ESAIM: COCV 32 (2026) 45
https://doi.org/10.1051/cocv/2026026

ESAIM: Control, Optimisation and Calculus of Variations
www.esaim-cocv.org

MEAN FIELD OPTIMIZATION PROBLEMS: STABILITY RESULTS AND LAGRANGIAN DISCRETIZATION

Kang Liu[1,*] and Laurent Pfeiffer[2]

Abstract. We formulate and investigate a convex optimization problem defined on a set of probability measures μ with prescribed marginal m, which we call Mean Field Optimization (MFO) problem. The cost function depends on an aggregate term, defined as the expectation of μ with respect to a contribution function. This problem is of particular interest in the context of Lagrangian potential mean field games and their discretization. We provide a first-order optimality condition and prove strong duality. We investigate stability properties of the MFO problem with respect to the prescribed marginal, from both primal and dual perspectives. In our stability analysis, we propose a method for recovering an approximate solution to an MFO problem with the help of an approximate solution to an MFO with a different marginal m, typically an empirical distribution. We combine this method with the stochastic Frank–Wolfe algorithm of [Bonnans *et al. SIAM J. Optim.* **33** (2023) 3083–3113] to derive a complete resolution method.

Mathematics Subject Classification. 49N80, 49Q22, 90C31, 90C46.

Received January 10, 2025. Accepted April 1, 2026.

1. Introduction

This article is dedicated to a general class of optimization problems involving probability measures with a prescribed marginal m. We will refer to them as Mean Field Optimization (MFO) problems. They typically arise in multi-agent optimization problems, for which a mean-field formulation of the problem, involving the probability distribution of the decisions of the agents (rather than an enumeration of them), is not only meaningful but also provides us with convexity properties of great numerical interest.

The first ambition of our work is to provide a general framework for the formulation of such situations. For the sake of clarity, we introduce here the MFO problems investigated in this work. We refer the reader to Section 2.3 for a complete description of the required assumptions. Let X and Y be two complete and separable metric spaces and let $\mathcal{H}$ be a separable Hilbert space. Let Z be a closed subset of $X \times Y$ and let m be a probability measure on X. We consider the following problem, parametrized by m:

$$\inf_{\mu \in \mathcal{P}_m(Z)} \; f\left(\int_Z g(x,y) d\mu(x,y) \right), \tag{$\mathrm{P}_m^{\mathrm{K}}$}$$

Keywords and phrases: Optimization with probability measures, potential mean field games, non-atomic games, stability analysis, Frank–Wolfe algorithm.

[1] Université Bourgogne Europe, CNRS, Institut de Mathématiques de Bourgogne, 21000, Dijon, France.
[2] Université Paris-Saclay, CNRS, CentraleSupélec, Inria, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France.

* Corresponding author: kang.liu@u-bourgogne.fr

where $g\colon Z \to \mathcal{H}$ is a Borel measurable function and $f\colon \mathcal{H} \to \mathbb{R}$ is a convex function. The admissible set $\mathcal{P}_m(Z)$ is the set of all probability measures on Z whose marginal distribution on X is m. A rigorous definition of $\mathcal{P}_m(Z)$ is provided in Section 2.3. Our model allows for heterogeneity within the agents, which is modeled by some parameter $x \in X$. The decision variables of the agents are generically denoted by $y \in Y$ and the probability measure μ represents the distribution of the parameter-decision pairs (x, y) of our agents. The distribution of the parameters of the agents is given by m, that is why we impose that μ has its first marginal equal to m. At an abstract level, we can interpret the term $\int_Z g d\mu$ as a common good, obtained by aggregating the contributions of all the agents.

Motivation

Our original interest for MFO problems comes from non-atomic games with a convex potential structure, for which finding a Nash Equilibrium (NE) is equivalent to solving an MFO problem. Indeed, the first-order optimality condition (Cor. 3.4) derived in our abstract framework is consistent with the standard characterization of NE in [1–3], while covering a more general class of models thanks to our abstract framework. Concrete examples with be provided and discussed in Section 8. Among these games, we have a special interest for Lagrangian Mean Field Games (MFGs) [1, 3–6], in which the agents each optimize the trajectory of some dynamical system and are parametrized by their initial condition. A concrete formulation of such Lagrangian MFGs is presented in Section 8.2. Problem ($\mathrm{P}^{\mathrm{K}}_m$) also arises in energy management problems, more specifically, in problems involving many small consumption (or production) units, for example electrical cars. For such problems, the parameter x could model any relevant characteristic of the cars (such as their charging capacity) while the variable y describes their charging profiles. The common good is typically the total energy consumption on a given time interval and the function f may be designed so as to penalize the distance to a reference consumption profile. We refer the reader to [7–9] for references containing concrete models. Note that in these references, the problem appears in the Monge form discussed below, with m an empirical measure.

Finally, let us mention the close connection between MFO problems and supervised learning through the training of neural networks with one hidden layer. In the mean-field (infinite-width) relaxation of this training problem [10–12], the variable μ describes the probability distribution of the neurons' weights. In this setting, the marginal space X essentially reduces to a single point, so the corresponding MFO formulation is considerably simpler. Regarding numerical algorithms, the methods (Alg. 3) considered in this article can be compared with classical training procedures, such as stochastic gradient descent, used in the above references. A systematic comparison appears particularly interesting and is left for future work. For other applications of MFO problems in learning, we refer to [13, 14] and the references therein.

Numerical approach

The numerical resolution of Problem ($\mathrm{P}^{\mathrm{K}}_m$) poses two main difficulties: the numerical manipulation of probability measures and the treatment of the marginal constraint. Let us focus on the first difficulty by supposing momentarily that m is a Dirac measure located at some point x^* and that $Z = \{x^*\} \times Y$, so that the problem ($\mathrm{P}^{\mathrm{K}}_m$) can simply be written as an optimization problem on $\mathcal{P}(Y)$:

$$\inf_{\nu \in \mathcal{P}(Y)} \; f\left(\int_Y g(x^*, y) d\nu(y)\right).$$

Unless Y is finite, $\mathcal{P}(Y)$ is an infinite dimensional set. A first approach would consist in discretizing the set Y, which would preserve the convexity of the problem. This approach suffers from the curse of dimensionality, since it requires an exponential number of points with respect to the dimension of Y. A second approach would consist in representing the probability measure as the empirical mean of a set of N points to be optimized. The clear drawback of this approach is the loss of convexity of the discretized problem; yet we mention that it proves efficient in the context of supervised learning problem [10]. In the general context of MFO problems, the Frank–Wolfe (FW) algorithm is a particularly advantageous algorithm. It produces a finitely supported

approximate solution and leverages the convexity of the original problem (without using a coarse discretization of Y). More specifically, it generates after k iterations an $\mathcal{O}(1/k)$-optimal solution supported by at most k points. The FW algorithm applies to our discretized problem below, as outlined in Algorithm 2. Moreover, it achieves an $\mathcal{O}(1/k)$-convergence rate, as shown in Lemma 7.5.

Coming back to the case of a general marginal, we propose to discretize m with an empirical measure m_N associated with N points, namely

$$m_N \coloneqq \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i},$$

and to solve:

$$\inf_{\mu \in \mathcal{P}_{m_N}(Z)} f\left(\int_Z g d\mu\right). \tag{$\mathrm{P}^{\mathrm{K}}_{m_N}$}$$

This idea was already proposed in [6], in an MFG context. The choice of the points $(x_i)_{i=1,\dots,N}$ will be discussed more in details in Section 7.2. We will present theoretical guarantees from [15] for a constructive approach based on quantization.

By the disintegration theorem, ($\mathrm{P}^{\mathrm{K}}_{m_N}$) is equivalent to a problem involving N probability measures, see (7.1). A direct implementation of the FW algorithm to ($\mathrm{P}^{\mathrm{K}}_{m_N}$) would lead to an approximate solution possibly involving kN points after k iterations of the algorithm, which may cause memory issues as k grows large. We will see that the Stochastic Frank–Wolfe algorithm, which we introduced and analyzed in [16], allows to obtain an approximate solution of ($\mathrm{P}^{\mathrm{K}}_{m_N}$) relying on only N support points, to the price of an additional error term of order $\mathcal{O}(1/N)$ in the main convergence result, see Lemma 7.6 and Remark 7.7. Let us mention that the Frank–Wolfe algorithm (also called conditional gradient method) was already applied to potential MFGs, see for example [17–19]. It can be seen as a generalization of the fictitious play, investigated in particular in [20, 21].

Stability results

We first derive a first-order necessary and sufficient optimality condition for ($\mathrm{P}^{\mathrm{K}}_{m}$) in Theorem 3.3, relying on the convexity of f and suitable regularity assumptions on g. This yields a fixed-point characterization of primal solutions in Corollary 3.4. Such a structure is reminiscent of characterizations of Nash equilibria in nonatomic games and of Lagrangian MFG models; see, *e.g.*, [1–3]. We also establish existence of a solution under a tightness assumption on minimizing sequences in Proposition 3.5.

We then study the stability of ($\mathrm{P}^{\mathrm{K}}_{m}$) with respect to the marginal m, in the topology induced by the Wasserstein–1 (Kantorovich–Rubinstein) distance (see Sect. 2.1 for the definition). This analysis is motivated by the numerical approach described above, where m is discretized. A key ingredient is our *bridging method* (Alg. 1), which constructs an approximate solution for a given marginal from an approximate solution corresponding to a nearby marginal. A diagrammatic illustration in the empirical case is provided in Remark 4.4. Our main stability estimate is Theorem 4.6, where we prove that the value function is Lipschitz continuous with respect to the Wasserstein–1 distance between marginals.

Finally, we introduce a dual formulation of ($\mathrm{P}^{\mathrm{K}}_{m}$) and prove strong duality in Proposition 6.2. We further show that the (unique) dual solution depends Hölder-continuously on m. These results lead to an explicit first-variation formula for the value function of the primal problems; see Proposition 6.5.

Kantorovich and Monge formulations

In the formulation of Problem ($\mathrm{P}_m^{\mathrm{K}}$), two agents with the same parameter x may take different decisions. We can impose that two agents with the same parameter x take the same decision $T(x)$. This leads to the problem

$$\inf_{T\in\mathcal{T}} f\left(\int_Z g(x,T(x))dm(x)\right), \tag{$\mathrm{P}_m^{\mathrm{M}}$}$$

where $\mathcal{T}$ denotes the set of Borel mappings $T\colon X\to Y$, such that $(x,T(x))\in Z$, for m-a.e. $x\in X$. We will refer to Problems ($\mathrm{P}_m^{\mathrm{K}}$) and ($\mathrm{P}_m^{\mathrm{M}}$) as the Kantorovich and Monge formulations, respectively, by analogy with the classical terminology of optimal transportation. Our last theoretical contribution is an upper bound on the difference of the values of the two formulations. In particular, when m is absolutely continuous, there is no gap between the Kantorovich and Monge formulations (see Cor. 5.7), which is similar with the result established in [22] for optimal transportation.

Organization

In Section 2, we present some notations and results in measure theory and set-valued functions, the rigorous description of the data of problem ($\mathrm{P}_m^{\mathrm{K}}$). Section 3 is dedicated to the primal problem: We provide a first-order optimality condition and an existence result. We perform in Section 4 a stability analysis for the primal problem, based on our bridging method. In Section 5, we analyze the gap between ($\mathrm{P}_m^{\mathrm{K}}$) and its Monge counterpart ($\mathrm{P}_m^{\mathrm{M}}$). In Section 6, we formulate the dual problem of ($\mathrm{P}_m^{\mathrm{K}}$), we prove strong duality, and we prove the stability of the dual solution. We provide our numerical method in Section 7. We perform in Section 8 some numerical simulations for a Lagrangian MFG model taken from [23] and for a congestion problem.

2. Preliminaries

2.1. Results in measure theory

A metric space is called a *Polish* space if it is complete and separable. Let X be a Polish space equipped with a metric d_X, and let $\mathcal{X}$ be a σ-algebra on X. The Borel σ-algebra on X is denoted by $\mathcal{B}^X$. Given any measure m on $\mathcal{X}$, we refer to the triplet $(X,\mathcal{X},m)$ as a measure space. Measure spaces are said to be complete if for any $A\in\mathcal{X}$ with $m(A)=0$ and for any subset B of A, we have $B\in\mathcal{X}$. We define

$$\mathcal{P}(X) \coloneqq \big\{m \text{ is a positive Borel measure on } X, \text{ and } m(X)=1\big\};$$

$$\mathcal{P}^1(X) \coloneqq \left\{m\in\mathcal{P}(X)\,\middle|\,\exists\, x_0\in X \text{ such that } \int_X d_X(x,x_0)dm<+\infty\right\}.$$

Let δ_x denote the Dirac measure at point x. We denote by $\mathcal{P}_\delta(\Omega)$ the set of finitely supported probability measures, defined by

$$\mathcal{P}_\delta(X) \coloneqq \left\{\sum_{k=1}^K \omega_k\delta_{x_k}\,\middle|\, K\in\mathbb{N},\ (\omega_k)_{k=1}^K\in(\mathbb{R}_+)^K,\ (x_k)_{k=1}^K\in X^K,\ \sum_{k=1}^K\omega_k=1\right\}.$$

In particular, we call $m\in\mathcal{P}_\delta(X)$ an *empirical distribution* if $\omega_k=1/K$ for $k=1,\dots,K$.

The set $\mathcal{P}(X)$ is endowed with the narrow topology. We say that a sequence $(m_n)_{n\geq1}$ in $\mathcal{P}(X)$ narrowly converges to some $m\in\mathcal{P}(X)$ if for any bounded and continuous function $F\colon X\to\mathbb{R}$,

$$\lim_{n\to+\infty}\int_X F dm_n=\int_X F dm.$$

The space $\mathcal{P}^1(X)$ is endowed with the *Kantorovich–Rubinstein Distance*,

$$d_1(m_0, m_1) \coloneqq \sup_{F \in \mathrm{Lip}_1(X)} \int_\Omega F d(m_0 - m_1),$$

where $\mathrm{Lip}_1(X)$ is the set of all 1-Lipschitz continuous functions on X. For any $m \in \mathcal{P}(X)$, the support of m is defined by

$$\mathrm{supp}(m) \coloneqq \big\{x \in X \mid m(V) > 0 \text{ for all open set } V \text{ such that } x \in V\big\}. \tag{2.1}$$

Lemma 2.1. *Let $m \in \mathcal{P}(X)$. Let $F\colon X \to \mathbb{R}_+$ be a Borel measurable function. Assume that*

$$\int_X F dm = 0.$$

Then $F = 0$, m-a.e. Moreover, if $F^{-1}(\{0\})$ is closed, then $\mathrm{supp}(m) \subseteq F^{-1}(\{0\})$.

Proof. The fact that $F = 0$, m-a.e., is from [24], Theorem 1.39(a). Now, let $F^{-1}(\{0\})$ be closed. Suppose that there exists $x \in \mathrm{supp}(m)$ such that $x \notin F^{-1}(\{0\})$. Since $F^{-1}(\{0\})$ is closed, there exists an open neighborhood V of x such that $F(x) > 0$, for all $x \in V$. By the definition of the support of a probability measure, we have $m(V) > 0$. Therefore, $\int_X F dm \geq \int_V F dm > 0$, contradiction. □

2.2. Results about set-valued functions

In this subsection, we consider a metric space X equipped with a metric d_X, a σ-algebra $\mathcal{X}$ on X, and a measure m on $\mathcal{X}$. Additionally, we fix a Polish space Y with a metric d_Y, and we denote the Borel σ-algebra on Y by $\mathcal{B}^Y$. We call F a set-valued function from X to Y if $F(x) \subseteq Y$ for all $x \in X$, denoted by $X \rightsquigarrow Y$ for short. The graph of F is defined by

$$\mathrm{Graph}(F) \coloneqq \{(x, y) \in X \times Y \mid y \in F(x)\}.$$

We say that F has closed (non-empty) images, if for any $x \in X$, $F(x)$ is closed (non-empty) in Y.

Let us give some definitions concerning regularity properties of set-valued functions, which are from [25], Definitions 1.4.1, 1.4.2, 1.4.5, and 8.1.1.

Definition 2.2. Let $F\colon X \rightsquigarrow Y$ be a set-valued function with non-empty images.

1. (Lower semi-continuity). The set-valued function F is *lower semi-continuous* at point $x \in X$ if for any $y \in F(x)$ and any sequence $(x_n \in X)_{n\geq 1}$ converging to x, there exists $y_n \in F(x_n)$ converging to y. The set-valued function F is said to be lower semi-continuous if it is lower semi-continuous at each point $x \in X$.
2. (Upper semi-continuity). The set-valued function F is *upper semi-continuous* at point $x \in X$ if for any neighborhood $\mathcal{U}$ of $F(x)$, there exists $\eta > 0$ such that for any $x' \in B_X(x, \eta)$, we have

   $$F(x') \subseteq \mathcal{U}.$$

   The set-valued function F is said to be upper semi-continuous if it is upper semi-continuous at each point $x \in X$.
3. (Lipschitz continuity). When X and Y are normed vector spaces, we say that F is L-Lipschitz continuous on X, for some $L > 0$, if for any $x_1, x_2 \in X$,

   $$F(x_1) \subseteq F(x_2) + B_Y(0, L d_X(x_1, x_2)).$$

   Here $B_Y(0, r)$ denotes the closed ball in Y centered at 0 with radius $r > 0$.

4. (Measurability). The set-valued function F is *measurable* if the inverse image of any open subset $\mathcal{O}$ of Y is measurable, *i.e.*,

$$F^{-1}(\mathcal{O}) \coloneqq \{x \in X \mid F(x) \cap \mathcal{O} \neq \emptyset\} \in \mathcal{X}.$$

Below, we present a key result on the measurable selection of set-valued functions, which is fundamental in the proofs of our article.

Theorem 2.3 (Aumann [26])**.** *Let $F\colon X \rightsquigarrow Y$ be a set-valued function with non-empty images. Let m be a σ-finite measure on $(X, \mathcal{X})$. If the graph of F, $\mathrm{Graph}(F)$, is $\mathcal{X} \otimes \mathcal{B}^Y$-measurable, then there exists a function $f\colon X \to Y$ such that:*

- *f is $(\mathcal{X}, \mathcal{B}^Y)$-measurable, and*
- *$f(x) \in F(x)$ for m-almost every $x \in X$.*

2.3. Data setting and technical lemmas

Recall the MFO problem ($\mathrm{P}_m^{\mathrm{K}}$). We consider the following setting:

- Two Polish spaces and their Borel σ-algebras: $(X, \mathcal{B}^X)$ and $(Y, \mathcal{B}^Y)$.
- A probability distribution on X: $m \in \mathcal{P}(X)$.
- A set-valued function $F\colon X \rightsquigarrow Y$ with a closed graph and non-empty images. Let

$$Z \coloneqq \mathrm{Graph}(F), \qquad Z_x \coloneqq F(x),\ \forall x \in X.$$

- The admissible set of probability measures:

$$\mathcal{P}_m(Z) \coloneqq \{\mu \in \mathcal{P}(Z) \mid \pi_x \# \mu = m\},$$

  where $\pi_x\colon Z \to X$, $(x, y) \mapsto x$.
- A separable Hilbert space: $\mathcal{H}$.
- Two Borel measurable functions: $g\colon Z \to \mathcal{H}$ and $f\colon \mathcal{H} \to \mathbb{R}$.

The integral $\int_Z g d\mu$ in ($\mathrm{P}_m^{\mathrm{K}}$) should be interpreted in the Bochner integration sense. We refer to [27], Appendix E for Bochner integrable functions.

Lemma 2.4. *If there exists a constant $M > 0$ such that $\|g(z)\| \leq M$ for any $z \in Z$, then the function g is Bochner integrable with respect to any $\mu \in \mathcal{P}(Z)$, i.e., $\int_Z g d\mu$ exists. Moreover, for any $\lambda \in \mathcal{H}$, we have*

$$\left\langle \lambda, \int_Z g d\mu \right\rangle = \int_Z \langle \lambda, g \rangle d\mu.$$

As a consequence, for any $\mu_1, \mu_2 \in \mathcal{P}(Z)$, we have

$$\left\langle \int_Z g d\mu_1 , \int_Z g d\mu_2 \right\rangle = \int_Z \int_Z \langle g(x), g(y) \rangle d\mu_1(x) d\mu_2(y).$$

Proof. As $\mathcal{H}$ is separable, the function g is strongly measurable. Moreover, as the constant function M is Bochner integrable with respect to any $\mu \in \mathcal{P}(Z)$, and $|g(z)| \leq M$ for any $z \in Z$, it follows from [27], Proposition E.2, Theorem E.6 that g is Bochner integrable with respect to any $\mu \in \mathcal{P}(Z)$. Therefore, we can apply [27], Proposition E.11 to obtain the first equality of this lemma. The second equality is obtained by applying twice the first one. □

In the next theorem, we say that a map $\nu\colon X \mapsto \nu_x \in \mathcal{P}(Y)$ is Borel measurable if for any Borel set $Y_0 \subseteq Y$, the map $x \mapsto \nu_x(Y_0)$ is Borel measurable.

Theorem 2.5 (Disintegration theorem)**.** *Let $\nu\colon x \in X \mapsto \nu_x \in \mathcal{P}(Y)$ be a Borel map, such that for m-a.e. $x \in X$, $\mathrm{supp}(\nu_x) \subseteq Z_x$. Then the mapping*

$$\phi \longmapsto \int_X \int_Y \phi(x,y)\, d\nu_x(y)\, dm(x), \qquad \text{for every Borel measurable } \phi : Z \to \mathbb{R}_+,$$

defines a unique probability measure $\mu \in \mathcal{P}_m(Z)$.

