Full text available at: http://dx.doi.org/10.1561/2400000038

Constrained Reinforcement Learning with Average Reward Objective: Model-Based and Model-Free Algorithms

Other titles in Foundations and Trends® in Optimization

Stochastic Optimization Methods for Policy Evaluation in Reinforcement Learning
Yi Zhou and Shaocong Ma
ISBN: 978-1-63828-370-6

Atomic Decomposition via Polar Alignment: The Geometry of Structured Optimization
Zhenan Fan, Halyun Jeong, Yifan Sun and Michael P. Friedlander
ISBN: 978-1-68083-742-1

Optimization Methods for Financial Index Tracking: From Theory to Practice
Konstantinos Benidis, Yiyong Feng and Daniel P. Palomar
ISBN: 978-1-68083-464-2

The Many Faces of Degeneracy in Conic Optimization
Dmitriy Drusvyatskiy and Henry Wolkowicz
ISBN: 978-1-68083-390-4

Constrained Reinforcement Learning with Average Reward Objective: Model-Based and Model-Free Algorithms

Vaneet Aggarwal
Purdue University
vaneet@purdue.edu

Washim Uddin Mondal
Indian Institute of Technology Kanpur
wmondal@iitk.ac.in

Qinbo Bai
Purdue University
bai113@purdue.edu

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Optimization

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

V. Aggarwal *et al.*. *Constrained Reinforcement Learning with Average Reward Objective: Model-Based and Model-Free Algorithms*. Foundations and Trends® in Optimization, vol. 6, no. 4, pp. 193–298, 2024.

ISBN: 978-1-63828-397-3

Foundations and Trends® in Optimization
Volume 6, Issue 4, 2024

Editorial Scope

Foundations and Trends® in Optimization publishes survey and tutorial articles in the following topics:

- algorithm design, analysis, and implementation (especially, on modern computing platforms
- models and modeling systems, new optimization formulations for practical problems
- applications of optimization in machine learning, statistics, and data analysis, signal and image processing, computational economics and finance, engineering design, scheduling and resource allocation, and other areas

Information for Librarians

Foundations and Trends® in Optimization, 2024, Volume 6, 4 issues. ISSN paper version 2167-3888. ISSN online version 2167-3918. Also available as a combined paper and online subscription.

Contents

Constrained Reinforcement Learning with Average Reward Objective: Model-Based and Model-Free Algorithms

Vaneet Aggarwal[1], Washim Uddin Mondal[2] and Qinbo Bai[1]

[1]*Purdue University, USA; vaneet@purdue.edu, bai113@purdue.edu*
[2]*Indian Institute of Technology Kanpur, India; wmondal@iitk.ac.in*

ABSTRACT

Reinforcement Learning (RL) serves as a versatile framework for sequential decision-making, finding applications across diverse domains such as robotics, autonomous driving, recommendation systems, supply chain optimization, biology, mechanics, and finance. The primary objective of these applications is to maximize the average reward. Real-world scenarios often necessitate adherence to specific constraints during the learning process.

This monograph focuses on the exploration of various model-based and model-free approaches for Constrained RL within the context of average reward Markov Decision Processes (MDPs). The investigation commences with an examination of model-based strategies, delving into two foundational methods – optimism in the face of uncertainty and posterior sampling. Subsequently, the discussion transitions to parametrized model-free approaches, where the primal

Vaneet Aggarwal, Washim Uddin Mondal and Qinbo Bai (2024), "Constrained Reinforcement Learning with Average Reward Objective: Model-Based and Model-Free Algorithms", Foundations and Trends® in Optimization: Vol. 6, No. 4, pp 193–298. DOI: 10.1561/2400000038.

dual policy gradient-based algorithm is explored as a solution for constrained MDPs. The monograph provides regret guarantees and analyzes constraint violation for each of the discussed setups.

For the above exploration, we assume the underlying MDP to be ergodic. Further, this monograph extends its discussion to encompass results tailored for weakly communicating MDPs, thereby broadening the scope of its findings and their relevance to a wider range of practical scenarios.

