Proceedings of Machine Learning Research vol 125:1–9, 2020

Learning Zero-Sum Simultaneous-Move Markov Games Using Function Approximation and Correlated Equilibrium

Qiaomin Xie QIAOMIN.XIE@CORNELL.EDU
School of Operations Research and Information Engineering, Cornell Univerisity

Yudong Chen YUDONG.CHEN@CORNELL.EDU
School of Operations Research and Information Engineering, Cornell Univerisity

Zhaoran Wang ZHAORANWANG@GMAIL.COM
Department of Industrial Engineering and Management Sciences, Northwestern University

Zhuoran Yang ZY6@PRINCETON.EDU
Department of Operations Research and Financial Engineering, Princeton University

Abstract

In this work, we develop provably efficient reinforcement learning algorithms for two-player zero-sum Markov games with simultaneous moves. We consider a family of Markov games where the reward function and transition kernel possess a linear structure. Two settings are studied: In the offline setting, we control both players and the goal is to find the Nash Equilibrium efficiently by minimizing the worst-case duality gap. In the online setting, we control a single player and play against an arbitrary opponent; the goal is to minimize the regret. For both settings, we propose an optimistic variant of the least-squares minimax value iteration algorithm. We show that our algorithm is computationally efficient and provably achieves an $\widetilde{O}(\sqrt{d^3H^3T})$ upper bound on the duality gap and regret, without requiring additional assumptions on the sampling model. We highlight that our setting requires overcoming several new challenges that are absent in MDPs or turn-based Markov games. In particular, to achieve optimism under the simultaneous-move games, we construct both upper and lower confidence bounds of the value function, and then derive the optimistic policy by solving a general-sum matrix game with these bounds as the payoff matrices. As finding the Nash Equilibrium of this general-sum game is computationally hard, our algorithm instead solves for a Coarse Correlated Equilibrium (CCE), which can be obtained efficiently via linear programming. To our best knowledge, such a CCE-based mechanism for implementing optimism has not appeared in the literature and might be of interest in its own right.[1]

1. Introduction

Reinforcement learning is typically modeled as a Markov Decision Process (MDP), where an agent aims to learn the optimal policy via interaction with the environment. In Multi-agent reinforcement learning (MARL), multiple agents interact with each other and the underlying environment, and their goal is to optimize their individual returns. This problem is often formulated as a Markov game (Shapley, 1953), a generalization of the MDP model. Powered by function approximation techniques such as deep neural networks, MARL has recently enjoyed tremendous empirical successes across a variety of real-world applications, including the game of Go (Silver et al., 2016), real-time strategy games (Vinyals et al., 2019), Texas Hold'em poker (Moravvcík et al., 2017;

[1] Extended abstract. Full version appears as [arXiv 2002.07066, v3]

Brown and Sandholm, 2018), autonomous driving (Shalev-Shwartz et al., 2016), and learning communication and emergent behaviors (Foerster et al., 2016; Lowe et al., 2017; Bansal et al., 2017).

In contrast to the vibrant empirical study, theoretical understanding of MARL is relatively inadequate. Most existing work on Markov games assumes access to either a sampling oracle or a well-explored behavioral policy, which fails to capture the exploration-exploitation tradeoff that is fundamental in real-world applications. Moreover, these results mostly focus on the relatively simple turn-based setting. An exception is the work Wei et al. (2017), which extends the UCRL2 algorithm (Jaksch et al., 2010) to zero-sum simultaneous-move Markov games. However, their approach explicitly estimates the transition model and thus only works in the tabular setting. Problems with complicated state spaces and transitions necessitate the use of function approximation architectures. In this regard, a fundamental question is left open: Can we design a provably efficient reinforcement learning algorithm for Markov games under the function approximation setting?

