IEEE CDC 2022 (Cancún)

Ergodic Control of a Heterogeneous Population and application to Electricity Pricing

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

December 7, 2022

In this talk

- Study of a *Mean-field MDP* for *heterogeneous* population
- Solutions via an *ergodic eigenproblem*
- Refined *Policy Iteration* Algorithm *à la* Howard and resolution of high-dimensional instances
- Application to *electricity pricing*:
  $\rightarrow$ Optimality of *periodic promotions* for important *switching costs*

Section 1

Definition of the model

MDP - Homogeneous population

A Markov Decision Process (MDP) is represented by a 4-tuple $\mathcal{M} = (\mathcal{X}, \mathcal{A}, P(a), \theta(a))$, where

- $\mathcal{X} = \{1, \dots, N\}$ is the *state* space,
- $\mathcal{A}$ is the *action* space,
- $P(a) \in \mathbb{R}^{N \times N}$ is the *transition matrix* associated with action $a \in \mathcal{A}$,
- $\theta(a) \in \mathbb{R}^{N}$ is the *instantaneous reward* to be in a given state due to action $a \in \mathcal{A}$.

(Bilevel) interpretation:

1. A *controller* chooses an action a,
2. An *agent* is influenced by this action:
   he moves from n to m with probability $P(a)_{n,m}$,
3. The *controller*'s reward is $\theta(a)_n$.

I-agent MDP - Homogeneous population

A I-agent Markov Decision Process (MDP) is represented by a 5-tuple $(\mathcal{X}, \mathcal{A}, P(a), \theta(a), I)$, where

- $\mathcal{X} = \{1, \dots, N\}$ is the *state* space,
- $\mathcal{A}$ is the *action* space,
- $P(a) \in \mathbb{R}^{N \times N}$ is the *transition matrix* associated with action $a \in \mathcal{A}$,
- $\theta(a) \in \mathbb{R}^{N}$ is the *instantaneous reward* to be in a given state due to action $a \in \mathcal{A}$.

(Bilevel) interpretation:

1. A *controller* chooses an action a,
2. Each *agent* $i \in [I]$ is influenced by this action:
   he moves from n_i to m_i with probability $P(a)_{n_i, m_i}$,
3. The *controller*'s reward is $\frac{1}{I} \sum_{i \in [I]} \theta(a)_{n_i}$.

I-agent MDP - Homogeneous population

A I-agent Markov Decision Process (MDP) is represented by a 5-tuple $(\mathcal{X}, \mathcal{A}, P(a), \theta(a), I)$, where

- $\mathcal{X} = \{1, \dots, N\}$ is the *state* space,
- $\mathcal{A}$ is the *action* space,
- $P(a) \in \mathbb{R}^{N \times N}$ is the *transition matrix* associated with action $a \in \mathcal{A}$,
- $\theta(a) \in \mathbb{R}^{N}$ is the *instantaneous reward* to be in a given state due to action $a \in \mathcal{A}$.

(Bilevel) interpretation:

1. A *controller* chooses an action a,
2. Each *agent* $i \in [I]$ is influenced by this action:
   he moves from n_i to m_i with probability $P(a)_{n_i, m_i}$,
3. The *controller*'s reward is $\frac{1}{I} \sum_{i \in [I]} \theta(a)_{n_i}$.

Remark: The I-agent MDP is equivalent to a standard MDP with

- state space: $\mathcal{X}^I$,
- transition matrix $Q(a) = \operatorname{diag}(P(a), \dots, P(a)) \in \mathbb{R}^{N^I \times N^I}$.

Lifted MDP - Homogeneous population

We define the *lifted MDP* associated with $\mathcal{M}$ as the *deterministic* MDP $(\mathcal{P}(\mathcal{X}), \mathcal{A}, T(a), r(a))$, where

- $\mathcal{P}(\mathcal{X}) = \Delta_N$ is the set of *probability measures* on $\mathcal{X}$,
- $T(a) := [\mu \in \Delta_N \mapsto \mu P(a)]$ is the *transition function* which gives the next state for action a,
- $r(a) := \left[\mu \in \Delta_N \mapsto \langle \theta(a), \mu \rangle_N\right]$ is the *expected* instantaneous reward according to a given measure due to action a.

Proposition (Mean-field MDP, see Motte and Pham, 2019)

For an infinite number of *indistinguishable* players ($I \to \infty$), the I-player MDP corresponds to the lifted MDP.

