STATISTICAL SCIENCE

Volume 40, Number 4 November 2025

Special Issue on Reinforcement Learning

Statistical Science **[ISSN 0883-4237 (print); ISSN 2168-8745 (online)]**, Volume 40, Number 4, November 2025. Published quarterly by the Institute of Mathematical Statistics, 9760 Smith Road, Waite Hill, Ohio 44094, USA. Periodicals postage paid at Cleveland, Ohio and at additional mailing offices.

POSTMASTER: Send address changes to *Statistical Science*, Institute of Mathematical Statistics, Dues and Subscriptions Office, PO Box 729, Middletown, Maryland 21769, USA.

Statistical Science **Volume 40, Number 4 (515–691) November 2025**

Volume 40 Number 4 November 2025

Special Issue on Reinforcement Learning

Statistical Science
2025, Vol. 40, No. 4, 515–516
https://doi.org/10.1214/25-STS1020

Introduction to the Special Issue on Reinforcement Learning

Nan Jiang and Ambuj Tewari

REFERENCES

[1] BERTSEKAS, D. P. and TSITSIKLIS, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA.

[2] SILVER, D., HUANG, A., MADDISON, C. J., GUEZ, A., SIFRE, L., DRIESSCHE, G. V. D., SCHRITTWIESER, J., ANTONOGLOU, I., PANNEERSHELVAM, V. et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature* **529** 484–489. https://doi.org/10.1038/nature16961

[3] SUTTON, R. S. and BARTO, A. G. (1998). *Reinforcement Learning*: *An Introduction*, MIT Press, Cambridge, MA.

[4] SUTTON, R. S. and BARTO, A. G. (2018). *Reinforcement Learning*: *An Introduction*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3889951

Nan Jiang is Associate Professor, Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, Illinois 61801, USA (e-mail: nanjiang@illinois.edu). Ambuj Tewari is Professor, Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109, USA (e-mail: tewaria@umich.edu).

Statistical Science
2025, Vol. 40, No. 4, 517–545
https://doi.org/10.1214/25-STS1001

Sample-Based Planning and Learning with Function Approximation

Tor Lattimore and Csaba Szepesvári

Abstract. We give a short tutorial on sample-based planning and learning with function approximation for reinforcement learning. The focus is on explaining the standard setups and algorithm design principles. The core algorithmic ideas explained are approximate backwards induction, the (Bellman) eluder dimension and the optimism principle.

Key words and phrases: Reinforcement learning, planning.

REFERENCES

[1] Abbasi-Yadkori, Y., Bartlett, P., Bhatia, K., Lazic, N., Szepesvari, C. and Weisz, G. (2019). Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning* 3692–3702. PMLR.

[2] Abbasi-Yadkori, Y., Pál, D. and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems* 2312–2320. Curran Associates, Red Hook.

[3] Abbasi-Yadkori, Y. and Szepesvári, C. (2011). Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the* 24*th Annual Conference on Learning Theory* 1–26. JMLR Workshop and Conference Proceedings.

[4] Agarwal, A., Henaff, M., Kakade, S. and Sun, W. (2020). Pc-pg: Policy cover directed exploration for provable policy gradient learning. *Adv. Neural Inf. Process. Syst.* **33** 13399–13412.

[5] Agarwal, A., Jiang, N., Kakade, S. M. and Sun, W. (2022). Reinforcement Learning: Theory and Algorithms.

[6] Agarwal, A., Kakade, S. M., Lee, J. D. and Mahajan, G. (2021). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *J. Mach. Learn. Res.* **22** Paper No. 98, 76. MR4279749

[7] Antos, A., Szepesvári, C. and Munos, R. (2006). Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. In *Learning Theory. Lecture Notes in Computer Science* **4005** 574–588. Springer, Berlin. MR2280632 https://doi.org/10.1007/11776420_42

[8] Bagnell, J., Kakade, S. M., Schneider, J. and Ng, A. (2003). Policy search by dynamic programming. In *Advances in Neural Information Processing Systems* (S. Thrun, L. Saul and B. Schölkopf, eds.) **16**. MIT Press, Cambridge.

[9] Bastani, H., Bayati, M. and Khosravi, K. (2021). Mostly exploration-free algorithms for contextual bandits. *Manag. Sci.* **67** 1329–1349.

[10] Bittanti, S. and Campi, M. C. (2006). Adaptive control of linear time invariant systems: The "bet on the best" principle. *Commun. Inf. Syst.* **6** 299–320. MR2346930 https://doi.org/10.4310/cis.2006.v6.n4.a3

[11] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities*: *A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

[12] Boutilier, C., Dearden, R., Goldszmidt, M. et al. (1995). Exploiting structure in policy construction. In *IJCAI* **14** 1104–1113.

[13] Brukhim, N., Dudik, M., Pacchiano, A. and Schapire, R. E. (2023). A unified model and dimension for interactive estimation. *Adv. Neural Inf. Process. Syst.* **36** 64589–64617.

[14] Campi, M. C. (1997). Achieving optimality in adaptive control: The "bet on the best" approach. In *Proceedings of the* 36*th IEEE Conference on Decision and Control* **5** 4671–4676.

[15] Chow, C.-S. and Tsitsiklis, J. N. (1989). The complexity of dynamic programming. *J. Complexity* **5** 466–488. MR1028908 https://doi.org/10.1016/0885-064X(89)90021-6

[16] Dann, C., Lattimore, T. and Brunskill, E. (2017). Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. *Adv. Neural Inf. Process. Syst.* **30**.

[17] Dann, C., Mohri, M., Zhang, T. and Zimmert, J. (2021). A provably efficient model-free posterior sampling method for episodic reinforcement learning. *Adv. Neural Inf. Process. Syst.* **34** 12040–12051.

[18] Du, S., Kakade, S., Lee, J., Lovett, S., Mahajan, G., Sun, W. and Wang, R. (2021). Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning* 2826–2836. PMLR.

[19] Du, S. S., Kakade, S. M., Wang, R. and Yang, L. F. (2019). Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*.

[20] Foster, D. J., Kakade, S., Qian, J. and Rakhlin, A. (2021). The statistical complexity of interactive decision making. arXiv preprint. Available at arXiv:2112.13487.

[21] Foster, D. J., Rakhlin, A., Simchi-Levi, D. and Xu, Y. (2020). Instance-dependent complexity of contextual bandits and reinforcement learning: a disagreement-based perspective. arXiv preprint. Available at arXiv:2010.03104.

[22] Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J. and Schapire, R. E. (2017). Contextual decision pro-

Tor Lattimore is a research scientist, Google DeepMind, London, United Kingdom (e-mail: lattimore@google.com). Csaba Szepesvári is research scientist, Google DeepMind, Edmonton, Ontario, Canada and a professor, the Department of Computing Science, University of Alberta, Edmonton, Ontario, Canada (e-mail: szepi@google.com).

cesses with low Bellman rank are PAC-learnable. In *International Conference on Machine Learning* 1704–1713. PMLR.

[23] JIN, C., LIU, Q. and MIRYOOSEFI, S. (2021). Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. *Adv. Neural Inf. Process. Syst.* **34** 13406–13418.

[24] JIN, C., YANG, Z., WANG, Z. and JORDAN, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory* 2137–2143. PMLR.

[25] KEARNS, M. and SINGH, S. (2002). Near-optimal reinforcement learning in polynomial time. *Mach. Learn.* **49** 209–232.

[26] KIEFER, J. and WOLFOWITZ, J. (1960). The equivalence of two extremum problems. *Canad. J. Math.* **12** 363–366. MR0117842 https://doi.org/10.4153/CJM-1960-030-4

[27] KUMAR, P. R. and BECKER, A. (1982). A new family of optimal adaptive controllers for Markov chains. *IEEE Trans. Automat. Control* **27** 137–146. MR0673081 https://doi.org/10.1109/TAC.1982.1102878

[28] LAI, T. L. and ROBBINS, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.* **6** 4–22. MR0776826 https://doi.org/10.1016/0196-8858(85)90002-8

[29] LATTIMORE, T. and SZEPESVÁRI, C. (2019). Learning with Good Feature Representations in Bandits and in RL with a Generative Model. Available at arXiv:1911.07676.

[30] LATTIMORE, T. and SZEPESVÁRI, C. (2020). *Bandit Algorithms*. Cambridge Univ. Press, Cambridge.

[31] LI, G., KAMATH, P., FOSTER, D. J. and SREBRO, N. (2022). Understanding the eluder dimension. In *Proceedings of the* 36*th International Conference on Neural Information Processing Systems*. *NIPS'*22.

[32] LI, S. and YANG, L. (2023). Horizon-free learning for Markov decision processes and games: Stochastically bounded rewards and improved bounds. In *International Conference on Machine Learning* 20221–20252. PMLR.

[33] LI, Y., WANG, R. and YANG, L. F. (2022). Settling the horizon-dependence of sample complexity in reinforcement learning. In 2021 *IEEE* 62*nd Annual Symposium on Foundations of Computer Science—FOCS* 2021 965–976. IEEE Comput. Soc., Los Alamitos, CA. MR4399750 https://doi.org/10.1109/FOCS52979.2021.00097

[34] LIU, B., XIE, Q. and MODIANO, E. (2019). Reinforcement learning for optimal control of queueing systems. In 2019 57*th Annual Allerton Conference on Communication, Control, and Computing* (*Allerton*) 663–670. IEEE.

[35] LU, X. (2020). *Information-Directed Sampling for Reinforcement Learning*. Stanford Univ. Press, Stanford.

[36] LU, X., VAN ROY, B., DWARACHERLA, V., IBRAHIMI, M., OSBAND, I., WEN, Z. et al. (2023). Reinforcement learning, bit by bit. *Found. Trends Mach. Learn.* **16** 733–865.

[37] MUNOS, R. and SZEPESVÁRI, C. (2008). Finite-time bounds for fitted value iteration. *J. Mach. Learn. Res.* **9** 815–857. MR2417255

[38] OSBAND, I. and VAN ROY, B. (2014). Model-based reinforcement learning and the eluder dimension. *Adv. Neural Inf. Process. Syst.* **27**.

[39] RUSSO, D. and VAN ROY, B. (2013). Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems* 2256–2264. Curran Associates, Red Hook.

[40] RUST, J. (1997). Using randomization to break the curse of dimensionality. *Econometrica* **65** 487–516. MR1445620 https://doi.org/10.2307/2171751

[41] SCHERRER, B. (2013). On the Performance Bounds of some Policy Search Dynamic Programming Algorithms. CoRR.

[42] SHALEV-SHWARTZ, S. and BEN-DAVID, S. (2009). *Understanding Machine Learning. From Theory to Algorithms*. Cambridge Univ. Press, Cambridge.

[43] SHARIFF, R. and SZEPESVÁRI, C. (2020). Efficient planning in large MDPs with weak linear function approximation. *Adv. Neural Inf. Process. Syst.* **33** 19163–19174.

[44] SUTTON, R. S. and BARTO, A. G. (2018). *Reinforcement Learning*: *An Introduction*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3889951

[45] SZEPESVÁRI, C. (2001). Efficient approximate planning in continuous space Markovian decision problems. *AI Commun.* **14** 163–176. MR1866295

[46] SZEPESVÁRI, C. (2022). *Algorithms for Reinforcement Learning*. *Synthesis Lectures on Artificial Intelligence and Machine Learning* **9**. Springer, Cham. Reprint of the 2010 original. MR4647095 https://doi.org/10.1007/978-3-031-01551-9

[47] TODD, M. J. (2016). *Minimum-Volume Ellipsoids*: *Theory and Algorithms*. *MOS-SIAM Series on Optimization* **23**. SIAM, Philadelphia, PA. MR3522166 https://doi.org/10.1137/1.9781611974386.ch1

[48] WANG, R., DU, S. S., YANG, L. F. and KAKADE, S. M. (2020). Is long horizon reinforcement learning more difficult than short horizon reinforcement learning? arXiv preprint. Available at arXiv:2005.00527.

[49] WANG, R., SALAKHUTDINOV, R. and YANG, L. F. (2020). Provably efficient reinforcement learning with general value function approximation. arXiv preprint. Available at arXiv:2005.10804.

[50] WEISZ, G., AMORTILA, P., JANZER, B., ABBASI-YADKORI, Y., JIANG, N. and SZEPESVÁRI, C. (2021). On query-efficient planning in MDPs under linear realizability of the optimal state-value function. In *Conference on Learning Theory* 4355–4385. PMLR.

[51] WEISZ, G., AMORTILA, P. and SZEPESVÁRI, C. (2021). Exponential lower bounds for planning in MDPs with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*. *Proc. Mach. Learn. Res.* (*PMLR*) 1237–1264. PMLR.

[52] WEISZ, G., GYÖRGY, A., KOZUNO, T. and SZEPESVÁRI, C. (2022). Confident approximate policy iteration for efficient local planning in q^π-realizable MDPs. *Adv. Neural Inf. Process. Syst.* **35** 25547–25559.

[53] WEISZ, G., GYÖRGY, A. and SZEPESVARI, C. (2023). Online RL in linearly q^π-realizable MDPs is as easy as in linear MDPs if you learn what to ignore. In *Advances in Neural Information Processing Systems* (A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt and S. Levine, eds.) **36** 59172–59205. Curran Associates, Red Hook.

[54] WEISZ, G., SZEPESVÁRI, C. and GYÖRGY, A. (2022). TensorPlan and the few actions lower bound for planning in MDPs under linear realizability of optimal value functions. In *International Conference on Algorithmic Learning Theory* 1097–1137. PMLR.

[55] WEN, Z. and VAN ROY, B. (2017). Efficient reinforcement learning in deterministic systems with value function generalization. *Math. Oper. Res.* **42** 762–782. MR3685265 https://doi.org/10.1287/moor.2016.0826

[56] YIN, D., HAO, B., ABBASI-YADKORI, Y., LAZIĆ, N. and SZEPESVÁRI, C. (2022). Efficient local planning with linear function approximation. In *International Conference on Algorithmic Learning Theory* 1165–1192. PMLR.

[57] ZANETTE, A. and BRUNSKILL, E. (2019). Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning* 7304–7312. PMLR.

[58] ZANETTE, A., LAZARIC, A., KOCHENDERFER, M. and BRUNSKILL, E. (2020). Learning near optimal policies with low inherent Bellman error. In *International Conference on Machine Learning* 10978–10989. PMLR.

[59] ZANETTE, A., LAZARIC, A., KOCHENDERFER, M. J. and BRUNSKILL, E. (2019). Limiting extrapolation in linear approximate value iteration. *Adv. Neural Inf. Process. Syst.* **32**.

[60] ZHANG, X., SONG, Y., UEHARA, M., WANG, M., AGARWAL, A. and SUN, W. (2022). Efficient reinforcement learning in block mdps: A model-free representation learning approach. In *International Conference on Machine Learning* 26517–26547. PMLR.

[61] ZHANG, Z., JI, X. and DU, S. (2022). Horizon-free reinforcement learning in polynomial time: The power of stationary policies. In *Conference on Learning Theory* 3858–3904. PMLR.

[62] ZIMMERT, J. and LATTIMORE, T. (2022). Return of the bias: Almost minimax optimal high probability bounds for adversarial linear bandits. In *Proceedings of Thirty Fifth Conference on Learning Theory* (P.-L. Loh and M. Raginsky, eds.). *Proceedings of Machine Learning Research* **178** 3285–3312. PMLR.

Statistical Science
2025, Vol. 40, No. 4, 546–569
https://doi.org/10.1214/25-STS1008

On the Statistical Complexity for Offline and Low-Adaptive Reinforcement Learning with Structures

Ming Yin, Mengdi Wang and Yu-Xiang Wang

Abstract. This article reviews the recent advances on the statistical foundation of reinforcement learning (RL) in the offline and low-adaptive settings. We will start by arguing why offline RL is the appropriate model for almost any real-life ML problems, even if they have nothing to do with the recent AI breakthroughs that use RL. Then we will zoom into two fundamental problems of offline RL: offline policy evaluation (OPE) and offline policy learning (OPL). It may be surprising to people that tight bounds for these problems were not known even for tabular and linear cases until recently. We delineate the differences between worst-case minimax bounds and instance-dependent bounds. We also cover key algorithmic ideas and proof techniques behind near-optimal instance-dependent methods in OPE and OPL. Finally, we discuss the limitations of offline RL and review a burgeoning problem of *low-adaptive exploration* which addresses these limitations by providing a sweet middle ground between offline and online RL.

Key words and phrases: Sample complexity, offline reinforcement learning, low-adaptive exploration.

REFERENCES

[1] ABBASI-YADKORI, Y., PÁL, D. and SZEPESVÁRI, C. (2011). Improved algorithms for linear stochastic bandits. *Adv. Neural Inf. Process. Syst.* **24**.

[2] AGARWAL, A., KAKADE, S. and YANG, L. F. (2020). Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory* 67–83. PMLR.

[3] ANAHTARCI, B., KARIKSIZ, C. D. and SALDI, N. (2023). Q-learning in regularized mean-field games. *Dyn. Games Appl.* **13** 89–117. MR4550415 https://doi.org/10.1007/s13235-022-00450-2

[4] ANITESCU, M. (2000). Degenerate nonlinear programming with a quadratic growth condition. *SIAM J. Optim.* **10** 1116–1135. MR1777083 https://doi.org/10.1137/S1052623499359178

[5] ANTOS, A., SZEPESVÁRI, C. and MUNOS, R. (2007). Fitted q-iteration in continuous action-space mdps. *Adv. Neural Inf. Process. Syst.* **20**.

[6] ASADI, K., LIU, Y., SABACH, S., YIN, M. and FAKOOR, R. (2024). Learning the target network in function space. In *International Conference on Machine Learning*.

[7] AUER, P. (2002). Finite-time analysis of the multiarmed bandit problem.