1

Introduction

Reinforcement Learning (RL) describes a class of problems where an agent repeatedly interacts with an unknown environment. The environment possesses a state that changes as a result of the action executed by the agent according to some pre-determined but unknown probability law. The environment also generates feedback, which is often called the reward. The agent's goal is to choose a sequence of actions (based on the sequence of observed states and rewards) that maximizes the expected cumulative sum of rewards obtained via this procedure. This model has found its application in a wide array of areas, ranging from networking to transportation to robotics to epidemic control [1], [20], [36], [39], [45], [48]. RL problems are typically analyzed via three distinct setups−episodic, infinite horizon discounted reward, and infinite horizon average reward. In an episodic setup, the environment restores its initial state after a certain number of interactions. Examples include video game-based applications where the learner restarts the game after either winning or losing it. In a discounted setup, the learner aims to maximize the expected *discounted* sum of rewards. The underlying philosophy is that the current reward, in certain applications, is deemed more valuable than the rewards obtained in the future. This idea naturally fits into financial applications where the reward (money) loses value over time due to inflation. The average reward setup, on the contrary, places both the current and future rewards on the same footing and aims to maximize the expected average reward computed over an infinitely long time horizon. The basic premise of the infinite horizon average reward setup aligns with most practical scenarios due to its ability to capture essential long-term behaviors. Some applications in real life require the learning procedure to respect the boundaries of certain constraints. In an epidemic control setup, for example, vaccination policies must take the supply shortage (budget constraint) into account. Such restrictive decision-making routines are described by a constrained Markov Decision Process (CMDP) [6], [15], [50]. This monograph aims to provide the key approaches to tackle CMDP with an average reward objective.

To gain more insight into CMDPs, consider a wireless sensor network where a device aims to update a server with its sensed values. At time t, the sensor can either choose to send a packet which, upon successful transmission, fetches a reward of one unit or to queue the packet and obtain a zero reward. However, communicating a packet results in p_t power consumption. The success probability of the intended packet is decided via a pre-determined but unknown function of p_t and the current wireless channel condition, s_t. The goal is to send as many packets as possible while keeping the average power consumption, $\sum_{t=1}^{T} p_t/T$, within some limit, say C. The *state* of the environment can be described by the pair (s_t, q_t) where s_t, as stated above, is the channel condition, and q_t is the queue length at time t. To limit the power consumption, the agent may choose to transmit packets when the channel condition is good or when the queue length grows beyond a certain threshold. The agent aims to learn the policies in an *online manner* which requires efficiently balancing exploration of state-space and exploitation of the estimated system dynamics [62].

Similar to the example above, many applications require keeping some costs low while simultaneously maximizing the rewards [10]. This monograph discusses model-based and model-free algorithms for the CMDP learning problem described above. A model-based algorithm aims to learn the optimal policy by creating a good estimate of the state-transition function of the underlying CMDP. The caveat of the model-based approach is the large memory requirement to store the estimated parameters which effectively curtails its applicability to large state space CMDPs. The alternative strategy, known as the model-free approach, either directly estimates the policy function or maintains an estimate of the Q function, which is subsequently used for policy generation [66]. Model-free algorithms typically demand lower memory and computational resources than their model-based counterparts.

The problem setup, where the system dynamics are known, is extensively studied [10]. For a constrained setup, the optimal policy is possibly stochastic [10], [57]. Even though the problem has been widely studied in episodic and discounted reward setups [13], [15], [26], [35], [72], the focus of this monograph is on the average reward setup, thus providing a comprehensive study of the state of the art in the area.

1.1 Section Organization

In Section 2, we consider a model-based approach for learning CMDPs with average reward and costs. We discuss posterior sampling-based and optimism-based algorithms. We demonstrate $\widetilde{O}(\sqrt{T})$ objective regret and zero constraint violation for both of them. The presented results follow the recent works of Agarwal *et al.* [6], [7].

