial case $\mathbf{M}_{i_1\dots i_\mathcal{T}} = 0$ holds. Otherwise, setting the first variation equal to 0 gives

$$0 = \mathbf{C}_{i_1\dots i_\mathcal{T}} + \epsilon\log(\mathbf{M}_{i_1\dots i_\mathcal{T}}) - \sum_{t\in\mathcal{V}} \lambda_t^{(i_t)} - \sum_{(t_1,t_2)\in\mathcal{E}} \Lambda_{t_1,t_2}^{(i_{t_1},i_{t_2})}$$

from which it can be seen that the optimum is such that $\mathbf{M}_{i_1\dots i_\mathcal{T}} > 0$. Moreover, solving for $\mathbf{M}_{i_1\dots i_\mathcal{T}}$ gives that $\mathbf{M} = \mathbf{K}\odot\mathbf{U}$, where $\mathbf{K}$ and $\mathbf{U}$ are given in (7). Note that this form for $\mathbf{M}$ also holds for the elements of $\mathbf{C}$ that are infinite. Plugging this back into $\mathcal{L}(\mathbf{M},\lambda,\Lambda)$ we get that $\inf_{\mathbf{M}\geq 0}\mathcal{L}(\mathbf{M},\lambda,\Lambda) = -\epsilon\langle \mathbf{K},\mathbf{U}\rangle + N^\mathcal{T}\epsilon$, which, after removing the constant, together with (9) gives the dual problem (6).

Finally, for improved readability the detailed proof of that there is no duality gap is deferred to Lemma A.2 in Appendix A. ∎

By using the change of variables implicit in (7b), problem (6) can be expressed equivalently as

$$\begin{aligned}\sup_{\substack{u_t\in\mathbb{R}^N_+,\ t\in\mathcal{V}\\ U_{t_1,t_2}\in\mathbb{R}^{N\times N}_+,\ (t_1,t_2)\in\mathcal{E}}} \quad &-\epsilon\langle \mathbf{K},\mathbf{U}\rangle - \sum_{t\in\mathcal{V}} (g^{(t)})^*\big(-\epsilon\log(u_t)\big)\\ &- \sum_{(t_1,t_2)\in\mathcal{E}} (f^{(t_1,t_2)})^*\big(-\epsilon\log(U_{t_1,t_2})\big).\end{aligned} \tag{10}$$

Moreover, under a Slater-type condition of for the primal problem, i.e., that the relative interior (denoted ri)[4] of the effective domains of the cost functions in (4) have a nonempty intersection, we have that the suprema in (6) and (10) are attained.

Assumption B. *Assume that there exists an $\mathbf{M} > 0$ such that $\langle \mathbf{C},\mathbf{M}\rangle < \infty$, and with marginals and bimarginals $(\mu_t)_{t\in\mathcal{V}}$ and $(R_{t_1,t_2})_{(t_1,t_2)\in\mathcal{E}}$ satisfying* (5b) *and* (5c)*, respectively, so that*

- *for all $g^{(t)}$ and $f^{(t_1,t_2)}$ that are polyhedral, $\mu_t \in dom(g^{(t)})$ and $R_{t_1,t_2} \in dom(f^{(t_1,t_2)})$,*
- *for all $g^{(t)}$ and $f^{(t_1,t_2)}$ that are not polyhedral, $\mu_t \in ri(dom(g^{(t)}))$ and $R_{t_1,t_2} \in ri(dom(f^{(t_1,t_2)}))$.*

Corollary III.6. *Given Assumption B the conclusions of Theorem III.5 holds, with the addition that the dual* (6) *is guaranteed to have a nonempty set of optimal solutions.*

Proof: The result follows from [69, Ch. 29 and 30]. ∎

Nevertheless, even if the Slater-type condition in Assumption B is not fulfilled, the form $\mathbf{M} = \mathbf{K}\odot\mathbf{U}$ will be important in deriving a convergent algorithm for solving (5).

Remark III.7. *For the unregularized problem, i.e., problem* (4) *without the term $\epsilon D(\mathbf{M})$, the dual can be obtained by a slight*

[4]The relative interior of a set A consists of all points in A that are interior when A is regarded as a subset of its affine hull, see [69, Ch. 6].

modification of the above argument. More precisely, it is given by

$$\begin{aligned}\sup_{\substack{\lambda_t\in\mathbb{R}^N,\ t\in\mathcal{V}\\ \Lambda_{t_1,t_2}\in\mathbb{R}^{N\times N},\ (t_1,t_2)\in\mathcal{E}}} \quad & -\sum_{t\in\mathcal{V}}(g^{(t)})^*(-\lambda_t) \\ & -\sum_{(t_1,t_2)\in\mathcal{E}}(f^{(t_1,t_2)})^*(-\Lambda_{t_1,t_2}) \\ \text{subject to} \quad & \sum_{t\in\mathcal{V}}\lambda_t^{(i_t)}+\sum_{(t_1,t_2)\in\mathcal{E}}\Lambda_{t_1,t_2}^{(i_{t_1},i_{t_2})}\leq \mathbf{C}_{i_1\dots i_{\mathcal{T}}}.\end{aligned} \tag{11}$$

In fact, the term $-\epsilon\langle\mathbf{K},\mathbf{U}\rangle$ in (6) *can be interpreted as a barrier term for the constraint in* (11) *in the sense that when $\epsilon\to 0$ then the term $-\epsilon\langle\mathbf{K},\mathbf{U}\rangle\to-\infty$ if the constraint in* (11) *is not fulfilled. Nevertheless, for the unregularized problem one need to impose further assumptions in order to guarantee the existence of an optimal solution. To see this, note that for a problem instance where $g^{(t)}(\mu_t)\equiv 0$, $t\in\mathcal{V}$, $f^{(t_1,t_2)}(R_{t_1,t_2})\equiv 0$, $(t_1,t_2)\in\mathcal{E}$, and where $\mathbf{C}$ has at least one element which is negative, the primal problem is unbounded from below.*

C. Coordinate ascent iterations for solving the dual problem

In this section we derive an efficient solution method for (5), based on performing coordinate ascent in the dual problem (6) (or, equivalently, in (10)). To this end, let $\phi((\lambda_t)_{t\in\mathcal{V}},(\Lambda_{t_1,t_2})_{(t_1,t_2)\in\mathcal{E}})$ denote the objective function in the dual problem (6). Given an iterate $((\lambda_t^k)_{t\in\mathcal{V}},(\Lambda_{t_1,t_2}^k)_{(t_1,t_2)\in\mathcal{E}})$, in a coordinate ascent step we cyclically select an element $j\in\mathcal{V}$ or $(j_1,j_2)\in\mathcal{E}$ and compute an update to the corresponding variable by taking λ_j^{k+1} to be in

$$\operatorname*{arg\,max}_{\lambda_j\in\mathbb{R}^N}\quad \phi(\lambda_j,(\lambda_t^k)_{t\in\mathcal{V}\setminus\{j\}},(\Lambda_{t_1,t_2}^k)_{(t_1,t_2)\in\mathcal{E}}), \tag{12a}$$

or Λ_{j_1,j_2}^{k+1} to be in

$$\operatorname*{arg\,max}_{\Lambda_{j_1,j_2}\in\mathbb{R}^{N\times N}}\phi(\Lambda_{j_1,j_2},(\lambda_t^k)_{t\in\mathcal{V}},(\Lambda_{t_1,t_2}^k)_{(t_1,t_2)\in\mathcal{E}\setminus\{(j_1,j_2)\}}), \tag{12b}$$

respectively, while taking $\lambda_t^{k+1}=\lambda_t^k$ and $\Lambda_{t_1,t_2}^{k+1}=\Lambda_{t_1,t_2}^k$ for all other elements. In order for this to be a well-defined algorithm, we need that the set of maximizing arguments in (12) is always nonempty. To guarantee this, we impose the following assumption (which is milder than Assumption B).

Assumption C. *Assume that $\mathbf{C}<\infty$ and that for each index $j\in\mathcal{V}$ there exists a $\mu_j>0$ so that*

- *if $g^{(j)}$ is polyhedral, then $\mu_j\in dom(g^{(j)})$,*
- *if $g^{(j)}$ is not polyhedral, then $\mu_j\in ri(dom(g^{(j)}))$,*

and analogously for each index $(j_1,j_2)\in\mathcal{E}$, R_{j_1,j_2}, and $f^{(j_1,j_2)}$.

Lemma III.8. *Under Assumptions A and C, the subproblems in* (12) *always have a nonempty set of maximizers.*

Proof: To prove the lemma, we restrict our attention to one subproblem of the form (12a); for subproblems of the form (12b) it follows analogously. Now, note that problem (12a) can be see as the Lagrangian dual of the primal problem

$$\begin{aligned}\operatorname*{minimize}_{\mathbf{M}\in\mathbb{R}_+^{N^{\mathcal{T}}},\ \mu_j\in\mathbb{R}^N}\quad & \langle\mathbf{C},\mathbf{M}\rangle+\epsilon D(\mathbf{M})+g^{(j)}(\mu_j)-\sum_{t\in\mathcal{V}\setminus\{j\}}(\lambda_t^k)^T P_t(\mathbf{M}) \\ & -\sum_{(t_1,t_2)\in\mathcal{E}}\operatorname{tr}[(\Lambda_{t_1,t_2}^k)^T P_{t_1,t_2}(\mathbf{M})] \\ \text{subject to}\quad & P_j(\mathbf{M})=\mu_j.\end{aligned}$$

Moreover, using Assumption C we have that $\mu_j>0$ and $\mathbf{M}=\mu_j\otimes(\otimes_{t\in\mathcal{V}\setminus\{j\}}\mathbf{1})>0$ is a point fulfilling Slater's condition for the above problem. Therefore, following [69, Ch. 29 and 30] we have that strong duality holds between these two problems, and in particular that the dual (12a) has a nonempty set of maximizers (cf. [75, Lem. 3.1]). ■

By the above lemma, the coordinate ascent steps in (12) are well-defined. Moreover, since each problem is concave and unconstrained, the optimal solution is where the subgradient is zero. To compute the subgradients, first note that

$$P_j(\mathbf{K}\odot\mathbf{U})\oslash u_j=\sum_{i_1,\dots,i_{\mathcal{T}}\setminus i_j}\mathbf{K}_{i_1\dots i_{\mathcal{T}}}\prod_{t\in\mathcal{V}\setminus\{j\}}u_t^{(i_t)}\prod_{(t_1,t_2)\in\mathcal{E}}U_{t_1,t_2}^{(i_{t_1},i_{t_2})}$$

is a well-defined vector which is independent of u_j. We therefore define

$$w_j:=P_j(\mathbf{K}\odot\mathbf{U})\oslash u_j, \tag{13a}$$

and note that this means that $P_j(\mathbf{K}\odot\mathbf{U})=u_j\odot w_j$. Analogously, we also define

$$W_{j_1,j_2}:=P_{j_1,j_2}(\mathbf{K}\odot\mathbf{U})\oslash U_{j_1,j_2}, \tag{13b}$$

which in the same way is a well-defined matrix, independent of U_{j_1,j_2}, and hence $P_{j_1,j_2}(\mathbf{K}\odot\mathbf{U})=U_{j_1,j_2}\odot W_{j_1,j_2}$.

Next, note that

$$\frac{\partial}{\partial\lambda_j^{(i_j)}}\langle\mathbf{K},\mathbf{U}\rangle=-\exp\left(\lambda_j^{(i_j)}/\epsilon\right)w_j^{(i_j)}=-u_j^{(i_j)}w_j^{(i_j)}$$

with $\mathbf{K}$ and $\mathbf{U}$ given as in (7) and w_j as in (13a). Thus, in each update of the variable λ_j one has to solve the inclusion problem

$$0\in\partial_{\lambda_j}\phi=-\exp(\lambda_j/\epsilon)\odot w_j+\partial(g^{(j)})^*(-\lambda_j), \tag{14a}$$

where ∂_{λ_j} denotes the subdifferential with respect to λ_j. By an analogous derivation, in each update of the variable Λ_{j_1,j_2} one has to solve the inclusion problem

$$\begin{aligned}0\in\partial_{\Lambda_{j_1,j_2}}\phi=&-\exp(\Lambda_{j_1,j_2}/\epsilon)\odot W_{j_1,j_2}\\ &+\partial(f^{(j_1,j_2)})^*(-\Lambda_{j_1,j_2}).\end{aligned} \tag{14b}$$

These inclusions, and hence the updates, can be reformulated in terms of the transformed dual variables u_j and U_{j_1,j_2}, in which case they read

$$0\in -u_j\odot w_j+\partial(g^{(j)})^*\big(-\epsilon\log(u_j)\big), \tag{15a}$$

$$0\in -U_{j_1,j_2}\odot W_{j_1,j_2}+\partial(f^{(j_1,j_2)})^*\big(-\epsilon\log(U_{j_1,j_2})\big). \tag{15b}$$

This is summarized in Algorithm 1. However, note that directly computing w_j and W_{j_1,j_2} needed in (15) by brute-force is computationally demanding, and effectively numerically infeasible for large-scale problems. Therefore, from this perspective Algorithm 1 is an "abstract algorithm". Nevertheless, for many graph structures it is possible to compute the projections efficiently by sequentially eliminating the modes of the tensor, see [2], [6], [29], [39]–[43], [72]. In fact, storing and using intermediate results of eliminated modes, the procedure can also be understood as a message-passing scheme [43]. Moreover, the examples of applications presented in Sections IV and V can be effectively solved in this way, and for each of these we present an implementable algorithm (see Algorithms 2 and 3). Finally, under relatively mild assumptions, Algorithm 1 is convergent in the following sense.