[8] AZAR, M. G., MUNOS, R. and KAPPEN, H. J. (2013). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Mach. Learn.* **91** 325–349. MR3064431 https://doi.org/10.1007/s10994-013-5368-1

[9] AZAR, M. G., OSBAND, I. and MUNOS, R. (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning* 263–272. PMLR.

[10] BAI, Y., JONES, A., NDOUSSE, K., ASKELL, A., CHEN, A., DASSARMA, N., DRAIN, D., FORT, S., GANGULI, D. et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. Technical Report.

[11] BAI, Y., XIE, T., JIANG, N. and WANG, Y.-X. (2019). Provably efficient q-learning with low switching cost. *Adv. Neural Inf. Process. Syst.* **32**.

[12] BELLMAN, R. (1966). Dynamic programming. *Science* **153** 34–37.

[13] CANDÈS, E. J. (2006). Modern statistical estimation via oracle inequalities. *Acta Numer.* **15** 257–325. MR2269743 https://doi.org/10.1017/S0962492906230010

[14] CESA-BIANCHI, N., DEKEL, O. and SHAMIR, O. (2013). Online learning with switching costs and other adaptive adversaries. *Adv. Neural Inf. Process. Syst.* **26**.

[15] CHEN, J. and JIANG, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International Con-*

Ming Yin is a Postdoc at the Department of Electrical and Computer Engineering at Princeton University, Princeton, New Jersey 08544, USA (e-mail: my0049@princeton.edu). Mengdi Wang is an Associate Professor at the Department of Electrical and Computer Engineering at Princeton University, Princeton, New Jersey 08544, USA (e-mail: mengdiw@princeton.edu). Yu-Xiang Wang is an Associate Professor at the Halıcıoğlu Data Science Institute at University of California, San Diego, La Jolla, California 92093, USA (e-mail: yuxiangw@ucsd.edu).

ference on Machine Learning 1042–1051. PMLR.

[16] CHRISTIANO, P. F., LEIKE, J., BROWN, T., MARTIC, M., LEGG, S. and AMODEI, D. (2017). Deep reinforcement learning from human preferences. *Adv. Neural Inf. Process. Syst.* **30**.

[17] DANN, C., MARINOV, T. V., MOHRI, M. and ZIMMERT, J. (2021). Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. *Adv. Neural Inf. Process. Syst.* **34** 1–12.

[18] DI, Q., ZHAO, H., HE, J. and GU, Q. (2023). Pessimistic nonlinear least-squares value iteration for offline reinforcement learning. arXiv preprint. Available at arXiv:2310.01380.

[19] DUAN, Y., JIA, Z. and WANG, M. (2020). Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning* 2701–2709. PMLR.

[20] DUDÍK, M., LANGFORD, J. and LI, L. (2011). Doubly robust policy evaluation and learning. arXiv preprint. Available at arXiv:1103.4601.

[21] ERNST, D., GEURTS, P. and WEHENKEL, L. (2005). Tree-based batch mode reinforcement learning. *J. Mach. Learn. Res.* **6** 503–556. MR2249830

[22] FAN, J., WANG, Z., XIE, Y. and YANG, Z. (2020). A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control* 486–489. PMLR.

[23] FAWZI, A., BALOG, M., HUANG, A., HUBERT, T., ROMERA-PAREDES, B., BAREKATAIN, M., NOVIKOV, A., RUIZ, F. J. R., SCHRITTWIESER, J. et al. (2022). Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature* **610** 47–53.

[24] FISHER, R. A. (1925). Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society* **22** 700–725. Cambridge Univ. Press, Cambridge.

[25] FUJIMOTO, S., HOOF, H. and MEGER, D. (2018). Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning* 1587–1596. PMLR.

[26] GAO, M., XIE, T., DU, S. S. and YANG, L. F. (2021). A provably efficient algorithm for linear Markov decision process with low switching cost. arXiv preprint. Available at arXiv:2101.00494.

[27] GAO, Z., HAN, Y., REN, Z. and ZHOU, Z. (2019). Batched multi-armed bandits problem. *Adv. Neural Inf. Process. Syst.* **32**.

[28] GELADA, C. and BELLEMARE, M. G. (2019). Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence* **33** 3647–3655.

[29] GILBOA, I. and SCHMEIDLER, D. (1989). Maxmin expected utility with nonunique prior. *J. Math. Econom.* **18** 141–153. MR1000102 https://doi.org/10.1016/0304-4068(89)90018-9

[30] GORDON, G. J. (1999). Approximate Solutions to Markov Decision Processes. Carnegie Mellon Univ.

[31] HAIDER, M., YIN, M., ZHANG, M., GUPTA, A., ZHU, J. and WANG, Y.-X. (2024). Networkgym: Reinforcement learning environments for multi-access traffic management in network simulation. Advances in Neural Information Processing Systems (NeurIPS 2024)-Dataset and Benchmark.

[32] HALLAK, A. and MANNOR, S. (2017). Consistent on-line off-policy evaluation. In *International Conference on Machine Learning* 1372–1383. PMLR.

[33] HAO, B., JI, X., DUAN, Y., LU, H., SZEPESVARI, C. and WANG, M. (2021). Bootstrapping fitted q-evaluation for off-policy inference. In *International Conference on Machine Learning* 4074–4084. PMLR.

[34] HASANBEIG, M., ABATE, A. and KROENING, D. (2018). Logically-constrained neural fitted q-iteration. arXiv preprint. Available at arXiv:1809.07823.

[35] HE, J., ZHOU, D. and GU, Q. (2021). Nearly minimax optimal reinforcement learning for discounted mdps. *Adv. Neural Inf. Process. Syst.* **34** 22288–22300.

[36] HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. MR1995826 https://doi.org/10.1111/1468-0262.00442

[37] HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460

[38] HUANG, J., CHEN, J., ZHAO, L., QIN, T., JIANG, N. and LIU, T.-Y. (2022). Towards deployment-efficient reinforcement learning: Lower bound and optimality. In *International Conference on Learning Representations*.

[39] HÜLLERMEIER, E. and WAEGEMAN, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Mach. Learn.* **110** 457–506. MR4234924 https://doi.org/10.1007/s10994-021-05946-3

[40] JIANG, N. and LI, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning* 652–661. PMLR.

[41] JIANG, N. and XIE, T. (2024). Offline reinforcement learning in large state spaces: Algorithms and guarantees. *Statist. Sci.*

[42] JIANG, N. and XIE, T. (2024). Offline reinforcement learning in large state spaces: Algorithms and guarantees.

[43] JIN, Y., YANG, Z. and WANG, Z. (2021). Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning* 5084–5096. PMLR.

[44] KALLUS, N. and UEHARA, M. (2020). Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *J. Mach. Learn. Res.* **21** Paper No. 167, 63. MR4209453

[45] KALLUS, N. and UEHARA, M. (2022). Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Oper. Res.* **70** 3282–3302. MR4538517

[46] KEARNS, M. and SINGH, S. (2002). Near-optimal reinforcement learning in polynomial time. *Mach. Learn.* **49** 209–232.

[47] KOSOROK, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. *Springer Series in Statistics*. Springer, New York. MR2724368 https://doi.org/10.1007/978-0-387-74978-5

[48] KOSTRIKOV, I., FERGUS, R., TOMPSON, J. and NACHUM, O. (2021). Offline reinforcement learning with Fisher divergence critic regularization. In *International Conference on Machine Learning* 5774–5783. PMLR.

[49] KUMAR, A., ZHOU, A., TUCKER, G. and LEVINE, S. (2020). Conservative q-learning for offline reinforcement learning. *Adv. Neural Inf. Process. Syst.* **33** 1179–1191.

[50] LAFFERTY, J., LIU, H. and WASSERMAN, L. (2008). *Minimax Theory*. *Lecture Notes on Statistical Machine Learning*. Available at http://www.stat.cmu.edu/~larry/=sml/Minimax.pdf.

[51] LATTIMORE, T. and SZEPESVÁRI, C. (2020). *Bandit Algorithms*. Cambridge Univ. Press, Cambridge.

[52] LE, H., VOLOSHIN, C. and YUE, Y. (2019). Batch policy learning under constraints. In *International Conference on Machine Learning* 3703–3712. PMLR.

[53] LEVINE, S., KUMAR, A., TUCKER, G. and FU, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv preprint. Available at arXiv:2005.01643.

[54] LI, G., WEI, Y., CHI, Y., GU, Y. and CHEN, Y. (2020). Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Adv. Neural Inf. Process. Syst.* **33** 12861–12872.

[55] LI, J., ZHANG, E., YIN, M., BAI, Q., WANG, Y.-X. and WANG, W. Y. (2023). Offline reinforcement learning with closed-form policy improvement operators. In *International Conference on Machine Learning* 20485–20528. PMLR.

[56] LI, L., CHU, W., LANGFORD, J. and WANG, X. (2011). Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* 297–306.

[57] LIU, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. *Springer Series in Statistics*. Springer, New York. MR1842342

[58] LIU, Q., LI, L., TANG, Z. and ZHOU, D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Adv. Neural Inf. Process. Syst.* **31**.

[59] LIU, Y., SWAMINATHAN, A., AGARWAL, A. and BRUNSKILL, E. (2020). Off-policy policy gradient with stationary distribution correction. In *Uncertainty in Artificial Intelligence* 1180–1190. PMLR.

[60] LIU, Y., SWAMINATHAN, A., AGARWAL, A. and BRUNSKILL, E. (2020). Provably good batch off-policy reinforcement learning without great exploration. *Adv. Neural Inf. Process. Syst.* **33** 1264–1274.

[61] LIZOTTE, D. J., BOWLING, M. and MURPHY, S. A. (2012). Linear fitted-Q iteration with multiple reward functions. *J. Mach. Learn. Res.* **13** 3253–3295. MR3005887

[62] LYU, J., MA, X., LI, X. and LU, Z. (2022). Mildly conservative q-learning for offline reinforcement learning. *Adv. Neural Inf. Process. Syst.* **35** 1711–1724.

[63] MAILLARD, O.-A., MANN, T. A. and MANNOR, S. (2014). How hard is my mdp?" the distribution-norm to the rescue". *Adv. Neural Inf. Process. Syst.* **27**.

[64] MANKOWITZ, D. J., MICHI, A., ZHERNOV, A., GELMI, M., SELVI, M., PADURARU, C., LEURENT, E., IQBAL, S., LESPIAU, J.-B. et al. (2023). Faster sorting algorithms discovered using deep reinforcement learning. *Nature* **618** 257–263.

[65] MAO, H., NETRAVALI, R. and ALIZADEH, M. (2017). Neural adaptive video streaming with pensieve. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication* 197–210.

[66] MATSUSHIMA, T., FURUTA, H., MATSUO, Y., NACHUM, O. and GU, S. (2021). Deployment-efficient reinforcement learning via model-based offline optimization. In *International Conference on Learning Representations*.

[67] MCLEISH, D. L. (1974). Dependent central limit theorems and invariance principles. *Ann. Probab.* **2** 620–628. MR0358933 https://doi.org/10.1214/aop/1176996608

[68] MIN, Y., WANG, T., ZHOU, D. and GU, Q. (2021). Variance-aware off-policy evaluation with linear function approximation. *Adv. Neural Inf. Process. Syst.* **34** 7598–7610.

[69] MNIH, V., KAVUKCUOGLU, K., SILVER, D., RUSU, A. A., VENESS, J., BELLEMARE, M. G., GRAVES, A., RIEDMILLER, M., FIDJELAND, A. K. et al. (2015). Human-level control through deep reinforcement learning. *Nature* **518** 529–533.

[70] MOONEY, C. Z., DUVAL, R. D. and DUVALL, R. (1993). *Bootstrapping*: *A Nonparametric Approach to Statistical Inference*, *Vol.* 95. Sage, Thousand Oaks.

[71] MUNOS, R. and SZEPESVÁRI, C. (2008). Finite-time bounds for fitted value iteration. *J. Mach. Learn. Res.* **9** 815–857. MR2417255

[72] MURPHY, S. A., VAN DER LAAN, M. J. and ROBINS, J. M. (2001). Marginal mean models for dynamic regimes. *J. Amer. Statist. Assoc.* **96** 1410–1423. MR1946586 https://doi.org/10.1198/016214501753382327

[73] NACHUM, O., CHOW, Y., DAI, B. and LI, L. (2019). Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Adv. Neural Inf. Process. Syst.* **32**.

[74] NACHUM, O., DAI, B., KOSTRIKOV, I., CHOW, Y., LI, L. and SCHUURMANS, D. (2019). Algaedice: Policy gradient from arbitrary experience. arXiv preprint. Available at arXiv:1912.02074.

[75] NEMATI, S., GHASSEMI, M. M. and CLIFFORD, G. D. (2016). Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. In 2016 38*th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (*EMBC*) 2978–2981. IEEE Press, London.

[76] NGUYEN-TANG, T. and ARORA, R. (2024). On sample-efficient offline reinforcement learning: Data diversity, posterior sampling and beyond. *Adv. Neural Inf. Process. Syst.* **36**.

[77] NGUYEN-TANG, T., YIN, M., GUPTA, S., VENKATESH, S. and ARORA, R. (2023). On instance-dependent bounds for offline reinforcement learning with linear function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence* **37** 9310–9318.

[78] OUYANG, L., WU, J., JIANG, X., ALMEIDA, D., WAINWRIGHT, C., MISHKIN, P., ZHANG, C., AGARWAL, S., SLAMA, K. et al. (2022). Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35** 27730–27744.

[79] PERCHET, V., RIGOLLET, P., CHASSANG, S. and SNOWBERG, E. (2016). Batched bandit problems. *Ann. Statist.* **44** 660–681. MR3476613 https://doi.org/10.1214/15-AOS1381

[80] POWELL, W. B. (2007). *Approximate Dynamic Programming*: *Solving the Curses of Dimensionality*. *Wiley Series in Probability and Statistics*. Wiley-Interscience, Hoboken, NJ. MR2347698 https://doi.org/10.1002/9780470182963

[81] PRECUP, D. (2000). Eligibility traces for off-policy policy evaluation. In *Computer Science Dept. Faculty Publication Series* 80.

[82] PUTERMAN, M. L. (1990). Markov decision processes. In *Stochastic Models*. *Handbooks Oper. Res. Management Sci.* **2** 331–434. North-Holland, Amsterdam. MR1100755 https://doi.org/10.1016/S0927-0507(05)80172-0

[83] QIAO, D. and WANG, Y.-X. (2023). Near-optimal deployment efficiency in reward-free reinforcement learning with linear function approximation. In *International Conference on Learning Representations*.

[84] QIAO, D., YIN, M., MIN, M. and WANG, Y.-X. (2022). Sample-efficient reinforcement learning with loglog (t) switching cost. In *International Conference on Machine Learning* 18031–18061. PMLR.

[85] QIAO, D., YIN, M. and WANG, Y.-X. (2024). Logarithmic switching cost in reinforcement learning beyond linear mdps. ISIT-2024.

[86] RASHIDINEJAD, P., ZHU, B., MA, C., JIAO, J. and RUSSELL, S. (2022). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *IEEE Trans. Inf. Theory* **68** 8156–8196. MR4544936 https://doi.org/10.1109/tit.2022.3185139

[87] REN, T., LI, J., DAI, B., DU, S. S. and SANGHAVI, S. (2021). Nearly horizon-free offline reinforcement learning. *Adv. Neural Inf. Process. Syst.* **34** 15621–15634.

[88] RIEDMILLER, M. (2005). Neural fitted q iteration–first experiences with a data efficient neural reinforcement learning method. In *European Conference on Machine Learning* 317–328. Springer, Berlin.

[89] SCHULMAN, J., LEVINE, S., ABBEEL, P., JORDAN, M. and MORITZ, P. (2015). Trust region policy optimization. In *International Conference on Machine Learning* 1889–1897. PMLR.

[90] SILVER, D., HUANG, A., MADDISON, C. J., GUEZ, A., SIFRE, L., VAN DEN DRIESSCHE, G., SCHRITTWIESER, J., ANTONOGLOU, I., PANNEERSHELVAM, V. et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature* **529** 484–489.

[91] SILVER, D., SCHRITTWIESER, J., SIMONYAN, K., ANTONOGLOU, I., HUANG, A., GUEZ, A., HUBERT, T., BAKER, L., LAI, M. et al. (2017). Mastering the game of go without human knowledge. *Nature* **550** 354–359.

[92] SIMCHOWITZ, M. and JAMIESON, K. G. (2019). Non-asymptotic gap-dependent regret bounds for tabular mdps. *Adv. Neural Inf. Process. Syst.* **32**.

[93] STIENNON, N., OUYANG, L., WU, J., ZIEGLER, D., LOWE, R., VOSS, C., RADFORD, A., AMODEI, D. and CHRISTIANO, P. F. (2020). Learning to summarize with human feedback. *Adv. Neural Inf. Process. Syst.* **33** 3008–3021.

[94] SUTTON, R. S. and BARTO, A. G. (2018). *Reinforcement Learning: An Introduction*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3889951

[95] SZEPESVÁRI, C. and MUNOS, R. (2005). Finite time bounds for sampling based fitted value iteration. In *Proceedings of the 22nd International Conference on Machine Learning* 880–887.

[96] TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data*. *Springer Series in Statistics*. Springer, New York. MR2233926

[97] UEHARA, M., HUANG, J. and JIANG, N. (2020). Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning* 9659–9668. PMLR.

[98] UEHARA, M. and SUN, W. (2021). Pessimistic model-based offline reinforcement learning under partial coverage. arXiv preprint. Available at arXiv:2107.06226.

[99] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

[100] WAGENMAKER, A. and PACCHIANO, A. (2023). Leveraging offline data in online reinforcement learning. In *International Conference on Machine Learning* 35300–35338. PMLR.

[101] WAGENMAKER, A. J., SIMCHOWITZ, M. and JAMIESON, K. (2022). Beyond no regret: Instance-dependent pac reinforcement learning. In *Conference on Learning Theory* 358–418. PMLR.