Lifted MDP - Homogeneous population

We define the *lifted MDP* associated with $\mathcal{M}$ as the *deterministic* MDP $(\mathcal{P}(\mathcal{X}), \mathcal{A}, T(a), r(a))$, where

- $\mathcal{P}(\mathcal{X}) = \Delta_N$ is the set of *probability measures* on $\mathcal{X}$,
- $T(a) := [\mu \in \Delta_N \mapsto \mu P(a)]$ is the *transition function* which gives the next state for action a,
- $r(a) := \left[\mu \in \Delta_N \mapsto \langle \theta(a), \mu \rangle_N\right]$ is the *expected* instantaneous reward according to a given measure due to action a.

Proposition (Mean-field MDP, see Motte and Pham, 2019)

For an infinite number of *indistinguishable* players ($I \to \infty$), the I-player MDP corresponds to the lifted MDP.

The matrix $P(a)$ is *no longer* the Markov kernel but *describes the dynamics* of the lifted MDP.

Model – Ergodic control on the lifted MDP

1. *Heterogeneous population*: each cluster $k \in [K]$ represents a proportion ρ_k of the overall pop.
2. *Distribution*: $\mu_t^k \in \Delta_N$ the distribution of the population of cluster k over $[N]$.
3. *Reward*:
$$r : (a_t, \mu_t) \mapsto \sum_{k \in [K]} \rho_k \left\langle \theta^k(a_t), \mu_t^k \right\rangle_N$$
4. *Transition*: $\mu_t^k = \mu_{t-1}^k P^k(a_t)$
5. *Controller's objective* (average long-term reward):
$$g^*(\mu_0) = \sup_{\pi \in \Pi} \liminf_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} r(\pi_t(\mu_t), \mu_t) \ . \qquad \text{(AvR)}$$

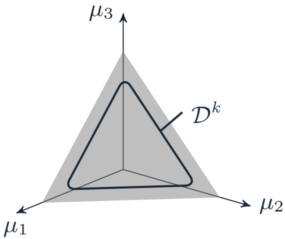
Model – Ergodic control on the lifted MDP

1. *Heterogeneous population*: each cluster $k \in [K]$ represents a proportion ρ_k of the overall pop.
2. *Distribution*: $\mu_t^k \in \Delta_N$ the distribution of the population of cluster k over $[N]$.
3. *Reward*:
$$r : (a_t, \mu_t) \mapsto \sum_{k \in [K]} \rho_k \left\langle \theta^k(a_t), \mu_t^k \right\rangle_N$$
4. *Transition*: $\mu_t^k = \mu_{t-1}^k P^k(a_t)$
5. *Controller's objective* (average long-term reward):
$$g^*(\mu_0) = \sup_{\pi \in \Pi} \liminf_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} r(\pi_t(\mu_t), \mu_t) \ . \tag{AvR}$$

Assumptions:

- $(A1)$ $a \mapsto P^k(a)$ is continuous,
- $(A2)$ There exists L such that for any sequence of actions $(a_1, \dots, a_L) \in \mathcal{A}^L$, $\prod_{i=1}^{L} P(a_i) \gg 0$,
- $(A2')$ For any action $a \in \mathcal{A}$, $P(a) \gg 0$,
- $(A3)$ $\exists M_r$ such that, $|\theta^{kn}(a)| \leq M_r$ for every $k \in [K]$, $n \in [N]$ and $a \in \mathcal{A}$.

Ergodic control



Let $\mathcal{D}^k := \text{vex}\left(\{\mu^k P_L^k(a) \mid a \in \mathcal{A}, \mu^k \in \Delta_N\}\right)$,
and $\mathcal{D} = \bigtimes_{k\in[K]} \mathcal{D}^k$.

Lemma

Let $(A1)$ – $(A2)$ hold. Then $\mathcal{D}^k \subseteq \text{relint } \Delta_N^K$.
Moreover, for $t \geq 1$, $\mu_t \in \mathcal{D}$ for any policy $\pi \in \Pi$.

For $v : \Delta_N^K \to \mathbb{R}$, the *Bellman operator* $\mathcal{B}$ is

$$\mathcal{B}\, v(\mu) = \max_{a\in\mathcal{A}}\{r(x,\mu) + v(\mu P(a))\} \ .$$

Theorem

Let $(A1)$ – $(A2)$ hold. Then, the *ergodic eigenproblem*

$$g\,\mathbb{1}_{\mathcal{D}} + h = \mathcal{B}\, h$$

admits a solution $g^* \in \mathbb{R}$ and h^* Lipschitz and convex on $\mathcal{D}$.
Moreover, g^* satisfies (AvR), and $a^*(\cdot) \in \arg\max \mathcal{B}\, h^*$ defines an *optimal policy*.

