Ergodic control of a heterogeneous population and application to electricity pricing

Quentin Jacquet, Wim van Ackooij, Clémence Alasseur, Stéphane Gaubert

Ergodic control of a heterogeneous population and application to electricity pricing

Quentin Jacquet, Wim van Ackooij, Clémence Alasseur and Stéphane Gaubert

***Abstract*— We consider a control problem for a heterogeneous population composed of agents able to switch at any time between different options. The controller aims to maximize an average gain per time unit, supposing that the population is of infinite size. This leads to an ergodic control problem for a "mean-field" Markov Decision Process in which the state space is a product of simplices, and the population evolves according to controlled linear dynamics. By exploiting contraction properties of the dynamics in Hilbert's projective metric, we prove that the infinite-dimensional ergodic eigenproblem admits a solution and show that the latter is in general non unique. This allows us to obtain optimal strategies, and to quantify the gap between steady-state strategies and optimal ones. In particular, we prove in the one-dimensional case that there exist cyclic policies – alternating between discount and profit taking stages – which secure a greater gain than constant-price policies. On numerical aspects, we develop a policy iteration algorithm with "on-the-fly" generated transitions, specifically adapted to decomposable models, leading to substantial memory savings. We finally apply our results on realistic instances coming from an electricity pricing problem encountered in the retail markets, and numerically observe the emergence of cyclic promotions for sufficient inertia in the customer behavior.**

I. Introduction

A. Ergodic control for mean-field MDPs

Many control problems involve a large number of rational agents, reacting to the decisions of a controller such as in finance [BHP99; BR11], routing problems [CS17] or epidemiology [Lee+21] . To overcome the intractability that appears when the number of individuals grows, mean-field control have been introduced, see e.g. [BFY13]. Assuming that the agents in the population are indistinguishable, the fundamental idea is to apply a mean-field type approximation and to show that looking at the population distribution (instead of each individual state) is sufficient. In particular, early convergence results (of order $1\sqrt{N}$) for the N-cooperative agents control problem to the mean-field limit have been proved by Gast and Gaujal [GG10] for discounted horizons. Motte and Pham [MP22] generalize the latter results in the presence of common-noise. We focus here on the *ergodic control problem* (i.e., infinite undiscounted horizon and average long-term rewards). In this context, it is showed in [Bäu23] that the optimal policy in the mean-field limit is ε-optimal for the N-agents discounted version when the size of the population is large and the discount factor is close to one.

Q. Jacquet, W. van Ackooij and C. Alasseur are with EDF R&D Saclay, Palaiseau, France {quentin.jacquet, wim.van-ackooij, clemence.alasseur}@edf.fr

Q. Jacquet and S. Gaubert are with INRIA, CMAP, Ecole Polytechnique, CNRS, Palaiseau, France stephane.gaubert@inria.fr

The ergodic control problem for a Markov decision process with Bellman operator $\mathcal{B}$, on a compact state space $\mathcal{D}$, is classically studied by means of the ergodic eigenproblem

$$g1_{\mathcal{D}} + h = \mathcal{B}h \ , \tag{1}$$

in which h is a bounded function on the state space, called the bias or potential, and g is a real constant. We refer to [HL96] for background on the topic. If the ergodic eigenproblem is solvable, then, g yields the optimal mean payoff per time unit, and it is independent of the initial state. Moreover, an optimal policy can be obtained by selecting maximizing actions in the expression of $\mathcal{B}h$. When the state and action spaces are finite, the ergodic eigenproblem is well understood, in particular, a solution does exist if every policy yields a unichain transition matrix (i.e., a matrix with a unique final class), see e.g. [Put94].

Here, the ergodic control problem arises from a mean-field approximation (called *lifted* MDP in [MP22]). The state space is therefore infinite and corresponds to the space of probability distributions over the choices proposed to a representative agent, typically for a finite number of choices $\{1, \dots, n\}$, $\mathcal{D}$ is the probability simplex $\Delta_n = \{\mu \in \mathbb{R}_+^n : \sum_{i=1}^n \mu_i = 1\}$. In this context of infinite state space, the solution to the ergodic eigenproblem is a more difficult question [KM97; Fat22; MN02; AGN11; AGN15]. This is particularly the case in absence of common-noise – where the lifted MDP is of *deterministic* nature – owing to the lack of regularizing effect coming from stochasticity.

B. Contributions

We consider a population of agents, that have different types/preferences. Each agent chooses between several options, taking into account the actions made by the controller, who aims at optimizing a mean reward per time unit. This is represented by a discrete-time ergodic control problem, in which the state –the population– belongs to a product of simplices. We suppose that the population evolves according to the Fokker-Planck equation of a controlled Markov chain. In this work, we directly study the "mean-field" model where the population is supposed to be of infinite size. This choice is motivated by our application on the French electricity market where the population is in fact the whole set of French households (around 30 millions), leading to intractable model without such mean-field hypothesis. Our first main result, Theorem 1, shows that the ergodic eigenproblem does admit a solution (in the presence of common noise or not). This entails that the value of the ergodic control problem is independent of the initial state, and this also allows us to

determine optimal stationary strategies. Theorem 1 requires a primitivity assumption on the semigroup of transition matrices; it applies in particular to positive transition matrices, such as the ones arising from logit based models. The proof relies on contraction properties of the dynamics in Hilbert's projective metric, which allow us to establish compactness estimates which guarantee the existence of a solution.

In order to numerically solve the ergodic eigenproblem, we develop a policy iteration algorithm building on [Coc+98; DF68] but especially intended for *decomposable* state spaces (e.g., for populations of independent customers) where transitions can be generated on-the-fly using pre-computed local information, see Algorithm 2. This refined version is afterwards compared with existing approaches, see Section VI, and reveals drastic computational time reductions with respect to the value-iteration algorithm and considerable memory gains in comparison with off-the-shelf policy iteration procedures.

In addition, we study stationary pricing strategies. Owing to the contraction properties of the dynamics, these are such that the population distribution converges to a stationary state. Then, we refine a result from [Fly79], providing a bound on the loss of optimality arising from the restriction to stationary pricing strategy. We define a family of Lagrangian functions, whose duality gap provides an explicit bound on the optimality loss, see Proposition 3. In particular, a zero duality gap guarantees that stationary pricing policies are optimal.

Finally, we apply our results to a problem of electricity pricing, inspired by a real case study (French contracts). An essential feature of this model is to take into account the *inertia* of customers, i.e., their tendency to keep their current contract even if it is not the best offer. This is represented by a logit-based stochastic transition model with switching costs. Theorem 3 provides a closed-form formula for the stationary distribution, which allows us to determine steady-state policies by reduction to a single-level problem. We also obtain qualitative results through majorization concepts [MOA11], showing that the addition of inertia in the model leads to a more concentrated distribution of the population, see Proposition 4. We present numerical tests on examples of dimension 2 and 4. These reveal the emergence of optimal cyclic policies for large switching costs, recovering the empirical notion of "promotions" of [DHR09] and [PE17].

C. Related works

As mentioned above, we allow here the ergodic eigenproblem to be of deterministic nature, more degenerate than its stochastic analogue studied in the context of average cost Markov Decision Processes, see e.g. [Ara+93] and the references therein. The main classical approach to show the existence of a solution to the infinite-dimensional ergodic eigenproblem relies on a Doeblin-type (or minorization) condition. The latter entails a contraction property of a Markov semi-group acting on spaces of measures, as well as the contraction of the Bellman operator with respect to the Hilbert pseudo-norm. It implies not only the existence but also the *uniqueness* of the ergodic eigenvector (up to a constant). This method has been used in [Kur89; HL96], and more recently in the works of Biswas [Bis15] and Wiecek [Wie19] in the mean-field context. We also refer the reader to [BCG19; GQ14] for background on Doeblin (and the more general Dobrushin) type conditions. In our setting, this approach does not apply. In fact, we provide an explicit counter example showing that the eigenvector may not be unique, see Example 1, and this entails that our existence result cannot be obtained using a Doeblin-type approach.

Instead, we exploit here the contraction properties of the dynamics to show that the family of value functions of the associated discounted problem is equi-Lipschitz. Then, following a now classical approach of Lions, Papanicolaou and Varadan [LPV87], a solution of the ergodic eigenproblem is obtained by a vanishing discount limit. The use of contraction ideas is partly inspired by a previous work of Calvez, Gabriel and Gaubert [CGG14], tackling a different problem (growth maximization). Also, [CGG14] deals with a PDE rather than discrete setting. Our result may also be compared with those of Bäuerle [Bäu23], in which the equi-Lipschitz property of the value function is supposed a priori.

In the deterministic setting, the ergodic eigenproblem is a special case of the "max-plus" or "tropical" infinite dimensional spectral problem [KM97; AGW09], or of the eigenproblem studied in discrete weak-KAM and Aubry Mather theory [Fat22; GT11]. Basic spectral theory results require the Bellman operator to be compact. This holds under demanding "controllability" conditions (see e.g. [KM97, Theorem 3.6]), not satisfied in our setting. Extensions of these results rely on quasi-compactness techniques [MN02; AGN11].

We also note that the ergodic eigenproblem, in the special deterministic "0-player case", has been studied under the name of cohomological equation in the field of dynamical systems. The existence of a regular solution is generally a difficult question, a series of results going back to the work of Livšic [Liv72], show that a Hölder continuous eigenvector does exists if the payment function is Hölder continuous, and if the dynamics is given by an Anosov diffeomorphism. The latter condition requires the tangent bundle of the state space to split in two components, on which the dynamics is either uniformly expanding or uniformly contracting. Here, we establish a "one player" version, but requiring a uniform contraction assumption.

On the computational aspect, a standard approach to solve the ergodic eigenproblem is to use the relative value-iteration (RVI) algorithm which goes back to White [Whi63]. Its convergence requires a demanding primitivity condition, see [FST78]. It has been remarked in [GS20] that this can be relaxed by combining RVI with Krasnoselskii-Mann damping. In the worstcase, ε^{-2} iterations are needed to solve the problem with a precision ε, see e.g. (Algorithm 1). Here, in the present mean-field case, an essential step is to discretize the state space, wich is a product of simplices. Recently, in the model-free context and for infinite dis-

counted horizons, RVI algorithms have been used by Carmona, Laurière and Tan [CLT21, Algo. 1], also considering a beforehand discretization. A different class of algorithm rely on policy-iteration (PI), still relying on a discretization. In the deterministic case, PI can be implemented by a fast graph algorithm, see e.g. [Coc+98]. Here, we exploit the decomposability of the transition matrix (the dynamics of the populations are only coupled by the control) to obtain a more scalable method. Different refinements of policy iteration have been developed by Festa [Fes18] using domain decomposition to obtain parallel Howard's algorithm, as well as memory space gain. Bayraktar, Bäuerle and Kara [BBK23] prove convergence results of the solution of the discretize version to the continuous one for discounted payoff by exhibiting regret bounds between the approximated mean-field MDP obtained by a semi-Lagrangian discretization (nearest neighbor) and the true infinite-population case. To this purpose, they fully exploit the Lipschitz property of the optimal value function [BBK23, Lemma 6] by supposing that the dynamics is contracting for the Wasserstein metric.

Finally, on the application side, we aim at analyzing the impact of switching costs on the retail electricity market, and especially on the optimal behavior of the retailers. In Economics, consumers are often supposed to be fully rational, and their reactions to price to be instantaneous. However, many studies highlight that switching costs and limited awareness conjointly lead to inertia in retail electricity markets, which hinders efficient choices, see [HMP17; NMS19; DW19]. Inertia in imperfect markets impacts the decision of the providers and modifies their pricing strategies. Studies tend to show the importance of promotions in the pricing behaviors of firms, see [HP10; AR12]. In particular, empirical analyses show how the depth and frequency of promotions are linked with the level of inertia. Here, we study the problem through a mathematical angle using mean-field MDPs. In comparison with the work of Pavlidis and Ellickson [PE17] on multiproduct pricing, we also consider logit-based transitions but we focus on long-term average rewards and reinforce the theoretical understanding of the model by identifying the optimal steady-states and studying the emergence of cyclic policies.

This article is organized as follows. In Section II, we first define the model and prove the results on the ergodic eigenproblem (existence and non-uniqueness). We then present two iterative algorithms to solve this fixed-point problem in Section III. We study steady-states and their optimality in Section IV, and illustrate the electricity application in Section V.

An initial account of some of the present results appeared in the conference paper [Jac+22].

II. Ergodic control

A. Notation

We denote by $\langle \cdot, \cdot \rangle_n$ the scalar product on $\mathbb{R}^n$, and for any x and y in $\mathbb{R}^n$, $x \vee y$ represents the elementwise maximum between x and y. We recall that the probability simplex of $\mathbb{R}^n$ is denoted by $\Delta_n = \{\mu \in \mathbb{R}^n_+ : \sum_{i=1}^n \mu_i = 1\}$. Besides, we denote by $\mathrm{sp}(f) := \max_{x\in E} f(x) - \min_{x\in E} f(x)$ the span of the function $f : E \to \mathbb{R}$. We say that a matrix P is positive, and we write $P \gg 0$, if all the coefficients of P are positive. The set of convex functions with finite real values on a space K is denoted by $\mathrm{Vex}\, K$, and the convex hull of a set K is denoted by $\mathrm{vex}\, K$. Moreover, the set of Lipschitz function on E is denoted by $\mathrm{Lip}(E)$, and the relative interior of a set E is denoted by $\mathrm{relint}(E)$.

