Projet de thèse

Algorithmes stochastiques pour les jeux à somme nulle avec paiement ergodique

English title: Stochastic algorithms for repeated 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/`.

Type de financement : – Concours Institut Polytechnique de Paris ou école membre.

Résumé : L'objet de cette thèse est de construire et d'étudier un algorithme stochastique pour la résolution de jeux à somme nulle avec paiement ergodique et espace d'états infini. On s'intéresse plus particulièrement aux jeux qui apparaissent dans l'étude des jeux en information partielle, ou celle de problèmes de croissance en dynamique des populations, ou celle de problèmes de jeux de multiplication matricielle. L'algorithme à construire devrait être une extension de la méthode numérique probabiliste max-plus (tropicale) introduite initiallement pour résoudre les équations d'Hamilton-Jacobi-Bellman. Pour étudier sa convergence et sa complexité, on étudiera d'abord les algorithmes déterministes de type itération sur les valeurs ou sur les politiques. On pourra aussi considérer, et comparer à, des extensions, au cas de jeux à 2 joueurs, des algorithmes concentrés en un point utilisés pour résoudre les POMDP, ou de l'apprentissage par renforcement. Pour ces travaux, on fera appel en particulier à des techniques de théorie de Perron-Frobenius non linéaire appliquées aux opérateurs de la programmation dynamique.

Abstract : The aim of this thesis is mainly to construct and study a stochastic algorithm allowing to solve zero-sum two-player games with an infinite state space. We shall consider particularly games arising when studying partial observation zero-sum games, or growth of population dynamics, or matrix multiplication games. The algorithm to be constructed should be an extension of the numerical max-plus probabilistic method developped initially for solving Hamilton-Jacobi-Bellman PDE. In order to study its convergence and complexity, one shall first study deterministic algorithms such as relative value iteration and policy iteration algorithms. One can also consider or compare with any extension to the two-player case of the point based methods used for solving POMDP, or of reinforcement learning. For this studies, one shall use in particular nonlinear Perron-Frobenius theory techniques applied to dynamic programming operators.

Context: Stochastic (zero-sum) repeated games were introduced by Shapley [33], where two players with conflicting objectives dynamically interact in a stochastic environment. Shapley originally considered games with infinite horizon *discounted payoff*. Then, Gillette [18] 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 [26].

The one-player case is part of Markov Decision Processes, stochastic control or multi-stage stochastic optimization, and was extensively studied after the development of dynamic programming by Bellman [13].

A remarkable subclass of stochastic mean-payoff games arises when imposing *ergodicity* or *irreducibility* conditions. Such conditions entail that the value of the game is independent of the initial state. In the finite state and action setting, this implies the existence of a solution to the nonlinear eigenproblem $T(u) = \lambda + u$, in which $u \in \mathbb{R}^n$ is a non-linear eigenvector, $\lambda \in \mathbb{R}$ is a non-linear eigenvalue, and T is the Shapley/dynamic programming operator of the game, which is a self-map of $\mathbb{R}^n$, where n is the number of states of the game. Then λ provides the constant value of the mean-payoff game. This equation is called the *ergodic equation*.

In the one-player case, White [38] introduced *relative value iteration*, which consist in fixed point iterations for the oprator T up to additive constants: $x_{k+1} = T(x_k) - \lambda_k$, $\lambda_k \in \mathbb{R}$. This solves the ergodic equation under a primitivity assumption. In [8], this assumption was relaxed by combining the relative value iteration with Krasnoselkii–Mann damping [21, 23]. The resulting algorithm solves turn-based and concurrent mean-payoff games and converges under general ergodicity conditions.

Another usual algorithm to solve one-player or turn-based perfect information games is the policy iteration algorithm, which goes back to Howard, Hoffman and Karp, and was later shown to be similar to the simplex algorithm for an associated linear program. Ye [39] 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 [19]. However, turn-based games with mean-payoff and finite state and action spaces belong to the complexity class NP $\cap$ coNP [15, 41] but are not known to be polynomial-time solvable (see also [10]).

