# Lyapunov-Based Sample Complexity Analysis for Weakly-Coupled MDPs

**Tianhao Wu** TIANHAO.WU@WISC.EDU
*Department of Industrial and Systems Engineering, University of Wisconsin–Madison*
**Matthew Zurek** MATTHEW.ZUREK@WISC.EDU
*Department of Computer Sciences, University of Wisconsin–Madison*
**Weina Wang** WEINAW@CS.CMU.EDU
*Computer Science Department, Carnegie Mellon University*
**Qiaomin Xie** QIAOMIN.XIE@WISC.EDU
*Department of Industrial and Systems Engineering, University of Wisconsin–Madison*

**Editors:** Steve Hanneke and Tor Lattimore

We study the sample complexity of learning in average-reward weakly-coupled Markov decision processes (WCMDPs) and Restless Bandits (RBs) under a generative model. Naive reduction to a tabular MDP leads to high complexity bounds as the state-action space is exponentially large in the number of arms $N$. By exploiting the weakly coupled structure, we show that near-optimal policies can be learned with sample and computational complexities that are polynomial in $N$. Specifically, we analyze the plug-in approach, which applies an efficient planning algorithm to an empirical model estimated from data. For fully heterogeneous WCMDPs, we establish the first finite-sample PAC guarantee with polynomial complexity and an $O(1/\sqrt{N})$ optimality gap. For homogeneous RBs, we further prove that a smaller optimality gap is achievable under mild structural assumptions.

Our main results can be summarized as follows.

**Theorem 1 (Informal)** *Consider an $N$-armed WCMDP or RB satisfying specific assumptions. Let $\widehat{\pi}$ be the learned policy obtained by applying a reference planning algorithm to the empirical model and $n$ be the number of samples per state-action pair. Then, for any initial state $\boldsymbol{s}_0$,*

$$\rho^\star(\boldsymbol{s}_0) - \rho^{\widehat{\pi}}(\boldsymbol{s}_0) \le \widetilde{O}\left(N/\sqrt{n}\right) + \varepsilon(N).$$

*Here $\varepsilon(N) = O(1/\sqrt{N})$ when the ID policy is used as the reference policy for WCMDP, and $\varepsilon(N) = O(\exp(-cN))$ when the two-set policy is used as the reference policy for RB.*

A main technical contribution of our work is a novel framework based on Lyapunov drift transfer. Classical approaches use simulation lemmas that rely on difficult-to-control bias function. We replace the bias function with an explicitly constructed Lyapunov function. Specifically, we first perform a Lyapunov drift analysis of the planning algorithm in the true system, and then employ a drift-transfer technique to analyze the empirical system by bounding the norm of the Lyapunov function. This framework applies to any planning algorithm admitting a drift analysis, providing a powerful tool for average-reward weakly-coupled systems.

Another technical contribution is a perturbation analysis for the LP relaxation used in RBs. The two-set policy depends on several structural properties of this LP, such as the support pattern, the neutral state, local stability, and the ergodicity of the induced single-armed Markov chain. We show that these properties are preserved when the empirical transition kernel is close to the true kernel. We also bound the distance between true and perturbed LP solutions. This result may be useful more broadly for analyzing LP-based policies under model perturbations.[1]

1. Extended abstract. Full version can be found at [arXiv:2606.14095, v2].

## Acknowledgments

T. Wu and Q. Xie are supported in part by National Science Foundation (NSF) Grants CNS-1955997, ECCS-2339794 and ECCS-2432546. T. Wu is also partially supported by NSF Award DMS-2023239. W. Wang is supported in part by the NSF Grants ECCS-2145713, CCF-2403194, CCF-2428569, and ECCS-2432545. Part of this research was performed while M. Zurek was visiting the Institute for Mathematical and Statistical Innovation (IMSI), which is supported by the NSF (Grant No. DMS-2425650). M. Zurek also acknowledges support by a Cisco Systems Fellowship and by NSF grant CCF-2233152.

## References

Alekh Agarwal, Sham Kakade, and Lin F. Yang. Model-Based Reinforcement Learning with a Generative Model is Minimax Optimal, April 2020. URL http://arxiv.org/abs/1906.03804. arXiv:1906.03804 [cs, math, stat] version: 3.