The *Hilbert projective metric* d_H on $\mathbb{R}^n_{>0}$ is defined as $d_H(u, v) = \max_{1\le i,j\le n} \log(\frac{u_i}{v_i}\frac{v_j}{u_j})$, see [LN09]. It is such that $d_H(u, v) = 0$ iff the vectors u and v are proportional, hence, the name "projective". For a set $E \subseteq \mathbb{R}^n_{>0}$, we denote by $\mathrm{Diam}_H(E) := \max_{u,v\in E} d_H(u, v)$ the diameter of the set E, and for a matrix $P \in \mathbb{R}^{n\times n}$ we denote by $\mathrm{Diam}_H(P) := \max_{1\le i,j\le n} d_H(P_i, P_j)$ the *diameter* of P, where P_i denotes the ith row of P. This can be seen to coincide with the diameter, in Hilbert's projective metric, of the image of the set $\mathbb{R}^n_{>0}$ by the matrix P, see Th. A.6.2, *ibid.*

Finally, for a sequence $(a_t)_{t\ge1}$, we respectively denote by $a_{s:t}$ and $a_{:t}$ the subsequences $(a_\tau)_{s\le\tau\le t}$ and $(a_\tau)_{1\le\tau\le t}$.

B. Model

We consider a large population model composed of K clusters of indistinguishable individuals. Each cluster $k \in [K] := \{1, \dots, K\}$ represents a proportion ρ_k of the overall population, and is supposed to react *independently* from the other clusters.

Let $\mathcal{X}$ and $\mathcal{A}$ be respectively the state and action spaces. We suppose in the sequel that $\mathcal{X}$ is finite and w.l.o.g. $\mathcal{X} = \{1, 2, \dots, N\}$. We suppose also that $\mathcal{A}$ is a compact set (in Section V, $\mathcal{A}$ will be an explicit subset of $\mathbb{R}^N$).

For any time $t \ge 0$ and any cluster k, we denote by $x^k_t \in [N]$ the stochastic choice made by a representative agent of cluster k at time t. The distribution of the population of cluster k over $[N]$ is then denoted by $\mu^k_t = \left(\mathbb{P}\left[x^k_t = i\right]\right)_{i\in[N]} \in \Delta_N$. We suppose that the dynamics of the process μ^k is given by a discrete time (linear) equation of the form

$$\mu^k_t = \mu^k_{t-1} P^k(a_t, \xi_t) \ , \tag{2}$$

where P^k is the Markov transition matrix of the underlying process x^k.

In the first instance, we consider that the latter matrix is impacted by an exogenous process (*common-noise*), independent of the initial state μ_0, and represented by a sequence of independent and identically distributed (i.i.d.) random variables $\{\xi_t\}$ with values in some space Ξ and common distribution σ.

At every time $t \ge 1$, a controller chooses an action $a_t \in \mathcal{A}$. She obtains a stochastic reward $r : \mathcal{A} \times \Delta_N^K \times \Xi \to \mathbb{R}$ defined as

$$r : (a_t, \mu_t, \xi_t) \mapsto \sum_{k\in[K]} \rho_k \left\langle \theta^k(a_t, \xi_t), \mu^k_t \right\rangle_N \ , \tag{3}$$

where $\theta^k(a, \xi) \in \mathbb{R}^N$ is the vector whose entry n represents the unitary reward for the controller coming from an individual of cluster k in state n, for realization ξ and after executing action a.

The semi-flow ϕ describing the dynamics of the state μ is then defined by a function depending on the past actions and past realizations of the common-noise:

$$\phi_t(a_{:t}, \xi_{:t}, \mu_0) := \mu_t \ .$$

We also denote by Π the set of policies. Then, for a given policy $\pi = \{\pi_t\}_{t\geq 1}$, the action taken by the controller at time t is $a_t = \pi_t(\mu_t)$.

In the sequel, the following assumptions will be used:

(A1) The transition $(a, \xi) \mapsto P^k(a, \xi)$ is a continuous function for any k.

(A2) There exists $L \in \mathbb{N}$ such that for any sequence of actions $a_{:L} \in \mathcal{A}^L$, any sequence $\xi_{:L}$ and cluster k, $\prod_{l\in[L]} P^k(a_i, \xi_i) \gg 0$.

Recall that in Perron-Frobenius theory, a nonnegative matrix M is said to be *primitive* if there is an index l such that $M^l \gg 0$, see [BP94, Ch. 2]. Assumption (A2) holds in particular under the following elementary condition:

(A2') For any $a \in \mathcal{A}$, cluster k and $\xi \in \Xi$, $P^k(a, \xi) \gg 0$.

We will also make the following assumption:

(A3) There exists a constant M_r such that, $|\theta^{kn}(a, \xi)| \leq M_r$ for every $k \in [K]$, $n \in [N]$, $a \in \mathcal{A}$ and $\xi \in \Xi$.

Condition (A2) has appeared in [Gau96] in the context of semigroup theory, it can be checked algorithmically by reduction to a problem of decision for finite semigroups, see Rk. 3.8, *ibid.* Observe that (A3) is very reasonable in practice.

We equip the product of simplices Δ_N^K with the norm $\|\mu\| := \sum_{k=1}^K \|\mu^k\|_1$. It follows from (A3) that for any action a and realization $\xi \in \Xi$, the total reward function $\mu \mapsto r(a, \mu, \xi)$ is a M_r-Lipschitz real-valued function from $(\Delta_N^K, \|\cdot\|)$ to $(\mathbb{R}, |\cdot|)$.

C. Optimality criteria

We suppose that the controller aims to maximize her average long-term reward, i.e.,

$$g^*(\mu_0) = \sup_{\pi\in\Pi} \liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} r(\pi_t(\mu_t), \mu_t) \ , \qquad (4)$$

where $r(a, \mu) = \int_\Xi r(a, \mu, \xi) d\sigma(\xi)$. Starting from μ_0, the population distribution will evolve in Δ_N^K through the dynamics described in Equation (2) according to a policy $\pi \in \Pi$. Nonetheless, with the assumptions we made, we next show that the dynamics effectively evolves on a particular subset of Δ_N^K.

Let $Q_L^k(a_{:L}) := \prod_{l\in[L]} P^k(a_l)$ be the transition matrix over L time steps, and $\mathcal{D}_L$ be defined as $\mathcal{D}_L = \bigtimes_{k\in[K]} \mathcal{D}_L^k$ where

$$\mathcal{D}_L^k = \operatorname{vex}\left(\left\{ \mu^k Q_L^k(a_{:L}, \xi_{:L}) \;\middle|\; \begin{array}{l} a_{:L} \in \mathcal{A}^L, \\ \mu^k \in \Delta_N, \\ \xi_{:L} \in \Xi^L \end{array} \right\} \right) \ . \qquad (5)$$

Lemma 1: Let (A1)-(A2') hold. Then $\mathcal{D}$ is a compact set included in the relative interior of Δ_N^K. Moreover, for $t \geq 1$, $\mu_t \in \mathcal{D}$ for any policy $\pi \in \Pi$.

Proof: The set $\{\mu^k Q^k(a_{:L}, \xi_{:L}) \mid (a_{:L}, \mu^k, \xi_{:L}) \in \mathcal{A}^L \times \Delta_N \times \Xi^L\}$ is compact, by continuity of $(a, \mu, \xi) \mapsto \mu Q^k(a, \xi)$ and compactness of Δ_N, $\mathcal{A}$ and Ξ. Therefore, $\mathcal{D}_L$ is compact as it is the convex hull of a compact set in finite dimension. Then, the positiveness of Q^k implies that $\mathcal{D}_L^k \subset \operatorname{relint}(\Delta_N)$. Moreover, by property of the semiflow, $\phi_t(a_{:t}, \xi_{:L}, \mu_0) = \phi_L\left(a_{t-L+1:t}, \xi_{t-L+1:t}, \phi_{t-L}(a_{:t-L}, \xi_{:t-L}, \mu_0)\right) \in \mathcal{D}_L$. ∎

We recall that the relative interior of the simplex, equipped with Hilbert's projective metric, is a complete metric space, on which the Hilbert's metric topology is the same as the Euclidean topology, see [LN09, § 2.5]. Hence, under (A1) and (A2), $(\mathcal{D}_L, d_H)$ is a complete metric space. We also recall *Birkhoff theorem*, which shows that every matrix $Q \gg 0$ is a contraction in Hilbert's projective metric, i.e., for all $\xi, a \in \Xi \times \mathcal{A}$ and $\mu, \nu \in \mathcal{D}$,

$$d_H(\mu P(a, \xi), \nu P(a, \xi)) \leq \kappa(P(a, \xi)) d_H(\mu, \nu) \ , \qquad (6)$$

where

$$\kappa(Q) := \tanh\left(\operatorname{Diam}_H(Q) \,/\, 4\right) < 1 \ ,$$

see [LN09, Appendix A]. This property applies to the transition matrix $P^k(a)$ under (A2'), or to Q_L^k under (A2).

D. Ergodic eigenproblem

For any real-valued function $v : \Delta_N^K \to \mathbb{R}$ and discount factor $\alpha \in]0, 1]$, we define the Bellman operator $\mathcal{B}_\alpha$ as

$$\mathcal{B}_\alpha\, v(\mu) = \max_{a\in\mathcal{A}} \int_\Xi \left[r(a, \mu, \xi) + \alpha v(\mu P(a, \xi)) \right] d\sigma(\xi) \ . \qquad (7)$$

For $\alpha = 1$, we simply write $\mathcal{B} = \mathcal{B}_1$. For $\alpha < 1$, we denote by v_α the solution of $\mathcal{B}_\alpha\, v = v$, which can be obtained as the limit of the sequence $(v_\alpha^j)_{j\in\mathbb{N}}$ where $v_\alpha^{j+1} = \mathcal{B}_\alpha\, v_\alpha^j$ and $v_\alpha^j \equiv 0$. This result follows from the fact that the Bellman operator $\mathcal{B}_\alpha$ is a sup-norm contraction, for $\alpha < 1$. A first observation is that $\mu \mapsto (\mathcal{B}\, v)(\mu)$ is convex for any real-valued convex function v. Indeed, the transition dynamics (2) is linear in μ, as well as the reward (3); therefore, for any $a \in \mathcal{A}$, the expression under the maximum is convex in μ, and since the maximization preserves the convexity, the observation is established. For a feedback policy π, we also define $\mathcal{B}^\pi$ the Kolmogorov operator such that $\mathcal{B}^\pi\, v\,(\mu) = \int_\Xi \left[r(\pi(\mu), \mu, \xi) + v(\mu P(\pi(\mu), \xi)) \right] d\sigma(\xi)$.

1) Existence of a solution: As mentioned previously, neither the minorization condition [Kur89; HL96; Wie19] nor the controllability condition [KM97] apply in our situation. Instead, we exploit here the contraction properties of the dynamics, with respect to Hilbert's projective metric, together with the vanishing discount approach, to show the existence. First let us show a preliminary result on metrics comparison:

Lemma 2: Let $\mathcal{D} \subset \operatorname{relint}(\Delta_n)$, $n \in \mathbb{N}$ and $x, y \in \mathcal{D}$. Then,

$$n \|x - y\|_\infty \leq d_H(x, y) \Upsilon(\operatorname{Diam}_H(\mathcal{D})) \qquad (8)$$

where $\Upsilon(d) = \frac{1}{d} e^d (e^d - 1)$.

Proof: We use the results in [AGN15]: Lemma 2.3 shows that for any vectors $u, x, y \in \mathcal{D}$ such that there exist $a, b > 0$ satisfying $ax \leq u \leq bx$ and $ay \leq u \leq by$, we have the following inequality:

$$\|x - y\|_u \leq \left(e^{d_T(x,y)} - 1\right) e^{\max(d_T(x,u), d_T(y,u))} \ ,$$

where d_T denotes the Thompson distance, and $\|z\|_u = \inf\{a > 0 \mid -au \leq z \leq au\}$. In particular, by choosing $u = (1/n, \dots 1/n)$ as the center of the simplex, $\|\cdot\|_u = n \|\cdot\|_\infty$. Moreover, $d_T(\cdot, \cdot) \leq d_H(\cdot, \cdot)$ on $\mathrm{relint}(\Delta_N^K)$, see [AGN15, Eq. 2.4]. Therefore,

$$\begin{aligned} n \|x - y\|_\infty &\leq \left(e^{d_H(x,y)} - 1\right) e^{\max(d_H(x,u), d_H(y,u))} \\ &\leq \left(e^{d_H(x,y)} - 1\right) e^{\mathrm{Diam}_H(\mathcal{D})} \ . \end{aligned}$$

We easily conclude using the fact that $f : x \mapsto e^x - 1$ is a convex function, and so for all $0 \leq x \leq \bar{x}$, $f(x) \leq x \frac{e^{\bar{x}} - 1}{\bar{x}}$. ∎

Applying Lemma 2, we obtain that $\mu \mapsto r(a, \mu, \xi)$ is Lipschitz of constant $M_r^{\mathcal{D}} := \frac{1}{K} M_r \Upsilon(\mathrm{Diam}_H(\mathcal{D}_L))$ for the Hilbert metric.

