SIAM J. Control Optim.
Vol. 62, No. 4, pp. 2176–2202

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 problem. This type of problem, which can be seen as a generalization of multimarginal optimal transport problems with graph-structured costs, appears in many applications. Examples are unbalanced optimal transport and multispecies potential mean field games, where the latter is a class of nonlinear density control problems. The method we develop 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 it is possible to derive efficient methods for computing these projections, and here we specifically consider the graph structure that occurs in multispecies potential mean field games. We also illustrate the methodology on a numerical example from this problem class.

Key words. tensor optimization, large-scale convex optimization, optimal transport, Sinkhorn algorithm, unbalanced optimal transport, potential mean field games

MSC codes. 90C06, 90C25, 90C35, 90C46, 49Q22, 91A15, 49M25, 93E20

DOI. 10.1137/23M1571587

1. 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 [29] and road networks with self-driving cars [56]. There are also many such problems in biology, ecology, and social sciences, including, e.g., cell, animal, and human populations [73]. A major challenge is understanding and controlling the macroscopic behavior of such complex large-scale systems, but since the number of agents in such systems is often too large to enable modeling of 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 great interest in control and estimation of densities, including swarm control [46, 68], modeling and control of epidemics

*Received by the editors May 9, 2023; accepted for publication (in revised form) March 11, 2024; published electronically July 16, 2024.
https://doi.org/10.1137/23M1571587
Funding: This work was supported by the Swedish Research Council (VR) under grant 2020-03454, KTH Digital Futures, the NSF under grants 1942523 and 2206576, the Knut and Alice Wallenberg Foundation under grants KAW 2018.0349 and KAW 2021.0274, and by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.
†Department of Mathematical Sciences, Chalmers University of Technology, and the University of Gothenburg, Gothenburg, Sweden (axelri@chalmers.se).
‡Signal Processing Laboratory, LTS 4, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland (isabel.haasler@epfl.ch).
§School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (yongchen@gatech.edu).
¶Division of Optimization and Systems Theory, Department of Mathematics, KTH Royal Institute of Technology, Stockholm, Sweden (johan.karlsson@math.kth.se).

[50], and covariance control in stochastic systems [15]. One key result is that certain density control problems of first-order integrators can be seen as optimal transport problems [6]. 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 [12, 16, 41].

While some density flow problems can be viewed as two-marginal optimal transport problems, many problems involve using a time grid in order to model, e.g., congestion, instantaneous costs, or observations [35]. For such problems it is natural to use versions of the multimarginal optimal transport problem. The latter is an optimization problem where a nonnegative tensor is sought to minimize a linear cost subject to constraints on the marginals, where the marginals are projections of the tensor on specific modes. When modeling the evolution of a system on a time grid, the marginals represent the distributions at different time points $j = 1, \dots, \mathcal{T}$. More specifically, for such control problems with identical and indistinguishable agents, the problem can be separated 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., [19, 27, 37]). However, when the agents have heterogeneous dynamics or objectives, the distribution at a given time does not contain all necessary information about the past. 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 nodes [35], and Euler flow problems can be seen as a special case of this [7]. By introducing an additional marginal representing different types of agents, we can also formulate and solve multispecies dynamic flow problems and large multicommodity problems [36]. The resulting optimization problems are large-scale problems, but algorithms have been developed to solve this type of structured multimarginal optimal transport problem [2, 7, 27, 28, 35, 36, 37, 38, 39, 67]. These extend Sinkhorn's method, developed for solving the bimarginal problem, in which an entropy regularization is added to the cost function and an approximate solution is computed by using coordinate ascent in the dual problem [24, 61]. Interestingly, in the bimarginal problem the entropy regularization term can also be interpreted as introducing stochasticity in the dynamics of the subsystems, and the entropy-regularized problem can in fact be shown to be equivalent to the Schrödinger bridge problem [18, 51, 52]. This connection has also been extended to the multimarginal setting [37]. Moreover, in this context it is also interesting to note that the algorithms developed for solving this type of structured multimarginal problem are closely related to the unified propagation and scaling algorithm for inference in graphical models [69].

Many of the problems in the previous paragraph can be formulated as optimal transport problems with fixed marginals. Nevertheless, in many situations it is also natural to consider problems where the marginals are not exactly known. A common strategy is then to penalize deviations from some given marginals [5, 8, 20, 21, 34, 45, 53, 62]. This is sometimes referred to as unbalanced optimal transport. The cost functions associated with the nonfixed marginals are often convex, but standard convex optimization methods in general do not scale to this type of large-scale problem. In this paper, we develop a theoretical framework for a type of convex structured tensor optimization problem, which is a generalization of graph-structured multimarginal optimal transport problems, along with numerical solution methods and convergence results for these. We also illustrate how this type of problem can be used to model and solve multispecies potential mean field games; this gives a solution method which is a generalization of the method for potential mean field games developed in [8]. An important observation is that the dual problem has a decomposable form and can be efficiently solved by using dual coordinate ascent [45] (cf. [61]). 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, which thus alleviates the computational bottleneck of the algorithm.

The outline of the paper is as follows: In section 2 we introduce some background material on optimal transport and convex optimization. The main results are presented in section 3, 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 4 we develop an algorithm for solving multispecies potential mean field games. This is done by casting the problem as a graph-structured tensor optimization problem and then specializing the general algorithm to the particular instance. In that section, we also present a numerical example to illustrate the use and performance of the algorithm. Finally, section 5 contains some concluding remarks. Some proofs are deferred to Appendix A for improved readability. This paper builds on [63], where we presented an algorithm, without proof of convergence, for the multispecies mean field game in a simplified setting (see also Remark 4.3).

2. Background.

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

2.1. The graph-structured multimarginal optimal transport problem.

The optimal transport problem seeks a transport plan for moving mass from an initial distribution to a target distribution with minimum cost. This topic has been extensively studied; see, e.g., the monograph [72] and references therein. An extension of this problem is the multimarginal optimal transport problem, in which a minimum-cost transport plan between several distributions is sought [7, 27, 32, 57, 60, 65, 66]. 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, the elements $\mathbf{M}^{(i_1 \dots i_{\mathcal{T}})}$ 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 elements of this vector are defined as

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

[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.

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 multimarginal optimal transport problem can thus be formulated as

$$\underset{\mathbf{M}\in\mathbb{R}_+^{N^{\mathcal{T}}}}{\text{minimize}} \quad \langle \mathbf{C},\mathbf{M}\rangle \tag{2.1a}$$

$$\text{subject to} \quad P_j(\mathbf{M})=\mu_j, \quad j\in\Gamma. \tag{2.1b}$$

Problem (2.1) 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 [24] (see also [61]). This perturbed problem can be solved by using the so-called Sinkhorn iterations.[2] The approach has been extended to the multimarginal setting [7, 27, 57]; however, in this case it only partly alleviates the computational difficulty. More precisely, in the multimarginal 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 $\mathbf{M}=\mathbf{K}\odot\mathbf{U}$ (see [7, 27]), 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 logarithm of the variables u_j correspond to the Lagrangian dual variables in a relaxation of the entropy-regularized version of (2.1). Moreover, the (multimarginal) Sinkhorn iterations iteratively update u_j to match the given marginals as follows:

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

where $\oslash$ denotes elementwise division. However, in the multimarginal 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 [54]. 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

$$\underset{\mathbf{M}\in\mathbb{R}_+^{N^{\mathcal{T}}}}{\text{minimize}} \quad \langle \mathbf{C},\mathbf{M}\rangle+\epsilon D(\mathbf{M}) \tag{2.2a}$$

$$\text{subject to} \quad P_t(\mathbf{M})=\mu_t, \quad t\in\tilde{\mathcal{V}}, \tag{2.2b}$$

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

$$\mathbf{C}^{(i_1\dots i_{\mathcal{T}})}=\sum_{(t_1,t_2)\in\mathcal{E}}C_{t_1,t_2}^{(i_{t_1},i_{t_2})}, \tag{2.3}$$

[2] In fact, the iterations have been discovered in different settings and therefore also have many different names; see, e.g., [18, 48].

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

where $C_{t_1,t_2} \in \mathbb{R}^{N\times N}$, which in particular means that the linear cost term in (2.2a) takes the form $\langle \mathbf{C}, \mathbf{M}\rangle = \sum_{(t_1,t_2)\in\mathcal{E}} \langle C_{t_1,t_2}, P_{t_1,t_2}(\mathbf{M})\rangle$. 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 (2.2) with a cost tensor structured according to (2.3) is called an (entropy-regularized) graph-structured multimarginal optimal transport problem [28, 36, 37]. Moreover, for many graph structures, the projections $P_t(\mathbf{M})$ and $P_{t_1,t_2}(\mathbf{M})$ can be efficiently computed (see, e.g, [2, 7, 27, 28, 35, 36, 37, 38, 39, 67]), and hence the Sinkhorn iterations can be used to efficiently solve such problems.

2.2. 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, 64]. 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 $\mathrm{epi}(f) := \{(x,\eta)\in \mathbb{R}^n\times\mathbb{R} \mid f(x)\le \eta\}$, and f is called convex if $\mathrm{epi}(f)\subset\mathbb{R}^{n+1}$ is a convex set. A function f is *lower-semicontinuous* if and only if $\mathrm{epi}(f)$ is closed [64, Thm. 7.1]. The *effective domain* of f is defined as $\mathrm{dom}(f) := \{x\in\mathbb{R}^n \mid f(x)<\infty\}$, and f is called *proper* if $f(x) > -\infty$ for all $x\in\mathbb{R}^n$ and $\mathrm{dom}(f)\neq\emptyset$. 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 $\mathrm{epi}(f)$ 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 *cofinite* if $\mathrm{epi}(f)$ contains no nonvertical half-lines, which is equivalent to the fact that f^* is finite everywhere, i.e., that $\mathrm{dom}(f^*)=\mathbb{R}^n$ [64, Cor. 13.3.1]. The *subdifferential* of a function f in a point x is the set $\partial f(x) := \{u\in\mathbb{R}^n \mid \langle y-x,u\rangle + f(x)\le f(y) \text{ for all } y\in\mathbb{R}^n\}$, 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 $\mathrm{int}(\mathrm{dom}(f))$, i.e., on the interior of the effective domain; and (ii) $\lim_{\ell\to\infty}\|\nabla f(x_\ell)\|\to\infty$ for any sequence $\{x_\ell\}_\ell\subset \mathrm{int}(\mathrm{dom}(f))$ that either converges to the boundary of $\mathrm{int}(\mathrm{dom}(f))$ or is such that $\|x_\ell\|\to\infty$. An operator $A:\mathbb{R}^n\to\mathbb{R}^n$ is called strongly monotone if there exists a $\gamma>0$ such that $\langle Ax-Ay, x-y\rangle \ge \gamma\|x-y\|^2$ for all $x,y\in\mathbb{R}^n$. Finally, let $\{x_\ell\}_\ell\subset\mathbb{R}^n$ be a sequence converging to some $\bar{x}\in\mathbb{R}^n$. The sequence is said to converge Q-linearly if there exists a $\gamma\in(0,1)$ such that $\|x_{\ell+1}-\bar{x}\|\le\gamma\|x_\ell-\bar{x}\|$, and the sequence is said to converge R-linearly[4] if there exists a sequence of nonnegative numbers $\{\gamma_\ell\}_\ell\subset\mathbb{R}_+$ converging Q-linearly to zero and such that $\|x_\ell-\bar{x}\|\le\gamma_\ell$ for all ℓ [59, sect. 9.2], [58, pp. 619–620].