In Section 3, we consider a model-free approach for learning CMDP via general parametrization. General parameterization indexes the policies by finite-dimensional parameters (e.g., weights of neural networks) to accommodate large state spaces. The learning is manifested by updating these parameters using policy gradient (PG)-type algorithms. This section primarily follows the works of Bai *et al.* [16], [17] and presents an algorithm that achieves $\widetilde{O}(T^{4/5})$ objective regret and constraint violation. Note that general parameterization subsumes the tabular setup. Moreover, the best-known regret bound achieved by any tabular model-free algorithm for average-reward CMDPs is $\tilde{\mathcal{O}}(T^{5/6})$ [66] which is worse than the above result in terms of orders. Due to this reason, we do not present any algorithm specific to the tabular model-free setup.

In the previous sections, we assumed the underlying CMDP to be ergodic. In Section 4, we go beyond this assumption to consider weakly communicating CMDPs. Note that the class of weakly communicating

CMDPs contains the set of ergodic CMDPs, and it is the largest class for which one can hope to establish theoretical guarantees for all instances [18], [40]. This section presents the model-based approach of Chen *et al.* [23] and proves $\tilde{\mathcal{O}}\left(T^{2/3}\right)$ objective regret and constraint violation. We note that no known model-free algorithm currently exists that guarantees a sublinear regret and constraint violation for weakly communicating CMDPs. This leaves multiple open questions.

1.2 Some Useful Inequalities

In this section, we provide some important inequalities for random variables, some of which will be used in this monograph.

Lemma 1.1 (Jensen's Inequality). Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function, and let X be a random variable. If $E[X]$ is finite, then

$$f(E[X]) \leq E[f(X)].$$

Lemma 1.2 (Cauchy-Schwarz Inequality [30]). For any vectors $\mathbf{u}$ and $\mathbf{v}$ in a real or complex inner product space, the Cauchy-Schwarz Inequality holds:

$$|\langle \mathbf{u}, \mathbf{v} \rangle|^2 \leq \langle \mathbf{u}, \mathbf{u} \rangle \cdot \langle \mathbf{v}, \mathbf{v} \rangle.$$

Lemma 1.3. [21, Lemma 30] For a random variable X such that $|X| \leq C$ almost surely, we have: $\text{VAR}[X^2] \leq 4C^2 \text{VAR}[X]$.

Lemma 1.4 (Azuma-Hoeffding's Inequality [60]). Let $X_1, \cdots, X_n$ be a Martingale difference sequence such that $|X_i| \leq c$ almost surely for all $i \in \{1, 2, \cdots, n\}$, then,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2}{2nc^2}\right) \tag{1.1}$$

Lemma 1.5 (Any interval Azuma's inequality, [23]). Let $\{X_i\}_{i=1}^{\infty}$ be a martingale difference sequence and $|X_i| \leq B$ almost surely. Then with probability at least $1-\delta$, for any l, n: $\left|\sum_{i=l}^{l+n-1} X_i\right| \leq B\sqrt{2n \ln \frac{4(l+n-1)^3}{\delta}}$.

Lemma 1.6. [22, Lemma 38] Let $\{X_i\}_{i=1}^{\infty}$ be a martingale difference sequence adapted to the filtration $\{\mathcal{F}_i\}_{i=0}^{\infty}$ and $|X_i| \leq B$ for some $B > 0$. Then with probability at least $1-\delta$, for all $n \geq 1$ simultaneously,

$$\left|\sum_{i=1}^{n} X_i\right| \leq 3\sqrt{\sum_{i=1}^{n} \mathbb{E}[X_i^2|\mathcal{F}_{i-1}] \ln\frac{4B^2n^3}{\delta}} + 2B\ln\frac{4B^2n^3}{\delta}.$$

Lemma 1.7. [68] Let p be an m-dimensional distribution and $\bar{p}$ be its empirical estimate obtained by averaging over n samples. Then, $\|p-\bar{p}\|_1 \leq \sqrt{m\ln\frac{2}{\delta}/n}$ with probability at least $1-\delta$.