Let us define the optimal infinite horizon discounted objective v_α, defined as

$$v_\alpha(\mu_0) = \sup_{\pi \in \Pi} \sum_{t \geq 1} \alpha^{t-1} r(\pi_t(\mu_t), \mu_t) \ , \tag{9}$$

where α is the discount factor and μ_0 is the initial distribution. As a consequence of Lemma 2, we obtain that the value functions of the discounted problems constitute an equi-Lipschitz family:

Lemma 3 (Equi-Lipschitz property): Assume that (A1)-(A3) hold. Then, $(v_\alpha)_{\alpha \in (0,1)}$ is $\left(\frac{M_r^{\mathcal{D}}}{1-\kappa}\right)$-equi-Lipschitz on $\mathcal{D}_L$ for the Hilbert metric, i.e., for all $\mu_0, \nu_0 \in \mathcal{D}_L$,

$$|v_\alpha(\mu_0) - v_\alpha(\nu_0)| \leq \frac{M_r^{\mathcal{D}}}{1 - \kappa} d_H(\mu^0, \nu^0) \ .$$

Proof: We first make the proof under the stronger assumption (A2'), and then deduce the general result.

We denote by $(v_\alpha^j)_{j \in \mathbb{N}}$ the sequence defined as $v_\alpha^{j+1} = \mathcal{B}_\alpha v_\alpha^j$ and $v_\alpha^0 \equiv 0$. Let us assume that for a given $j \in \mathbb{N}$, v_α^j is M_α^j-Lipschitz w.r.t the Hilbert metric , i.e.,

$$\left|v_\alpha^j(\mu) - v_\alpha^j(\nu)\right| \leq M_\alpha^j d_H(\mu, \nu) \ .$$

Then, for $\mu, \nu \in \mathcal{D}_1 \subset \Delta_K^N$, we have:

$$\begin{aligned} &\left|\mathcal{B}_\alpha v_\alpha^j(\mu) - \mathcal{B}_\alpha v_\alpha^j(\nu)\right| \\ &\leq \int_\Xi |r(a, \mu, \xi) - r(a, \nu, \xi)| \, d\sigma(\xi) \\ &\quad + \alpha \int_\Xi \left|v_\alpha^j(\mu P(a, \xi)) - v_\alpha^j(\nu P(a, \xi))\right| d\sigma(\xi) \\ &\leq M_\alpha^{j+1} d_H(\mu, \nu), \end{aligned}$$

with $M_\alpha^{j+1} = M_r^{\mathcal{D}} + \alpha\kappa M_\alpha^j$. Therefore, for all $j \in \mathbb{N}$, $M_\alpha^j \leq M := \frac{M_r^{\mathcal{D}}}{1-\kappa}$, which is independent of j and α. So, at the limit, v_α is M-equi-Lipschitz w.r.t. the Hilbert pseudo-metric.

To deduce the general result with (A2), we define

- $\widetilde{\mathcal{A}} := \mathcal{A}^L$, $\widetilde{\Xi} := \Xi^L$, $\tilde{\alpha} := \alpha^L$,
- $\tilde{\phi}_\tau(\tilde{a}_{:\tau}, \tilde{\xi}_{:\tau}, \mu_0) := \mu_0 \prod_{1 \leq t \leq \tau} Q(\tilde{a}_t, \tilde{\xi}_t)$,
- $\tilde{r}(a_{:L}, \mu, \xi_{:L}) := \sum_{l \in [L]} \alpha^{l-1} r(a_l, \phi_l(a_{:l}, \mu, \xi_{:l}), \xi_l)$,
- and

$$\tilde{\mathcal{B}}_\alpha v(\mu) = \max_{\tilde{a} \in \widetilde{\mathcal{A}}} \left\{ \begin{array}{l} \displaystyle\int_{\widetilde{\Xi}} \left[\tilde{r}(\tilde{a}, \mu, \tilde{\xi}) + v(\nu)\right] d\tilde{\sigma}(\tilde{\xi}) \\ \text{s.t } \nu = \mu Q_L(\tilde{a}, \tilde{\xi}) \end{array} \right\}$$

.

and observe that

$$v_\alpha(\mu_0) = \sum_{\tau \geq 1} \tilde{\alpha}^{\tau - 1} \tilde{r}(\tilde{a}_\tau, \tilde{\phi}_\tau(\tilde{a}_{:\tau}, \tilde{\xi}_{:\tau}, \mu_0)) \ .$$

We have rescaled the time (τ instead of t) so that the transition matrix between time τ and time $\tau+1$ is $Q_L(\tilde{a}_\tau, \tilde{\xi}_\tau)$. One τ-time step corresponds to L t-time steps. As the transition $Q_L(\tilde{a}, \tilde{\xi})$ is now positive, the proof is exactly the same as before, in the τ-time space. ∎

Remark 1: The result remains if the common noise ξ_t is controlled, i.e., depends on the action a_t. However, if ξ_t depends on the state, then there is no guarantee that M_α^{j+1} is bounded. In the latter case, it would require $\left\|v_\alpha^j\right\|_\infty$ to be uniformly bounded, which is not guaranteed (when $\alpha \to 1$, it goes to infinity in general).

We are now able to prove the main result:

Theorem 1 (Existence of a solution): Assume that (A1)-(A3) hold. Then, the ergodic eigenproblem

$$g \, 1_{\mathcal{D}_L} + h = \mathcal{B} \, h \tag{10}$$

admits a solution $h^* \in \mathrm{Lip}(\mathcal{D}_L) \cap \mathrm{Vex}(\mathcal{D}_L)$ and $g^* \in \mathbb{R}$.

Proof: Let us define a reference distribution $\overline{\mu} \in \Delta_N^K$, $g_\alpha^* = (1 - \alpha) v_\alpha(\overline{\mu})$ and $h_\alpha^* = v_\alpha - v_\alpha(\overline{\mu}) \, 1_{\mathcal{D}_L}$. Then, as v_α is equi-Lipschitz on $\mathcal{D}_L$ (Lemma 3), h_α^* is equi-bounded and equi-Lipschitz on $\mathcal{D}_L$ (in particular equi-continuous). By the Arzelà-Ascoli theorem, $h_\alpha^* \to h^* \in \mathcal{C}^0(\mathcal{D}_L)$.

Finally, from the discounted reward approach, we get $\mathcal{B}(\alpha v_\alpha) = v_\alpha$, therefore

$$\frac{g_\alpha^*}{1 - \alpha} \, 1_{\mathcal{D}_L} + h_\alpha^* = \mathcal{B}\left(\frac{\alpha g_\alpha^*}{1 - \alpha} \, 1_{\mathcal{D}_L} + \alpha h_\alpha^*\right) \ .$$

By the additive homogeneity property of the Bellman function, $g_\alpha^* \, 1_{\mathcal{D}_L} + h_\alpha^* = \mathcal{B}(\alpha h_\alpha^*)$. The fixed-point equation (10) is then obtained by continuity of the Bellman operator $\mathcal{B}$.

To conclude, h^* is convex since v_α is convex and the pointwise convergence preserves the convexity. ∎

Proposition 1: For any solution (g^*, h^*) of (10), g^* satisfies (4), and a maximizer $a^*(\cdot) \in \arg\max \mathcal{B} \, h^*$ defines an optimal *stationary* policy for the average gain problem.

Proof: Let $\pi \in \Pi$ be a policy. By definition, for every t, $\mathcal{B}^{\pi_t} h^* \leq \mathcal{B} \, h^* = h^* + g^* \, 1_{\mathcal{D}_L}$. Therefore, iterating the Kolmogorov operator, we obtain

$$(\mathcal{B}^{\pi_1} \circ \dots \circ \mathcal{B}^{\pi_t}) h^* \leq h^* + t g^* \, 1_{\mathcal{D}_L} \ .$$

Let $\underline{h}^* := \min_{\mu \in \mathcal{D}_L} h^*(\mu)$ be the minimum of h^*. Then, $0_{\mathcal{D}_L} \leq h - \underline{h}^* \, 1_{\mathcal{D}_L}$, and so $(\mathcal{B}^{\pi_1} \circ \dots \circ \mathcal{B}^{\pi_t})(0_{\mathcal{D}_L}) \leq h^* + (t g^* - \underline{h}^*) \, 1_{\mathcal{D}_L}$. Finally,

$$\liminf_{t \to \infty} \frac{1}{t} (\mathcal{B}^{\pi_1} \circ \dots \circ \mathcal{B}^{\pi_t})(0_{\mathcal{D}_L})(\mu_0) \leq g^* \ .$$

Any strategy has an average reward lower than g^*. As we have proved that the bias function h^* is continuous on $\mathcal{D}_L$, a maximizer $a^*(\mu)$ can be found for any state μ, and so playing the strategy $a^*(\mu)$ achieves the best possible average gain g^*. ■

In particular, the constant g^* in (10) is unique, and it coincides with the optimal average long-term reward, for all choices of the initial state μ_0. However, even if the payoff g^* is unique, the bias function h^* is not (and so neither is the optimal policy).

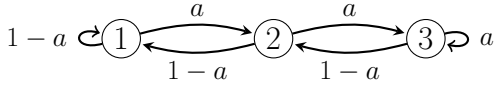
Remark 2: In [Bäu23], Bäuerle also used the vanishing discount approach, but here we do not assume a priori the equi-boundedness of the optimal discounted objective functions v_α. Instead, using a contraction argument on the dynamics, we obtained that $(v_\alpha)_{\alpha\in(0,1)}$ is equi-Lipschitz (see Lemma 3). In particular, it entails that any optimal eigenvector h^* is Lipschitz (and not only upper semi-continuous).

In the sequel, we restrict the study to deterministic problem (absence of common noise, i.e., Ξ is reduced to a singleton).

2) Non-uniqueness of the solution: As discussed in the introduction, classical approaches to the infinite dimensional ergodic problem rely on a geometric ergodicity/Doeblin type condition. This condition entails that the bias function is unique up to an additive constant. We next show that under (A1) to (A3), the bias function may not be unique, implying that the present results cannot be derived from such approaches.

To get a non-unique bias, we will construct instances where there exist several "attractor" states, and where a family of strategies can be found so that each of them secures the optimal mean payoff. Then, different attractors lead to different bias functions. To illustrate this fact, we introduce in Example 1 a deterministic model satisfying (A1) to (A3). Note that taking the same dynamics as in the example but without node 2 can also lead to a non-unique solution of the eigenproblem as long as we allow for a more general form of reward $r(a,\mu)$. Here, we aim at fitting exactly with our application case by considering that the reward function satisfies Eq. (3).

Example 1 (Non-uniqueness of the eigenvector):



Let us consider the dynamical system described by the following transition matrix:

$$P(a) = \begin{bmatrix} 1-a & a & 0 \\ 1-a & 0 & a \\ 0 & 1-a & a \end{bmatrix},$$

where the a is supposed to belong to the action space $\mathcal{A}$, which is of the form

$$\mathcal{A} = [a_0, a_1]\ , \quad 0 < a_0 < 1/2 \ \text{ and } \ a_1 = 1 - a_0\ .$$

We consider the following unitary reward $\theta(\cdot)$:

$$\theta(a)_n = \begin{cases} 1-a \text{ if } n=1, \\ 0 \quad \text{if } n=2, \\ a \quad \text{if } n=3, \end{cases}$$

The reward function is then $r(a,\mu) = (1-a)\mu_1 + a\mu_3$ for any $\mu \in \Delta_3$, and

$$r(a, \mu P(a)) = (1-a)^2(1-\mu_3) + a^2(1-\mu_1)\ .$$

In the sequel, we work in the sub-simplex $\Delta_3^{\leqslant} := \{(x,y) \in \mathbb{R}^2_{\geq 0} \mid x+y \leq 1\}$, considering that μ_2 can be reconstructed as $\mu_2 = 1-\mu_1-\mu_3$.

a) The associated ergodic eigen problem.: For any real-valued function $v : \Delta_3^{\leqslant} \to \mathbb{R}$, let us define the Bellman operator $\mathcal{B}$ as

$$\mathcal{B}\, v(\mu_1,\mu_3) = \max_{a\in\mathcal{A}} \left\{ \begin{array}{l} (1-a)^2(1-\mu_3) + a^2(1-\mu_1) \\ + v\left((1-a)(1-\mu_3), a(1-\mu_1)\right) \end{array} \right\}\ .$$