3. Convex graph-structured tensor optimization. In this work, we consider a family of optimization problems that generalizes problems of the form (2.2). To this end, let $\mathcal{G}=(\mathcal{V},\mathcal{E})$ be a connected graph with $\mathcal{T}=|\mathcal{V}|$ nodes, and let $\mathbf{C}\in\bar{\mathbb{R}}^{N^\mathcal{T}}$ be a cost tensor that takes the form (2.3). The convex graph-structured tensor optimization problems of interest here are problems of the form

$$\text{(3.1)} \qquad \underset{\mathbf{M}\in\mathbb{R}_+^{N^\mathcal{T}}}{\text{minimize}} \quad \langle \mathbf{C},\mathbf{M}\rangle + \epsilon D(\mathbf{M}) + \sum_{t\in\mathcal{V}} g_t(P_t(\mathbf{M})) + \sum_{(t_1,t_2)\in\mathcal{E}} f_{t_1,t_2}(P_{t_1,t_2}(\mathbf{M})),$$

[4]This is a slightly weaker notion of convergence, compared to Q-linear, that "is concerned with the overall rate of decrease in the error, rather than the decrease over each individual step of the algorithm" [58, pp. 619–620].

where g_t and f_{t_1,t_2} are proper, convex, and lower-semicontinuous functions; further assumptions on these functions will be imposed where needed. The reason for our interest in problems of the form (3.1) is that a number of different applications can be modeled as such problems. In particular, this is true for convex dynamic network flow problems (cf. [36]) and potential multispecies mean field games. The latter is studied in detail in section 4.

Remark 3.1. To see that problems of the form (3.1) are generalizations of the graph-structured multimarginal optimal transport problem (2.2), let $\iota_A(\cdot)$ denote the indicator function on the set $A \subseteq \mathbb{R}^n$, i.e., the function

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

and note that this function is proper, convex, and lower-semicontinuous if and only if A is a nonempty, closed, convex set. Now, (2.2) is recovered from (3.1) by taking $g_t(P_t(\mathbf{M})) = \iota_{\{\mu_t\}}(P_t(\mathbf{M}))$ 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 $(t_1,t_2) \in \mathcal{E}$. Other particular cases of interest are unbalanced versions of (2.2) [5] or versions of (2.2) where some of the equality constraints are replaced by inequality constraints; cf. [36].

Remark 3.2. In problem (3.1) 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 (2.3) to be the zero-matrix.

Note that (3.1) 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 (3.1) 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. [28, 36, 39]), and when the functions g_t and f_{t_1,t_2} are in some sense simple.

3.1. An equivalent problem and existence of solution. We first introduce and analyze a problem that is equivalent to (3.1) 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}, $(t_1,t_2) \in \mathcal{E}$, we can rewrite problem (3.1) as

$$\underset{\substack{\mathbf{M} \in \mathbb{R}_+^{N^{\mathcal{T}}},\, \mu_t \in \mathbb{R}^N,\, t \in \mathcal{V} \\ R_{t_1,t_2} \in \mathbb{R}^{N \times N},\, (t_1,t_2) \in \mathcal{E}}}{\text{minimize}} \quad \langle \mathbf{C}, \mathbf{M} \rangle + \epsilon D(\mathbf{M}) + \sum_{t \in \mathcal{V}} g_t(\mu_t) + \sum_{(t_1,t_2) \in \mathcal{E}} f_{t_1,t_2}(R_{t_1,t_2}) \tag{3.2a}$$

$$\text{subject to} \quad P_t(\mathbf{M}) = \mu_t,\ t \in \mathcal{V}, \tag{3.2b}$$

$$P_{t_1,t_2}(\mathbf{M}) = R_{t_1,t_2},\ (t_1,t_2) \in \mathcal{E}. \tag{3.2c}$$

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

Assumption 3.3. Assume that all elements of 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 (3.2) with finite objective function value, i.e., a nonnegative tensor M such that $\langle \mathrm{C}, \mathrm{M}\rangle < \infty$, and with marginals and bimarginals as in (3.2b)–(3.2c), 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, these assumptions ensures that (3.2) has an optimal solution, as stated in the following lemma.

LEMMA 3.4. *If Assumption* 3.3 *holds, then there exists a unique optimal solution to problem* (3.2).

Proof. See Appendix A. ☐

Remark 3.5. A necessary condition for Assumption 3.3 to hold is that there exist vectors $\mu_t \in \mathbb{R}^N_+ \cap \mathrm{dom}(g_t)$ for all $t \in \mathcal{V}$, matrices $R_{t_1,t_2} \in \mathbb{R}^{N\times N}_+ \cap \mathrm{dom}(f_{t_1,t_2})$ 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 3.3. More precisely, the existence of marginals and bimarginals that are consistent with each other does not, in general, guarantee that there exists a tensor that matches the marginals and bimarginals. A counterexample can be found in [37, Rem. 3].

3.2. Form of the optimal solution and Lagrangian dual. Next, we derive the Lagrangian dual of (3.2) and show that there is no duality gap between the primal and dual problems.

THEOREM 3.6. *A Lagrangian dual of* (3.2) *is, up to a constant, given by*

$$\sup_{\substack{\lambda_t \in \mathbb{R}^N,\ t\in\mathcal{V}\\ \Lambda_{t_1,t_2}\in\mathbb{R}^{N\times N},\ (t_1,t_2)\in\mathcal{E}}} -\epsilon\langle \mathrm{K},\mathrm{U}\rangle - \sum_{t\in\mathcal{V}} (g_t)^*(-\lambda_t) - \sum_{(t_1,t_2)\in\mathcal{E}} (f_{t_1,t_2})^*(-\Lambda_{t_1,t_2}), \tag{3.3}$$

where K *and* U *are given by*

$$\mathrm{K}^{(i_1\dots i_\mathcal{T})} = \exp(-\mathbf{C}^{(i_1\dots i_\mathcal{T})}/\epsilon), \tag{3.4a}$$

(3.4b)
$$\mathrm{U}^{(i_1\dots i_\mathcal{T})} = \prod_{t\in\mathcal{V}} u_t^{(i_t)} \prod_{(t_1,t_2)\in\mathcal{E}} U_{t_1,t_2}^{(i_{t_1},i_{t_2})} = \prod_{t\in\mathcal{V}} \exp\left(\lambda_t^{(i_t)}/\epsilon\right) \prod_{(t_1,t_2)\in\mathcal{E}} \exp\left(\Lambda_{t_1,t_2}^{(i_{t_1},i_{t_2})}/\epsilon\right).$$

Moreover, under Assumption 3.3*, the minimum in* (3.2) *equals the supremum in* (3.3) *(up to the discarded constant). Finally, if the dual* (3.3) *has an optimal solution, then the optimal solution to the primal problem takes the form* $\mathrm{M}^\star = \mathrm{K}\odot\mathrm{U}^\star$*, where* $\mathrm{U}^\star$ *is obtained via* (3.4b) *from an optimal solution to* (3.3).

Proof. Relaxing each of the constraints (3.2b) and (3.2c) with multipliers $\lambda_t \in \mathbb{R}^N$ and $\Lambda_{t_1,t_2} \in \mathbb{R}^{N\times N}$, respectively, we get the Lagrangian

$$\begin{aligned}L(\mathbf{M},\mu,R,\lambda,\Lambda) := \langle \mathbf{C},\mathbf{M}\rangle + \epsilon D(\mathbf{M}) + \sum_{t\in\mathcal{V}} g_t(\mu_t) + \sum_{(t_1,t_2)\in\mathcal{E}} f_{t_1,t_2}(R_{t_1,t_2})\\ + \sum_{t\in\mathcal{V}} \lambda_t^T(\mu_t - P_t(\mathbf{M})) + \sum_{(t_1,t_2)\in\mathcal{E}} \mathrm{tr}[\Lambda_{t_1,t_2}^T(R_{t_1,t_2} - P_{t_1,t_2}(\mathbf{M}))],\end{aligned} \tag{3.5}$$

where μ denotes $(\mu_t)_{t\in\mathcal{V}}$, and similar notation is used 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; an analogous result follows for f_{t_1,t_2} and the inf over R_{t_1,t_2}. This means that

$$\begin{aligned}&\inf_{\mathbf{M}\geq 0,\mu,R} L(\mathbf{M},\mu,R,\lambda,\Lambda)\\ &= \inf_{\mathbf{M}\geq 0} \mathcal{L}(\mathbf{M},\lambda,\Lambda) - \sum_{t\in\mathcal{V}}(g_t)^*(-\lambda_t) - \sum_{(t_1,t_2)\in\mathcal{E}} (f_{t_1,t_2})^*(-\Lambda_{t_1,t_2}),\end{aligned} \tag{3.6}$$

where $\mathcal{L}(\mathbf{M},\lambda,\Lambda) := \langle \mathbf{C},\mathbf{M}\rangle + \epsilon D(\mathbf{M}) - \sum_{t\in\mathcal{V}}\lambda_t^T P_t(\mathbf{M}) - \sum_{(t_1,t_2)\in\mathcal{E}} \mathrm{tr}[\Lambda_{t_1,t_2}^T P_{t_1,t_2}(\mathbf{M})]$. Noticing that $\lambda_t^T P_t(\mathbf{M}) = \sum_{i_t=1}^N \lambda_t^{(i_t)} \sum_{i_1,\dots,i_\mathcal{T}\setminus\{i_t\}} \mathbf{M}^{(i_1\dots i_\mathcal{T})} = \sum_{i_1,\dots,i_\mathcal{T}} \lambda_t^{(i_t)} \mathbf{M}^{(i_1\dots i_\mathcal{T})}$ and that $\mathrm{tr}[\Lambda_{t_1,t_2}^T P_{t_1,t_2}(\mathbf{M})] = \sum_{i_1,\dots,i_\mathcal{T}} \Lambda_{t_1,t_2}^{(i_{t_1},i_{t_2})}\mathbf{M}^{(i_1\dots i_\mathcal{T})}$, we see that $\mathcal{L}(\mathbf{M},\lambda,\Lambda)$ 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 then follows that $\mathbf{M}^{(i_1\dots i_\mathcal{T})} > 0$. Moreover, solving for $\mathbf{M}^{(i_1\dots i_\mathcal{T})}$ gives $\mathbf{M} = \mathbf{K}\odot\mathbf{U}$, where $\mathbf{K}$ and $\mathbf{U}$ are given as in (3.4). 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 (3.6) gives the dual problem (3.3). Finally, for improved readability, the detailed proof of the fact that there is no duality gap is deferred to Lemma A.2 in Appendix A. ∎