Lemma 1.8. [25, Theorem D.3] Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of i.i.d random variables with expectation μ and $X_n \in [0, B]$ almost surely. Then with probability at least $1-\delta$, for any $n \geq 1$:

$$\begin{aligned}
&\left|\sum_{i=1}^{n}(X_i-\mu)\right| \\
&\leq \min\left\{2\sqrt{B\mu n\ln\frac{2n}{\delta}} + B\ln\frac{2n}{\delta}, 2\sqrt{B\sum_{i=1}^{n}X_i\ln\frac{2n}{\delta}} + 7B\ln\frac{2n}{\delta}\right\}.
\end{aligned}$$

Lemma 1.9. [25, Lemma D.4] and [24, Lemma E.2] Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of random variables w.r.t to the filtration $\{\mathcal{F}_i\}_{i=0}^{\infty}$ and $X_i \in [0, B]$ almost surely. Then with probability at least $1-\delta$, for all $n \geq 1$ simultaneously:

$$\begin{aligned}
\sum_{i=1}^{n}\mathbb{E}[X_i|\mathcal{F}_{i-1}] &\leq 2\sum_{i=1}^{n}X_i + 4B\ln\frac{4n}{\delta}, \\
\sum_{i=1}^{n}X_i &\leq 2\sum_{i=1}^{n}\mathbb{E}[X_i|\mathcal{F}_{i-1}] + 8B\ln\frac{4n}{\delta}.
\end{aligned}$$

References

[1] A. O. Al-Abbasi, A. Ghosh, and V. Aggarwal, "Deeppool: Distributed model-free algorithm for ride-sharing using deep reinforcement learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 12, 2019, pp. 4714–4727.

[2] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "Optimality and approximation with policy gradient methods in markov decision processes," in *Proceedings of Thirty Third Conference on Learning Theory*, J. Abernethy and S. Agarwal, Eds., ser. Proceedings of Machine Learning Research, vol. 125, pp. 64–66, PMLR, Sep. 2020. URL: http://proceedings.mlr.press/v125/agarwal20a.html.

[3] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "On the theory of policy gradient methods: Optimality, approximation, and distribution shift," *The Journal of Machine Learning Research*, vol. 22, no. 1, 2021, pp. 4431–4506.

[4] M. Agarwal and V. Aggarwal, "Reinforcement learning for joint optimization of multiple rewards," *Journal of Machine Learning Research*, vol. 24, no. 49, 2023, pp. 1–41.

[5] M. Agarwal, V. Aggarwal, and T. Lan, "Multi-objective reinforcement learning with non-linear scalarization," in *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 9–17, 2022.

[6] M. Agarwal, Q. Bai, and V. Aggarwal, “Concave utility reinforcement learning with zero-constraint violations,” *Transactions on Machine Learning Research*, 2022.

[7] M. Agarwal, Q. Bai, and V. Aggarwal, “Regret guarantees for model-based reinforcement learning with long-term average constraints,” in *Uncertainty in Artificial Intelligence*, PMLR, pp. 22–31, 2022.

[8] S. Agrawal and R. Jia, “Optimistic posterior sampling for reinforcement learning: Worst-case regret bounds,” in *Advances in Neural Information Processing Systems*, pp. 1184–1194, 2017.

[9] E. Altman and A. Schwartz, “Adaptive control of constrained markov chains,” *IEEE Transactions on Automatic Control*, vol. 36, no. 4, 1991, pp. 454–462. DOI: 10.1109/9.75103.

[10] E. Altman, *Constrained Markov decision processes*, vol. 7. CRC Press, 1999.

[11] Q. Bai, M. Agarwal, and V. Aggarwal, “Joint optimization of concave scalarized multi-objective reinforcement learning with policy gradient based algorithm,” *Journal of Artificial Intelligence Research*, vol. 74, 2022, pp. 1565–1597.

[12] Q. Bai, V. Aggarwal, and A. Gattami, “Provably sample-efficient model-free algorithm for mdps with peak constraints,” *Journal of Machine Learning Research*, vol. 24, no. 60, 2023, pp. 1–25.