[102] WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics: A Non-asymptotic Viewpoint*. *Cambridge Series in Statistical and Probabilistic Mathematics* **48**. Cambridge Univ. Press, Cambridge. MR3967104 https://doi.org/10.1017/9781108627771

[103] WANG, T., ZHOU, D. and GU, Q. (2021). Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. *Adv. Neural Inf. Process. Syst.* **34** 13524–13536.

[104] WANG, X., CUI, Q. and DU, S. S. (2022). On gap-dependent bounds for offline reinforcement learning. *Adv. Neural Inf. Process. Syst.* **35** 14865–14877.

[105] WANG, Y.-X., AGARWAL, A. and DUDIK, M. (2017). Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning* 3589–3597. PMLR.

[106] WU, J., BRAVERMAN, V. and YANG, L. (2022). Gap-dependent unsupervised exploration for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics* 4109–4131. PMLR.

[107] WU, Y., TUCKER, G. and NACHUM, O. (2019). Behavior regularized offline reinforcement learning. arXiv preprint. Available at arXiv:1911.11361.

[108] XIAO, C., WU, Y., MEI, J., DAI, B., LATTIMORE, T., LI, L., SZEPESVARI, C. and SCHUURMANS, D. (2021). On the optimality of batch policy optimization algorithms. In *International Conference on Machine Learning* 11362–11371. PMLR.

[109] XIE, T., CHENG, C.-A., JIANG, N., MINEIRO, P. and AGARWAL, A. (2021). Bellman-consistent pessimism for offline reinforcement learning. *Adv. Neural Inf. Process. Syst.* **34** 6683–6694.

[110] XIE, T. and JIANG, N. (2020). Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence* 550–559. PMLR.

[111] XIE, T. and JIANG, N. (2021). Batch value-function approximation with only realizability. In *International Conference on Machine Learning* 11404–11413. PMLR.

[112] XIE, T., JIANG, N., WANG, H., XIONG, C. and BAI, Y. (2021). Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Adv. Neural Inf. Process. Syst.* **34** 27395–27407.

[113] XIE, T., MA, Y. and WANG, Y.-X. (2019). Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Adv. Neural Inf. Process. Syst.* **32**.

[114] XIONG, W., ZHONG, H., SHI, C., SHEN, C., WANG, L. and ZHANG, T. (2022). Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and Markov game. arXiv preprint. Available at arXiv:2205.15512.

[115] YIN, M., BAI, Y. and WANG, Y.-X. (2021). Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics* 1567–1575. PMLR.

[116] YIN, M., BAI, Y. and WANG, Y.-X. (2021). Near-optimal offline reinforcement learning via double variance reduction. *Adv. Neural Inf. Process. Syst.* **34** 7677–7688.

[117] YIN, M., DUAN, Y., WANG, M. and WANG, Y.-X. (2022). Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. arXiv preprint. Available at arXiv:2203.05804.

[118] YIN, M., WANG, M. and WANG, Y.-X. (2023). Offline reinforcement learning with differentiable function approximation is provably efficient. In *International Conference on Learning Representations*.

[119] YIN, M. and WANG, Y.-X. (2020). Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics* 3948–3958. PMLR.

[120] YIN, M. and WANG, Y.-X. (2021). Optimal uniform ope and model-based offline reinforcement learning in time-homogeneous, reward-free and task-agnostic settings. *Adv. Neural Inf. Process. Syst.* **34** 12890–12903.

[121] YIN, M. and WANG, Y.-X. (2021). Towards instance-optimal offline reinforcement learning with pessimism. *Adv. Neural Inf. Process. Syst.* **34** 4065–4078.

[122] ZANETTE, A. (2021). Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl. In *International Conference on Machine Learning* 12287–12297. PMLR.

[123] ZANETTE, A. and BRUNSKILL, E. (2019). Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning* 7304–7312. PMLR.

[124] ZHAN, W., HUANG, B., HUANG, A., JIANG, N. and LEE, J. (2022). Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory* 2730–2775. PMLR.

[125] ZHANG, R., DAI, B., LI, L. and SCHUURMANS, D. (2020). Gendice: Generalized offline estimation of stationary values. In *International Conference on Learning Representations*.

[126] ZHANG, R., ZHANG, X., NI, C. and WANG, M. (2022). Off-policy fitted q-evaluation with differentiable function approximators: Z-estimation and inference theory. In *International Conference on Machine Learning* 26713–26749. PMLR.

[127] ZHANG, Y., BAI, Y. and JIANG, N. (2023). Offline learning in Markov games with general function approximation. In *International Conference on Machine Learning* 40804–40829. PMLR.

[128] ZHANG, Z., CHEN, Y., LEE, J. and DU, S. S. (2025). Settling the sample complexity of online reinforcement learning. *J. ACM* **72** Art. 22, 63. MR4926208

[129] ZHANG, Z., DU, S. and JI, X. (2021). Near optimal reward-free reinforcement learning. In *Proceedings of the* 38*th International Conference on Machine Learning* **139** 12402–12412.

[130] ZHANG, Z., JIANG, Y., ZHOU, Y. and JI, X. (2022). Near-optimal regret bounds for multi-batch reinforcement learning. *Adv. Neural Inf. Process. Syst.* **35** 24586–24596.

[131] ZHAO, H., HE, J. and GU, Q. (2024). A nearly optimal and low-switching algorithm for reinforcement learning with general function approximation. *Adv. Neural Inf. Process. Syst.*

[132] ZHAO, X., XIA, L., ZHANG, L., DING, Z., YIN, D. and TANG, J. (2018). Deep reinforcement learning for page-wise recommendations. In *Proceedings of the* 12*th ACM Conference on Recommender Systems* 95–103.

[133] ZHOU, D., GU, Q. and SZEPESVARI, C. (2021). Nearly minimax optimal reinforcement learning for linear mixture Markov decision processes. In *Conference on Learning Theory* 4532–4576. PMLR.

Statistical Science
2025, Vol. 40, No. 4, 570–596
https://doi.org/10.1214/25-STS1000

Offline Reinforcement Learning in Large State Spaces: Algorithms and Guarantees

Nan Jiang and Tengyang Xie

Abstract. This article introduces the theory of offline reinforcement learning in large state spaces, where good policies are learned from historical data without online interactions with the environment. Key concepts introduced include expressivity assumptions on function approximation (e.g., Bellman-completeness vs. realizability) and data coverage (e.g., all-policy vs. single-policy coverage). A rich landscape of algorithms and results is described, depending on the assumptions one is willing to make and the sample and computational complexity guarantees one wishes to achieve. We also discuss open questions and connections to adjacent areas.

Key words and phrases: Offline reinforcement learning, function approximation.

REFERENCES

[1] AFSAR, M. M., CRUMP, T. and FAR, B. (2022). Reinforcement learning based recommender systems: A survey. *ACM Comput. Surv.* **55** 1–38.

[2] AGARWAL, A., KAKADE, S., KRISHNAMURTHY, A. and SUN, W. (2020). FLAMBE: Structural Complexity and Representation Learning of Low Rank MDPs. arXiv preprint. Available at arXiv:2006.10814.

[3] AGARWAL, A., KAKADE, S. M., LEE, J. D. and MAHAJAN, G. (2020). Optimality and approximation with policy gradient methods in Markov decision processes. In *Conference on Learning Theory* 64–66. PMLR.

[4] AMATO, C. (2024). (A Partial Survey of) Decentralized, Cooperative Multi-Agent Reinforcement Learning. arXiv preprint. Available at arXiv:2405.06161.

[5] AMORTILA, P., JIANG, N. and XIE, T. (2020). A variant of the Wang–Foster–kakade lower bound for the discounted setting. arXiv preprint. Available at arXiv:2011.01075.

[6] ANTOS, A., SZEPESVÁRI, C. and MUNOS, R. (2006). Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. In *Learning Theory. Lecture Notes in Computer Science* **4005** 574–588. Springer, Berlin. MR2280632 https://doi.org/10.1007/11776420_42

[7] AUER, P., CESA-BIANCHI, N. and FISCHER, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* **47** 235–256.

[8] BAI, C., WANG, L., YANG, Z., DENG, Z.-H., GARG, A., LIU, P. and WANG, Z. (2022). Pessimistic bootstrapping for uncertainty-driven offline reinforcement learning. In *International Conference on Learning Representations*.

[9] BAIRD, L. (1995). Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings* 1995 30–37. Elsevier, Amsterdam.

[10] BARRETO, A., PRECUP, D. and PINEAU, J. (2011). Reinforcement learning using kernel-based stochastic factorization. *Adv. Neural Inf. Process. Syst.* **24**.

[11] BARRETO, A. M. S., PINEAU, J. and PRECUP, D. (2014). Policy iteration based on stochastic factorization. *J. Artificial Intelligence Res.* **50** 763–803. MR3254852 https://doi.org/10.1613/jair.4301

[12] BASSEN, J., BALAJI, B., SCHAARSCHMIDT, M., THILLE, C., PAINTER, J., ZIMMARO, D., GAMES, A., FAST, E. and MITCHELL, J. C. (2020). Reinforcement learning for the adaptive scheduling of educational activities. In *Proceedings of the* 2020 *CHI Conference on Human Factors in Computing Systems* 1–12.

[13] BENNETT, A. and KALLUS, N. (2024). Proximal reinforcement learning: Efficient off-policy evaluation in partially observed Markov decision processes. *Oper. Res.* **72** 1071–1086. MR4761145

[14] BERTSEKAS, D. P. and TSITSIKLIS, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA.

[15] BHARDWAJ, M., XIE, T., BOOTS, B., JIANG, N. and CHENG, C.-A. (2023). Adversarial model for offline reinforcement learning. In *Thirty-Seventh Conference on Neural Information Processing Systems*.

[16] BRUNS-SMITH, D. and ZHOU, A. (2023). Robust fitted-q-evaluation and iteration under sequentially exogenous unobserved confounders. arXiv preprint. Available at arXiv:2302.00662.

[17] BRUNS-SMITH, D. A. (2021). Model-free and model-based policy evaluation when causality is uncertain. In *International Conference on Machine Learning* 1116–1126. PMLR.

[18] CHEN, J. and JIANG, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the* 36*th International Conference on Machine Learning* 1042–1051.

Nan Jiang is an Assistant Professor of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois 60606 USA (e-mail: nanjiang@illinois.edu). Tengyang Xie is Assistant Professor of Computer Science, University of Wisconsin-Madison, Madison, Wisconsin 53706 USA (e-mail: tx@cs.wisc.edu).

[19] CHEN, J. and JIANG, N. (2022). Offline reinforcement learning under value and density-ratio realizability: The power of gaps. In *Uncertainty in Artificial Intelligence* 378–388. PMLR.

[20] CHENG, C.-A., XIE, T., JIANG, N. and AGARWAL, A. (2022). Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning*.

[21] CUI, Q. and DU, S. S. (2022). When are offline two-player zero-sum Markov games solvable? *Adv. Neural Inf. Process. Syst.* **35** 25779–25791.

[22] CUI, Q. and DU, S. S. (2022). Provably efficient offline multi-agent reinforcement learning via strategy-wise bonus. *Adv. Neural Inf. Process. Syst.* **35** 11739–11751.

[23] DAI, B., SHAW, A., LI, L., XIAO, L., HE, N., LIU, Z., CHEN, J. and SONG, L. (2018). Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning* 1133–1142.

[24] DUAN, Y., JIA, Z. and WANG, M. (2020). Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning* 2701–2709. PMLR.

[25] ERNST, D., GEURTS, P. and WEHENKEL, L. (2005). Tree-based batch mode reinforcement learning. *J. Mach. Learn. Res.* **6** 503–556. MR2249830

[26] FAN, J., WANG, Z., XIE, Y. and YANG, Z. (2020). A theoretical analysis of deep Q-learning. In *Learning for Dynamics and Control* 486–489. PMLR.

[27] FARAHMAND, A. and SZEPESVÁRI, C. (2011). Model selection in reinforcement learning. *Mach. Learn.* **85** 299–332. MR3108236 https://doi.org/10.1007/s10994-011-5254-7

[28] FARAHMAND, A.-M., SZEPESVÁRI, C. and MUNOS, R. (2010). Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems* 568–576.

[29] FENG, Y., LI, L. and LIU, Q. (2019). A kernel loss for solving the Bellman equation. In *Advances in Neural Information Processing Systems* 15430–15441.

[30] FENG, Y., REN, T., TANG, Z. and LIU, Q. (2020). Accountable off-policy evaluation with kernel Bellman statistics. In *Proceedings of the* 37*th International Conference on Machine Learning* (*ICML*-20).

[31] FOSTER, D. J., KAKADE, S. M., QIAN, J. and RAKHLIN, A. (2021). The statistical complexity of interactive decision making. arXiv preprint. Available at arXiv:2112.13487.

[32] FUJIMOTO, S., MEGER, D. and PRECUP, D. (2019). Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning* 2052–2062.

[33] GORDON, G. J. (1995). Stable function approximation in dynamic programming. In *Proceedings of the Twelfth International Conference on Machine Learning* 261–268.

[34] GRETTON, A., SMOLA, A., HUANG, J., SCHMITTFULL, M., BORGWARDT, K. and SCHÖLKOPF, B. (2008). Covariate shift by kernel mean matching.

[35] HAFNER, D., LILLICRAP, T., BA, J. and NOROUZI, M. (2019). Dream to control: Learning behaviors by latent imagination. arXiv preprint. Available at arXiv:1912.01603.

[36] HAZAN, E. et al. (2016). Introduction to online convex optimization. *Found. Trends Optim.* **2** 157–325.

[37] HUANG, A., CHEN, J. and JIANG, N. (2023). Reinforcement learning in low-rank MDPs with density features. In *Proceedings of the* 40*th International Conference on Machine Learning*. *Proceedings of Machine Learning Research* **202** 13710–13752. PMLR.

[38] HUANG, A., GHAVAMZADEH, M., JIANG, N. and PETRIK, M. (2023). Non-adaptive online finetuning for offline reinforcement learning. In *NeurIPS* 2023 *Workshop on Generalization in Planning*.

[39] HUANG, A., ZHAN, W., XIE, T., LEE, J. D., SUN, W., KRISHNAMURTHY, A. and FOSTER, D. J. (2024). Correcting the mythos of kl-regularization: Direct alignment without overoptimization via chi-squared preference optimization. arXiv preprint. Available at arXiv:2407.13399.

[40] JIA, Z., RAKHLIN, A., SEKHARI, A. and WEI, C.-Y. (2024). Offline reinforcement learning: Role of state aggregation and trajectory data. arXiv preprint. Available at arXiv:2403.17091.

[41] JIANG, N. (2024). A Note on Loss Functions and Error Compounding in Model-based Reinforcement Learning. arXiv preprint. Available at arXiv:2404.09946.

[42] JIANG, N. and HUANG, J. (2020). Minimax value interval for off-policy evaluation and policy optimization. *Adv. Neural Inf. Process. Syst.* **33**.

[43] JIANG, N., KRISHNAMURTHY, A., AGARWAL, A., LANGFORD, J. and SCHAPIRE, R. E. (2017). Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the* 34*th International Conference on Machine Learning* **70** 1704–1713.

[44] JIANG, N. and LI, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the* 33*rd International Conference on Machine Learning* **48** 652–661.

[45] JIN, C., LIU, Q. and MIRYOOSEFI, S. (2021). Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Adv. Neural Inf. Process. Syst.* **34** 13406–13418.

[46] JIN, C., YANG, Z., WANG, Z. and JORDAN, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory* 2137–2143. PMLR.

[47] JIN, Y., YANG, Z. and WANG, Z. (2020). Is pessimism provably efficient for offline RL? arXiv preprint. Available at arXiv:2012.15085.

[48] KAKADE, S. and LANGFORD, J. (2002). Approximately optimal approximate reinforcement learning. In *Proceedings of the* 19*th International Conference on Machine Learning* **2** 267–274.

[49] KAKADE, S. M. (2001). A natural policy gradient. *Adv. Neural Inf. Process. Syst.* **14**.

[50] KALLUS, N. and ZHOU, A. (2020). Confounding-robust policy evaluation in infinite-horizon reinforcement learning. *Adv. Neural Inf. Process. Syst.* **33** 22293–22304.

[51] KEARNS, M. and SINGH, S. (2002). Near-optimal reinforcement learning in polynomial time. *Mach. Learn.* **49** 209–232.

[52] KIDAMBI, R., RAJESWARAN, A., NETRAPALLI, P. and JOACHIMS, T. (2020). Morel: Model-based offline reinforcement learning. arXiv preprint. Available at arXiv:2005.05951.

[53] KUMAR, A., FU, J., SOH, M., TUCKER, G. and LEVINE, S. (2019). Stabilizing off-policy q-learning via bootstrapping error reduction. *Adv. Neural Inf. Process. Syst.* **32**.

[54] KUMAR, P. R. and VARAIYA, P. (2016). *Stochastic Systems*: *Estimation*, *Identification*, *and Adaptive Control*. *Classics in Applied Mathematics* **75**. SIAM, Philadelphia, PA. Reprint of 1986 edition. MR3471643 https://doi.org/10.1137/1.9781611974263

[55] LAZARIC, A., GHAVAMZADEH, M. and MUNOS, R. (2012). Finite-sample analysis of least-squares policy iteration. *J. Mach. Learn. Res.* **13** 3041–3074. MR2997720

[56] LERASLE, M. (2019). *Lecture Notes*: *Selected Topics on Robust Statistical Learning Theory*. arXiv preprint. Available at arXiv:1908.10761.

[57] LI, L., CHU, W., LANGFORD, J. and SCHAPIRE, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the* 19*th International Conference on World Wide Web* 661–670. ACM, New York.

[58] LI, L., WALSH, T. J. and LITTMAN, M. L. (2006). Towards a unified theory of state abstraction for MDPs. In *Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics* 531–539.