Deterministic MDP without controllability – the most degenerate case

	Time	Transitions	Assumption
Schweitzer, 1985	discrete	stochastic	unichain[1]
Biswas, 2015	discrete	stochastic	Doeblin / minorization[2]
Mallet-Paret and Nussbaum, 2002	discrete	deterministic	quasi-compactness
Fathi, 2010	continuous	deterministic	controlability[3]
Zavidovique, 2012	discrete	deterministic	controlability
Calvez et al., 2014	continuous	deterministic	contraction of the dynamics $(A2)$
This work	discrete	deterministic	contraction of the dynamics $(A2)$

Standard unichain/Doeblin type conditions entail that the eigenvector is *unique*, up to an additive constant, this is *no longer true* in our case.

[1]the Markov Chain induced by any deterministic stationary policy consists of a single recurrent class plus a –possibly empty– set of transient states (i.e., there exists a subset of states that are visited infinitely often with probability 1 independently of the starting state)

[2]for all state s, action a and measurable subset B of the state space, $P(B|x, a) \geq \epsilon\mu(B)$

[3]for every pair of states (s, s'), there exists an action a making s' accessible from s

Ergodic control – Sketch of the proof (existence)

We use a contraction argument directly on the dynamics (*not on* the Bellman Operator):
Let d_H be the Hilbert's projective metric defined as

$$d_H(u, v) = \max_{1 \le i,j \le n} \log \left(\frac{u_i}{v_i} \frac{v_j}{u_j} \right) \quad .$$

Under $(A1)$ – $(A2)$, $(\mathcal{D}, d_H)$ is a complete metric space.

Birkhoff theorem

Every matrix $Q \gg 0$ is a contraction in Hilbert's projective metric, i.e.,

$$\forall \mu, \nu \in (\mathbb{R}^N_{>0}), \; d_H(\mu Q, \nu Q) \le \kappa_Q d_H(\mu, \nu) \quad ,$$

where $\kappa_Q := \tanh\left(\mathrm{Diam}_H(Q) \,/\, 4\right) < 1$.

We then use the method of *vanishing discount approach* (Lions et al., 1987):

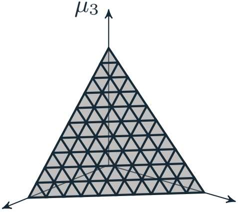
$\rightarrow$ the family of α-discounted objective function $\left(V_\alpha(\cdot)\right)_\alpha$ is *equi-Lipschitz*, which entails the existence of the eigenvector by a *compactness* argument.

Section 2

Algorithms

Relative Value Iteration with Krasnoselskii-Mann damping

- Regular grid Σ of the simplex Δ_N^K,
- Bellman Operator $\mathcal{B}^\Sigma$ using Freudenthal triangulation (Lovejoy, 1991).



Algorithm RVI with Mann-type iterates

Require: Σ, $\mathcal{B}^\Sigma$, $\hat{h}_0$

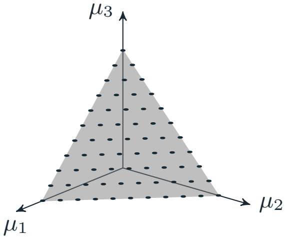
1: $v_{max} \leftarrow -\infty$
2: Initialize $\hat{h} = \hat{h}_0$, $\hat{h}'(\mu) = \mathcal{B}^\Sigma\, \hat{h}$
3: **while** $\mathrm{sp}(\hat{h}' - \hat{h}) > \epsilon$ **do**
4: $\quad \hat{h} \leftarrow (\,\hat{h}' - \max\{\hat{h}'\}e + \hat{h}\,)/2$
5: $\quad \hat{h}'(\hat{\mu}) \leftarrow (\mathcal{B}^\Sigma\, \hat{h})(\hat{\mu})$ for all $\hat{\mu} \in \Sigma$
6: **end while**
7: $\hat{g} \leftarrow (\,\max(\hat{h}' - \hat{h}) + \min(\hat{h}' - \hat{h})\,)/2$
8: **return** $\hat{g}, \hat{h}$

Proposition (Gaubert and Stott, 2020)

Convergence time of RVI = $O(\epsilon^{-2})$

Policy Iteration

- Regular grid Σ of the simplex Δ_N^K,
- Bellman Operator $\mathcal{B}^{\Sigma}$ using semi-lagrangian discretization.