Theorem III.9. *Given that Assumptions A and C hold, and assume further that*

1) *$g^{(t)}$, $t\in\mathcal{V}$, and $f^{(t_1,t_2)}$, $(t_1,t_2)\in\mathcal{E}$, are all continuous on $dom(g^{(t)})$ and $dom(f^{(t_1,t_2)})$, respectively,*
2) *for all $g^{(t)}$, $t\in\mathcal{V}$, and $f^{(t_1,t_2)}$, $(t_1,t_2)\in\mathcal{E}$, that are not polyhedral, the feasible point in Assumption A is such that $\mu_t\in ri(dom(g^{(t)}))$ and $R_{t_1,t_2}\in ri(dom(f^{(t_1,t_2)}))$, respectively.*

Algorithm 1 Generalized Sinkhorn method for solving (5).

Give: graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, cost tensor $\mathbf{C}$ that decouples according to $\mathcal{G}$, functions $(g^{(t)})^*$, for $t \in \mathcal{V}$, and $(f^{(t_1,t_2)})^*$, for $(t_1, t_2) \in \mathcal{E}$, nonnegative initial guesses $(u_t^0)_{t\in\mathcal{V}}$ and $(U_{t_1,t_2}^0)_{(t_1,t_2)\in\mathcal{E}}$.
$k = 0$
while Not converged **do**
  $k = k + 1$
  for $j \in \mathcal{V}$ and $(j_1, j_2) \in \mathcal{E}$ **do**
    Update u_j^k by solving (15a) with w_j as in (13a).
    Update U_{j_1,j_2}^k by solving (15b) with W_{j_1,j_2} as in (13b).
  end for
end while
return $(u_t^k)_{t\in\mathcal{V}}$ and $(U_{t_1,t_2}^k)_{(t_1,t_2)\in\mathcal{E}}$.

Let $(u_t^k)_{t\in\mathcal{V}}$ and $(U_{t_1,t_2}^k)_{(t_1,t_2)\in\mathcal{E}}$ be the iterates of Algorithm 1 at iteration k, and let $\mathbf{U}^k$ be the corresponding tensor as in (7b)*. Moreover, let $\mathbf{M}^k = \mathbf{K} \odot \mathbf{U}^k$. Then $(\mathbf{M}^k)_k$ is a bounded sequence that converges to the optimal solution to* (5)*. Furthermore, if the set of optimal solutions to* (6) *is nonempty and bounded, then $((u_t^k)_{t\in\mathcal{V}}, (U_{t_1,t_2}^k)_{(t_1,t_2)\in\mathcal{E}})_k$ is a bounded sequence and every cluster point is an optimal solution to* (10)*.*

Proof: To prove the theorem, we define

$$h(\mathbf{M}) := \langle \mathbf{C}, \mathbf{M}\rangle + \epsilon D(\mathbf{M}).$$

This is a strictly convex function, since the entropy term is strictly convex [4, Ex. 9.35]. Moreover $\mathrm{dom}(h) = \mathbb{R}_+^{N^{\mathcal{T}}}$, and hence polyhedral. Next, the Fenchel conjugate of h can be obtained by using [4, Ex. 13.2] and [4, Prop 13.23], which gives

$$\begin{aligned} h^*(\mathbf{T}) &= -\epsilon \sum_{i_1\dots i_{\mathcal{T}}} \exp((\mathbf{T}_{i_1\dots i_{\mathcal{T}}} - \mathbf{C}_{i_1\dots i_{\mathcal{T}}})/\epsilon) - 1 \\ &= -\epsilon\langle \mathbf{K}, \exp(\mathbf{T}/\epsilon)\rangle + n^{\mathcal{T}}\epsilon, \end{aligned}$$

(compare with the expression $\inf_{\mathbf{M}} \mathcal{L}(\mathbf{M}, \lambda, \Lambda) = -\epsilon\langle \mathbf{K}, \mathbf{U}\rangle + N^{\mathcal{T}}\epsilon$ in the proof of Theorem III.5). Hence, h is co-finite. Therefore, following along the lines of [75, Sec. 6], we have that $(\mathbf{M}^k)_k$ is a bounded sequence and that every cluster point is an optimal solution to (5). In particular, [75, Thm. 3.1] imposes some slightly stronger assumptions,[5] but it is readily checked in all places where these stronger assumptions are invoked that the same conclusions hold true in this particular case under the weaker assumptions. For brevity, we omit the details.

Next, since $(\mathbf{M}^k)_k$ is a bounded sequence and every cluster point is optimal to (5), by the uniqueness of the optimal solution $\mathbf{M}^\star$ the sequence must converge to it. To see this, for $\delta > 0$ let $B_\delta(\mathbf{M}^\star)$ denote a ball of radius δ around $\mathbf{M}^\star$. If there is an infinite number of points of $(\mathbf{M}^k)_k$ outside of $B_\delta(\mathbf{M}^\star)$, since the corresponding subsequence is also bounded there is a converging subsubsequence, i.e., the subsubsequence has a cluster point; call this cluster point $\mathbf{M}^\infty$. Since this is also a cluster point of the original sequence, we must have $\mathbf{M}^\infty = \mathbf{M}^\star$, but we must also have that $\mathbf{M}^\infty \notin B_\delta(\mathbf{M}^\star)$, which is a contradiction. Therefore, for every $\delta > 0$ there can only be a finite number of points in the sequence $(\mathbf{M}^k)_k$ that do not belong to $B_\delta(\mathbf{M}^\star)$, and hence the entire sequence converges to $\mathbf{M}^\star$.

Finally, the last statement of the theorem follows similarly from [75, Thm. 3.1(b)]. ∎

[5]More precisely, to directly apply the result in [75, Thm. 3.1], we must assume that the feasible point in Assumption A is such that $\mathbf{M} > 0$; see [75, Ass. B] where "f_0" corresponds to $\langle \mathbf{C}, \mathbf{M}\rangle + \epsilon D(\mathbf{M})$. For an example of where this weaker assumption is indeed used, see Example III.12.

The above theorem guarantees convergence, but does not guarantee how fast the iterates converge. In particular, in order to guarantee R-linear convergence (for the definition, see, e.g., [62, Sec. 9.2] or [60, pp. 619-620]) we need to impose further assumptions on the functions involved.

Theorem III.10. *Given Assumption A, further assume that there exists an $\mathbf{M} > 0$ with marginals and bimarginals $(\mu_t)_{t\in\mathcal{V}}$ and $(R_{t_1,t_2})_{(t_1,t_2)\in\mathcal{E}}$ satisfying* (5b) *and* (5c)*, respectively, and that all functions $g^{(t)}$ and $f^{(t_1,t_2)}$ are such that either*

i) the function is a polyhedral indicator function and $\mu_t \in \mathrm{dom}(g^{(t)})$ or $R_{t_1,t_2} \in \mathrm{dom}(f^{(t_1,t_2)})$, respectively, or

ii) the function is co-finite, essentially smooth, continuous on the effective domain, and the gradient operator is strongly monotone and Lipschitz continuous on any compact convex subset of the interior of the effective domain, and so that $\mu_t \in \mathrm{int}(\mathrm{dom}(g^{(t)}))$ or $R_{t_1,t_2} \in \mathrm{int}(\mathrm{dom}(f^{(t_1,t_2)}))$, respectively.

Under these assumptions, let $(u_t^k)_{t\in\mathcal{V}}$ and $(U_{t_1,t_2}^k)_{(t_1,t_2)\in\mathcal{E}}$ be the iterates of Algorithm 1, and let $\mathbf{M}^k = \mathbf{K}\odot\mathbf{U}^k$. Then $\mathbf{M}^k \to \mathbf{M}^\star$ at least R-linearly, where $\mathbf{M}^\star$ is the unique optimal solution to (5)*, and the cost function in* (10)*, evaluated in $(u_t^k)_{t\in\mathcal{V}}$ and $(U_{t_1,t_2}^k)_{(t_1,t_2)\in\mathcal{E}}$, converges to the optimal value of* (5) *at least R-linearly.*

Proof: Assume first that all functions are as in ii). In this case, note that (5a) is separable in the different variables, and that E in [57, Eq. (1.1)] is of the form

$$E^T = \left[\begin{array}{ccc|ccc} \mathtt{P}_1^T & \dots & \mathtt{P}_{\mathcal{T}}^T & \mathtt{P}_{1,2}^T & \dots & \mathtt{P}_{\mathcal{T},\mathcal{T}-1}^T \\ \hline & -I & & & 0 & \\ \hline & 0 & & & -I & \end{array}\right]$$

where $\mathtt{P}_t$ is a matrix so that $\mathtt{P}_t\mathrm{vec}(\mathbf{M})$ is the projection on the tth marginal and $\mathtt{P}_{t_1,t_2}$ is a matrix such that $\mathtt{P}_{t_1,t_2}\mathrm{vec}(\mathbf{M})$ is the projection on the (t_1, t_2)-bimarginal. This means that $\mathtt{P}_t^T$ and $\mathtt{P}_{t_1,t_2}^T$ are the corresponding back-projections. Now, under the given assumptions the results in [57, Thm 6.1] are directly applicable.

In the case that some of the functions are of the form as in i), this cost function can be replaced by a finite number of inequality constraints. By adding the corresponding inequalities in the matrix E above, the above argument show R-linear convergence of the algorithm. ∎

Remark III.11. *One of the assumptions in Theorem III.10 is that all functions $g^{(t)}$ and $f^{(t_1,t_2)}$ (that are not polyhedral indicator functions) are such that they are differentiable on the interior of their effective domains. Under this assumption, all inclusions in* (14) *and* (15) *are in fact equalities on the interior of the effective domain.*

Example III.12. *Here, we consider a small bimarginal example to illustrate some of the differences between the results presented so far. To this end, let $\mathbf{M}, \mathbf{C} \in \mathbb{R}^{2\times 2}$, and consider the problem*

$$\underset{\mathbf{M}\in\mathbb{R}_+^{2\times 2}}{\text{minimize}}\ D(\mathbf{M}) \ \text{ subject to } \ P_1(\mathbf{M}) \le \begin{bmatrix}1\\2\end{bmatrix},\ P_{12}(\mathbf{M}) \ge \begin{bmatrix}1 & 0\\0 & 0\end{bmatrix},$$

where we for simplicity have taken $\mathbf{C} = 0$ and $\epsilon = 1$. The two constraints together imply that $\mathbf{M}_{12} = 0$ for any feasible solution, and hence neither the conditions in Assumption B nor the ones in Theorem III.10 are fulfilled. Nevertheless, the conditions in Assumption A are fulfilled, and hence the problem has a unique optimal solution (Lemma III.3). Moreover, the conditions in Assumption C are fulfilled, and hence each step in the algorithm is therefore well-defined (Lemma III.8). In fact, the conditions in Theorem III.9 are fulfilled, which guarantees that the dual ascent algorithm is converging to the optimal solution; the latter is given by

$$\mathbf{M}^\star = \begin{bmatrix}1 & 0\\1 & 1\end{bmatrix}.$$

Moreover, for suitable initial conditions the coordinate ascent method gives the iterates

$$u_1^k = \begin{bmatrix} 1/(\exp(k)+1) \\ 1 \end{bmatrix}, \qquad U_{1,2}^k = \begin{bmatrix} \exp(k) & 1 \\ 1 & 1 \end{bmatrix},$$

and the corresponding dual cost converges towards the optimal value as $k \to \infty$. However, the dual problem does not attain an optimal solution since $(U_{1,2}^k)_k$ diverges. Finally, by evaluating $\|\mathbf{M}^k - \mathbf{M}^\star\|_2$ it can be seen that in fact the iterates converge R-linearly, which indicates that there might be room for improvement with respect to the conditions in Theorem III.10.

As a final remark, note that Assumption B implicitly and Assumption C explicitly enforce that we must have $\mathbf{C} < \infty$. Similarly, the functions $g^{(t)}$ and $f^{(t_1,t_2)}$ must have effective domains that include marginals and bimarginals that are elementwise strictly positive, and hence they cannot, e.g., be indicator functions on singletons with zero elements. For some applications this is not fulfilled, and in particularly this is the case for the examples in Sections IV and V. Nevertheless, the assumptions can be weakened somewhat to accommodate for this, similar to [41, Sec. 4.1]. More specifically, if any element $\mathbf{C}_{i_1 \dots i_\mathcal{T}} = \infty$, then we can fix $\mathbf{M}_{i_1 \dots i_\mathcal{T}} = 0$ and remove it from the set of variables. This means that $\mathbf{M}$ is technically no longer a tensor, but the marginal and bimarginal projections can still be defined, and the above derivations carry over to this setting. Similarly, if $\operatorname{dom}(g^{(j)})$ is such that $\mu_j^{(i_j)} = 0$, then we can remove all the variables $\mathbf{M}_{i_1 \dots i_\mathcal{T}}$ with indices $\{(i_1, \dots, i_{j-1}, i_j, i_{j+1}, \dots, i_\mathcal{T}) \mid i_t = 1, \dots, N \text{ for } t \neq j\}$, and analogously for $f^{(j_1,j_2)}$ and the bimarginals. From the perspective of Algorithm 1, it is interesting to note that in the first case $\mathbf{K}_{i_1 \dots i_\mathcal{T}} = 0$, and in the second case we can take $u_j^{(i_j)} = 0$.