Conversely, for any $\mu \in \mathcal{P}_m(Z)$, there exists an m-a.e. uniquely determined Borel map $\nu\colon X \to \mathcal{P}(Y)$ such that for a.e. $x \in X$, $\mathrm{supp}(\nu_x) \subseteq Z_x$ and such that for any Borel measurable $\phi\colon Z \to \mathbb{R}_+$,

$$\int_Z \phi d\mu = \int_X \int_{Z_x} \phi(x,y) d\mu_x(y) dm(x). \tag{2.2}$$

Moreover, for a.e. $x \in X$, μ_x is uniquely determined.

Proof. The first part of the theorem is justified in [28], Section 5.3, expect the fact that $\mathrm{supp}(\mu) \subseteq Z$. Take a point $(x,y) \in X \times Y \backslash Z$. As Z is closed, there are two neighborhoods N_x and N_y of resp. x and y, such that $N_x \times N_y \cap Z = \emptyset$. This implies that for any $x \in N_x$, $N_y \cap Z_x = \emptyset$ and thus $\mu(N_x \times N_y) = 0$, which proves that (x,y) does not lie in the support of μ. The second part of the theorem is established in [28], Theorem 5.3.1. □

Remark 2.6. It is easy to verify that (2.2) also holds true for functions ϕ which are bounded from below, adding to them a sufficiently large constant.

Remark 2.7. Formula (2.2) extends to bounded functions ϕ valued in $\mathcal{H}$ since for all $\lambda \in \mathcal{H}$,

$$\Big\langle \lambda, \int_Z \phi d\mu \Big\rangle = \int_Z \langle \lambda, \phi \rangle d\mu = \int_X \int_{Z_x} \langle \lambda, \phi \rangle d\nu_x dm = \Big\langle \lambda, \int_X \int_{Z_x} \phi d\nu_x dm \Big\rangle.$$

As a result, we obtain an equivalent reformulation of Problem ($\mathrm{P}_m^{\mathrm{K}}$):

$$\inf_{\nu\colon X \to \mathcal{P}(Y)} f\left(\int_X g(x,y) d\nu_x(y) dm(x) \right), \quad \text{subject to: } \mathrm{supp}(\nu_x) \subseteq Z_x, \text{ for } m\text{-a.e. } x \in X, \tag{2.3}$$

where the infimum is taken with respect to Borel measurable maps $\nu\colon X \to \mathcal{P}(Y)$.

3. Optimality condition

3.1. Assumptions and constants

To simplify the presentation of the assumptions and the results of the article, we introduce the following (set-valued) functions, parameterized by $\lambda \in \mathcal{H}$:

- $g_\lambda\colon Z \to \mathbb{R}$ and $u_\lambda\colon X \to \mathbb{R}$,

$$g_\lambda(x,y) = \langle \lambda\,,\, g(x,y) \rangle\,, \qquad u_\lambda(x) = \inf_{y \in Z_x} g_\lambda(x,y);$$

- $G_\lambda\colon X \rightsquigarrow \mathbb{R}$ and $\mathbf{BR}_\lambda\colon X \rightsquigarrow Y$,

$$G_\lambda(x) = \{ g_\lambda(x,y) \mid y \in Z_x \}\,, \qquad \mathbf{BR}_\lambda(x) = \operatorname*{argmin}_{y \in Z_x} g_\lambda(x,y).$$

Assumption A. The following holds:

1. The function g is bounded. The function f is convex and differentiable, and ∇f is Lipschitz continuous with modulus L.
2. Let $\mathcal{H}_f \coloneqq \nabla f(\mathcal{H})$. Fixing any $\lambda \in \mathcal{H}_f$, we have:
   - the function g_λ is lower semi-continuous;
   - the set-valued function G_λ is lower semi-continuous;
   - the set-valued function $\mathbf{BR}_\lambda$ has non-empty images.

Three useful constants below are defined, following Assumption A:

$$M \coloneqq \sup_{z\in Z} \|g(z)\|, \qquad D \coloneqq \sup_{z_1,z_2\in Z} \|g(z_1)-g(z_2)\|^2, \qquad C \coloneqq \sup_{\mu\in\mathcal{P}(Z)} \left\|\nabla f\left(\int_Z g d\mu\right)\right\|. \tag{3.1}$$

In Section 6, we will consider the dual problem of ($\mathrm{P}_m^{\mathrm{K}}$). For the analysis of Section 6, Assumption A needs to be strengthened as follows:

Assumption A*. Assumption A(1) holds true and Assumption A(2) holds true for all $\lambda \in \mathrm{dom}(f^*)$, where f^* is the Fenchel conjugate of f.

Remark 3.1. We provide the following comments on our assumptions and their relation to the literature:

1. **Strengthening from Assumption A to Assumption A*.** Assumption A* is stronger than Assumption A since
$$\mathcal{H}_f = \nabla f(\mathcal{H}) \subseteq \mathrm{dom}(f^*).$$
2. **A Lipschitz estimate.** Under Assumption A, for any $\mu_0, \mu_1 \in \mathcal{P}(Z)$,
$$\left| f\left(\int_Z g\, d\mu_1\right) - f\left(\int_Z g\, d\mu_0\right)\right| \le C \left\| \int_Z g\, d(\mu_1-\mu_0)\right\|,$$
where C is defined in (3.1).
3. **Relation to existing regularity assumptions in the literature.** The convexity of f and the Lipschitz continuity of ∇f are standard in the analysis of first-order methods, such as gradient descent [29] and the FW algorithm [30]. Related optimization problems over measures are studied in [31–33] within the general framework
$$\inf_{\mu\in\mathcal{M}} F(K\mu) + R(\mu),$$
where $\mathcal{M}$ is a space of Radon measures, K is a linear operator into a Hilbert space, F plays a role analogous to our f, and R is a Tikhonov type regularization term. In these works, comparable smoothness assumptions on F are imposed; see, *e.g.*, [31], Assessment 2.2 and [33], Assessment 5.1, as well as the quadratic choices of F considered in [34].

   In contrast, our setting does not include an explicit coercive regularization term $R(\mu)$, which motivates the boundedness assumption on g. More importantly, our model incorporates the marginal constraint $\pi_x\#\mu = m$, which is natural in the MFG interpretation introduced in Section 8.2 and leads to the additional structural conditions in Assumption A(2).

3.2. First-order-optimality condition

The following lemma plays a key role in proving the first-order optimality condition for ($\mathrm{P}_m^{\mathrm{K}}$).

Lemma 3.2. *Let Assumption A hold true. For any $\lambda \in \mathcal{H}_f$, we have*

$$\inf_{\mu \in \mathcal{P}_m(Z)} \int_Z g_\lambda d\mu = \int_X u_\lambda dm.$$

Here we present a proof of Lemma 3.2 for the case where m has finite support, that is, $m \in \mathcal{P}_\delta(X)$. This particular case provides us with insight into the general proof, and proves beneficial for resolving the discretized problem introduced in Section 7.

Proof of Lemma 3.2 when $m \in \mathcal{P}_\delta(X)$. Fix any $\mu \in \mathcal{P}_m(Z)$. Since g is bounded over Z, the function g_λ is bounded from below. By Lemma 2.5 and Remark 2.6, we have

$$\int_Z g_\lambda d\mu = \int_X \int_{Z_x} g_\lambda(x,y) d\mu_x(y) dm(x) \geq \int_X u_\lambda dm,$$

where the second inequality follows from the definition of u_λ.

Let us prove the converse inequality. Let us fix $m \in \mathcal{P}_\delta(X)$. Let $K \in \mathbb{N}$, let $(x_k)_{k=1,\dots,K} \in X^K$ and let $(\omega_k)_{k=1,\dots,K} \in \mathbb{R}_+^K$ be such that $\sum_{k=1}^K \omega_k = 1$ and $m = \sum_{k=1}^K \omega_k \delta_{x_k}$. For any $k = 1, \dots, K$, let $y_k \in \mathbf{BR}_\lambda(x_k)$. Let us define $\tilde{\mu} = \sum_{k=1}^K \omega_k \delta_{(x_k, y_k)}$. Clearly $\tilde{\mu} \in \mathcal{P}_m(Z)$. Moreover,

$$\int_Z g_\lambda d\tilde{\mu} = \sum_{k=1}^K \omega_k g_\lambda(x_k, y_k) = \sum_{k=1}^K \omega_k u_\lambda(x_k) = \int_X u_\lambda dm.$$

The conclusion follows, moreover, $\tilde{\mu}$ minimizes $\int_Z g_\lambda d\mu$ over $\mathcal{P}_m(Z)$. □

In the general case, we need to apply measurable selection results from Theorem 2.3 to $\mathbf{BR}_\lambda$. The complete proof is given in Appendix A.

Theorem 3.3 (First-order optimality condition)**.** *Let Assumption A(1) hold true. Let $\bar{\mu} \in \mathcal{P}_m(Z)$ and $\bar{\lambda} = \nabla f\left(\int_Z g d\bar{\mu}\right)$. Consider the following three assertions:*

1. *The measure $\bar{\mu}$ is a solution of problem $(\mathrm{P}_m^{\mathrm{K}})$;*
2. *$\int_Z g_{\bar{\lambda}} d\bar{\mu} = \inf_{\mu \in \mathcal{P}_m(Z)} \int_Z g_{\bar{\lambda}} d\mu$;*
3. *$\mathrm{supp}(\bar{\mu}_x) \subseteq \mathbf{BR}_{\bar{\lambda}}(x)$, m-a.e., where $\bar{\mu}_x$ is defined by the disintegration theorem.*

Then, assertions (1) and (2) are equivalent. Moreover, under Assumption A(2), assertions (1), (2), and (3) are equivalent.

Proof. **Step 1**. (Equivalence between (1) and (2)). We first prove that (1) ⇒ (2). Suppose that $\bar{\mu}$ is a solution of problem $(\mathrm{P}_m^{\mathrm{K}})$. Take an arbitrary $\mu \in \mathcal{P}_m(Z)$. Then, for any $\alpha \in [0,1]$, we have

$$\begin{aligned} f\left(\int_Z g d\bar{\mu}\right) &\leq f\left(\int_Z g d(\bar{\mu} + \alpha(\mu - \bar{\mu}))\right) \\ &\leq f\left(\int_Z g d\bar{\mu}\right) + \alpha \left\langle \bar{\lambda}, \int_Z g d(\mu - \bar{\mu}) \right\rangle + \frac{\alpha^2 L D}{2}, \end{aligned}$$

where the second inequality follows from the Lipschitz-continuity of ∇f and the definition of D. Therefore

$$0 \leq \left\langle \bar{\lambda}, \int_Z g d(\mu - \bar{\mu}) \right\rangle + \frac{\alpha L D}{2}$$

Let α go to 0. We obtain that

$$\left\langle \bar{\lambda}, \int_Z g d\bar{\mu} \right\rangle = \inf_{\mu \in \mathcal{P}_m(Z)} \left\langle \bar{\lambda}, \int_Z g d\mu \right\rangle. \tag{3.2}$$

This implies (2) by the definition of $g_{\bar{\lambda}}$.

We now prove (2) $\Rightarrow$ (1). Let (2) hold true. We obtain (3.2) by the definition of $g_{\bar{\lambda}}$. The convexity of f implies that for any $\mu \in \mathcal{P}_m(Z)$,

$$f\left(\int_Z g d\mu\right) \geq f\left(\int_Z g d\bar{\mu}\right) + \left\langle \bar{\lambda}, \int_Z g d\mu - \int_Z g d\bar{\mu} \right\rangle \geq f\left(\int_Z g d\bar{\mu}\right).$$

Therefore, $\bar{\mu}$ is a solution of problem ($\mathrm{P}_m^{\mathrm{K}}$).

Step 2. (Equivalence between (2) and (3)). By Theorem 2.5, we have

$$\int_Z g_{\bar{\lambda}} d\bar{\mu} = \int_X \int_{Z_x} g_{\bar{\lambda}}(x,y) d\bar{\mu}_x(y) dm(x).$$

By Lemma 3.2, we have

$$\inf_{\mu \in \mathcal{P}_m(Z)} \int_Z g_{\bar{\lambda}} d\mu = \int_X u_{\bar{\lambda}} dm.$$

Therefore, assertion (2) is equivalent to

$$\int_X \int_{Z_x} g_{\bar{\lambda}}(x,y) d\bar{\mu}_x(y) dm(x) = \int_X u_{\bar{\lambda}} dm. \tag{3.3}$$

Let (3) hold true. It follows that $\int_{Z_x} g_{\bar{\lambda}}(x,y) d\bar{\mu}_x(y) = u_{\bar{\lambda}}(x)$, m-a.e., which implies (3.3).

Let (2) hold true. We obtain (3.3). The function $x \mapsto \left(\int_{Z_x} g_{\bar{\lambda}}(x,y) d\bar{\mu}_x(y)\right) - u_{\bar{\lambda}}(x)$ is nonnegative, for m-a.e. $x \in X$, by the definition of $u_{\bar{\lambda}}$. By (3.3), its integral is null, thus, as a consequence of Lemma 2.1, we have

$$\int_{Z_x} g_{\bar{\lambda}}(x,y) d\bar{\mu}_x(y) = u_{\bar{\lambda}}(x) = \inf_{y \in Z_x} g_{\bar{\lambda}}(x,y), \qquad m\text{-a.e.} \tag{3.4}$$

Fix $x \in X$ such that equality holds in (3.4). Consider the map $y \in Z_x \mapsto g_{\bar{\lambda}}(x,y) - u_{\bar{\lambda}}(x)$. It is nonnegative, with a null integral, and $\mathbf{BR}_{\bar{\lambda}}(x)$ is non-empty and closed. Then assertion (3) follows with Lemma 2.1. □

Corollary 3.4. *Under Assumption A, $\bar{\mu}$ is a solution of ($\mathrm{P}_m^{\mathrm{K}}$) if and only if the following equilibrium equation is satisfied:*

$$\begin{cases} \bar{\lambda} = \nabla f\left(\int_Z g d\bar{\mu}\right), \\ \mathrm{supp}(\bar{\mu}_x) \subseteq \mathbf{BR}_{\bar{\lambda}}(x), \quad m\text{-a.e.} \end{cases} \tag{3.5}$$

Proof. This is a consequence of Theorem 3.3. □

The conditions in (3.5) admit a game-theoretic interpretation as the NE conditions of a non-atomic game in which agents interact through the aggregate variable $\bar{\lambda}$. The relation $\bar{\lambda} = \nabla f(\int_Z g d\bar{\mu})$ shows how $\bar{\lambda}$ results from the collective behavior of the agents, while the relation $\mathrm{supp}(\bar{\mu}_x) \subseteq \mathbf{BR}_{\bar{\lambda}}(x)$ shows that the agents behave optimally, for some criterion that depends on $\bar{\lambda}$. We will discuss some more concrete examples in Section 8.

3.3. Existence of a solution under tightness assumptions

We denote by $\mathbf{val}(\mathrm{P}_m^{\mathrm{K}})$ the value of problem $(\mathrm{P}_m^{\mathrm{K}})$. We can easily deduce from Assumption A that $\mathbf{val}(\mathrm{P}_m^{\mathrm{K}}) > -\infty$. The following proposition demonstrates the existence of a solution to problem $(\mathrm{P}_m^{\mathrm{K}})$ under some additional assumptions.

Proposition 3.5 (Existence)**.** *Let Assumption A hold true. Let $(\mu_n)_{n\geq 1}$ be a minimizing sequence for problem $(\mathrm{P}_m^{\mathrm{K}})$. Suppose that $\{\mu_n\}_{n\geq 1}$ is tight in $\mathcal{P}(Z)$, i.e. for any $\epsilon > 0$, there exists a compact subset K_ϵ of Z such that*

$$\mu_n(K_\epsilon) \geq 1 - \epsilon, \qquad \forall n \geq 1.$$

Then every accumulation point of $\{\mu_n\}_{n\geq 1}$ for the narrow topology (there exists at least one) is a solution of $(\mathrm{P}_m^{\mathrm{K}})$.

Proof. By *Prokhorov's* theorem [35], p. 43, the set $\{\mu_n\}_{n\geq 1}$ is relatively compact with respect to the narrow topology. Without loss of generality, suppose that $(\mu_n)_{n\geq 1}$ narrowly converges to some $\bar{\mu} \in \mathcal{P}(Z)$. The set $\mathcal{P}_m(Z)$ is closed with respect to narrow topology by [3], Proposition 2.4. This implies that $\bar{\mu} \in \mathcal{P}_m(Z)$. Let $\bar{\lambda} = \nabla f(\int_Z g d\bar{\mu})$. Since f is convex, we have

$$f\left(\int_Z g d\mu_n\right) \geq f\left(\int_Z g d\bar{\mu}\right) + \int_Z g_{\bar{\lambda}} d(\mu_n - \bar{\mu}). \tag{3.6}$$

Since $g_{\bar{\lambda}} \colon Z \to \mathbb{R}$ is lower semi-continuous and bounded from below by Assumption A, we deduce the following inequality from [35], Lemma 4.3:

$$\liminf_{n\to+\infty} \int_Z g_{\bar{\lambda}} d(\mu_n - \bar{\mu}) \geq 0.$$

In inequality (3.6), letting n go to infinity, by the definition of μ_n, we have

$$\mathbf{val}(\mathrm{P}_m^{\mathrm{K}}) = \liminf_{n\to+\infty} f\left(\int_Z g d\mu_n\right) \geq f\left(\int_Z g d\bar{\mu}\right) \geq \mathbf{val}(\mathrm{P}_m^{\mathrm{K}}).$$

Therefore, $\bar{\mu}$ is a solution of problem $(\mathrm{P}_m^{\mathrm{K}})$. □

Remark 3.6. Let us comment on the tightness assumption for $\{\mu_n\}_{n\geq 1}$ in Proposition 3.5. A trivial sufficient condition is that Z is compact, in which case every family of probability measures on Z is tight.

Another convenient sufficient condition is the presence of a coercive term (*e.g.* a Tikhonov-type penalty) in the objective, as in the optimization problems over measures considered in [31–33].

4. Stability analysis and bridging method

In this section, we study the stability of the primal problem $(\mathrm{P}_m^{\mathrm{K}})$ with respect to its parameter m. We need the following assumptions (recall the data setting introduced in Section 2.3).

Assumption B. The following holds:

1. The space X is a closed subset of a separable Banach space;
2. The set Z_x is compact for any $x \in X$ and the set-valued function $F \colon X \rightsquigarrow Y$ is upper semi-continuous;
3. There exists $L_g \geq 0$ such that the set-valued function

$$\mathcal{Z} \colon X \rightsquigarrow \mathcal{H},\ x \mapsto \{g(x,y) \mid y \in Z_x\} \tag{4.1}$$

is L_g-Lipschitz on X.

Remark 4.1. Assume that X is a finite set. Then Assumption B is satisfied as soon as Z_x is compact for every $x \in X$. Indeed, point (1) holds trivially, and points (2)-(3) are automatic, since there exists a minimal strictly positive distance between two distinct points of X. In particular, Assumption B holds for the discretized MFO problems ($\mathrm{P}^{\mathrm{K}}_{m_N}$) studied in Section 7, where m is replaced by an empirical measure m_N and X can be restricted to $\mathrm{supp}(m_N)$. Consequently, the stability results established in this section apply directly to the discretized problems.

Let m_0 and m_1 lie in $\mathcal{P}(X)$. We consider the following two instances of ($\mathrm{P}^{\mathrm{K}}_m$) with $m = m_0$ and $m = m_1$ respectively:

$$\inf_{\mu \in \mathcal{P}_{m_0}(Z)} f\left(\int_Z g d\mu\right); \tag{$\mathrm{P}^{\mathrm{K}}_{m_0}$}$$

$$\inf_{\mu \in \mathcal{P}_{m_1}(Z)} f\left(\int_Z g d\mu\right). \tag{$\mathrm{P}^{\mathrm{K}}_{m_1}$}$$

Suppose that we have an (approximate) solution of problem ($\mathrm{P}^{\mathrm{K}}_{m_0}$), denoted by μ_0. Our goal is to propose a constructive method for finding a candidate $\mu_1 \in \mathcal{P}_{m_1}(Z)$ that is close to μ_0. We call it the bridging method. The method proceeds into two steps.

- In the first step, we assign to the agents a new parameter. More precisely, given μ_0 and a transportation plan ρ between m_0 and m_1, we construct a measure $\nu \in \mathcal{P}(Z \times X)$ such that $d\nu(x_0, y_0, x_1)$ describes the proportion of agents with parameter x_0 and decision variable y_0 to whom we assign the parameter x_1.
- The second step of the method is a correction step: the decision variable y_0 of an agent with parameter x_0 may not stay feasible when x_0 is changed to x_1 (*i.e.*, possibly $y_0 \notin Z_{x_1}$). So we need to construct a mapping that corrects the decision into a new one that is feasible.

To formalize the first step, we introduce some notation. We set

$$\Pi(m_0, m_1) = \{\rho \in \mathcal{P}(X \times X) \mid \pi_{x_0}\#\rho = m_0,\ \pi_{x_1}\#\rho = m_1\},$$

where $\pi_{x_0} \colon (x_0, x_1) \in X \times X \mapsto x_0$ and $\pi_{x_1} \colon (x_0, x_1) \in X \times X \mapsto x_1$. We also consider the functions

$$\pi_{x_0,x_1} \colon (x_0, y_0, x_1) \in Z \times X \mapsto (x_0, x_1) \quad \text{and} \quad \pi_{x_0,y_0} \colon (x_0, y_0, x_1) \in Z \times X \mapsto (x_0, y_0).$$

The following classical result is known as the gluing lemma [35], Chapter 1. It provides us with the desired measure ν announced in the first step of the bridging method.