In Example 1, the transition $a \in \mathcal{A} \mapsto P(a)$ is linear. Moreover, the transition matrix over two time steps is then

$$(P(a))^2 = \begin{bmatrix} (1-a)^2 + a(1-a) & a(1-a) & a^2 \\ (1-a)^2 & 2a(1-a) & a^2 \\ (1-a)^2 & a(1-a) & a^2 + a(1-a) \end{bmatrix}$$

and has positive coefficients. Therefore, the transition matrix $P(a)$ satisfies the primitivity assumption (A2) for all $a \in \mathcal{A}$. Using Theorem 1, the ergodic eigenproblem

$$g 1_{\mathcal{D}_1} + h = \mathcal{B}\, h \tag{11}$$

admits a solution $h^* \in \mathrm{Lip}(\mathcal{D}_1) \cap \mathrm{Vex}(\mathcal{D}_1)$ and $g^* \in \mathbb{R}$, where $\mathcal{D}_1$ is defined one can construct the effective domain $\mathcal{D}_1$ as in (5). As the quantity in the maximum is convex, for any convex function $v : \mathcal{D}_1 \to \mathbb{R}$, the maximum value in $\mathcal{B}\, v$ is obtained for $a = a_0$ or $a = a_1$. Therefore, in the sequel, we restrict wlog the state space to be $\mathcal{A} = \{a_0, a_1\}$.

b) Steady states.: Let $k \in \{0,1\}$. The equilibrium distribution $\hat{\mu}$ achieved by a constant decision a_k is given by the equation $\hat{\mu} P(a_k) = \hat{\mu}$, which has a unique solution:

$$\hat{\mu}_1^k = \frac{(1-a_k)^2}{1-a_k(1-a_k)},\ \hat{\mu}_3^k = \frac{a_k^2}{1-a_k(1-a_k)}\ . \tag{12}$$

c) Bias function for the Kolmogorov operator.: Let us define $\mathcal{B}^k$ the Kolmogorov operator associated to the constant strategy $\pi : \mu \mapsto a_k$, i.e.,

$$\begin{aligned} \mathcal{B}^k\, v(\mu_1,\mu_3) =& (1-a_k)^2(1-\mu_3) + a_k^2(1-\mu_1) \\ & + v\left((1-a_k)(1-\mu_3), a_k(1-\mu_1)\right)\ . \end{aligned}$$

Then, the linear function $h^k(\mu_1,\mu_3) = \alpha^k\mu_1 + \beta^k\mu_3$ and the gain g^k are solutions of

$$h^k(\mu_1,\mu_3) + g^k = \mathcal{B}^k\, h^k(\mu_1,\mu_3),\ (\mu_1,\mu_3) \in \mathcal{D}_1 \tag{13}$$

if and only g^k, α^k and β^k satisfy the following system

$$\begin{cases} g^k = (1-a_k)^2 + a_k^2 + (1-a_k)\alpha^k + a_k\beta^k \\ \alpha^k = -a_k^2 - a_k\beta^k \\ \beta^k = -(1-a_k)^2 - (1-a_k)\alpha^k \end{cases},$$

where the unique solution of the latter system is given by

$$g^k = \frac{a_k^3 + (1-a_k)^3}{1 - a_k(1-a_k)},$$
$$\alpha^k = \frac{a_k(1-a_k)^2 - a_k^2}{1 - a_k(1-a_k)}, \tag{14}$$
$$\beta^k = \frac{(1-a_k)a_k^2 - (1-a_k)^2}{1 - a_k(1-a_k)} \quad .$$

Note that $g^0 = g^1$ since $a_0 + a_1 = 1$, and we simply denoted it by g^*.

d) Solution for the ergodic eigen problem.: We now exhibit a family of eigenvectors where each of them constitutes a solution to the ergodic eigenproblem associated with Example 1:

Theorem 2 (Non-uniqueness of the eigenvector): Let $v^k : \mathcal{D} \to \mathbb{R}$, $k \in \{0,1\}$, be defined as

$$v^k(\mu_1,\mu_3) = \hat{h}^{k0}(\mu_1,\mu_3) \vee \hat{h}^{k1}(\mu_1,\mu_3) \vee \hat{h}^{k2}(\mu_1,\mu_3) \quad , \tag{15}$$

with

- $\hat{h}^{ij}(\cdot,\cdot) := h^j(\cdot,\cdot) - h^j(\hat{\mu}_1^i, \hat{\mu}_3^i),\ i,j \in \{0,1\}$,
- $\hat{h}^{k2}(\cdot,\cdot) := \mathcal{B}^{1-k}\,\hat{h}^{kk}(\cdot,\cdot) - g^*$.

Then, for any $\lambda \in [0,1]$, the couple (v^λ, g^*) is solution the ergodic eigenproblem (11) – corresponding to Example 1 – with g^* defined in (14) and

$$v^\lambda(\mu_1,\mu_3) := \left(v^0(\mu_1,\mu_3) - \tfrac{\lambda}{1-\lambda}\right) \vee \left(v^1(\mu_1,\mu_3) - \tfrac{1-\lambda}{\lambda}\right) \quad .$$

Proof: As first observation, the couple (h^k, g^*), solution of (13), is not solution of the ergodic eigenproblem (11). Therefore, let us try to construct a solution as a mixture of h^0 and h^1. To this purpose, let us define the function the function $u^0 : \mathcal{D}_1 \to \mathbb{R}$ as

$$u^0(\mu_1,\mu_3) = \hat{h}^{00}(\mu_1,\mu_3) \vee \hat{h}^{01}(\mu_1,\mu_3) \quad .$$

For $(\mu_1,\mu_3) \in \mathcal{D}_1$, the value of $\mathcal{B}\,u^0(\mu_1,\mu_3)$ is given by the maximum of 4 quantities:

(i) $\mathcal{B}^0\,\hat{h}^{00}(\mu_1,\mu_3) = \hat{h}^{00}(\mu_1,\mu_3) + g^*$,
(ii) $\mathcal{B}^0\,\hat{h}^{01}(\mu_1,\mu_3)$,
(iii) $\mathcal{B}^1\,\hat{h}^{00}(\mu_1,\mu_3)$,
(iv) $\mathcal{B}^1\,\hat{h}^{01}(\mu_1,\mu_3) = \hat{h}^{01}(\mu_1,\mu_3) + g^*$,

The equality in (i) and (iv) comes from the fact that $\hat{h}^{00}$ and $\hat{h}^{01}$ are solutions for (13). Besides, we can prove using basic algebra that

$$\mathcal{B}^0\,\hat{h}^{01}(\mu_1,\mu_3) - \mathcal{B}^1\,\hat{h}^{00}(\mu_1,\mu_3) = g^*(\hat{\mu}_3^0 - \hat{\mu}_1^0) \le 0 \quad .$$

Therefore, the maximum is obtained either with (i), (iii) or (iv). We consider now the function

$$v^0(\mu_1,\mu_3) = \hat{h}^{00}(\mu_1,\mu_3) \vee \hat{h}^{01}(\mu_1,\mu_3) \vee \hat{h}^{02}(\mu_1,\mu_3) \quad , \tag{16}$$

with $\hat{h}^{02}(\mu_1,\mu_3) := \mathcal{B}^1\,\hat{h}^{00}(\mu_1,\mu_3) - g^*$. By construction, $v^0 = \mathcal{B}\,u^0 - g^*$. Moreover, one can show that $\mathcal{B}\,\hat{h}^{02}(\mu_1,\mu_3) - g^* \le v^0(\mu_1,\mu_3)$. Therefore, for all $(\mu_1,\mu_3) \in \mathcal{D}_1$,

$$\mathcal{B}\,v^0(\mu_1,\mu_3) = \mathcal{B}\,u^0(\mu_1,\mu_3) = v^0(\mu_1,\mu_3) - g^* \quad .$$

As a conclusion, (v^0, g^*) is a solution of (11).

By symmetry of the problem, we can construct the function $v^1(\mu_1,\mu_3) = v^0(\mu_3,\mu_1)$, and (v^1, g^*) is a different solution of (11). Finally, each max-plus combination of v^0 and v^1 also constitutes a solution of the ergodic eigenproblem. ■

We display in Figure 1 the eigenvector v^0, $v^{1/2}$ and v^1, obtained numerically (using the RVI procedure, see Algorithm 1), as with the eigenvector v^0, obtained theoretically (see above).

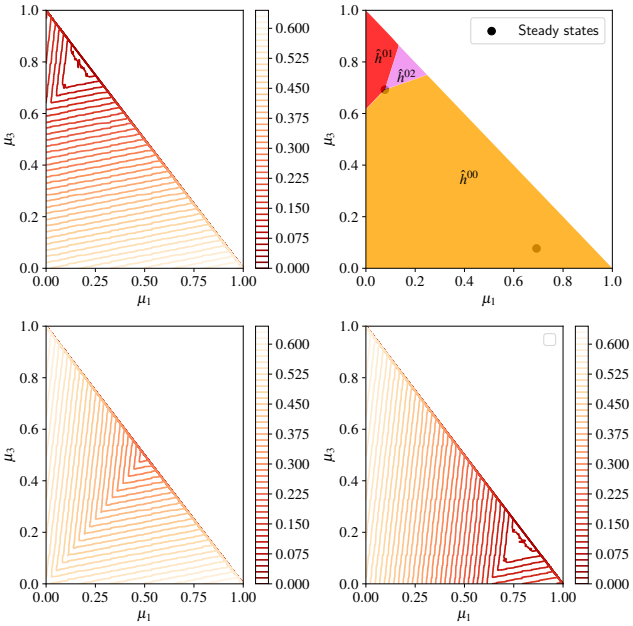


Fig. 1: Eigenvectors for Example 1 with $\mathcal{A} = [0.25, 0.75]$. (Upper left): the eigenvector v^0 (obtained by the RVI algorithm, see Section III). (Upper right): the theoretical v^0 found in (16), showing the two steady states to which each of optimal strategies converges. (Lower left): the eigenvector $v^{1/2}$. (Lower right): the eigenvector v^1.

III. Numerical resolution

We present in this section two iterative algorithms in order to numerically solve the ergodic eigenproblem (10).

A. Relative Value Iteration with Krasnoselskii-Mann damping

Relative Value Iteration (RVI) has been extensively studied to solve unichain finite-state MDP [Put94; Ber98]. Simplicial state-spaces appear in particular in the definition of *belief state* for partially observable MDP [Hau00]. For such continuous state-spaces, a discretization must be done as a prerequisite to RVI algorithm. Here, we define a regular grid Σ of the simplex Δ_N^K, and $\mathcal{B}^\Sigma$ the Bellman Operator with a linear point approximation on the grid Σ, achieved by a Freudenthal triangulation [Lov91]. With this framework, we have the following property:

Proposition 2 ([Hau00], Thm 12): For any $v \in \mathrm{Vex}(\Delta_N^K)$,

$$\mathcal{B}\,v \le \mathcal{B}^\Sigma\,v \quad .$$

As the bias function $\hat{h}$ is convex at each iteration, the solution return by Algorithm 1 provides a gain which is an upper bound of the optimal gain g^*.

Algorithm 1 RVI with Mann-type iterates

Require: Grid Σ, Bellman operator $\mathcal{B}^\Sigma$, initial function $\hat{h}_0$
Initialize $\hat{h} = \hat{h}_0$, $\hat{h}'(\mu) = \mathcal{B}^\Sigma \hat{h}$
while $\mathrm{sp}(\hat{h}' - \hat{h}) > \epsilon$ **do**
  $\hat{h} \leftarrow (\hat{h}' - \max\{\hat{h}'\}e + \hat{h})/2$
  $\hat{h}'(\hat{\mu}) \leftarrow (\mathcal{B}^\Sigma \hat{h})(\hat{\mu})$ for all $\hat{\mu} \in \Sigma$
end while
$\hat{g} \leftarrow (\max(\hat{h}' - \hat{h}) + \min(\hat{h}' - \hat{h}))/2$
return $\hat{g}, \hat{h}$

In Algorithm 1, we use, following [GS20], a mixture of the classical relative value iteration algorithm [Put94] with a *Krasnoselskii-Mann* damping. As detailed in [GS20] (Th. 9 and Coro 13), it follows from a theorem of Ishikawa that the sequence of bias function $\hat{h}$ does converge, and it follows from a theorem of Baillon and Bruck that $\hat{g}$ provides an ϵ approximation of the optimal average cost g^* after $O(1/\epsilon^2)$ iterations.

B. Howard algorithm with on-the-fly transition generation

We focus here on an other class of iterative methods to solve MDPs, namely *policy iteration* (PI) algorithms, initiated by Howard (see e.g [DF68] of [Put94]). For deterministic Markov decision processes, a combinatorial implementation of Howard algorithm, with a linear-time per policy, was given in [Coc+98]. We refine the latter algorithm, with a method adapted to "decomposable" state spaces.

Let $\Lambda = (\hat{\mu}_i)_{i\in[M]}$ be a *local* semi-Lagrangian discretization of the simplex Δ_N of size $M := |\Lambda|$. We refer the grid to be *local*, since the discretization is done for the probability space of one sub-population and not on the *global* probability space Δ_N^K. The *global* discretization is then

$$\Sigma = (\hat{\mu}_{\vec{i}_1}, \dots, \hat{\mu}_{\vec{i}_K})_{\vec{i}\in[M]^K} \ .$$

We define the local transition operator $T^{\Lambda,k} : (i,a) \in [M] \times \mathcal{A} \mapsto \arg\min_{j\in[M]} \|\hat{\mu}_i P^k(a) - \hat{\mu}_j\|_\infty$. For each $k \in [K]$, this operator can be computed in a preprocessing step, and stored in $O(M \times |\mathcal{A}|)$. Note that contrary to the RVI algorithm – where a Freudenthal triangulation is performed during the computation of $\mathcal{B}^\Sigma$ – the transition operator is here approximated by finding the closest discretization point (in the L_∞-norm) to the real next state.