By using the change of variables implicit in (3.4b), problem (3.3) can be expressed equivalently as

$$\begin{aligned}\sup_{\substack{u_t\in\mathbb{R}_+^N,\, t\in\mathcal{V}\\ U_{t_1,t_2}\in\mathbb{R}_+^{N\times N},\,(t_1,t_2)\in\mathcal{E}}} \quad & -\epsilon\langle\mathbf{K},\mathbf{U}\rangle - \sum_{t\in\mathcal{V}}(g_t)^*\big(-\epsilon\log(u_t)\big)\\ & - \sum_{(t_1,t_2)\in\mathcal{E}} (f_{t_1,t_2})^*\big(-\epsilon\log(U_{t_1,t_2})\big).\end{aligned} \tag{3.7}$$

Moreover, under a Slater-type condition for the primal problem, i.e., that the relative interiors (denoted ri)[5] of the effective domains of the cost functions in (3.1) have a nonempty intersection, we have that the suprema in (3.3) and (3.7) are attained.

[5]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 [64, Chap. 6].

Assumption 3.7. Assume that there exists an $\mathbf{M} > 0$ such that $\langle \mathrm{C}, \mathrm{M} \rangle < \infty$, with marginals $(\mu_t)_{t\in\mathcal{V}}$ and bimarginals $(R_{t_1,t_2})_{(t_1,t_2)\in\mathcal{E}}$ satisfying (3.2b) and (3.2c), respectively, so that

- for all g_t and f_{t_1,t_2} that are polyhedral, $\mu_t \in \mathrm{dom}(g_t)$ and $R_{t_1,t_2} \in \mathrm{dom}(f_{t_1,t_2})$;
- for all g_t and f_{t_1,t_2} that are not polyhedral, $\mu_t \in \mathrm{ri}(\mathrm{dom}(g_t))$ and $R_{t_1,t_2} \in \mathrm{ri}(\mathrm{dom}(f_{t_1,t_2}))$.

COROLLARY 3.8. *Given Assumption* 3.7 *the conclusions of Theorem* 3.6 *hold, with the addition fact that the dual* (3.3) *is guaranteed to have a nonempty set of optimal solutions.*

Proof. The result follows from [64, Chaps. 29 and 30]. ∎

Even if the Slater-type condition in Assumption 3.7 is not fulfilled, the form $\mathbf{M} = \mathbf{K} \odot \mathbf{U}$ will be important in deriving a convergent algorithm for solving (3.2).

3.3. Coordinate ascent iterations for solving the dual problem. In this section we derive an efficient solution method for (3.2), based on performing coordinate ascent in the dual problem (3.3) (or, equivalently, in (3.7)). 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 (3.3). 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

$$\underset{\lambda_j\in\mathbb{R}^N}{\arg\max} \quad \phi(\lambda_j, (\lambda_t^k)_{t\in\mathcal{V}\setminus\{j\}}, (\Lambda_{t_1,t_2}^k)_{(t_1,t_2)\in\mathcal{E}}), \tag{3.8a}$$

or Λ_{j_1,j_2}^{k+1} to be in

$$\underset{\Lambda_{j_1,j_2}\in\mathbb{R}^{N\times N}}{\arg\max} \phi(\Lambda_{j_1,j_2}, (\lambda_t^k)_{t\in\mathcal{V}}, (\Lambda_{t_1,t_2}^k)_{(t_1,t_2)\in\mathcal{E}\setminus\{(j_1,j_2)\}}), \tag{3.8b}$$

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 the set of maximizing arguments in (3.8) to always be nonempty. To guarantee this, we impose the following assumption (which is milder than Assumption 3.7).

Assumption 3.9. Assume that $\mathrm{C} < \infty$. Moreover, assume that for each index $j \in \mathcal{V}$, there exists a $\mu_j > 0$ so that

- if g_j is polyhedral, then $\mu_j \in \mathrm{dom}(g_j)$;
- if g_j is not polyhedral, then $\mu_j \in \mathrm{ri}(\mathrm{dom}(g_j))$;

and that analogous assumptions hold for each index $(j_1, j_2) \in \mathcal{E}$, R_{j_1,j_2}, and f_{j_1,j_2}.

LEMMA 3.10. *Under Assumptions* 3.3 *and* 3.9*, the subproblems in* (3.8) *always have a nonempty set of maximizers.*

Proof. To prove the lemma, we restrict our attention to one subproblem of the form (3.8a); for subproblems of the form (3.8b) the proof follows analogously. Now, note that problem (3.8a) can be see as the Lagrangian dual of the primal problem

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

Moreover, using Assumption 3.9 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 [64, Chaps. 29 and 30] we have that strong duality holds between these two problems, and in particular that the dual (3.8a) has a nonempty set of maximizers (cf. [71, Lem. 3.1]). ◻

By the above lemma, the coordinate ascent steps in (3.8) 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

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

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

$$w_j := P_j(\mathbf{K} \odot \mathbf{U}) \oslash u_j \tag{3.9a}$$

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

$$W_{j_1,j_2} := P_{j_1,j_2}(\mathbf{K} \odot \mathbf{U}) \oslash U_{j_1,j_2}, \tag{3.9b}$$

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

Next, note that $\partial\langle \mathbf{K}, \mathbf{U}\rangle / \partial\lambda_j^{(i_j)} = -\exp\left(\lambda_j^{(i_j)}/\epsilon\right) w_j^{(i_j)} = -u_j^{(i_j)} w_j^{(i_j)}$ with $\mathbf{K}$ and $\mathbf{U}$ given as in (3.4) and w_j as in (3.9a). Thus, in each update of the variable λ_j one has to solve the inclusion problem

$$0 \in \partial_{\lambda_j}\phi = -\exp(\lambda_j/\epsilon) \odot w_j + \partial(g_j)^*(-\lambda_j), \tag{3.10a}$$

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

$$0 \in \partial_{\Lambda_{j_1,j_2}}\phi = -\exp(\Lambda_{j_1,j_2}/\epsilon) \odot W_{j_1,j_2} + \partial(f_{j_1,j_2})^*(-\Lambda_{j_1,j_2}). \tag{3.10b}$$

To verify that the two equalities in (3.10) hold, see, e.g., [4, Cor. 16.38]. 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

$$0 \in -u_j \odot w_j + \partial(g_j)^*\big(-\epsilon\log(u_j)\big), \tag{3.11a}$$

$$0 \in -U_{j_1,j_2} \odot W_{j_1,j_2} + \partial(f_{j_1,j_2})^*\big(-\epsilon\log(U_{j_1,j_2})\big). \tag{3.11b}$$

This is summarized in Algorithm 3.1. However, note that directly computing w_j and W_{j_1,j_2} needed in (3.11) by brute-force is computationally demanding and, effectively, numerically infeasible for large-scale problems. Therefore, from this perspective, Algorithm 3.1 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, 7, 27, 35, 36, 37, 38, 39, 67]. In particular, in section 4 we show how this is done for the application of multispecies potential mean field games (see Algorithm 4.1). Moreover, by storing and using intermediate results of eliminated modes, we can understand the procedure also as a message-passing scheme [39]. Finally, under relatively mild assumptions, Algorithm 3.1 is convergent in the following sense.

Algorithm 3.1 Generalized Sinkhorn method for solving (3.2).

1: 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 $(U_{t_1,t_2}^0)_{(t_1,t_2)\in\mathcal{E}}$.
2: $k=0$
3: **while** Not converged **do**
4: $\quad k=k+1$
5: $\quad$ **for** $j\in\mathcal{V}$ and $(j_1,j_2)\in\mathcal{E}$ **do**
6: $\qquad$ Update u_j^k by solving (3.11a) with w_j as in (3.9a).
7: $\qquad$ Update U_{j_1,j_2}^k by solving (3.11b) with W_{j_1,j_2} as in (3.9b).
8: $\quad$ **end for**
9: **end while**
10: **return** $(u_t^k)_{t\in\mathcal{V}}$ and $(U_{t_1,t_2}^k)_{(t_1,t_2)\in\mathcal{E}}$.

THEOREM 3.11. *Given Assumptions* 3.3 *and* 3.9, *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* 3.3 *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.*

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* 3.1 *at iteration* k, *and let* U^k *be the corresponding tensor as in* (3.4b). *Moreover, let* $\mathrm{M}^k=\mathrm{K}\odot\mathrm{U}^k$. *Then* $(\mathrm{M}^k)_k$ *is a bounded sequence that converges to the optimal solution to* (3.2). *Furthermore, if the set of optimal solutions to* (3.3) *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* (3.7).

Proof. To prove the theorem, let $h(\mathbf{M}):=\langle\mathbf{C},\mathbf{M}\rangle+\epsilon D(\mathbf{M})$, which is a strictly convex function (cf. [4, Ex. 9.35]). Moreover, $\mathrm{dom}(h)=\mathbb{R}_+^{N^\mathcal{T}}$ and hence is polyhedral. Next, we observe that h is cofinite, since the Fenchel conjugate of h is given by[6]

$$h^*(\mathbf{T})=-\epsilon\sum_{i_1,\dots,i_\mathcal{T}}\exp((\mathbf{T}^{(i_1\dots i_\mathcal{T})}-\mathbf{C}^{(i_1\dots i_\mathcal{T})})/\epsilon)-1=-\epsilon\langle\mathbf{K},\exp(\mathbf{T}/\epsilon)\rangle+N^\mathcal{T}\epsilon;$$

see [4, Ex. 13.2 and Prop. 13.23]. Therefore, following along the lines of [71, sect. 6], we have that $(\mathbf{M}^k)_k$ is a bounded sequence and that every cluster point is an optimal solution to (3.2). In particular, [71, Thm. 3.1] imposes some slightly stronger assumptions,[7] but it is readily checked in all places where these stronger assumptions are invoked that the same conclusions hold in this particular case under the weaker assumptions. For brevity, we omit the details of the modifications needed.