Konstantin Avrachenkov, Vivek S Borkar, and Pratik Shah. Lagrangian index policy for restless bandits with average reward. *arXiv preprint arXiv:2412.12641*, 2024.

Konstantin E Avrachenkov and Vivek S Borkar. Whittle index based q-learning for restless bandits with average reward. *Automatica*, 139:110186, 2022.

Arpita Biswas, Gaurav Aggarwal, Pradeep Varakantham, and Milind Tambe. Learning index policies for restless bandits with application to maternal healthcare. 2021.

Craig Boutilier and Tyler Lu. Budget allocation using weakly coupled, constrained markov decision processes. In *UAI*, 2016.

David B Brown and James E Smith. Index policies and performance bounds for dynamic selection problems. *Management Science*, 66(7):3029–3050, 2020.

David B Brown and Jingwei Zhang. Dynamic programs with shared resources and signals: Dynamic fluid policies and asymptotic optimality. *Operations Research*, 70(5):3015–3033, 2022.

David B Brown and Jingwei Zhang. Fluid policies, reoptimization, and performance guarantees in dynamic resource allocation. *Operations Research*, 73(2):1029–1045, 2025.

Wenhan Dai, Yi Gai, Bhaskar Krishnamachari, and Qing Zhao. The non-bayesian restless multi-armed bandit: A case of near-logarithmic regret. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2940–2943. IEEE, 2011.

Ibrahim El Shar and Daniel Jiang. Weakly coupled deep q-networks. *Advances in Neural Information Processing Systems*, 36:43931–43950, 2023.

Nicolas Gast, Bruno Gaujal, and Chen Yan. Reoptimization nearly solves weakly coupled markov decision processes. *arXiv preprint arXiv:2211.01961*, 2022.

Nicolas Gast, Bruno Gaujal, and Chen Yan. Linear program-based policies for restless bandits: Necessary and sufficient conditions for (exponentially fast) asymptotic optimality. *Mathematics of Operations Research*, 49(4):2468–2491, November 2024. ISSN 1526-5471. doi: 10.1287/moor.2022.0101. URL http://dx.doi.org/10.1287/moor.2022.0101.

Kevin D Glazebrook, HM Mitchell, and PS Ansell. Index policies for the maintenance of a collection of machines by a set of repairmen. *European Journal of Operational Research*, 165(1): 267–284, 2005.

Jeffrey Thomas Hawkins. *A Langrangian decomposition approach to weakly coupled dynamic optimization problems and its applications*. PhD thesis, Massachusetts Institute of Technology, 2003.

David J Hodge and Kevin D Glazebrook. On the asymptotic optimality of greedy index heuristics for multi-action restless bandits. *Advances in Applied Probability*, 47(3):652–667, 2015.

Yige Hong, Qiaomin Xie, Yudong Chen, and Weina Wang. Achieving exponential asymptotic optimality in average-reward restless bandits without global attractor assumption. *arXiv preprint arXiv:2405.17882*, 2024a.

Yige Hong, Qiaomin Xie, Yudong Chen, and Weina Wang. Unichain and aperiodicity are sufficient for asymptotic optimality of average-reward restless bandits. *arXiv preprint arXiv:2402.05689*, 2024b.

Weici Hu and Peter Frazier. An asymptotically optimal index policy for finite-horizon restless bandits. *arXiv preprint arXiv:1707.00205*, 2017.

Yujia Jin and Aaron Sidford. Efficiently Solving MDPs with Stochastic Mirror Descent, August 2020. URL http://arxiv.org/abs/2008.12776. arXiv:2008.12776.

Yujia Jin and Aaron Sidford. Towards tight bounds on the sample complexity of average-reward mdps. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5055–5064. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jin21b.html.

Young Hun Jung, Marc Abeille, and Ambuj Tewari. Thompson sampling in non-episodic restless bandits. *arXiv preprint arXiv:1910.05654*, 2019.

Michael Kearns and Satinder Singh. Finite-Sample Convergence Rates for Q-Learning and Indirect Algorithms. In *Advances in Neural Information Processing Systems*, volume 11. MIT Press, 1998. URL https://proceedings.neurips.cc/paper/1998/hash/99adff456950dd9629a5260c4de21858-Abstract.html.