[59] LIU, P., ZHAO, L., AGARWAL, S., LIU, J., HUANG, A., AMORTILA, P. and JIANG, N. (2025). Model Selection for Off-policy Evaluation: New Algorithms and Experimental Protocol. arXiv preprint. Available at arXiv:2502.08021.

[60] LIU, Q., LI, L., TANG, Z. and ZHOU, D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems* 5356–5366.

[61] MANDLEKAR, A., XU, D., WONG, J., NASIRIANY, S., WANG, C., KULKARNI, R., FEI-FEI, L., SAVARESE, S., ZHU, Y. et al. (2022). What matters in learning from offline human demonstrations for robot manipulation. In *Conference on Robot Learning* 1678–1690. PMLR.

[62] MNIH, V., KAVUKCUOGLU, K., SILVER, D., RUSU, A. A., VENESS, J., BELLEMARE, M. G., GRAVES, A., RIEDMILLER, M., FIDJELAND, A. K. et al. (2015). Human-level control through deep reinforcement learning. *Nature* **518** 529–533.

[63] MUNOS, R. (2007). Performance bounds in L_p-norm for approximate value iteration. *SIAM J. Control Optim.* **46** 541–561. MR2309039 https://doi.org/10.1137/040614384

[64] MUNOS, R. and SZEPESVÁRI, C. (2008). Finite-time bounds for fitted value iteration. *J. Mach. Learn. Res.* **9** 815–857. MR2417255

[65] MUTTI, M., DE SANTI, R., DE BARTOLOMEIS, P. and RESTELLI, M. (2023). Convex reinforcement learning in finite trials. *J. Mach. Learn. Res.* **24** Paper No. [250], 42. MR4633977

[66] NACHUM, O., CHOW, Y., DAI, B. and LI, L. (2019). Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Adv. Neural Inf. Process. Syst.* **32**.

[67] NACHUM, O. and DAI, B. (2020). Reinforcement learning via Fenchel–Rockafellar duality. arXiv preprint. Available at arXiv:2001.01866.

[68] NIE, A., REUEL, A.-K. and BRUNSKILL, E. (2023). Understanding the impact of reinforcement learning personalization on subgroups of students in math tutoring. In *International Conference on Artificial Intelligence in Education* 688–694 Springer, Berlin.

[69] OZDAGLAR, A. E., PATTATHIL, S., ZHANG, J. and ZHANG, K. (2023). Revisiting the linear-programming framework for offline rl with general function approximation. In *International Conference on Machine Learning* 26769–26791. PMLR.

[70] PATTERSON, A., WHITE, A. and WHITE, M. (2022). A generalized projected Bellman error for off-policy value estimation in reinforcement learning. *J. Mach. Learn. Res.* **23** Paper No. [145], 61. MR4577097

[71] PERDOMO, J. C., KRISHNAMURTHY, A., BARTLETT, P. and KAKADE, S. (2023). A complete characterization of linear estimators for offline policy evaluation. *J. Mach. Learn. Res.* **24** Paper No. [284], 50. MR4664721

[72] PIRES, B. A. and SZEPESVÁRI, C. (2012). Statistical linear estimation with penalized estimators: an application to reinforcement learning. arXiv preprint. Available at arXiv:1206.6444.

[73] PRECUP, D., SUTTON, R. S. and SINGH, S. P. (2000). Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning* 759–766.

[74] PUTERMAN, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. *Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Wiley, New York. A Wiley-Interscience Publication. MR1270015

[75] ROSS, S., GORDON, G. and BAGNELL, D. (2011). A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* 627–635.

[76] SCHULMAN, J., WOLSKI, F., DHARIWAL, P., RADFORD, A. and KLIMOV, O. (2017). Proximal policy optimization algorithms. arXiv preprint. Available at arXiv:1707.06347.

[77] SHI, C., UEHARA, M., HUANG, J. and JIANG, N. (2022). A minimax learning approach to off-policy evaluation in confounded partially observable Markov decision processes. In *International Conference on Machine Learning* 20057–20094. PMLR.

[78] SHIRANTHIKA, C., CHEN, K.-W., WANG, C.-Y., YANG, C.-Y., SUDANTHA, B. and LI, W.-F. (2022). Supervised optimal chemotherapy regimen based on offline reinforcement learning. *IEEE J. Biomed. Health Inform.* **26** 4763–4772.

[79] SILVER, D., SCHRITTWIESER, J., SIMONYAN, K., ANTONOGLOU, I., HUANG, A., GUEZ, A., HUBERT, T., BAKER, L., LAI, M. et al. (2017). Mastering the game of go without human knowledge. *Nature* **550** 354–359.

[80] SONG, Y., ZHOU, Y., SEKHARI, A., BAGNELL, D., KRISHNAMURTHY, A. and SUN, W. (2022). Hybrid RL: Using both offline and online data can make RL efficient. In *The Eleventh International Conference on Learning Representations*.

[81] SUTTON, R. S. and BARTO, A. G. (2018). *Reinforcement Learning: An Introduction*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3889951

[82] TANG, S. and WIENS, J. (2021). Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In *Machine Learning for Healthcare Conference* 2–35. PMLR.

[83] TENNENHOLTZ, G., SHALIT, U. and MANNOR, S. (2020). Off-policy evaluation in partially observable environments. In *Proceedings of the AAAI Conference on Artificial Intelligence* **34** 10276–10283.

[84] THOMAS, P. S. and BRUNSKILL, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning* 2139–2148.

[85] TSITSIKLIS, J. N. and VAN ROY, B. (1996). Feature-based methods for large scale dynamic programming. *Mach. Learn.* **22** 59–94.

[86] UEHARA, M., HUANG, J. and JIANG, N. (2020). Minimax weight and Q-function learning for off-policy evaluation. In *Proceedings of the 37th International Conference on Machine Learning* 1023–1032.

[87] UEHARA, M., KIYOHARA, H., BENNETT, A., CHERNOZHUKOV, V., JIANG, N., KALLUS, N., SHI, C. and SUN, W. (2022). Future-Dependent Value-Based Off-Policy Evaluation in POMDPs. arXiv preprint. Available at arXiv:2207.13081.

[88] UEHARA, M., ZHANG, X. and SUN, W. (2022). Representation learning for online and offline RL in low-rank MDPs. In *International Conference on Learning Representations*.

[89] VAN HASSELT, H., DORON, Y., STRUB, F., HESSEL, M., SONNERAT, N. and MODAYIL, J. (2018). Deep reinforcement learning and the deadly triad. arXiv preprint. Available at arXiv:1812.02648.

[90] VINYALS, O., BABUSCHKIN, I., CZARNECKI, W. M., MATHIEU, M., DUDZIK, A., CHUNG, J., CHOI, D. H., POWELL, R., EWALDS, T. et al. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* **575** 350–354.

[91] WAGENMAKER, A. and PACCHIANO, A. (2023). Leveraging offline data in online reinforcement learning. In *International Conference on Machine Learning* 35300–35338. PMLR.

[92] WANG, R., FOSTER, D. P. and KAKADE, S. M. (2020). What are the statistical limits of offline RL with linear function approximation? arXiv preprint. Available at arXiv:2010.11895.

[93] WANG, R., WU, Y., SALAKHUTDINOV, R. and KAKADE, S. (2021). Instabilities of offline rl with pre-trained neural representation. In *International Conference on Machine Learning* 10948–10960. PMLR.

[94] WIESEMANN, W., KUHN, D. and RUSTEM, B. (2013). Robust Markov decision processes. *Math. Oper. Res.* **38** 153–183. MR3029483 https://doi.org/10.1287/moor.1120.0566

[95] XIAO, C., WU, Y., MEI, J., DAI, B., LATTIMORE, T., LI, L., SZEPESVARI, C. and SCHUURMANS, D. (2021). On the optimality of batch policy optimization algorithms. In *International Conference on Machine Learning* 11362–11371. PMLR.

[96] XIE, T., CHENG, C.-A., JIANG, N., MINEIRO, P. and AGARWAL, A. (2021). Bellman-consistent pessimism for offline reinforcement learning. arXiv preprint. Available at arXiv:2106.06926.

[97] XIE, T., FOSTER, D. J., BAI, Y., JIANG, N. and KAKADE, S. M. (2023). The role of coverage in online reinforcement learning. In *The Eleventh International Conference on Learning Representations*.

[98] XIE, T. and JIANG, N. (2020). Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence* 550–559. PMLR.

[99] XIE, T. and JIANG, N. (2021). Batch value-function approximation with only realizability. In *International Conference on Machine Learning* 11404–11413. PMLR.

[100] XIE, T., JIANG, N., WANG, H., XIONG, C. and BAI, Y. (2021). Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Adv. Neural Inf. Process. Syst.* **34** 27395–27407.

[101] XIE, T., MA, Y. and WANG, Y.-X. (2019). Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Adv. Neural Inf. Process. Syst.* **32**.

[102] YANG, L. F. and WANG, M. (2019). Sample-optimal parametric Q-learning with linear transition models. arXiv preprint. Available at arXiv:1902.04779.

[103] YE, C., XIONG, W., ZHANG, Y., JIANG, N. and ZHANG, T. (2024). A theoretical analysis of Nash learning from human feedback under general kl-regularized preference. arXiv preprint. Available at arXiv:2402.07314.

[104] YIN, M. and WANG, Y.-X. (2020). Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics* 3948–3958. PMLR.

[105] YU, T., THOMAS, G., YU, L., ERMON, S., ZOU, J., LEVINE, S., FINN, C. and MA, T. (2020). Mopo: Model-based offline policy optimization. arXiv preprint. Available at arXiv:2005.13239.

[106] ZAHAVY, T., O'DONOGHUE, B., DESJARDINS, G. and SINGH, S. (2021). Reward is enough for convex mdps. *Adv. Neural Inf. Process. Syst.* **34** 25746–25759.

[107] ZANETTE, A. (2020). Exponential lower bounds for batch reinforcement learning: Batch RL can be exponentially harder than online RL. arXiv preprint. Available at arXiv:2012.08005.

[108] ZANETTE, A. (2023). When is realizability sufficient for off-policy reinforcement learning? In *International Conference on Machine Learning* 40637–40668. PMLR. Ann Arbor, MI.

[109] ZANETTE, A., WAINWRIGHT, M. J. and BRUNSKILL, E. (2021). Provable benefits of actor-critic methods for offline reinforcement learning. *Adv. Neural Inf. Process. Syst.* **34** 13626–13640.

[110] ZHAN, W., HUANG, B., HUANG, A., JIANG, N. and LEE, J. (2022). Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory* 2730–2775. PMLR.

[111] ZHANG, S. and JIANG, N. (2021). Towards hyperparameter-free policy selection for offline reinforcement learning. *Adv. Neural Inf. Process. Syst.* **34** 12864–12875.

[112] ZHANG, T. (2006). From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *Ann. Statist.* **34** 2180–2210. MR2291497 https://doi.org/10.1214/009053606000000704

[113] ZHANG, Y., BAI, Y. and JIANG, N. (2023). Offline learning in Markov games with general function approximation. In *International Conference on Machine Learning* 40804–40829. PMLR.

[114] ZHANG, Y. and JIANG, N. (2024). On the curses of future and history in future-dependent value functions for off-policy evaluation. arXiv preprint. Available at arXiv:2402.14703.

[115] ZHAO, Y., KOSOROK, M. R. and ZENG, D. (2009). Reinforcement learning design for cancer clinical trials. *Stat. Med.* **28** 3294–3315. MR2750277 https://doi.org/10.1002/sim.3720

[116] ZHAO, Y., ZENG, D., SOCINSKI, M. A. and KOSOROK, M. R. (2011). Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics* **67** 1422–1433. MR2872393 https://doi.org/10.1111/j.1541-0420.2011.01572.x

[117] ZHENG, G., ZHANG, F., ZHENG, Z., XIANG, Y., YUAN, N. J., XIE, X. and LI, Z. (2018). DRN: A deep reinforcement learning framework for news recommendation. In *Proceedings of the 2018 World Wide Web Conference* 167–176.

[118] ZHOU, R., LIU, T., CHENG, M., KALATHIL, D., KUMAR, P. and TIAN, C. (2024). Natural actor-critic for robust reinforcement learning with function approximation. *Adv. Neural Inf. Process. Syst.* **36**.

[119] ZHU, H., RASHIDINEJAD, P. and JIAO, J. (2023). Importance weighted actor-critic for optimal conservative offline reinforcement learning. *Adv. Neural Inf. Process. Syst.* **36**.

Statistical Science
2025, Vol. 40, No. 4, 597–622
https://doi.org/10.1214/25-STS989

The Central Role of the Loss Function in Reinforcement Learning

Kaiwen Wang, Nathan Kallus and Wen Sun

Abstract. This paper illustrates the central role of loss functions in data-driven decision making, providing a comprehensive survey on their influence in cost-sensitive classification (CSC) and reinforcement learning (RL). We demonstrate how different regression loss functions affect the sample efficiency and adaptivity of value-based, decision-making algorithms. Across multiple settings, we prove that algorithms using the binary cross-entropy loss achieve first-order bounds scaling with the optimal policy's cost and are much more efficient than the commonly used squared loss. Moreover, we prove that distributional algorithms using the maximum likelihood loss achieve second-order bounds scaling with the policy variance and are even sharper than first-order bounds. This in particular proves the benefits of distributional RL. We hope that this paper serves as a guide analyzing decision-making algorithms with varying loss functions, and can inspire the reader to seek out better loss functions to improve any decision-making algorithm.

Key words and phrases: First-order (small-loss) and second-order (variance-dependent) bounds, RL with function approximation, distributional RL.

REFERENCES

[1] ADLER, B., AGARWAL, N., AITHAL, A., ANH, D. H., BHATTACHARYA, P., BRUNDYN, A., CASPER, J., CATANZARO, B., CLAY, S. et al. (2024). Nemotron-4 340B Technical Report. arXiv preprint. Available at arXiv:2406.11704.

[2] AGARWAL, A., JIANG, N., KAKADE, S. M. and SUN, W. (2019). Reinforcement learning: Theory and algorithms. CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep 32 96.

[3] AGARWAL, A., KAKADE, S., KRISHNAMURTHY, A. and SUN, W. (2020). Flambe: Structural complexity and representation learning of low rank mdps. In *Advances in Neural Information Processing Systems* **33** 20095–20107.

[4] AGARWAL, A., SONG, Y., SUN, W., WANG, K., WANG, M. and ZHANG, X. (2023). Provable benefits of representational transfer in reinforcement learning. In *The Thirty Sixth Annual Conference on Learning Theory* 2114–2187. PMLR.

[5] AUDIBERT, J.-Y. and TSYBAKOV, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.* **35** 608–633. MR2336861 https://doi.org/10.1214/009053606000001217

[6] AUER, P., CESA-BIANCHI, N., FREUND, Y. and SCHAPIRE, R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM J. Comput.* **32** 48–77. MR1954855 https://doi.org/10.1137/S0097539701398375

[7] AYOUB, A., WANG, K., LIU, V., ROBERTSON, S., MCINERNEY, J., LIANG, D., KALLUS, N. and SZEPESVARI, C. (2024). Switching the loss reduces the cost in batch reinforcement learning. In *Forty-First International Conference on Machine Learning*.

[8] BALL, P. J., SMITH, L., KOSTRIKOV, I. and LEVINE, S. (2023). Efficient online reinforcement learning with offline data. In *International Conference on Machine Learning* 1577–1594. PMLR.

[9] BAS-SERRANO, J., CURI, S., KRAUSE, A. and NEU, G. (2021). Logistic Q-learning. In *International Conference on Artificial Intelligence and Statistics* 3610–3618. PMLR.

[10] BELLEMARE, M. G., CANDIDO, S., CASTRO, P. S., GONG, J., MACHADO, M. C., MOITRA, S., PONDA, S. S. and WANG, Z. (2020). Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature* **588** 77–82.

[11] BELLEMARE, M. G., DABNEY, W. and MUNOS, R. (2017). A distributional perspective on reinforcement learning. In *International Conference on Machine Learning* 449–458. PMLR.

[12] BELLEMARE, M. G., DABNEY, W. and ROWLAND, M. (2023). *Distributional Reinforcement Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR4649766

[13] BENNETT, A., KALLUS, N., OPRESCU, M., SUN, W. and WANG, K. (2024). Efficient and sharp off-policy evaluation in robust Markov decision processes. In *Advances in Neural Information Processing Systems*.

[14] CHANG, J., WANG, K., KALLUS, N. and SUN, W. (2022). Learning Bellman complete representations for offline policy evaluation. In *International Conference on Machine Learning* 2938–2971. PMLR.

Kaiwen Wang is a Ph.D. Candidate, Computer Science Department, Cornell Tech, New York, New York 10044, USA (e-mail: kw437@cornell.edu). Nathan Kallus is an Associate Professor, Associate Professor of Operations Research and Information Engineering, Cornell Tech, New York, New York 10044, USA (e-mail: kallus@cornell.edu). Wen Sun is an Assistant Professor, Computer Science Department, Cornell Tech, New York, New York 10044, USA (e-mail: ws455@cornell.edu).

[15] CHEN, J. and JIANG, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning* 1042–1051. PMLR.

[16] CHENG, C.-A., XIE, T., JIANG, N. and AGARWAL, A. (2022). Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning* 3852–3878. PMLR.

[17] DABNEY, W., OSTROVSKI, G., SILVER, D. and MUNOS, R. (2018). Implicit quantile networks for distributional reinforcement learning. In *International Conference on Machine Learning* 1096–1105. PMLR.

[18] DANN, C., JIANG, N., KRISHNAMURTHY, A., AGARWAL, A., LANGFORD, J. and SCHAPIRE, R. E. (2018). On oracle-efficient pac rl with rich observations. In *Advances in Neural Information Processing Systems* **31**.