The *global* transition can then be obtained *on-the-fly*, i.e., for any action $a \in \mathcal{A}$ and global index $\vec{i} \in [M]^K$, $T^\Sigma(\vec{i}, a)$ can be recomputed whenever it is required in the algorithm knowing the sub-transition $T^{\Lambda,k}(\vec{i}_k, a)$ for all $k \in \mathbb{N}$:

$$T^\Sigma : (\vec{i}, a) \in [M]^K \times \mathcal{A} \mapsto (T^{\Lambda,k}(\vec{i}_k, a))_{k\in[K]} \ . \quad (17)$$

Remark 3: A complete storage of T^Σ would lead to a memory occupation in $O(M^K \times |\mathcal{A}|)$, whereas the storage of all $T^{\Lambda,k}$, $k \in [K]$, is in $O(K \times M \times |\mathcal{A}|)$.

Algorithm 2 Howard Algorithm with on-the-fly transition generation

Require: Local grid Λ, family of local transitions $(T^{\Lambda,k})_{k\in[K]}$, initial decision vector $\hat{d}'$
do
  $\hat{d} \leftarrow \hat{d}'$
  $\hat{g}, \hat{h}$ solution of ▷ Policy Evaluation
$$\begin{cases} \hat{g} + \hat{h}_{\vec{i}} = r(\hat{d}_{\vec{i}}, \hat{\mu}_{\vec{i}}) + \hat{h}_{\vec{j}}, \ \vec{i} \in \Sigma \\ \vec{j} = T^\Sigma(\vec{i}, \hat{d}_{\vec{i}}) \end{cases}$$
  for $\vec{i} \in \Sigma$ **do** ▷ Policy Improvement
    $\hat{d}'_{\vec{i}} \leftarrow \arg\min_{a\in\mathcal{A}} \begin{Bmatrix} r(a, \hat{\mu}_{\vec{i}}) + \hat{h}_{\vec{j}} \\ \text{s.t.} \vec{j} = T^\Sigma(\vec{i}, a) \end{Bmatrix}$
  end for
while $\hat{d}' \neq \hat{d}$
return $\hat{g}, \hat{d}$

Algorithm 2 shows the Howard algorithm with on-the-fly transition generation. It consists in alternating a policy evaluation step with a policy improvement step. We implemented a parallelized version of this algorithm[1] by adapting the code of [Coc+98], initially intended for computing spectral elements in max-plus algebra. The algorithm is known to have experimentally a superlinear convergence which, in finite action-space setting, is reached in finitely many steps, see e.g.[Put94]. Despite the decomposable transition, all the subpopulations $k \in [K]$ are linked together through a common policy. In the implementation, both the policy $\hat{d}$ and the bias function $\hat{h}$ depend on the global state associated to index $\vec{i} \in [M]^K$. Therefore, the memory needed to run the algorithm is still exponential in the number of segments – in $O(M^K)$ – but would have been worst with stored global transition T^Σ – in $O(M^K \times |\mathcal{A}|)$ – the action space being very large in general, see Remark 3. We provide in Table I below benchmarks showing the gain (speedup and memory usage) brought by this approach.

IV. Steady-state optimality

A. Definition

It is of interest to investigate cases in which the dynamic problem reduces to a static one. In fact, in some cases the optimal stationary policy may be a simple policy that attracts the system to a steady-state ("get there, stay there" – [Fly79]). For instance, Bauerle [Bäu23] derives a class of mean-field MDPs solvable by a static program.

Definition 1: Let $\mathcal{S} = \{(a,\mu) \in \mathcal{A} \times \Delta_N^K \,|\, \mu = \mu P(a)\}$ be the action-space domain of stationary probabilities. Then, $\mu \in \Delta_N^K$ is a *steady-state* if there exists $a \in \mathcal{A}$ such that $(a,\mu) \in \mathcal{S}$.

If (A2) holds, then for any cluster k and any price $a \in \mathcal{A}$, the Markov chain induced by the transition matrix $P^k(a)$ has a unique stationary distribution. We denote by $\overline{\mu}(\cdot) : \mathcal{A} \mapsto \Delta_N^K$ the mapping sending an action to the stationary distribution it induces.

[1]Available at `https://gitlab.com/these_tarif/ergodic_inertia`

Definition 2: The *optimal steady-state gain* $\overline{g}$ is defined as

$$\overline{g} := \max_{(a,\mu)\in\mathcal{S}} r(a,\mu) \ . \tag{18}$$

If (A2) holds, (18) is in general a static nonconvex maximization problem over the actions. Nonetheless, we can expect to solve it efficiently in the case where $\overline{\mu}(\cdot)$ is analytically known, see e.g. Section V. Maximizers $\overline{a}$ are called *optimal steady-state price*, they correspond to a steady-state distribution $\overline{\mu}(\overline{a})$.

B. Optimality gap

In this section we introduce a class of Lagrangian functions designed so that each dual problem turns out to be an upper bound of g^*. This extends the result of [Fly79] involving usual Lagrangian functions. We use here a more general Lagrangian, depending on the choice of a non-linear function φ. This leads to much tighter bounds, and allows us to prove the optimality of a steady-state strategy whenever a zero duality gap is obtained. Let Φ be defined as

$$\Phi = \{\varphi : \Delta_N^K \to \Delta_N^K \text{ injective and bounded}\} \ .$$

For a given function $\varphi \in \Phi$, we define the Lagrangian function $\mathcal{L}^{(\varphi)} : (\mathcal{A}, \Delta_N^K, \mathbb{R}^{KN}) \to \mathbb{R}$ by

$$\mathcal{L}^{(\varphi)}(a,\mu,\lambda) := r(a,\mu P(a)) + \langle \lambda, \varphi\left(\mu P(a)\right) - \varphi(\mu)\rangle_{KN} \ .$$

As a direct consequence of the injectivity of φ, we obtain that for any given $\varphi \in \Phi$,

$$\overline{g} = \max_{(a,\mu)\in\mathcal{A}\times\Delta_N^K} \inf_{\lambda\in\mathbb{R}^{KN}} \mathcal{L}^{(\varphi)}(a,\mu,\lambda) \ .$$

We also define the dual problem $g^{(\varphi)}$ as

$$g^{(\varphi)} := \inf_{\lambda\in\mathbb{R}^{KN}} \max_{(a,\mu)\in\mathcal{A}\times\Delta_N^K} \mathcal{L}^{(\varphi)}(a,\mu,\lambda) \ . \tag{19}$$

The following result bounds the suboptimality gap induced by the restriction to steady-state policies. It will be applied in Figure 3 to study the optimality of such policies, in our application.

Proposition 3: With (g^*, h^*) solution of (10) and $\overline{g}$ defined in (18),

$$\overline{g} \le g^* \le g^{(\varphi)}, \ \forall \varphi \in \Phi \ .$$

Proof: The proof extends the arguments in [Fly79, Remark 5.1] to nonlinear functions $\varphi \in \Phi$.

First, from the geometrical convergence of the dynamics (see Equation (6)), the valid strategy consisting in executing action $\overline{a}$ each period of time induces an average reward of $\overline{g}$, regardless the initial distribution. Therefore, $\overline{g} \le g^*$.

Then, for $\epsilon > 0$, there exists λ^ϵ such that for any $(a,\mu) \in \mathcal{A}\times\Delta_N^k$,

$$r(a,\mu P(a)) + \langle \lambda^\epsilon, \varphi(\mu P(a)) - \varphi(\mu)\rangle_{KN} \le g^{(\varphi)} + \epsilon \ .$$

We construct a sequence of decision $a_1, \dots, a_T$ leading to distribution $\mu_1, \dots, \mu_T$. Then, at each period t,

$$r(a_t,\mu_t) + \langle \lambda^\epsilon, \varphi(\mu_t) - \varphi(\mu_{t-1})\rangle_{KN} \le g^{(\varphi)} + \epsilon \ .$$

Therefore, we take the mean over $t = 1, \dots, T$ to recover the average reward criteria:

$$\frac{1}{T}\sum_{t=1}^{T} r(a_t,\mu_t) + \frac{1}{T}\langle \lambda^\epsilon, \varphi(\mu_T) - \varphi(\mu_0)\rangle_{KN} \le g^{(\varphi)} + \epsilon \ .$$

The second term converges to zero when $T \to \infty$ as we suppose that φ is bounded on the simplex. So,

$$\liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} r(a_t,\mu_t) \le g^{(\varphi)} + \epsilon \ .$$

The latter inequality is valid for any $\epsilon > 0$, and any sequence of action $(a_t)_{t\in\mathbb{N}}$, so $g^* \le g^{(\varphi)}$. ∎

We define the duality gap $\delta_{\mathcal{L}^{(\varphi)}}$ as

$$\delta_{\mathcal{L}^{(\varphi)}} := g^{(\varphi)} - \overline{g}.$$

As an immediate consequence of Proposition 3, if there exists $\varphi \in \Phi$ such that $\delta_{\mathcal{L}^{(\varphi)}} = 0$, then $g^* = \overline{g}$, and the dynamic program 4 reduces to the static optimization program (18). Depending on the problem parameters, the duality gap may, or may not, vanish, see Figure 3.

V. Application to electricity pricing

A. Description

We suppose that an electricity provider has N-1 different types of offers and that a study has distinguished beforehand K customer segments, assuming that customers of a given segment have approximately the same behavior. Given a segment k and an offer $n \in [N\text{-}1]$, the *reservation price* R^{kn} is the maximum price that customers of this segment are willing to spend on n, and E^{kn} is the (fixed) quantity a customer of segment k will purchase if he chooses n. The *utility* for these customers is linear and is defined as

$$U^{kn}(a) := R^{kn} - E^{kn}a^n \ .$$

where a^n is the price for one unit of product n. The action space is then a compact subset of $\mathbb{R}^{N-1}$.

To model the competition between the provider and the other providers of the market, consumers have an alternative option (state of index N). We suppose that this alternative offer is fixed over time (for example a regulated contract). Then, under this assumption, it can be modelized w.l.o.g. by a null utility for each cluster ($U^{kN} = 0$).

If a customer of segment k chooses the contract $n < N$ at price a^n, then the provider receives $E^{kn}a^n$ from the electricity consumption of the customer and has an induced cost of C^{kn}. Note that the cost should depend on the quantity E^{kn}, but as it is supposed to be a parameter, we omit this dependency. The (linear) reward for the provider is then

$$\theta^{kn}(a) = E^{kn}a^n - C^{kn}, \ n < N, \quad \theta^{kN} = 0 \ .$$

We suppose that the transition probability follows a logit response, see e.g. [PE17]:

$$[P^k(a)]_{n,m} = \frac{e^{\beta[U^{km}(a) + \gamma^{kn}\,1_{m=n}]}}{\sum_{l\in[N]} e^{\beta[U^{kl}(a) + \gamma^{kn}\,1_{l=n}]}} \ , \tag{20}$$

where the parameter γ^{kn} is the cost for segment k to switch from contract n to another one, and β is the intensity of the choice (it can represent a "rationality parameter"). One can easily check that (A1)-(A3) are satisfied.

In the no-switching-cost case ($\gamma = 0$), we say that the customers response is *instantaneous*, and corresponds to the classical logit distribution, see e.g. [Tra09]:

$$\mu_L^{kn} = e^{\beta U^{kn}(a)} \; / \sum_{l\in[N]} e^{\beta U^{kl}(a)} \quad . \tag{21}$$

B. Steady-states

The application scope of the transition model we defined in (20) is broader than electricity pricing. For this specific kernel, we derive a closed-form expression for the stationary distributions, fully characterized by the instantaneous response:

Theorem 3: Given a constant action a, the distribution μ_t^k converges to $\overline{\mu}^k(a)$, defined as

$$\overline{\mu}^{kn}(a) = \frac{\eta^{kn}(a)\mu_L^{kn}(a)}{\sum_{l\in[N]} \eta^{kl}(a)\mu_L^{kl}(a)} \quad . \tag{22}$$

where $\eta^{kn}(a) := 1 + \left[e^{\beta\gamma^{kn}} - 1\right] \mu_L^{kn}(a)$, and μ_L is defined in (21).