Lemma 4.2. *Given $\mu_0 \in \mathcal{P}_{m_0}(Z)$ and $\rho \in \Pi(m_0, m_1)$, there exists $\nu \in \mathcal{P}(Z \times X)$ such that*

$$\pi_{x_0,x_1}\#\nu = \rho \quad \textit{and} \quad \pi_{x_0,y_0}\#\nu = \mu_0.$$

The correction mapping used for the second step of our method will be chosen as a measurable selection of the set-valued function $S \colon Z \times X \rightsquigarrow Y$, defined by

$$S(x_0, y_0, x_1) = \{y_1 \in Z_{x_1} \mid \|g(x_1, y_1) - g(x_0, y_0)\| \leq L_g d_X(x_0, x_1)\}. \tag{4.2}$$

Lemma 4.3. *Let Assumption B hold true. Fix any $\nu \in \mathcal{P}(Z \times X)$. Then, the set-valued function S has a measurable selection function $s \colon Z \times X \to Y$ such that, for ν-a.e. $(x_0, y_0, x_1) \in Z \times X$, we have $(x_1, s(x_0, y_0, x_1)) \in Z$*

and

$$\|g(x_1, s(x_0, y_0, x_1)) - g(x_0, y_0)\| \leq L_g d_X(x_0, x_1).$$

The proof of the lemma is postponed to Appendix B.
Algorithm 1 summarizes our bridging method.

Algorithm 1: Bridging method.

Input: $m_0, m_1 \in \mathcal{P}^1(X)$, and $\mu_0 \in \mathcal{P}_{m_0}(Z)$.
Step 1. Find a transportation plan $\rho \in \Pi(m_0, m_1)$.
Step 2. Find $\nu \in \mathcal{P}(Z \times X)$ such that $\pi_{x_0,y_0}\#\nu = \mu_0$ and $\pi_{x_0,x_1}\#\nu = \rho$.
Step 3. Choose a ν-measurable selection $s : Z \times X \to Y$ of S (Lem. 4.3) and define

$$\tilde{s} : Z \times X \to Z, \qquad \tilde{s}(x_0, y_0, x_1) \coloneqq \big(x_1,\, s(x_0, y_0, x_1)\big). \tag{4.3}$$

Step 4. Set $\mu_1 = \tilde{s}\#\nu \in \mathcal{P}_{m_1}(Z)$, where $\tilde{s}$ is defined in (4.3).
Output: μ_1.

Remark 4.4. We illustrate the main idea of the bridging method in a simple discrete setting. Consider the empirical measures

$$m_0 = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}, \qquad \mu_0 = \frac{1}{N}\sum_{i=1}^{N} \delta_{(x_i,y_i)}, \qquad m_1 = \frac{1}{N}\sum_{i=1}^{N} \delta_{x'_i}.$$

Assume that an optimal Monge transport map $T : \{x_i\}_{i=1}^N \to \{x'_i\}_{i=1}^N$ attains the Wasserstein–1 distance between m_0 and m_1. The bridging construction consists in moving each x_i to $\tilde{x}_i := T(x_i)$ and selecting a corresponding $\tilde{y}_i \in Z_{\tilde{x}_i}$ such that

$$\big\|g(x_i, y_i) - g(\tilde{x}_i, \tilde{y}_i)\big\| \leq L_g\, d_X(x_i, \tilde{x}_i),$$

which is guaranteed by Assumption B(3). This yields the transported measure

$$\mu_1 = \frac{1}{N}\sum_{i=1}^{N} \delta_{(\tilde{x}_i,\tilde{y}_i)}.$$

An illustrative diagram is provided in Figure 1.

We have the following result.

Lemma 4.5. *Let Assumptions A-B hold true. Let μ_1 denote the output of Algorithm 1. Then $\mu_1 \in \mathcal{P}_{m_1}(Z)$ and*

$$\left|f\left(\int_Z g d\mu_1\right) - f\left(\int_Z g d\mu_0\right)\right| \leq C L_g \int_{X\times X} d_X(x_0, x_1) d\rho(x_0, x_1), \tag{4.4}$$

where C is the constant defined in (3.1) and ρ is the transport plan in step 1 of Algorithm 1.

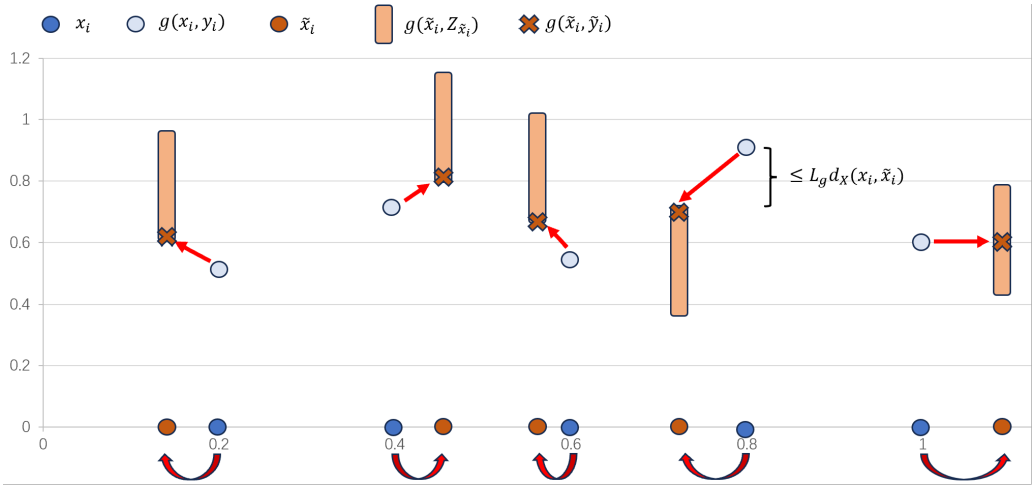


FIGURE 1. Illustration of the bridging method for an empirical measure case.

Proof. Since $\nu \in \mathcal{P}(Z \times X)$ and $\tilde{s}\colon Z \times X \to Z$, we have that $\mu_1 = \tilde{s}\#\nu$ lies in $\mathcal{P}(Z)$. Next we show that $\pi_x\#\mu_1 = m_1$. We have

$$\pi_x\#\mu_1 = \pi_x\#(s\#\nu) = (\pi_x \circ s)\#\nu.$$

Next it is direct to verify that $\pi_x \circ \tilde{s} = \pi_{x_1} \circ \pi_{x_0,x_1}$. Moreover, since $\pi_{x_0,x_1}\#\nu = \rho$, we have that

$$(\pi_x \circ \tilde{s})\#\nu = (\pi_{x_1} \circ \pi_{x_0,x_1})\#\nu = \pi_{x_1}\#(\pi_{x_0,x_1}\#\nu) = \pi_{x_1}\#\rho = m_1,$$

as ρ lies in $\mathcal{P}(m_0, m_1)$.

It remains to prove the estimate (4.4). We have that $\int_Z g d\mu_1 = \int_{Z\times X} g \circ s d\nu$ and $\int_Z g d\mu_0 = \int_{Z\times X} g d\nu$. Using Lemma 4.3, we deduce that

$$\Big\| \int_Z g d\mu_1 - \int_Z g d\mu_0 \Big\| \leq \int_Z \|g \circ \tilde{s} - g\| d\nu(x_0, y_0, x_1) \leq L_g \int_{X\times X} d_X(x_0, x_1) d\rho(x_0, x_1).$$

The announced inequality follows with Remark 3.1(2). □

Obviously it is of interest to consider a transportation plan that minimizes the right-hand side of estimate (4.4), *i.e.*, a solution to the transportation problem

$$\inf_{\rho \in \Pi(m_0,m_1)} \int_{X\times X} d_X(x, x') d\rho(x, x'). \tag{OT}$$

It is well-known that if m_0 and m_1 lie in $\mathcal{P}^1(X)$, then $d_1(m_0, m_1) = \mathbf{val}(\mathrm{OT})$ (see, *e.g.*, [35], Rem. 6.5). This yields the following stability estimate.

Theorem 4.6. *Let Assumptions A-B hold true. Assume that $m_0, m_1 \in \mathcal{P}^1(Z)$. Then*

$$|\mathbf{val}(\mathrm{P}^{\mathrm{K}}_{m_0}) - \mathbf{val}(\mathrm{P}^{\mathrm{K}}_{m_1})| \leq C L_g d_1(m_0, m_1). \tag{4.5}$$

Proof. Let $\mu_0 \in \mathcal{P}_{m_0}(Z)$. Let $\rho \in \Pi(m_0, m_1)$ and let $\mu_1 \in \mathcal{P}_{m_1}(Z)$ denote the output of Algorithm 1. Then it holds that

$$\mathbf{val}(\mathrm{P}^{\mathrm{K}}_{m_1}) \leq f\left(\int_Z g \,\mathrm{d}\mu_1\right) \leq f\left(\int_Z g \,\mathrm{d}\mu_0\right) + CL_g \int_{X\times X} d_X(x_0, x_1) d\rho(x_0, x_1).$$

Minimizing the right-hand side of the above inequality with respect to μ_0 and ρ, we deduce that

$$\mathbf{val}(\mathrm{P}^{\mathrm{K}}_{m_1}) \leq \mathbf{val}(\mathrm{P}^{\mathrm{K}}_{m_0}) + CL_g d_1(m_0, m_1).$$

Exchanging m_0 and m_1, we obtain (4.5). □

Corollary 4.7. *Let Assumptions A-B hold true. Assume that $m_0, m_1 \in \mathcal{P}^1(Z)$. Let μ_0 be an ε_1-solution of $(\mathrm{P}^{\mathrm{K}}_{m_0})$ and let ρ be an ε_2-solution of (OT). Then the output of Algorithm 1 is $\varepsilon_1 + CL_g(\varepsilon_2 + 2d_1(m_0, m_1))$ optimal, i.e.,*

$$f\left(\int_Z g d\mu_1\right) - \mathbf{val}(\mathrm{P}^{\mathrm{K}}_{m_1}) \leq \varepsilon_1 + CL_g(\varepsilon_2 + 2d_1(m_0, m_1)).$$

Proof. Using Lemma 4.5 and Theorem 4.6, we obtain that

$$\begin{aligned} f\left(\int_Z g \,\mathrm{d}\mu_1\right) - \mathbf{val}(\mathrm{P}^{\mathrm{K}}_{m_1}) \leq & \left[f\left(\int_Z g d\mu_1\right) - f\left(\int_Z g d\mu_0\right)\right] \\ & + \left[f\left(\int_Z g d\mu_0\right) - \mathbf{val}(\mathrm{P}^{\mathrm{K}}_{m_0})\right] + \left[\mathbf{val}(\mathrm{P}^{\mathrm{K}}_{m_0}) - \mathbf{val}(\mathrm{P}^{\mathrm{K}}_{m_1})\right] \\ \leq & CL_g(d_1(m_0, m_1) + \varepsilon_2) + \varepsilon_1 + CL_g d_1(m_0, m_1), \end{aligned}$$

as was to be shown. □

5. Relaxation gap between the Monge and Kantorovich formulations

We discuss in this section the relationship between our MFO problem $(\mathrm{P}^{\mathrm{K}}_m)$ and its "Monge" counterpart $(\mathrm{P}^{\mathrm{M}}_m)$, defined in the introduction. Recall that $\mathcal{T}$ is the set of Borel mappings $T\colon X \to Y$, such that $(x, T(x)) \in Z$, for m-a.e. $x \in X$. Therefore, for $T \in \mathcal{T}$, the measure defined by $(I, T)\#m$ lies in $\mathcal{P}_m(Z)$ and it holds that $f(\int_Z g d\mu) = f(\int_X g \circ (I, T) dm)$. It follows that

$$\mathbf{val}(\mathrm{P}^{\mathrm{K}}_m) \leq \mathbf{val}(\mathrm{P}^{\mathrm{M}}_m).$$

The purpose of this section is to establish some upper bounds for the following quantity, which we will refer to as the relaxation gap:

$$\mathbf{gap}(m) = \mathbf{val}(\mathrm{P}^{\mathrm{M}}_m) - \mathbf{val}(\mathrm{P}^{\mathrm{K}}_m) \geq 0.$$

The following result is a gap estimate in the case where m is a discrete measure. Its proof is a direct adaption of [16], Proposition 2.6. Similar results and techniques have been applied to determine the universal approximation rate in machine learning (see [36], Thm. 1 and [37], Lem. 1), and these can be traced back to general results from functional analysis [38], Lemma 2. The following lemma involves the Lipschitz-modulus L of ∇f and the constant D, defined as the square of the diameter of the image of g (see (3.1)).

Lemma 5.1. *Let Assumption A hold true. Assume that m is of the form $m = \sum_{i=1}^N \alpha_i \delta_{x_i}$, where $\alpha_i \geq 0$, for all $i = 1, ..., N$ and $\sum_{i=1}^N \alpha_i = 1$. Then*

$$\mathbf{gap}(m) \leq \frac{1}{2} LD \|\alpha\|_2^2,$$

where $\|\alpha\|_2$ denotes the Euclidean norm of the vector $\alpha = (\alpha_1, \ldots, \alpha_N)$.

Proof. As m is a discrete distribution, the Monge problem amounts to optimize, for each i, the decision (say y_i) to be taken by agents with parameter x_i. Similarly, the Kantorovich formulation amounts to optimize, for each i, the distribution of the decisions (say ν_i) to be taken by the agents with parameters x_i (see Rem. 2.7). This yields the following equivalent formulations of Problems ($\mathrm{P}_m^{\mathrm{K}}$) and ($\mathrm{P}_m^{\mathrm{M}}$):

$$\mathbf{val}(\mathrm{P}_m^{\mathrm{M}}) = \inf_{\substack{y=(y_i)_{i=1,\ldots,N} \\ y \in \prod_{i=1}^N Z_{x_i}}} f\left(\sum_{i=1}^N \alpha_i g(x_i, y_i)\right), \tag{5.1}$$

$$\mathbf{val}(\mathrm{P}_m^{\mathrm{K}}) = \inf_{\substack{\nu=(\nu_i)_{i=1,\ldots,N} \\ \nu \in \prod_{i=1}^N \mathcal{P}(Z_{x_i})}} f\left(\sum_{i=1}^N \alpha_i \int_{Z_{x_i}} g(x_i, \cdot) d\nu_i\right). \tag{5.2}$$

Let $\varepsilon > 0$. Take an ε-optimal candidate ν for (5.2). Denote $z = \sum_{i=1}^N \alpha_i \int_{Z_{x_i}} g(x_i, \cdot) d\nu_i$. We have $f(z) \leq \mathbf{val}(\mathrm{P}_m^{\mathrm{K}}) + \varepsilon$. As ∇f is L-Lipschitz continuous, we have, for any $y \in \prod_{i=1}^N Z_{x_i}$,

$$\begin{aligned}
\mathbf{val}(\mathrm{P}_m^{\mathrm{M}}) &\leq f\left(\sum_{i=1}^N \alpha_i g(x_i, y_i)\right) \\
&\leq f(z) + \left\langle \nabla f(z), \sum_{i=1}^N \alpha_i g(x_i, y_i) - z \right\rangle + \frac{L}{2} \left\| \left(\sum_{i=1}^N \alpha_i g(x_i, y_i)\right) - z \right\|^2.
\end{aligned}$$

Next we define $\hat{g}_i$ as the map $y_i \mapsto g(x_i, y_i) - \int_{Z_{x_i}} g(x_i, \cdot) d\nu_i$. Using this definition and the one of z, the above inequality writes

$$\begin{aligned}
\mathbf{val}(\mathrm{P}_m^{\mathrm{M}}) &\leq f(z) + \left\langle \nabla f(z), \sum_{i=1}^N \alpha_i \hat{g}_i(y_i) \right\rangle + \frac{L}{2} \left\| \sum_{i=1}^N \alpha_i \hat{g}_i(y_i) \right\|^2 \\
&= f(z) + \left\langle \nabla f(z), \sum_{i=1}^N \alpha_i \hat{g}_i(y_i) \right\rangle + \frac{L}{2} \sum_{i=1}^N \alpha_i^2 \|\hat{g}_i(y_i)\|^2 + \frac{L}{2} \sum_{i=1}^N \sum_{j \neq i} \alpha_i \alpha_j \langle \hat{g}_i(y_i), \hat{g}_j(y_j) \rangle.
\end{aligned} \tag{5.3}$$

The rest of the proof consists of three steps, which we do not detail: (i) integrate (5.3) in y with respect to $\nu := \nu_1 \times \ldots \times \nu_N$, (ii) use the fact that $\|\hat{g}_i(y_i)\|^2 \leq D$, for any y_i, and that $\int_{Z_{x_i}} \hat{g}_i d\nu_i = 0$, (iii) use the inequality $f(z) \leq \mathbf{val}(\mathrm{P}_m^{\mathrm{K}}) + \varepsilon$ and pass to the limit when $\varepsilon \downarrow 0$. □

We extend the above gap estimate to the case of general probability distributions m with the help of a stability estimate for the Monge formulation, derived with an adaptation of the bridging method (Alg. 1). The following lemma shows how to generate a good candidate T_1 to ($\mathrm{P}_{m_1}^{\mathrm{M}}$) out of a candidate T_0 to $\mathrm{P}_{m_0}^{\mathrm{M}}$, when m_0 is the push-forward of m_1. The construction relies on the measurable selection provided by Lemma 4.3.

Lemma 5.2. *Let Assumptions A and B hold true. Let $m_1 \in \mathcal{P}(X)$. Let $U\colon X \to X$ be a Borel map and let $m_0 = U\#m_1$. Let $T_0 \in \mathcal{T}$ and define the probability measure $\nu \in \mathcal{P}(Z \times X)$ by*

$$\nu \coloneqq \big(U,\, T_0 \circ U,\, \mathrm{Id}\big)\#m_1.$$

Let $s : Z \times X \to Y$ be a measurable selection of S associated with ν given by Lemma 4.3, and define

$$T_1(x_1) \coloneqq s\big(U(x_1),\, T_0(U(x_1)),\, x_1\big) \qquad \text{for } m_1\text{-a.e. } x_1 \in X.$$

Then

$$f\left(\int_Z g \circ (I, T_1) dm_1\right) \leq f\left(\int_Z g \circ (I, T_0) dm_0\right) + CL_g \int_X d_X(x_1, U(x_1)) dm_1(x_1), \tag{5.4}$$

Proof. Lemma 4.3 ensures that $T_1 \in \mathcal{T}$ and that for ν-a.e. $(x_0, y_0, x_1) \in Z \times X$,

$$\|g(x_1, s(x_0, y_0, x_1)) - g(x_0, y_0))\| \leq C d_X(x_0, x_1).$$

We apply this inequality with $x_0 = U(x_1)$ and $y_0 = T_0(x_0)$. For this special choice, we have $s(x_0, y_0, x_1) = T_1(x_1)$ by definition of T_1. We deduce that for m_1-a.e. $x_1 \in X$,

$$\|g \circ (I, T_1)(x_1) - g \circ (I, T_0) \circ U(x_1)\| \leq C d_X(x_1, U(x_1)).$$

Since $m_0 = U\#m_1$, we obtain that

$$\begin{aligned}\left\|\int_Z g \circ (I, T_1) dm_1 - \int_Z g \circ (I, T_0) dm_0\right\| &\leq \int_Z \big\|g \circ (I, T_1) - g \circ (I, T_0) \circ U\big\| dm_1 \\ &\leq C \int_Z d_X(x_1, U(x_1)) dm_1(x_1).\end{aligned}$$

The announced inequality follows with the Lipschitzianity of f (Rem. 3.1(2)). □

Remark 5.3. Lemma 5.2 is nothing but a reformulation of Lemma 4.5 in the Monge formalism. Indeed, in the framework of Lemma 5.2, one can define $\rho = (U, I)\#m_1 \in \Pi(m_0, m_1)$ and $\mu_0 = (I, T_0)\#m_0 \in \mathcal{P}_Z(m_0)$. Then we have $\mu_0 = (U, T_0 \circ U)\#m_1$ and thus the measure $\nu = (U, T_0 \circ U, I)\#m_1$ realizes a gluing between μ_0 and ρ; moreover, the output of Algorithm 1 is simply $(I, T_1)\#m_1$ and estimate (5.4) can be recovered as a consequence of (4.4).

Corollary 5.4. *For any $m \in \mathcal{P}(X)$, for any Borel map $U\colon X \to X$, we have*

$$\mathbf{val}(\mathrm{P}^{\mathrm{M}}_{U\#m}) \leq \mathbf{val}(\mathrm{P}^{\mathrm{M}}_m) + CL_g \int_X d_X(x, U(x)) dm(x).$$

Proof. For an arbitrary $\varepsilon > 0$, take an ε-solution T to Problem $(\mathrm{P}^{\mathrm{M}}_m)$. Then Lemma 5.2 yields

$$\begin{aligned}\mathbf{val}(\mathrm{P}^{\mathrm{M}}_{U\#m}) &\leq f\left(\int_Z g \circ (I, T) dm\right) + CL_g \int_X d_X(x, U(x)) dm(x) \\ &\leq \mathbf{val}(\mathrm{P}^{\mathrm{M}}_m) + \varepsilon + CL_g \int_X d_X(x, U(x)) dm(x).\end{aligned}$$

The result follows, passing to the limit as $\varepsilon \downarrow 0$. □

Definition 5.5. We call a centered partition any tuple P of the form $(\Omega_i, x_i)_{i=1,\dots,N} \in (2^X \times X)^N$, where $(\Omega_i)_{i=1,\dots,N}$ is a Borel-measurable partition of X. Given a centered partition $P = (\Omega_i, x_i)_{i=1,\dots,N}$, we define the mapping $U[P]\colon X \to X$ by

$$U[P](x) = x_i,$$

where i is the unique index in $\{1, \dots, N\}$ such that $x \in \Omega_i$. When m is also given, we define the vector $\alpha[P, m] = (\alpha_i[P, m])_{i=1,\dots,N} \in \mathbb{R}^N$ by

$$\alpha_i[P, m] = m(\Omega_i).$$

Note that the cardinality of a centered partition is not prescribed, also we do not require that the points x_i are two-by-two different or that they lie in Ω_i. Note also that for a centered partition $P = (\Omega_i, x_i)_{i=1,\dots,N}$, we have the identity $U[P]\#m = \sum_{i=1}^N \alpha_i[P, m]\delta_{x_i}$.