In this paper, we provide an affirmative answer to this question for two-player zero-sum Markov games with simultaneous moves and a linear structure. In particular, we study an episodic setting, where each episode consists of H timesteps and the players act simultaneously at each timestep. Upon reaching the H-th timestep, the episode terminates and players replay the game again by starting a new episode. Here, the players have no knowledge of the system model (i.e., the transition kernel) nor access to a sampling oracle that returns the next state and rewards for an arbitrary state-action pair. Therefore, the players have to learn the system from data by playing the game sequentially through each episode and repeatedly for multiple episodes. More specifically, we study episodic Markov games under both the offline and online settings. In the offline setting, both players are controlled by a central learner, and the goal is to find an approximate Nash Equilibrium of the game, with the approximation error measured by a notion of duality gap. In the online setting, we control one of the players and play against an opponent who implements an arbitrary policy. Our goal is to minimize the total regret, defined as the difference between the cumulative return of the controlled player and its optimal achievable return when the opponent plays the best response policy. Both settings are generalizations of the regret minimization problem for MDPs.

Furthermore, to incorporate function approximation, we consider Markov games with a linear structure, motivated by the linear MDP model recently studied in Jin et al. (2019). In particular, we assume that both the transition kernel and the reward admit a d-dimensional linear representation in a known feature mapping, which can be potentially nonlinear in its inputs. For both the online and offline settings, we propose the first provably efficient reinforcement learning algorithm without additional assumptions on the sampling model. Our algorithm is an Optimistic version of Minimax Value Iteration (OMNI-VI) with least squares estimation—a model-free approach—which constructs upper confidence bounds of the optimal action-value function to promote exploration. We show that the OMNI-VI algorithm is computationally efficient, and it provably achieves an $\widetilde{O}(\sqrt{d^3H^3T})$ regret in the online setting and a similar duality gap guarantee in the offline setting, where $\widetilde{O}$ omits logarithmic terms. Note that the bounds do not depend on the cardinalities of the state and action spaces, which can be very large or even infinite. When specialized to MDPs, our results recover the regret bounds established in Jin et al. (2019) and are thus near-optimal.

We emphasize that the Markov game model poses several new and fundamental challenges that are absent in MDPs and arise due to subtle game-theoretic considerations. Addressing these challenges require several new ideas, which we summarize as follows.

1. **Optimism via General-Sum Games.** In the offline simultaneous-move setting, implementing the optimism principle for *both* players requires constructing both upper and lower confi-

dence bounds (UCB and LCB) for the optimal value function of the game. Doing so necessitates as an algorithmic subroutine of finding the solution of a *general-sum* (matrix) game where the two players' payoff functions correspond to the upper and lower bounds for the action-value (or Q) functions of the original Markov game, even though the latter is zero-sum to begin with. This stands in sharp contrast of turn-based games (Hansen et al., 2013; Jia et al., 2019; Sidford et al., 2019), in which each turn only involves constructing an UCB for one player.

2. **Using Correlated Equilibrium.** Finding the Nash equilibrium (NE) of a general-sum matrix game, however, is computationally hard in general (Daskalakis et al., 2009; Chen et al., 2009). Our second critical observation is that it suffices to find a *Coarse Correlated Equilibrium (CCE)* (Moulin and Vial, 1978) of the game. Originally developed in algorithmic game theory, CCE is a tractable notion of equilibrium that strictly generalizes NE. In contrast to NE, a CCE can be found efficiently even for general-sum games (Papadimitriou and Roughgarden, 2008). Moreover, our analysis shows that using any CCE of the matrix general-sum game are sufficient for ensuring optimism for the original Markov game.

3. **Concentration and Game Stability.** The last challenge is more technical, arising in the analysis of the algorithm where we need to establish certain uniform concentration bounds for the CCEs. As we elaborate later, the CCEs of a general-sum game are *unstable* (i.e., not Lipschitz) with respect to the payoff matrices. Therefore, standard approaches for proving uniform concentration, such as covering/ε-net arguments, fail fundamentally. We overcome this issue by carefully *stabilizing* the algorithm, for which we make use of an ε-net *in the algorithm*. Moreover, we show that this can be done in a computationally efficient way.