D. Extension to multiple costs on each marginal

In some problems, certain marginals or bimarginals are associated with cost functions for which there is no simple Fenchel conjugate. However, sometimes these functions can be split into simpler components. In particular, this is often the case when a marginal is associated with both a cost and has to satisfy an inequality constraint. To handle such cases, we consider a modified version of problem (5) that takes the form

$$\begin{aligned} \underset{\substack{\mathbf{M} \in \mathbb{R}_+^{N^\mathcal{T}},\ \mu_{t,k_1} \in \mathbb{R}_+^N, \\ R_{t_1,t_2,k_2} \in \mathbb{R}_+^{N \times N} \\ t \in \mathcal{V} \text{ and } k_1 = 1,\dots,\kappa_1 \\ (t_1,t_2) \in \mathcal{E} \text{ and } k_2 = 1,\dots,\kappa_2}}{\text{minimize}} \quad & \langle \mathbf{C}, \mathbf{M} \rangle + \epsilon D(\mathbf{M}) + \sum_{t \in \mathcal{V}} \sum_{k_1=1}^{\kappa_1} g_{k_1}^{(t)}(\mu_{t,k_1}) \\ & + \sum_{(t_1,t_2) \in \mathcal{E}} \sum_{k_2=1}^{\kappa_2} f_{k_2}^{(t_1,t_2)}(R_{t_1,t_2,k_2}) \qquad (16) \\ \text{subject to} \quad & P_t(\mathbf{M}) = \mu_{t,k_1},\ k_1 = 1, \dots, \kappa_1,\ t \in \mathcal{V} \\ & P_{t_1,t_2}(\mathbf{M}) = R_{t_1,t_2,k_2},\ k_2 = 1, \dots, \kappa_2, \\ & \qquad (t_1,t_2) \in \mathcal{E}. \end{aligned}$$

For ease of notation, we have the same number of functions κ_1 and κ_2 associated with each marginal and bimarginal, respectively, however this can easily be relaxed. Moreover, note that the constraints implicitly ensure that for any feasible point we have $\mu_{t,k_1} = \mu_{t,k_1'}$ for all $k_1, k_1' = 1, \dots, \kappa_1$ and all $t \in \mathcal{V}$, and similarly for the bimarginals. Next, by modifying the arguments in the previous sections it is straightforward to derive a Lagrangian dual of (16). In particular, similar to before, if the dual problem has an optimal solution then the optimal solution to (16) is of the form $\mathbf{M} = \mathbf{K} \odot \mathbf{U}$, where $\mathbf{K}$ is as before (see (7a)) and $\mathbf{U}$ has the form

$$\begin{aligned} \mathbf{U}_{i_1 \dots i_\mathcal{T}} &= \left(\prod_{t \in \mathcal{V}} \prod_{k_1=1}^{\kappa_1} u_{t,k_1}^{(i_t)} \right) \left(\prod_{(t_1,t_2) \in \mathcal{E}} \prod_{k_2=1}^{\kappa_2} U_{t_1,t_2,k_2}^{(i_{t_1},i_{t_2})} \right) \\ &= \left(\prod_{t \in \mathcal{V}} \prod_{k_1=1}^{\kappa_1} \exp\left(\lambda_{t,k_1}^{(i_t)}/\epsilon\right) \right) \left(\prod_{(t_1,t_2) \in \mathcal{E}} \prod_{k_2=1}^{\kappa_2} \exp\left(\Lambda_{t_1,t_2,k_2}^{(i_{t_1},i_{t_2})}/\epsilon\right) \right). \end{aligned}$$

Note that this structure is similar to the one in (7b). In fact, it can be interpreted as splitting u_t as $u_t = \odot_{k_1=1}^{\kappa_1} u_{t,k_1}$, and similarly $U_{t_1,t_2} = \odot_{k_2=1}^{\kappa_2} U_{t_1,t_2,k_2}$. Moreover, this means that the coordinate ascent inclusion for $u_{j,\tilde{k}_1}$ is given by

$$0 \in -u_{j,\tilde{k}_1} \odot \Big(\bigodot_{k_1 \neq \tilde{k}_1} u_{j,k_1} \Big) \odot w_j + \partial (g_{\tilde{k}_1}^{(j)})^* \big(-\epsilon \log(u_{j,\tilde{k}_1}) \big),$$

where w_j is defined analogously to (13a) as

$$w_j = P_j(\mathbf{K} \odot \mathbf{U}) \oslash \Big(\bigodot_{k_1=1}^{\kappa_1} u_{j,k_1} \Big) \qquad (17)$$

Similar expressions hold for the inclusion problem for U_{t_1,t_2,k_2}. Furthermore, reexamining the proof of Theorem III.9 and III.10, it can be readily seen that by modifying the assumptions accordingly, i.e., stating the assumptions in terms of each function $g_{k_1}^{(t)}$ and $f_{k_2}^{(t_1,t_2)}$, the results can be extended to this setting. For brevity, we omit explicitly stating these results.

Finally, by reexamining the argument of sequentially eliminating the modes of the tensor as in [29], [40], [41], one can see that the efficiency in computing w_j and W_{j_1,j_2} in (17) only depends on the underlying graph structure $(\mathcal{V}, \mathcal{E})$, and not on the number of cost functions associated with each marginal and bimarginal. Therefore, we can still efficiently solve the inclusions for "simple functions" and graph structures for which the projections can be easily computed.

IV. Convex dynamic network flow problems

Given a network $G = (V, E)$ with nodes V and edges E, a minimum-cost flow problem is to determine the cheapest way to transport a commodity from a set of sources $S^+ \subset V$ to a set of sinks $S^- \subset V$. In the simplest case, when the cost is specified as a linear cost c_e per unit for using edge $e \in E$ in the transportation, and the flow on each edge e is limited by the capacity $d_e \in \mathbb{R}_+$, the problem can be formulated as a linear programming problem [10]. However, in many cases the costs associated with using edges are in fact nonlinear functions of the flows, i.e., given by some $g_e : \mathbb{R}_+ \to \mathbb{R}$ for $e \in E$. Here, we will restrict our attention to the case when g_e is convex for all $e \in E$, cf. [10], [61], [63]. In a dynamic network flow problem, the travel times on the edges are also taken into account [3], [32], [35], [73]. These types of flow problems have many applications, for example in production planning, vehicle routing, and scheduling [25], [61], [66]. A common way to solve dynamic network flow problems is to formulate a corresponding static problem on the time-expanded network [35]. However, in many applications the time-expanded network becomes large and hence intractable to work with. Here, we will develop a (approximate) solution algorithm for convex dynamic network flow problems by casting it as a graph-structured tensor-optimization problem (4) and using the results in Section III. For results and algorithms corresponding to the linear case, see [41].

A. Solution method for convex dynamic network flow problems

On a network where the flow time on all edges are the same,[6] one way to formulate a dynamic network flow problem is as follows

[6]Note that this is a standard assumption, and can (at least approximately) be achieved by introducing intermediate edges and nodes.

[41]: consider the time points $t = 1, \dots, \mathcal{T}$, let $N_1 := |E|$ be the number of edges in the network, $N_2 := |S^+|$ be the number of sources, and $N_3 := |S^-|$ be the number of sinks. Furthermore, let $N := N_1 + N_2 + N_3$ and consider the tensor $\mathbf{M} \in \mathbb{R}_+^{N^{\mathcal{T}}}$. This means that the projection $P_t(\mathbf{M}) = \mu_t \in \mathbb{R}_+^N$ can be interpreted as the vector representing the flow at time t (N_1 elements), together with the amount that is left in the sources (N_2 elements) and the amount that has reached the sinks (N_3 elements) at that time point t, for $t = 1, \dots, \mathcal{T}$. Since the initial and final distribution of the commodity are known and fixed, the marginals μ_1 and $\mu_{\mathcal{T}}$ are given.

Next, note that the tensor $\mathbf{M}$ represents all flows which are possible from a combinatorial perspective. However, the set of feasible flows is much smaller. In particular, a flow needs to start in a source, end in a sink, and can only take a path which is feasible according to the network G. This flow feasibility is implicitly imposed using the cost tensor. To this end, let $C \in \bar{\mathbb{R}}_+^{N \times N}$ be a cost matrix that encodes the topology of the network G. This means that $C_{ik} = 0$ if state i connects to state k, i.e., 1) if the indices i and k represent edges connected via a node, 2) if i represents a source node and k an edge connected to it, or 3) if k represents a sink node and i an edge connected to it. Otherwise, $C_{ik} = \infty$. Expressed differently,

$$C_{ik} = \begin{cases} 0 & \text{if } i \in E \cup S^+ \text{ connects to } k \in E \cup S^+ \cup S^-, \\ 0 & \text{if } i \in E \cup S^+ \cup S^- \text{ connects to } k \in E \cup S^-, \\ \infty & \text{else.} \end{cases} \tag{18a}$$

Now, let

$$\mathbf{C}_{i_1 \dots i_{\mathcal{T}}} = \sum_{t=1}^{\mathcal{T}-1} C_{i_t i_{t+1}}. \tag{18b}$$

This means that

$$\langle \mathbf{C}, \mathbf{M} \rangle = \sum_{i_1, \dots, i_{\mathcal{T}}} \left(\sum_{t=1}^{\mathcal{T}-1} C_{i_t i_{t+1}} (P_{t,t+1}(\mathbf{M}))_{i_t, i_{t+1}} \right),$$

and hence $\langle \mathbf{C}, \mathbf{M} \rangle$ is finite if and only if $\mathbf{M}$ has support only on feasible flows. To see this, note that the bimarginal projection $P_{t,t+1}(\mathbf{M})$ describes how the flow at time t transitions to the flow at time $t+1$, i.e., the element $(P_{t,t+1}(\mathbf{M}))_{i_t, i_{t+1}}$ is the amount of flow that transitions from state i_t (i.e., the corresponding edge, source or sink) at time t to state i_{t+1} at time $t+1$. Since by construction we have that $C_{i_t, i_{t+1}}$ is ∞ for transitions which are not compatible with the network G, we have that $\langle \mathbf{C}, \mathbf{M} \rangle = \infty$ if and only if the flow described by $\mathbf{M}$ is not compatible with G. With this in mind, it is also clear that one can model linear costs associated with using edges, or for staying in sources or sinks, by changing the corresponding zero-values of the cost matrix in (18a) to non-zero values.

Finally, defining the convex function $g : \mathbb{R}_+^N \to \mathbb{R}$ by $g : x \mapsto \sum_{e \in E} g_e(x_e)$, where $g_e : \mathbb{R}_+ \to \mathbb{R}$ is the convex cost associate with edge $e \in E$, the convex dynamic network flow problem can be formulated as

$$\underset{\mathbf{M} \in \mathbb{R}_+^{N^{\mathcal{T}}}}{\text{minimize}} \;\; \langle \mathbf{C}, \mathbf{M} \rangle + \sum_{t=2}^{\mathcal{T}-1} g(P_t(\mathbf{M})) \tag{19a}$$

$$\text{subject to} \;\; P_1(\mathbf{M}) = \mu_1, \quad P_{\mathcal{T}}(\mathbf{M}) = \mu_{\mathcal{T}}. \tag{19b}$$

This is a problem of the form (4), with no costs imposed on the bimarginals and with underlying graph structure $\mathcal{G}$ given by a path-graph, except that it does not have an entropy term $\epsilon D(\mathbf{M})$. The latter can be introduced in order to derive efficient algorithms for approximately solving (19). However, it has also been shown that the introduction of the entropy term in (19) can be interpreted as finding robust transport plans [20]–[22]. In any case, the corresponding projections needed to solve the inclusion problems (15) can be efficiently computed [29, Prop. 2] (see also [40], [41]), and hence Algorithm 1 can be adapted to solve the entropy-regularized version of (19).

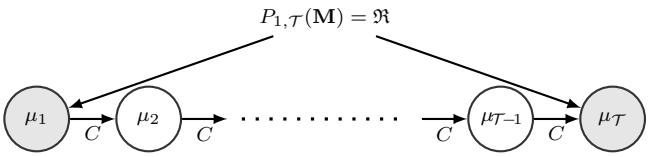


Fig. 1: Illustration of the graph $\mathcal{G}$ for the convex dynamic network flow problem with origin-destination constraint. Grey circles correspond to known densities, and white circles correspond to densities which are to be optimized over.

B. Convex dynamic network flow problems with origin-destination constraint

The above formulation (19) of a convex dynamic network flow problem finds an optimal way of steering the commodity from the sources to the sinks. However, the formulation only models aggregate distributions and flows, and does not take any individual behaviors into account. While this is adequate when all agents in the ensemble are indistinguishable, in some applications the latter is not true. One example is in traffic networks, where agents have specific destinations. One way to solve this is to consider multi-commodity flow problems [11], [41], [49]. However, following along the lines of [39], origin-destination constraints can also be included in this framework by simply changing the constraints in (19b) to a bimarginal constraint.