Jackson A Killian, Arpita Biswas, Sanket Shah, and Milind Tambe. Q-learning lagrange policies for multi-action restless bandits. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 871–881, 2021.

Jackson A Killian, Lily Xu, Arpita Biswas, and Milind Tambe. Restless and uncertain: Robust policies for restless bandits via deep multi-agent reinforcement learning. In *Uncertainty in Artificial Intelligence*, pages 990–1000. PMLR, 2022.

Jongmin Lee, Mario Bravo, and Roberto Cominetti. Near-optimal sample complexity for mdps via anchoring. *arXiv preprint arXiv:2502.04477*, 2025.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the Sample Size Barrier in Model-Based Reinforcement Learning with a Generative Model. In *Advances in Neural Information Processing Systems*, volume 33, pages 12861–12872. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/96ea64f3a1aa2fd00c72faacf0cb8ac9-Abstract.html.

Tianjiao Li, Feiyang Wu, and Guanghui Lan. Stochastic First-Order Methods for Average-Reward Markov Decision Processes. *Mathematics of Operations Research*, December 2024. ISSN 0364-765X. doi: 10.1287/moor.2022.0241. URL https://pubsonline.informs.org/doi/full/10.1287/moor.2022.0241. Publisher: INFORMS.

Haoyang Liu, Keqin Liu, and Qing Zhao. Logarithmic weak regret of non-bayesian restless multi-armed bandit. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1968–1971. IEEE, 2011.

Nicolas Meuleau, Milos Hauskrecht, Kee-Eung Kim, Leonid Peshkin, Leslie Pack Kaelbling, Thomas L Dean, and Craig Boutilier. Solving very large weakly coupled markov decision processes. *AAAI/IAAI*, 8:2, 1998.

Carl D. Meyer, Jr. The Role of the Group Generalized Inverse in the Theory of Finite Markov Chains. *SIAM Review*, 17(3):443–464, July 1975. ISSN 0036-1445. doi: 10.1137/1017044. URL https://epubs.siam.org/doi/10.1137/1017044. Publisher: Society for Industrial and Applied Mathematics.

Khaled Nakhleh, Santosh Ganji, Ping-Chun Hsieh, I Hou, Srinivas Shakkottai, et al. Neurwin: Neural whittle index network for restless bandits via deep rl. *Advances in Neural Information Processing Systems*, 34:828–839, 2021.

Gergely Neu and Nneka Okolo. Dealing with unbounded gradients in stochastic saddle-point optimization, June 2024. URL http://arxiv.org/abs/2402.13903. arXiv:2402.13903 [cs, math, stat] version: 2.

Ronald Ortner, Daniil Ryabko, Peter Auer, and Rémi Munos. Regret bounds for restless markov bandits. In *International conference on algorithmic learning theory*, pages 214–228. Springer, 2012.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Francisco Robledo, Vivek Borkar, Urtzi Ayesta, and Konstantin Avrachenkov. Qwi: Q-learning with whittle index. *ACM SIGMETRICS Performance Evaluation Review*, 49(2):47–50, 2022.

Francisco Robledo, Urtzi Ayesta, and Konstantin Avrachenkov. Deep reinforcement learning for weakly coupled mdp's with continuous actions. In *International Conference on Analytical and Stochastic Modeling Techniques and Applications*, pages 67–80. Springer, 2024.

Cem Tekin and Mingyan Liu. Online learning of rested and restless bandits. *IEEE Transactions on Information Theory*, 58(8):5588–5611, 2012.

Adrienne Tuynman, Rémy Degenne, and Emilie Kaufmann. Finding good policies in average-reward Markov Decision Processes without prior knowledge, May 2024. URL http://arxiv.org/abs/2405.17108. arXiv:2405.17108 [cs].

Ina Maria Verloop. Asymptotically optimal priority policies for indexable and nonindexable restless bandits. 2016.

Sofía S Villar. Indexability and optimal index policies for a class of reinitialising restless bandits. *Probability in the engineering and informational sciences*, 30(1):1–23, 2016.