The above results and algorithms concern problems with a finite state space. When the state space is infinite, for instance the finite dimensonal space $\mathbb{R}^d$, 1) the ergodic equation may not have a solution even under ergodic conditions; 2) if the ergodic equation has a solution, one may apply the above algorithms to a space discretization of the ergodic equation, but the resulting algorithm will suffer from the curse of dimensionality. For discounted or finite horizon one-player problems, several techniques have been introduced to bypass this curse of dimensionality, among them are

- Optimization only along an "optimal trajectory": Stochastic Dual Dynamic Programming for convex problems [27, 28]; Point based methods for POMDP (Partially Observable Markov Decision Processes) [32, 22]. See a comparison in [3].
- Max-plus or tropical numerical methods developped initially in the case of time continuous deterministic one-player problems [24, 29], and then in the stochastic case [25, 5], and in the discrete time context [17, 4]. Some of them are also stochastic.

Problems with an infinite state space arise for instance when considering Partially Observable Markov Decision Processes (POMDP). Indeed, when the state space is finite, a discounted POMDP can be solved by computing the solution of the dynamic programming equation of a one-player game with perfect information on the space of beliefs [12, 36]. Similarly, for mean-payoff problems, the existence of a solution to the ergodic dynamic programming equation of the MDP on the space of beliefs allows one to compute the value and the policies/strategies of the POMDP (see for instance [16]). One may also consider zero-sum games in which the players have only partial observations on the state, but share the same observations (signals). Then, their beliefs will coincide and the problem also reduces to a zero-sum game with mean-payoff on the space of beliefs.

A particular class of POMDP is obtained when the parameters of the MDP (that is the transition probabilities and the rewards) are not known, but the state can be observed. This is the framework of reinforment learning [37]. In this context, if the state process can be simulated, one can use stochastic value iteration algorithms like in [34, 14].

Problems with an infinite state space arise also when solving the general matrix multiplication games introduced by Asarin et al in [11] in which the mean-payoff is the growth of a product of matrices: $\limsup_{k\to\infty} \|A_1 \dots A_k\|^{1/k}$. When the matrices have nonnegative entries and the rows of matrices can be selected independently, this is the entropy games introduced also in [11]. Entropy games capture a variety of applications, arising in risk sensitive control [20, 9], portfolio optimization [2], growth maximization and population dynamics [35, 31, 30, 40]. In [1], it is shown 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. In the general case, matrix multiplication games can be solved using an infinite state space, such as the positive cone or the cone of positive definite matrices. This is what is done in [6, 7], while considering more general games involving nonlinear nonexpansive dynamics instead of matrices. Note however, that in this case, the existence of a solution to the ergodic equation only hold in some special situations. In [7], relative value iterations with a Krasnoselkii-Mann damping are used to compute the value in some particular situations.

Proposed work : Several questions arise in the study of mean-payoff games with infinite state space:

1. Is there a solution to the ergodic equation ? We already said that this is not always possible.
2. Can we characterize the value of the game with a weaker condition than the ergodic equation, such as the supremum of the subeigenvalues (the solutions of $\rho + u \leq T(u)$)
3. Is the value continuous in the parameters ? Counter example exist, see [7].
4. Is the value approximable ?
5. What is the complexity of this approximation ?
6. Construct a deterministic algorithm approximating the value, and study the convergence. By deterministic, we mean with a bound on the error valid for all instances.
7. Construct a stochastic algorithm approximating the value, and determine the relation between a bound on the error and the probability for such a bound to hold.

The first questions have been solved in some particular cases. This thesis will focus on the last question, assuming existence of a solution to the ergodic equation and continuity of the value, and keeping in mind the exemples of games with infinite state space already described above, that is the ones arising when studying partial observation zero-sum games, or growth of population dynamics, or matrix multiplication games.

For the stochastic algorithm, we think to consider any extension of the probabilistic max-plus methods used for solving Hamilton-Jacobi-Bellman Partial differential equations, see [5, 4]. In the probabilistic max-plus method, the value function is approximated by a supremum of quadratic forms and this approximation is computed inductively by combining sampling and regressions over the sets of quadratic forms approximating the application of the dynamic programming operator on the supremum of quadratic forms. Although the method may be compared to a specialized neural network approximation, the method does not use any nonlinear regression. There are several directions of extensions of the method to the two-player case : keeping the supremum representation of the value function or replace it by a mix between suprema and infima.