To this end, note that, as mentioned above, the bimarginal projection $P_{t_1,t_2}(\mathbf{M})$ describes how the flow at time t_1 transitions to the flow at time t_2. In particular, this means that $P_{1,\mathcal{T}}(\mathbf{M})$ describes how the initial flow distribution transitions to the final flow distribution. Introducing the origin-destination matrix $\mathfrak{R} \in \mathbb{R}_+^{N \times N}$ (cf. [76]) with elements

$$\mathfrak{R}_{ik} = \begin{cases} \xi_{ik} & \text{if } i \in S^+ \text{ and } k \in S^- \\ 0 & \text{else,} \end{cases}$$

by imposing the bimarginal constraint $P_{1,\mathcal{T}}(\mathbf{M}) = \mathfrak{R}$ the element $\xi_{ik} \geq 0$ represents how much of the commodity that starts in source $i \in S^+$ will go to sink $k \in S^-$. Therefore, an entropy-regularized convex dynamic network flow problem with origin-destination constraint matrix $\mathfrak{R}$ can be formulated as

$$\underset{\mathbf{M} \in \mathbb{R}_+^{N^{\mathcal{T}}}}{\text{minimize}} \;\; \langle \mathbf{C}, \mathbf{M} \rangle + \epsilon D(\mathbf{M}) + \sum_{t=2}^{\mathcal{T}-1} g(P_t(\mathbf{M})) \tag{20a}$$

$$\text{subject to} \;\; P_{1,\mathcal{T}}(\mathbf{M}) = \mathfrak{R}. \tag{20b}$$

The underlying graph structure $\mathcal{G}$ for (20) is no longer a path-graph, but contains a cycle. The latter is illustrated in Figure 1. Nevertheless, the projections needed to solve the inclusion problems (15) can be efficiently computed as follows.

Proposition IV.1 ([39, Thm. 2 and Cor. 1])**.** *For $\epsilon > 0$, let $\mathbf{K} = \exp(-\mathbf{C}/\epsilon)$ and let $K = \exp(-C/\epsilon)$, with $\mathbf{C}$ and C defined as in* (18)*, and let*

$$\mathbf{U}_{i_1 \dots i_{\mathcal{T}}} = U_{i_1, i_{\mathcal{T}}} \prod_{t=2}^{\mathcal{T}-1} u_t^{(i_t)}.$$

Define

$$\hat{\Psi}_j = \begin{cases} K, & j = 2, \\ \hat{\Psi}_{j-1} \mathrm{diag}(u_{j-1}) K, & j = 3, \dots, \mathcal{T}, \end{cases}$$

Algorithm 2 Method for solving the entropy-regularized convex dynamic network flow problem (20).

Input: Initial guess $u_2, \dots, u_{\mathcal{T}-1}$, U

$\Psi_{\mathcal{T}-1} \leftarrow K$
for $j = \mathcal{T}-2, \dots, 1$ **do**
  $\Psi_j \leftarrow K\mathrm{diag}(u_{j+1})\Psi_{j+1}$
end for
while Not converged **do**
  $U \leftarrow \mathfrak{R} \oslash \Psi_1$
  $\hat{\Psi}_2 \leftarrow K$
  for $j = 2, \dots, \mathcal{T}-1$ **do**
    $w_j \leftarrow (\hat{\Psi}_j^T \odot (\Psi_j U^T))\mathbf{1}$
    Update u_j by $0 = -u_j \odot w_j + \partial g^*(-\epsilon \log(u_j))$.
    $\hat{\Psi}_{j+1} \leftarrow \hat{\Psi}_j \mathrm{diag}(u_j) K$
  end for
  $\Psi_{\mathcal{T}-1} \leftarrow K$
  for $j = \mathcal{T}-2, \dots, 1$ **do**
    $\Psi_j \leftarrow K\mathrm{diag}(u_{j+1})\Psi_{j+1}$
  end for
end while

Output: $u_2, \dots, u_{\mathcal{T}-1}$, U

and

$$\Psi_j = \begin{cases} K, & j = \mathcal{T}-1, \\ K\mathrm{diag}(u_{j+1})\Psi_{j+1}, & j = 1, \dots, \mathcal{T}-2. \end{cases}$$

Then the projections of the tensor $\mathbf{K} \odot \mathbf{U}$ *are given by*

$$P_{1,\mathcal{T}}(\mathbf{K} \odot \mathbf{U}) = U \odot \Psi_1,$$
$$P_j(\mathbf{K} \odot \mathbf{U}) = u_j \odot \left(\hat{\Psi}_j^T \odot (\Psi_j U^T)\right)\mathbf{1},$$

for $j = 2, \dots, \mathcal{T}-1$.

Finally, Algorithm 1 can now be adapted to solve (20) efficiently; the latter is given in Algorithm 2. The origin-destination constraint is a first step towards introducing heterogeneity among the transported commodity. A larger degree of heterogeneity can be introduced by considering the multi-commodity case. The model and solution algorithm described above can be extended to this case, along the same lines as in Section V-C (cf. [41]).

C. Numerical examples

In this section we consider numerical examples of convex dynamic network flow problems with origin-destination constraint. To this end, consider the transportation network $G = (V, E)$ depicted in Figure 2a. This network has 150 undirected edges, and hence the corresponding directed network has $|E| = 2 \cdot 150 = 300$ edges. Moreover, the network has $|V| = 57$ nodes and we consider the flow over $\mathcal{T} - 1 = 29$ time steps. Next, on each edge we introduce a maximum capacity given by d_e which is equal to 20 times the length of the edge for normal edges, and 100 times the length of the edge for the thick edges, since the latter represent highways.

1) Performance comparison: In this numerical experiment, as cost on an edge $e \in E$ we consider the convex function $g_e : x_e \mapsto (x_e/d_e)^2 + I_{[0,d_e]}(x_e)$, i.e., a quadratic function normalized with the capacity on the edge and a hard constraint on not exceeding maximum capacity on the edge. We consider a problem with an origin-destination constraint, which can be modeled as a multi-commodity flow problem when formulated on a time-expanded network. More precisely, for each commodity, there is one source node in which we let $10|V| = 570$ units start, and 10 units of the commodity are sent to each node in the network. No two commodities share the same source node, and we solve the problem for $1, 2, \dots, 57$ commodities. This can be formulated as an instance of the problem (20), where the origin-destination matrix $\mathfrak{R}$ has elements $\xi_{ik} = 10$ for all $i \in S^+ \subseteq V$ and all $k \in S^- = V$. Finally, we also introduce an incentive for goods to arrive early by in the cost matrix in (18a) setting $C_{ii} = -0.01$ for $i \in S^-$; this can be handled in a formulation on a time-expanded network by introducing extra nodes and edges, where the edges are associated with a linear cost of -0.01.

TABLE I: List of some functions together with their Fenchel conjugates. In particular, for the p-norm, $p \in (1, \infty)$ and $1/p + 1/q = 1$.

Function $f(x)$	Fenchel conjugate $f^*(x^*)$
0	$I_{\{0\}}(x^*)$
$I_{\{\hat{x}\}}(x)$	$\langle x^*, \hat{x} \rangle$
$I_{[\alpha,\beta]}(x)$	$\sum_{i=1}^{n} \left(x^*_{(i)} \beta_i I_{\mathbb{R}_+}(x^*_{(i)}) + x^*_{(i)} \alpha_i I_{\mathbb{R}_-}(x^*_{(i)}) \right)$
$\sigma \lVert x - y \rVert_p^p$	$\langle x^*, y \rangle + \frac{1}{q\, \sigma^{q-1}\, p^{q-1}} \lVert x^* \rVert_q^q$
$\frac{x}{\beta - x} + I_{[0,\beta]}(x)$	$\begin{cases} 0 & \text{if } x^* \le 1/\beta \\ x^*\beta - 2\sqrt{x^*\beta} + 1 & \text{if } x^* \ge 1/\beta \end{cases}$

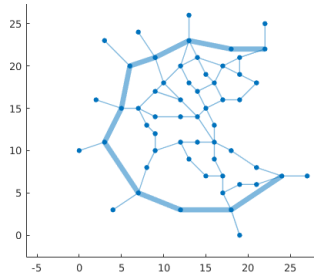
We use Algorithm 2, where the latter is adapted as in Section III-D to handle both the quadratic cost and the capacity constraint; for details on the corresponding Fenchel conjugates, see Table I. The solution times for the proposed method is compared with the solution times obtained when solving the problem on a time-expanded network using the commercial solver Gurobi [38]. Solution times for both methods, with a varying number of commodities, are shown in Figure 2b. As can be seen in the figure, our method outperforms Gurobi as the number of commodities increases.[7]

2) Illustration with non-quadratic cost: Next, we illustrate a solution on the same transportation network G but for a problem with non-quadratic cost on the edges. More precisely, as cost on each edge $e \in E$, we consider the convex function

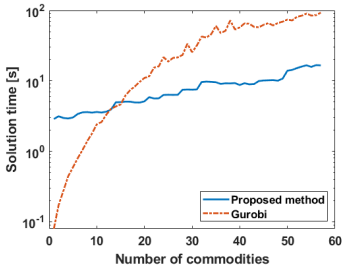
$$g_e : x_e \mapsto \frac{x_e}{d_e - x_e} + I_{[0,d_e]}(x_e). \tag{21}$$

This is used to minimize the total congestion in the network, cf. [63]. We use a similar setup as in the previous example: the origin-destination matrix $\mathfrak{R}$ is taken to have nonzero elements $\xi_{ik} = 10$ for all $i \in S^+ = V$ and all $k \in S^- = V$, which means that $10|V| = 570$ units start in each node, and that each node receives $10|V| = 570$ units; for the cost matrix in (18a) we set $C_{ii} = -0.01$ for $i \in S^-$ as an incentive for goods to arrive early. The problem is solved for $\epsilon = 10^{-2}$, using Algorithm 2. For details on the Fenchel conjugate of (21), see Table I. The flow at a number of different time points is illustrated in Figure 2c. In particular, in Figure 2c the width of each edge is proportional to the logarithm of $1+$ the flow on that edge at that time point. Moreover, the maximum capacity utilization, i.e., the value $\max_{e \in E}\{x_e/d_e\}$, varies between 0.8106 and 0.8339 for time points $t = 2, \dots, 29$.

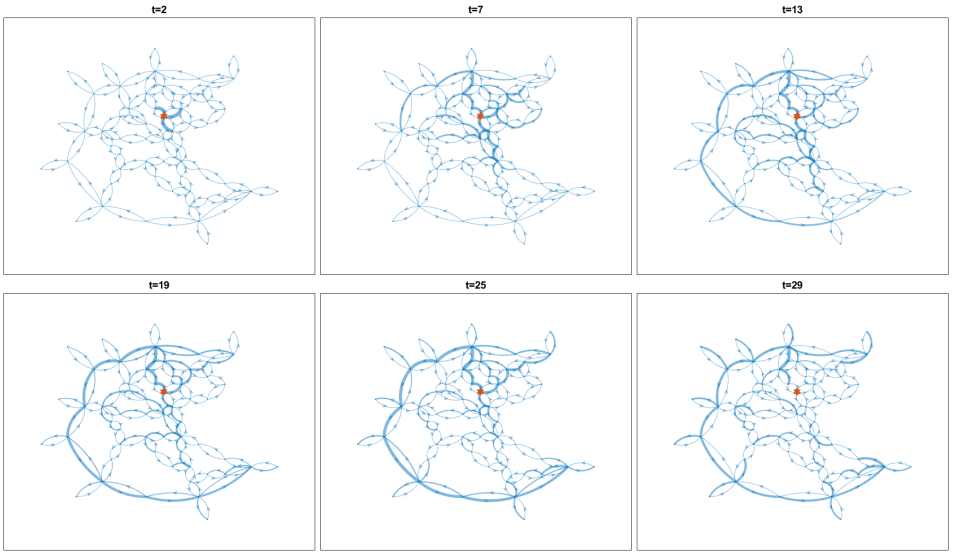
[7]Simulations are run on a Dell OptiPlex7080 with Intel(R) Core(TM) i7-10700 CPU and 32GB of RAM. Moreover, in the comparison it should also be noted that the proposed method is implemented in Matlab, while the Gurobi back-end is implemented in C.



(a) Flow network G.



(b) Solution times for the proposed method and for Gurobi, when solving the problem in Section IV-C1.



(c) An illustration of the optimal flow, using the setup in Section IV-C2, from one of the sources at a number of different time points. The width of each edge is proportional to the logarithm of $1+$ the flow on that edge at the given time point.

Fig. 2: Figures for the numerical examples in Section IV-C.

V. Multi-species potential mean field games

An important tool for analyzing and controlling systems of systems, which has emerged during the last decades, is mean field games [12], [28], [45]–[47], [52]. Mean field games are models of dynamic games where each player's action is negligible to other players at the individual level, but where the actions are significant when aggregated. A subclass of such games are potential mean field games. These can be seen as density control problems, where the density abides to a controlled Fokker-Planck equation with distributed control [52]. This type of control problems have been studied in, e.g., [8], [14], [18]. An important generalization of mean field games is the multi-species setting, where the population consists of several different types of agents or species [1], [9], [24], [46], [50], [52]. In this section, we show that discretizations of potential multi-species mean field games take the form of a convex graph-structured tensor optimization problem (5). By also deriving efficient methods for computing the corresponding projections needed in Algorithm 1, we here develop an efficient numerical solution algorithm for solving such problems. In order to do so, we will first consider the nonlinear density control problem obtained in the single-species setting, and its corresponding discretization.