Theorem 5.6. *Let Assumptions A and B hold true. For any $m \in \mathcal{P}(X)$, it holds that*

$$\mathbf{gap}(m) \le \inf_P \left(\frac{1}{2} LD \|\alpha[P, m]\|_2^2 + 2CL_g \int_X d_X(x, U[P](x)) dm(x) \right),$$

where the infimum is considered with respect to all centered partitions.

Proof. Let P be a centered partition. We have: $\mathbf{gap}(m) = (a) + (b) + (c)$ with

$$(a) = \mathbf{val}(\mathrm{P}_m^{\mathrm{M}}) - \mathbf{val}(\mathrm{P}_{U[P]\#m}^{\mathrm{M}}), \quad (b) = \mathbf{gap}(U[P]\#m), \quad (c) = \mathbf{val}(\mathrm{P}_{U[P]\#m}^{\mathrm{K}}) - \mathbf{val}(\mathrm{P}_m^{\mathrm{K}}).$$

The three terms can be respectively bounded with Corollary 5.4, Lemma 5.1, and Theorem 4.6. This yields

$$\begin{aligned}
(a) &\le CL_g \int_X d_X(x, U[P](x)) dm(x) \\
(b) &\le \frac{1}{2} LD \|\alpha[P, m]\|_2^2 \\
(c) &\le CL_g d_1(m, U[P]\#m) \le CL_g \int_X d_X(x, U[P](x)) dm(x).
\end{aligned}$$

Combining the three above inequalities and minimizing the result with respect to P, we obtain the desired inequality. □

Corollary 5.7. *Let Assumptions A and B hold true. In addition, assume that X is a compact set of $\mathbb{R}^n$ and that m is absolutely continuous with respect to the Lebesgue measure. Then* $\mathbf{gap}(m) = 0$.

Proof. Let us fix $\varepsilon > 0$. By [39], Theorem 3.5, there exists $\delta > 0$ such that for all Borel set Ω, if $\int_\Omega \mathrm{d}x \le \delta$, then $m(\Omega) \le \varepsilon$. Let $\delta' = \min\{\varepsilon, \delta\}^{1/n}$. Since X is a compact subset of $\mathbb{R}^n$, there exists a finite partition $P = (\Omega_i, x_i)_{i=1,\dots,N}$ where $\Omega_i = (x_i + B) \cap X$, and where $x_i \in \delta'\mathbb{Z}^n$ and $B = [0, \delta')^n$. We have $m(\Omega_i) \le \varepsilon$, for all $i = 1, \dots, N$. Therefore,

$$\|\alpha[P, m]\|_2^2 = \sum_{i=1}^N m(\Omega_i)^2 \le \max_{i=1,\dots,N} m(\Omega_i) \left(\sum_{i=1}^N m(\Omega_i) \right) \le \varepsilon.$$

Moreover, we have $d_X(x, U[P](x)) \leq \delta'$, for all $x \in X$. Then Theorem 5.6 implies that

$$\mathbf{gap}(m) \leq \frac{1}{2} LD\varepsilon + 2CL_g \varepsilon^{1/n}.$$

Passing to the limit as $\varepsilon \downarrow 0$, the result follows. □

Remark 5.8. Corollary 5.7 can be generalized to the case where X is a compact set of $\mathbb{R}^n$ and $m = \sum_{i=1}^{N} \alpha_i \delta_{x_i} + \alpha_{N+1} \tilde{m}$, where $\tilde{m} \in \mathcal{P}(X)$ is absolutely continuous with respect to the Lebesgue measure, $\alpha_i \geq 0$, for all $i = 1, \ldots, N+1$, and $\sum_{i=1}^{N+1} \alpha_i = 1$. It is easy to show that in this case,

$$\mathbf{gap}(m) \leq \frac{1}{2} LD \sum_{i=1}^{N} \alpha_i^2.$$

6. Duality analysis

6.1. The dual problem

This section is dedicated to the duality analysis of the primal problem ($\mathrm{P}_m^{\mathrm{K}}$). In the sequel of this section, let Assumptions $\mathbf{A}^*$ and B hold true. Consider the equivalent formulation of problem of ($\mathrm{P}_m^{\mathrm{K}}$),

$$\inf_{\mu \in \mathcal{P}_m(Z),\, \beta \in \mathcal{H}} f(\beta), \qquad \text{s.t. } \beta = \int_Z g d\mu. \tag{$\tilde{\mathrm{P}}_m$}$$

The *Lagrangian* $\mathcal{L}\colon \mathcal{H}^2 \times \mathcal{P}_m(Z) \to \mathbb{R}$ associated with ($\tilde{\mathrm{P}}_m$) writes,

$$\mathcal{L}(\lambda, \beta, \mu) = f(\beta) + \left\langle \lambda, \int_Z g d\mu - \beta \right\rangle.$$

Then, the dual problem of ($\tilde{\mathrm{P}}_m$) is,

$$\sup_{\lambda \in \mathcal{H}} \inf_{\beta \in \mathcal{H},\, \mu \in \mathcal{P}_m(Z)} \mathcal{L}(\lambda, \beta, \mu) = \sup_{\lambda \in \mathcal{H}} \left(-f^*(\lambda) + \inf_{\mu \in \mathcal{P}_m(Z)} \int_Z \langle \lambda, g(z) \rangle \, d\mu(z) \right), \tag{6.1}$$

where f^* is the Fenchel conjugate of f. For any $\lambda \in \mathcal{H}$, since g is bounded over Z, the second term $\inf_{\mu \in \mathcal{P}_m(Z)} \int_Z \langle \lambda, g(z) \rangle \, d\mu(z)$ is finite. Therefore, it suffices to study (6.1) for $\lambda \in \mathrm{dom}(f^*)$, *i.e.*,

$$\sup_{\lambda \in \mathrm{dom}(f^*)} \left(-f^*(\lambda) + \inf_{\mu \in \mathcal{P}_m(Z)} \int_Z g_\lambda d\mu \right).$$

The result of Lemma 3.2 holds true for all $\lambda \in \mathrm{dom}(f^*)$ under Assumption $\mathbf{A}^*$. Applying it to the previous problem, we obtain the following equivalent dual problem:

$$-\inf_{\lambda \in \mathrm{dom}(f^*)} \mathcal{D}_m(\lambda) \coloneqq f^*(\lambda) - \int_X u_\lambda dm. \tag{D_m}$$

Lemma 6.1. *The function $\mathcal{D}_m$ is strongly convex with modulus $1/L$. As a consequence, problem (D_m) has a unique solution, denoted by $\lambda^*(m)$. Moreover, there exists a constant C^* independent of m such that*

$$\|\lambda^*(m)\| \leq C^*.$$

Proof. Since ∇f is L-Lipschitz continuous, we know that f^* is strongly convex with modulus $1/L$ (*i.e.* $f^* - 1/(2L)\|\cdot\|^2$ is convex) (see [40], Thm. 18.15). Let us consider $u_\lambda(x)$ as a function of λ while fixing any $x \in X$. By definition, $\lambda \mapsto u_\lambda(x)$ is the infimum of a family of affine functions (with respect to λ), thus it is concave with respect to λ. Consequently, $-\int_X u_\lambda dm$ is convex with respect to λ. Therefore, $\mathcal{D}_m$ is $1/L$-strongly convex. Additionally, $\mathrm{dom}(f^*)$ is both convex and closed. These properties guarantee the existence and uniqueness of the minimizer $\lambda^*(m)$.

Since M is an upper bound of $\|g(z)\|$, it follows that for all $\lambda \in \mathcal{H}$:

$$-M\|\lambda\| \leq \inf_{y\in Z_x} -\|\lambda\|\|g(x,y)\| \leq u_\lambda(x) \leq \sup_{y\in Z_x} \|\lambda\|\|g(x,y)\| \leq M\|\lambda\|.$$

Let $\lambda_0 \in \mathrm{dom}(f^*)$. As $\mathcal{D}_m(\lambda^*(m)) \leq \mathcal{D}_m(\lambda_0)$, we can derive the following inequalities:

$$f^*(\lambda_0) + M\|\lambda_0\| \geq \mathcal{D}_m(\lambda_0) \geq \mathcal{D}_m(\lambda^*(m)) \geq f^*(\lambda^*(m)) - M\|\lambda^*(m)\|.$$

The strong convexity of f^* yields that

$$\frac{1}{2L}\|\lambda^*(m) - \lambda_0\|^2 + \langle p_0, \lambda^*(m) - \lambda_0\rangle \leq f^*(\lambda^*(m)) - f^*(\lambda_0),$$

where $p_0 \in \partial f^*(\lambda_0)$. Combining the two above inequalities, we obtain:

$$\frac{1}{2L}\|\lambda^*(m) - \lambda_0\|^2 + \langle p_0, \lambda^*(m) - \lambda_0\rangle \leq M(\|\lambda^*(m)\| + \|\lambda_0\|).$$

The announced result follows, with $C^* = 3\|\lambda_0\| + 2L(M + \|p_0\|)$. □

6.2. Strong duality

Let us now prove the strong duality principle between ($\mathrm{P}^{\mathrm{K}}_m$) and ($\mathrm{D}_m$), *i.e.*, $\mathbf{val}(\mathrm{P}^{\mathrm{K}}_m) = \mathbf{val}(\mathrm{D}_m)$. We apply the Fenchel–Rockafellar theorem [41].

Proposition 6.2. *The following statements hold:*

1. $\mathbf{val}(\mathrm{P}^{\mathrm{K}}_m) = \mathbf{val}(\mathrm{D}_m)$;
2. *Let μ be a solution of ($\mathrm{P}^{\mathrm{K}}_m$) and let $\lambda^*(m)$ be a solution of (D_m). Then*

$$\lambda^*(m) = \nabla f\left(\int_Z g\, d\mu\right).$$

Proof. Let $G_m := \{\int_\mu g d\mu \mid \mu \in \mathcal{P}_m(Z)\} \subseteq \mathcal{H}$ and let $\overline{G}_m$ be its closure in $\mathcal{H}$. Consider the convex optimization problem on $\mathcal{H}$

$$\inf_{z\in\mathcal{H}} \; f(z) + \chi_{\overline{G}_m}(z) \;=\; \inf_{z\in\overline{G}_m} \; f(z). \tag{6.2}$$

Since f is continuous, we have $\inf_{z\in\overline{G}_m} f(z) = \inf_{z\in G_m} f(z)$.

The Fenchel dual of (6.2) reads

$$\sup_{\lambda\in\mathcal{H}} \; -f^*(\lambda) - \chi^*_{\overline{G}_m}(-\lambda). \tag{6.3}$$

Moreover,

$$-\chi^*_{\overline{G}_m}(-\lambda) = \inf_{z\in\overline{G}_m} \langle \lambda, z\rangle = \inf_{z\in G_m} \langle \lambda, z\rangle = \inf_{\mu\in\mathcal{P}_m(Z)} \Big\langle \lambda, \int_Z g\, d\mu \Big\rangle,$$

since $z \mapsto \langle \lambda, z\rangle$ is continuous on $\mathcal{H}$ and thus attains the same infimum on a set and its closure. Therefore, $\mathbf{val}(\mathrm{D}_m) = \mathbf{val}(6.3)$ and $\mathbf{val}(\mathrm{P}^{\mathrm{K}}_m) = \mathbf{val}(6.2)$.

We now apply Fenchel–Rockafellar to (6.2). The function f is convex and continuous. The indicator $\chi_{\overline{G}_m}$ is convex and lower semicontinuous since $\overline{G}_m$ is convex and closed. Finally, $\overline{G}_m$ is nonempty (because $\mathcal{P}_m(Z) \neq \emptyset$), hence $\mathrm{dom}(f) - \mathrm{dom}(\chi_{\overline{G}_m}) = \mathcal{H} - \overline{G}_m = \mathcal{H}$ and thus $0 \in \mathrm{int}(\mathcal{H} - \overline{G}_m)$. The Fenchel–Rockafellar theorem yields

$$\mathbf{val}(6.2) = \mathbf{val}(6.3),$$

hence $\mathbf{val}(\mathrm{P}^{\mathrm{K}}_m) = \mathbf{val}(\mathrm{D}_m)$.

Next, assume that μ solves $(\mathrm{P}^{\mathrm{K}}_m)$ and $\lambda^*(m)$ solves (D_m). By strong duality and the above identification of the dual,

$$-f^*(\lambda^*(m)) + \inf_{z\in\overline{G}_m} \langle \lambda^*(m), z\rangle = f\left(\int_Z g\, d\mu\right),$$

and since $\int_Z g\, d\mu \in G_m \subset \overline{G}_m$, we have

$$-f^*(\lambda^*(m)) + \Big\langle \lambda^*(m), \int_Z g\, d\mu \Big\rangle = f\left(\int_Z g\, d\mu\right).$$

This is exactly Fenchel's equality at $z = \int_Z g\, d\mu$, hence $\lambda^*(m) = \nabla f(z)$, *i.e.*, $\lambda^*(m) = \nabla f\big(\int_Z g\, d\mu\big)$. □

6.3. Stability of the dual solution

Lemma 6.3. *For any $\lambda_1, \lambda_2 \in \mathcal{H}$ and $x_1, x_2 \in X$, it holds that*

$$|u_{\lambda_1}(x_1) - u_{\lambda_2}(x_2)| \leq L_g\|\lambda_1\| d_X(x_1, x_2) + M\|\lambda_1 - \lambda_2\|.$$

Proof. By the triangle inequality,

$$|u_{\lambda_1}(x_1) - u_{\lambda_2}(x_2)| \leq |u_{\lambda_1}(x_1) - u_{\lambda_1}(x_2)| + |u_{\lambda_1}(x_2) - u_{\lambda_2}(x_2)|.$$

By the definition of u_λ, we have

$$u_{\lambda_1}(x_1) - u_{\lambda_1}(x_2) = \inf_{y_1\in Z_{x_1}} \langle \lambda_1\,,\, g(x_1, y_1)\rangle - \inf_{y_2\in Z_{x_2}} \langle \lambda_1\,,\, g(x_2, y_2)\rangle.$$

Let $\tilde{y}_2^\epsilon$ be an ϵ-minimizer of $\inf_{y_2\in Z_{x_2}} \langle \lambda_1\,,\, g(x_2, y_2)\rangle$, with $\epsilon > 0$. By the Lipschitz continuity of $\mathcal{Z}$, there exists $\tilde{y}_1^\epsilon \in Z_{x_1}$ such that

$$\|g(x_1, \tilde{y}_1^\epsilon) - g(x_2, \tilde{y}_2^\epsilon)\| \leq L_g d_X(x_1, x_2).$$

By the Cauchy-Schwarz inequality, we have

$$u_{\lambda_1}(x_1) - u_{\lambda_1}(x_2) \leq \langle \lambda_1\,,\, g(x_1, \tilde{y}_1^\epsilon) - g(x_2, \tilde{y}_2^\epsilon)\rangle + \epsilon \leq \|\lambda_1\| L_g d_X(x_1, x_2) + \epsilon.$$

By the arbitrariness of ϵ, we have $|u_{\lambda_1}(x_1) - u_{\lambda_1}(x_2)| \leq \|\lambda_1\| L_g d_X(x_1, x_2)$.

On the other hand,

$$u_{\lambda_1}(x_2) - u_{\lambda_2}(x_2) = \inf_{y \in Z_{x_2}} \langle \lambda_1 \, , \, g(x_2, y) \rangle - \inf_{y \in Z_{x_2}} \langle \lambda_2 \, , \, g(x_2, y) \rangle \leq \sup_{y \in Z_{x_2}} \langle \lambda_1 - \lambda_2 \, , \, g(x_2, y) \rangle.$$

By the Cauchy–Schwarz inequality and the definition of M, we have that

$$u_{\lambda_1}(x_2) - u_{\lambda_2}(x_2) \leq M \|\lambda_1 - \lambda_2\|.$$

The conclusion follows. □

Lemma 6.4 (Stability of the dual problem)**.** *For any m_0, $m_1 \in \mathcal{P}(\Omega)$, we have*

$$|\mathcal{D}_{m_0}(\lambda^*(m_0)) - \mathcal{D}_{m_1}(\lambda^*(m_1))| \leq C^* L_g \, d_1(m_0, m_1), \tag{6.4}$$

$$\|\lambda^*(m_0) - \lambda^*(m_1)\|^2 \leq 2C^* L_g L \, d_1(m_0, m_1), \tag{6.5}$$

where C^* is the a priori bound of $\|\lambda^*(\cdot)\|$ obtained in Lemma 6.1

Proof. According to Lemma 6.1, we know that $\|\lambda^*(m_0)\|$ and $\|\lambda^*(m_1)\|$ are smaller than C^*. Then, by Lemma 6.3, $u_{\lambda^*(m_0)}(x)$ and $u_{\lambda^*(m_1)}(x)$ are $(C^* L_g)$-Lipschitz continuous with respect to x. Hence,

$$\begin{aligned} \mathcal{D}_{m_0}(\lambda^*(m_0)) &= f^*(\lambda^*(m_0)) - \int_X u_{\lambda^*(m_0)}(x) dm_0(x) \\ &= f^*(\lambda^*(m_0)) - \int_X u_{\lambda^*(m_0)}(x) dm_1(x) + \int_X u_{\lambda^*(m_0)}(x) d(m_1 - m_0)(x) \\ &\geq \mathcal{D}_{m_1}(\lambda^*(m_0)) - C^* L_g d_1(m_0, m_1), \end{aligned} \tag{6.6}$$

where the third line is by the definition of the Kantorovich–Rubinstein distance. Since $\lambda^*(m_1)$ minimizes $\mathcal{D}_{m_1}$ and since $\mathcal{D}_{m_1}$ is $1/L$-strongly convex, we have

$$\mathcal{D}_{m_1}(\lambda^*(m_0)) \geq \mathcal{D}_{m_1}(\lambda^*(m_1)) + \frac{1}{2L} \|\lambda^*(m_0) - \lambda^*(m_1)\|^2. \tag{6.7}$$

Combining (6.6) and (6.7), we obtain that

$$\mathcal{D}_{m_0}(\lambda^*(m_0)) \geq \mathcal{D}_{m_1}(\lambda^*(m_1)) + \frac{1}{2L} \|\lambda^*(m_0) - \lambda^*(m_1)\|^2 - C^* L_g d_1(m_0, m_1).$$

In particular, we have $\mathcal{D}_{m_1}(\lambda^*(m_1)) - \mathcal{D}_{m_0}(\lambda^*(m_0)) \leq C^* L_g d_1(m_0, m_1)$. Exchanging the positions of m_0 and m_1 in (6.6), we obtain

$$\mathcal{D}_{m_1}(\lambda^*(m_1)) \geq \mathcal{D}_{m_0}(\lambda^*(m_0)) + \frac{1}{2L} \|\lambda^*(m_0) - \lambda^*(m_1)\|^2 - C^* L_g d_1(m_0, m_1). \tag{6.8}$$

Inequality (6.4) follows immediately and (6.5) is deduced by summing (6.6)–(6.8). □

6.4. First variation of the value function

The value function of problem ($\mathrm{P}_m^{\mathrm{K}}$) is defined by

$$V \colon \mathcal{P}^1(X) \to \mathbb{R}, \ m \mapsto \mathbf{val}(\mathrm{P}_m^{\mathrm{K}}).$$

Our goal is to characterize the first variation of V. Define the following function:

$$v\colon \mathcal{P}^1(X)\times X\to\mathbb{R},\ (m,x)\mapsto u_{\lambda^*(m)}(x).$$

Proposition 6.5. *For any $m_0,m_1\in\mathcal{P}^1(X)$, we have*

$$\lim_{t\to 0^+}\frac{V(m_0+t(m_1-m_0))-V(m_0)}{t}=\int_X v(m_0,x)d(m_1-m_0)(x).$$

As a consequence, v is the first variation of V, i.e.,

$$V(m_1)-V(m_0)=\int_{t=0}^1\int_X v(m_0+t(m_1-m_0),x)d(m_1-m_0)(x)\mathrm{d}t.$$

Proof. For $t\in[0,1]$, set $m_t\coloneqq m_0+t(m_1-m_0)$ and let $\lambda_t\coloneqq\lambda^*(m_t)$. By the strong duality (Prop. 6.2), we have that for any $m\in\mathcal{P}^1(X)$,

$$V(m)=-\mathcal{D}_m(\lambda^*(m)),\qquad\text{with}\quad \mathcal{D}_m(\lambda)=f^*(\lambda)-\int_X u_\lambda(x)\,dm(x).$$

Since λ_t minimizes $\mathcal{D}_{m_t}$, we have

$$\mathcal{D}_{m_t}(\lambda_t)\le\mathcal{D}_{m_t}(\lambda_0),\qquad\text{hence}\qquad V(m_t)\ge-\mathcal{D}_{m_t}(\lambda_0).$$

Using the linear dependence of $\mathcal{D}_m(\lambda)$ on m, this yields

$$V(m_t)-V(m_0)\ge-\mathcal{D}_{m_t}(\lambda_0)+\mathcal{D}_{m_0}(\lambda_0)=\int_X u_{\lambda_0}(x)\,d(m_t-m_0)(x)=t\int_X v(m_0,x)\,d(m_1-m_0)(x).$$

Similarly, since λ_0 minimizes $\mathcal{D}_{m_0}$, we have $\mathcal{D}_{m_0}(\lambda_0)\le\mathcal{D}_{m_0}(\lambda_t)$, *i.e.*, $V(m_0)\ge-\mathcal{D}_{m_0}(\lambda_t)$. Therefore,

$$V(m_t)-V(m_0)\le-\mathcal{D}_{m_t}(\lambda_t)+\mathcal{D}_{m_0}(\lambda_t)=\int_X u_{\lambda_t}(x)\,d(m_t-m_0)(x)=t\int_X v(m_t,x)\,d(m_1-m_0)(x).$$

Altogether, for all $t\in(0,1]$,

$$\int_X v(m_0,x)\,d(m_1-m_0)(x)\ \le\ \frac{V(m_t)-V(m_0)}{t}\ \le\ \int_X v(m_t,x)\,d(m_1-m_0)(x).\tag{6.9}$$

By Lemmas 6.3–6.4, the map $(m,x)\mapsto v(m,x)$ is continuous on $\mathcal{P}^1(X)\times X$ (for (d_1,d_X)). Hence

$$\int_X v(m_t,x)\,d(m_1-m_0)(x)\longrightarrow\int_X v(m_0,x)\,d(m_1-m_0)(x)\qquad\text{as } t\to 0^+.$$

Letting $t\to 0^+$ in (6.9) gives the claimed right derivative formula.