Since $(\mathbf{M}^k)_k$ is a bounded sequence and every cluster point is optimal to (3.2), by the uniqueness of the optimal solution $\mathbf{M}^\star$ the sequence must converge to it. To see this, note that since $(\mathbf{M}^k)_k$ is bounded, if it does not converge, then it must have at least two cluster points. This is a contradiction, since every cluster point must be optimal, and the optimal solution is unique. Finally, the last statement of the theorem follows similarly from [71, Thm. 3.1(b)]. ∎

[6] 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 3.6.

[7] More precisely, to directly apply the result in [71, Thm. 3.1], we must assume that the feasible point in Assumption 3.3 is such that $\mathbf{M}>0$; see [71, Assump. 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 3.14.

The above theorem guarantees convergence but does not guarantee how fast the iterates converge. In particular, in order to guarantee R-linear convergence we need to impose further assumptions on the functions involved.

THEOREM 3.12. *Given Assumption* 3.3*, 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* (3.2b) *and* (3.2c)*, respectively, and that all functions* g_t *and* f_{t_1,t_2} *are such that*

(i) *either the function is a polyhedral indicator function and* $\mu_t \in dom(g_t)$ *or* $R_{t_1,t_2} \in dom(f_{t_1,t_2})$*, respectively;*

(ii) *or the function is cofinite, essentially smooth, and 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 such that* $\mu_t \in int(dom(g_t))$ *or* $R_{t_1,t_2} \in int(dom(f_{t_1,t_2}))$*, 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* 3.1*, and let* $\mathrm{M}^k = \mathrm{K} \odot \mathrm{U}^k$*. Then* $\mathrm{M}^k \to \mathrm{M}^\star$ *at least R-linearly, where* $\mathrm{M}^\star$ *is the unique optimal solution to* (3.2)*, and the cost function in* (3.7)*, 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* (3.2) *at least R-linearly.*

Proof. Assume first that all functions are as in (ii). In this case, note that (3.2a) is separable in the different variables, and that E in [55, 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 [55, Thm 6.1] are directly applicable.

In the case when some of the functions are of the form 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 same arguments as before shows R-linear convergence of the algorithm. ∎

Remark 3.13. One assumption in Theorem 3.12 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 (3.10) and (3.11) are in fact equalities on the interior of the effective domain.

Example 3.14. To illustrate some of the differences between the results presented so far, here we consider a small bimarginal example. 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 for simplicity we have taken $\mathbf{C} = 0$ and $\epsilon = 1$. The two constraints together imply that $\mathbf{M}^{(1,2)} = 0$ for any feasible solution, and hence neither the conditions in Assumption 3.7 nor the ones in Theorem 3.12 are fulfilled. Nevertheless, the conditions in Assumption 3.3 are fulfilled, and hence the problem has a unique optimal solution (Lemma 3.4); the latter is given by

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

Moreover, the conditions in Assumption 3.9 are fulfilled, and hence each step in the algorithm is therefore well-defined (Lemma 3.10). In fact, the conditions in Theorem 3.11 are fulfilled, which guarantees that the dual ascent algorithm is converging to the optimal solution. 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 $(U^{(1,2)})^k$ diverges as $k \to \infty$. Finally, by evaluating $\|\mathbf{M}^k - \mathbf{M}^\star\|_2$ we i can see that in fact the iterates converge R-linearly, which indicates that there might be room for improvement with respect to the conditions in Theorem 3.12.

As a final remark, we note that Assumptions 3.7 and 3.9 both enforce that we must have $\mathbf{C} < \infty$—the first one implicitly and the second one explicitly. 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 particular this is the case for the example in section 4. Nevertheless, the assumptions can be weakened somewhat to accommodate for this, similarly to [36, sect. 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 $\mathrm{dom}(g_j)$ 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 also do so analogously for f_{j_1,j_2} and the bimarginals. From the perspective of Algorithm 3.1, 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 $u_j^{(i_j)} = 0$.

3.4. Extension to multiple costs on each marginal. In some problems, marginals and bimarginals can be associated with multiple functions, typically when they are both associated with a cost and an inequality constraint. To handle such cases, we consider a modified version of problem (3.2) that takes the form[8]

$$\begin{aligned} \underset{\substack{\mathbf{M}\in\mathbb{R}_+^{N^\mathcal{T}},\ \mu_{t,k_1}\in\mathbb{R}_+^N,\\ R_{t_1,t_2,k_2}\in\mathbb{R}_+^{N\times N}\\ t\in\mathcal{V} \text{ and } k_1=1,\dots,\kappa_1\\ (t_1,t_2)\in\mathcal{E} \text{ and } k_2=1,\dots,\kappa_2}}{\text{minimize}} \quad & \langle \mathbf{C},\mathbf{M}\rangle + \epsilon D(\mathbf{M}) + \sum_{t\in\mathcal{V}}\sum_{k_1=1}^{\kappa_1} g_{t,k_1}(\mu_{t,k_1}) \\ & + \sum_{(t_1,t_2)\in\mathcal{E}}\sum_{k_2=1}^{\kappa_2} f_{t_1,t_2,k_2}(R_{t_1,t_2,k_2}) \\ \text{subject to} \quad & P_t(\mathbf{M}) = \mu_{t,k_1},\ k_1 = 1,\dots,\kappa_1,\ t\in\mathcal{V} \\ & P_{t_1,t_2}(\mathbf{M}) = R_{t_1,t_2,k_2},\ k_2=1,\dots,\kappa_2,\ (t_1,t_2)\in\mathcal{E}. \end{aligned} \tag{3.12}$$

By modifying the arguments in the previous sections, it is straightforward to derive a Lagrangian dual of (3.12) and to see that if the dual problem has an optimal solution, then the optimal solution to (3.12) is of the form $\mathbf{M} = \mathbf{K}\odot\mathbf{U}$, where

[8]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 $\mu_{t,k_1} = \mu_{t,k_1'}$, for all $k_1, k_1' = 1,\dots,\kappa_1$ and all $t\in\mathcal{V}$, for any feasible point, and that similar relations hold for the bimarginals.

$$\mathbf{U}^{(i_1 \dots i_{\mathcal{T}})} = \left(\prod_{t \in \mathcal{V}} \prod_{k_1=1}^{\kappa_1} u_{t,k_1}^{(i_t)} \right) \left(\prod_{(t_1,t_2) \in \mathcal{E}} \prod_{k_2=1}^{\kappa_2} U_{t_1,t_2,k_2}^{(i_{t_1}, i_{t_2})} \right)$$

and where $\mathbf{K}$ is as before (see (3.4a)). This 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

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

where w_j is defined analogously to (3.9a) as

$$w_j = P_j(\mathbf{K} \odot \mathbf{U}) \oslash \left(\bigodot_{k_1=1}^{\kappa_1} u_{j,k_1} \right). \tag{3.13}$$

Similar expressions hold for the inclusion problem for $U_{t_1,t_2,\tilde{k}_2}$. Furthermore, reexamining the proof of Theorems 3.11 and 3.12, we can readily see that by modifying the assumptions accordingly, 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 [27, 36, 37], one can see that the efficiency in computing w_j in (3.13) (and also W_{j_1,j_2}) 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.

4. Multispecies 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 [11, 26, 42, 43, 44, 49]. 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 by a controlled Fokker–Planck equation with distributed control [49]. This type of control problem has been studied in, e.g., [9, 13, 17]. An important generalization of mean field games is the multispecies setting, where the population consists of several different types of agents or species [1, 10, 22, 43, 47, 49]. In this section, we show that discretizations of potential multispecies mean field games take the form of a convex graph-structured tensor optimization problem (3.2). By also deriving efficient methods for computing the corresponding projections needed in Algorithm 3.1, 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.

4.1. 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

$$dx(t) = f(x(t))dt + B(x(t))\big(v(x(t),t)dt + \sqrt{\epsilon} dw\big), \tag{4.1}$$

with initial condition $x(0) = x_0 \sim \rho_0(x)$, where w is an m-dimensional Wiener process. More precisely, assume that $f : X \to \mathbb{R}^n$ and $B : X \to \mathbb{R}^{n \times m}$ are continuously differentiable with bounded derivatives, in which case, under suitable conditions on the (Markovian) feedback v, there exists a unique solution to (4.1) almost surely; see, e.g, [30, Thm. V.4.1] and [9, pp. 7–8]. Moreover, under suitable regularity conditions [9, 13], the density $\rho(t,\cdot)$ which describes the distribution of particles at time point t exists and is the solution of a controlled Fokker–Planck equation (cf. [3, p. 72]). A potential mean field game can then be reformulated as the density optimal control problem [49]

$$\text{(4.2a)} \qquad \underset{\rho,v}{\text{minimize}} \ \int_0^1\int_X \frac{1}{2}\|v\|^2\rho dx dt + \int_0^1 \mathcal{F}_t(\rho(t,\cdot))dt + \mathcal{G}(\rho(1,\cdot))$$

$$\text{(4.2b)} \qquad \text{subject to } \frac{\partial \rho}{\partial t} + \nabla\cdot((f+Bv)\rho) - \frac{\epsilon}{2}\sum_{i,k=1}^{n}\frac{\partial^2(\sigma_{ik}\rho)}{\partial x_i \partial x_k} = 0,$$

$$\text{(4.2c)} \qquad \rho(0,\cdot) = \rho_0.$$

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 piecewise continuous with respect to t.

To discretize problem (4.2), 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 (4.1). 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 (4.2b) and (4.2c). Moreover, let $\mathcal{P}^0$ denote the corresponding (uncontrolled) Wiener process with initial density ρ_0. By the Girsanov theorem (see, e.g., [31, pp. 156–157] and [25, p. 321]), we have that

$$\text{(4.3)} \qquad \frac{1}{2}\int_X\int_0^1 \|v\|^2\rho dt dx = \frac{1}{2}\mathbb{E}_{\mathcal{P}^v}\left\{\int_0^1 \|v\|^2 dt\right\} = \epsilon \mathrm{KL}(\mathcal{P}^v\|\mathcal{P}^0)$$

where $\mathrm{KL}(\cdot\|\cdot)$ is the Kullback–Leibler divergence; see, e.g., [8, 14, 17, 33, 51, 52]. To ensure that (4.3) holds, it is important that the control signal and the noise enter the system through the same channel, as in (4.1) [15, 16]. Moreover, the link between stochastic control and entropy provided by (4.3) has recently led to several novel applications of optimal control [12, 14, 15, 16, 18].