A. The single-species problem

Let $X \subset \mathbb{R}^n$ be a state space, and consider a set of infinitesimal agents on X which obeys the (Itô) stochastic differential equation

$$dx(t) = f(x(t))dt + B(x(t))\big(v(x(t),t)dt + \sqrt{\epsilon}dw\big), \qquad (22)$$

subject to the initial condition $x(0) = x_0 \sim \rho_0(x)$. More precisely, we assume that $f : X \to \mathbb{R}^n$ and $B : X \to \mathbb{R}^{n\times n}$ are continuously differentiable with bounded derivatives, in which case, under suitable conditions on the (Markovian) feedback v, there exists a unique solution to (22) a.s., see, e.g, [33, Thm. V.4.1], [8, pp. 7-8]. Moreover, under suitable regularity conditions [8], [14] the density $\rho(t,\cdot)$ which describe the distribution of particles at time point t exists and is the solution of a controlled Fokker-Planck equation (cf. [67, p. 72]). A potential mean field game can then be reformulated as the density optimal control problem [52]

$$\underset{\rho,v}{\text{minimize}} \quad \int_0^1\int_X \frac{1}{2}\|v\|^2\rho dx dt + \int_0^1 \mathcal{F}_t(\rho(t,\cdot))dt + \mathcal{G}(\rho(1,\cdot)) \qquad (23a)$$

$$\text{subject to} \quad \frac{\partial\rho}{\partial t} + \nabla\cdot((f+Bv)\rho) - \frac{\epsilon}{2}\sum_{i,k=1}^{n}\frac{\partial^2(\sigma_{ik}\rho)}{\partial x_i \partial x_k} = 0 \qquad (23b)$$

$$\rho(0,\cdot) = \rho_0. \qquad (23c)$$

Here, $\sigma(x) := B(x)B(x)^T$. Moreover, $\mathcal{F}_t$ and $\mathcal{G}$ are functionals on $L_2 \cap L_\infty$, and we assume that they are proper, convex, and lower-semicontinuous. We also assume that $\mathcal{F}_t$ is piece-wise continuous with respect to t.

To discretize problem (23), we rewrite it as a problem over path space. To this end, let $\mathcal{P}^v$ denote the distribution on path space, i.e., a probability distribution over $C([0,1],X)$:= the set of continuous functions from $[0,1]$ to X, induced by the controlled process (22). In particular, this means that for the marginal of $\mathcal{P}^v$ corresponding to time t, denoted $\mathcal{P}^v_t$, we have that $\mathcal{P}^v_t = \rho(t,\cdot)$, where ρ is the solution to (23b) and (23c). Moreover, let $\mathcal{P}^0$ denote the corresponding (uncontrolled) Wiener process with initial density ρ_0. By the Girsanov theorem (see, e.g., [34, pp. 156-157], [27, p. 321]), we have that

$$\frac{1}{2}\int_X\int_0^1 \|v\|^2\rho dt dx = \frac{1}{2}\mathbb{E}_{\mathcal{P}^v}\left\{\int_0^1 \|v\|^2 dt\right\} = \epsilon \mathrm{KL}(\mathcal{P}^v\|\mathcal{P}^0) \qquad (24)$$

where $\mathrm{KL}(\cdot\|\cdot)$ is the Kullback-Leibler divergence, see, e.g., [7], [15], [18], [37], [53], [54]. To ensure that (24) holds, it is important that the control signal and the noise enter the system through the same channel, as in (22) [16], [17]. Moreover, the link between stochastic control and entropy provided by (24) has recently led to several novel applications of optimal control [13], [15]–[17], [19].

By using (24), the problem (23) can be reformulated as

$$\begin{aligned} \underset{\mathcal{P}^v}{\text{minimize}} \quad & \epsilon\,\mathrm{KL}(\mathcal{P}^v\|\mathcal{P}^0) + \int_0^1 \mathcal{F}_t(\mathcal{P}^v_t)dt + \mathcal{G}(\mathcal{P}^v_1) \\ \text{subject to} \quad & \mathcal{P}^v_0 = \rho_0. \end{aligned}$$

Next, we discretize this problem in both time and space. More precisely, discretizing over time into the time points $0, \Delta t, 2\Delta t, \dots, 1$, where $\Delta t = 1/\mathcal{T}$, and over space into the grid points $x_1, \dots, x_N$, the problem becomes

$$\begin{aligned} \underset{\substack{\mathbf{M}\in\mathbb{R}_+^{N^{\mathcal{T}+1}} \\ \mu_1,\dots,\mu_{\mathcal{T}}\in\mathbb{R}_+^N}}{\text{minimize}} \quad & \langle \mathbf{C}, \mathbf{M}\rangle + \epsilon D(\mathbf{M}) + \Delta t \sum_{j=1}^{\mathcal{T}-1} F_j(\mu_j) + G(\mu_{\mathcal{T}}) && \text{(25a)} \\ \text{subject to} \quad & P_j(\mathbf{M}) = \mu_j,\ j = 1, 2, \dots, \mathcal{T}, && \text{(25b)} \\ & P_0(\mathbf{M}) = \mu_0. && \text{(25c)} \end{aligned}$$

Here, $\mathbf{M}$ is a nonnegative $(\mathcal{T}+1)$-mode tensor that represents the flow of the agents, μ_0 is a discrete approximation of ρ_0, and μ_j is the distribution of agents at time point j. Moreover, $\mathbf{C}$ is a $(\mathcal{T}+1)$-mode tensor that represents the cost of moving agents. Since the cost of moving agents depends only on the current time step, the cost tensor takes the form

$$\mathbf{C}_{i_0,\dots,i_{\mathcal{T}}} = \sum_{j=0}^{\mathcal{T}-1} C_{i_j, i_{j+1}}, \tag{26a}$$

where C is a $N \times N$ matrix defining the transition costs. More precisely, the elements C_{ik} are the (optimal) cost of moving mass from discretization point x_i to discretization point x_k in one time step, given by

$$C_{ik} \;=\; \begin{cases} \underset{v\in L_2([0,\Delta t])}{\text{minimize}} & \displaystyle\int_0^{\Delta t} \frac{1}{2}\|v\|^2 dt \\ \text{subject to} & \dot{x} = f(x) + B(x)v \\ & x(0) = x_i, \quad x(\Delta t) = x_k. \end{cases} \tag{26b}$$

The optimal control problem (26b) can typically not be solved analytically, except in the linear-quadratic case. Nevertheless, a numerical solution to the problem suffices, and the computation of the cost function C can be done off-line before solving (25). In order to guarantee that the elements (26b) are all finite, we typically impose the following assumption: *that the deterministic counterpart to system* (22) *is controllable in the (rather strong) sense that for all* $x_0, x_1 \in X$ *and for all* $t > 0$ *there exists a control signal in* $L_2([0,t])$ *that transitions the system from the initial state* $x(0) = x_0$ *to the final state* $x(t) = x_1$. By allowing the matrix C to have elements with value ∞, this assumption may be relaxed. However, in this case one has to assure that (25) has a feasible solution with finite object function value, i.e., that there is an $\mathbf{M} \in \mathbb{R}_+^{N^{\mathcal{T}+1}}$ that fulfills (25b) and (25c) and is such that (25a) is finite (cf. Assumption A).

Finally, note that the problem (25)–(26) is a convex graph-structured tensor optimization problem of the form (5) on a path-graph.

Remark V.1. *Another solution method for solving problems of the form* (25)*, for agents that follow the dynamics of a first-order integrator, has been presented in [7]. The two methods are similar, and the main difference is that the computational method developed in [7] is based on a variable elimination technique, in contrast to the belief-propagation-type technique used here; see the discussion just before Theorem III.9.*

B. The multi-species problem

A multi-species mean field game is an extension of mean field games to a set of heterogeneous agents, and the idea was already presented in the seminal work [46], [52]. Here, we consider a multi-species potential mean field game which has L different populations, each of which can be associated with different costs and constraints, and where each infinitesimal agent of species ℓ obeys the dynamics

$$dx_\ell(t) = f(x_\ell)dt + B(x_\ell)\big(v_\ell dt + \sqrt{\epsilon} dw_\ell\big).$$

Next, let $\rho_\ell(t,\cdot)$ denote the distribution of species ℓ at time point t, and note that a multi-species potential mean field game can, analogously to the single species game, be formulated as an optimal control problem over densities. More precisely, the problem of interest here takes the form

$$\begin{aligned} \underset{\rho,\rho_\ell,v_\ell}{\text{minimize}} \quad & \int_0^1 \int_X \sum_{\ell=1}^L \frac{1}{2}\|v_\ell\|^2 \rho_\ell \, dx dt + \int_0^1 \mathcal{F}_t(\rho(t,\cdot))dt + \mathcal{G}(\rho(1,\cdot)) \\ & + \sum_{\ell=1}^L \left(\int_0^1 \mathcal{F}_t^\ell(\rho_\ell(t,\cdot))dt + \mathcal{G}^\ell(\rho_\ell(1,\cdot)) \right) && \text{(27a)} \\ \text{subject to} \quad & \frac{\partial \rho_\ell}{\partial t} + \nabla \cdot ((f(x) + B(x)v_\ell)\rho_\ell) \\ & - \frac{\epsilon}{2} \sum_{i,k=1}^n \frac{\partial^2(\sigma_{ik}\rho_\ell)}{\partial x_i \partial x_k} = 0, \quad \ell = 1, \dots L, && \text{(27b)} \\ & \rho_\ell(0,\cdot) = \rho_{0,\ell}, \quad \rho(t,x) = \sum_{\ell=1}^L \rho_\ell(t,x), && \text{(27c)} \end{aligned}$$

where we impose the same assumptions on $\mathcal{F}_t^\ell$ and $\mathcal{G}^\ell$ as on $\mathcal{F}_t$ and $\mathcal{G}$, respectively. The functionals $\int_0^1 \mathcal{F}_t(\cdot)dt$ and $\mathcal{G}(\cdot)$ are the cooperative part of the cost, which connects the different species. In particular, for $\mathcal{F}_t \equiv 0$, $\mathcal{G} \equiv 0$, (27) reduces to L independent single-species problems. Moreover, the functionals $\int_0^1 \mathcal{F}_t^\ell(\cdot)dt$ and $\mathcal{G}^\ell(\cdot)$ are the ones that give rise to the heterogeneity among the species.

C. Numerical algorithm for solving the multi-species problem

To derive a numerical algorithm for solving (27), analogously to the single-species problem we first discretize the problem over time and space. To this end, by adapting the arguments in the previous section, we arrive at the discrete problem

$$\begin{aligned} \underset{\substack{\mathbf{M}_\ell,\mu_j,\mu_j^{(\ell)} \\ j=1,\dots,\mathcal{T} \\ \ell=1,\dots,L}}{\text{minimize}} \quad & \sum_{\ell=1}^L \big(\langle \mathbf{C}, \mathbf{M}_\ell\rangle + \epsilon D(\mathbf{M}_\ell)\big) + \Delta t \sum_{j=1}^{\mathcal{T}-1} F_j(\mu_j) + G(\mu_{\mathcal{T}}) \\ & + \sum_{\ell=1}^L \left(\Delta t \sum_{j=1}^{\mathcal{T}-1} F_j^\ell(\mu_j^{(\ell)}) + G^\ell(\mu_{\mathcal{T}}^{(\ell)}) \right) && \text{(28a)} \\ \text{subject to} \quad & P_j(\mathbf{M}_\ell) = \mu_j^{(\ell)},\ j = 1, \dots, \mathcal{T},\ \ell = 1, \dots, L, && \text{(28b)} \\ & P_0(\mathbf{M}_\ell) = \mu_{0,\ell},\ \ell = 1, \dots, L, && \text{(28c)} \\ & \sum_{\ell=1}^L \mu_j^{(\ell)} = \mu_j,\ j = 0, \dots, \mathcal{T} && \text{(28d)} \end{aligned}$$

where $\mathbf{C}$ still has the form (26), and where $\mu_{0,\ell}$ are discrete approximations of $\rho_{0,\ell}$. In particular, note that the second line in the cost (28a) is the discretization of the second line in (27a). Moreover, also note that (28) consists of L coupled multi-marginal optimal transport problems, coupled via the constraint (28d) and the cost imposed on μ_j, for $j = 1, \dots, \mathcal{T}$, in (28a).