Jinghan Wang, Mengdi Wang, and Lin F. Yang. Near Sample-Optimal Reduction-based Policy Learning for Average Reward MDP, December 2022. URL http://arxiv.org/abs/2212.00603. arXiv:2212.00603 [cs].

Shengbo Wang, Jose Blanchet, and Peter Glynn. Optimal Sample Complexity for Average Reward Markov Decision Processes. October 2023. URL https://openreview.net/forum?id=jOm5p3q7c7.

Shengbo Wang, Jose Blanchet, and Peter Glynn. Optimal sample complexity for average reward markov decision processes, 2024a. URL https://arxiv.org/abs/2310.08833.

Shengbo Wang, Jose Blanchet, and Peter Glynn. Optimal Sample Complexity for Average Reward Markov Decision Processes, February 2024b. URL http://arxiv.org/abs/2310.08833. arXiv:2310.08833.

Siwei Wang, Longbo Huang, and John Lui. Restless-ucb, an efficient and low-complexity algorithm for online restless bandits. *Advances in Neural Information Processing Systems*, 33: 11878–11889, 2020.

Richard R Weber and Gideon Weiss. On an index policy for restless bandits. *Journal of applied probability*, 27(3):637–648, 1990.

Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, 25(A):287–298, 1988.

Guojun Xiong and Jian Li. Finite-time analysis of whittle index based q-learning for restless multi-armed bandits with neural network function approximation. *Advances in Neural Information Processing Systems*, 36:29048–29073, 2023.

Guojun Xiong, Shufan Wang, and Jian Li. Learning infinite-horizon average-reward restless multi-action bandits via index awareness. *Advances in Neural Information Processing Systems*, 35: 17911–17925, 2022.

Guojun Xiong, Shufan Wang, Jian Li, and Rahul Singh. Whittle index-based q-learning for wireless edge caching with linear function approximation. *IEEE/ACM Transactions on Networking*, 32 (5):4286–4301, 2024.

Chen Yan, Weina Wang, and Lei Ying. Achieving "$\tilde{o}(1/n)$" optimality gap in restless bandits through gaussian approximation. *arXiv preprint arXiv:2410.15003*, 2024.

Zhe Yu, Yunjian Xu, and Lang Tong. Deadline scheduling as restless bandits. *IEEE Transactions on Automatic Control*, 63(8):2343–2358, 2018.

Xiangcheng Zhang, Yige Hong, and Weina Wang. Projection-based lyapunov method for fully heterogeneous weakly-coupled mdps. *arXiv preprint arXiv:2502.06072*, 2025.

Xiangyu Zhang and Peter I. Frazier. Restless Bandits with Many Arms: Beating the Central Limit Theorem, July 2021. URL http://arxiv.org/abs/2107.11911. arXiv:2107.11911 [cs, math].

Xiangyu Zhang and Peter I. Frazier. Near-optimality for infinite-horizon restless bandits with many arms, March 2022. URL http://arxiv.org/abs/2203.15853. arXiv:2203.15853 [cs, math].

Jiahong Zhou, Shunhui Mao, Guoliang Yang, Bo Tang, Qianlong Xie, Lebin Lin, Xingxing Wang, and Dong Wang. Rl-mpca: A reinforcement learning based multi-phase computation allocation approach for recommender systems. In *Proceedings of the ACM Web Conference 2023*, pages 3214–3224, 2023.

Matthew Zurek and Yudong Chen. The Plug-in Approach for Average-Reward and Discounted MDPs: Optimal Sample Complexity Analysis, October 2024. URL http://arxiv.org/abs/2410.07616. arXiv:2410.07616 [cs].

Matthew Zurek and Yudong Chen. Span-agnostic optimal sample complexity and oracle inequalities for average-reward rl. *arXiv preprint arXiv:2502.11238*, 2025a.

Matthew Zurek and Yudong Chen. Span-Based Optimal Sample Complexity for Weakly Communicating and General Average Reward MDPs. *Advances in Neural Information Processing Systems*, 37:33455–33504, January 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3acbe9dc3a1e8d48a57b16e9aef91879-Abstract-Conference.html.