By using (4.3), the problem (4.2) can be reformulated as

$$\text{(4.4a)} \qquad \underset{\mathcal{P}^v}{\text{minimize}} \quad \epsilon\,\mathrm{KL}(\mathcal{P}^v\|\mathcal{P}^0) + \int_0^1 \mathcal{F}_t(\mathcal{P}^v_t)dt + \mathcal{G}(\mathcal{P}^v_1)$$

$$\text{(4.4b)} \qquad \text{subject to} \quad \mathcal{P}^v_0 = \rho_0.$$

Next, we discretize this problem in both time and space. More precisely, discretizing over time into the time points $0, \Delta t, 2\Delta t, \ldots, 1$, where $\Delta t = 1/\mathcal{T}$, and over space into the grid points $x_1,\ldots,x_N$, we get that $\mathcal{P}^v$ becomes a tensor $\mathbf{M} \in \mathbb{R}_+^{N^{\mathcal{T}+1}}$, i.e., a nonnegative $(\mathcal{T}+1)$-mode tensor that represents the flow of the agents. More precisely, $\mathbf{M}^{(i_0\ldots i_\mathcal{T})}$ is the discrete approximation corresponding to the probability of a sample path that passes through the discrete states $x_{i_0},\ldots,x_{i_\mathcal{T}}$ at the corresponding discrete time instances. Similarly, $\mathcal{P}^0$ becomes a nonnegative $(\mathcal{T}+1)$-mode tensor of probabilities corresponding to the evolution of the (uncontrolled) Wiener process. For reasons that will be clear shortly, we call this tensor $\mathbf{K}$ and let $\mathbf{K}^{(i_0\ldots i_\mathcal{T})} = \gamma\exp(-\mathbf{C}^{(i_0\ldots i_\mathcal{T})}/\epsilon)$ for some tensor $\mathbf{C}$ and where $\gamma > 0$ is a normalizing constant so that $\sum_{i_0,\dots,i_\mathcal{T}} \mathbf{K}^{(i_0\dots i_\mathcal{T})} = 1$. The Kullback–Leibler divergence can then be discretized as

$$\begin{aligned}\epsilon \mathrm{KL}(\mathcal{P}^v \| \mathcal{P}^0) &\approx \epsilon \sum_{i_0,\dots,i_\mathcal{T}} \log\left(\frac{\mathbf{M}^{(i_0\dots i_\mathcal{T})}}{\mathbf{K}^{(i_0\dots i_\mathcal{T})}}\right) \mathbf{M}^{(i_0\dots i_\mathcal{T})} \\ &= \epsilon \sum_{i_0,\dots,i_\mathcal{T}} \log\left(\mathbf{M}^{(i_0\dots i_\mathcal{T})}\right) \mathbf{M}^{(i_0\dots i_\mathcal{T})} - \epsilon \sum_{i_0,\dots,i_\mathcal{T}} \log(\gamma) \mathbf{M}^{(i_0\dots i_\mathcal{T})} \\ &\quad - \epsilon \sum_{i_0,\dots,i_\mathcal{T}} \log\left((\exp(-\mathbf{C}^{(i_0\dots i_\mathcal{T})}/\epsilon)\right) \mathbf{M}^{(i_0\dots i_\mathcal{T})} \\ &= \epsilon D(\mathbf{M}) + \text{constant} - \epsilon \sum_{i_0,\dots,i_\mathcal{T}} \frac{-\mathbf{C}^{(i_0\dots i_\mathcal{T})}}{\epsilon} \mathbf{M}^{(i_0\dots i_\mathcal{T})} \\ &= \epsilon D(\mathbf{M}) + \text{constant} + \langle \mathbf{C}, \mathbf{M} \rangle.\end{aligned}$$

Discarding the constants, the discretized version of (4.4), and thus the discretized version of (4.2), therefore becomes

$$\begin{aligned} &\underset{\substack{\mathbf{M}\in\mathbb{R}_+^{N^{\mathcal{T}+1}} \\ \mu_1,\dots,\mu_\mathcal{T}\in\mathbb{R}_+^N}}{\text{minimize}} && \langle \mathbf{C}, \mathbf{M} \rangle + \epsilon D(\mathbf{M}) + \Delta t \sum_{j=1}^{\mathcal{T}-1} F_j(\mu_j) + G(\mu_\mathcal{T}) && \text{(4.5a)} \\ &\text{subject to} && P_j(\mathbf{M}) = \mu_j,\ j = 1, 2, \dots, \mathcal{T}, && \text{(4.5b)} \\ & && P_0(\mathbf{M}) = \mu_0. && \text{(4.5c)} \end{aligned}$$

Here, μ_0 is a discrete approximation of ρ_0, and μ_j is the distribution of agents at time point j. All that is left to do is to derive the form of the $(\mathcal{T}+1)$-mode tensor $\mathbf{C}$, which in (4.5) can be seen to correspond to a cost of moving agents. To this end, first note that since $\mathcal{P}^0$ is the probability distribution on path space corresponding to a time-homogeneous Markov process, we have that

$$\mathbf{K}^{(i_0\dots i_\mathcal{T})} = \prod_{j=0}^{\mathcal{T}-1} K^{(i_j, i_{j+1})}$$

where K is an $N \times N$ matrix defining the transition probabilities between the discrete states in one time step. This in turn means that the cost tensor takes the form

$$\mathbf{C}^{(i_0\dots i_\mathcal{T})} = \sum_{j=0}^{\mathcal{T}-1} C^{(i_j, i_{j+1})}, \tag{4.6a}$$

where C is an $N \times N$ matrix defining the transition costs between time points in (4.5). More precisely, we approximate the elements $C^{(i,k)}$ as the optimal cost of moving mass from discretization point x_i to discretization point x_k in one time step, given by

$$C^{(i,k)} = \begin{cases} \underset{v\in L_2([0,\Delta t])}{\text{minimize}} & \int_0^{\Delta t} \frac{1}{2}\|v\|^2 dt \\ \text{subject to} & \dot{x} = f(x) + B(x)v, \\ & x(0) = x_i, \quad x(\Delta t) = x_k. \end{cases} \tag{4.6b}$$

This approximation is motivated by the fact that for small time steps (corresponding to a small noise level in a time-rescaled version of the problem) there is a concentration of the probability around the trajectories that are solutions to the corresponding optimal control problem [40] (cf. [25, sect. 5], [16, sect. IV], and [70, Thm. 2]). Intuitively, this makes sense since for small time steps, the transitions are approximately Gaussian. However, the "distance" in the transition is no longer measured in the

Euclidean norm but instead by the optimal control cost, since a lower control cost means that the system is "easier" to steer between the two states, and hence it is more likely that the system will make that transition.

The optimal control problem (4.6b) typically cannot 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 (4.5).

Finally, note that the problem (4.5)–(4.6) is a convex graph-structured tensor optimization problem of the form (3.2) on a path graph. In order to guarantee that (4.5) has an optimal solution, one must guarantee that it 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 (4.5b) and (4.5c) and is such that (4.5a) is finite (cf. Assumption 3.3). A sufficient condition for this to hold is that the functions F_j, for $j = 1, \ldots, \mathcal{T} - 1$, and G are finite on all of $\mathbb{R}_+^N$, and that the elements (4.6b) are all finite. That latter is true if the deterministic counterpart to system (4.1) 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 $x(t) = x_1$. Two examples of classes of systems that have this property are controllable linear systems and systems where $B(x)$ is square and invertible for all x.

Remark 4.1. Another solution method for solving problems of the form (4.5), for agents that follow the dynamics of a first-order integrator, has been presented in [8]. The two methods are similar, and the main difference is that the computational method developed in [8] is based on a variable elimination technique, in contrast to the belief-propagation-type technique used here; see the discussion just before Theorem 3.11.

4.2. The multispecies problem. A multispecies 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 works [43, 49]. Here, we consider a multispecies 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

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

with initial condition $x_\ell(0) = x_{\ell,0} \sim \rho_{\ell,0}(x)$. Next, let $\rho_\ell(t,\cdot)$ denote the distribution of species ℓ at time point t, and note that a multispecies 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

$$\begin{aligned}
&\underset{\rho,\rho_\ell,v_\ell}{\text{minimize}} \quad \int_0^1\!\!\int_X \sum_{\ell=1}^L \frac{1}{2}\|v_\ell\|^2 \rho_\ell \, dx dt + \int_0^1 \mathcal{F}_t(\rho(t,\cdot))dt + \mathcal{G}(\rho(1,\cdot)) \\
&\qquad\qquad + \sum_{\ell=1}^L \left(\int_0^1 \mathcal{F}_{\ell,t}(\rho_\ell(t,\cdot))dt + \mathcal{G}_\ell(\rho_\ell(1,\cdot)) \right) && \text{(4.7a)}\\
&\text{subject to} \quad \frac{\partial \rho_\ell}{\partial t} + \nabla \cdot ((f(x) + B(x)v_\ell)\rho_\ell) \\
&\qquad\qquad - \frac{\epsilon}{2} \sum_{i,k=1}^n \frac{\partial^2(\sigma_{ik}\rho_\ell)}{\partial x_i \partial x_k} = 0, \quad \ell = 1, \ldots, L, && \text{(4.7b)}\\
&\qquad\qquad \rho_\ell(0,\cdot) = \rho_{\ell,0}, \quad \rho(t,x) = \sum_{\ell=1}^L \rho_\ell(t,x), && \text{(4.7c)}
\end{aligned}$$

where we impose the same assumptions on $\mathcal{F}_{\ell,t}$ 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 $\mathcal{F}_t \equiv 0$, $\mathcal{G} \equiv 0$, (4.7) reduces to L independent single-species problems. Moreover, the functionals $\int_0^1 \mathcal{F}_{\ell,t}(\cdot)dt$ and $\mathcal{G}_\ell(\cdot)$ are the ones that give rise to the heterogeneity among the species.