Define $\bar V(t)\coloneqq V(m_t)$ and

$$\bar v(t)\coloneqq\int_X v(m_t,x)\,d(m_1-m_0)(x).$$

Applying (6.9) with base point m_t and increment $m_{t+h} - m_t = h(m_1 - m_0)$ yields, for any $h > 0$,

$$\bar{v}(t) \leq \frac{\bar{V}(t+h) - \bar{V}(t)}{h} \leq \bar{v}(t+h).$$

Since $\bar{v}$ is continuous, it follows that $\bar{V}$ is differentiable on $[0,1]$ with $\bar{V}'(t) = \bar{v}(t)$. The fundamental theorem of calculus [24], Theorem 7.21 then gives

$$V(m_1) - V(m_0) = \bar{V}(1) - \bar{V}(0) = \int_0^1 \bar{v}(t)\,\mathrm{d}t,$$

which is exactly the announced formula. □

7. Numerical approach

7.1. General approach

We present in this section our numerical method for solving ($\mathrm{P}^{\mathrm{K}}_m$). The main difficulty lies in the fact that the problem is infinite dimensional (unless Z is a finite set). Thus we propose a first simplification of the problem, motivated by the bridging method, and consisting in discretizing the measure m: we replace it by an empirical measure m_N of the form $m_N = \sum_{i=1}^N \delta_{x_i}/N$. We discuss in Section 7.2 the effect of such a discretization and provide quantization results of the literature.

We next discuss in Section 7.3 the resolution of the discretized problem ($\mathrm{P}^{\mathrm{K}}_{m_N}$). Remark 2.7 shows that ($\mathrm{P}^{\mathrm{K}}_{m_N}$) is equivalent to

$$\inf_{(\nu_1,\dots,\nu_N)\in\prod_{i=1}^N \mathcal{P}(Z_{x_i})} f\left(\frac{1}{N}\sum_{i=1}^N \int_X g(x_i,y)d\nu_i(y)\right). \tag{7.1}$$

We discuss the interest of solving it with the Frank–Wolfe algorithm. Our main motivation lies in the following two features:

- parallelizability: each iteration of the algorithm amounts to solve N independent problems of the form $\inf_{y_i\in Z_{x_i}} \langle \lambda, g_i(x_i,y_i)\rangle$, for a given $\lambda \in \mathcal{H}$.
- sparsity: at iteration k, the candidate solutions ν_i have a finite support, of cardinality atmost k.

The sparsity property of the Frank–Wolfe algorithm is limited by the fact that a large number of iterations may be required. For large values of k, the storage of kN support points may be problematic. On the other hand, when N is large, the relaxation gap (between ($\mathrm{P}^{\mathrm{K}}_{m_N}$) and its Monge counterpart) is small, of order $1/N$. We describe in Section 7.4 a method for solving ($\mathrm{P}^{\mathrm{K}}_{m_N}$) with a convergence guaranty, that both leverages the convexity of the problem and only requires the manipulation of Monge candidates, thus significantly mitigates the storage issue mentioned above (at most $2N$ points must be stored during the execution of the method). The method is taken from another work of ours [16] and is called Stochastic Frank–Wolfe algorithm. Another interest of the method is that when m_N can be expressed as the pushforward of the original distribution m, an approximate Monge solution can be obtained for ($\mathrm{P}^{\mathrm{K}}_m$).

7.2. Discretization

In view of estimate (4.5), one should look for an empirical distribution $m_N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ that is as close as possible to m for the d_1-distance. This problem is commonly known as the optimal quantization problem, and for detailed information on this topic, we refer to [42]. Here, we present a slightly modified version of an optimal

quantization result obtained in [15], Proposition 12. For any subset A of X, we denote by $r_N(A)$ the minimum radius r required to cover A with N closed balls of radius r. It is defined by

$$r_N(A) \coloneqq \inf_{x \in A^N} \min \left\{ r \geq 0 \,\middle|\, A \subseteq \bigcup_{i=1}^{N} B_X(x_i, r) \right\}.$$

The upper box-counting dimension (or the upper Minkowski dimension) of A [43], p. 41 is defined as follows:

$$\bar{D}(A) \coloneqq \inf \left\{ \bar{D} > 0 \,\middle|\, \exists \bar{C} > 0 \text{ such that } r_N(A) \leq \bar{C} N^{-1/\bar{D}},\ \forall N \in \mathbb{N}_+ \right\}.$$

Lemma 7.1. *Let $m \in \mathcal{P}^1(X)$, and let $A \subseteq X$ be the support of m. There exists a sequence $(m_N)_{N \geq 1}$ of empirical distributions on X such that the following holds:*

1. *If $\bar{D}(A) > 1$, then there exists a constant $\tilde{C}_1$ such that for any $N \geq 1$,*

$$d_1(m, m_N) \leq \tilde{C}_1 N^{-\frac{1}{\bar{D}(A)}}.$$

2. *If $\bar{D}(A) = 1$, then there exists a constant $\tilde{C}_2$ such that for any $N \geq 1$,*

$$d_1(m, m_N) \leq \tilde{C}_2 N^{-1} \log N.$$

3. *If $\bar{D}(A) < 1$, then there exists a constant $\tilde{C}_3$ such that for any $N \geq 1$,*

$$d_1(m, m_N) \leq \tilde{C}_3 N^{-1}.$$

Proof. This follows from the proof presented in [15], Proposition 12, with the only difference being that in the final inequality, we employ the triangle inequality for the d_1-distance instead of the Minkowski inequality for the Wasserstein-2 distance. □

Remark 7.2. If A is a subset of a smooth d-dimensional submanifold of a Euclidean space, then $\bar{D}(A) \leq d$. This estimate is deduced from [43], p. 48 (i)–(ii).

7.3. Frank–Wolfe algorithm

We discuss in this subsection (and the next one) the resolution of the problem ($\mathrm{P}^{\mathrm{K}}_{m_N}$), written in its equivalent form (7.1). For general convex optimization problems, the Frank–Wolfe algorithm relies on the resolution of a sequence of linearized problems, obtained by replacing the cost function of the problem by a first-order Taylor approximation of it. In the context of problem (7.1), the linearized problem is of the general form:

$$\inf_{\nu \in \prod_{i=1}^N \mathcal{P}(Z_{x_i})} \sum_{i=1}^{N} \left\langle \lambda, \int_Z g_i d\nu_i \right\rangle, \tag{7.2}$$

for some $\lambda \in \nabla f(\mathcal{H})$. The following result, which coincides with Lemma 3.2 in the simple case of a discrete distribution m, is straightforward. It shows that (7.2) amounts to solve N independent problems, posed on the sets Z_{x_i}.

Lemma 7.3. *Let Assumption A hold true. For any $\lambda \in \nabla f(\mathcal{H})$, the tuple $(\mathbf{BR}_\lambda(x_i))_{i=1,\dots,N}$ is a solution to (7.2).*

Algorithm 2: Frank–Wolfe algorithm.

Initialization: $\nu^0 \in \prod_{i=1}^N \mathcal{P}(Z_{x_i})$. Set $K \geq 1$.
for $k = 0, 1, 2, \ldots, K-1$ **do**
  Step 1: resolution of the linearized problem.
  Compute $\lambda^k = \nabla f \left(\frac{1}{N} \sum_{i=1}^N \int_{Z_{x_i}} g(x_i, \cdot) d\mu^k \right)$.
  for $i = 1, \ldots, N$ **do**
    Compute $y_i^k = \mathbf{BR}_{\lambda^k}(x_i)$.
  end for
  Step 2: update.
  Choose $\omega_k \in [0, 1]$.
  Set $\nu^{k+1} = (1 - \omega_k)\nu^k + \omega_k \left(\delta_{y_1^k}, \ldots, \delta_{y_N^k} \right)$.
end for

Remark 7.4. If we take $\omega_k = 1/(k+1)$ for all k, then it is easy to see that $\nu_i^K = \frac{1}{K} \sum_{k=0}^{K-1} \delta_{y_i^k}$. We recover the fictitious play of [20], applied to the Lagrangian discretization of first-order MFGs.

Lemma 7.5. *Let Assumption A hold true. In Algorithm 2, take $\omega_k = 2/(k+2)$ for all k. Then for any $K \geq 1$,*

$$f\left(\int_Z g d\mu^K \right) - \mathbf{val}(\mathrm{P}^{\mathrm{K}}_{m_N}) \leq \frac{2LD}{K}.$$

Proof. This is a consequence of [16], Proposition 3.4. □

7.4. Stochastic Frank–Wolfe algorithm

We discuss in this Subsection the computation of an approximate solution to $(\mathrm{P}^{\mathrm{K}}_{m_N})$ in a Monge form. Lemma 5.1 indeed implies that for any $\varepsilon > \frac{LD}{2N}$, there exists an N-uplet $(y_1, \ldots, y_N)$ in $\prod_{i=1}^N Z_{x_i}$ such that

$$f\left(\frac{1}{N} \sum_{i=1}^N g(x_i, y_i) \right) \leq \mathbf{val}(\mathrm{P}^{\mathrm{K}}_{m_N}) + \varepsilon, \tag{7.3}$$

since we have $\|(1/N, \ldots, 1/N)\|_2^2 = 1/N$. The Stochastic Frank–Wolfe algorithm, introduced in [16] is a stochastic algorithm whose efficiency is ensured in expectation: after $2N$ iterations, the expectation of the optimality gap is bounded by $2LD/N$ (see Lem. 7.6).

The Stochastic Frank–Wolfe algorithm is described in Algorithm 3. In the statement of the algorithm, $\mathrm{Bern}(\omega)$ denotes the Bernoulli distribution with a parameter $\omega \in [0, 1]$.

We briefly comment here on Algorithm 3 and refer the reader to [16] for a more detailed description and possible variants and improvements. At the beginning of iteration k, a candidate $y^k = (y_1^k, \ldots, y_N^k)$ to optimality is available, corresponding to the N-uplet $(\delta_{y_1^k}, \ldots, \delta_{y_N^k})$. Having in mind this correspondence, we see that Step 1 of Algorithm 3 is the same as Step 1 of Algorithm 2. The main difference lies in the second step. For Algorithm 3, n_k candidates $\hat{y}^{k,j}$, $j = 1, \ldots, n_k$ are computed with stochastic simulations: for each i, $\hat{y}_i^{k,j}$ is equal to y_i^k (the i-th coordinate of the previous candidate) with probability $1 - \omega_k$ and $\hat{y}_i^{k,j}$ is equal to $\bar{y}_i^k$ with probability ω_k. Very roughly speaking, the convex combination of probability measures is transformed into a convex combination "in probability".

We have the following convergence result.

Algorithm 3: Stochastic Frank–Wolfe algorithm.

Initialization: $y^0 \in \prod_{i=1}^N Z_{x_i}$. Set $K \geq 1$.
for $k = 0, 1, 2, \ldots, K-1$ **do**
  Step 1: resolution of the linearized problem.
  Compute $\lambda^k = \nabla f(\frac{1}{N}\sum_{i=1}^N g(x_i, y_i^k))$.
  for $i = 1, 2, \ldots, N$ **do**
    Find $\bar{y}_i^k \in \mathbf{BR}_{\lambda^k}(x_i)$.
  end for
  Step 2: update.
  Choose $n_k \in \mathbb{N}^*$. Set $\omega_k = 2/(k+2)$.
  for $j = 1, 2, \ldots, n_k$ **do**
    for $i = 1, 2, \ldots, N$ **do**
      Simulate $P_i^{k,j} \sim \mathrm{Bern}(\omega_k)$, independently of all previously defined random variables.
      Set $\hat{y}_i^{k,j} = (1 - P_i^{k,j}) y_i^k + P_i^{k,j} \bar{y}_i^k$.
    end for
    Define $\hat{y}^{k,j} = (\hat{y}_i^{k,j})_{i=1,\ldots,N}$.
  end for
  Find $y^{k+1} \in \mathrm{argmin}\left\{ f(\frac{1}{N}\sum_{i=1}^N g(x_i, y_i)) \,\middle|\, y \in \{\hat{y}^{k,j},\, j = 1, 2, \ldots, n_k\}\right\}$.
end for

Lemma 7.6. *Let Assumption A hold true. In Algorithm 3, at each iteration k, take any number n_k of stochastic simulations. Then for any $K = 1, 2, \ldots, 2N$,*

$$\mathbb{E}\left[f\left(\frac{1}{N} \sum_{i=1}^N g(x_i, y_i^K) \right) \right] - \mathbf{val}(\mathrm{P}^{\mathrm{K}}_{m_N}) \leq \frac{4LD}{K}.$$

Proof. This is from [16], Theorem 3.7. □

Remark 7.7. Lemma 7.6 provides an expected convergence result for Algorithm 3. We complete it with a probability estimate, based on [16], Corollary 3.8. Recall the constants D and C defined in (3.1). Fix any $A > 0$ and any $K = 1, \ldots, 2N$. Choose integers $n_k \geq 1$ such that

$$n_k \geq \frac{Ak^2}{N}, \qquad k = 1, \ldots, K.$$

Then, it holds that

$$\mathbb{P}\left[f\left(\frac{1}{N} \sum_{i=1}^N g(x_i, y_i^K) \right) \leq \mathbf{val}(\mathrm{P}^{\mathrm{K}}_{m_N}) + \frac{4LD + C\sqrt{D}}{K} \right] \geq 1 - \exp\left(-\frac{A}{12} \right).$$

We make the following comments:

1. For fixed A and N, the lower bound on n_k is very mild at early iterations (and may even be 1), but it grows quadratically in k. Intuitively, at the beginning the optimization error dominates, whereas in later iterations the sampling noise becomes the main bottleneck; increasing n_k is then necessary to control the stochastic error.
2. For fixed N, increasing A raises the required n_k at each iteration, while improving the success probability, which approaches 1 as A grows.

3. For fixed A, increasing N reduces the required sample size n_k at each iteration while maintaining the same high-probability guarantee. This indicates that larger N improves the algorithm's performance.

We conclude this subsection with two estimates concerning the combination of the Stochastic Frank–Wolfe and the bridging method, under Assumptions A-B. There are two possibilities to retrieve an approximate solution of the original problem ($\mathrm{P}_m^{\mathrm{K}}$), using the outcome y^K of Algorithm 3 after K iterations, for $1 \leq K \leq 2N$ and for arbitrary numbers $n_k \geq 1$ of simulations.

- Let $\mu_N^K = \frac{1}{N}\sum_{i=1}^N \delta_{(x_i, y_i^K)}$. Apply Algorithm 1, with $m_0 = m_N$, $m_1 = m$, $\mu_0 = \mu_N^K$ and $\rho \in \Pi(m_N, m)$. Calling $\tilde{\mu}^K$ the result, we have the following estimate from Lemmas 4.5 and 7.6:

$$\mathbb{E}\left[f\left(\int_Z g d\tilde{\mu}^K\right)\right] - \mathbf{val}(\mathrm{P}_m^{\mathrm{K}}) \leq \frac{4LD}{K} + L_g C \int_{X\times X} d_X(x_0, x_1) d\rho(x_0, x_1). \tag{7.4}$$

- In the case where $m_N = U\#m$ (for example, under the assumptions of Corollary 5.7), an approximate solution in Monge form T^K is obtained by defining first $I\colon x \in X \mapsto I(x) \in \{1, \dots, N\}$, where $I(x)$ is such that $U(x) = x_{I(x)}$. Then T^K is defined by given by

$$T^K(x) = s(x_{I(x)}, y^K_{I(x)}, x)$$

and we have the following estimate from Lemmas 5.2 and 7.6:

$$\mathbb{E}\left[f\left(\int_X g \circ (I, T^K) d\tilde{\mu}^K\right)\right] - \mathbf{val}(\mathrm{P}_m^{\mathrm{K}}) \leq \frac{4LD}{K} + L_g C \int_X d_X(x, U(x)) dm(x). \tag{7.5}$$

Remark 7.8. It is worth noting that the error estimates in (7.4) and (7.5) comprise two terms. The first term, $4LD/K$, is independent of the sampling (discretization) of m, while the second term is independent of the optimization algorithm. This general principle is commonly referred to as mesh independence in numerical analysis.

7.5. Numerical perspectives through semi-discrete optimal transportation

We provide here a constructive approach from semi-discrete optimal transportation to obtain the transport mapping U used in (7.5). Assume that X is a closed subset of a Euclidean space (d_X is the Euclidean distance $\|\cdot\|$) and $m \in \mathbb{L}^\infty(X)$. Following [44], we define the following function on $\mathbb{R}^N$, referred to as Kantorovich's functional:

$$\psi(W) \coloneqq \int_X \min_i \left(\|x_i - x\|^2 + W_i\right) dm(x) - \frac{1}{N}\sum_{i=1}^N W_i.$$

It is proved in [44], Theorem 1.1 that $\psi(W)$ is concave and C^1. Let W^* be a global maximizer of ψ. Excluding a negligible set with respect to m, we can define the following centered partition $P = (\Omega_i, x_i)_{i=1,\dots,N}$ associated with W^*:

$$\Omega_i \coloneqq \{x \in X \mid \|x - x_i\| + W_i^* \leq \|x - x_j\| + W_j^*, \text{ for } j = 1, \dots, N\}.$$

This centered partition P is commonly referred to in the literature as the Laguerre tessellation [44] or the power diagram [45] associated with W^* (reducing to the Voronoi diagram when $W^* = 0$). Then, we define the mapping U as in Definition 5.5:

$$U(x) = x_i, \quad \text{for } x \in \Omega_i.$$

It follows from [44], Corollary 1.2 and [22] that

$$\int_X \|x - U(x)\|^2 dm(x) = \mathbb{W}_2^2(m, m_N),$$

where $\mathbb{W}_2$ is the Wasserstein-2 distance. Combining this with the estimate above this subsection and applying the Cauchy–Schwarz inequality, we obtain that

$$\mathbb{E}\left[f\left(\int_X g \circ (I, T^K) d\tilde{\mu}^K\right)\right] - \mathbf{val}(\mathrm{P}_m^{\mathrm{K}}) \leq \frac{4LD}{K} + L_g C\, \mathbb{W}_2(m, m_N).$$

Finally, by applying the quantization result (in the expectation sense) from [46], we obtain that

$$\mathbb{E}_{\mathrm{Algo}, m_N \sim m}\left[f\left(\int_X g \circ (I, T^K) d\tilde{\mu}^K\right)\right] - \mathbf{val}(\mathrm{P}_m^{\mathrm{K}}) \leq \frac{4LD}{K} + L_g \tilde{C} \times \begin{cases} N^{-1/2}, & \text{if } d \leq 3, \\ N^{-1/2} \log N, & \text{if } d = 4, \\ N^{-2/d}, & \text{otherwise.} \end{cases}$$

Here, the constant $\tilde{C}$ is independent of N, d denotes the dimension of X, and $\mathbb{E}_{\mathrm{Algo}, m_N \sim m}$ represents the expectation with respect to the stochastic steps in Algorithm 3 and the random points x_i, which are i.i.d. samples drawn from m.

Remark 7.9 (Numerical algorithms for finding W^*)**.** The key step in constructing the centered partition P using the scheme presented above is to determine a maximizer W^* of Kantorovich's functional ψ. Since ψ is concave and continuously differentiable, several numerical methods have been proposed for computing W^*. For instance, the damped Newton algorithm introduced in [44] is shown to have a linear convergence rate under certain regularity assumptions on X. Additionally, multiscale approaches in [47] and [45] can also been utilized. A sensitivity analysis of the ultimate error in our scheme and the inexactness of W^* presents an interesting perspective.

