Projet de thèse

Jeux à somme nulle avec paiement ergodique et information partielle

English title: Partially Observable Markov Decision Processes and zero-sum games with mean-payoff

Diplôme préparé : Thèse de doctorat de l'Institut Polytechnique de Paris (EDMH), mention Mathématiques Appliquées. Établissement opérateur d'inscription : École polytechnique.

Encadrement/Supervision: La thèse sera effectuée au sein de l'équipe "Tropical", commune au CMAP, École polytechnique, CNRS, Institut polytechnique de Paris et à l'Inria Saclay Île-de-France, et située: CMAP, Ecole polytechnique, Route de Saclay, 91128 Palaiseau Cedex, France.

Elle sera dirigée par Marianne Akian (DR INRIA), e-mail: marianne.akian@inria.fr,

Web page: `http://www.cmap.polytechnique.fr/~akian/`.

Plusieurs coencadrements seront possibles en particulier par Bruno Ziliotto (Toulouse School of Economics) et Guillaume Vigeral (Paris 1). Le doctorant bénéficiera aussi du contexte scientifique de l'ANR ZADyG (2026–2029), à l'interface des Mathématiques Appliquées et de l'Informatique, ainsi que de son soutien financier.

Type de financement : – Financement par l'ANR ZADyG via un contrat doctoral avec l'Inria Saclay Île-de-France.

Résumé : L'objet de cette thèse est d'étudier des processus de décision markoviens ou des jeux à somme nulle avec paiement ergodique et information partielle. Pour le cas à un joueur, le problème se ramène à un jeux en information parfaite sur l'espace des croyances (à un joueur), et si de plus le joueur est aveugle, on obtient un problème déterministe. Dans le cas à deux joueurs, les joueurs peuvent partager la même information, ou alors un joueur a toute l'information alors que l'autre non.

Dans toutes ces situations, il faudra d'abord étudier les jeux d'un point de vue théorique (existence de la valeur du jeu à l'équilibre, caractérisation de la valeur comme solution d'une équation de la programmation dynamique ergodique, existence de strategies optimales). Ensuite, il faudra les étudier d'un point de vue algorithmique, par exemple en étudiant les algorithmes d'itération sur les valeurs relative, ou d'itération sur les politiques, ou des méthodes "ponctuelles".

Abstract : The aim of this thesis is to study ergodic Partially Observable Markov Decision Processes or zero-sum games. In the one player case, this problem can be reduced to an ergodic one-player game with perfect information on the space of beliefs, and if the state is unobservable (that is the player is blind), the reduced problem is deterministic. In the two player case, the two players may share a common information, or one player has full information whereas the other not.

In all these situations, we shall first study the game from a theoretical point of view (existence of the value of the game, characterization of this value as the solution of an ergodic dynamic programming equation, existence of optimal strategies). Then, we shall study the game from the algorithmic point of view, for instance by studying the relative value iteration and policy iteration algorithms and also point based methods.

Context: Stochastic (zero-sum) repeated games were introduced by Shapley [22], where two players with conflicting objectives dynamically interact in a stochastic environment. Shapley originally considered games with infinite horizon *discounted payoff*. Then, Gillette [11] studied games in which each player optimizes a *mean payoff* (average reward per time unit).

These games depend on the information structure. In *turn-based* games, two players play sequentially, alternating moves, or choices of an action, being aware of the previous decision of the other player. In *concurrent games*, at each stage, the two players choose simultaneously one action, being unaware of the choice of the other player at the same stage. Turn-based games are equivalent to a subclass of concurrent games (in which in each state, one of the two players is a dummy). The existence of the value for concurrent stochastic mean-payoff games is a celebrated result of Mertens and Neyman [18].

The one-player case is part of stochastic control or multi-stage stochastic optimization, and was extensively studied after the development of dynamic programming by Bellman [8].