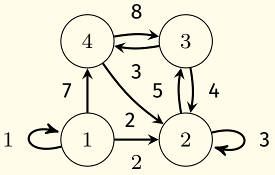
On-the-fly generation of transitions, refining (C.-Terrasson et al., 1998).

↪ solve the spectral problem
$$\max_{1 \leq j \leq n} (A_{ij} + x_j) = \lambda + x_i \ .$$

↪ the transition is *decomposed* on each segment

Example

$$A = \begin{bmatrix} 1 & 2 & \mathbb{0} & 7 \\ \mathbb{0} & 3 & 5 & \mathbb{0} \\ \mathbb{0} & 4 & \mathbb{0} & 3 \\ \mathbb{0} & 2 & 8 & \mathbb{0} \end{bmatrix}$$



Proposition

PI has *finite* time convergence

Numerical results

Instance[4]	(node, arcs)	RVI-KM	PI[5]	This work[6]
$K = 2, N = 2$ $\delta_\mu = 1/50$	$(7.4\ 10^5, 6.9\ 10^8)$	7h 15Mo	390s 13Go	70s 103Mo

[4] K: segments, N: contracts, δ_μ: discretization's precision (for each dimension)
[5] Cochet-Terrasson et al., 1998
[6] Each method ran on a 10 threads on a laptop i7-1065G7 CPU@1.30GHz.

Section 3

Application to electricity pricing

And if consumers *do not immediately* react ?

Intuition (Dubé et al., 2010; Horsky and Pavlidis, 2010)

"I switch to a new contract if there is a *sufficient* difference with my current offer."

Image from https://www.sketchbubble.com/en/presentation-switching-costs.html

Model

An electricity provider has N-1 different types of offers.
Given k and an offer $n \in [N\text{-}1]$, we know

- *Reservation price* R^{kn}: max. price that k want to spend on n,
- *Energy consumption* E^{kn}: fixed consumption if k chooses n,
- *Utility* $U^{kn}(a) := R^{kn} - E^{kn} a^n$, where a^n is the price for one unit of n.

Consumers have an alternative option (state of index N):
$\rightarrow$ fixed offer over time (regulated contract) with $U^{kN} = 0$.

The (linear) reward for the provider is then

$$\theta^{kn}(a) = \underbrace{E^{kn} a^n}_{\text{electricity invoice}} - \underbrace{C^{kn}}_{\text{cost}}, \; n < N, \quad \theta^{kN} = 0 \; .$$

Model

An electricity provider has N-1 different types of offers.
Given k and an offer $n \in [N\text{-}1]$, we know

- *Reservation price* R^{kn}: max. price that k want to spend on n,
- *Energy consumption* E^{kn}: fixed consumption if k chooses n,
- *Utility* $U^{kn}(a) := R^{kn} - E^{kn} a^n$, where a^n is the price for one unit of n.

Consumers have an alternative option (state of index N):
$\rightarrow$ fixed offer over time (regulated contract) with $U^{kN} = 0$.

The (linear) reward for the provider is then

$$\theta^{kn}(a) = \underbrace{E^{kn} a^n}_{\text{electricity invoice}} - \underbrace{C^{kn}}_{\text{cost}}, \; n < N, \quad \theta^{kN} = 0 \; .$$

Assumption: The transition probability follows a *logit response*, see e.g. Pavlidis and Ellickson, 2017:

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

- γ^{kn} is the cost for segment k to *switch* from contract n to another one,
- β is the intensity of the choice (it can represent a "*rationality* parameter").

Steady-states

Theorem

Given a constant *action* a*, the distribution sequence* $\left(\mu_t^k\right)_t$ *converges to* $\overline{\mu}^k(a)$*, defined as*

$$\overline{\mu}^{kn}(a) = \frac{\eta^{kn}(a)\mu_L^{kn}(a)}{\sum_{l\in[N]} \eta^{kl}(a)\mu_L^{kl}(a)} \ . \tag{1}$$

where $\eta^{kn}(a) := 1 + \left[e^{\beta\gamma^{kn}} - 1\right] \mu_L^{kn}(a)$*, and*

$$\mu_L^{kn} = e^{\beta\, U^{kn}(a)} \ / \ \sum_{l\in[N]} e^{\beta\, U^{kl}(a)} \ . \tag{2}$$

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

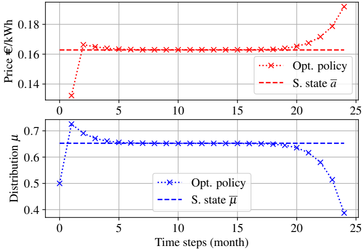
$$\overline{g} = \max_{a\in\mathcal{A}} r(a, \overline{\mu}(a)) \ . \tag{3}$$