8. Examples and numerical results

8.1. The traffic assignment problem

The traffic assignment problem is a non-atomic game whose potential formulation takes the form of problem ($\mathrm{P}_m^{\mathrm{K}}$). We describe it briefly in this subsection. Consider a finite set of nodes $\mathcal{N}$ and a finite set of edges $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$. The model is a static model that describes how the agents move on the network, taking into account their origins and destinations as well as the congestion on each arc.

We fix a subset X of $\mathcal{N} \times \mathcal{N}$. Each parameter $x = (x_1, x_2) \in X$ represents an origin-destination pair. Next, we denote by Y the set of subsets of $\mathcal{E}$. For each x, we consider a set of possible paths connecting x_1 and x_2, denoted Z_x. Mathematically, we simply describe a path as a subset of $\mathcal{E}$, so $Z_x \subseteq Y$.

For the definition of the potential problem, we define $\mathcal{H} = \mathbb{R}^{\mathcal{E}}$. The function g is defined by $g\colon (x, y) \in Z \mapsto (g(x, y)_e)_{e \in \mathcal{E}} \in \mathcal{H}$, where

$$g(x, y)_e = \begin{cases} 1, & \text{if } e \in y, \\ 0, & \text{otherwise.} \end{cases}$$

In words, $g(x, y)_e = 1$ if the edge e belongs to the path y, 0 otherwise. Next, we fix a family of functions $\phi_e\colon [0, \infty) \to [0, \infty)$, parametrized by $e \in \mathcal{E}$. We assume that these functions are non-decreasing and we fix a

primitive Φ_e for each of them. The functions Φ_e are convex. Finally, we define $f\colon \mathcal{H} \to \mathbb{R}$ by

$$f(q) = \sum_{e\in\mathcal{E}} \Phi_e(q_e).$$

With these definitions at hand, it remains to interpret the optimality conditions for the associated MFO problem. Let us consider $\bar{\mu} \in \mathcal{P}_m(Z)$ and let $\bar{\lambda} = \nabla f(\int_Z g d\bar{\mu}) \in \mathcal{H}$. A direct calculation shows that

$$\bar{\lambda}_e = \phi_e(q_e), \quad \text{where: } q_e = \sum_{\substack{(x,y)\in Z\\ e\in y}} \bar{\mu}(x,y).$$

We can interpret q_e as the proportion of agents using the edge e. We interpret ϕ_e as a function that gives the travelling time on the edge e in function of the congestion q_e. So here the dual variable $\bar{\lambda}$ has a natural interpretation as a vector containing all the travelling times of the network. Finally, for any $x \in X$, we have

$$\inf_{y\in Z_x} \langle \bar{\lambda}, g(x,y)\rangle = \inf_{y\in Z_x} \sum_{e\in y} \bar{\lambda}_e.$$

Here $\sum_{e\in y} \bar{\lambda}_e$ describes the total duration of the path y. As a consequence, solving the MFO problem is equivalent to find $\bar{\mu}$ such that for any x, $\bar{\mu}_x$ is supported by the optimal paths (among those connecting x_1 to x_2), the travel time of a path y begin defined by the above relations. This notion of equilibrium is known as Wardrop equilibrium in the literature. Using the MFO setting, we recover the well-known equivalence between Wardrop equilbria and their potential formulation, see [48], Chapter 3. We mention here that the modelling is different (but equivalent) to the standard one in which one rather describes the distribution of the agents with respect to the edges, instead of using the distribution with respect to the paths. We also note that the Frank–Wolfe algorithm is a very standard algorithm for solving those problems, see [48], Section 5.2.

8.2. Lagrangian mean-field games

Model

We present here a class of *potential* Lagrangian mean-field games (MFGs) that can be cast within the MFO framework. There is a large amount of literature on Lagrangian MFGs, we refer the reader to [1, 3–6] and the references therein.

Fix a final time $T > 0$, a compact set $\Omega \subset \mathbb{R}^d$ constraining the trajectories, and a bound $M > 0$ on the admissible speed. Let $\mathrm{AC}([0,T];\mathbb{R}^d)$ denote the space of absolutely continuous curves on $[0,T]$. Define the path space

$$Y \coloneqq \Big\{ y \in \mathrm{AC}([0,T];\mathbb{R}^d) \;\Big|\; y(t) \in \Omega,\ \forall t \in [0,T] \Big\}, \qquad Z_x \coloneqq \Big\{ y \in Y \;\Big|\; y(0) = x,\ \|\dot{y}\|_{L^\infty(0,T)} \le M \Big\}, \qquad x \in \Omega,$$

and set

$$X \coloneqq \Omega, \qquad Z \coloneqq \big\{ (x,y) \in X \times Y :\ y \in Z_x \big\}.$$

Let $m \in \mathcal{P}(\Omega)$ denote the distribution of initial states.

Let $L : \mathbb{R}^d \to \mathbb{R}$ be a running cost, $h : \mathbb{R}^d \to \mathbb{R}^k$ a feature map, and $\Phi : \mathbb{R}^k \to \mathbb{R}$ a potential. The associated potential Lagrangian MFG can be written as

$$\inf_{\mu\in\mathcal{P}_m(Z)} \left\{ \int_Z \left(\int_0^T L(\dot{y}(t))\,\mathrm{d}t \right) d\mu(x,y) \;+\; \int_0^T \Phi\Big(\int_Z h\big(y(t)\big)\, d\mu(x,y) \Big)\,\mathrm{d}t \right\}. \tag{8.1}$$

We briefly explain the interpretation of (8.1). The measure $\mu \in \mathcal{P}_m(Z)$ represents the joint law of an initial condition $x \sim m$ and an admissible trajectory $y \in Z_x$. The first term averages the individual running costs over the population, while the second term is a potential interaction cost that depends on the aggregated feature of the agents' positions at each time t.

Equation (8.1) fits ($\mathrm{P}_m^{\mathrm{K}}$). Indeed, define the Hilbert space $\mathcal{H} := \mathbb{R} \times L^2([0,T];\mathbb{R}^k)$ and the mapping $g : Z \to \mathcal{H}$ by

$$g(x,y) := \big(g_1(x,y), g_2(x,y)\big), \qquad g_1(x,y) := \int_0^T L(\dot y(t))\,\mathrm{d}t, \qquad g_2(x,y) := h \circ y,$$

and define $f : \mathcal{H} \to \mathbb{R}$ by

$$f(q_1,q_2) := q_1 + \int_0^T \Phi\big(q_2(t)\big)\,\mathrm{d}t.$$

Then, for any $\mu \in \mathcal{P}_m(Z)$,

$$f\left(\int_Z g\,d\mu\right) = \int_Z \left(\int_0^T L(\dot y(t))\,\mathrm{d}t\right) d\mu(x,y) \;+\; \int_0^T \Phi\left(\int_Z h\big(y(t)\big)\,d\mu(x,y)\right) \mathrm{d}t,$$

where $\int_Z g\,d\mu = (\int_Z g_1\,d\mu,\ \int_Z g_2\,d\mu)$ and the second integral is understood in the Bochner sense in $L^2([0,T];\mathbb{R}^k)$. Consequently, (8.1) is exactly of the form of ($\mathrm{P}_m^{\mathrm{K}}$).

Verification of the abstract assumptions

Assumption A and Assumption A* hold for the Lagrangian MFG model (8.1) provided that $\Omega \subset \mathbb{R}^d$ is compact and the admissible trajectories satisfy the speed constraint $\|\dot y\|_{L^\infty} \le M$, and that the data satisfy:

1. $L : \mathbb{R}^d \to \mathbb{R}$ is convex and lower semicontinuous, and $\sup_{\|v\|\le M} |L(v)| < +\infty$;
2. $h : \mathbb{R}^d \to \mathbb{R}^k$ is continuous;
3. $\Phi : \mathbb{R}^k \to \mathbb{R}$ is convex and C^1, and $\nabla\Phi$ is Lipschitz with constant L_Φ.

For instance, in the linear–quadratic setting, one may take L to be quadratic, h to be affine, and Φ to be a quadratic potential; these choices clearly satisfy the above conditions.

Let us verify Assumption A(1). The mapping $g = (g_1, g_2)$ is bounded on Z. Indeed, for any $(x,y) \in Z$ we have $\|\dot y\|_{L^\infty} \le M$, hence

$$|g_1(x,y)| = \left|\int_0^T L(\dot y(t))\,\mathrm{d}t\right| \le T \sup_{\|v\|\le M} |L(v)|, \qquad \|g_2(x,y)\|_{L^2} = \|h \circ y\|_{L^2} \le \sqrt{T} \sup_{z\in\Omega} \|h(z)\|.$$

Moreover, the functional $f(q_1,q_2) = q_1 + \int_0^T \Phi(q_2(t))\,\mathrm{d}t$ is convex and differentiable on $\mathcal{H} = \mathbb{R} \times L^2([0,T];\mathbb{R}^k)$, with

$$\nabla f(q_1,q_2) = \big(1, \nabla\Phi(q_2(\cdot))\big).$$

Since the Nemytskii operator $q_2 \mapsto \nabla\Phi(q_2(\cdot))$ is L_Φ-Lipschitz on $L^2([0,T];\mathbb{R}^k)$, it follows that ∇f is Lipschitz on $\mathcal{H}$ with modulus $L = L_\Phi$.

Let us now verify Assumption A(2) and Assumption A*. Let $\lambda \in \mathrm{dom}(f^*)$. Then necessarily $\lambda = (1, \lambda_2)$ with $\lambda_2 \in L^2([0,T];\mathbb{R}^k)$, and

$$g_\lambda(x,y) = \langle \lambda, g(x,y)\rangle_{\mathcal{H}} = g_1(x,y) + \langle \lambda_2, h \circ y\rangle_{L^2}.$$

Since h is continuous on the compact set Ω, the map $y \mapsto h \circ y$ is continuous with respect to the supremum norm on Y, and therefore $(x,y) \mapsto \langle \lambda_2, h \circ y\rangle_{L^2}$ is continuous on Z. Furthermore, by convexity and lower semicontinuity of L, the integral functional $y \mapsto \int_0^T L(\dot y(t))\,\mathrm{d}t$ is lower semicontinuous on the equi-Lipschitz class $\{\|\dot y\|_{L^\infty} \le M\}$. Consequently, g_λ is lower semicontinuous on Z.

Finally, for each $x \in X$, the set Z_x is compact in $C([0,T];\mathbb{R}^d)$ by the Arzelà–Ascoli theorem (boundedness follows from $y(t) \in \Omega$ and equicontinuity from the uniform Lipschitz bound). Hence $g_\lambda(x,\cdot)$ attains its minimum on Z_x, so $\mathbf{BR}_\lambda(x) \neq \emptyset$. Moreover, G_λ is lower semicontinuous since it admits the continuous selection $x \mapsto g_\lambda(x, y_x)$, where $y_x(t) \equiv x$ is the constant admissible trajectory. This proves Assumption A(2) for all $\lambda \in \mathrm{dom}(f^*)$, *i.e.* Assumption A*.

We finally verify Assumption B. The first two items of Assumption B can be easily checked with the arguments already presented above. For point (3), *i.e.*, the Lipschitz continuity of $\mathcal{Z}$, we need further assumptions on the constraint set Ω and h:

1. Ω is convex;
2. h is Lipschitz on Ω.

Let $D_\Omega := \mathrm{diam}(\Omega)$ and $L_\infty := \sup_{\|v\|\le M} |L(v)| < \infty$. Fix $x_0, x_1 \in \Omega$ and $y_0 \in Z_{x_0}$, and set $\delta := \|x_1 - x_0\|/M \le D_\Omega/M$. Define $y_1 \in Z_{x_1}$ by

$$y_1(t) := \begin{cases} x_1 + \dfrac{t}{\delta}(x_0 - x_1), & t \in [0,\delta], \\ y_0(t-\delta), & t \in [\delta, T]. \end{cases}$$

If $\delta \ge T$, then we take the first dynamic in the whole time horizon $[0,T]$. Then $y_1(0) = x_1$, $y_1(t) \in \Omega$ (by convexity of Ω), and $\|\dot y_1\|_{L^\infty} \le M$. Moreover,

$$|g_1(x_1,y_1) - g_1(x_0,y_0)| \le \int_0^\delta |L(\dot y_1(t))|\,\mathrm{d}t + \int_{T-\delta}^T |L(\dot y_0(t))|\,\mathrm{d}t \le 2L_\infty\,\delta = \frac{2L_\infty}{M}\,\|x_1 - x_0\|.$$

For the feature term, using Lipschitz continuity of h and $\|y_1(t) - y_0(t)\| \le 2\|x_1 - x_0\|$ on $[0,\delta]$ and $\|y_1(t) - y_0(t)\| = \|y_0(t-\delta) - y_0(t)\| \le M\delta = \|x_1 - x_0\|$ on $[\delta, T]$, we obtain

$$\|g_2(x_1,y_1) - g_2(x_0,y_0)\|_{L^2} = \|h\circ y_1 - h \circ y_0\|_{L^2} \le L_h\left(2\sqrt{\delta} + \sqrt{T}\right)\|x_1 - x_0\| \le C_h\,\|x_1 - x_0\|,$$

where $C_h := L_h\big(2\sqrt{D_\Omega/M} + \sqrt{T}\big)$. Combining the two estimates yields

$$\|g(x_1,y_1) - g(x_0,y_0)\|_{\mathcal{H}} \le L_g\,\|x_1 - x_0\|$$

for some $L_g > 0$ depending only on $(M, T, \Omega, L_\infty, L_h)$, which proves the L_g-Lipschitz property of $\mathcal{Z}$.

In summary, the Lagrangian MFG model considered here satisfies all the assumptions required throughout the theoretical part of the article. Note that the set Z is compact (by the Arzelà–Ascoli theorem), so the existence of a solution is also ensured (see Rem. 3.6).

Optimality conditions

Let us write the optimality conditions corresponding to Lagrangian MFG. Let $\bar{\mu} \in \mathcal{P}_m(Z)$ and let $\bar{\lambda} = \nabla f(\int_Z g d\bar{\mu})$. Then, $\bar{\lambda} = (\bar{\lambda}_1, \bar{\lambda}_2)$, with $\bar{\lambda}_1 = 1$ and

$$\bar{\lambda}_2(t) = \nabla\Phi\left(\int_Z h(y(t))d\bar{\mu}(x,y)\right).$$

For any $t \in [0,T]$, denote by $e_t \colon Z \to \mathbb{R}^d$ the mapping defined by $e_t(x,y) = y(t)$. Then $\bar{\lambda}_2$ is equivalently defined by

$$\bar{\lambda}_2(t) = \nabla\Phi\left(\int_\Omega h(y')d\bar{m}_t(y')\right), \quad \text{where: } m_t = e_t\#\bar{\mu}.$$

In this context, the best-response mapping $\mathbf{BR}_{\bar{\lambda}}(x)$ in (3.5) is equivalent to solving the following optimal control problem:

$$\inf_{y\in Z_x} \int_0^T \left[L(\dot{y}(t)) + \left\langle \nabla\Phi\left(\textstyle\int_\Omega h d\bar{m}_t\right), h(y(t))\right\rangle\right] \mathrm{d}t.$$

Let us note that the above problem is not convex in general. It could be solved by dynamic programming in the situation where the dimension of y is moderate.

8.3. Numerical results for a competition problem with a non-renewable resource

Model

We consider a Lagrangian MFG in which the agents exploit their own stock of an exhaustible resource. The model is taken from [23]. We fix a time horizon $[0,T]$ where $T \in [0,+\infty)$ (the case $T = \infty$ investigated in [23] is not considered here). The state variable of a representative agent is the level of the stock of resource at any time, denoted $(X_t^q)_{t\in[0,T]}$ and the control is the speed of extraction at any time, denoted q. The dynamic of a given producer with an initial position $x_0 \geq 0$ is described as follows:

$$X_t^q \coloneqq x_0 - \int_0^t q_\tau d\tau, \qquad t \in [0,T],$$

where $q_t \geq 0$, for any $t \in [0,T]$. We impose that $X_T^q \geq 0$, which implies that $X_t^q \geq 0$ at any time.

We define the set of aggregate production, denoted as $\mathcal{G}$, by

$$\mathcal{G} \coloneqq \left\{Q \in \mathbb{L}^2([0,T],\mathbb{R}) \mid 0 \leq Q(t) \leq \frac{1}{2}, \forall t \in [0,T]\right\}.$$

The price of the resource for this representative producer depends on its extracting speed and an aggregate production $Q \in \mathcal{G}$,

$$p_t \coloneqq 1 - q_t - \epsilon Q_t, \qquad t \in [0,T],$$

where $\epsilon \in (0,1)$ is a constant. The gain of this representative producer writes,

$$\int_0^T \mathrm{e}^{-rt} q_t(1 - q_t - \epsilon Q_t)\mathrm{d}t.$$

where $r \geq 0$ is a discount rate. Therefore, given an aggregate production $Q \in \mathcal{G}$ and an initial position $x_0 \geq 0$, we can formulate an optimal control problem associated with this representative producer,

$$\begin{cases} \inf\limits_{q \in \mathcal{G}} & J^Q(q) \coloneqq \displaystyle\int_0^T \mathrm{e}^{-rt} q_t (q_t - 1 + \epsilon Q_t) \mathrm{d}t; \\ \text{s.t.} & \displaystyle\int_0^T q_t \mathrm{d}t \leq x_0. \end{cases} \tag{8.2}$$

Lemma 8.1. *Problem (8.2) has a unique solution $q^Q(x_0)$. Moreover, $0 \leq q^Q(x_0)(t) \leq \frac{1}{2}$, for a.e. $t \in (0,T)$.*

Proof. It is easy to see that $\mathcal{G}$ is a non-empty and convex subset of $\mathbb{L}^2([0,T],\mathbb{R})$. Following [24], Theorem 3.12, if $(f_n \in \mathcal{G})_{n\geq 1}$ converges to f in $\mathbb{L}^2$ sense, then there exists a subsequence of $(f_n)_{n\geq 1}$ converges to f a.e. As a consequence, f lies in $\mathcal{G}$. Therefore, $\mathcal{G}$ is closed. Furthermore, by Hölder's inequality, we obtain that $\{q \in \mathbb{L}^2([0,T],\mathbb{R}) \mid \int_0^T q_t \mathrm{d}t \leq x_0\}$ is non-empty, convex and closed in $\mathbb{L}^2([0,T],\mathbb{R})$. It follows that the admissible set of problem (8.2) is non-empty, closed and convex in Hilbert space $\mathbb{L}^2([0,T],\mathbb{R})$. On the other hand, the cost function $J^Q(\cdot)$ is strongly convex. Then the existence of the solution of (8.2) comes from [49], Corollary 3.23 and the uniqueness is by the strong convexity of J^Q.

Let q be the solution to (8.2). Define $q'(t) = \min\{q(t), \frac{1}{2}\}$, for a.e. $t \in (0,T)$. Since $q' \leq q$, q' is also feasible for problem (8.2). Moreover, the running cost $q_0 \mapsto q_0(q_0 - 1 + \varepsilon Q_t)$ is increasing for $q_0 \geq \frac{1}{2}$. As a consequence, $J^Q(q') \leq J^Q(q)$. Therefore, q' is optimal, and since the solution is unique, we have $q = q'$, which proves that $q \leq \frac{1}{2}$. □

Let $m \in \mathcal{P}([0,+\infty))$ denote the distribution of the initial conditions of the producers. The aggregate production rate corresponding to q^Q is given by

$$Q_t^Q \coloneqq \int_0^\infty q_t^Q(x_0) dm(x_0), \qquad \forall t \in [0,T].$$

Following [23], we call NE a solution Q^* to the fix-point problem:

$$Q^* = Q^{Q^*}, \qquad Q^* \in \mathcal{G}. \tag{8.3}$$

Potential problem

In this paragraph, we find an optimization problem associated with the fixed point problem (8.3), which is a particular case of problem ($\mathrm{P}_m^{\mathrm{K}}$). Let us specify metric spaces and admissible sets in ($\mathrm{P}_m^{\mathrm{K}}$) associated with (8.3):

$$X = [0,\infty), \qquad Y = \mathcal{G}, \qquad F(x) = \left\{ q \in \mathcal{G} \mid \int_0^T q_t \mathrm{d}t \leq x \right\}, \qquad Z = \mathrm{Graph}(F), \qquad Z_x = F(x).$$

Let us define the separable Hilbert space $\mathbb{L}^2_{\mathrm{e}^{-rt}}([0,T])$ [24], Example 4.5(b):

$$\mathbb{L}^2_{\mathrm{e}^{-rt}}([0,T]) \coloneqq \left\{ \zeta \colon [0,T] \to \mathbb{R} \text{ is Lebesgue measurable} \,\middle|\, \int_0^T \mathrm{e}^{-rt} |\zeta(t)|^2 \mathrm{d}t < +\infty \right\},$$

with a scalar product,

$$\langle f_1, f_2 \rangle_{\mathbb{L}^2_{\mathrm{e}^{-rt}}([0,T])} = \int_0^T \mathrm{e}^{-rt} f_1(t) f_2(t) \mathrm{d}t.$$

It is easy to check that $Y = \mathcal{G} \subseteq \mathbb{L}^2_{\mathrm{e}^{-rt}}([0,T])$. Then, in ($\mathrm{P}^{\mathrm{K}}_m$), we set $\mathcal{H} = \mathbb{R} \times \mathbb{L}^2_{\mathrm{e}^{-rt}}([0,T])$,

$$g\colon Z \to \mathcal{H}, (x,q) \mapsto \left(\int_0^T \mathrm{e}^{-rt}(q_t^2 - q_t)\mathrm{d}t,\ q \right),$$
$$f\colon \mathcal{H} \to \mathbb{R}, (y_1, y_2) \mapsto y_1 + \frac{\epsilon}{2}\|y_2\|^2_{\mathbb{L}^2_{\mathrm{e}^{-rt}}([0,T])}.$$

Therefore, problem ($\mathrm{P}^{\mathrm{K}}_m$) associated with (8.3) writes:

$$\inf_{\mu \in \mathcal{P}_m(Z)} \int_Z \int_0^T \mathrm{e}^{-rt}(q_t^2 - q_t)\mathrm{d}t d\mu(x,q) + \frac{\epsilon}{2}\int_0^T \mathrm{e}^{-rt}\left(\int_Z q_t d\mu(x,q)\right)^2 \mathrm{d}t. \tag{8.4}$$

Proposition 8.2. *If $\bar{\mu}$ is a solution of problem (8.4), then $Q^* = \int_Z q d\bar{\mu}(x,q)$ is a NE of the optimal exploitation of exhaustible resources problem, i.e., Q^* is a solution of (8.3).*

Proof. Let us first check that Assumption A holds true for problem (8.4). It is easy to see that Assumption A(1) and the first and the third points in Assumption A(2) are true by the continuity of g and Lemma 8.1. Let us prove that G_λ is lower semi-continuous for any $\lambda \in \mathcal{H}_f$. This is a consequence of the claim that the set-valued function $\mathcal{Z}\colon X \rightsquigarrow \mathcal{H}$, $x \mapsto \{g(x,y) \mid y \in Z_x\}$ is Hölder continuous. To see this, we fix any $x_1 < x_2$ in X. If $q \in Z_{x_1}$, then we have immediately that $q \in Z_{x_2}$. This implies that $Z_{x_1} \subseteq Z_{x_2}$. On the other hand, let $q \in Z_{x_2}$. We construct $q' \in Z_{x_1}$ by the following method:

$$q'_t = \begin{cases} q_t, & \text{if } \int_0^t q_\tau d\tau \le x_1; \\ 0, & \text{otherwise.} \end{cases}$$

As a consequence, we have that $\|q' - q\|_{\mathbb{L}^1([0,T])} \le x_2 - x_1$. Therefore, by Hölder's inequality,

$$\|q' - q\|^2_{\mathbb{L}^2_{\mathrm{e}^{-rt}}([0,T])} \le \|\mathrm{e}^{-rt}(q' - q)\|_{\mathbb{L}^\infty([0,T])}\|q' - q\|_{\mathbb{L}^1([0,T])} \le x_2(x_2 - x_1).$$

This implies that $Z_{x_2} \subseteq Z_{x_1} + \mathcal{B}_Y(0, \sqrt{x_2(x_2 - x_1)})$. Therefore, Assumption A follows.