Proof: In the proof, we forget the dependence on k and a. The stationary probability is defined as $\forall m \in [N]$, $\mu^m \left[1 - P^{mm}\right] = \sum_{n\neq m} \mu^n P^{nm}$. We can then replace by the definition of the probabilities (20) to obtain

$$\mu^m \left[\frac{\sum_{l\neq m} e^{\beta U^m}}{\sum_l e^{\beta[U^l + 1_{l=m}\,\gamma^m]}}\right] = \sum_{n\neq m} \mu^n \left[\frac{e^{\beta U^m}}{\sum_l e^{\beta[U^l + 1_{l=n}\,\gamma^n]}}\right] \quad .$$

Defining $\tilde{\mu}^n := \frac{\mu^n}{\sum_l e^{\beta[U^l + 1_{l=n}\,\gamma^n]}}$, we obtain

$$\forall m \in [N],\; \tilde{\mu}^m \sum_{l\neq m} e^{\beta U^l} = e^{\beta U^m} \sum_{l\neq m} \tilde{\mu}^l \quad .$$

The solution $\tilde{\mu}^n := \lambda e^{\beta U^n}, n \in [N]$ is then a valid solution, and the constant λ is chosen so that $\sum_{l\in[N]} \mu^l = 1$:

$$\begin{aligned} \overline{\mu}^{kn}(a) &= \lambda e^{\beta U^{kn}(a)} \sum_{m\in[N]} e^{\beta[U^{kn}(a) + 1_{m=n}\,\gamma^{kn}]} \\ \lambda^{-1} &= \sum_{n\in[N]} e^{\beta U^{kn}(a)} \sum_{m\in[N]} e^{\beta[U^{km}(a) + 1_{m=n}\,\gamma^{kn}]} \end{aligned} \tag{23}$$

Finally, $\eta^{kn} = \sum_l e^{\beta[U^{kl} + 1_{l=n}\,\gamma^{kn}]} / \sum_l e^{\beta U^{kl}}$. We recover the definition of $\overline{\mu}$ (23). ∎

As a consequence, the optimal steady-state can be found by solving

$$\overline{g} = \max_{a\in\mathcal{A}} r(a, \overline{\mu}(a)) \quad . \tag{24}$$

Problem (24) has no guarantee to be convex. However, it is a box-constrained smooth optimization problem which can be much more efficiently solved (at least up to local maximum) than the original time-dependent problem.

In addition, if we suppose that $\gamma^{kn} = \gamma^k > 0$ for all n, then for any $a \in \mathcal{A}$, we get the two following properties as immediate consequence of Theorem 3:

- $\lim_{\gamma^k\to 0} \overline{\mu}^k(a) = \mu_L^k(a)$,
- $(\overline{\mu}^{kn})$ and (μ_L^{kn}) are sorted in the same order.

We now aim to compare the steady-state $\overline{\mu}^k(a)$ with the logit distribution $\mu_L^k(a)$ using the majorization theory:

Definition 3 (Majorization,[MOA11]): For a vector $a \in \mathbb{R}^d$, we denote by $a^\downarrow \in \mathbb{R}^d$ the vector with the same components, but sorted in descending order. Given $a, b \in \Delta_d$, we say that a majorizes b from below written $a \succ b$ iff

$$\sum_{i=1}^k a_i^\downarrow \geq \sum_{i=1}^k b_i^\downarrow \quad \text{for } k = 1, \ldots, d \quad .$$

Proposition 4 (Majorization property of the steady-state): Let $k \in [K]$ and $a \in \mathcal{A}$ be given. Suppose that $\gamma^{kn} = \gamma^k > 0$ for all $n \in [N]$, then the stationary distribution majorizes the instantaneous logit response i.e.,

$$\overline{\mu}^k(a) \succ \mu_L^k(a) \quad . \tag{25}$$

Proof: Let us suppose that we reorder the probabilities (and the η) such that they are sorted in the decreasing order.

$$\begin{aligned} \left(\sum_{m=1}^n \overline{\mu}^m\right)^{-1} &= \frac{\sum_{l=1}^n \eta^l \mu_L^l + \sum_{l=n+1}^N \eta^l \mu_L^l}{\sum_{m=1}^n \eta^m \mu_L^m} \\ &= 1 + \frac{\sum_{l=n+1}^N \eta^l \mu_L^l}{\sum_{m=1}^n \eta^m \mu_L^m} \\ &\leq 1 + \frac{\sum_{l=n+1}^N \mu_L^l}{\sum_{m=1}^n \mu_L^m} = \left(\sum_{m=1}^n \mu_L^m\right)^{-1} \quad . \end{aligned}$$

The inequality comes from the sorting of η, and the last equality from $\sum \mu_L = 1$. ∎

Proposition 4 establishes a qualitative feature of this model: if the price is kept constant over time, then, in the model with inertia, the stationary distribution of the population *majorizes* the one obtained in the corresponding logit-model without inertia. Recalling that the majorization order expresses a form of dispersion, this means that inertia increases the concentration of the population on its favorite offers.

Lemma 4: Let us consider a and b in Δ_d. If $a \succ b$, then for all i, $a_i \leq d b_i$.

Proof: $a_i^\downarrow \leq \sum_{j=i}^d a_j^\downarrow = 1 - \sum_{j=1}^{i-1} a_j^\downarrow \leq 1 - \sum_{j=1}^{i-1} b_j^\downarrow = \sum_{j=i}^d b_j^\downarrow \leq (d - i + 1) b_i^\downarrow \leq d b_i^\downarrow$. ∎

Proposition 5 (Boundedness of the steady-state gain): Even with $\mathcal{A} = \mathbb{R}^{N-1}$, the optimal steady-state gain $\overline{g}$ is bounded independently of γ.

Proof: Suppose that the optimal steady-state gain is

attained for an action a, then

$$\begin{aligned}
\overline{g} &= \sum_{k\in[K]} \rho_k \sum_{n\in[N]} (E^{kn}a^n - C^{kn})\overline{\mu}^{kn}(a) \\
&\leq \max_{k,n}(R^{kn} - C^{kn}) + \sum_{k\in[K]} \rho_k \sum_{n\in[N]} (E^{kn}a^n - R^{kn})\overline{\mu}^{kn}(a) \\
&\leq \max_{k,n}(R^{kn} - C^{kn}) + \sum_{\substack{k\in[K] \\ U^{kn}(a)<0}} \rho_k \left\langle -U^k(a), \overline{\mu}^k(a)\right\rangle_N \\
&\leq \max_{k,n}(R^{kn} - C^{kn}) + N \sum_{\substack{k\in[K] \\ U^{kn}(a)<0}} \rho_k \left\langle -U^k(a), \mu_L^k(a)\right\rangle_N \\
&\leq \max_{k,n}(R^{kn} - C^{kn}) + \frac{N}{\beta e} \ .
\end{aligned}$$

The third inequality comes from Lemma 4. For the fourth one, since the logit expression contains a no-purchase option, $\mu_L^{kn} \leq \frac{1}{1+e^{-\beta U^{kn}(a)}}$. To conclude, it remains to see that $1 + e^{\beta z} \geq (\beta e)z$ for all z. ∎

Proposition 5 proves that the optimal steady-state gain cannot diverge to infinity when the inertia grows. This qualitative result is no longer true for the optimal strategy (which may be a periodic sequence of actions instead of a single constant one), see Section VII-A.

VI. Numerical results

The numerical results were obtained on a laptop i7-1065G7 CPU@1.30GHz. We solved the problem up to dimension 4 (2 provider offers, 2 clusters) with high precision ($\delta_\mu = 50$ points for each dimension, 1.6 million discretization points, precision $\epsilon = 10^{-5}$) in 7 hours for RVI algorithm and in 70 seconds for the Howard algorithm adapted to decomposable state-spaces (both methods parallelized on 8 threads), see Table I. The Policy Iteration algorithm adapted to decomposable state-spaces (Algorithm 2) induces drastic computational time reductions with respect to the value-iteration algorithm and considerable memory gains in comparison with standard policy iteration procedure.

Instance	(node, arcs)	RVI	PI [Coc+98]	This work
$K = 1, N = 1$ $\delta_\mu = 1/2000$	(2e3, 2.5e6)	70s 0.8Mo	1s 30Mo	0.2s 9Mo
$K = 2, N = 2$ $\delta_\mu = 1/50$	(7.4e5, 6.9e8)	7h 15Mo	390s 13Go	70s 103Mo

TABLE I: Comparison RVI / Howard

We provide running times that include the graph building step (which is a very costly operation for high dimensional graph in the standard PI algorithm. Each method ran on 8 threads.

In order to visualize qualitative results, we focus on the minimal non-trivial example (1 offer and 1 cluster). Note that the conclusions we draw from this example remain valid for the case 2 offers / 2 clusters. We use data of realistic orders of magnitude: we consider a population that checks monthly the market offers and consumes $E = 500$kWh each month. The provider competes with a regulated offer of 0.17€/kWh (inducing a reservation price of 85€), and has a cost of 0.13€/kWh. We suppose that the prices are freely chosen by the provider in the range 0.08-0.22€/kWh. The intensity parameter β is fixed to 0.1.

Numerical experiments in Fig. 3-2 emphasize the role of the switching cost. There exists a threshold – around $\gamma = 22$ in Fig. 3 – above which the steady-state policy become dominated by a cyclic strategy, where a period of promotion is periodically applied to recover a sufficient market share (period of 7 time steps on this example, see Fig. 2b and Fig. 2d). Below this threshold, the optimal policy has an attractor point which is exactly the best steady-state price, see Fig. 2c. The finite horizon policy is therefore a "turnpike" like strategy [Dam+14]: we rapidly converge to the steady-state and diverge at the end of the horizon, see Fig. 2a. Fig. 3 highlights that the adding of a convex function φ strengthens the upper bound, so that the optimality of the steady-state strategy is guaranteed up to γ around 19.

VII. Study of the minimal non-trivial model

Let us study the simple (yet non-trivial) case where the company has 1 contract ($N = 1$) and the population is homogeneous ($K = 1$). Numerical results have been shown in previous section.

In this setting, the probability μ to choose the retailer contract lies in the segment $[0, 1]$. For the finite-horizon setting, the toy model is therefore defined as

$$\max_{a_1,\dots,a_T\in\mathcal{A}^T} \left\{ \begin{array}{l} \displaystyle\sum_{t=1}^{T}(a_t - C)\mu_t \\ \text{s. t. } \begin{bmatrix}\mu_t & 1\text{-}\mu_t\end{bmatrix} = \begin{bmatrix}\mu_{t-1} & 1\text{-}\mu_{t-1}\end{bmatrix} P(a_t) \end{array} \right\} \tag{26}$$

with

$$P(a_t) = \begin{bmatrix} \frac{e^{\beta\gamma}e^{-\beta(a_t-R)}}{1+e^{\beta\gamma}e^{-\beta(a_t-R)}} & \frac{1}{1+e^{\beta\gamma}e^{-\beta(a_t-R)}} \\ \frac{e^{-\beta(a_t-R)}}{e^{\beta\gamma}+e^{-\beta(a_t-R)}} & \frac{e^{\beta\gamma}}{e^{\beta\gamma}+e^{-\beta(a_t-R)}} \end{bmatrix}$$

In the sequel, the data is $C = 2$, $R = 3$, $\beta = 3$, $T = 45$.

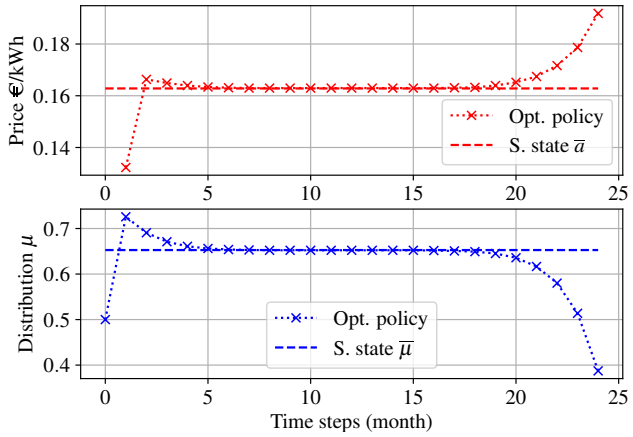
A. Cycling strategies

Figure 2 suggests a threshold (in terms of switching cost intensity) that separates the decision behavior into two different regimes : the convergence to a steady state for low switching costs intensity and the convergence to periodic strategies above the threshold (see Figure 2b). Therefore, in order to better understand this cycling behavior, we define the set of periodic strategies in the one-dimensional case as follows:

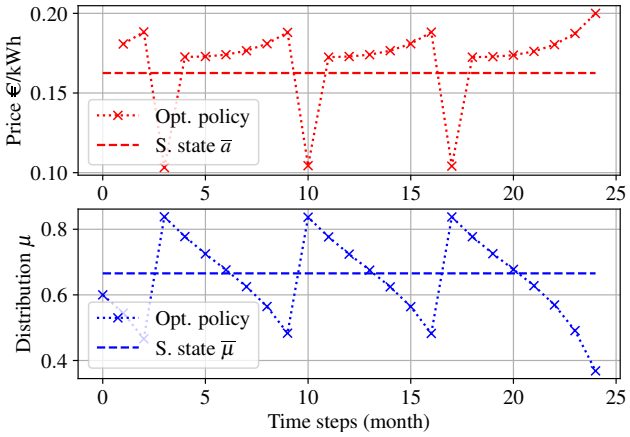
Definition 4: A *τ-cycle* is a cycling strategy of τ time steps, defined by the customer response $(\mu_0, \dots, \mu_{\tau-1}, \mu_\tau)$, with the cycling condition $\mu_0 = \mu_\tau$. For any τ-cycle l, we denote by

(i) $\overline{\mu}[l] = \frac{1}{\tau}\sum_{t=1}^{\tau}\mu_t$ and $V[l] = \frac{1}{\tau}\sum_{t=1}^{\tau}(\mu_t - \overline{\mu})^2$ the mean and the variance of the customer distribution over the cycle,

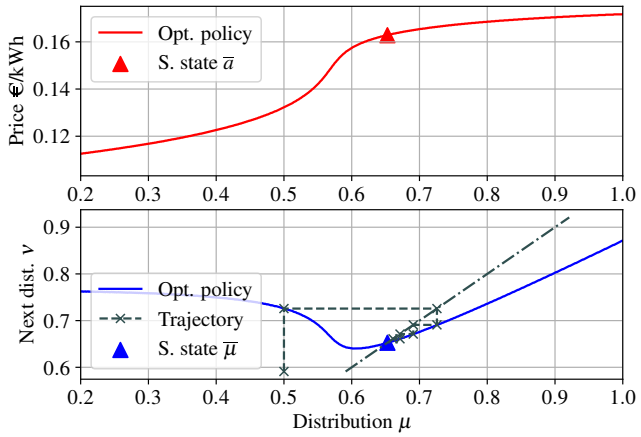
(ii) $g[l] = \frac{1}{\tau}\sum_{t=1}^{\tau}(a_t - C)\mu_t$ the gain (mean profit over the cycle).