4.3. Numerical algorithm for solving the multispecies problem. To derive a numerical algorithm for solving (4.7), analogously to the single-species problem we first discretize the problem over time and space. To this end, by adapting the arguments in subsection 4.1, we arrive at the discrete problem

$$\begin{aligned}
\underset{\substack{\mathbf{M}_\ell,\, \mu_j,\, \mu_{\ell,j} \\ j=1,\dots,\mathcal{T} \\ \ell=1,\dots,L}}{\text{minimize}} \quad & \sum_{\ell=1}^{L} \big(\langle \mathbf{C}, \mathbf{M}_\ell\rangle + \epsilon D(\mathbf{M}_\ell)\big) + \Delta t \sum_{j=1}^{\mathcal{T}-1} F_j(\mu_j) + G(\mu_{\mathcal{T}}) \\
& + \sum_{\ell=1}^{L} \left(\Delta t \sum_{j=1}^{\mathcal{T}-1} F_{\ell,j}(\mu_{\ell,j}) + G_\ell(\mu_{\ell,\mathcal{T}}) \right) && \text{(4.8a)} \\
\text{subject to} \quad & P_j(\mathbf{M}_\ell) = \mu_{\ell,j},\ j=1,\dots,\mathcal{T},\ \ell=1,\dots,L, && \text{(4.8b)} \\
& P_0(\mathbf{M}_\ell) = \mu_{\ell,0},\ \ell=1,\dots,L, && \text{(4.8c)} \\
& \sum_{\ell=1}^{L} \mu_{\ell,j} = \mu_j,\ j=0,\dots,\mathcal{T}, && \text{(4.8d)}
\end{aligned}$$

where $\mathbf{C}$ still has the form (4.6), and where $\mu_{\ell,0}$ are discrete approximations of $\rho_{\ell,0}$. In particular, note that the second line in the cost (4.8a) is the discretization of the second line in (4.7a). Moreover, also note that (4.8) consists of L coupled graph-structured tensor optimization problems, coupled via the constraint (4.8d) and the cost imposed on μ_j, for $j=1,\dots,\mathcal{T}$, in (4.8a).

Next, we reformulate (4.8) into a single graph-structured tensor optimization problem (cf. [36]). 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 (4.8d), while the bimarginal projection $P_{-1,j}(\mathbf{M})$ gives the $L\times N$ matrix $[\mu_{1,j},\dots,\mu_{L,j}]^T$. By introducing

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

the constraint (4.8c) 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

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

and similarly for $\mathscr{G}^L$, the last term in the cost (4.8a) can be written as functions 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

$$\underset{\substack{\mathbf{M},\,\mu_j,\,R_{-1,j}\\ j=1,\ldots,\mathcal{T}}}{\text{minimize}}\ \langle \tilde{\mathbf{C}}, \mathbf{M}\rangle + \epsilon D(\mathbf{M}) + \Delta t \sum_{j=1}^{\mathcal{T}-1} F_j(\mu_j) + G(\mu_{\mathcal{T}}) + \sum_{j=1}^{\mathcal{T}-1} \mathscr{F}_j^L(R_{-1,j}) + \mathscr{G}^L(R_{-1,\mathcal{T}}) \tag{4.9a}$$

$$\text{subject to}\ \ P_j(\mathbf{M}) = \mu_j, \quad j=1,\ldots,\mathcal{T}, \tag{4.9b}$$

$$P_{-1,j}(\mathbf{M}) = R_{-1,j}, \quad j=1,\ldots,\mathcal{T}, \tag{4.9c}$$

$$P_{-1,0}(\mathbf{M}) = \mathfrak{R}_{-1,0}, \tag{4.9d}$$

where

$$\tilde{\mathbf{C}}^{(\ell i_0 \ldots i_{\mathcal{T}})} = \sum_{j=0}^{\mathcal{T}-1} C^{(i_j, i_{j+1})}. \tag{4.9e}$$

The problem (4.9) is readily seen to be a graph-structured tensor optimization problem of the form (3.2) and hence can be solved using Algorithm 3.1. In particular, the iterates of the transport plan produced by Algorithm 3.1 are of the form $\mathbf{M}^k = \mathbf{K} \odot \mathbf{U}^k$, where $\mathbf{K} = \exp(-\tilde{\mathbf{C}}/\epsilon)$ and

$$\mathbf{U}^{(\ell i_0 \ldots i_{\mathcal{T}})} = U_{-1,0}^{(\ell, i_0)} \prod_{j=1}^{\mathcal{T}} U_{-1,j}^{(\ell, i_j)} \prod_{j=1}^{\mathcal{T}} u_j^{(i_j)}. \tag{4.10}$$

The underlying graph-structure is illustrated in Figure 1, and by adapting the arguments in [36], marginal and bimarginal projections needed in the inclusion problems (3.11) can be computed efficiently as follows.

THEOREM 4.2. *Let* $\mathrm{K} = \exp(-\tilde{\mathrm{C}}/\epsilon)$, *with* $\tilde{\mathrm{C}}$ *defined as in* (4.9e) *and* $\epsilon > 0$, *and let* U *be as in* (4.10). *Define* $K = \exp(-C/\epsilon)$, *and let*

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

and

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

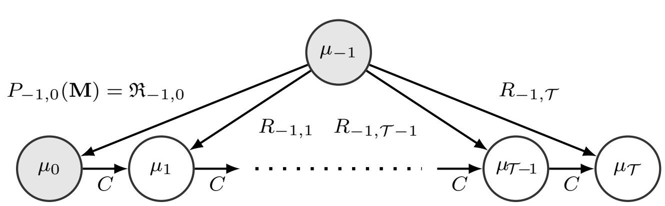


FIG. 1. *Illustration of the graph $\mathcal{G}$ for the multispecies density optimal control problem. Gray circles correspond to known densities, and white circles correspond to densities which are to be optimized over.*

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

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

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

Proof. See Appendix A. ∎

Finally, using Theorem 4.2 and specializing Algorithm 3.1 to solving the particular problem (4.9), an algorithm for solving discretized multispecies potential mean field games is given in Algorithm 4.1.

Remark 4.3. The algorithms in [63] are special instances of Algorithm 4.1. In particular, if $\mathscr{F}_j^L(\cdot) = \langle C_j, \cdot\rangle$ for some $C_j \in \mathbb{R}^{L\times N}$, then $(\mathscr{F}_j^L)^*(\cdot) = \iota_{\{C_j\}}(\cdot)$. 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 [63, Alg. 1]. On the other hand, if $\mathscr{G}^L(\cdot) = \iota_{\{\mathfrak{R}_{-1,\mathcal{T}}\}}(\cdot)$ 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 hence can be removed. Moreover, $(\mathscr{G}^L)^*(\cdot) = \langle \mathfrak{R}_{-1,\mathcal{T}}, \cdot\rangle$, from which we recover [63, Alg. 2].

4.4. Numerical example. In this section we demonstrate Algorithm 4.1 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

Algorithm 4.1 Method for solving the multispecies potential mean field game (4.9).

Given: Initial guess $u_1, \dots, u_{\mathcal{T}}$, $U_{-1,0}, \dots, U_{-1,\mathcal{T}}$
while Not converged **do**
  $\Psi_{\mathcal{T}-1} \leftarrow U_{-1,\mathcal{T}}\,\mathrm{diag}(u_{\mathcal{T}})K^T$
  for $j = \mathcal{T} - 2, \dots, 0$ **do**
    $\Psi_j \leftarrow (\Psi_{j+1} \odot U_{-1,j+1})\mathrm{diag}(u_{j+1})K^T$
  end for
  $U_{-1,0} \leftarrow \mathfrak{R}_{-1,0} \oslash \Psi_0$
  $\hat{\Psi}_1 \leftarrow U_{-1,0}K$
  for $j = 1, \dots, \mathcal{T} - 1$ **do**
    $W_{-1,j} \leftarrow (\hat{\Psi}_j \odot \Psi_j)\mathrm{diag}(u_j)$, and update $U_{-1,j}$ by solving (3.11b)
    $w_j \leftarrow (\hat{\Psi}_j \odot \Psi_j \odot U_{-1,j})^T\mathbf{1}$, and update u_j by solving (3.11a)
    $\hat{\Psi}_{j+1} \leftarrow (\hat{\Psi}_j \odot U_{-1,j})\mathrm{diag}(u_j)K$
  end for
  $W_{-1,\mathcal{T}} \leftarrow \hat{\Psi}_{\mathcal{T}}\mathrm{diag}(u_{\mathcal{T}})$, and update $U_{-1,\mathcal{T}}$ by solving (3.11b)
  $w_{\mathcal{T}} \leftarrow (\hat{\Psi}_{\mathcal{T}} \odot U_{-1,\mathcal{T}})^T\mathbf{1}$, and update $u_{\mathcal{T}}$ by solving (3.11a)
end while
return $u_1, \dots, u_{\mathcal{T}}$, $U_{-1,0}, \dots, U_{-1,\mathcal{T}}$

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 (4.6b), 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

$$\begin{aligned}
&\underset{\substack{\mathbf{M}_\ell \in \mathbb{R}_+^{(100^2)^{40}}, \\ \mu_{\ell,j} \in \mathbb{R}_+^{100^2} \\ j=1,\dots,39,\ \ell=1,2,3,4}}{\text{minimize}} && \sum_{\ell=1}^{4} \Big(\langle \mathbf{C}, \mathbf{M}_\ell \rangle + \epsilon D(\mathbf{M}_\ell) \Big) + \sum_{j=1}^{39} \langle c_3, \mu_{3,j} \rangle \\
&\text{(4.11a)} && + 0.1 \sum_{j=1}^{39} \|\mu_{4,j} - \tilde{\nu}\|_2^2 + 3\|\mu_{19} - \tilde{\mu}_1\|_2^2 + 3\|\mu_{39} - \tilde{\mu}_2\|_2^2 \\
&\text{(4.11b)} \quad \text{subject to} && P_j(\mathbf{M}_\ell) = \mu_{\ell,j}, \quad j = 0, \dots, 39,\ \ell = 1,2,3,4, \\
&\text{(4.11c)} && \sum_{\ell=1}^{4} \mu_{\ell,j} = \mu_j, \quad j = 0, \dots, 39, \\
&\text{(4.11d)} && \mu_j \le \kappa_j, \quad j = 1, \dots, 39, \\
&\text{(4.11e)} && \mu_{1,j} \le \tilde{\kappa}, \quad j = 1, \dots, 39.
\end{aligned}$$

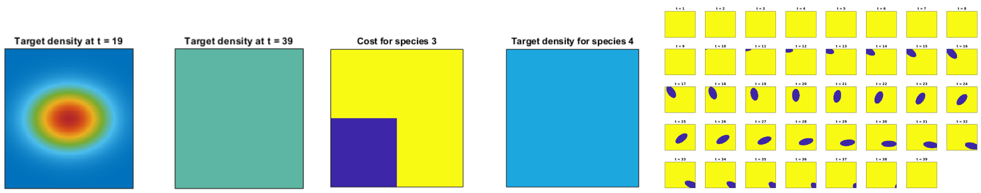
Here, $\tilde{\mu}_1$ and $\tilde{\mu}_2$ are the two distributions given in Figure 2(a). Moreover, the linear cost c_3, associated with species 3, and the target distribution $\tilde{\nu}$, associated with species 4, are both given in Figure 2(b).[9] Finally, for the capacity constraint (4.11d), κ_j is illustrated in Figure 2(c), while for the capacity constraint (4.11e), $\tilde{\kappa}$ is zero in the lower half of the domain and infinite for the upper half.