Another class of games consists of *entropy games*, introduced by Asarin, Cervelle, Degorre, Dima, Horn and Kozyakin as an interesting category of "matrix multiplication games" [6]. Entropy games capture a variety of applications, arising in risk sensitive control [13, 4], portfolio optimization [2], growth maximization and population dynamics [23, 20, 19, 27]. In [1], Akian, Gaubert, Grand-Clément and Guillaud showed that entropy games are actually special cases of stochastic mean-payoff games, in which action spaces are infinite sets (simplices), and payments are given by Kullback-Leibler divergences.

Usual algorithms to solve one-player or turn-based perfect information games include the (relative) value iteration and the policy iteration algorithms. Policy iteration algorithm goes back to Howard, Hoffman and Karp, and was later shown to be similar to the simplex algorithm for an associated linear program. Ye [26] showed that the policy iteration algorithm for one-player games with a fixed discount factor is strongly polynomial. This was also extended to two-player turn based zero-sum games [12].

However, turn-based games with mean-payoff and finite state and action spaces belong to the complexity class NP $\cap$ coNP [10, 28] but are not known to be polynomial-time solvable (see also [5]). Similarly, entropy games belong to the class NP $\cap$ coNP as shown in [6]. Concurrent games are hard to solve exactly: the value is an algebraic number whose degree may be exponential in the number of states [14].

In the undiscounted one-player case, White [25] introduced *relative value iteration*, which consist in fixed point iterations up to additive constants. This solves the dynamic programming equation under a primitivity assumption. In [3], this assumption was relaxed by combining the relative value iteration with Krasnoselkii–Mann damping [15, 17], which applies under a weaker ergodicity assumption. Moreover, a similar algorithm was obtained to solve entropy games.

In all the above variants of games, the players may have only partial observations of the state and of the actions of the other player. In the one player case, this leads to the so called Partially Observable Markov Decision Processes (POMDP), a particular case of which is the unobservable (blind) case. In the two player case, the two players may share a common observation or not.

When the state space is finite, an infinite horizon discounted POMDP can be reduced to the dynamic programming equation of a one-player game with perfect information on the space of beliefs [7, 24]. Then, value iterations can be applied after a discretization of the state space. However, such an algorithm suffers from curse of dimensionality. Therefore, the main approximation algorithm consists in point based methods [21]. The complexity of such algorithms depends on the "diameter" of the set of reachable beliefs [16].

In [9], unobservable Markov Decision Processes with mean-payoff are studied from the point of view of decidability. However, approximation of the value of mean-payoff one-player games or two player games with incomplete information does not seem to have been considered in the literature.

Proposed work : The aim of the thesis is to study theoretically and algorithmically ergodic Partially Observable Markov Decision Processes or zero-sum games. One will consider variants of this problem depending in particular on the information structure (unobservable (blind) games, perfect information from one player,...), and on the ergodicity properties of the transition probability matrices. One first goal is to give sufficient conditions for the existence of a value and strategies/policies. Then, the convergence and/or the complexity of relative value iteration and policy iteration algorithms will be studied.

Prerequisites : A Master in Applied mathematics, in particular in optimization or games. It would be better if a Master thesis is done on the same subject. Do not hesitate to contact us for this.

References

[1] M. Akian, S. Gaubert, J. Grand-Clément, and J. Guillaud. The operator approach to entropy games. *Theory of Computing Systems*, 63:1089–1130, 2019.

[2] M. Akian, A. Sulem, and M. I. Taksar. Dynamic optimization of long-term growth rate for a portfolio with transaction costs and logarithmic utility. *Mathematical Finance*, 11(2):153–188, April 2001.

[3] Marianne Akian, Stéphane Gaubert, Ulysse Naepels, and Basile Terver. Solving irreducible stochastic mean-payoff games and entropy games by relative krasnoselskii-mann iteration. volume 272. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023. See also arXiv:2305.02458.

[4] V. Anantharam and V. S. Borkar. A variational formula for risk-sensitive reward. *SIAM J. Contro Optim.*, 55(2):961–988, 2017.

[5] D. Andersson and P. B. Miltersen. The complexity of solving stochastic games on graphs. In *Proceedings of the 20th International Symposium on Algorithms and Computation (ISAAC)*, volume 5878 of *Lecture Notes in Comput. Sci.*, pages 112–121. Springer, 2009.