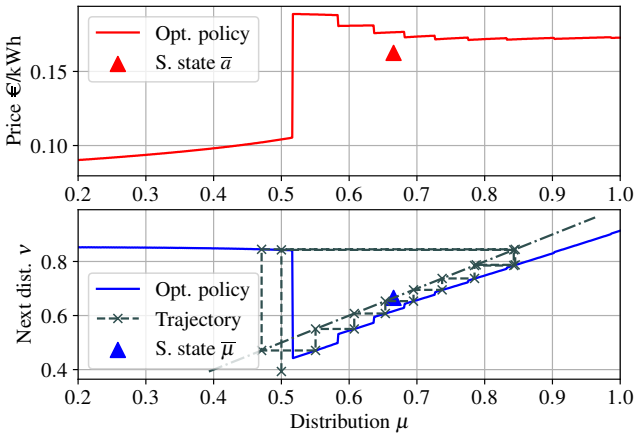
(a) Optimal finite horizon trajectory (provider action and customer distribution) for *low* switching cost.



(b) Optimal finite horizon trajectory (provider action and customer distribution) for *high* switching cost.



(c) Optimal decision for the long-run average reward (provider action and next customer distribution) for *low* switching cost. Graphical iteration is drawn in dotted lines.



(d) Optimal decision for the long-run average reward (provider action and next customer distribution) for *high* switching cost. Graphical iteration is drawn in dotted lines.

Fig. 2: Numerical results for both the finite horizon and long-term average reward criteria. *Low* (resp. *high*) switching cost stands for $\gamma = 20$ (resp. 25).

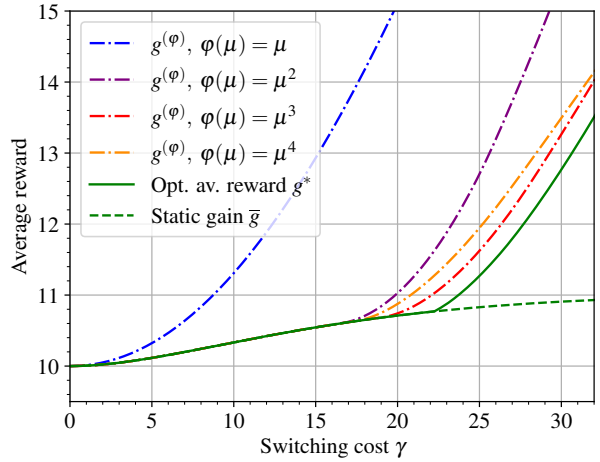


Fig. 3: Optimal gain g^* for a range of switching costs, along with lower bound $\overline{g}$ and upper bounds $g^{(\varphi)}$, $\varphi(\cdot) = (\cdot)^{1,2,3,4}$.

Proposition 6: Let $\gamma > 0$, knowing μ_{t-1} and μ_t in $[0, 1]$, there exists a unique a_t verifying the constraint in (26), defined as

$$\hat{a}_t := e^{-\beta(a_t - R)} = \frac{2\mu_t - \kappa_t + \sqrt{(2\mu_t - \kappa_t)^2 + 4\hat{\gamma}^2\mu_t(1-\mu_t)}}{2\hat{\gamma}(1-\mu_t)} \tag{27}$$

where $\hat{\gamma} = e^{\beta\gamma}$ and $\kappa_t = 1 + (\hat{\gamma}^2 - 1)(\mu_{t-1} - \mu_t)$.

Proof: From (26), one obtains the following equation:

$$\mu_t = \left[\frac{\hat{\gamma}\hat{a}_t}{1 + \hat{\gamma}\hat{a}_t} - \frac{\hat{a}_t}{\hat{\gamma} + \hat{a}_t}\right]\mu_{t-1} + \frac{\hat{a}_t}{\hat{\gamma} + \hat{a}_t} \ ,$$

that can be equivalently written as a second-order equation: $0 = \hat{a}_t^2\left[\hat{\gamma}(\mu_t - 1)\right] + \hat{a}_t\left[2\mu_t - \kappa_t\right] + \left[\hat{\gamma}\mu_t\right]$ of discriminant $\Delta = (2\mu_t - \kappa_t)^2 + 4\hat{\gamma}^2\mu_t(1-\mu_t) \geq 0$. ■

Corollary 1: As a special case of Proposition 6,

(i) if $\gamma = 0$, $\hat{a}_t = \frac{\mu_t}{1-\mu_t}$,

(ii) the steady-state policy that converges to $\mu \in]0, 1[$ is obtained by fixing the price to

$$\hat{a} = \frac{2\mu - 1 + \sqrt{(2\mu - 1)^2 + 4\hat{\gamma}^2\mu(1-\mu)}}{2\hat{\gamma}(1-\mu)} \ . \tag{28}$$

Proof: Items (i) and (ii) are obtained with $\kappa_t = 1$, either with $\hat{\gamma} = 1$ or $\mu_{t-1} = \mu_t$. ■

Proposition 6 gives an explicit expression of the (unique) action that allows a transition between state μ_{t-1} and μ_t. The uniqueness can be extended to transitions $\mu_t = \mu_{t-1}P(a)$ in higher dimension, but the explicit characterization of the action is not straightforward. We now want to compare the gain over a τ-cycle l and the steady-state gain $\overline{g}$. A first result is readily obtained in absence of switching costs, i.e., $\gamma = 0$, showing that constant-price policies are in this case optimal:

Proposition 7 (Gain without switching cost): Suppose that $\gamma = 0$, then, the optimal steady-state policy induces a gain greater than the one achieved by any τ-cycle l of at least $\frac{V[l]}{\beta}$, i.e.,

$$g[l] \leq \overline{g} - \frac{V[l]}{\beta} \ .$$

As a consequence, the optimal cycle corresponds to a constant-price policy.

Proof: Using Corollary 1, $a_t = R - \frac{1}{\beta}\log\left(\frac{\mu_t}{1-\mu_t}\right)$, and the mean profit of a τ-cycle l is

$$g[l] = (R-C)\overline{\mu}[l] - \frac{1}{\beta\tau}\sum_{t=1}^{\tau}\mu_t \log\left(\frac{\mu_t}{1-\mu_t}\right) \ .$$

The function $\mu \mapsto \mu\log\left(\frac{\mu}{1-\mu}\right)$ is strongly convex of modulus 1. Therefore, using Jensen's inequality for strongly convex function, see e.g. [MN10], we obtain that

$$g[l] \leq (R-C)\overline{\mu}[l] - \frac{1}{\beta}\overline{\mu}[l]\log\left(\frac{\overline{\mu}[l]}{1-\overline{\mu}[l]}\right) - \frac{V[l]}{\beta} \leq \overline{g} - \frac{V[l]}{\beta} \ .$$

■

Let us specialize the τ-cycles to a particular sub-class:

Definition 5: A *(s,S,τ)-cycle* is a specific τ-cycle, in which $\mu_t = S + \frac{s-S}{\tau}t, t \leq \tau$.

Proposition 8: Let us consider a (s,S,τ)-cycle l. Then,

$$g[l] \geq \frac{s(\tau-1)-S}{\tau}\gamma + O(1) \text{ as } \gamma \to \infty \ .$$

As a consequence, there exists a threshold $\Gamma > 0$ such that for any $\gamma \geq \Gamma$, the optimal steady-state policy is dominated by a (s,S,τ)-cycle.

Proof: Recalling that $\sqrt{a^2+b} \leq |a| + \frac{b}{2|a|}$, we have $\sqrt{(2\mu-\kappa)^2 + 4\hat{\gamma}^2\mu(1-\mu)} \leq |2\mu-\kappa| + \frac{\hat{\gamma}^2\mu(1-\mu)}{|2\mu-\kappa|}$. We first look at a period $1 \leq t < \tau$ where $\mu_{t-1} - \mu_t = \frac{S-s}{\tau}$. As we suppose that $\gamma \to \infty$, $\hat{\gamma} \geq \sqrt{1+\frac{\tau}{S-s}}$ and so $\kappa \geq 2\mu$. Therefore,

$$\hat{a} \leq \frac{\hat{\gamma}\mu}{1 + (\hat{\gamma}^2-1)\frac{S-s}{\tau} - 2\mu}$$

and

$$a \geq R + \frac{1}{\beta}\log\left(\frac{(\hat{\gamma}^2-1)(S-s)-\tau}{\hat{\gamma}\tau}\right)$$
$$\simeq \frac{1}{\beta}\log(\hat{\gamma}) + O(1) = \gamma + O(1) \ .$$

If now we look at the last period $t = \tau$. Then, as we suppose that $\gamma \to \infty$, $\hat{\gamma} \geq \sqrt{1+\frac{1}{S-s}}$. Therefore,

$$\hat{a} \leq \frac{1 + (\hat{\gamma}^2-1)(S-s)}{\hat{\gamma}(1-y)} + \frac{\hat{\gamma}}{(\hat{\gamma}^2-1)(S-s)-1}$$

and

$$a \geq R - \frac{1}{\beta}\log\left(\frac{1+(\hat{\gamma}^2\text{-}1)(S-s)}{\hat{\gamma}(1-y)} + \frac{\hat{\gamma}}{(\hat{\gamma}^2\text{-}1)(S-s)-1}\right)$$
$$\simeq -\gamma + O(1) \ .$$

The mean profit is finally bounded by below : $g[l] \geq \frac{1}{\tau}\left[(\tau-1)s - S\right]\gamma + O(1)$. To conclude, any (s,S,τ)-cycle satisfying $\tau \geq 1 + \frac{S}{s}$ induces a mean profit that diverges with respect to γ. In the meantime, the steady-state optimum is bounded, see Proposition 5, and so dominated for sufficiently large switching cost γ.

■

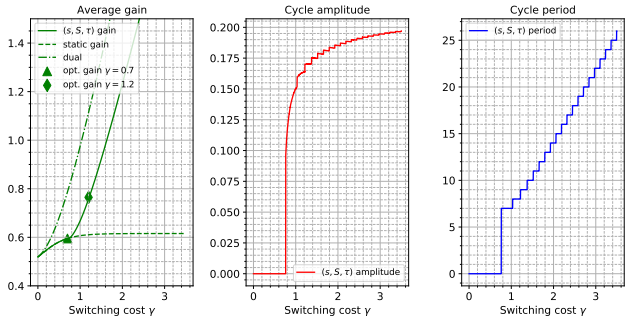


Fig. 4: Evolution of the optimal (s,S,τ)-cycle for a range of switching cost values

The left (resp. middle, right) panel shows the gain (resp. cycle amplitude, cycle period) of the optimal (s,S,τ)-cycle. The steady-state gain $\overline{g}$ is displayed for comparison, as well as the optimal gain obtained in Figure 2. A kink appears at $\gamma \simeq 0.762$, indicating the separation of the cycling behavior from the steady-state behavior.

In Figure 4, we compute the optimal (s,S,τ)-cycle, by iterating over the possible values of s, S, and τ for each given value of γ. Before the kink, the optimal cycle is in reality the constant-price strategy (cycle of amplitude 0), and after this point, there exists cycle of positive amplitude that outperforms the steady-state strategy. The results found in Figure 2 for a broader class of cycles are consistent, and the (s,S,τ)-cycles are good approximations of the optimal policy.

VIII. Conclusion

We developed an ergodic control model to represent the evolution of a large population of customers, able to actualize their choices at any time. Using qualitative properties of the population dynamics (contraction in Hilbert's projective metric), we showed the existence of a solution to the ergodic eigenproblem (in the presence of noise or in the deterministic setting), which we applied to a problem of electricity pricing. A numerical study reveals the existence of optimal cyclic

promotion mechanisms, that have already been observed in economics, and we proved this behavior on a toy example. We also quantified the sub-optimality of constant-price strategy in terms of a specific duality gap.