Next, we reformulate (28) into one single entropy-regularized multi-marginal transport problem (cf. [41]). To this end, let $\mathbf{M} \in \mathbb{R}^{L \times N^{\mathcal{T}+1}}$ be the $(\mathcal{T}+2)$-mode tensor such that $\mathbf{M}_{\ell i_0 \dots i_{\mathcal{T}}} = (\mathbf{M}_\ell)_{i_0 \dots i_{\mathcal{T}}}$, i.e., $\mathbf{M}_{\ell i_0 \dots i_{\mathcal{T}}}$ is the amount of mass of species ℓ that moves along the path $x_{i_0}, \dots, x_{i_{\mathcal{T}}}$. For this tensor $\mathbf{M}$, we will use the index -1 to denote the "species index". This means that $(P_{-1}(\mathbf{M}))_\ell = \sum_{i_0,\dots,i_{\mathcal{T}}} (\mathbf{M}_\ell)_{i_0 \dots i_{\mathcal{T}}}$, for $\ell = 1, \dots, L$, and hence the elements of the additional marginal $\mu_{-1} \in \mathbb{R}_+^L$ are the total mass of the densities of the different species. Moreover, this means that $P_j(\mathbf{M})$ is the total distribution μ_j at time $j\Delta t$, as defined by (28d), while the bimarginal projection $P_{-1,j}(\mathbf{M})$ gives the $L \times N$ matrix $[\mu_j^{(1)}, \dots, \mu_j^{(L)}]^T$. By introducing the matrix

$$\mathfrak{R}^{(-1,0)} = [\mu_{0,1}, \dots, \mu_{0,L}]^T \in \mathbb{R}_+^{L \times N},$$

the constraint (28c) can be imposed by requiring that $P_{-1,0}(\mathbf{M}) = \mathfrak{R}^{(-1,0)}$. Next, by defining the functions $\mathscr{F}_j^L : \mathbb{R}^{L\times N} \to \mathbb{R}$ as

$$\mathscr{F}_j^L(R^{(-1,j)}) = \sum_{\ell=1}^{L} \Delta t F_j^\ell(\mu_j^{(\ell)}) \quad j = 1,\dots,\mathcal{T},$$

and similarly for $\mathscr{G}^L$, the last term in the cost (28a) can be written as functionals applied to the bimarginal projections. Finally, by noting that $\sum_{\ell=1}^{L} D(\mathbf{M}_\ell) = D(\mathbf{M})$, we can write the problem as

$$\begin{aligned}
\underset{\substack{\mathbf{M},\mu_j,R^{(-1,j)}\\ j=1,\dots,\mathcal{T}}}{\text{minimize}} \quad & \langle \tilde{\mathbf{C}}, \mathbf{M}\rangle + \epsilon D(\mathbf{M}) + \Delta t \sum_{j=1}^{\mathcal{T}-1} F_j(\mu_j) + G(\mu_\mathcal{T}) \\
& + \sum_{j=1}^{\mathcal{T}-1} \mathscr{F}_j^L(R^{(-1,j)}) + \mathscr{G}^L(R^{(-1,\mathcal{T})}) && \text{(29a)}\\
\text{subject to} \quad & P_j(\mathbf{M}) = \mu_j, \quad j = 1,\dots,\mathcal{T}, && \text{(29b)}\\
& P_{-1,j}(\mathbf{M}) = R^{(-1,j)}, \quad j = 1,\dots,\mathcal{T}, && \text{(29c)}\\
& P_{-1,0}(\mathbf{M}) = \mathfrak{R}^{(-1,0)} && \text{(29d)}
\end{aligned}$$

where

$$\tilde{\mathbf{C}}_{\ell i_0 \dots i_\mathcal{T}} = \sum_{j=0}^{\mathcal{T}-1} C_{i_j, i_{j+1}}. \quad \text{(29e)}$$

The problem (29) is readily seen to be a graph-structured entropy-regularized multi-marginal optimal transport problem of the form (5), and can hence be solved using Algorithm 1. In particular, the iterates of the transport plan produced by Algorithm 1 are of the form $\mathbf{M}^k = \mathbf{K} \odot \mathbf{U}^k$, where $\mathbf{K} = \exp(-\tilde{\mathbf{C}}/\epsilon)$ and where

$$\mathbf{U}_{\ell i_0 \dots i_\mathcal{T}} = U_{-1,0}^{(\ell,i_0)} \prod_{j=1}^{\mathcal{T}} U_{-1,j}^{(\ell,i_j)} \prod_{j=1}^{\mathcal{T}} u_j^{(i_j)}. \quad \text{(30)}$$

The underlying graph-structure is illustrated in Figure 3, and by adapting the arguments in [41], marginal and bimarginal projections needed in the inclusion problems (15) can be computed efficiently as follows.

Theorem V.2. *Let $\mathbf{K} = \exp(-\tilde{\mathbf{C}}/\epsilon)$, with $\tilde{\mathbf{C}}$ defined as in* (29e) *and $\epsilon > 0$, and let $\mathbf{U}$ be as in* (30)*. Define $K = \exp(-C/\epsilon)$, and let*

$$\hat{\Psi}_j = \begin{cases} U_{-1,0}K, & j = 1,\\ \left(\hat{\Psi}_{j-1} \odot U_{-1,j-1}\right) \operatorname{diag}(u_{j-1})K, & j = 2,\dots,\mathcal{T}, \end{cases}$$

and

$$\Psi_j = \begin{cases} U_{-1,\mathcal{T}} \operatorname{diag}(u_\mathcal{T})K^T, & j = \mathcal{T}-1,\\ (\Psi_{j+1} \odot U_{-1,j+1}) \operatorname{diag}(u_{j+1})K^T, & j = 0,\dots,\mathcal{T}-2. \end{cases}$$

Then we have the following expressions for projections of the tensor $\mathbf{K} \odot \mathbf{U}$

$$\begin{aligned}
&P_{-1,0}(\mathbf{K}\odot\mathbf{U}) = U_{-1,0} \odot \Psi_0,\\
&P_{-1,j}(\mathbf{K}\odot\mathbf{U}) = \hat{\Psi}_j \odot \Psi_j \odot U_{-1,j}\operatorname{diag}(u_j),\\
&P_{-1,\mathcal{T}}(\mathbf{K}\odot\mathbf{U}) = U_{-1,\mathcal{T}}\operatorname{diag}(u_\mathcal{T}) \odot \hat{\Psi}_\mathcal{T},\\
&P_\mathcal{T}(\mathbf{K}\odot\mathbf{U}) = u_\mathcal{T} \odot \left(\hat{\Psi}_\mathcal{T} \odot U_{-1,\mathcal{T}}\right)^T \mathbf{1},\\
&P_j(\mathbf{K}\odot\mathbf{U}) = u_j \odot \left(\hat{\Psi}_j \odot \Psi_j \odot U_{-1,j}\right)^T \mathbf{1},
\end{aligned}$$

for $j = 1,\dots,\mathcal{T}-1$.

Proof: See Appendix A. ∎

Finally, using Theorem V.2 and specializing Algorithm 1 to solving the particular problem (29), an algorithm for solving discretized multi-species potential mean field games is given in Algorithm 3.

Remark V.3. *The algorithms in [68] are special instances of Algorithm 3. In particular, if $\mathscr{F}_j^L(\cdot) = \langle C_j, \cdot\rangle$ for some $C_j \in \mathbb{R}^{L\times N}$, then $(\mathscr{F}_j^L)^*(\cdot) = I_{\{C_j\}}(\cdot)$. Hence, $U_{-1,j}$ must equal $K_j := \exp(-C_j/\epsilon)$. Similarly, if $\mathscr{G}^L(\cdot) = \langle C_\mathcal{T}, \cdot\rangle$, we get that $U_{-j,\mathcal{T}}$ must be equal to $K_\mathcal{T}$, from which we recover [68, Alg. 1]. On the other hand, if $\mathscr{G}^L(\cdot) = I_{\{\mathfrak{R}^{(-1,\mathcal{T})}\}}(\cdot)$ for some given $\mathfrak{R}^{(-1,\mathcal{T})}$, then the marginal $\mu_\mathcal{T}$ is also known and any cost associated with it is a constant and can hence be removed. Moreover, $(\mathscr{G}^L)^*(\cdot) = \langle \mathfrak{R}^{(-1,\mathcal{T})}, \cdot\rangle$, from which we recover [68, Alg. 2].*

Algorithm 3 Method for solving the multi-species potential mean field game (29).

Input: Initial guess $u_1,\dots,u_\mathcal{T}$, $U_{-1,0},\dots,U_{-1,\mathcal{T}}$
$\Psi_{\mathcal{T}-1} \leftarrow U_{-1,\mathcal{T}} \operatorname{diag}(u_\mathcal{T})K^T$
for $j = \mathcal{T}-2,\dots,0$ **do**
  $\Psi_j \leftarrow (\Psi_{j+1} \odot U_{-1,j+1})\operatorname{diag}(u_{j+1})K^T$
end for
while Not converged **do**
  $U_{-1,0} \leftarrow \mathfrak{R}^{(-1,0)} \oslash \Psi_0$
  $\hat{\Psi}_1 \leftarrow U_{-1,0}K$
  for $j = 1,\dots,\mathcal{T}-1$ **do**
    $W_{-1,j} \leftarrow (\hat{\Psi}_j \odot \Psi_j)\operatorname{diag}(u_j)$
    Update $U_{-1,j}$ by $0 = -U_{-1,j} \odot W_{-1,j} + \partial(\mathscr{F}_j^L)^*(-\epsilon\log(U_{-1,j}))$.
    $w_j \leftarrow (\hat{\Psi}_j \odot \Psi_j \odot U_{-1,j})^T\mathbf{1}$
    Update u_j by $0 = -u_j \odot w_j + \partial(\Delta t F_j)^*(-\epsilon\log(u_j))$.
    $\hat{\Psi}_{j+1} \leftarrow (\hat{\Psi}_j \odot U_{-1,j})\operatorname{diag}(u_j)K$
  end for
  $W_{-1,\mathcal{T}} \leftarrow \hat{\Psi}_\mathcal{T}\operatorname{diag}(u_\mathcal{T})$
  Update $U_{-1,\mathcal{T}}$ by $0 = -U_{-1,\mathcal{T}} \odot W_{-1,\mathcal{T}} + \partial(\mathscr{G}^L)^*(-\epsilon\log(U_{-1,\mathcal{T}}))$.
  $w_\mathcal{T} \leftarrow (\hat{\Psi}_\mathcal{T} \odot U_{-1,\mathcal{T}})^T\mathbf{1}$
  Update $u_\mathcal{T}$ by $0 = -u_\mathcal{T} \odot w_\mathcal{T} + \partial G^*(-\epsilon\log(u_\mathcal{T}))$.
  $\Psi_{\mathcal{T}-1} \leftarrow U_{-1,\mathcal{T}} \operatorname{diag}(u_\mathcal{T})K^T$
  for $j = \mathcal{T}-1,\dots,1$ **do**
    $\Psi_{j-1} \leftarrow (\Psi_j \odot U_{-1,j})\operatorname{diag}(u_j)K^T$
  end for
end while
Output: $u_1,\dots,u_\mathcal{T}$, $U_{-1,0},\dots,U_{-1,\mathcal{T}}$

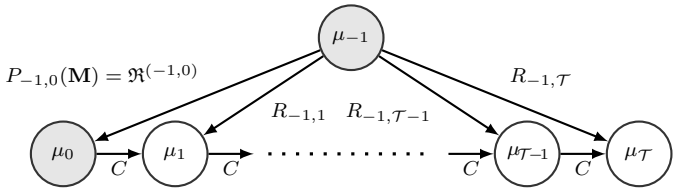


Fig. 3: Illustration of the graph $\mathcal{G}$ for the multi-species density optimal control problem. Grey circles correspond to known densities, and white circles correspond to densities which are to be optimized over.

D. Numerical example

In this section we demonstrate Algorithm 3 on a two-dimensional numerical example with $L = 4$ different species. To this end, we consider the state space $[0,3] \times [0,3]$ and uniformly discretize it into 100×100 grid points; the latter are denoted $x_{i,k}$ for $i,k = 1,\dots,100$. No points are placed on the boundary of the state space, which means that the cell size is $\Delta x = 0.03^2$. Moreover, time is discretized into $\mathcal{T}+1 = 40$ time steps, i.e., with a discretization size $\Delta t = 1/39$ and with time index $j = 0,\dots,39$. The dynamics of

the agents is taken to be $f(x) \equiv 0$ and $B(x) = I$. This means that the cost matrix, with elements (26b), is time-independent and given by $C = [\|x_{i_1,k_1} - x_{i_2,k_2}\|^2]_{i_1,i_2,k_1,k_2=1}^{100}$. This corresponds to the squared Wasserstein-2 distance on the discrete grid.

For $\epsilon = 10^{-2}$, we consider the discrete problem

$$\begin{aligned}
\underset{\substack{\mathbf{M}_\ell \in \mathbb{R}_+^{100^{40}}, \\ \mu_j^{(\ell)} \in \mathbb{R}_+^{100} \\ j=1,\dots,39,\ \ell=1,2,3,4}}{\text{minimize}} \quad & \sum_{\ell=1}^{4} \Big(\langle \mathbf{C}, \mathbf{M}_\ell \rangle + \epsilon D(\mathbf{M}_\ell) \Big) \\
& + \sum_{j=1}^{39} \langle c_3, \mu_j^{(3)} \rangle + 0.1 \sum_{j=1}^{39} \|\mu_j^{(4)} - \tilde{\mu}^{(4)}\|_2^2 \\
& + 3\|\mu_{19} - \tilde{\mu}_1\|_2^2 + 3\|\mu_{39} - \tilde{\mu}_2\|_2^2 && \text{(31a)} \\
\text{subject to} \quad & P_j(\mathbf{M}_\ell) = \mu_j^{(\ell)}, \quad j = 0, \dots, 39, \\
& \qquad\qquad \ell = 1, 2, 3, 4, && \text{(31b)} \\
& \sum_{\ell=1}^{4} \mu_j^{(\ell)} = \mu_j, \quad j = 0, \dots, 39, && \text{(31c)} \\
& \mu_j \le \kappa_j, \quad j = 1, \dots, 39, && \text{(31d)} \\
& \mu_j^{(1)} \le \kappa^{(1)}, \quad j = 1, \dots, 39. && \text{(31e)}
\end{aligned}$$

Here, $\tilde{\mu}_1$ and $\tilde{\mu}_2$ are the two distributions given in Figure 4a. Moreover, the linear cost c_3, associated with species 3, and the target distribution $\tilde{\mu}^{(4)}$, associated with species 4, are both given in Figure 4b.[8] Finally, for the capacity constraint (31d), κ_j is illustrated in Figure 4c, while for the capacity constraint (31e), $\kappa^{(1)}$ is zero in the lower half of the domain and infinite for the upper half.