[13] Q. Bai, A. S. Bedi, M. Agarwal, A. Koppel, and V. Aggarwal, “Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 3682–3689, 2022.

[14] Q. Bai, A. S. Bedi, M. Agarwal, A. Koppel, and V. Aggarwal, “Achieving zero constraint violation for concave utility constrained reinforcement learning via primal-dual approach,” *Journal of Artificial Intelligence Research*, vol. 78, 2023, pp. 975–1016.

[15] Q. Bai, A. S. Bedi, and V. Aggarwal, “Achieving zero constraint violation for constrained reinforcement learning via conservative natural policy gradient primal-dual algorithm,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[16] Q. Bai, W. U. Mondal, and V. Aggarwal, “Learning Infinite Horizon Average Reward Constrained Markov Decision Processes with Primal-Dual Policy Gradient Algorithm,” *arXiv preprint arXiv:2402.02042*, 2024.

[17] Q. Bai, W. U. Mondal, and V. Aggarwal, “Regret analysis of policy gradient algorithm for infinite horizon average reward markov decision processes,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[18] P. L. Bartlett and A. Tewari, “Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps,” in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 35–42, 2009.

[19] K. Brantley, M. Dudik, T. Lykouris, S. Miryoosefi, M. Simchowitz, A. Slivkins, and W. Sun, “Constrained episodic reinforcement learning in concave-convex and knapsack settings,” *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 16 315–16 326.

[20] C.-L. Chen, H. Zhou, J. Chen, M. Pedramfar, V. Aggarwal, T. Lan, Z. Zhu, C. Zhou, T. Gasser, P. M. Ruiz, *et al.*, “Two-tiered online optimization of region-wide datacenter resource allocation via deep reinforcement learning,” *arXiv preprint arXiv:2306.17054*, 2023.

[21] L. Chen, M. Jafarnia-Jahromi, R. Jain, and H. Luo, “Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path,” *Advances in Neural Information Processing Systems*, 2021.

[22] L. Chen, R. Jain, and H. Luo, “Improved no-regret algorithms for stochastic shortest path with linear mdp,” in *International Conference on Machine Learning*, PMLR, pp. 3204–3245, 2022.

[23] L. Chen, R. Jain, and H. Luo, “Learning infinite-horizon average-reward Markov decision process with constraints,” in *Proceedings of the 39th International Conference on Machine Learning*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., ser. Proceedings of Machine Learning Research, vol. 162, pp. 3246–3270, PMLR, 17–23 Jul 2022.

[24] A. Cohen, Y. Efroni, Y. Mansour, and A. Rosenberg, "Minimax regret for stochastic shortest path," *Advances in Neural Information Processing Systems*, 2021.

[25] A. Cohen, H. Kaplan, Y. Mansour, and A. Rosenberg, "Near-optimal regret bounds for stochastic shortest path," in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, pp. 8210–8219, PMLR, 2020.

[26] D. Ding, X. Wei, Z. Yang, Z. Wang, and M. Jovanovic, "Provably efficient safe exploration via primal-dual policy optimization," in *International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 3304–3312, 2021.

[27] D. Ding, K. Zhang, T. Basar, and M. Jovanovic, "Natural policy gradient primal-dual method for constrained markov decision processes," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[28] D. Ding, K. Zhang, J. Duan, T. Başar, and M. R. Jovanović, *Convergence and sample complexity of natural policy gradient primal-dual methods for constrained mdps*, 2023. arXiv: 2206.02346 [math.OC].

[29] R. Dorfman and K. Y. Levy, "Adapting to mixing time in stochastic optimization with markovian data," in *International Conference on Machine Learning*, PMLR, pp. 5429–5446, 2022.

[30] S. S. Dragomir, "A survey on cauchy-bunyakovsky-schwarz type discrete inequalities," *J. Inequal. Pure Appl. Math*, vol. 4, no. 3, 2003, pp. 1–142.