The present model has connections with partially observable MDPs, in which the state space is also a simplex. We plan to explore such connections in future work. We also aim at analyzing the problem through the weak KAM angle. In particular, we could expect to obtain a turnpike-like property when the Aubry set (to which the dynamics converge under any optimal policy) is reduced to a singleton. Finally, the approximation error induced by the discretization should be explored. In particular, exploiting the contraction of the dynamics may allow us to obtain convergence ratios, using similar arguments as in [BBK23].

REFERENCES

[AGN11] M. Akian, S. Gaubert, and R. Nussbaum. "A Collatz-Wielandt characterization of the spectral radius of order-preserving homogeneous maps on cones". 2011. eprint: 1112.5968.

[AGN15] M. Akian, S. Gaubert, and R. Nussbaum. "Uniqueness of the fixed point of nonexpansive semidifferentiable maps". In: *Transactions of the American Mathematical Society* 368.2 (Feb. 2015), pp. 1271–1320. DOI: 10.1090/s0002-9947-2015-06413-7.

[AGW09] M. Akian, S. Gaubert, and C. Walsh. "The max-plus Martin boundary". In: *Doc. Math* 14 (2009), pp. 195–240.

[AR12] W. J. Allender and T. J. Richards. "Brand Loyalty and Price Promotion Strategies: An Empirical Analysis". In: *Journal of Retailing* 88.3 (Sept. 2012), pp. 323–342. DOI: 10.1016/j.jretai.2012.01.001.

[Ara+93] A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh, and S. I. Marcus. "Discrete-Time Controlled Markov Processes with Average Cost Criterion: A Survey". In: *SIAM Journal on Control and Optimization* 31.2 (Mar. 1993), pp. 282–344. DOI: 10.1137/0331018.

[BCG19] V. Bansaye, B. Cloez, and P. Gabriel. "Ergodic Behavior of Non-conservative Semigroups via Generalized Doeblin's Conditions". In: *Acta Applicandae Mathematicae* 166.1 (Apr. 2019), pp. 29–72. DOI: 10.1007/s10440-019-00253-5.

[Bäu23] N. Bäuerle. "Mean Field Markov Decision Processes". In: *Applied Mathematics and Optimization* 88.1 (Apr. 2023). DOI: 10.1007/s00245-023-09985-1.

[BR11] N. Bäuerle and U. Rieder. *Markov Decision Processes with Applications to Finance*. Springer Berlin Heidelberg, 2011. DOI: 10.1007/978-3-642-18324-9.

[BBK23] E. Bayraktar, N. Bauerle, and A. D. Kara. *Finite Approximations for Mean Field Type Multi-Agent Control and Their Near Optimality*. 2023. arXiv: 2211.09633 [math.OC].

[BFY13] A. Bensoussan, J. Frehse, and P. Yam. *Mean Field Games and Mean Field Type Control Theory*. Springer New York, 2013. DOI: 10.1007/978-1-4614-8508-7.

[BP94] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Society for Industrial and Applied Mathematics, Jan. 1994. DOI: 10.1137/1.9781611971262.

[Ber98] D. P. Bertsekas. "A New Value Iteration method for the Average Cost Dynamic Programming Problem". In: *SIAM Journal on Control and Optimization* 36.2 (Mar. 1998), pp. 742–759. DOI: 10.1137/s0363012995291609.

[BHP99] T. Bielecki, D. Hernández-Hernández, and S. R. Pliska. "Risk sensitive control of finite state Markov chains in discrete time, with applications to portfolio management". In: *Mathematical Methods of Operations Research (ZOR)* 50.2 (Oct. 1999), pp. 167–188. DOI: 10.1007/s001860050094.

[Bis15] A. Biswas. *Mean Field Games with Ergodic cost for Discrete Time Markov Processes*. 2015. DOI: 10.48550/ARXIV.1510.08968. arXiv: 1510.08968 [math.OC].

[Cab09] L. Cabral. "Small switching costs lead to lower prices". In: *Journal of Marketing Research* 46.4 (2009), pp. 449–451.

[CS17] D. Calderone and S. Shankar. "Infinite-horizon average-cost Markov decision process routing games". In: *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, Oct. 2017. DOI: 10.1109/itsc.2017.8317849.

[CGG14] V. Calvez, P. Gabriel, and S. Gaubert. "Non-linear eigenvalue problems arising from growth maximization of positive linear dynamical systems". In: *Proceedings of the 53rd IEEE Annual Conference on Decision and Control (CDC), Los Angeles*. 2014, pp. 1600–1607.

[CLT21] R. Carmona, M. Laurière, and Z. Tan. *Model-Free Mean-Field Reinforcement Learning: Mean-Field MDP and Mean-Field Q-Learning*. 2021. arXiv: 1910.12802 [math.OC].

[Coc+98] J. Cochet-Terrasson, G. Cohen, S. Gaubert, M. McGettrick, and J.-P. Quadrat. "Numerical Computation of Spectral Elements in Max-Plus Algebra". In: *IFAC Proceedings Volumes* 31.18 (July 1998), pp. 667–674.

[Dam+14] T. Damm, L. Grüne, M. Stieler, and K. Worthmann. "An Exponential Turnpike Theorem for Dissipative Discrete Time Optimal Control Problems". In: *SIAM Journal on Control and Optimization* 52.3 (Jan. 2014), pp. 1935–1957. DOI: 10.1137/120888934.

[DF68] E. V. Denardo and B. L. Fox. "Multichain Markov Renewal Programs". In: *SIAM Journal on Applied Mathematics* 16.3 (May 1968), pp. 468–487. DOI: 10.1137/0116038.

[DW19] L. Dressler and S. Weiergraber. "Alert the inert! switching costs and limited awareness in retail electricity markets". 2019.

[DHR09] J.-P. Dubé, G. J. Hitsch, and P. E. Rossi. "Do Switching Costs Make Markets Less Competitive?" In: *Journal of Marketing Research* 46.4 (Aug. 2009), pp. 435–445. DOI: 10.1509/jmkr.46.4.435.

[Fat22] A. Fathi. "The weak-KAM theorem in Lagrangian dynamics". Book to appear. 2022.

[FST78] A. Federgruen, P. Schweitzer, and H. Tijms. "Contraction mappings underlying undiscounted Markov decision problems". In: *Journal of Mathematical Analysis and Applications* 65.3 (Oct. 1978), pp. 711–730. DOI: 10.1016/0022-247x(78)90174-9.

[Fes18] A. Festa. "Domain decomposition based parallel Howard's algorithm". In: *Mathematics and Computers in Simulation* 147 (May 2018), pp. 121–139. DOI: 10.1016/j.matcom.2017.04.008.

[Fly79] J. Flynn. "Steady State Policies for Deterministic Dynamic Programs". In: *SIAM Journal on Applied Mathematics* 37.1 (Aug. 1979), pp. 128–147. DOI: 10.1137/0137009.

[GT11] E. Garibaldi and P. Thieullen. “Minimizing orbits in the discrete Aubry–Mather model”. In: *Nonlinearity* 24 (2011), pp. 563–611.

[GG10] N. Gast and B. Gaujal. “A mean field approach for optimization in discrete time”. In: *Discrete Event Dynamic Systems* 21.1 (Oct. 2010), pp. 63–101. DOI: 10.1007/s10626-010-0094-3.

[Gau96] S. Gaubert. “On the burnside problem for semigroups of matrices in the (max, +) algebra”. In: *Semigroup Forum* 52.1 (Dec. 1996), pp. 271–292. DOI: 10.1007/bf02574104.

[GQ14] S. Gaubert and Z. Qu. “Dobrushin’s Ergodicity Coefficient for Markov Operators on Cones”. In: *Integral Equations and Operator Theory* 81.1 (Nov. 2014), pp. 127–150. DOI: 10.1007/s00020-014-2193-2.

[GS20] S. Gaubert and N. Stott. “A convergent hierarchy of non-linear eigenproblems to compute the joint spectral radius of nonnegative matrices”. In: *Mathematical Control & Related Fields* 10.3 (2020), pp. 573–590. DOI: 10.3934/mcrf.2020011.

[Hau00] M. Hauskrecht. “Value-Function Approximations for Partially Observable Markov Decision Processes”. In: *Journal of Artificial Intelligence Research* 13 (Aug. 2000), pp. 33–94. DOI: 10.1613/jair.678.

[HL96] O. Hernández-Lerma and J. B. Lasserre. *Discrete-Time Markov Control Processes*. Springer New York, 1996. DOI: 10.1007/978-1-4612-0729-0.

[HP10] D. Horsky and P. Pavlidis. “Brand Loyalty Induced Price Promotions: An Empirical Investigation”. In: *SSRN Electronic Journal* (2010). DOI: 10.2139/ssrn.1674765.

[HMP17] A. Hortaçsu, S. A. Madanizadeh, and S. L. Puller. “Power to Choose? An Analysis of Consumer Inertia in the Residential Electricity Market”. In: *American Economic Journal: Economic Policy* 9.4 (Nov. 2017), pp. 192–226. DOI: 10.1257/pol.20150235.

[Jac+22] Q. Jacquet, W. van Ackooij, C. Alasseur, and S. Gaubert. “Ergodic control of a heterogeneous population and application to electricity pricing”. In: *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, Dec. 2022. DOI: 10.1109/cdc51059.2022.9992336.

[KM97] V. N. Kolokoltsov and V. P. Maslov. *Idempotent analysis and its applications*. Vol. 401. Mathematics and its Applications. Dordrecht: Kluwer Academic Publishers Group, 1997, pp. xii+305. ISBN: 0-7923-4509-6.

[Kur89] M. Kurano. “The Existence of a Minimum Pair of State and Policy for Markov Decision Processes under the Hypothesis of Doeblin”. In: *SIAM Journal on Control and Optimization* 27.2 (Mar. 1989), pp. 296–307. DOI: 10.1137/0327016.

[Lee+21] W. Lee, S. Liu, H. Tembine, W. Li, and S. Osher. “Controlling Propagation of Epidemics via Mean-Field Control”. In: *SIAM Journal on Applied Mathematics* 81.1 (Jan. 2021), pp. 190–207. DOI: 10.1137/20m1342690.

[LN09] B. Lemmens and R. Nussbaum. *Nonlinear Perron–Frobenius Theory*. Cambridge University Press, 2009. DOI: 10.1017/cbo9781139026079.

[LPV87] P.-L. Lions, G. Papanicolaou, and S. Varadhan. “Homogenization of Hamilton-Jacobi equation”. Jan. 1987.

[Liv72] A. N. Livšic. “Cohomology of dynamical systems”. In: *Mathematics of the USSR-Izvestiya* 6.6 (Dec. 1972), pp. 1278–1301. DOI: 10.1070/im1972v006n06abeh001919.

[Lov91] W. S. Lovejoy. “Computationally Feasible Bounds for Partially Observed Markov Decision Processes”. In: *Operations Research* 39.1 (Feb. 1991), pp. 162–175. DOI: 10.1287/opre.39.1.162.

[MN02] J. Mallet-Paret and R. Nussbaum. “Eigenvalues for a Class of Homogeneous Cone Maps Arising from Max-Plus Operators”. In: *Discrete and Continuous Dynamical Systems* 8.3 (2002), pp. 519–562.

[MOA11] A. W. Marshall, I. Olkin, and B. C. Arnold. *Inequalities: Theory of Majorization and Its Applications*. Springer New York, 2011. DOI: 10.1007/978-0-387-68276-1.

[MN10] N. Merentes and K. Nikodem. “Remarks on strongly convex functions”. In: *Aequationes mathematicae* 80.1-2 (Sept. 2010), pp. 193–199. DOI: 10.1007/s00010-010-0043-0.

[MP22] M. Motte and H. Pham. “Mean-field Markov decision processes with common noise and open-loop controls”. In: *The Annals of Applied Probability* 32.2 (Apr. 2022). DOI: 10.1214/21-aap1713.

[NMS19] T. Ndebele, D. Marsh, and R. Scarpa. “Consumer switching in retail electricity markets: Is price all that matters?” In: *Energy Economics* 83 (Sept. 2019), pp. 88–103. DOI: 10.1016/j.eneco.2019.06.012.

[PE17] P. Pavlidis and P. B. Ellickson. “Implications of parent brand inertia for multiproduct pricing”. In: *Quantitative Marketing and Economics* 15.4 (July 2017), pp. 369–407. DOI: 10.1007/s11129-017-9187-8.

[Put94] M. L. Puterman. *Markov Decision Processes*. Wiley, Apr. 1994. DOI: 10.1002/9780470316887.

[Tra09] K. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2009.

[Whi63] D. White. “Dynamic programming, Markov chains, and the method of successive approximations”. In: *Journal of Mathematical Analysis and Applications* 6.3 (June 1963), pp. 373–376. DOI: 10.1016/0022-247x(63)90017-9.

[Wie19] P. Wiecek. “Discrete-Time Ergodic Mean-Field Games with Average Reward on Compact Spaces”. In: *Dynamic Games and Applications* 10.1 (Feb. 2019), pp. 222–256. DOI: 10.1007/s13235-019-00296-1.