The multi-marginal optimal transport reformulation of (31) was solved using Algorithm 3. The latter is adapted as in Section III-D to handle both the costs on the total marginals in (31a) and the inequality constraints in (31d); details on the Fenchel conjugates of the functions involved can be found in Table I. Results are shown in Figure 5, where the initial distributions $\mu_0^{(\ell)}$ for the different agents can be seen in the left-most column (showing time point $j = 0$).

VI. Conclusions

In this paper we have seen that graph structured tensor optimization problems naturally appear in several areas in systems and control. We have developed numerical algorithms for these problems based on dual coordinate ascent that utilize the fact that the dual problems decouple according to the graph structure. We also showed that under mild conditions these algorithms are globally convergent, and in certain cases the convergence is R-linear. We believe that these methods are useful for addressing many other types of problems, e.g., in flow problems where the nodes or edges also have dynamics (cf. [25]). Moreover, we also believe that these methods can be extended to handle, e.g., multi-species potential mean field games where each species also has different dynamics.

Appendix

A. Deferred proofs

Proof of Lemma III.3: By Assumption A, there is a feasible point to (5) with finite objective function value, and since problems (5) and (4) are equivalent, this means that the objective function in (4) is proper. To show that the minimum for the latter is attained, note that $g^{(t)}$, $t \in \mathcal{V}$, and $f^{(t_1,t_2)}$, $(t_1,t_2) \in \mathcal{E}$, are all proper, convex, and lower-semicontinuous, and hence they all have a continuous affine minorant [4, Thm. 9.20]. However, and since the entropy term $\epsilon D(\mathbf{M})$ is radially unbounded and grows faster towards ∞ than linearly, we have that the objective function in (4) is radially unbounded. Since the entire objective function is also proper, convex, and lower-semicontinuous, the minimum is attained [69, Thm. 27.2], and it is unique since $D(\mathbf{M})$ (and hence the entire objective function in (4)) is strictly convex. ∎

[8] Note that $\tilde{\mu}_2$ and $\tilde{\mu}^{(4)}$ are uniform distributions. The former has the same total mass as the total distribution μ_0, and the latter the same as $\mu_0^{(4)}$.

Lemma A.1. *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be proper, convex, and lower-semicontinuous, then $ri(dom(f^*)) \neq \emptyset$.*

Proof: Since f is proper, convex, and lower-semicontinuous, so is f^* [4, Cor. 13.38]. $\mathrm{dom}(f^*)$ is therefore nonempty, and by [4, Prop. 8.2] it is convex. Using [4, Fact 6.14(i)], the result follows. ∎

Lemma A.2. *There is no duality gap between* (5) *and* (6)*.*

Proof: To prove it, we derive a Lagrangian dual to an equivalent problem to (6), and show that for the latter strong duality holds with (5). To this end, note that a problem with the same set of globally optimal solutions as (6) is the constrained optimization problem

$$\begin{aligned}
\sup_{\mathbf{U},\lambda,\Lambda} \quad & -\epsilon\langle \mathbf{K}, \mathbf{U} \rangle - \sum_{t\in\mathcal{V}} (g^{(t)})^*(-\lambda_t) - \sum_{(t_1,t_2)\in\mathcal{E}} (f^{(t_1,t_2)})^*(-\Lambda_{t_1,t_2}) \\
\text{s.t.} \quad & \log(\mathbf{U}_{i_1\dots i_\mathcal{T}}) = \frac{1}{\epsilon}\left(\sum_{t\in\mathcal{V}} \lambda_t^{(i_t)} + \sum_{(t_1,t_2)\in\mathcal{E}} \Lambda_{t_1,t_2}^{(i_{t_1},i_{t_2})} \right).
\end{aligned}$$

However, the latter is nonconvex due to the nonaffine equality constraint. Nevertheless, since $\mathbf{K} \geq 0$ the cost function is nonincreasing in $\mathbf{U}$, and since the logarithm is a monotone increasing function, the above problem has the same globally optimal solution as the relaxed, convex problem with the equality changed for an inequality $\geq$. Moreover, for this convex problem, by using Lemma A.1 it is easily seen that Slater's condition is fulfilled, and hence strong duality holds. Next, relaxing the convex inequality constraints with multipliers $\mathbf{Q}_{i_1\dots i_\mathcal{T}} \geq 0$ we get the Lagrangian

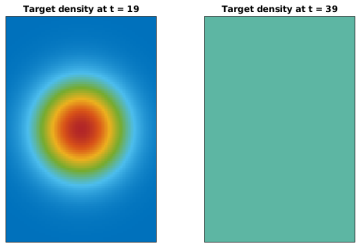
$$\begin{aligned}
& -\epsilon\langle \mathbf{K}, \mathbf{U} \rangle - \sum_{t\in\mathcal{V}} (g^{(t)})^*(-\lambda_t) - \sum_{(t_1,t_2)\in\mathcal{E}} (f^{(t_1,t_2)})^*(-\Lambda_{t_1,t_2}) \\
& + \sum_{i_1\dots i_\mathcal{T}} \mathbf{Q}_{i_1\dots i_\mathcal{T}} \left(\log(\mathbf{U}_{i_1\dots i_\mathcal{T}}) - \frac{1}{\epsilon}\left(\sum_{t\in\mathcal{V}} \lambda_t^{(i_t)} + \sum_{(t_1,t_2)\in\mathcal{E}} \Lambda_{t_1,t_2}^{(i_{t_1},i_{t_2})} \right) \right)
\end{aligned}$$

which separates over λ_t, Λ_{t_1,t_2}, and $\mathbf{U}$. Moreover, we have that $\sum_{i_1\dots i_\mathcal{T}} \mathbf{Q}_{i_1\dots i_\mathcal{T}} \frac{1}{\epsilon}\lambda_t^{(i_t)} = \langle 1/\epsilon P_t(\mathbf{Q}), \lambda_t \rangle$, and therefore when taking the supremum over λ_t we get

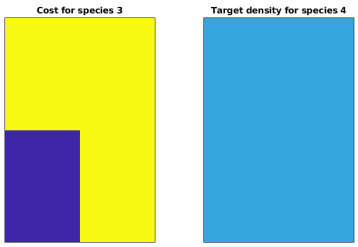
$$\begin{aligned}
\sup_{\lambda_t \in \mathbb{R}^N} -(g^{(t)})^*(-\lambda_t) - \langle 1/\epsilon P_t(\mathbf{Q}), \lambda_t \rangle &= (g^{(t)})^{**}(1/\epsilon P_t(\mathbf{Q})) \\
&= g^{(t)}(1/\epsilon P_t(\mathbf{Q})),
\end{aligned}$$

where the last equality follows from [4, Thm. 13.37]; an analogous result holds for $(f^{(t_1,t_2)})^*$ and Λ_{t_1,t_2}. The remaining part of the Lagrangian is $\sup_{\mathbf{U}\in\mathbb{R}^{N^\mathcal{T}}} -\epsilon\langle \mathbf{K}, \mathbf{U} \rangle + \langle \mathbf{Q}, \log(\mathbf{U}) \rangle$, and to find this supremum we first note that if $\mathbf{K}_{i_1\dots i_\mathcal{T}} = 0$, then we must have $\mathbf{Q}_{i_1\dots i_\mathcal{T}} = 0$ or else the cost function is unbounded. For all other elements, we take the derivative with respect to $\mathbf{U}_{i_1\dots i_\mathcal{T}}$ and set it equal to zero, from which it follows that $\mathbf{U}_{i_1\dots i_\mathcal{T}} = \mathbf{Q}_{i_1\dots i_\mathcal{T}}/(\epsilon\mathbf{K}_{i_1\dots i_\mathcal{T}}) > 0$, which is hence the supremum. Plugging this back into the cost, we get

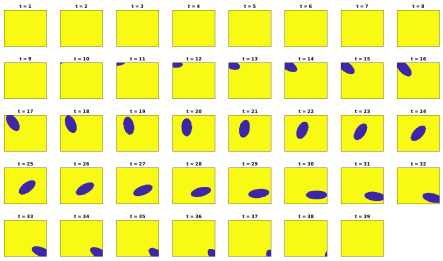
$$\begin{aligned}
& -\epsilon\langle \mathbf{K}, \mathbf{U} \rangle + \langle \mathbf{Q}, \log(\mathbf{U}) \rangle \\
& \quad = \sum_{i_1\dots i_\mathcal{T}} -\mathbf{Q}_{i_1\dots i_\mathcal{T}} + \langle \mathbf{Q}, \log(\mathbf{Q}) \rangle - \langle \mathbf{Q}, \log(\epsilon\mathbf{K}) \rangle \\
& \quad = \sum_{i_1\dots i_\mathcal{T}} -\mathbf{Q}_{i_1\dots i_\mathcal{T}} + \langle \mathbf{Q}, \log(\mathbf{Q}) - \log(\epsilon) \rangle + (1/\epsilon)\langle \mathbf{Q}, \mathbf{C} \rangle,
\end{aligned}$$



(a) Target densities $\tilde{\mu}_1$ (left) and $\tilde{\mu}_2$ (right) for the total density at time points $j = 19$ and $j = 39$, respectively.



(b) The left plot shows the linear cost c_3 for species 3, where blue means cost 0 and yellow means a cost of $390\Delta x\Delta t$. The right plots shows the target distributions $\tilde{\mu}^{(4)}$ for species 4.



(c) Illustration of the capacity constraints κ_j at the different time points j: blue means zero capacity (obstacle) while yellow means infinite capacity.

Fig. 4: Figures describing the setup in the numerical example in Section V-D.

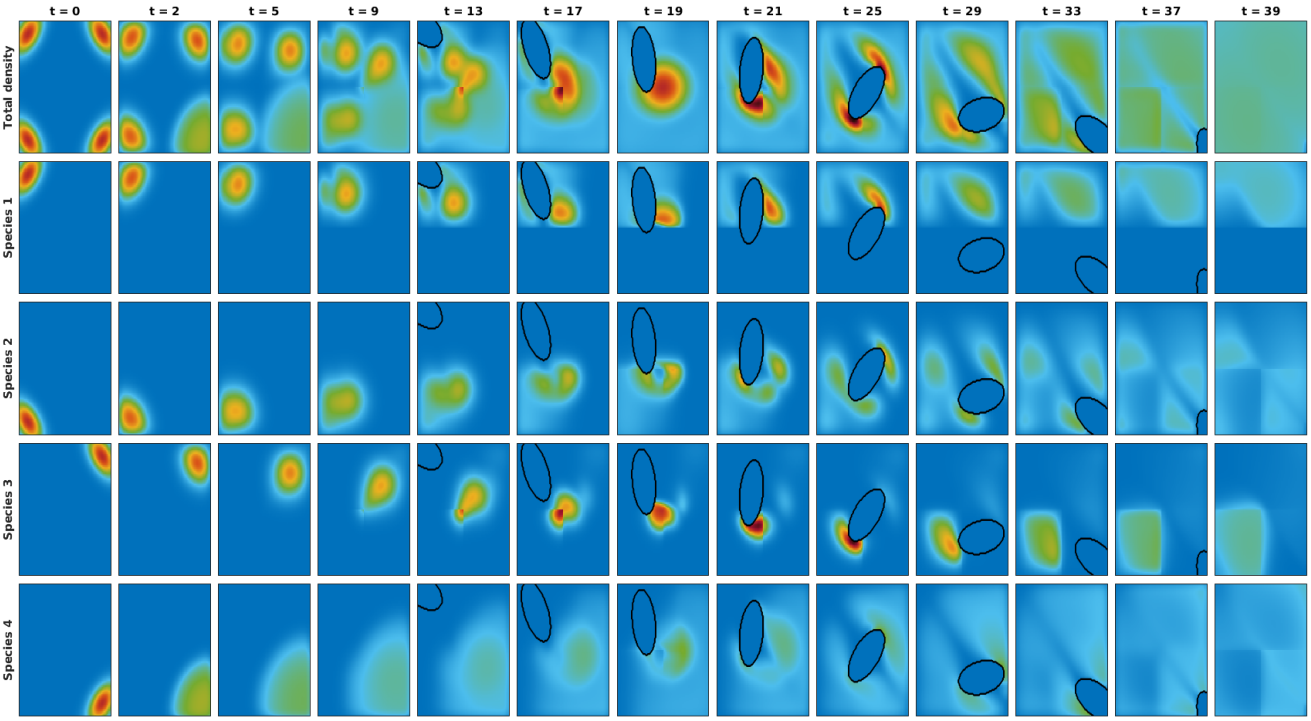


Fig. 5: Time evolution of total density and densities of the individual species, for the numerical example in Section V-D.

together with the constraints that $\mathbf{Q}_{i_1\dots i_\mathcal{T}} = 0$ if $\mathbf{K}_{i_1\dots i_\mathcal{T}} = 0$. But for any element such that $\mathbf{K}_{i_1\dots i_\mathcal{T}} = 0$ we have that $\mathbf{C}_{i_1\dots i_\mathcal{T}} = \infty$, and the constraints can thus be removed since they are implicitly enforced by the cost function. Therefore, with the change of variable $\mathbf{Q} = \epsilon\mathbf{M}$ we recover, up to a constant, the primal problem (4). Since (4) has the same optimal value as (5), the result follows. ∎