The graph-structured tensor optimization reformulation of (4.11) was solved using Algorithm 4.1. The latter is adapted as in section 3.4 to handle both the costs on the total marginals in (4.11a) and the inequality constraints in (4.11d); details on the Fenchel conjugates of the functions involved can be found in Appendix B. Results are shown in Figure 2(d), where the initial distributions $\mu_{\ell,0}$ for the different agents can be seen in the leftmost column (showing time point $j = 0$).

5. 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. This framework can also be used to solve convex multicommodity dynamic network flow problems akin to the ones studied in [36]. Moreover, we believe that these methods are useful for addressing many other types of problems, e.g., flow problems where the nodes or edges also have dynamics (cf. [23]). Moreover, we also believe that these methods can be extended to handle, e.g., multispecies potential mean field games where each species also has different dynamics.

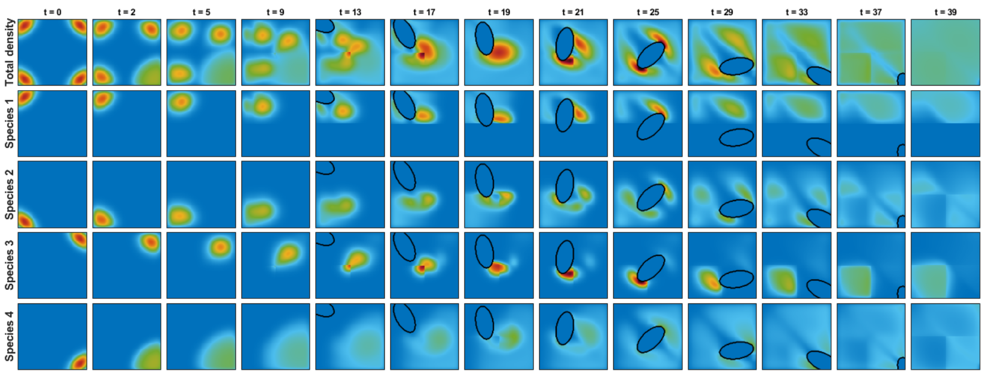
[9]Note that $\tilde{\mu}_2$ and $\tilde{\nu}$ are uniform distributions. The former has the same total mass as the total distribution μ_0, and the latter has the same as $\mu_{4,0}$.



(a) Target densities for total density.

(b) Species-dependent cost and constraint.

(c) Illustration of the capacity constraints.



(d) Optimal solution.

Fig. 2. *Figures describing the setup in the numerical example in section* 4.4. (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) *Illustration of species-dependent cost and constraint: 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* $\tilde{\nu}$ *for species* 4*.* (c) *The capacity constraint* κ_j *at the different time points* j*: blue means zero capacity (obstacle) while yellow means infinite capacity.* (d) *The optimal solution, illustrated as time evolution of total density and densities of the individual species. (See online version for color.)*

Appendix A. Deferred proofs.

Proof of Lemma 3.4. By Assumption 3.3, there is a feasible point to problem (3.2) with finite objective function value, and since problems (3.2) and (3.1) are equivalent, this means that the objective function in (3.1) 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, since the entropy term $\epsilon D(\mathbf{M})$ is radially unbounded and grows faster than linearly towards ∞, we therefore have that the objective function in (3.1) is radially unbounded. Since the entire objective function is also proper, convex, and lower-semicontinuous, the minimum is attained [64, Thm. 27.2], and it is unique since $D(\mathbf{M})$ (and hence the entire objective function in (3.1)) is strictly convex. ∎

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 the function f is proper, convex, and lower-semicontinuous, so is the function 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* (3.2) *and* (3.3).

Proof. To prove the lemma, we derive a Lagrangian dual of an equivalent problem to (3.3) and show that for the latter, strong duality holds with (3.2). To this end, note that a problem with the same set of globally optimal solutions as (3.3) is the constrained optimization problem

$$\sup_{\mathbf{U},\lambda,\Lambda} \; -\epsilon\langle \mathbf{K},\mathbf{U}\rangle - \sum_{t\in\mathcal{V}} (g_t)^*(-\lambda_t) - \sum_{(t_1,t_2)\in\mathcal{E}} (f_{t_1,t_2})^*(-\Lambda_{t_1,t_2})$$

$$\text{subject to} \quad \log(\mathbf{U}^{(i_1\dots i_\mathcal{T})}) = \frac{1}{\epsilon}\left(\sum_{t\in\mathcal{V}} \lambda_t^{(i_t)} + \sum_{(t_1,t_2)\in\mathcal{E}} \Lambda_{t_1,t_2}^{(i_{t_1},i_{t_2})}\right).$$

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 to 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

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

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

$$\sup_{\lambda_t\in\mathbb{R}^N} -(g_t)^*(-\lambda_t) - \langle 1/\epsilon P_t(\mathbf{Q}),\lambda_t\rangle = (g_t)^{**}(1/\epsilon P_t(\mathbf{Q})) = g_t(1/\epsilon P_t(\mathbf{Q})),$$

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

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

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 (3.1). Since (3.1) has the same optimal value as (3.2), the result follows. ∎

Proof of Theorem 4.2. 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 (4.10), this means that

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

where

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

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

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. ∎

Appendix B. Fenchel conjugates of some functions. In all the examples below, let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$.

Example B.1. Let $\alpha, \beta \in \bar{\mathbb{R}}$, $\alpha_i \le \beta_i$, for $i = 1, \dots, n$, and $[\alpha, \beta] := \{y \in \mathbb{R}^n \mid \alpha^{(i)} \le y^{(i)} \le \beta^{(i)}, i = 1, \dots, n\}$. For a set $A \subset \mathbb{R}$, let $\mathcal{I}_A$ be the characteristic function $\mathcal{I}_A(x) = 1$ if $x \in A$ and 0 otherwise. The Fenchel conjugate of $f(x) = \iota_{[\alpha,\beta]}(x)$ is $f^*(x^*) = \sum_{i=1}^n \left((x^*)^{(i)} \beta^{(i)} \mathcal{I}_{\mathbb{R}_+}((x^*)^{(i)}) + (x^*)^{(i)} \alpha^{(i)} \mathcal{I}_{\mathbb{R}_-}((x^*)^{(i)}) \right)$.

Example B.2. Let $p \in (1, \infty)$, let $\sigma > 0$, and let $y \in \mathbb{R}^n$. The Fenchel conjugate of $f(x) = \sigma \|x - y\|_p^p$ is $f^*(x^*) = \langle x^*, y \rangle + \frac{1}{q\,\sigma^{q-1} p^{q-1}} \|x^*\|_q^q$, where $1/p + 1/q = 1$.

Example B.3. Let $\beta \in \mathbb{R}^n$, and let $\beta_i > 0$ for $i = 1, \dots, n$. The Fenchel conjugate of $f(x) = x \oslash (\beta - x) + \iota_{[0,\beta]}(x)$ is $f^*(x^*) = \sum_{i=1}^n f_i^*(x_i^*)$, where $f_i^*(x_i^*) = 0$ if $x_i^* \le 1/\beta_i$ and $f_i^*(x_i^*) = x_i^* \beta_i - 2\sqrt{x_i^* \beta_i} + 1$ if $x_i^* > 1/\beta_i$.

REFERENCES

[1] Y. Achdou, M. Bardi, and M. Cirant, *Mean field games models of segregation*, Math. Models Methods Appl. Sci., 27 (2017), pp. 75–113.

[2] J. M. Altschuler and E. Boix-Adsera, *Polynomial-time algorithms for multimarginal optimal transport problems with structure*, Math. Program., 199 (2023), pp. 1107–1178.

[3] K. Åström, *Introduction to Stochastic Control Theory*, Dover, Mineola, NY, 2006, Unabridged republication of original published by Academic Press, New York, 1970.

[4] H. Bauschke and P. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2nd ed., Springer, Cham, 2017.

[5] F. Beier, J. von Lindheim, S. Neumayer, and G. Steidl, *Unbalanced multi-marginal optimal transport*, J. Math. Imaging Vision, 65 (2023), pp. 394–413.

[6] J.-D. Benamou and Y. Brenier, *A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem*, Numer. Math., 84 (2000), pp. 375–393.

[7] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré, *Iterative Bregman projections for regularized transportation problems*, SIAM J. Sci. Comput., 37 (2015), pp. A1111–A1138, https://doi.org/10.1137/141000439.