Impact of switching costs γ on toy model

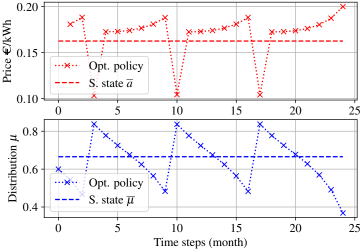
low γ → high γ

"Turnpike" like strategy:
Attraction to a steady-state

Cyclic strategy:
A promotion is periodically applied



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



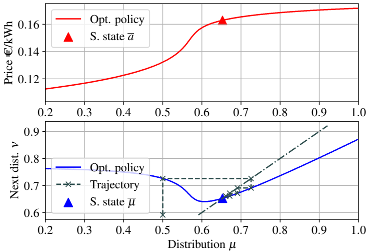
(b) Optimal finite horizon trajectory (provider action and customer distribution) for *high* switching cost.

$\hookrightarrow$ Confirms *optimality of periodic promotions*, already observed in Economics, see e.g. Horsky and Pavlidis, 2010.

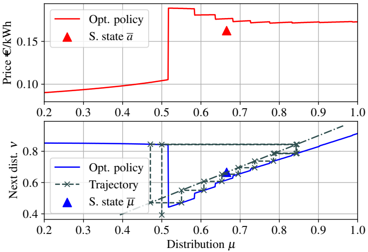
Impact of switching costs γ on toy model

<u>"Turnpike" like strategy</u>:
Attraction to a steady-state

<u>Cyclic strategy</u>:
A promotion is periodically applied



(a) Optimal decision for the long-run average reward (provider action and next customer distribution)



(b) Optimal decision for the long-run average reward (provider action and next customer distribution)

$\hookrightarrow$ Confirms *optimality of periodic promotions*, already observed in Economics, see e.g. Horsky and Pavlidis, 2010.

Conclusion and Perspectives

Conclusion

- Resolution of deterministic lifted MDP using a eigenproblem representation
- Refinement of Policy Iteration for Heterogeneous populations
- Application to electricity pricing, and highlight of the switching cost's impact

Perspectives

- Conditions for the convergence to a steady-state
- Links between dissipativity condition (control theory) and strict subsolutions (weak-KAM theory)
- Study of other transitions (non logit-based)

References

Schweitzer, P. J. (1985). On undiscounted markovian decision processes with compact action spaces. RAIRO - Operations Research - Recherche Opérationnelle, 19(1), 71–86.

Lions, P.-L., Papanicolaou, G., & Varadhan, S. (1987). Homogenization of hamilton-jacobi equation.

Lovejoy, W. S. (1991). Computationally feasible bounds for partially observed markov decision processes. Operations Research, 39(1), 162–175.

Cochet-Terrasson, J., Cohen, G., Gaubert, S., McGettrick, M., & Quadrat, J.-P. (1998). Numerical computation of spectral elements in max-plus algebra. IFAC Proceedings Volumes, 31(18), 667–674.

Mallet-Paret, J., & Nussbaum, R. (2002). Eigenvalues for a class of homogeneous cone maps arising from max-plus operators. Discrete and Continuous Dynamical Systems, 8(3), 519–562.

Dubé, J.-P., Hitsch, G. J., & Rossi, P. E. (2010). State dependence and alternative explanations for consumer inertia. The RAND Journal of Economics, 41(3), 417–445.

Fathi, A. (2010). The weak-KAM theorem in lagrangian dynamics [Book to appear].

References

Horsky, D., & Pavlidis, P. (2010). Brand loyalty induced price promotions: An empirical investigation. SSRN Electronic Journal.

Zavidovique, M. (2012). Strict sub-solutions and mañé potential in discrete weak KAM theory. Commentarii Mathematici Helvetici, 1–39.

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

Biswas, A. (2015). Mean field games with ergodic cost for discrete time markov processes.

Pavlidis, P., & Ellickson, P. B. (2017). Implications of parent brand inertia for multiproduct pricing. Quantitative Marketing and Economics, 15(4), 369–407.

Motte, M., & Pham, H. (2019). Mean-field markov decision processes with common noise and open-loop controls.

Gaubert, S., & Stott, N. (2020). A convergent hierarchy of non-linear eigenproblems to compute the joint spectral radius of nonnegative matrices. Mathematical Control & Related Fields, 10(3), 573–590.