Let $\bar{\mu}$ be a solution of problem (8.4), $\bar{\lambda} = \nabla f(\int_Z g d\bar{\mu})$ and $Q^* = \int_Z q d\bar{\mu}(x,q)$. By the definitions of f and g, we obtain that $\bar{\lambda} = (1, \epsilon Q^*)$, moreover,

$$g_{\bar{\lambda}}(x,q) = \int_0^T \mathrm{e}^{-rt} q_t(q_t - 1 + \epsilon Q^*_t)\mathrm{d}t.$$

By Lemma 8.1, $\mathbf{BR}_{\bar{\lambda}}(x_0) = \{q^{Q^*}(x_0)\}$ for any $x_0 \in X$. By Corollary 3.4, we have that $(\bar{\lambda}, \bar{\mu})$ satisfies the following equilibrium equation:

$$\begin{cases} \bar{\lambda} = \left(1\,,\, \epsilon \int_Z q d\bar{\mu}\right) \\ \bar{\mu}_x = \delta_{q^{Q^*}(x)}, \quad m\text{-a.e.} \end{cases}$$

Combining with Theorem 2.5, we obtain that $\int_Z q d\bar{\mu} = \int_X q^{Q^*}(x) dm(x)$. Recall that $Q^* = \int_Z q d\bar{\mu}$, then (8.3) follows. □

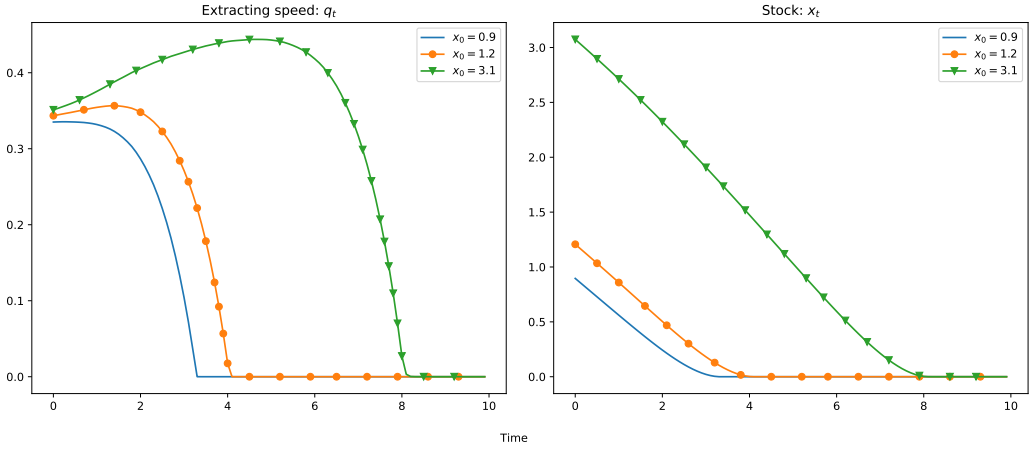


FIGURE 2. Extracting speeds and stocks of three producers with initial stocks: 0.9, 1.2, and 3.1.

Numerical simulations

Let the initial measure m be an exponential distribution with parameter $a \geq 0$, *i.e.*, $dm(x) = ae^{-ax}dx$ for all $x \geq 0$. Let us independently sample the distribution m for N times, denoting the samples by $x_1, x_2, ..., x_N$, and $m_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$. The time space $[0, T]$ is discretized with a step size $\Delta t = T/M$ for some $M \geq 1$. Then, a totally discretized problem associated with (8.4) writes:

$$\begin{cases} \inf_{q \in \mathbb{R}^{N \otimes M}} & J_N(q) := \frac{\Delta t}{N} \sum_{i=1}^{N} \sum_{t=0}^{M-1} e^{-rt\Delta t}(q_{i,t}^2 - q_{i,t}) + \frac{\epsilon \Delta t}{2} \sum_{t=0}^{M-1} e^{-rt} \left(\frac{1}{N} \sum_{i=1}^{N} q_{i,t} \right)^2, \\ \text{such that} & q_i \in S^M(x_i) := \{q \in [0, 1/2]^M \mid \Delta t \sum_{t=0}^{M-1} q_t \leq x_i\}, \qquad i = 1, 2, \ldots, N. \end{cases} \tag{8.5}$$

We apply Algorithm 3 to solve (8.5). At each iteration, the evaluation of a best-response, for each producer i amounts to solve a problem of the following form:

$$\begin{cases} \inf_{q_i \in \mathbb{R}^M} & \Delta t \sum_{t=0}^{M-1} e^{-rt\Delta t} q_{i,t}(q_{i,t} - 1 + \epsilon Q_t), \\ \text{such that} & q_i \in S^M(x_i), \end{cases} \tag{8.6}$$

for a given $Q \in [0, 1/2]^M$. This problem is a convex quadratic programming problem in $\mathbb{R}^M$ that can be dealt with by some solvers, such as GUROBI [50].

For the resolution of the problem, we chose the following parameters: $T = 10$, $\epsilon = r = a = 1$, $N = 100$, $M = 100$, $K = 100$, $n_k = 10$, for all k. Figure 2 shows the extracting speeds and the stocks of three producers with initial stocks: 0.9, 1.2, and 3.1. From Figure 2, we see that the producers with the higher initial stock have the same extracting speed as those with a lower initial stock, at the beginning. However, as the smaller agents exhaust their resource, the larger ones progressively raise their extraction speed. Once the extraction speed reaches its maximum value, it rapidly decreases to zero. These observations are consistent with the findings of [23], Section 3.3.

To study the error caused by sampling, we independently sample the exponential distribution m for $100 * N$ times, and group them into batches of N. The empirical distribution corresponding to each batch is set as the initial distribution. Then we apply Algorithm 3 to compute Q^* corresponding to each initial distribution. In Figure 3, we show the mean and standard deviation of the results of the 100 simulations.

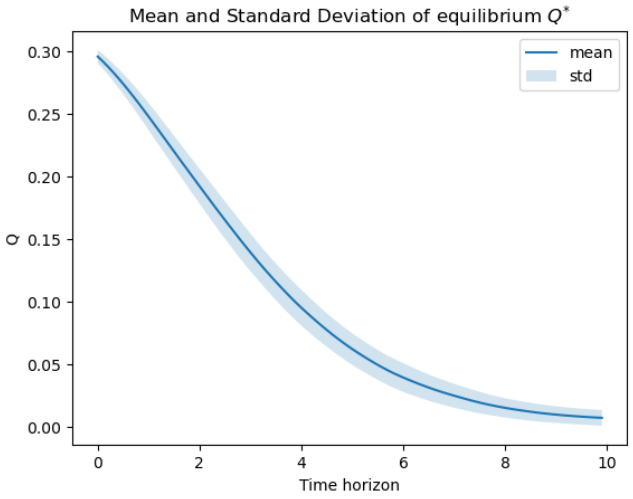


FIGURE 3. Mean and standard deviation of the equilibria of 100 batches.

8.4. Numerical results for a congestion game

Model

Consider a second numerical example within the context of the minimal-time deterministic MFG. We set the following parameters: the state space is fixed as $[0,1]$, a maximum duration is denoted by $T>0$, and an upper bound for the speed is given by $\bar{V}>0$. In this particular example, the dynamics governing each player are characterized by the set Z, defined as:

$$Z=\left\{(x,\gamma)\in[0,1]\times \mathrm{AC}([0,T]) \mid \dot{\gamma}_t=v_t,\ \gamma_0=x,\ 0\le v_t\le \bar{V}\right\}.$$

The objective for the players in this example is to reach the target point 1 as soon as possible, while simultaneously ensuring that the density at each point does not become excessively high. To quantify this, we introduce the congestion function $\tilde{\mathcal{F}}\colon \mathcal{P}([0,1])\to\mathbb{R}$, which is defined as follows:

$$\tilde{\mathcal{F}}(m)=\begin{cases}\int_0^1 m(x)^2\mathrm{d}x, & \text{if } m\ll \mathrm{d}x,\\ +\infty, \text{otherwise.}\end{cases}$$

Given an initial distribution $m_0\in\mathcal{P}([0,1])$, the resulting deterministic MFG problem can be expressed as follows:

$$\inf_{\mu\in\mathcal{P}_{m_0}(Z)} J(\mu)\coloneqq \int_Z\int_0^T \mathbb{I}_{[0,1]}(\gamma_t)\mathrm{d}t d\mu(x,\gamma)+\alpha\int_0^T \tilde{\mathcal{F}}(e_t\#\pi_2\#\mu)\mathrm{d}t, \tag{8.7}$$

where $\alpha>0$ is a penalty parameter.

Regularization

Note that the congestion function $\tilde{\mathcal{F}}$ does not fit to the framework studied in this article. To address this, we begin by approximating $\tilde{\mathcal{F}}$ with a function that aligns with our framework. We achieve this by partitioning the interval $[0,1]$ into $J\in\mathbb{N}_+$ small, uniform subintervals: $I_1,\dots,I_J$, where $I_j=[(j-1)\Delta x, j\Delta x]$, and $\Delta x=1/J$.

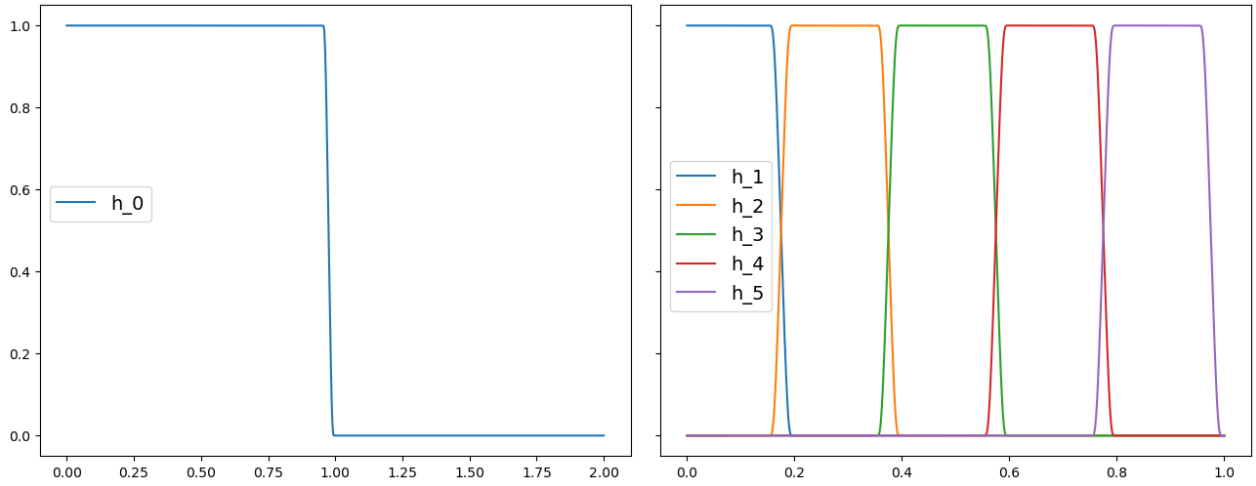


FIGURE 4. Example of h_0 and $\{h_j\}_{j=1,\dots,J}$ with $J = 5$ and $k = 20$.

Subsequently, we approximate $\tilde{\mathcal{F}}(m)$ as follows:

$$\tilde{\mathcal{F}}(m) \approx \Delta x \sum_{j=1}^{J} \left(\frac{1}{\Delta x} \int_0^1 \mathbb{I}_{I_j}(x) dm(x) \right)^2.$$

To facilitate the execution of the numerical experiments, we replace the indicator function $\mathbb{I}_{I_j}$ by some smooth functions. Let $k \geq J$ be a positive integer. We introduce two smooth functions, denoted as $\varphi_k \in \mathcal{C}^\infty(\mathbb{R})$ and $\phi_{k,\Delta x} \in \mathcal{C}^\infty(\mathbb{R})$. These functions are parametrized by the variables k and Δx, and are defined as follows:

$$\varphi_k(x) = \begin{cases} 0, & \text{if } x <= 0, \\ \frac{1}{1+\mathrm{e}^{1/kx - 1/(1-kx)}}, & \text{if } 0 < x < \frac{1}{k}, \\ 1 & \text{otherwise,} \end{cases}$$

$$\phi_{k,\Delta x}(x) = \begin{cases} 0, & \text{if } x \leq -\frac{1}{k}, \\ \varphi_k(x + 1/k), & \text{if } -\frac{1}{k} < x < 0, \\ 1 & \text{if } 0 \leq x \leq \Delta x - \frac{1}{k}, \\ 1 - \varphi_k(x - \Delta x + \frac{1}{k}), & \text{otherwise.} \end{cases}$$

Then, we approximate $\mathbb{I}_{I_j}$ by $h_j \colon \mathbb{R}_+ \to [0,1], x \mapsto \phi_{k,\Delta x}(x - (j-1)\Delta x)$ for $j = 1, \dots, J$, and $\mathbb{I}_{[0,1]}$ by $h_0 \colon \mathbb{R}_+ \to [0,1]$,

$$h_0(x) = \begin{cases} 1, & \text{if } 0 \leq x < 1 - \frac{1}{k}, \\ 1 - \varphi_k\left(x - 1 + \frac{1}{k}\right), & \text{otherwise.} \end{cases}$$

An important property of $\{h_j\}_{j=1,\dots,J}$ is that $\sum_{j=1}^{J} h_j(x) = h_0(x)$, which is 1 for any $x \in [0, 1 - 1/k]$, see Figure 4. The resulting approximated MFO problem associated with (8.7) is,

$$\inf_{\mu \in \mathcal{P}_{m_0}(Z)} \int_Z \int_0^T h_0(\gamma_t) \mathrm{d}t d\mu(x,\gamma) + \frac{\alpha}{\Delta x} \sum_{j=1}^{J} \int_0^T \left(\int_Z h_j(\gamma_t) d\mu(x,\gamma) \right)^2 \mathrm{d}t. \tag{8.8}$$

Discretization and numerical results

To proceed with our numerical experiments, we discretize the time horizon $[0,T]$ into M steps, each of duration $\Delta t = T/M$. Additionally, we discretize the initial distribution m_0 by $m_N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$. We formulate the fully discretized problem associated with (8.8) as follows:

$$\begin{cases} \inf_{(\gamma^i)_{i=1,\dots,N}} & \frac{\Delta t}{N}\sum_{i=1}^N \sum_{t=0}^{M-1} h_0(\gamma_t^i) + \frac{\alpha \Delta t}{\Delta x}\sum_{j=1}^J \sum_{t=0}^{M-1}\left(\frac{1}{N}\sum_{i=1}^N h_j(\gamma_t^i)\right)\Big)^2, \\ \text{such that} & 0 \le \gamma_{t+1}^i - \gamma_t^i \le \bar{V}\Delta t,\ \gamma_0^i = x_i, \text{ for } t = 0,\dots,M, \text{ and } i = 1,\dots,N. \end{cases} \tag{8.9}$$

Therefore, given some $(\bar{\gamma}^i)_{i=1,\dots,N}$ satisfying the constraint in (8.9), the sub-problem for player i is

$$\begin{cases} \inf_{\gamma^i} & \Delta t \sum_{t=0}^{M-1} h_0(\gamma_t^i) + \frac{2\alpha\Delta t}{\Delta x}\sum_{j=1}^J \sum_{t=0}^{M-1} \bar{y}_{j,t} h_j(\gamma_t^i), \\ \text{such that} & 0 \le \gamma_{t+1}^i - \gamma_t^i \le \bar{V}\Delta t,\ \gamma_0^i = x_i, \text{ for } t = 0,\dots,M, \end{cases} \tag{8.10}$$

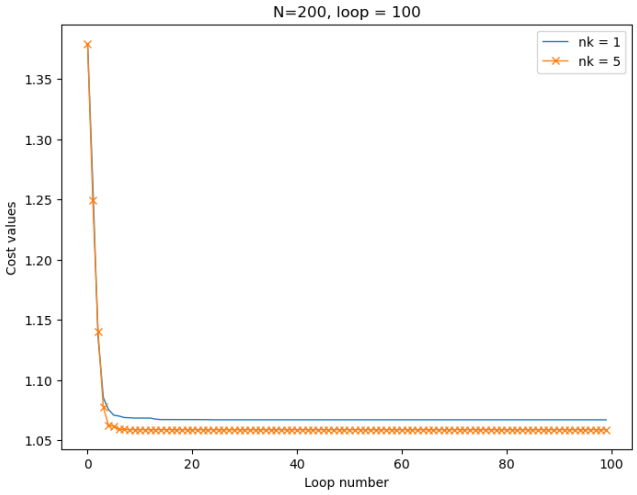


FIGURE 5. Convergence results of Algorithm 3.