[6] E. Asarin, J. Cervelle, A. Degorre, C. Dima, F. Horn, and V. Kozyakin. Entropy games and matrix multiplication games. In *Proc. of the 33rd Int'l Symposium on Theoretical Aspects of Computer Science (STACS)*, volume 47 of *LIPIcs. Leibniz Int. Proc. Inform.*, pages 11:1–11:14. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2016.

[7] Karl Johan Åström. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205, 1965.

[8] R. Bellman. The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503–515, 1954.

[9] Krishnendu Chatterjee, David Lurie, Raimundo Saona, and Bruno Ziliotto. Ergodic unobservable mdps: Decidability of approximation, 2024. Preprint arXiv:2405.12583.

[10] A. Condon. The complexity of stochastic games. *Inform. and Comput.*, 96(2):203–224, 1992.

[11] D. Gillette. Stochastic games with zero stop probabilities. In M. Dresher, A. W. Tucker, and P. Wolfe, editors, *Contributions to the Theory of Games III*, volume 39 of *Ann. of Math. Stud.*, pages 179–188. Princeton University Press, Princeton, NJ, 1957.

[12] T. D. Hansen, P. B. Miltersen, and U. Zwick. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *J. ACM*, 60:1–16, 2013.

[13] R. A. Howard and J. E. Matheson. Risk-sensitive Markov decision processes. *Management Science*, 18(7):356–369, 1972.

[14] C. Ickstadt, Th. Theobald, and E. Tsigaridas. Semidefinite games. *International Journal of Game Theory*, pages 1–31, 2024.

[15] M. A. Krasnosel'skiĭ. Two remarks on the method of successive approximations. *Uspekhi Matematicheskikh Nauk*, 10:123–127, 1955.

[16] Hanna Kurniawati, David Hsu, and Wee Sun Lee. Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces. In *Robotics: Science and systems*, volume 2008. Zurich, Switzerland., 2008.

[17] W. R. Mann. Mean value methods in iteration. *Proceedings of the American Mathematical Society*, 4:506–510, 1953.

[18] J.-F. Mertens and A. Neyman. Stochastic games. *Internat. J. Game Theory*, 10(2):53–66, 1981.

[19] U. G. Rothblum. Multiplicative Markov decision chains. *Mathematics of Operations Research*, 9(1):6–24, 1984.

[20] U. G. Rothblum and P. Whittle. Growth optimality for branching Markov decision chains. *Mathematics of Operations Research*, 7(4):582–601, 1982.

[21] Guy Shani, Joelle Pineau, and Robert Kaplow. A survey of point-based POMDP solvers. *Autonomous Agents and Multi-Agent Systems*, 27(1):1–51, 2013.

[22] L. S. Shapley. Stochastic games. *Proc. Natl. Acad. Sci. USA*, 39(10):1095–1100, 1953.

[23] K. Sladký. *On dynamic programming recursions for multiplicative Markov decision chains*, pages 216–226. Springer Berlin Heidelberg, Berlin, Heidelberg, 1976.

[24] Edward J Sondik. The optimal control of partially observable markov processes over the infinite horizon: Discounted costs. *Operations research*, 26(2):282–304, 1978.

[25] D.J White. Dynamic programming, Markov chains, and the method of successive approximations. *Journal of Mathematical Analysis and Applications*, 6(3):373–376, 1963.

[26] Y. Ye. The simplex and policy-iteration methods are strongly polynomial for the Markov decision problem with a fixed discount rate. *Mathematics of Operations Research*, 36(4):593–603, November 2011.

[27] W. H. M. Zijm. Asymptotic expansions for dynamic programming recursions with general nonnegative matrices. *J. Optim. Theory Appl.*, 54(1):157–191, 1987.

[28] U. Zwick and M. Paterson. The complexity of mean payoff games on graphs. *Theoret. Comput. Sci.*, 158(1–2):343–359, 1996.