[19] DANN, C., MANSOUR, Y., MOHRI, M., SEKHARI, A. and SRIDHARAN, K. (2022). Guarantees for epsilon-greedy reinforcement learning with function approximation. In *International Conference on Machine Learning* 4666–4689. PMLR.

[20] DEVROYE, L. and LUGOSI, G. (2001). *Combinatorial Methods in Density Estimation*. *Springer Series in Statistics*. Springer, New York. MR1843146 https://doi.org/10.1007/978-1-4613-0125-7

[21] FAREBROTHER, J., ORBAY, J., VUONG, Q., TAIGA, A. A., CHEBOTAR, Y., XIAO, T., IRPAN, A., LEVINE, S., CASTRO, P. S. et al. (2024). Stop regressing: Training value functions via classification for scalable deep RL. In *Forty-First International Conference on Machine Learning*.

[22] FENG, F., YIN, W., AGARWAL, A. and YANG, L. (2021). Provably correct optimization and exploration with non-linear policies. In *International Conference on Machine Learning* 3263–3273. PMLR.

[23] FOSTER, D., AGARWAL, A., DUDÍK, M., LUO, H. and SCHAPIRE, R. (2018). Practical contextual bandits with regression oracles. In *International Conference on Machine Learning* 1539–1548. PMLR.

[24] FOSTER, D. J., BLOCK, A. and MISRA, D. (2024). Is behavior cloning all you need? Understanding horizon in imitation learning. arXiv preprint. Available at arXiv:2407.15007.

[25] FOSTER, D. J., KAKADE, S. M., QIAN, J. and RAKHLIN, A. (2021). The statistical complexity of interactive decision making. arXiv preprint. Available at arXiv:2112.13487.

[26] FOSTER, D. J. and KRISHNAMURTHY, A. (2021). Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. In *Advances in Neural Information Processing Systems* **34** 18907–18919.

[27] FOSTER, D. J., KRISHNAMURTHY, A., SIMCHI-LEVI, D. and XU, Y. (2022). Offline reinforcement learning: Fundamental barriers for value function approximation. In *Conference on Learning Theory* 3489–3489. PMLR.

[28] FOSTER, D. J., RAKHLIN, A., SEKHARI, A. and SRIDHARAN, K. (2022). On the complexity of adversarial decision making. In *Advances in Neural Information Processing Systems* **35** 35404–35417.

[29] GAO, Z., CHANG, J. D., ZHAN, W., OERTELL, O., SWAMY, G., BRANTLEY, K., JOACHIMS, T., BAGNELL, J. A., LEE, J. D. et al. (2024). Rebel: Reinforcement learning via regressing relative rewards. arXiv preprint. Available at arXiv:2404.16767.

[30] HU, Y., KALLUS, N. and MAO, X. (2022). Fast rates for contextual linear optimization. *Manage. Sci.* **68** 4236–4245.

[31] IMANI, E. and WHITE, M. (2018). Improving regression performance with distributional losses. In *International Conference on Machine Learning* 2157–2166. PMLR.

[32] JASON, M. Y., JAYARAMAN, D. and BASTANI, O. (2022). Conservative offline distributional reinforcement learning. In *Advances in Neural Information Processing Systems*.

[33] JIA, Z., QIAN, J., RAKHLIN, A. and WEI, C.-Y. (2024). How does variance shape the regret in contextual bandits? In *Advances in Neural Information Processing Systems*.

[34] JIANG, N. and AGARWAL, A. (2018). Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference on Learning Theory* 3395–3398. PMLR.

[35] JIN, C., LIU, Q. and MIRYOOSEFI, S. (2021). Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. In *Advances in Neural Information Processing Systems* **34** 13406–13418.

[36] JIN, C., YANG, Z., WANG, Z. and JORDAN, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory* 2137–2143. PMLR.

[37] KAKADE, S. and LANGFORD, J. (2002). Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning* 267–274.

[38] KALLUS, N., MAO, X., WANG, K. and ZHOU, Z. (2022). Doubly robust distributionally robust off-policy evaluation and learning. In *International Conference on Machine Learning* 10598–10632. PMLR.

[39] KOLTER, J. (2011). The fixed points of off-policy TD. In *Advances in Neural Information Processing Systems* **24**.

[40] KONTOROVICH, A. (2024). Binomial small deviations. Tweet.

[41] KRISHNAMURTHY, A., AGARWAL, A., HUANG, T.-K., DAUMÉ, H. III and LANGFORD, J. (2019). Active learning for cost-sensitive classification. *J. Mach. Learn. Res.* **20** Paper No. 65, 50. MR3960919

[42] KRISHNAN, S., YANG, Z., GOLDBERG, K., HELLERSTEIN, J. and STOICA, I. (2018). Learning to optimize join queries with deep reinforcement learning. arXiv preprint. Available at arXiv:1808.03196.

[43] KUMAR, A., ZHOU, A., TUCKER, G. and LEVINE, S. (2020). Conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems* **33** 1179–1191.

[44] LIU, Q., CHUNG, A., SZEPESVÁRI, C. and JIN, C. (2022). When is partially observable reinforcement learning not scary? In *Conference on Learning Theory* 5175–5220. PMLR.

[45] LUEDTKE, A. and CHAMBAZ, A. (2017). Faster rates for policy learning. arXiv preprint. Available at arXiv:1704.06431.

[46] LYKOURIS, T., SRIDHARAN, K. and TARDOS, É. (2018). Small-loss bounds for online learning with partial information. In *Conference on Learning Theory* 979–986. PMLR.

[47] MA, Y., JAYARAMAN, D. and BASTANI, O. (2021). Conservative offline distributional reinforcement learning. In *Advances in Neural Information Processing Systems* **34** 19235–19247.

[48] MHAMMEDI, Z., BLOCK, A., FOSTER, D. J. and RAKHLIN, A. (2024). Efficient model-free exploration in low-rank mdps. In *Advances in Neural Information Processing Systems* **36**.

[49] MHAMMEDI, Z., FOSTER, D. J. and RAKHLIN, A. (2024). The power of resets in online reinforcement learning. arXiv preprint. Available at arXiv:2404.15417.

[50] MISRA, D., HENAFF, M., KRISHNAMURTHY, A. and LANGFORD, J. (2020). Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International Conference on Machine Learning* 6961–6971. PMLR.

[51] MNIH, V., KAVUKCUOGLU, K., SILVER, D., RUSU, A. A., VENESS, J., BELLEMARE, M. G., GRAVES, A., RIEDMILLER, M., FIDJELAND, A. K. et al. (2015). Human-level control through deep reinforcement learning. *Nature* **518** 529–533.

[52] MODI, A., CHEN, J., KRISHNAMURTHY, A., JIANG, N. and AGARWAL, A. (2024). Model-free representation learning and exploration in low-rank MDPs. *J. Mach. Learn. Res.* **25** Paper No. [6], 76. MR4723856

[53] MUNOS, R. and SZEPESVÁRI, C. (2008). Finite-time bounds for fitted value iteration. *J. Mach. Learn. Res.* **9** 815–857. MR2417255

[54] OUYANG, L., WU, J., JIANG, X., ALMEIDA, D., WAINWRIGHT, C., MISHKIN, P., ZHANG, C., AGARWAL, S., SLAMA, K. et al. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems* **35** 27730–27744.

[55] PACCHIANO, A. (2024). Second order bounds for contextual bandits with function approximation. arXiv preprint. Available at arXiv:2409.16197.

[56] RASHIDINEJAD, P., ZHU, B., MA, C., JIAO, J. and RUSSELL, S. (2022). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *IEEE Trans. Inf. Theory* **68** 8156–8196. MR4544936 https://doi.org/10.1109/tit.2022.3185139

[57] RUSSO, D. and VAN ROY, B. (2013). Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems* **26**.

[58] SONG, Y., ZHOU, Y., SEKHARI, A., BAGNELL, D., KRISHNAMURTHY, A. and SUN, W. (2023). Hybrid RL: Using both offline and online data can make RL efficient. In *The Eleventh International Conference on Learning Representations*.

[59] TSITSIKLIS, J. N. and VAN ROY, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Trans. Automat. Control* **42** 674–690. MR1454208 https://doi.org/10.1109/9.580874

[60] UEHARA, M. and SUN, W. (2022). Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*.

[61] WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics: A Non-asymptotic Viewpoint*. *Cambridge Series in Statistical and Probabilistic Mathematics* **48**. Cambridge Univ. Press, Cambridge. MR3967104 https://doi.org/10.1017/9781108627771

[62] WANG, J., WANG, K., LI, Y., KALLUS, N., TRUMMER, I. and SUN, W. (2024). JoinGym: An efficient join order selection environment. *Reinf. Learn. J.* **1**.

[63] WANG, K., KALLUS, N. and SUN, W. (2023). Near-minimax-optimal risk-sensitive reinforcement learning with cvar. In *International Conference on Machine Learning* 35864–35907. PMLR.

[64] WANG, K., KIDAMBI, R., SULLIVAN, R., AGARWAL, A., DANN, C., MICHI, A., GELMI, M., LI, Y., GUPTA, R. et al. (2024). Conditional language policy: A general framework for steerable multi-objective finetuning. In *Findings of Empirical Methods in Natural Language Processing*.

[65] WANG, K., LIANG, D., KALLUS, N. and SUN, W. (2024). A reductions approach to risk-sensitive reinforcement learning with optimized certainty equivalents. arXiv preprint. Available at arXiv:2403.06323.

[66] WANG, K., OERTELL, O., AGARWAL, A., KALLUS, N. and SUN, W. (2024). More benefits of being distributional: Second-order bounds for reinforcement learning. In *International Conference on Machine Learning*.

[67] WANG, K., ZHOU, K., WU, R., KALLUS, N. and SUN, W. (2023). The benefits of being distributional: Small-loss bounds for reinforcement learning. In *Advances in Neural Information Processing Systems* **36**.

[68] WANG, R., FOSTER, D. and KAKADE, S. M. (2021). What are the statistical limits of offline RL with linear function approximation? In *International Conference on Learning Representations*.

[69] WANG, Z., ZHOU, D., LUI, J. and SUN, W. (2024). Model-based RL as a minimalist approach to horizon-free and second-order bounds. arXiv preprint. Available at arXiv:2408.08994.

[70] WATKINS, C. J. and DAYAN, P. (1992). Q-learning. *Mach. Learn.* **8** 279–292.

[71] WEISZ, G., AMORTILA, P. and SZEPESVÁRI, C. (2021). Exponential lower bounds for planning in MDPs with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory. Proc. Mach. Learn. Res.* (*PMLR*) **132** 1237–1264. PMLR. MR4227360

[72] WU, R., UEHARA, M. and SUN, W. (2023). Distributional offline policy evaluation with predictive error guarantees. In *International Conference on Machine Learning* 37685–37712. PMLR.

[73] XIE, T., CHENG, C.-A., JIANG, N., MINEIRO, P. and AGARWAL, A. (2021). Bellman-consistent pessimism for offline reinforcement learning. In *Advances in Neural Information Processing Systems* **34** 6683–6694.

[74] XIE, T., FOSTER, D. J., BAI, Y., JIANG, N. and KAKADE, S. M. (2023). The role of coverage in online reinforcement learning. In *The Eleventh International Conference on Learning Representations*.

[75] ZHANG, S., LI, H., WANG, M., LIU, M., CHEN, P.-Y., LU, S., LIU, S., MURUGESAN, K. and CHAUDHURY, S. (2023). On the convergence and sample complexity analysis of deep Q-networks with ϵ-greedy exploration. In *Thirty-Seventh Conference on Neural Information Processing Systems*.

[76] ZHANG, T. (2023). *Mathematical Analysis of Machine Learning Algorithms*. Cambridge Univ. Press, Cambridge. https://doi.org/10.1017/9781009093057

[77] ZHOU, J. P., WANG, K., CHANG, J., GAO, Z., KALLUS, N., WEINBERGER, K. Q., BRANTLEY, K. and SUN, W. (2025). $Q\sharp$: Provably optimal distributional RL for LLM post-training. arXiv preprint. Available at arXiv:2502.20548.

Statistical Science
2025, Vol. 40, No. 4, 623–640
https://doi.org/10.1214/25-STS1012

Partially Observable RL: Benign Structures and Simple Generic Algorithms

Qinghua Liu and Chi Jin

Abstract. Partially observable Reinforcement Learning (RL) applications, where agents must make a series of decisions without complete knowledge of the underlying states of the system, are widespread. Partially observable RL can be notoriously difficult—well-known information-theoretic results show that learning partially observable Markov decision processes (POMDPs) requires an exponential number of samples in the worst case. However, this does not preclude the possibility of identifying rich subclasses of structured POMDPs for which learning remains feasible.

This survey aims to offer a high-level overview of recent advancements in learning structured POMDPs. We will identify clean and practical problem structures that facilitate sample-efficient learning. Additionally, we will introduce simple and generic algorithms for learning POMDPs under different structural conditions. Finally, we will provide a unified view of these results under the framework of well-conditioned predictive state representation, which further reveals new tractable classes of partially observable problems along the way.

Key words and phrases: Partially observable reinforcement learning.

REFERENCES

[1] AGARWAL, A., JIANG, N., KAKADE, S. M. and SUN, W. (2019). Reinforcement Learning: Theory and Algorithms. CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep, 32:96.

[2] AKKAYA, I., ANDRYCHOWICZ, M., CHOCIEJ, M., LITWIN, M., MCGREW, B., PETRON, A., PAINO, A., PLAPPERT, M., POWELL, G. et al. (2019). Solving Rubik's cube with a robot hand. arXiv preprint. Available at arXiv:1910.07113.

[3] ALTABAA, A. and YANG, Z. (2024). On the role of information structure in reinforcement learning for partially-observable sequential teams and games. arXiv preprint. Available at arXiv:2403.00993.

[4] ANANDKUMAR, A., GE, R., HSU, D., KAKADE, S. M. and TELGARSKY, M. (2014). Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.* **15** 2773–2832. MR3270750

[5] ÅSTRÖM, K. J. (1965). Optimal control of Markov processes with incomplete state information. *J. Math. Anal. Appl.* **10** 174–205. MR0173570 https://doi.org/10.1016/0022-247X(65)90154-X

[6] ÅSTRÖM, K. J. (2012). *Introduction to Stochastic Control Theory*. Courier Corporation.

[7] AUER, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.* **3** 397–422. MR1984023 https://doi.org/10.1162/153244303321897663

[8] AZAR, M. G., OSBAND, I. and MUNOS, R. (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning* 263–272. PMLR.

[9] AZIZZADENESHELI, K., LAZARIC, A. and ANANDKUMAR, A. (2016). Reinforcement learning of POMDPs using spectral methods. In *Conference on Learning Theory* 193–256. PMLR.

[10] BERNSTEIN, D. S., GIVAN, R., IMMERMAN, N. and ZILBERSTEIN, S. (2002). The complexity of decentralized control of Markov decision processes. *Math. Oper. Res.* **27** 819–840. MR1939179 https://doi.org/10.1287/moor.27.4.819.297

[11] BRAFMAN, R. I. and TENNENHOLTZ, M. (2002). R-MAX—a general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.* **3** 213–231. MR1971337 https://doi.org/10.1162/153244303765208377

[12] BROWN, N. and SANDHOLM, T. (2019). Superhuman AI for multiplayer poker. *Science* **365** 885–890. MR3966356 https://doi.org/10.1126/science.aay2400

[13] BRUNSKILL, E. and LI, L. (2013). Sample complexity of multi-task reinforcement learning. arXiv preprint. Available at arXiv:1309.6821.

[14] CAMPBELL, M., HOANE JR, A. J. and HSU, F.-H. (2002). Deep blue. *Artif. Intell.* **134** 57–83.

[15] CHADES, I., CARWARDINE, J., MARTIN, T., NICOL, S., SABBADIN, R. and BUFFET, O. (2012). Momdps: A solution for modelling adaptive management problems. In *Proceedings of the AAAI Conference on Artificial Intelligence* **26** 267–273.

[16] CHEN, F., BAI, Y. and MEI, S. (2022). Partially observable rl with b-stability: Unified structural condition and sharp sample-efficient algorithms. arXiv preprint. Available at arXiv:2209.14990.

Qinghua Liu is a postdoc researcher at Microsoft Research, NYC, USA (e-mail: qinghual@princeton.edu). Chi Jin is an assistant professor at Princeton University, Princeton, USA (e-mail: chij@princeton.edu).

[17] CHEN, F., MEI, S. and BAI, Y. (2025). Unified algorithms for RL with decision-estimation coefficients: PAC, reward-free, preference-based learning and beyond. *Ann. Statist.* **53** 426–456. MR4865022 https://doi.org/10.1214/24-aos2483

[18] CHEN, F., WANG, H., XIONG, C., MEI, S. and BAI, Y. (2023). Lower bounds for learning in revealing pomdps. In *International Conference on Machine Learning* 5104–5161. PMLR.

[19] GUO, Z. D., DOROUDI, S. and BRUNSKILL, E. (2016). A PAC RL algorithm for episodic POMDPs. In *Artificial Intelligence and Statistics* 510–518. PMLR.

[20] DU, S., KRISHNAMURTHY, A., JIANG, N., AGARWAL, A., DUDIK, M. and LANGFORD, J. (2019). Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning* 1665–1674. PMLR.

[21] EFRONI, Y. and JIN, C. (2022). Akshay Krishnamurthy, and Sobhan Miryoosefi. Provable reinforcement learning with a short-term memory. arXiv preprint. Available at arXiv:2202.03983.

[22] EVEN-DAR, E., KAKADE, S. M. and MANSOUR, Y. (2007). The value of observation for monitoring dynamic systems. In *IJCAI* 2474–2479.

[23] FOSTER, D. J., KAKADE, S. M., QIAN, J. and RAKHLIN, A. (2021). The statistical complexity of interactive decision making. arXiv preprint. Available at arXiv:2112.13487.

[24] GEER, S. A., VAN DE GEER, S. and WILLIAMS, D. (2000). *Empirical Processes in M-Estimation* **6**. Cambridge Univ. Press, Cambridge.