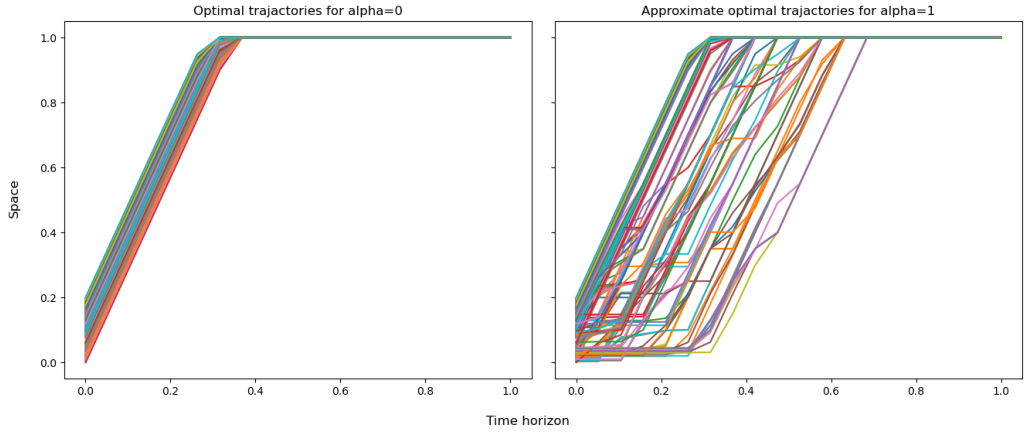
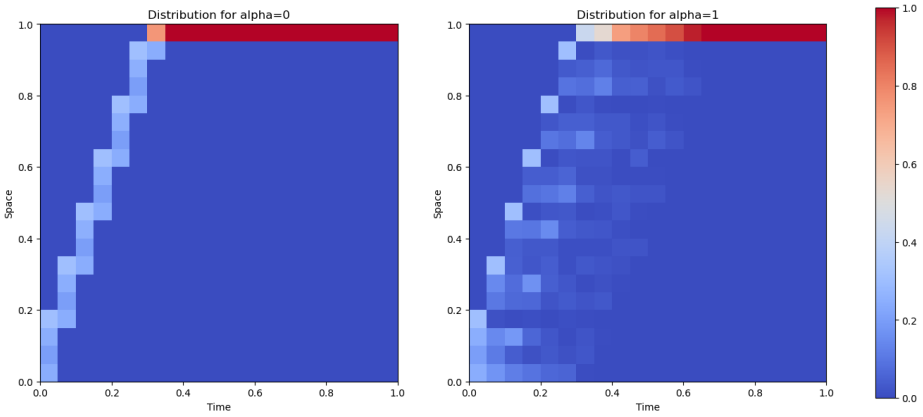
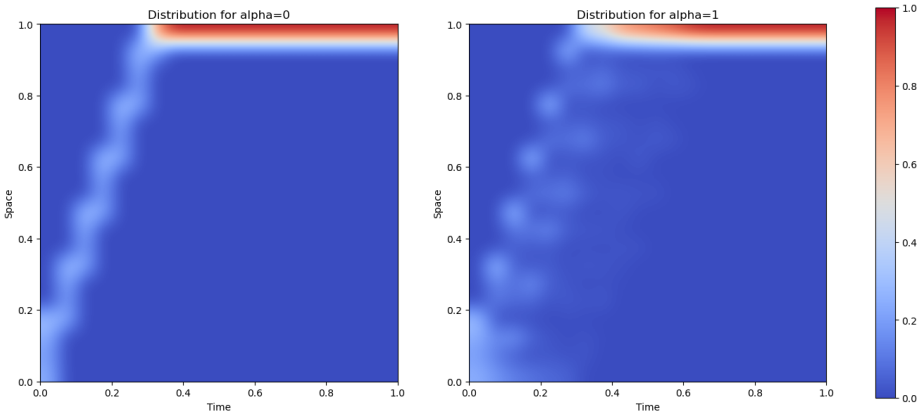


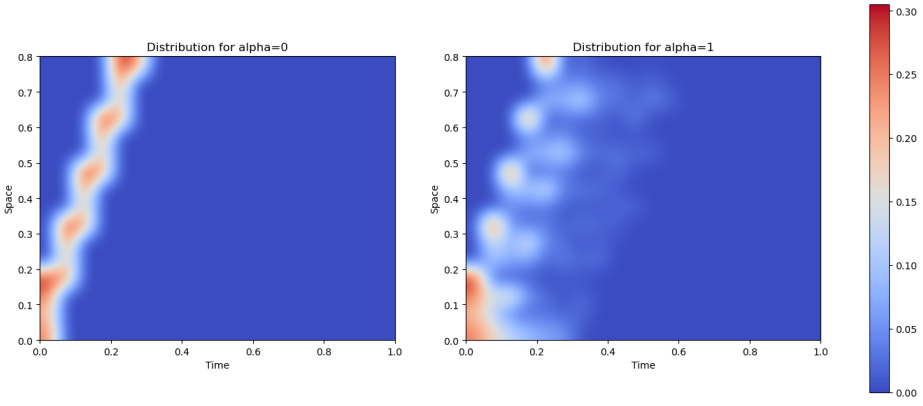
FIGURE 6. Optimal trajectory of each player: the case $\alpha = 0$ (left), the case $\alpha = 1$ (right).



(A) Comparison of distributions of positions at each time: the case $\alpha = 0$ (left), the case $\alpha = 1$ (right).



(B) Comparison of regularized distributions of positions at each time: the case $\alpha = 0$ (left), the case $\alpha = 1$ (right).



(C) Comparison of regularized distributions of positions in horizon $[0, 0.8]$ at each time: the case $\alpha = 0$ (left), the case $\alpha = 1$ (right).

FIGURE 7. Distributions.

where $\bar{y}_{j,t} = \frac{1}{N}\sum_{i=1}^{N} h_j(\bar{\gamma}_t^i)$. The sub-problem (8.10) is a finite-dimensional non-convex optimization problem, which is addressed by the open-source solver "scipy.optimize.minimize" [51].

Let us specify the parameters used in the numerical simulation of problem (8.8) as follows:

$$J = 5,\ k = 20,\ \bar{V} = 3,\ T = 1,\ \alpha = 1,\ M = 20,\ N = 200,\ m_0 = \mathrm{Uni}_{[0,0.2]},$$

where "Uni" represents the uniform distribution, and the points x_i are drawn from samples of m_0.

We first present in Figure 5 convergence results of Algorithm 3 for the discretized problem (8.9) in 100 iterations, utilizing parameter settings of $n_k = 1$ and 5. We see that in both choices, the algorithm converges to a local minimum very fast (fewer than 20 iterations). In Figure 6, we compare the optimal trajectories of γ^i for two cases: $\alpha = 0$ and $\alpha = 1$. It is evident that, in the case of $\alpha = 0$, the optimal strategy for each player is to move at the maximum speed, $\bar{V} = 3$, as there is no penalty for density. This is depicted in the left part of Figure 6. However, when $\alpha = 1$, players starting from greater initial positions choose to run at the maximum speed, whereas those with lower initial positions prefer to wait briefly to avoid congestion in density with those starting farther ahead. In Figure 7a, we draw the agents' state distributions at each time for both $\alpha = 0$ and $\alpha = 1$, along with a regularized version obtained through interpolation in Figure 7b. For a more detailed view of the density evolution before reaching the target, we further depict the restricted distributions within the spatial interval $[0, 0.8]$ in Figure 7c, using a distinct color scale.

Remark 8.3. Let us underline that for this example, the optimization problems involved in the evaluation of the best-response mapping are non-convex. As mentionned above, we address them with the open-source solver scipy.optimize.minimize whose default method for tackling constrained non-linear optimization problems is the SLSQP (Sequential Least SQuares Programming) algorithm, a quasi-Newton-type algorithm. Consequently, the quality of the initial guess plays a crucial role in the resolution of sub-problems. In the context of Algorithm 3, our experience shows that at iteration k, it is more efficient to initialise the evaluation of $\bar{y}_i^k \in \mathbf{BR}_{\lambda^k}(x_i)$ with y_i^k (rather than $\bar{y}_i^{k-1}$). We conjecture that the chance for the solver to generate a local solution is higher when initializing with $\bar{y}_i^{k-1}$.

9. Conclusion

We have provided a general framework for analyzing Mean Field Optimization problems. We have proposed a general method, based on an extension of the Frank–Wolfe algorithm for solving MFO problems, with a convergence guarantee, assuming that some best-response function can be efficiently computed (with a solver or with specific methods). Numerous extensions of the current setting could be considered. For example, one could formulate a stochastic setting with a random variable impacting all agents. In this setting the evaluation of λ (in the SFW algorithm) may require to use Monte-Carlo approximations, adding a new source of error in the general algorithm. One may also realize a general convergence analysis that would take into account the need to discretize the sets Z_x (in particular in the case of MFGs, where Z_x is an infinite dimension set). Finally, at a purely numerical level, we could investigate variants of the proposed method in which the distribution m is discretized progressively. This would reduce the number of subproblems to solve in the early iterations of the SFW algorithm. We also mention that the SFW is robust in the following sense: at the end of iteration k, if y^{k+1} is replaced by any other point yielding a reduction of the cost function, then the general convergence properties of the SFW algorithm are preserved. This fact could motivate the design of heuristic improvements on a case-by-case basis.

Data availability statement

The research data associated with this article are included in the article.

References

[1] P. Cannarsa and R. Capuani, Existence and uniqueness for mean field games with state constraints. *PDE Models Multi-agent Phenomena* (2018) 49–71.

[2] M.-W. Cheung and R. Lahkar, Nonatomic potential games: the continuous strategy case. *Games Econ. Behav.* **108** (2018) 341–362.

[3] F. Santambrogio and W. Shim, A Cucker–Smale inspired deterministic mean field game with velocity interactions. *SIAM J. Control Optim.* **59** (2021) 4155–4187.

[4] J.-D. Benamou, G. Carlier and F. Santambrogio, Variational mean field games. *Active Particles*, Vol. 1 of *Advances in Theory, Models, and Applications* (2017) 141–171.

[5] J.F. Bonnans, J. Gianatti and L. Pfeiffer, A lagrangian approach for aggregative mean field games of controls with mixed and final constraints. *SIAM J. Control Optim.* **61** (2023) 105–134.

[6] C. Sarrazin, Lagrangian discretization of variational mean field games. *SIAM J. Control Optim.* **60** (2022) 1365–1392.

[7] O. Beaude, P. Benchimol, S. Gaubert, P. Jacquot and N. Oudjane, A privacy-preserving method to optimize distributed resource allocation. *SIAM J. Optim.* **30** (2020) 2303–2336.

[8] P. Jacquot, O. Beaude, S. Gaubert and N. Oudjane, Analysis and implementation of an hourly billing mechanism for demand response management. *IEEE Trans. Smart Grid* **10** (2018) 4265–4278.

[9] K. Liu, N. Oudjane and L. Pfeiffer, Decomposed resolution of finite-state aggregative optimal control problems, in *2023 Proceedings of the Conference on Control and its Applications (CT)*. SIAM (2023) 56–63.

[10] L. Chizat and F. Bach, On the global convergence of gradient descent for over-parameterized models using optimal transport. *Adv. Neural Inform. Process. Syst.* **31** (2018) 3036–3046.

[11] K. Liu and E. Zuazua, Representation and regression problems in neural networks: Relaxation, generalization, and numerics. *Math. Models Methods Appl. Sci.* **35** (2025) 1471–1521.

[12] S. Mei, A. Montanari and P.-M. Nguyen, A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci. U.S.A.* **115** (2018) E7665–E7671.

[13] N. Boyd, G. Schiebinger and B. Recht, The alternating descent conditional gradient method for sparse inverse problems. *SIAM J. Optim.* **27** (2017) 616–639.

[14] M. Wang, Vanishing price of decentralization in large coordinative nonconvex optimization. *SIAM J. Optim.* **27** (2017) 1977–2009.

[15] Q. Mérigot and J.-M. Mirebeau, Minimal geodesics along volume-preserving maps, through semidiscrete optimal transport. *SIAM J. Numer. Anal.* **54** (2016) 3465–3492.

[16] Bonnans, K. Liu, N. Oudjane, L. Pfeiffer and C. Wan, Large-scale nonconvex optimization: randomization, gap estimation, and numerical resolution. *SIAM J. Optim.* **33** (2023) 3083–3113.

[17] M. Geist, J. Pérolat, M. Laurière, R. Elie, S. Perrin, O. Bachem, R. Munos and O. Pietquin, Concave utility reinforcement learning: The mean-field game viewpoint, in *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems* (2022) 489–497.

[18] P. Lavigne and L. Pfeiffer, Generalized conditional gradient and learning in potential mean field games. *Appl. Math. Optim.* **89** (2023) Article no 89.

[19] K. Liu and L. Pfeiffer, A mesh-independent method for second-order potential mean field games. *IMA J. Numer. Anal.* **45** (2025) 1226–1266.

[20] P. Cardaliaguet and S. Hadikhanloo, Learning in mean field games: the fictitious play. *ESAIM: Control Optim. Calc. Var.* **23** (2017) 569–591.

[21] S. Hadikhanloo and F.J. Silva, Finite mean field games: fictitious play and convergence to a first order continuous mean field game. *J. Math. Pures Appl.* **132** (2019) 369–397.

[22] Y. Brenier, Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math.* **305** (1987) 805–808.

[23] P. Graewe, U. Horst and R. Sircar, A maximum principle approach to a deterministic mean field game of control with absorption. *SIAM J. Control Optim.* **60** (2022) 3173–3190.

[24] W. Rudin, *Real and Complex Analysis*, vol. 3. McGraw-Hill (1987).

[25] J.-P. Aubin and Hélène Frankowska, *Set-valued Analysis*. Springer Science & Business Media (2009).

[26] R.J. Aumann, Measurable utility and the measurable choice problem, in *La Décision*. Colloque Internationaux du C.N.R.S., Paris (1969) 15–26.

[27] D.L. Cohn, *Measure Theory*, vol. 1. Springer (2013).
[28] L. Ambrosio, N. Gigli and G. Savaré, *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media (2005).
[29] Y. Nesterov, *Lectures on Convex Optimization*, vol. 137. Springer (2018).
[30] M. Jaggi, Revisiting Frank–Wolfe: Projection-free sparse convex optimization, in *International Conference on Machine Learning*. PMLR (2013) 427–435.
[31] K. Bredies, M. Carioni, S. Fanzon and D. Walter, Asymptotic linear convergence of fully-corrective generalized conditional gradient methods. *Math. Program.* **205** (2024) 135–202.
[32] K. Bredies and H.K. Pikkarainen, Inverse problems in spaces of measures. *ESAIM: Control Optim. Calc. Var.* **19** (2013) 190–218.
[33] K. Pieper and D. Walter, Linear convergence of accelerated conditional gradient algorithms in spaces of measures. *ESAIM: Control Optim. Calc. Var.* **27** (2021) 38.
[34] K. Bredies, D.A Lorenz and P. Maass, A generalized conditional gradient method and its connection to an iterative shrinkage method. *Computat. Optim. Appl.* **42** (2009) 173–193.
[35] C. Villani, *Optimal Transport: Old and New*, vol. 338. Springer (2009).
[36] C. Ma, L. Wu, and W. E, The Barron space and the flow-induced function spaces for neural network models. *Construct. Approx.* **55** (2022) 369–406.
[37] A.R. Barron, Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory* **39** (1993) 930–945.
[38] G. Pisier, Remarques sur un résultat non publié de B. Maurey. *Séminaire d'Analyse fonctionnelle (Ecole Polytechnique)* (1981) 1–12.
[39] G.B. Folland, *Real Analysis: Modern Techniques and their Applications*, vol. 40. John Wiley & Sons (1999).
[40] H.H. Bauschke and P.L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, vol. 408. Springer (2011).
[41] R.T. Rockafellar, *Convex Analysis*, vol. 11. Princeton University Press (1997).
[42] A. Gersho and R.M. Gray, *Vector Quantization and Signal Compression*, vol. 159. Springer Science & Business Media (2012).
[43] K. Falconer, Fractal Geometry: Mathematical Foundations and Applications. John Wiley & Sons (2004).
[44] J. Kitagawa, Q. Mérigot and B. Thibert, Convergence of a newton algorithm for semi-discrete optimal transport. *J. Eur. Math. Soc.* **21** (2019) 2603–2651.
[45] B. Lévy, A numerical algorithm for l_2 semi-discrete optimal transport in 3d. *ESAIM: Math. Model. Numer. Anal.* **49** (2015) 1693–1715.
[46] N. Fournier and A. Guillin, On the rate of convergence in wasserstein distance of the empirical measure. *Probab. Theory Related Fields* **162** (2015) 707–738.
[47] Q. Mérigot, A multiscale approach to optimal transport. *Comput. Graph. Forum* **30** (2011) 1583–1592.
[48] Y. Sheffi, *Urban Transportation Networks*, vol. 6. Prentice-Hall, Englewood Cliffs, NJ (1985).
[49] H. Brézis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, vol. 2. Springer (2011).
[50] LLC Gurobi Optimization, Gurobi optimizer reference manual (2018).
[51] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, *et al.* Scipy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods* **17** (2020) 261–272.
[52] H. Tong, Some characterizations of normal and perfectly normal spaces. *Duke Math. J.* **19** (1952) 289–292.

Appendix A. Proof of Lemma 3.2

This appendix is dedicated to the proof of Lemma 3.2. It relies on two intermediate results (stated in Lemmas A.1 and A.2). Let us denote by $\mathcal{B}^X$ and $\mathcal{B}^Y$ the Borel σ-algebra on X and Y, respectively.

Lemma A.1. *Let Assumption A hold true. For any $\lambda \in \mathcal{H}_f$, the set-valued function $\mathbf{BR}_\lambda$ has a closed graph.*

Proof. Let $x_k \in X$ converge to some $\bar{x} \in X$, and let $y_k \in \mathbf{BR}_\lambda(x_k)$ converge to some $\bar{y} \in Y$. We have to prove that $\bar{y} \in \mathbf{BR}_\lambda(\bar{x})$. First, we have $\bar{y} \in Z_{\bar{x}}$, since Z is closed. Fix any $y \in Z_{\bar{x}}$. Since G_λ is lower semi-continuous, there exists a sequence $(\hat{y}_k)_{k\in\mathbb{N}}$ in Z_{x_k} such that

$$g_\lambda(\bar{x}, y) = \lim_{k\to\infty} g_\lambda(x_k, \hat{y}_k).$$

By the lower semi-continuity of g_λ, we have

$$g_\lambda(\bar{x}, \bar{y}) \leq \liminf_{k\to\infty} g_\lambda(x_k, y_k).$$

Since $y_k \in \mathbf{BR}_\lambda(x_k)$ and $\hat{y}^k \in Z_{x_k}$, we have $g_\lambda(x_k, y_k) \leq g_\lambda(x_k, \hat{y}_k)$ for any k. Passing to the limit in this inequality (using the above inequalities), we deduce that $g_\lambda(\bar{x}, \bar{y}) \leq g_\lambda(\bar{x}, y)$. Thus, $\mathbf{BR}_\lambda$ has a closed graph. □

Note that a similar result to Lemma A.1 is presented in [1], Lemma 3.4.

Lemma A.2. *Let Assumption A hold true. The function u_λ is upper semi-continuous for any $\lambda \in \mathcal{H}_f$, thus Borel measurable.*

Proof. Let $\lambda \in \mathcal{H}_f$. Since g is bounded over Z, we have that $u_\lambda(x) > -\infty$ for any $x \in X$. Fix any $x \in X$. Let $y \in \mathbf{BR}_\lambda(x)$. Let $(x_n \in X)_{n\geq 1}$ be a sequence converging to x. By the lower semi-continuity of G_λ, there exists $y_n \in Z_{x_n}$ such that $g_\lambda(x, y) = \lim_{n\to\infty} g_\lambda(x_n, y_n)$. Therefore,

$$u_\lambda(x) = g_\lambda(x, y) = \lim_{n\to\infty} g_\lambda(x_n, y_n) \geq \limsup_{n\to\infty} u_\lambda(x_n).$$

We obtain the upper semi-continuuity of u_λ for any $\lambda \in \mathcal{H}_f$. Since any upper semi-continuous function defined on a metric space is the limit of a monotonically decreasing sequence of continuous functions [52], Theorem 3, we deduce that u_λ is Borel measurable. □

Proof of Lemma 3.2. Let us recall that we have already proved that

$$\inf_{\mu\in\mathcal{P}_m(Z)} \int_Z g_\lambda d\mu \geq \int_X u_\lambda dm.$$

To prove the result, it suffices to construct $\tilde{\mu} \in \mathcal{P}_m(Z)$ such that

$$\int_Z g_\lambda d\tilde{\mu} = \int_X u_\lambda dm. \tag{A.1}$$

By Assumption A, the set-valued function $\mathbf{BR}_\lambda \colon X \rightsquigarrow Y$ has non-empty closed images. By Lemma A.1, Graph($\mathbf{BR}_\lambda$) is closed in $X \times Y$, thus is a $\mathcal{B}^X \otimes \mathcal{B}^Y$-measurable set. By Theorem 2.3 (2), there exists a $(\mathcal{B}^X, \mathcal{B}^Y)$-measurable function $\mathbf{br}_\lambda \colon X \to Y$ such that

$$\mathbf{br}_\lambda(x) \in \mathbf{BR}_\lambda(x), \quad m\text{-a.e.}$$

We define $\mathcal{A} \colon X \to Z$, $x \mapsto (x, \mathbf{br}_\lambda(x))$ and

$$\tilde{\mu} = \mathcal{A}\#m.$$

It follows immediately that $\tilde{\mu} \in \mathcal{P}_m(Z)$, since $\pi_1 \circ \mathcal{A}$ is the identity function. Next, we prove (A.1). We indeed have

$$g_\lambda \circ \mathcal{A} = u_\lambda, \quad m\text{-a.e.}$$

Therefore, $g_\lambda \circ \mathcal{A}$ is Borel measurable by Lemma A.2. Since u_λ is bounded, we have

$$\int_X u_\lambda \, \mathrm{d}m = \int_X g_\lambda \circ \mathcal{A} \, \mathrm{d}m = \int_Z g_\lambda \, \mathrm{d}(\mathcal{A}\#m) = \int_Z g_\lambda \, \mathrm{d}\tilde{\mu},$$

which proves (A.1) and concludes the proof. □

Appendix B. Proof of Lemma 4.3

Assumption B(3) implies that S has nonempty values. Set $E \coloneqq Z \times X$ and consider the graph

$$\mathrm{Graph}(S) \coloneqq \{(x_0, y_0, x_1, y_1) \in E \times Y \mid y_1 \in S(x_0, y_0, x_1)\}.$$

By definition of S,

$$\mathrm{Graph}(S) = \Big\{(x_0, y_0, x_1, y_1) \in E \times Y \mid y_1 \in Z_{x_1},\ \|g(x_1, y_1) - g(x_0, y_0)\| \le L_g \, d_X(x_0, x_1)\Big\}.$$

Since $g : Z \to \mathcal{H}$ is Borel measurable and the operations $(u, v) \mapsto u - v$ and $w \mapsto \|w\|$ are continuous on $\mathcal{H}$, the map $(x_0, y_0, x_1, y_1) \mapsto \|g(x_1, y_1) - g(x_0, y_0)\|$ is Borel on $Z \times Z$. Moreover, $(x_0, x_1) \mapsto d_X(x_0, x_1)$ is continuous, hence the above inequality defines a Borel set. Therefore, $\mathrm{Graph}(S) \in \mathcal{B}(E) \otimes \mathcal{B}(Y)$.

Since ν is a probability measure on E (hence σ-finite) and Graph(S) is measurable, Theorem 2.3 yields a Borel measurable function $s : E \to Y$ such that $s(x_0, y_0, x_1) \in S(x_0, y_0, x_1)$ for ν-a.e. $(x_0, y_0, x_1) \in E$. Unfolding the definition of S concludes the proof.