Proof of Theorem V.2: Note that $\mathbf{K}_{\ell i_0\dots i_\mathcal{T}} = \prod_{t=0}^{\mathcal{T}-1} K_{i_t,i_{t+1}}$. Together with (30), this means that

$$(P_{-1,j}(\mathbf{K}\odot\mathbf{U}))_{\ell,i_j} = \sum_{\substack{i_0,\dots i_{j-1}\\ i_{j+1},\dots i_\mathcal{T}}} \left(\left(\prod_{t=0}^{\mathcal{T}-1} K_{i_t,i_{t+1}} U^{(\ell,i_0)}_{-1,0}\right)\right.$$
$$\left.\left(\prod_{t=1}^{\mathcal{T}} U^{(\ell,i_t)}_{-1,t}\right)\left(\prod_{t=1}^{\mathcal{T}} u_t^{(i_t)}\right)\right) = U^{(\ell,i_j)}_{-1,j} u_j^{(i_j)} \hat{\Psi}_j^{(\ell,j)} \Psi_j^{(\ell,j)},$$

where

$$\hat{\Psi}_j^{(\ell,i_j)} = \sum_{i_0,\dots i_{j-1}} U^{(\ell,i_0)}_{-1,0} K_{i_0,i_1} \prod_{t=1}^{j-1} U^{(\ell,i_t)}_{-1,t} u_t^{(i_t)} K_{i_t,i_{t+1}},$$

$$\Psi_j^{(\ell,i_j)} = \sum_{i_{j+1},\dots i_\mathcal{T}} U^{(\ell,i_\mathcal{T})}_{-1,\mathcal{T}} u_\mathcal{T}^{(i_\mathcal{T})} K_{i_{\mathcal{T}-1},i_\mathcal{T}} \prod_{t=j+1}^{\mathcal{T}-1} U^{(\ell,i_t)}_{-1,t} u_t^{(i_t)} K_{i_{t-1},i_t}.$$

A direct calculation gives that $\hat{\Psi}_j$ and Ψ_j above fulfill the recursive definitions in the theorem, which proves the form of the bimarginal projection for $j = 1,\dots,\mathcal{T}-1$. Next, the form of the bimarginal projections for $j = 0$ and $\mathcal{T}$ can be readily verified analogously. Finally, note that

$$(P_j(\mathbf{K}\odot\mathbf{U}))_{i_j} = \sum_{\ell=1}^{L} (P_{-1,j}(\mathbf{K}\odot\mathbf{M}))_{\ell,i_j},$$

which gives the result for the projections and proves the theorem. ∎

REFERENCES

[1] Y. Achdou, M. Bardi, and M. Cirant. Mean field games models of segregation. *Mathematical Models and Methods in Applied Sciences*, 27(01):75–113, 2017.

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

[3] J.E. Aronson. A survey of dynamic network flows. *Annals of Operations Research*, 20(1):1–66, 1989.

[4] H.H. Bauschke and P.L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, Cham, 2nd edition, 2017.

[5] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[6] 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(2):A1111–A1138, 2015.

[7] J.-D. Benamou, G. Carlier, S. Di Marino, and L. Nenna. An entropy minimization approach to second-order variational mean-field games. *Mathematical Models and Methods in Applied Sciences*, 29(08):1553–1583, 2019.

[8] A. Bensoussan, J. Frehse, and P. Yam. *Mean field games and mean field type control theory*. Springer, New York, NY, 2013.

[9] A. Bensoussan, T. Huang, and M. Laurière. Mean field control and mean field game models with several populations. *Minimax Theory and its Applications*, 3(2):173–209, 2018.

[10] D. Bertsekas. *Network optimization: continuous and discrete models*. Athena Scientific, 1998.

[11] D. Bertsimas and S. S. Patterson. The traffic flow management rerouting problem in air traffic control: A dynamic network flow approach. *Transportation Science*, 34(3):239–255, 2000.

[12] P.E. Caines, M. Huang, and R.P. Malhamé. Mean field games. In T. Başar and G. Zaccour, editors, *Handbook of Dynamic Game Theory*, pages 345–372. Springer, Cham, 2018.

[13] K. Caluya and A. Halder. Wasserstein proximal algorithms for the Schrödinger bridge problem: Density control with nonlinear drift. *IEEE Transactions on Automatic Control*, 2021.

[14] P. Cardaliaguet, P.J. Graber, A. Porretta, and D. Tonon. Second order mean field games with degenerate diffusion and local coupling. *Nonlinear Differential Equations and Applications NoDEA*, 22(5):1287–1317, 2015.

[15] Y. Chen, T.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(2):671–691, 2016.

[16] Y. Chen, T.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(5):1158–1169, 2016.

[17] Y. Chen, T.T. Georgiou, and M. Pavon. Optimal transport over a linear dynamical system. *IEEE Transactions on Automatic Control*, 62(5):2137–2152, 2017.

[18] Y. Chen, T.T. Georgiou, and M. Pavon. Steering the distribution of agents in mean-field games system. *Journal of Optimization Theory and Applications*, 179(1):332–357, 2018.

[19] Y. Chen, T.T. Georgiou, and M. Pavon. Stochastic control liasons: Richard Sinkhorn meets Gaspard Monge on a Schrödinger bridge. *SIAM Review*, 63(2):249–313, 2021.

[20] Y. Chen, T.T. Georgiou, M. Pavon, and A. Tannenbaum. Robust transport over networks. *IEEE Transactions on Automatic Control*, 62(9):4675–4682, 2016.

[21] Y. Chen, T.T. Georgiou, M. Pavon, and A. Tannenbaum. Efficient robust routing for single commodity network flows. *IEEE Transactions on Automatic Control*, 63(7):2287–2294, 2017.

[22] Y. Chen, T.T. Georgiou, M. Pavon, and A. Tannenbaum. Relaxed Schrödinger bridges and robust network routing. *IEEE Transactions on Control of Network Systems*, 7(2):923–931, 2019.

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

[24] M. Cirant. Multi-population mean field games systems with Neumann boundary conditions. *Journal de Mathématiques Pures et Appliquées*, 103(5):1294–1315, 2015.

[25] G. Como. On resilient control of dynamical flow networks. *Annual Reviews in Control*, 43:80–90, 2017.

[26] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2292–2300, 2013.

[27] P. Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1):313–329, 1991.

[28] B. Djehiche, A. Tcheukam, and H. Tembine. Mean-field-type games in engineering. *AIMS Electronics and Electrical Engineering*, 1(1):18–73, 2017.

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

[30] J. Fan, I. Haasler, J. Karlsson, and Y. Chen. On the complexity of the optimal transport problem with graph-structured cost. In *International Conference on Artificial Intelligence and Statistics*, pages 9147–9165. PMLR, 2022.

[31] H. Farhangi. The path of the smart grid. *IEEE Power Energy Mag.*, 8(1), 2010.

[32] L. Fleischer and M. Skutella. Minimum cost flows over time without intermediate storage. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 66–75, 2003.

[33] W.H. Fleming and R.W. Rishel. *Deterministic and stochastic optimal control*. Springer-Verlag, New York, N.Y., 1975.

[34] H. Föllmer. Random fields and diffusion processes. In P.-L. Hennequin, editor, *École d'Été de Probabilités de Saint-Flour XV–XVII, 1985–87*, volume 1362 of *Lecture Notes in Mathematics*, pages 101–203. Springer, Berlin, Heidelberg, 1988.

[35] L.R. Ford and D.R. Fulkerson. Constructing maximal dynamic flows from static flows. *Operations research*, 6(3):419–433, 1958.

[36] W. Gangbo and A. Święch. Optimal maps for the multidimensional Monge-Kantorovich problem. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 51(1):23–45, 1998.

[37] I. Gentil, C. Léonard, and L. Ripani. About the analogy between optimal transport and minimal entropy. *Annales de la Faculté des sciences de Toulouse: Mathématiques*, 26(3):569–600, 2017.

[38] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2022.

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

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

[41] I. Haasler, A. Ringh, Y. Chen, and J. Karlsson. Scalable computation of dynamic flow problems via multi-marginal graph-structured optimal transport. *arXiv preprint arXiv:2106.14485v1*, 2021.

[42] I. Haasler, A. Ringh, and J. Karlsson. Control and estimation of ensembles via strucutred optimal transport: A computational approach based on entropy-regularized multimarginal optimal transport. *IEEE Control Systems Magazine*, 41(4):50–69, 2021.

[43] I. Haasler, R. Singh, Q. Zhang, J. Karlsson, and Y. Chen. Multi-marginal optimal transport and probabilistic graphical models. *IEEE Transactions on Information Theory*, 67(7):4647–4668, 2021.

[44] A. Hindawi, J.-B. Pomet, and L. Rifford. Mass transportation with LQ cost functions. *Acta applicandae mathematicae*, 113(2):215–229, 2011.

[45] M. Huang, P.E. Caines, and R.P. Malhamé. Social optima in mean field LQG control: centralized and decentralized strategies. *IEEE Transactions on Automatic Control*, 57(7):1736–1751, 2012.

[46] M. Huang, R.P. Malhamé, and P.E. Caines. Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Communications in Information & Systems*, 6(3):221–252, 2006.

[47] B. Jovanovic and R.W. Rosenthal. Anonymous sequential games. *Journal of Mathematical Economics*, 17(1):77–87, 1988.

[48] 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.

[49] J. L. Kennington. A survey of linear cost multicommodity network flows. *Operations Research*, 26(2):209–236, 1978.

[50] A. Lachapelle and M.-T. Wolfram. On a mean field game approach modeling congestion and aversion in pedestrian crowds. *Transportation research part B: methodological*, 45(10):1572–1589, 2011.

[51] B. Lamond and N.F. Stewart. Bregman's balancing method. *Transportation Research Part B: Methodological*, 15(4):239–248, 1981.

[52] J.-M. Lasry and P.-L. Lions. Mean field games. *Japanese journal of mathematics*, 2(1):229–260, 2007.

[53] C. Léonard. From the Schrödinger problem to the Monge–Kantorovich problem. *Journal of Functional Analysis*, 262(4):1879–1920, 2012.

[54] C. Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete & Continuous Dynamical Systems - A*, 34(4):1533–1574, 2014.

[55] T. Lin, N. Ho, M. Cuturi, and M.I. Jordan. On the complexity of approximating multimarginal optimal transport. *Journal of Machine Learning Research*, 23(65):1–43, 2022.

[56] D.G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, New York, NY, 1969.

[57] Z.-Q. Luo and P. Tseng. On the convergence rate of dual ascent methods for linearly constrained convex minimization. *Mathematics of Operations Research*, 18(4):846–867, 1993.

[58] G. Meyer and S. Beiker, editors. *Road Vehicle Automation*. Springer, Cham, 2014.

[59] L. Nenna. *Numerical methods for multi-marginal optimal transportation*. PhD thesis, PSL Research University, 2016.

[60] J. Nocedal and S. Wright. *Numerical optimization*. Springer, New York, NY, 2nd edition, 2006.

[61] J.B. Orlin. Minimum convex cost dynamic network flows. *Mathematics of Operations Research*, 9(2):190–207, 1984.

[62] J.M. Ortega and W.C. Rheinboldt. *Iterative solution of nonlinear equations in several variables*. Academic Press, New York, NY, 1970.

[63] A. Ouorou, P. Mahey, and J.-Ph. Vial. A survey of algorithms for convex multicommodity flow problems. *Management science*, 46(1):126–147, 2000.

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

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

[66] W.B. Powell, P. Jaillet, and A. Odoni. Stochastic and dynamic networks and routing. *Handbooks in operations research and management science*, 8:141–295, 1995.

[67] K.J. Åström. *Introduction to stochastic control theory*. Dover, Mineola, NY, 2006. Unabridged republication of original published by Academic Press, 1970.

[68] A. Ringh, I. Haasler, Y. Chen, and J. Karlsson. Efficient computations of multi-species mean field games via graph-structured optimal transport. Accetpted to *IEEE 60th Conference on Decision and Control (CDC)*, 2021.

[69] R.T. Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, NJ, 1970.

[70] L. Rüschendorf. Optimal solutions of multivariate coupling problems. *Applicationes Mathematicae*, 23(3):325–338, 1995.

[71] L. Rüschendorf and L. Uckelmann. On the n-coupling problem. *Journal of multivariate analysis*, 81(2):242–258, 2002.

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

[73] M. Skutella. An introduction to network flows over time. In *Research trends in combinatorial optimization*, pages 451–482. Springer, 2009.

[74] Y.W. Teh and M. Welling. The unified propagation and scaling algorithm. *Advances in neural information processing systems*, pages 953–960, 2002.

[75] P. Tseng. Dual coordinate ascent methods for non-strictly convex minimization. *Mathematical programming*, 59(1-3):231–247, 1993.

[76] H.J. Van Zuylen and L.G. Willumsen. The most likely trip matrix estimated from traffic counts. *Transportation Research Part B: Methodological*, 14(3):281–293, 1980.

[77] C. Villani. *Topics in optimal transportation*. American Mathematical Society, Providence, RI, 2003.

[78] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440, 1998.