We discuss the above points in greater details after formally describing our algorithms. Note that our regret and duality gap bounds also imply polynomial sample complexity (or PAC) guarantees for learning the NEs of Markov games. Moreover, as turn-based games can be viewed as a special case of simultaneous games, where at each state the reward and transition kernel only depend on the action of one of the players, our algorithms and guarantees readily apply to the turn-based setting. To our best knowledge, our algorithm is the first provably efficient method for two-player zero-sum Markov games with simultaneous moves under the function approximation setting.

1.1. Related Work

There is a large body of literature on applying reinforcement learning methods to stochastic games. Under the tabular setting, the work in Littman (1994, 2001a,b); Greenwald et al. (2003); Hu and Wellman (2003); Grau-Moya et al. (2018) extends the Q-learning algorithm to zero/general-sum Markov games, and that in Perolat et al. (2018); Srinivasan et al. (2018) extends the actor-critic algorithm. Most of their convergence guarantees are asymptotic and rely on access to a sampling oracle. Particularly related to us is the work in Sidford et al. (2019), which proposes a variance-reduced minimax Q-learning algorithm with near-optimal sample complexity. The theoretical results therein also require a sampling oracle, and they focus turn-based games, a special case of simultaneous-move games. The work in Lagoudakis and Parr (2012); Perolat et al. (2015); Pérolat et al. (2016b,a,c); Yang et al. (2019) applies function approximation techniques to value-iteration methods and establishes finite-time convergence to the NEs of two-player zero-sum Markov games. Their results are based on the framework of fitted value-iteration (Munos and Szepesvári, 2008) and the availability of a well-explored behavioral policy. The recent work in Jia et al. (2019) studies turn-based zero-sum Markov games, where the transition model is assumed to be embedded in some d-dimensional feature space, extending the MDP model proposed by Yang and Wang (2019b). In summary, all of the work above either assumes a sampling oracle or a well explored behavioral policy for drawing transitions, therefore effectively bypassing the exploration issue.

Our work builds on a line of research on provably efficient methods for MDPs without additional assumptions on the sampling model. Most of the existing work focus on the tabular setting; see e.g., Strehl et al. (2006); Jaksch et al. (2010); Osband et al. (2014); Osband and Van Roy (2016); Azar et al. (2017); Dann et al. (2017); Agrawal and Jia (2017); Jin et al. (2018); Russo (2019). Under the function approximation setting, sample-efficient algorithms have been proposed using linear function approximators (Abbasi-Yadkori et al., 2019a,b; Jin et al., 2019; Yang and Wang, 2019a; Zanette et al., 2019; Du et al., 2019b; Cai et al., 2019; Wang et al., 2019), as well as nonlinear ones (Wen and Van Roy, 2017; Jiang et al., 2017; Dann et al., 2018; Du et al., 2019b; Dong et al., 2019; Du et al., 2019a). Among these results, our work is most related to Jin et al. (2019); Zanette et al. (2019); Cai et al. (2019), which consider linear MDP models and propose optimistic and randomized variants of least-squares value iteration (LSVI) as well as optimistic variants of proximal policy optimization (Schulman et al., 2017). Our linear Markov game model generalizes the MDP model considered in these papers, and our OMNI-VI algorithm can be viewed as a generalization of the optimistic LSVI method proposed in Jin et al. (2019). As mentioned before, the game structures in our problem pose fundamental challenges that are absent in MDPs, and thus their algorithms cannot be trivially extended to our game setting.