[31] Y. Efroni, S. Mannor, and M. Pirotta, "Exploration-exploitation in constrained mdps," *arXiv preprint arXiv:2003.02189*, 2020.

[32] R. Fruit, M. Pirotta, A. Lazaric, and R. Ortner, "Efficient bias-span-constrained exploration-exploitation in reinforcement learning," in *International Conference on Machine Learning*, PMLR, pp. 1578–1586, 2018.

[33] S. Ganesh and V. Aggarwal, "An accelerated multi-level monte carlo approach for average reward reinforcement learning with general policy parametrization," *arXiv*, 2024.

[34] S. Ganesh, W. U. Mondal, and V. Aggarwal, "Variance-reduced policy gradient approaches for infinite horizon average reward markov decision processes," *arXiv preprint arXiv:2404.02108*, 2024.

[35] A. Gattami, Q. Bai, and V. Aggarwal, "Reinforcement learning for constrained markov decision processes," in *International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 2656–2664, 2021.

[36] N. Geng, T. Lan, V. Aggarwal, Y. Yang, and M. Xu, "A multi-agent reinforcement learning perspective on distributed traffic engineering," in *2020 IEEE 28th International Conference on Network Protocols (ICNP)*, IEEE, pp. 1–11, 2020.

[37] A. Ghosh, X. Zhou, and N. Shroff, "Achieving sub-linear regret in infinite horizon average reward constrained mdp with linear function approximation," in *The Eleventh International Conference on Learning Representations*, 2023.

[38] H. Gong and M. Wang, "A duality approach for regret minimization in average-award ergodic markov decision processes," in *Learning for Dynamics and Control*, PMLR, pp. 862–883, 2020.

[39] G. Gonzalez, M. Balakuntala, M. Agarwal, T. Low, B. Knoth, A. W. Kirkpatrick, J. McKee, G. Hager, V. Aggarwal, Y. Xue, *et al.*, "Asap: A semi-autonomous precise system for telesurgery during communication delays," *IEEE Transactions on Medical Robotics and Bionics*, 2023.

[40] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," *Journal of Machine Learning Research*, vol. 11, no. Apr, 2010, pp. 1563–1600.

[41] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan, "Provably efficient reinforcement learning with linear function approximation," in *Proceedings of Thirty Third Conference on Learning Theory*, J. Abernethy and S. Agarwal, Eds., ser. Proceedings of Machine Learning Research, vol. 125, pp. 2137–2143, PMLR, Sep. 2020.

[42] J. Langford and S. Kakade, "Approximately optimal approximate reinforcement learning," in *Proceedings of ICML*, 2002.

[43] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[44] G. F. Lawler, *Introduction to stochastic processes*. Chapman and Hall/CRC, 2018.
[45] L. Ling, W. U. Mondal, and S. V. Ukkusuri, "Cooperating graph neural networks with deep reinforcement learning for vaccine prioritization," *IEEE Journal of Biomedical and Health Informatics*, 2024.
[46] T. Liu, R. Zhou, D. Kalathil, P. Kumar, and C. Tian, "Learning policies with zero or bounded constraint violation for constrained mdps," *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 17 183–17 193.
[47] Y. Liu, K. Zhang, T. Basar, and W. Yin, "An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods," *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 7624–7636.
[48] K. Manchella, M. Haliem, V. Aggarwal, and B. Bhargava, "Passgoodpool: Joint passengers and goods fleet management with reinforcement learning aided pricing, matching, and route planning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 4, 2021, pp. 3866–3877.
[49] W. U. Mondal and V. Aggarwal, "Improved sample complexity analysis of natural policy gradient algorithm with general parameterization for infinite horizon discounted reward markov decision processes," in *International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 3097–3105, 2024.
[50] W. U. Mondal and V. Aggarwal, "Sample-efficient constrained reinforcement learning with general parameterization," *arXiv preprint arXiv:2405.10624*, 2024.
[51] W. U. Mondal, V. Aggarwal, and S. V. Ukkusuri, "Mean-field control based approximation of multi-agent reinforcement learning in presence of a non-decomposable shared global state," *Transactions on Machine Learning Research*, 2023.
[52] I. Osband, D. Russo, and B. Van Roy, "(more) efficient reinforcement learning via posterior sampling," in *Advances in Neural Information Processing Systems*, pp. 3003–3011, 2013.