[8] J.-D. Benamou, G. Carlier, S. Di Marino, and L. Nenna, *An entropy minimization approach to second-order variational mean-field games*, Math. Models Methods Appl. Sci., 29 (2019), pp. 1553–1583.
[9] A. Bensoussan, J. Frehse, and P. Yam, *Mean Field Games and Mean Field Type Control Theory*, Springer, New York, 2013.
[10] A. Bensoussan, T. Huang, and M. Laurière, *Mean field control and mean field game models with several populations*, Minimax Theory Appl., 3 (2018), pp. 173–209.
[11] P. Caines, M. Huang, and R. Malhamé, *Mean field games*, in Handbook of Dynamic Game Theory, T. Başar and G. Zaccour, eds., Springer, Cham, 2018, pp. 345–372.
[12] K. F. Caluya and A. Halder, *Wasserstein proximal algorithms for the Schrödinger bridge problem: Density control with nonlinear drift*, IEEE Trans. Automat. Control, 67 (2021), pp. 1163–1178.
[13] P. Cardaliaguet, P. Graber, A. Porretta, and D. Tonon, *Second order mean field games with degenerate diffusion and local coupling*, NoDEA Nonlinear Differential Equations Appl., 22 (2015), pp. 1287–1317.
[14] Y. Chen, T. Georgiou, and M. Pavon, *On the relation between optimal transport and Schrödinger bridges: A stochastic control viewpoint*, J. Optim. Theory Appl., 169 (2016), pp. 671–691.
[15] Y. Chen, T. Georgiou, and M. Pavon, *Optimal steering of a linear stochastic system to a final probability distribution: Part* I, IEEE Trans. Automat. Control, 61 (2016), pp. 1158–1169.
[16] Y. Chen, T. Georgiou, and M. Pavon, *Optimal transport over a linear dynamical system*, IEEE Trans. Automat. Control, 62 (2017), pp. 2137–2152.
[17] Y. Chen, T. Georgiou, and M. Pavon, *Steering the distribution of agents in mean-field games system*, J. Optim. Theory Appl., 179 (2018), pp. 332–357.
[18] Y. Chen, T. T. Georgiou, and M. Pavon, *Stochastic control liaisons: Richard Sinkhorn meets Gaspard Monge on a Schrödinger bridge*, SIAM Rev., 63 (2021), pp. 249–313, https://doi.org/10.1137/20M1339982.
[19] Y. Chen and J. Karlsson, *State tracking of linear ensembles via optimal mass transport*, IEEE Control Syst. Lett., 2 (2018), pp. 260–265.
[20] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard, *Scaling algorithms for unbalanced optimal transport problems*, Math. Comp., 87 (2018), pp. 2563–2609.
[21] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard, *Unbalanced optimal transport: Dynamic and Kantorovich formulations*, J. Funct. Anal., 274 (2018), pp. 3090–3123.
[22] M. Cirant, *Multi-population mean field games systems with Neumann boundary conditions*, J. Math. Pures Appl., 103 (2015), pp. 1294–1315.
[23] G. Como, *On resilient control of dynamical flow networks*, Annu. Rev. Control, 43 (2017), pp. 80–90.
[24] M. Cuturi, *Sinkhorn distances: Lightspeed computation of optimal transport*, in Advances in Neural Information Processing Systems 26 (NIPS 2013), Curran Associates, Red Hook, NY, 2013, pp. 2292–2300.
[25] P. Dai Pra, *A stochastic control approach to reciprocal diffusion processes*, Appl. Math. Optim., 23 (1991), pp. 313–329.
[26] B. Djehiche, A. Tcheukam, and H. Tembine, *Mean-field-type games in engineering*, AIMS Electron. Electr. Eng., 1 (2017), pp. 18–73.
[27] 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 Process., 171 (2020), 107474.
[28] J. Fan, I. Haasler, J. Karlsson, and Y. Chen, *On the complexity of the optimal transport problem with graph-structured cost*, in Proceedings of the 25th International Conference on Artificial Intelligence and Statistics, Proc. Mach. Learn. Res. 151, PMLR, 2022, pp. 9147–9165.
[29] H. Farhangi, *The path of the smart grid*, IEEE Power Energy Mag., 8 (2010), pp. 18–28.
[30] W. Fleming and R. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.
[31] H. Föllmer, *Random fields and diffusion processes*, in École d'Été de Probabilités de Saint-Flour XV–XVII, 1985-87, Lecture Notes in Math. 1362, P.-L. Hennequin, ed., Springer, Berlin, Heidelberg, 1988, pp. 101–203.

[32] W. Gangbo and A. Świkech, *Optimal maps for the multidimensional Monge-Kantorovich problem*, Comm. Pure Appl. Math., 51 (1998), pp. 23–45.
[33] I. Gentil, C. Léonard, and L. Ripani, *About the analogy between optimal transport and minimal entropy*, Ann. Fac. Sci. Toulouse Math., 26 (2017), pp. 569–600.
[34] T. T. Georgiou, J. Karlsson, and M. S. Takyar, *Metrics for power spectra: An axiomatic approach*, IEEE Trans. Signal Process., 57 (2008), pp. 859–867.
[35] I. Haasler, Y. Chen, and J. Karlsson, *Optimal steering of ensembles with origin-destination constraints*, IEEE Control Syst. Lett., 5 (2020), pp. 881–886.
[36] I. Haasler, A. Ringh, Y. Chen, and J. Karlsson, *Scalable computation of dynamic flow problems via multimarginal graph-structured optimal transport*, Math. Oper. Res., 49 (2024), pp. 986–1011, https://doi.org/10.1287/moor.2021.0148.
[37] I. Haasler, A. Ringh, Y. Chen, and J. Karlsson, *Multimarginal optimal transport with a tree-structured cost and the Schrödinger bridge problem*, SIAM J. Control Optim., 59 (2021), pp. 2428–2453, https://doi.org/10.1137/20M1320195.
[38] I. Haasler, A. Ringh, and J. Karlsson, *Control and estimation of ensembles via structured optimal transport: A computational approach based on entropy-regularized multimarginal optimal transport*, IEEE Control Syst. Mag., 41 (2021), pp. 50–69.
[39] I. Haasler, R. Singh, Q. Zhang, J. Karlsson, and Y. Chen, *Multi-marginal optimal transport and probabilistic graphical models*, IEEE Trans. Inform. Theory, 67 (2021), pp. 4647–4668.
[40] O. Hijab, *Asymptotic Bayesian estimation of a first order equation with small diffusion*, Ann. Probab., 12 (1984), pp. 890–902.
[41] A. Hindawi, J.-B. Pomet, and L. Rifford, *Mass transportation with LQ cost functions*, Acta Appl. Math., 113 (2011), pp. 215–229.
[42] M. Huang, P. Caines, and R. Malhamé, *Social optima in mean field LQG control: Centralized and decentralized strategies*, IEEE Trans. Automat. Control, 57 (2012), pp. 1736–1751.
[43] M. Huang, R. Malhamé, and P. Caines, *Large population stochastic dynamic games: Closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle*, Commun. Inf. Syst., 6 (2006), pp. 221–252.
[44] B. Jovanovic and R. Rosenthal, *Anonymous sequential games*, J. Math. Econ., 17 (1988), pp. 77–87.
[45] J. Karlsson and A. Ringh, *Generalized Sinkhorn iterations for regularizing inverse problems using optimal mass transport*, SIAM J. Imaging Sci., 10 (2017), pp. 1935–1962, https://doi.org/10.1137/17M111208X.
[46] V. Krishnan and S. Martínez, *Distributed optimal transport for the deployment of swarms*, in Proceedings of the 2018 IEEE Conference on Decision and Control (CDC), Miami, FL, 2018, pp. 4583–4588.
[47] A. Lachapelle and M.-T. Wolfram, *On a mean field game approach modeling congestion and aversion in pedestrian crowds*, Trans. Res. Part B Methodol., 45 (2011), pp. 1572–1589.
[48] B. Lamond and N. Stewart, *Bregman's balancing method*, Trans. Res. Part B Methodol., 15 (1981), pp. 239–248.
[49] J.-M. Lasry and P.-L. Lions, *Mean field games*, Jpn. J. Math., 2 (2007), pp. 229–260.
[50] W. Lee, S. Liu, H. Tembine, W. Li, and S. Osher, *Controlling propagation of epidemics via mean-field control*, SIAM J. Appl. Math., 81 (2021), pp. 190–207, https://doi.org/10.1137/20M1342690.
[51] C. Léonard, *From the Schrödinger problem to the Monge–Kantorovich problem*, J. Funct. Anal., 262 (2012), pp. 1879–1920.
[52] C. Léonard, *A survey of the Schrödinger problem and some of its connections with optimal transport*, Discrete Contin. Dyn. Syst. A, 34 (2014), pp. 1533–1574.
[53] M. Liero, A. Mielke, and G. Savaré, *Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures*, Invent. Math., 211 (2018), pp. 969–1117.
[54] T. Lin, N. Ho, M. Cuturi, and M. Jordan, *On the complexity of approximating multimarginal optimal transport*, J. Mach. Learn. Res., 23 (2022), pp. 2835–2877.
[55] Z.-Q. Luo and P. Tseng, *On the convergence rate of dual ascent methods for linearly constrained convex minimization*, Math. Oper. Res., 18 (1993), pp. 846–867.
[56] G. Meyer and S. Beiker, eds., *Road Vehicle Automation*, Springer, Cham, 2014.
[57] L. Nenna, *Numerical Methods for Multi-marginal Optimal Transportation*, Ph.D. thesis, PSL Research University, 2016.
[58] J. Nocedal and S. Wright, *Numerical Optimization*, 2nd ed., Springer, New York, 2006.

[59] J. Ortega and W. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Classics in Appl. Math., 30, SIAM, Philadelphia, 2000. An unabridged republication of the work first published by Academic Press, New York, 1970.
[60] B. Pass, *Multi-marginal optimal transport: Theory and applications*, ESAIM Math. Model. Numer. Anal., 49 (2015), pp. 1771–1790.
[61] G. Peyré and M. Cuturi, *Computational optimal transport: With applications to data science*, Found. Trends Mach. Learn., 11 (2019), pp. 355–607.
[62] B. Piccoli and F. Rossi, *Generalized Wasserstein distance and its application to transport equations with source*, Arch. Ration. Mech. Anal., 211 (2014), pp. 335–358.
[63] A. Ringh, I. Haasler, Y. Chen, and J. Karlsson, *Efficient computations of multi-species mean field games via graph-structured optimal transport*, in Proceedings of the 2021 60th IEEE Conference on Decision and Control (CDC), Austin, TX, 2021, pp. 5261–5268.
[64] R. Rockafellar, *Convex Analysis*, Princeton Math. Ser., Princeton University Press, Princeton, NJ, 1970.
[65] L. Rüschendorf, *Optimal solutions of multivariate coupling problems*, Appl. Math., 23 (1995), pp. 325–338.
[66] L. Rüschendorf and L. Uckelmann, *On the n-coupling problem*, J. Multivariate Anal., 81 (2002), pp. 242–258.
[67] R. Singh, I. Haasler, Q. Zhang, J. Karlsson, and Y. Chen, *Inference with aggregate data in probabilistic graphical models: An optimal transport approach*, IEEE Trans. Automat. Control, 67 (2022), pp. 4483–4497.
[68] C. Sinigaglia, S. Bandyopadhyay, M. Quadrelli, and F. Braghin, *Optimal-transport-based control of particle swarms for orbiting rainbows concept*, J. Guidance Control Dyn., 44 (2021), pp. 2108–2117.
[69] Y. Teh and M. Welling, *The unified propagation and scaling algorithm*, in Advances in Neural Information Processing Systems 14 (NIPS 2001), MIT Press, Cambridge, MA, 2002, pp. 953–960.
[70] E. Todorov, *General duality between optimal control and estimation*, in Proceedings of the 2008 47th IEEE Conference on Decision and Control (CDC), Cancun, Mexico, 2008, pp. 4286–4292.
[71] P. Tseng, *Dual coordinate ascent methods for non-strictly convex minimization*, Math. Program., 59 (1993), pp. 231–247.
[72] C. Villani, *Topics in Optimal Transportation*, American Mathematical Society, Providence, RI, 2003.
[73] D. J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world' networks*, Nature, 393 (1998), 440.