Finally, we remark that work on provably sample efficient RL methods for Markov games is quite scarce. The only comparable work we are aware of is Wei et al. (2017), which proposes a model-based algorithm by extending the UCRL2 algorithm (Jaksch et al., 2010) for tabular MDPs to the game setting. Similarly to our work, they also consider both the online and offline settings. On the other hand, they only consider tabular setting, which is a special case of our linear model. Moreover, their model-based algorithm explicitly estimates the Markov transition kernel and relies on the complicated technique of Extended Value Iteration, whose computational cost is quite high as it requires augmenting the state/action spaces. In comparison, our algorithm is model-free in the sense that it directly estimates the value functions; moreover, the computational cost of our algorithm only depends on the dimension d of the feature and not the cardinality of the state space.

Acknowledgments

Q. Xie and Y. Chen would like to thank Siddhartha Banerjee for inspiring discussion.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz. POLITEX: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pages 3692–3702, 2019a.

Yasin Abbasi-Yadkori, Nevena Lazic, Csaba Szepesvari, and Gellert Weisz. Exploration-enhanced POLITEX. *arXiv preprint arXiv:1908.10479*, 2019b.

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pages 1184–1194, 2017.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org, 2017.

Trapit Bansal, Jakub Pachocki, Szymon Sidor, Ilya Sutskever, and Igor Mordatch. Emergent complexity via multi-agent competition. *arXiv preprint arXiv:1710.03748*, 2017.

Avrim Blum, MohammadTaghi Hajiaghayi, Katrina Ligett, and Aaron Roth. Regret minimization and the price of total anarchy. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 373–382, 2008.

Noam Brown and Tuomas Sandholm. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*, 2019.

Xi Chen, Xiaotie Deng, and Shang-Hua Teng. Settling the complexity of computing two-player Nash equilibria. *Journal of the ACM (JACM)*, 56(3):14, 2009.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5713–5723, 2017.

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient PAC rl with rich observations. In *Advances in Neural Information Processing Systems*, pages 1422–1432, 2018.

Constantinos Daskalakis, Paul W. Goldberg, and Christos H. Papadimitriou. The complexity of computing a Nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009.

Trevor Davis, Neil Burch, and Michael Bowling. Using response functions to measure strategy strength. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

Kefan Dong, Jian Peng, Yining Wang, and Yuan Zhou. $\sqrt{n}$-regret for learning in markov decision processes with function approximation and low Bellman rank. *arXiv preprint arXiv:1909.02506*, 2019.

Simon S Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudík, and John Langford. Provably efficient RL with rich observations via latent state decoding. *arXiv preprint arXiv:1901.09018*, 2019a.

Simon S Du, Yuping Luo, Ruosong Wang, and Hanrui Zhang. Provably efficient Q-learning with function approximation via distribution shift error checking oracle. In *Advances in Neural Information Processing Systems*, pages 8058–8068, 2019b.

Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in neural information processing systems*, pages 2137–2145, 2016.

Jordi Grau-Moya, Felix Leibfried, and Haitham Bou-Ammar. Balancing two-player stochastic games with soft q-learning. *arXiv preprint arXiv:1802.03216*, 2018.

Amy Greenwald, Keith Hall, and Roberto Serrano. Correlated Q-learning. In *International Conference on Machine Learning*, volume 20, page 242, 2003.

Thomas Dueholm Hansen, Peter Bro Miltersen, and Uri Zwick. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *Journal of the ACM (JACM)*, 60(1):1, 2013.

Junling Hu and Michael P Wellman. Nash Q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov):1039–1069, 2003.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro A Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. *arXiv preprint arXiv:1810.08647*, 2018.

Zeyu Jia, Lin F. Yang, and Mengdi Wang. Feature-based Q-learning for two-player stochastic games. *arXiv preprint arXiv:1906.00423*, 2019.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1704–1713. JMLR. org, 2017.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I. Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*, 2019.

Michail Lagoudakis and Ron Parr. Value function approximation in zero-sum Markov games. *arXiv preprint arXiv:1301.0580*, 2012.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *preprint*, page 28, 2018.

Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

Michael L. Littman. Friend-or-foe Q-learning in general sum games. In *International Conference on Machine Learning*, pages 322–328, 2001a.