As a first step of the thesis, one can study the convergence and complexity of deterministic algorithms that have already been considered in the literature or are simple adaptations of classical algorithms: relative value iteration and policy iteration algorithms. One can also study how to adapt deterministic and stochastic point based methods from 1-player to 2-player games, and reinforcement learning techniques like Q-learning, and compare them with the extended probabilistic max-plus method.

From a theoretical point of view, one shall build on the works on competitive spectral radii [6, 7], and the techniques developped there.

Prerequisites : A Master in Applied mathematics, in particular in optimization or games or stochastic processes.

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, Jean-Philippe Chancelier, Luz Pascal, and Benoît Tran. Tropical numerical methods for solving stochastic control problems. In *MTNS 2022 - 25th International Symposium on Mathematical Theory of Networks and Systems*, Bayreuth (DE), Germany, September 2022. `https://inria.hal.science/hal-03944216`.

[4] Marianne Akian, Jean-Philippe Chancelier, and Benoît Tran. A stochastic algorithm for deterministic multistage optimization problems. *Annals of Operations Research*, 345:1–38, January 2025. `https://arxiv.org/abs/1810.12870`.

[5] Marianne Akian and Eric Fodjo. From a monotone probabilistic scheme to a probabilistic max-plus algorithm for solving Hamilton-Jacobi-Bellman equations. In *Hamilton-Jacobi-Bellman equations*, volume 21 of *Radon Ser. Comput. Appl. Math.*, pages 1–23. De Gruyter, Berlin, 2018.

[6] Marianne Akian, Stéphane Gaubert, and Loïc Marchesini. The Competive Spectral Radius of Families of Nonexpansive Mappings. `https://arxiv.org/abs/2410.21097`, October 2024.

[7] Marianne Akian, Stéphane Gaubert, Loïc Marchesini, and Ian Morris. Continuity and approximability of competitive spectral radii. In *CDC 2025 - 64th IEEE Conference on Decision and Control*, Rio De Janeiro, Brazil, December 2025. `https://arxiv.org/abs/2505.22468`.

[8] 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.

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

[10] 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.

[11] 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.

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

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

[14] Mario Bravo and Roberto Cominetti. Stochastic fixed-point iterations for nonexpansive maps: convergence and error bounds. *SIAM J. Control Optim.*, 62(1):191–219, 2024.

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

[16] Emmanuel Fernández-Gaucherand, Aristotle Arapostathis, and Steven I. Marcus. On the average cost optimality equation and the structure of optimal policies for partially observable Markov decision processes. *Ann. Oper. Res.*, 29(1-4):439–469, 1991.

[17] Stéphane Gaubert and Nikolas Stott. A convergent hierarchy of non-linear eigenproblems to compute the joint spectral radius of nonnegative matrices. *Math. Control Relat. Fields*, 10(3):573–590, 2020.

[18] 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.

[19] 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.

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

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

[22] 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.

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

[24] W. M. McEneaney. A curse-of-dimensionality-free numerical method for solution of certain HJB PDEs. *SIAM J. Control Optim.*, 46(4):1239–1276, 2007.

[25] William M. McEneaney, Hidehiro Kaise, and Seung Hak Han. Idempotent method for continuous-time stochastic control and complexity attenuation. In *Proceedings of the 18th IFAC World Congress, 2011*, pages 3216–3221, Milano, Italie, 2011.

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

[27] M. V. F. Pereira and L. M. V. G. Pinto. Multi-stage stochastic optimization applied to energy planning. *Math. Programming*, 52(2, Ser. B):359–375, 1991.

[28] Andy Philpott, Vitor de Matos, and Erlon Finardi. On Solving Multistage Stochastic Programs with Coherent Risk Measures. *Operations Research*, 61(4):957–970, August 2013.

[29] Z. Qu. A max-plus based randomized algorithm for solving a class of HJB PDEs. In *53rd IEEE Conference on Decision and Control*, pages 1575–1580, December 2014.

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

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

[32] 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.

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

[34] Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. Variance reduced value iteration and faster algorithms for solving Markov decision processes. *Nav. Res. Logist.*, 70(5):423–442, 2023.

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

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

[37] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning. An introduction*. Adapt. Comput. Mach. Learn. Cambridge, MA: MIT Press, 2nd expanded and updated edition edition, 2018.

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

[39] 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.

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

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