[53] I. Osband and B. Van Roy, “Why is posterior sampling better than optimism for reinforcement learning?” In *International conference on machine learning*, PMLR, pp. 2701–2710, 2017.

[54] Y. Ouyang, M. Gagrani, A. Nayyar, and R. Jain, “Learning unknown markov decision processes: A Thompson sampling approach,” *Advances in neural information processing systems*, vol. 30, 2017.

[55] B. Patel, W. A. Suttle, A. Koppel, V. Aggarwal, B. M. Sadler, A. S. Bedi, and D. Manocha, “Global optimality without mixing time oracles in average-reward rl via multi-level actor-critic,” in *International Conference on Machine Learning*, 2024.

[56] F. Pesquerel and O.-A. Maillard, “Imed-rl: Regret optimal learning of ergodic markov decision processes,” in *NeurIPS 2022-Thirty-sixth Conference on Neural Information Processing Systems*, 2022.

[57] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[58] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, 1st. New York, NY, USA: John Wiley & Sons, Inc., 1994.

[59] D. Russo and B. Van Roy, “Learning to optimize via posterior sampling,” *Mathematics of Operations Research*, vol. 39, no. 4, 2014, pp. 1221–1243.

[60] R. J. Serfling, “Probability inequalities for the sum in sampling without replacement,” *The Annals of Statistics*, vol. 2, no. 1, 1974, pp. 39–48.

[61] L. Shani, Y. Efroni, A. Rosenberg, and S. Mannor, “Optimistic policy optimization with bandit feedback,” in *Proceedings of the 37th International Conference on Machine Learning*, pp. 8604–8613, 2020.

[62] R. Singh, A. Gupta, and N. B. Shroff, “Learning in markov decision processes under constraints,” *arXiv preprint arXiv:2002.12435*, 2020.

[63] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, “Policy gradient methods for reinforcement learning with function approximation,” *Advances in neural information processing systems*, vol. 12, 1999.

[64] L. Wang, Q. Cai, Z. Yang, and Z. Wang, "Neural policy gradient methods: Global optimality and rates of convergence," in *International Conference on Learning Representations*, 2020.
[65] C.-Y. Wei, M. J. Jahromi, H. Luo, H. Sharma, and R. Jain, "Model-free reinforcement learning in infinite-horizon average-reward markov decision processes," in *International conference on machine learning*, PMLR, pp. 10 170–10 180, 2020.
[66] H. Wei, X. Liu, and L. Ying, "A provably-efficient model-free algorithm for infinite-horizon average-reward constrained markov decision processes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.
[67] H. Wei, X. Liu, and L. Ying, "Triple-q: A model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation," in *International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 3274–3307, 2022.
[68] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. J. Weinberger, "Inequalities for the l1 deviation of the empirical distribution," *Hewlett-Packard Labs, Tech. Rep*, 2003.
[69] T. Yu, Y. Tian, J. Zhang, and S. Sra, "Provably efficient algorithms for multi-objective competitive rl," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, pp. 12 167–12 176, PMLR, 18–24 Jul 2021.
[70] J. Zhang, C. Ni, C. Szepesvari, M. Wang, *et al.*, "On the convergence and sample efficiency of variance-reduced policy gradient method," *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 2228–2240.
[71] Z. Zhang and Q. Xie, "Sharper model-free reinforcement learning for average-reward markov decision processes," in *The Thirty Sixth Annual Conference on Learning Theory*, PMLR, pp. 5476–5477, 2023.
[72] L. Zheng and L. Ratliff, "Constrained upper confidence reinforcement learning," in *Learning for Dynamics and Control*, PMLR, pp. 620–629, 2020.