Michael L. Littman. Value-function reinforcement learning in Markov games. *Cognitive Systems Research*, 2(1):55–66, 2001b.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in neural information processing systems*, pages 6379–6390, 2017.

A. Maitra and T. Parthasarathy. On stochastic games. *Journal of Optimization Theory and Applications*, 5(4):289–300, 1970.

A. Maitra and T. Parthasarathy. On stochastic games, ii. *Journal of Optimization Theory and Applications*, 8(2):154–160, 1971.

Matej Moravvcík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

Hervé Moulin and J-P Vial. Strategically zero-sum games: the class of games whose completely mixed equilibria cannot be improved upon. *International Journal of Game Theory*, 7(3-4):201–221, 1978.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(May):815–857, 2008.

Ian Osband and Benjamin Van Roy. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. *arXiv preprint arXiv:1402.0635*, 2014.

Christos H Papadimitriou and Tim Roughgarden. Computing correlated equilibria in multi-player games. *Journal of the ACM (JACM)*, 55(3):1–29, 2008.

Julien Perolat, Bruno Scherrer, Bilal Piot, and Olivier Pietquin. Approximate dynamic programming for two-player zero-sum Markov games. In *International Conference on Machine Learning*, pages 1321–1329, 2015.

Julien Pérolat, Bilal Piot, Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. Softened approximate policy iteration for markov games. 2016a.

Julien Pérolat, Bilal Piot, Bruno Scherrer, and Olivier Pietquin. On the use of non-stationary strategies for solving two-player zero-sum markov games. In *Artificial Intelligence and Statistics*, pages 893–901, 2016b.

Julien Pérolat, Florian Strub, Bilal Piot, and Olivier Pietquin. Learning Nash equilibrium for general-sum markov games from batch data. *arXiv preprint arXiv:1606.08718*, 2016c.

Julien Perolat, Bilal Piot, and Olivier Pietquin. Actor-critic fictitious play in simultaneous move multistage games. In *International Conference on Artificial Intelligence and Statistics*, pages 919–928, 2018.

Daniel Russo. Worst-case regret bounds for exploration via randomized value functions. In *Advances in Neural Information Processing Systems*, pages 14410–14420, 2019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

Lloyd S. Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10): 1095–1100, 1953.

Aaron Sidford, Mengdi Wang, Lin F Yang, and Yinyu Ye. Solving discounted stochastic two-player games with near-optimal time and sample complexity. *arXiv preprint arXiv:1908.11071*, 2019.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, and Marc Lanctot. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484, 2016.

Sriram Srinivasan, Marc Lanctot, Vinicius Zambaldi, Julien Pérolat, Karl Tuyls, Rémi Munos, and Michael Bowling. Actor-critic policy optimization in partially observable multiagent environments. In *Advances in neural information processing systems*, pages 3422–3435, 2018.

Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. PAC model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 881–888, 2006.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok, editors, *Compressed Sensing*, pages 210–268. Cambridge University Press, 2012. ISBN 9780511794308.

Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, oct 2019.

Yining Wang, Ruosong Wang, Simon S Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Online reinforcement learning in stochastic games. In *Advances in Neural Information Processing Systems*, pages 4987–4997, 2017.

Zheng Wen and Benjamin Van Roy. Efficient reinforcement learning in deterministic systems with value function generalization. *Mathematics of Operations Research*, 42(3):762–782, 2017.

Lin F Yang and Mengdi Wang. Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. *arXiv preprint arXiv:1905.10389*, 2019a.

Lin F Yang and Mengdi Wang. Sample-optimal parametric Q-learning with linear transition models. *arXiv preprint arXiv:1902.04779*, 2019b.

Zhuora Yang, Yuchen Xie, and Zhaoran Wang. A theoretical analysis of deep Q-learning. *arXiv preprint:1901.00137*, 2019.

Andrea Zanette, David Brandfonbrener, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. *arXiv preprint arXiv:1911.00567*, 2019.