[25] GOLOWICH, N., MOITRA, A. and ROHATGI, D. (2022). Learning in observable pomdps, without computationally intractable oracles. *Adv. Neural Inf. Process. Syst.* **35** 1458–1473.

[26] GOLOWICH, N., MOITRA, A. and ROHATGI, D. (2022). Planning in observable POMDPs in quasipolynomial time. arXiv preprint. Available at arXiv:2201.04735.

[27] GUO, J., CHEN, M., WANG, H., XIONG, C., WANG, M. and BAI, Y. (2023). Sample-efficient learning of pomdps with multiple observations in hindsight. arXiv preprint. Available at arXiv:2307.02884.

[28] GUO, J., LI, Z., WANG, H., WANG, M., YANG, Z. and ZHANG, X. (2023). Provably efficient representation learning with tractable planning in low-rank pomdp. In *International Conference on Machine Learning* 11967–11997. PMLR.

[29] HAUSKRECHT, M. and FRASER, H. (2000). Planning treatment of ischemic heart disease with partially observable Markov decision processes. *Artif. Intell. Med.* **18** 221–244.

[30] HO, B. L. and KALMAN, R. E. (1966). Effective construction of linear state-variable models from input/output functions: Die konstruktion von linearen modeilen in der darstellung durch zustandsvariable aus den beziehungen für ein-und ausgangsgrößen. *Automatisierungstechnik* **14** 545–548.

[31] HSU, D., KAKADE, S. M. and ZHANG, T. (2012). A spectral algorithm for learning hidden Markov models. *J. Comput. System Sci.* **78** 1460–1480. MR2926144 https://doi.org/10.1016/j.jcss.2011.12.025

[32] JAEGER, H. (1998). Discrete-time, Discrete-valued Observable Operator Models: a Tutorial. GMD-Forschungszentrum Informationstechnik Darmstadt, Germany.

[33] JAKSCH, T., ORTNER, R. and AUER, P. (2010). Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.* **11** 1563–1600. MR2645461

[34] JIANG, N., KRISHNAMURTHY, A., AGARWAL, A., LANGFORD, J. and SCHAPIRE, R. E. (2017). Contextual decision processes with low Bellman rank are PAC-learnable. In *International Conference on Machine Learning* 1704–1713. PMLR.

[35] JIN, C., ALLEN-ZHU, Z., BUBECK, S. and JORDAN, M. I. (2018). Is Q-learning provably efficient? *Adv. Neural Inf. Process. Syst.* **31**.

[36] JIN, C., KAKADE, S. M., KRISHNAMURTHY, A. and LIU, Q. (2020). Sample-efficient reinforcement learning of undercomplete POMDPs. *Adv. Neural Inf. Process. Syst.*

[37] JIN, C., LIU, Q. and MIRYOOSEFI, S. (2021). Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. *Adv. Neural Inf. Process. Syst.* **34**.

[38] JIN, C., LIU, Q., WANG, Y. and YU, T. (2024). V-learning—a simple, efficient, decentralized algorithm for multiagent reinforcement learning. *Math. Oper. Res.* **49** 2295–2322. MR4838262 https://doi.org/10.1287/moor.2021.0317

[39] KALMAN, R. E. (1960). A new approach to linear filtering and prediction problems. *J. Basic Eng.* **82** 35–45. MR3931993

[40] KREPS, D. M. (1990). *Game Theory and Economic Modelling*. Oxford Univ. Press, London.

[41] KRISHNAMURTHY, A., AGARWAL, A. and LANGFORD, J. (2016). PAC reinforcement learning with rich observations. *Adv. Neural Inf. Process. Syst.* **29**.

[42] KWON, J., EFRONI, Y., CARAMANIS, C. and MANNOR, S. (2021). Reinforcement learning in reward-mixing mdps. *Adv. Neural Inf. Process. Syst.* **34** 2253–2264.

[43] KWON, J., EFRONI, Y., CARAMANIS, C. and MANNOR, S. (2021). Rl for latent mdps: Regret guarantees and a lower bound. *Adv. Neural Inf. Process. Syst.* **34** 24523–24534.

[44] KWON, J., EFRONI, Y., CARAMANIS, C. and MANNOR, S. (2023). Reward-mixing mdps with few latent contexts are learnable. In *International Conference on Machine Learning* 18057–18082. PMLR.

[45] KWON, J., MANNOR, S., CARAMANIS, C. and EFRONI, Y. (2024). Rl in latent mdps is tractable: Online guarantees via off-policy evaluation. arXiv preprint. Available at arXiv:2406.01389.

[46] LALE, S., AZIZZADENESHELI, K., HASSIBI, B. and ANANDKUMAR, A. (2020). Logarithmic regret bound in partially observable linear dynamical systems. *Adv. Neural Inf. Process. Syst.* **33** 20876–20888.

[47] LATTIMORE, T. and SZEPESVÁRI, C. (2020). *Bandit Algorithms*. Cambridge Univ. Press, Cambridge.

[48] LEE, J., AGARWAL, A., DANN, C. and ZHANG, T. (2023). Learning in pomdps is sample-efficient with hindsight observability. In *International Conference on Machine Learning* 18733–18773. PMLR.

[49] LEURGANS, S. E., ROSS, R. T. and ABEL, R. B. (1993). A decomposition for three-way arrays. *SIAM J. Matrix Anal. Appl.* **14** 1064–1083. MR1238921 https://doi.org/10.1137/0614071

[50] LEVINSON, J., ASKELAND, J., BECKER, J., DOLSON, J., HELD, D., KAMMEL, S., KOLTER, J. Z., LANGER, D., PINK, O. et al. (2011). Towards fully autonomous driving: Systems and algorithms. In 2011 *IEEE Intelligent Vehicles Symposium* (*IV*) 163–168 IEEE Press, London.

[51] LITTMAN, M. and SUTTON, R. S. (2001). Predictive representations of state. *Adv. Neural Inf. Process. Syst.* **14**.

[52] LITTMAN, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings* 1994 157–163. Elsevier, Amsterdam.

[53] LIU, Q., CHUNG, A., SZEPESVARI, C. and JIN, C. When is partially observable reinforcement learning not scary? In *Proceedings of Thirty Fifth Conference on Learning Theory*. *Proceedings of Machine Learning Research* **178** 5175–5220. Available at https://proceedings.mlr.press/v178/liu22f.html.

[54] LIU, Q., NETRAPALLI, P., SZEPESVARI, C. and JIN, C. (2023). Optimistic MLE: A generic model-based algorithm for partially observable sequential decision making. In *STOC'23—Proceedings of the 55th Annual ACM Symposium on Theory of Computing* 363–376. ACM, New York. MR4617392 https://doi.org/10.1145/3564246.3585161

[55] LIU, Q., SZEPESVÁRI, C. and JIN, C. (2022). Sample-efficient reinforcement learning of partially observable Markov games. *Adv. Neural Inf. Process. Syst.* **35** 18296–18308.

[56] LIU, X. and ZHANG, K. (2023). Partially observable multi-agent rl with (quasi-) efficiency: The blessing of information sharing. In *International Conference on Machine Learning* 22370–22419. PMLR.

[57] LJUNG, L. (1998). System identification. In *Signal Analysis and Prediction* 163–173 Springer, Berlin.

[58] LU, M., MIN, Y., WANG, Z. and YANG, Z. (2022). Pessimism in the face of confounders: Provably efficient offline reinforcement learning in partially observable Markov decision processes. arXiv preprint. Available at arXiv:2205.13589.

[59] MANIA, H., TU, S. and RECHT, B. (2019). Certainty equivalence is efficient for linear quadratic control. *Adv. Neural Inf. Process. Syst.* **32**.

[60] MNIH, V., KAVUKCUOGLU, K., SILVER, D., GRAVES, A., ANTONOGLOU, I., WIERSTRA, D. and RIEDMILLER, M. (2013). Playing atari with deep reinforcement learning. arXiv preprint. Available at arXiv:1312.5602.

[61] MOSSEL, E. and ROCH, S. (2005). Learning nonsingular phylogenies and hidden Markov models. In *STOC'05: Proceedings of the 37th Annual ACM Symposium on Theory of Computing* 366–375. ACM, New York. MR2181638 https://doi.org/10.1145/1060590.1060645

[62] MUNDHENK, M., GOLDSMITH, J., LUSENA, C. and ALLENDER, E. (2000). Complexity of finite-horizon Markov decision process problems. *J. ACM* **47** 681–720. MR1866174 https://doi.org/10.1145/347476.347480

[63] OYMAK, S. and OZAY, N. (2019). Non-asymptotic identification of lti systems from a single trajectory. In 2019 *American Control Conference* (*ACC*) 5655–5661. IEEE Press, London.

[64] OYMAK, S. and OZAY, N. (2022). Revisiting Ho–Kalman-based system identification: Robustness and finite-sample analysis. *IEEE Trans. Automat. Control* **67** 1914–1928. MR4402414 https://doi.org/10.1109/tac.2021.3083651

[65] PAPADIMITRIOU, C. H. and TSITSIKLIS, J. N. (1987). The complexity of Markov decision processes. *Math. Oper. Res.* **12** 441–450. MR0906416 https://doi.org/10.1287/moor.12.3.441

[66] SHAPLEY, L. S. (1953). Stochastic games. *Proc. Natl. Acad. Sci. USA* **39** 1095–1100. MR0061807 https://doi.org/10.1073/pnas.39.10.1953

[67] SHI, C., UEHARA, M., HUANG, J. and JIANG, N. (2022). A minimax learning approach to off-policy evaluation in confounded partially observable Markov decision processes. In *International Conference on Machine Learning* 20057–20094. PMLR.

[68] SILVER, D., SCHRITTWIESER, J., SIMONYAN, K., ANTONOGLOU, I., HUANG, A., GUEZ, A., HUBERT, T., BAKER, L., LAI, M. et al. (2017). Mastering the game of go without human knowledge. *Nature* **550** 354–359.

[69] SIMCHOWITZ, M., SINGH, K. and HAZAN, E. (2020). Improper learning for non-stochastic control. In *Conference on Learning Theory* 3320–3436. PMLR.

[70] SINGH, S., JAMES, M. and RUDARY, M. (2012). Predictive state representations: a new theory for modeling dynamical systems. arXiv preprint. Available at arXiv:1207.4167.

[71] STEIMLE, L. N., KAUFMAN, D. L. and DENTON, B. T. (2021). Multi-model Markov decision processes. *IISE Trans.* **53** 1124–1139.

[72] SZEPESVÁRI, C. (2022). *Algorithms for Reinforcement Learning*. *Synthesis Lectures on Artificial Intelligence and Machine Learning* **9**. Springer, Cham. Reprint of the 2010 original. MR4647095 https://doi.org/10.1007/978-3-031-01551-9

[73] TIAN, Y., ZHANG, K., TEDRAKE, R. and SRA, S. (2023). Can direct latent model learning solve linear quadratic Gaussian control? In *Learning for Dynamics and Control Conference* 51–63. PMLR.

[74] TODOROV, E. and LI, W. (2005). A generalized iterative lqg method for locally-optimal feedback control of constrained nonlinear stochastic systems. In *Proceedings of the* 2005, *American Control Conference*, 2005 300–306. IEEE Press, New York.

[75] TSIAMIS, A. and PAPPAS, G. J. (2019). Finite sample analysis of stochastic system identification. In 2019 *IEEE 58th Conference on Decision and Control* (*CDC*) 3648–3654. IEEE Press, London.

[76] UEHARA, M., KIYOHARA, H., BENNETT, A., CHERNOZHUKOV, V., JIANG, N., KALLUS, N., SHI, C. and SUN, W. (2024). Future-dependent value-based off-policy evaluation in pomdps. *Adv. Neural Inf. Process. Syst.* **36**.

[77] UEHARA, M., SEKHARI, A., LEE, J. D., KALLUS, N. and SUN, W. (2022). Provably efficient reinforcement learning in partially observable dynamical systems. *Adv. Neural Inf. Process. Syst.* **35** 578–592.

[78] NEUMANN, J. V. (1928). Zur Theorie der Gesellschaftsspiele. *Math. Ann.* **100** 295–320. MR1512486 https://doi.org/10.1007/BF01448847

[79] VINYALS, O., BABUSCHKIN, I., CZARNECKI, W. M., MATHIEU, M., DUDZIK, A., CHUNG, J., CHOI, D. H., POWELL, R., EWALDS, T. et al. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* **575** 350–354.

[80] VLASSIS, N., LITTMAN, M. L. and BARBER, D. (2012). On the computational complexity of stochastic controller optimization in POMDPs. *ACM Trans. Comput. Theory* **4** 1–8.

[81] WANG, L., CAI, Q., YANG, Z. and WANG, Z. (2022). Embed to control partially observed systems: Representation learning with provable sample efficiency. arXiv preprint. Available at arXiv:2205.13476.

[82] ZHAN, W., UEHARA, M., SUN, W. and LEE, J. D. (2022). Pac reinforcement learning for predictive state representations. arXiv preprint. Available at arXiv:2207.05738.

[83] ZHANG, Y. and JIANG, N. (2024). On the curses of future and history in future-dependent value functions for off-policy evaluation. arXiv preprint. Available at arXiv:2402.14703.

[84] ZHENG, Y., FURIERI, L., KAMGARPOUR, M. and LI, N. (2021). Sample complexity of linear quadratic Gaussian (lqg) control for output feedback systems. In *Learning for Dynamics and Control* 559–570. PMLR.

[85] ZHONG, H., XIONG, W., ZHENG, S., WANG, L., WANG, Z., YANG, Z. and ZHANG, T. (2022). Gec: a unified framework for interactive decision making in mdp, pomdp, and beyond. arXiv preprint. Available at arXiv:2211.01962.

Statistical Science
2025, Vol. 40, No. 4, 641–655
https://doi.org/10.1214/25-STS998

The Theory of Online Control

Elad Hazan and Karan Singh

Abstract. This article introduces the reader to the fundamentals of a novel algorithmic paradigm in control of dynamical systems called *online control*. The new approach is closely related to the decision-making framework of online convex optimization. This methodology is applied together with convex relaxations to obtain new methods that yield provable guarantees even in the absence of distributional assumptions.

The differences between online control and other control frameworks stem from its unorthodox objective. In optimal control, robust control, and other control methodologies that assume stochastic noise, the goal is to design controllers that perform as well as an *ex ante* optimal strategy. In online control, both the cost functions as well as the perturbations to the dynamical model are chosen online by an adversary and, therefore, the optimal policy is not defined a priori. Instead, the objective of an online controller is to attain low regret against the best policy chosen in hindsight from a benchmark class of policies.

To achieve this objective, online control algorithms adapt the decision making framework of online convex optimization to stateful systems, and are based on iterative mathematical optimization algorithms. These algorithms confer finite-time regret bounds and computational complexity guarantees. After an introduction to the core algorithms and principles, we survey recent advances and extensions of this methodology to time-varying systems, deep neural controllers, partial observation, system identification, planning and model-free reinforcement learning.

Key words and phrases: Nonstochastic control, online learning, LQR.

REFERENCES

ABBASI-YADKORI, Y., BARTLETT, P. and KANADE, V. (2014). Tracking adversarial targets. In *International Conference on Machine Learning* 369–377. PMLR.

ABBASI-YADKORI, Y. and SZEPESVÁRI, C. (2011). Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the* 24*th Annual Conference on Learning Theory* 1–26.

AGARWAL, N., BULLINS, B., HAZAN, E., KAKADE, S. and SINGH, K. (2019). Online control with adversarial disturbances. In *International Conference on Machine Learning* 111–119. PMLR.

AGARWAL, N., HAZAN, E., MAJUMDAR, A. and SINGH, K. (2021). A regret minimization approach to iterative learning control. In *International Conference on Machine Learning* 100–109. PMLR.

AGARWAL, N., HAZAN, E. and SINGH, K. (2019). Logarithmic regret for online control. In *Advances in Neural Information Processing Systems* 10175–10184.

AGARWAL, N., SUO, D., CHEN, X. and HAZAN, E. (2023). Spectral state space models. arXiv preprint. Available at arXiv:2312.06837.

ANAVA, O., HAZAN, E. and MANNOR, S. (2013). Online Convex Optimization Against Adversaries with Memory and Application to Statistical Arbitrage. arXiv preprint. Available at arXiv:1302.6937.

ANDERSON, J., DOYLE, J. C., LOW, S. H. and MATNI, N. (2019). System level synthesis. *Annu. Rev. Control* **47** 364–393. MR3973221 https://doi.org/10.1016/j.arcontrol.2019.03.006

ANDREW, L., BARMAN, S., LIGETT, K., LIN, M., MEYERSON, A., ROYTMAN, A. and WIERMAN, A. (2013). A tale of two metrics: Simultaneous bounds on competitiveness and regret. In *Proceedings of the* 26*th Annual Conference on Learning Theory*. *Proceedings of Machine Learning Research* **30** 741–763. PMLR, Princeton, NJ.

ARORA, R., DEKEL, O. and TEWARI, A. (2012). Online bandit learning against an adaptive adversary: from regret to policy regret. arXiv preprint. Available at arXiv:1206.6400.

BABY, D. and WANG, Y.-X. (2022). Optimal dynamic regret in LQR control. *Adv. Neural Inf. Process. Syst.* **35** 24879–24892.

BERTSEKAS, D. P. (2007). *Dynamic Programming and Optimal Control. Vol. I*, 3rd ed. Athena Scientific, Belmont, MA.

BORODIN, A. and EL-YANIV, R. (1998). *Online Computation and Competitive Analysis*. Cambridge Univ. Press, New York. MR1617778

BRAHMBHATT, A., BUZAGLO, G., DRUCHYNA, S. and HAZAN, E. (2025a). *A New Approach to Controlling Linear Dynamical Sys-*

Elad Hazan is a Professor of Computer Science, Princeton University, Princeton, New Jersey 08540, USA (e-mail: ehazan@princeton.edu). Karan Singh is an Assistant Professor of Operations Research, Tepper School of Business, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA (e-mail: karansingh@cmu.edu).

tems.

BRAHMBHATT, A., BUZAGLO, G., DRUCHYNA, S. and HAZAN, E. (2025b). Efficient spectral control of partially observed linear dynamical systems. arXiv preprint. Available at arXiv:2505.20943.

CASSEL, A., COHEN, A. and KOREN, T. (2020). Logarithmic regret for learning linear quadratic regulators efficiently. In *International Conference on Machine Learning* 1328–1337. PMLR.

CASSEL, A. and KOREN, T. (2020). Bandit linear control. *Adv. Neural Inf. Process. Syst.* **33** 8872–8882.

CESA-BIANCHI, N. and LUGOSI, G. (2006). *Prediction, Learning, and Games*. Cambridge Univ. Press, Cambridge. MR2409394 https://doi.org/10.1017/CBO9780511546921

CHEN, X. and HAZAN, E. (2021). Black-box control for linear dynamical systems. In *Conference on Learning Theory* 1114–1143. PMLR.

CHEN, X. and HAZAN, E. (2023). Online control for meta-optimization. In *Thirty-Seventh Conference on Neural Information Processing Systems*.

COHEN, A., HASIDIM, A., KOREN, T., LAZIC, N., MANSOUR, Y. and TALWAR, K. (2018). Online linear quadratic control. In *International Conference on Machine Learning* 1029–1038.

COHEN, A., KOREN, T. and MANSOUR, Y. (2019). Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In *International Conference on Machine Learning* 1300–1309.

COHEN-KARLIK, E., MENUHIN-GRUMAN, I., GIRYES, R., COHEN, N. and GLOBERSON, A. (2022). Learning low dimensional state spaces with overparameterized recurrent neural nets. arXiv preprint. Available at arXiv:2210.14064.

DANIELY, A. and MANSOUR, Y. (2019). Competitive ratio vs regret minimization: Achieving the best of both worlds. In *Algorithmic Learning Theory* 2019. *Proc. Mach. Learn. Res.* (*PMLR*) **98** 333–368. PMLR. MR3932850

DEAN, S., MANIA, H., MATNI, N., RECHT, B. and TU, S. (2018). Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems* 4188–4197.

DEAN, S., MANIA, H., MATNI, N., RECHT, B. and TU, S. (2020). On the sample complexity of the linear quadratic regulator. *Found. Comput. Math.* **20** 633–679. MR4130536 https://doi.org/10.1007/s10208-019-09426-y

DOYLE, J. C., MATNI, N., WANG, Y.-S., ANDERSON, J. and LOW, S. (2017). System level synthesis: A tutorial. In 2017 *IEEE* 56*th Annual Conference on Decision and Control* (*CDC*) 2856–2867.

EVEN-DAR, E., KAKADE, S. M. and MANSOUR, Y. (2004). Experts in a Markov decision process. *Adv. Neural Inf. Process. Syst.* **17**.

FAN, M. K. H., TITS, A. L. and DOYLE, J. C. (1988). Robustness in the presence of joint parametric uncertainty and unmodeled dynamics. In 1988 *American Control Conference* 1195–1200. IEEE Comput. Soc., Los Alamitos.

FAZEL, M., GE, R., KAKADE, S. and MESBAHI, M. (2018). Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning* 1467–1476.

FIECHTER, C.-N. (1997). PAC adaptive control of linear systems. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory* 72–80.

FOSTER, D. J. and SIMCHOWITZ, M. (2020). Logarithmic regret for adversarial online control.

FURIERI, L., ZHENG, Y., PAPACHRISTODOULOU, A. and KAMGARPOUR, M. (2019). An input-output parametrization of stabilizing controllers: Amidst Youla and system level synthesis. *IEEE Control Syst. Lett.* **3** 1014–1019. MR4208882

GHAI, U., GUPTA, A., XIA, W., SINGH, K. and HAZAN, E. (2023). Online nonstochastic model-free reinforcement learning. arXiv preprint. Available at arXiv:2305.17552.

GHAI, U., SNYDER, D., MAJUMDAR, A. and HAZAN, E. (2021). Generating adversarial disturbances for controller verification. In *Learning for Dynamics and Control* 1192–1204. PMLR.

GOEL, G., AGARWAL, N., SINGH, K. and HAZAN, E. (2023). Best of both worlds in online control: Competitive ratio and policy regret. In *Learning for Dynamics and Control Conference* 1345–1356. PMLR.

GOEL, G. and HASSIBI, B. (2023). Competitive control. *IEEE Trans. Automat. Control* **68** 5162–5173. MR4639976 https://doi.org/10.1109/tac.2022.3218769

GOEL, G., LIN, Y., SUN, H. and WIERMAN, A. (2019). Beyond online balanced descent: An optimal algorithm for smoothed online optimization. *Adv. Neural Inf. Process. Syst.* **32**.

GOLOWICH, N., HAZAN, E., LU, Z., ROHATGI, D. and SUN, Y. J. (2024). Online control in population dynamics. arXiv preprint. Available at arXiv:2406.01799.

GRADU, P., HALLMAN, J. and HAZAN, E. (2020). Non-stochastic control with bandit feedback. *Adv. Neural Inf. Process. Syst.* **33** 10764–10774.

GRADU, P., HALLMAN, J., SUO, D., YU, A., AGARWAL, N., GHAI, U., SINGH, K., ZHANG, C., MAJUMDAR, A. et al. (2021). Deluca–a Differentiable Control Library: Environments, Methods, and Benchmarking. arXiv preprint. Available at arXiv:2102.09968.

GRADU, P., HAZAN, E. and MINASYAN, E. (2020). Adaptive regret for control of time-varying dynamics. arXiv preprint. Available at arXiv:2007.04393.

HARDT, M., MA, T. and RECHT, B. (2018). Gradient descent learns linear dynamical systems. *J. Mach. Learn. Res.* **19** Paper No. 29, 44. MR3862436

HAZAN, E. (2016). Introduction to online convex optimization. *Found. Trends Optim.* **2** 157–325.

HAZAN, E., AGARWAL, A. and KALE, S. (2007). Logarithmic regret algorithms for online convex optimization. *Mach. Learn.* **69** 169–192.

HAZAN, E., KAKADE, S. M. and SINGH, K. (2020). The nonstochastic control problem. In *Algorithmic Learning Theory. Proc. Mach. Learn. Res.* (*PMLR*) **117** 408–421. PMLR. MR4165397

HAZAN, E., LEE, H., SINGH, K., ZHANG, C. and ZHANG, Y. (2018). Spectral filtering for general linear dynamical systems. In *Advances in Neural Information Processing Systems* 4634–4643.

HAZAN, E. and SINGH, K. (2023). Introduction to online nonstochastic control.

HAZAN, E., SINGH, K. and ZHANG, C. (2017). Learning linear dynamical systems via spectral filtering. In *Advances in Neural Information Processing Systems* 6702–6712.

HU, Y., WIERMAN, A. and QU, G. (2022). On the sample complexity of stabilizing lti systems on a single trajectory. *Adv. Neural Inf. Process. Syst.* **35** 16989–17002.

KAILATH, T., SAYED, A. H. and HASSIBI, B. (2000). *Linear Estimation*. Prentice Hall, New York.

KALMAN, R. E. (1960). A new approach to linear filtering and prediction problems. *J. Basic Eng.* **82** 35–45.

KALMAN, R. E. et al. (1960). Contributions to the theory of optimal control. *Bol. Soc. Mat. Mex.* **5** 102–119. MR0127472

KOZDOBA, M., MARECEK, J., TCHRAKIAN, T. and MANNOR, S. (2019). On-line learning of linear dynamical systems: Exponential forgetting in Kalman filters. *Proc. AAAI Conf. Artif. Intell.* **33** 4098–4105.

LI, W. and TODOROV, E. (2004). Iterative linear quadratic regulator design for nonlinear biological movement systems. In *First International Conference on Informatics in Control, Automation and Robotics* 2 222–229. SciTePress.

LI, Y., DAS, S. and LI, N. (2021). Online optimal control with affine constraints. *Proc. AAAI Conf. Artif. Intell.* **35** 8527–8537.

LIN, Y., PREISS, J. A., ANAND, E., LI, Y., YUE, Y. and WIERMAN, A. (2024a). Online adaptive policy selection in time-varying systems: No-regret via contractive perturbations. *Adv. Neural Inf. Process. Syst.* **36**.

LIN, Y., PREISS, J. A., XIE, F., ANAND, E., CHUNG, S.-J., YUE, Y. and WIERMAN, A. (2024b). Online policy optimization in unknown nonlinear systems. In *The Thirty Seventh Annual Conference on Learning Theory* 3475–3522. PMLR.

LYAPUNOV, A. M. (1992). The general problem of the stability of motion. *Internat. J. Control* **55** 531–534. Translated by A. T. Fuller from Édouard Davaux's French translation (1907) of the 1892 Russian original, with an editorial (historical introduction) by Fuller, a biography of Lyapunov by V. I. Smirnov, and the bibliography of Lyapunov's works collected by J. F. Barrett, Lyapunov centenary issue. MR1154209 https://doi.org/10.1080/00207179208934253

MANIA, H., TU, S. and RECHT, B. (2019). Certainty equivalence is efficient for linear quadratic control. In *Advances in Neural Information Processing Systems* 10154–10164.

MARSDEN, A. and HAZAN, E. (2025). Dimension-free regret for learning asymmetric linear dynamical systems. arXiv preprint. Available at arXiv:2502.06545.

MINASYAN, E., GRADU, P., SIMCHOWITZ, M. and HAZAN, E. (2021). Online control of unknown time-varying dynamical systems. *Adv. Neural Inf. Process. Syst.* **34** 15934–15945.

MORI, K. (2004). Elementary proof of controller parametrization without coprime factorizability. *IEEE Trans. Automat. Control* **49** 589–592. MR2049818 https://doi.org/10.1109/TAC.2004.825618

ROCKAFELLAR, R. T. (1987). Linear-quadratic programming and optimal control. *SIAM J. Control Optim.* **25** 781–814. MR0885199 https://doi.org/10.1137/0325045

SHALEV-SHWARTZ, S. et al. (2012). Online learning and online convex optimization. *Found. Trends Mach. Learn.* **4** 107–194.

SIMCHOWITZ, M. (2020). Making non-stochastic control (almost) as easy as stochastic. *Adv. Neural Inf. Process. Syst.* **33** 18318–18329.

SIMCHOWITZ, M., BOCZAR, R. and RECHT, B. (2019). Learning linear dynamical systems with semi-parametric least squares. In *Conference on Learning Theory* 2714–2802. PMLR.

SIMCHOWITZ, M. and FOSTER, D. (2020). Naive exploration is optimal for online LQR. In *Proceedings of the* 37*th International Conference on Machine Learning* (H. D. III and A. Singh, eds.). *Proceedings of Machine Learning Research* **119** 8937–8948. PMLR.

SIMCHOWITZ, M., SINGH, K. and HAZAN, E. (2020). Improper learning for non-stochastic control. In *Conference on Learning Theory* 3320–3436. PMLR.

SLOTINE, J.-J. E. and LI, W. (1991). *Applied Nonlinear Control*. Prentice Hall, New York.

STENGEL, R. F. (1994). *Optimal Control and Estimation*. Dover, New York. Corrected reprint of the 1986 original. MR1298435

SUGGALA, A., SUN, Y. J., NETRAPALLI, P. and HAZAN, E. (2024). Second order methods for bandit optimization and control. arXiv preprint. Available at arXiv:2402.08929.

SUN, Y. J., NEWMAN, S. and HAZAN, E. (2023). Optimal rates for bandit nonstochastic control. arXiv preprint. Available at arXiv:2305.15352.

YOULA, D. C., JABR, H. A. and BONGIORNO, J. J. JR. (1976). Modern Wiener-Hopf design of optimal controllers. II. The multivariable case. *IEEE Trans. Automat. Control* **AC-21** 319–338. MR0446637 https://doi.org/10.1109/tac.1976.1101223

ZAMES, G. (1981). Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses. *IEEE Trans. Automat. Control* **26** 301–320. MR0613539 https://doi.org/10.1109/TAC.1981.1102603

ZHANG, R. C., ZHENG, Y., LI, W. and LI, N. (2023). On the relationship of optimal state feedback and disturbance response controllers. *IFAC-PapersOnLine* **56** 7503–7508.

ZHAO, P., WANG, Y.-X. and ZHOU, Z.-H. (2022). Non-stationary online learning with memory and non-stochastic control. In *International Conference on Artificial Intelligence and Statistics* 2101–2133. PMLR.

ZHENG, Y., FURIERI, L., PAPACHRISTODOULOU, A., LI, N. and KAMGARPOUR, M. (2021). On the equivalence of Youla, system-level, and input-output parameterizations. *IEEE Trans. Automat. Control* **66** 413–420. MR4210429 https://doi.org/10.1109/tac.2020.2979785

ZHOU, K., DOYLE, J. C. and GLOVER, K. (1996). *Robust and Optimal Control*. Prentice Hall, Upper Saddle River, NJ.

ZINKEVICH, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the* 20*th International Conference on Machine Learning* (*icml*-03) 928–936.

Statistical Science
2025, Vol. 40, No. 4, 656–670
https://doi.org/10.1214/25-STS999

Stochastic Approximation and Reinforcement Learning: The Interface and a Little Beyond

Vivek Borkar

Abstract. This article traces the coevolution of stochastic approximation and reinforcement learning. Beginning with a brief historical background on both, it will highlight the points of intersection and the spin-offs thereof, summarizing results in reinforcement learning that have benefited from stochastic approximation theory. The article ends by outlining further developments and challenges that remain.

Key words and phrases: Stochastic approximation, reinforcement learning, almost supermartingales, "ODE" approach, convergence.

REFERENCES

[1] ABOUNADI, J., BERTSEKAS, D. and BORKAR, V. S. (2001). Learning algorithms for Markov decision processes with average cost. *SIAM J. Control Optim.* **40** 681–698. MR1871450 https://doi.org/10.1137/S0363012999361974

[2] ABOUNADI, J., BERTSEKAS, D. P. and BORKAR, V. (2002). Stochastic approximation for nonexpansive maps: Application to *Q*-learning algorithms. *SIAM J. Control Optim.* **41** 1–22. MR1920154 https://doi.org/10.1137/S0363012998346621

[3] AGARWAL, A., KAKADE, S. M., LEE, J. D. and MAHAJAN, G. (2021). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *J. Mach. Learn. Res.* **22** Paper No. 98. MR4279749

[4] AMARI, S. I. (1998). Natural gradient works efficiently in learning. *Neural Comput.* **10** 251–276.

[5] ANANTHARAM, V. and BORKAR, V. S. (2012). Stochastic approximation with long range dependent and heavy tailed noise. *Queueing Syst.* **71** 221–242. MR2925797 https://doi.org/10.1007/s11134-012-9283-0

[6] ANDRIEU, C., MOULINES, É. and PRIOURET, P. (2005). Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.* **44** 283–312. MR2177157 https://doi.org/10.1137/S0363012902417267

[7] ARTUR, B., ERMOL'EV, Y. M. and KANIOVSKIĬ, Y. M. (1983). A generalized urn problem and its applications. *Cybernetics* **19** 61–71.

[8] AVRACHENKOV, K. E. and BORKAR, V. S. (2022). Whittle index based Q-learning for restless bandits with average reward. *Automatica J. IFAC* **139** Paper No. 110186. MR4383041 https://doi.org/10.1016/j.automatica.2022.110186

[9] AVRACHENKOV, K. E., BORKAR, V. S., DOLHARE, H. P. and PATIL, K. (2021). Full gradient DQN reinforcement learning: A provably convergent scheme. In *Modern Trends in Controlled Stochastic Processes. V.III. Theory and Applications* (A. Piunovskyi and Y. Zhang, eds.). *Emerg. Complex. Comput.* **41** 192–220. Springer, Cham. MR4437151 https://doi.org/10.1007/978-3-030-76928-4_10

[10] AZAR, M. G., MUNOS, R., GHAVAMZADEH, M. and KAPPEN, H. (2011). Speedy Q-learning. In *Proceedings of the 24th International Conference on Neural Information Processing Systems* 2411–2419.

[11] BAIRD, L. (1995). Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings* 30–37. Morgan Kaufmann, San Mateo.

[12] BARTO, A. G., SUTTON, R. S. and ANDERSON, C. W. (1983). Neuron-like adaptive elements that can solve difficult learning control problems. *IEEE Trans. Syst. Man Cybern.* **13** 834–846.

[13] BASU, A., BHATTACHARYYA, T. and BORKAR, V. S. (2008). A learning algorithm for risk-sensitive cost. *Math. Oper. Res.* **33** 880–898. MR2464648 https://doi.org/10.1287/moor.1080.0324

[14] BENAIM, M. (1996). A dynamical system approach to stochastic approximations. *SIAM J. Control Optim.* **34** 437–472. MR1377706 https://doi.org/10.1137/S0363012993253534

[15] BENAIM, M. (1999). Dynamics of stochastic approximation algorithms. In *Le Séminaire de Probabilités* (J. Azéma, M. Emery, M. Ledoux and M. Yor, eds.) *Springer Lecture Notes in Mathematics* **1709** 1–68. Springer, Berlin.

[16] BENAÏM, M., HOFBAUER, J. and SORIN, S. (2005). Stochastic approximations and differential inclusions. *SIAM J. Control Optim.* **44** 328–348. MR2177159 https://doi.org/10.1137/S0363012904439301

[17] BENAÏM, M., HOFBAUER, J. and SORIN, S. (2006). Stochastic approximations and differential inclusions. II. Applications. *Math. Oper. Res.* **31** 673–695. MR2281223 https://doi.org/10.1287/moor.1060.0213

[18] BENVENISTE, A., MÉTIVIER, M. and PRIOURET, P. (1990). *Adaptive Algorithms and Stochastic Approximations. Applications of Mathematics* (*New York*) **22**. Springer, Berlin. MR1082341 https://doi.org/10.1007/978-3-642-75894-2

[19] BENVENISTE, A. and RUGET, G. (1982). A measure of the tracking capability of recursive stochastic algorithms with constant gains. *IEEE Trans. Automat. Control* **27** 639–649. MR0680322 https://doi.org/10.1109/TAC.1982.1102981

[20] BERTSEKAS, D. P. (1998). A new value iteration method for the average cost dynamic programming problem. *SIAM J. Control Optim.* **36** 742–759. MR1616546 https://doi.org/10.1137/S0363012995291609

Vivek Borkar is Retired Professor, Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai 400076, India (e-mail: borkar.vs@gmail.com).

[21] BERTSEKAS, D. P. (2020). *Rollout, Policy Iteration, and Distributed Reinforcement Learning*. *Athena Scientific Optimization and Computation Series*. Athena Scientific, Belmont, MA. MR4496005

[22] BERTSEKAS, D. P. (2019). *Reinforcement Learning and Optimal Control*. *Athena Scientific Optimization and Computation Series*. Athena Scientific, Belmont, MA. MR4315932

[23] BERTSEKAS, D. P. and TSITSIKLIS, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA.

[24] BHANDARI, J. and RUSSO, D. (2024). Global optimality guarantees for policy gradient methods. *Oper. Res.* **72** 1906–1927. MR4818024

[25] BHANDARI, J., RUSSO, D. and SINGAL, R. (2021). A finite time analysis of temporal difference learning with linear function approximation. *Oper. Res.* **69** 950–973. MR4280425

[26] BHATNAGAR, S., BORKAR, V. S. and GUIN, S. (2023). Actor-critic or critic-actor? A tale of two time scales. *IEEE Control Syst. Lett.* **7** 2671–2676. MR4629278 https://doi.org/10.1109/lcsys.2023.3288931

[27] BHATNAGAR, S. and LAKSHMANAN, K. (2012). An online actor-critic algorithm with function approximation for constrained Markov decision processes. *J. Optim. Theory Appl.* **153** 688–708. MR2915592 https://doi.org/10.1007/s10957-012-9989-5

[28] BHATNAGAR, S., SUTTON, R. S., GHAVAMZADEH, M. and LEE, M. (2009). Natural actor-critic algorithms. *Automatica J. IFAC* **45** 2471–2482. MR2889304 https://doi.org/10.1016/j.automatica.2009.07.008

[29] BORKAR, V. S. (1997). Stochastic approximation with two time scales. *Systems Control Lett.* **29** 291–294. MR1432654 https://doi.org/10.1016/S0167-6911(97)90015-3

[30] BORKAR, V. S. (2000). Asynchronous stochastic approximation. *SIAM J. Control Optim.* **38** 662–663.

[31] BORKAR, V. S. (2001). A sensitivity formula for risk-sensitive cost and the actor-critic algorithm. *Systems Control Lett.* **44** 339–346. MR2021952 https://doi.org/10.1016/S0167-6911(01)00152-9

[32] BORKAR, V. S. (2002). On the lock-in probability of stochastic approximation. *Combin. Probab. Comput.* **11** 11–20. MR1888179 https://doi.org/10.1017/S0963548301004953

[33] BORKAR, V. S. (2002). *Q*-learning for risk-sensitive control. *Math. Oper. Res.* **27** 294–311. MR1908528 https://doi.org/10.1287/moor.27.2.294.324

[34] BORKAR, V. S. (2003). Avoidance of traps in stochastic approximation. *Systems Control Lett.* **50** 1–9. MR2011930 https://doi.org/10.1016/S0167-6911(03)00118-X

[35] BORKAR, V. S. (2005). An actor-critic algorithm for constrained Markov decision processes. *Systems Control Lett.* **54** 207–213. MR2115538 https://doi.org/10.1016/j.sysconle.2004.08.007

[36] BORKAR, V. S. (2006). Stochastic approximation with 'controlled Markov' noise. *Systems Control Lett.* **55** 139–145. MR2187843 https://doi.org/10.1016/j.sysconle.2005.06.005

[37] BORKAR, V. S. (2023). *Stochastic Approximation*: *A Dynamical Systems Viewpoint*, 2nd ed. *Texts and Readings in Mathematics* **48**. Springer, Singapore. MR4755318 https://doi.org/10.1007/978-981-99-8277-6

[38] BORKAR, V. S. and KONDA, V. R. (1997). The actor-critic algorithm as multi-time-scale stochastic approximation. *Sādhanā* **22** 525–543. MR1610035 https://doi.org/10.1007/BF02745577

[39] BORKAR, V. S. and MEYN, S. P. (2000). The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Control Optim.* **38** 447–469. MR1741148 https://doi.org/10.1137/S0363012997331639

[40] BORKAR, V. S. and MEYN, S. P. (2002). Risk-sensitive optimal control for Markov decision processes with monotone cost. *Math. Oper. Res.* **27** 192–209. MR1886226 https://doi.org/10.1287/moor.27.1.192.334

[41] BORKAR, V. S., PINTO, J. and PRABHU, T. (2009). A new learning algorithm for optimal stopping. *Discrete Event Dyn. Syst.* **19** 91–113. MR2478127 https://doi.org/10.1007/s10626-008-0055-2

[42] BORKAR, V. S. and SHAH, D. A. (2024). Remarks on differential inclusion limits of stochastic approximation. *Pure Appl. Funct. Anal.* **9** 645–653. MR4816646

[43] BRANDIÈRE, O. (1998). Some pathological traps for stochastic approximation. *SIAM J. Control Optim.* **36** 1293–1314. MR1618037 https://doi.org/10.1137/S036301299630759X

[44] BRANDIÈRE, O. and DUFLO, M. (1996). Les algorithmes stochastiques contournent-ils les pièges? *Ann. Inst. Henri Poincaré Probab. Stat.* **32** 395–427. MR1387397

[45] BUCKLEW, J. A., KURTZ, T. G. and SETHARES, W. A. (1993). Weak convergence and local stability properties of fixed step size recursive algorithms. *IEEE Trans. Inf. Theory* **39** 966–978. MR1237722 https://doi.org/10.1109/18.256503

[46] BUSH, R. R. and MOSTELLER, F. (1955). *Stochastic Models for Learning*. Wiley, New York.

[47] CAO, X.-R. (2007). *Stochastic Learning and Optimization*: *A Sensitivity-Based Approach*. Springer, New York. MR2414445 https://doi.org/10.1007/978-0-387-69082-7

[48] CHANDAK, S., BORKAR, V. S. and DODHIA, P. (2022). Concentration of contractive stochastic approximation and reinforcement learning. *Stoch. Syst.* **12** 411–430. MR4561276 https://doi.org/10.1287/stsy.2022.0097

[49] CHEN, H.-F. (2002). *Stochastic Approximation and Its Applications*. *Nonconvex Optimization and Its Applications* **64**. Kluwer Academic, Dordrecht. MR1942427

[50] CHEN, S., DEVRAJ, A., LU, F., BUSIC, A. and MEYN, S. P. (2020). Zap Q-learning with nonlinear function approximation. In *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems*.

[51] CHEN, Z., MAGULURI, S. T., SHAKKOTTAI, S. and SHANMUGAM, K. (2020). Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. *Adv. Neural Inf. Process. Syst.* **33** 8223–8234.

[52] CHEN, Z., MAGULURI, S. T., SHAKKOTTAI, S. and SHANMUGAM, K. (2024). A Lyapunov theory for finite-sample guarantees of asynchronous Q-learning and TD-learning variants. arXiv preprint. Available at arXiv:2102.01567.

[53] CHEN, Z., MAGULURI, S. T. and ZUBELDIA, M. (2025). Concentration of contractive stochastic approximation: Additive and multiplicative noise. *Ann. Appl. Probab.* **35** 1298–1352. MR4897761 https://doi.org/10.1214/24-aap2143

[54] CHRISTIAN, B. (2021). *The Alignment Problem*: *How Can Machines Learn Human Values?* Atlantic Books, London.

[55] CHUNG, K. L. (1954). On a stochastic approximation method. *Ann. Math. Statist.* **25** 463–483. MR0064365 https://doi.org/10.1214/aoms/1177728716

[56] DEREVITSKII, D. P. and FRADKOV, A. L. (1974). Two models for analyzing the dynamics of adaptation algorithms. *Autom. Remote Control* **35** 59–67.

[57] DEVRAJ, A. and MEYN, S. P. (2017). Zap Q-learning. In *Proceedings of the 31st Conference on Neural Information Processing Systems*. Long Beach, CA.

[58] DUFF, M. O. (1995). Q-learning for bandit problems. In *Machine Learning Proceedings* 1995, 209–217 Morgan Kaufman.

[59] DUFLO, M. (1997). *Random Iterative Models*. *Applications of Mathematics* (*New York*) **34**. Springer, Berlin. MR1485774 https://doi.org/10.1007/978-3-662-12880-0

[60] DUPUIS, P. (1988). Large deviations analysis of some recursive algorithms with state dependent noise. *Ann. Probab.* **16** 1509–1536. MR0958200

[61] DUPUIS, P. and KUSHNER, H. J. (1989). Stochastic approximation and large deviations: Upper bounds and w.p.1 convergence. *SIAM J. Control Optim.* **27** 1108–1135. MR1009340 https://doi.org/10.1137/0327059

[62] FABIAN, V. (1968). On asymptotic normality in stochastic approximation. *Ann. Math. Statist.* **39** 1327–1332. MR0231429 https://doi.org/10.1214/aoms/1177698258

[63] FATHI, M. and FRIKHA, N. (2013). Transport-entropy inequalities and deviation estimates for stochastic approximations schemes. *Electron. J. Probab.* **18** no. 67. MR3084653 https://doi.org/10.1214/EJP.v18-2586

[64] FILIPPOV, A. F. (1988). *Differential Equations with Discontinuous Righthand Sides*. *Mathematics and Its Applications* (*Soviet Series*) **18**. Kluwer Academic, Dordrecht. MR1028776 https://doi.org/10.1007/978-94-015-7793-9

[65] FLAXMAN, A. D., KALAI, A. T. and MCMAHAN, H. B. (2005). Online convex optimization in the bandit setting: Gradient descent without a gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 385–394. ACM, New York. MR2298287

[66] FRADKOV, A. L. (2011). Continuous-time averaged models of discrete-time stochastic systems: Survey and open problems. In *Proceedings of the* 50*th IEEE Conference on Decision and Control and European Control Conference* (*CDC-ECC*), 2076–2081. Orlando, FL.

[67] FRIKHA, N. and MENOZZI, S. (2012). Concentration bounds for stochastic approximations. *Electron. Commun. Probab.* **17** no. 47. MR2988393 https://doi.org/10.1214/ecp.v17-1952

[68] GITTINS, J. C. (1979). Bandit processes and dynamic allocation indices. *J. Roy. Statist. Soc. Ser. B, Methodol.* **41** 148–177. MR0547241

[69] GOSAVI, A. (2004). Reinforcement learning for long-run average cost. *European J. Oper. Res.* **155** 654–674. MR2054686 https://doi.org/10.1016/S0377-2217(02)00874-3

[70] GOSAVI, A. (2009). Reinforcement learning: A tutorial survey and recent advances. *INFORMS J. Comput.* **21** 178–192. MR2549123 https://doi.org/10.1287/ijoc.1080.0305

[71] HULT, H., LINDHE, A., NYQUIST, P. and WU, G. J. (2025). A weak convergence approach to large deviations for stochastic approximations. arXiv preprint. Available at arXiv:2502.02529.

[72] JAAKKOLA, T., JORDAN, M. I. and SINGH, S. P. (1994). On the convergence of stochastic iterative dynamic programming algorithms. *Neural Comput.* **6** 1185–1201.

[73] JOHN, I., KAMANCHI, C. and BHATNAGAR, S. (2020). Generalized speedy Q-learning. *IEEE Control Syst. Lett.* **4** 524–529. MR4211339

[74] JOSLIN, J. A. and HEUNIS, A. J. (2000). Law of the iterated logarithm for a constant-gain linear stochastic gradient algorithm. *SIAM J. Control Optim.* **39** 533–570. MR1788071 https://doi.org/10.1137/S0363012997331007

[75] KALYANAKRISHNAN, S., MISRA, N. and GOPALAN, A. (2016). Randomised procedures for initialising and switching actions in policy iteration. *Proc.* 30*th AAAI Conf. Artif. Intell.* **30** 3145–3151.

[76] KAMAL, S. (2010). On the convergence, lock-in probability, and sample complexity of stochastic approximation. *SIAM J. Control Optim.* **48** 5178–5192. MR2735521 https://doi.org/10.1137/090769363

[77] KAMAL, S. (2012). Stabilization of stochastic approximation by step size adaptation. *Systems Control Lett.* **61** 543–548. MR2910330 https://doi.org/10.1016/j.sysconle.2012.02.005

[78] KAMANCHI, C., DIDDIGI, R. B. and BHATNAGAR, S. (2020). Successive over-relaxation Q-learning. *IEEE Control Syst. Lett.* **4** 55–60. MR4211255

[79] KARANDIKAR, R. L. and VIDYASAGAR, M. (2021). Convergence of batch asynchronous stochastic approximation with applications to reinforcement learning. arXiv preprint. Available at arXiv:2109.03445.

[80] KARANDIKAR, R. L. and VIDYASAGAR, M. (2024). Convergence rates for stochastic approximation: Biased noise with unbounded variance, and applications. *J. Optim. Theory Appl.* **203** 2412–2450. MR4837247 https://doi.org/10.1007/s10957-024-02547-7

[81] KARMAKAR, P. and BHATNAGAR, S. (2018). Two time-scale stochastic approximation with controlled Markov noise and off-policy temporal-difference learning. *Math. Oper. Res.* **43** 130–151. MR3774637 https://doi.org/10.1287/moor.2017.0855

[82] KONDA, V. R. and BORKAR, V. S. (1999). Actor-critic-type learning algorithms for Markov decision processes. *SIAM J. Control Optim.* **38** 94–123. MR1740605 https://doi.org/10.1137/S036301299731669X

[83] KONDA, V. R. and TSITSIKLIS, J. N. (2003). On actor-critic algorithms. *SIAM J. Control Optim.* **42** 1143–1166. MR2044789 https://doi.org/10.1137/S0363012901385691

[84] KUSHNER, H. J. and CLARK, D. S. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. *Applied Mathematical Sciences* **26**. Springer, New York. MR0499560

[85] KUSHNER, H. J. and YIN, G. (1987). Asymptotic properties of distributed and communicating stochastic approximation algorithms. *SIAM J. Control Optim.* **25** 1266–1290. MR0905045 https://doi.org/10.1137/0325070

[86] LAI, T. L. (2003). Stochastic approximation. *Ann. Statist.* **31** 391–406.

[87] LAI, T. L. and YUAN, H. (2021). Stochastic approximation: From statistical origin to big-data, multidisciplinary applications. *Statist. Sci.* **36** 291–302. MR4255195 https://doi.org/10.1214/20-sts784

[88] LAUAND, C. K. and MEYN, S. P. (2024). Revisiting step-size assumptions in stochastic approximation. arXiv preprint. Available at arXiv:2405.17834.

[89] LEMMENS, B. and NUSSBAUM, R. (2012). *Nonlinear Perron–Frobenius Theory*. *Cambridge Tracts in Mathematics* **189**. Cambridge Univ. Press, Cambridge. MR2953648 https://doi.org/10.1017/CBO9781139026079

[90] LJUNG, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Trans. Automat. Control* **22** 551–575. MR0465458 https://doi.org/10.1109/tac.1977.1101561

[91] LU, F. and MEYN, S. P. (2023). Convex Q learning in a stochastic environment. In *Proceedings of the 62nd IEEE Conference on Decision and Control* 776–781. Singapore.

[92] MANNE, A. S. (1960). Linear programming and sequential decisions. *Manage. Sci.* **6** 259–267. MR0129022 https://doi.org/10.1287/mnsc.6.3.259

[93] MEERKOV, S. M. (1972). Simplified description of slow Markov walks. II. *Autom. Remote Control* **3** 404—414.

[94] MEYN, S. (2024). The projected Bellman equation in reinforcement learning. *IEEE Trans. Automat. Control* **69** 8323–8337. MR4835015 https://doi.org/10.1109/tac.2024.3409647

[95] MEYN, S. P. (2022). *Control Systems and Reinforcement Learning*. Cambridge Univ. Press, Cambridge, UK.

[96] MNIH, V., KAVUKCUOGLU, K., SILVER, D., RUSU, A. A., VENESS, J., BELLEMARE, M. G., GRAVES, A., RIEDMILLER, M., FIDJELAND, A. K. et al. (2015). Human-level control through deep reinforcement learning. *Nature* **518** 529–533.

[97] MOHARRAMI, M., MURTHY, Y., ROY, A. and SRIKANT, R. (2025). A policy gradient algorithm for the risk-sensitive exponential cost MDP. *Math. Oper. Res.* **50** 431–458. MR4873496 https://doi.org/10.1287/moor.2022.0139

[98] NAJIM, K. and POZNYAK, A. S. (2014). *Learning Automata: Theory and Applications*. Elsevier, Oxford.

[99] NARENDRA, K. S. and THATHACHAR, M. A. L. (1974). Learning automata—a survey. *IEEE Trans. Syst. Man Cybern.* **4** 323–334.

[100] NEDIĆ, A. and BERTSEKAS, D. P. (2003). Least squares policy evaluation algorithms with linear function approximation. *Discrete Event Dyn. Syst.* **13** 79–110.

[101] NEVELSON, M. and KHASMINSKII, R. (1976). *Stochastic Approximation and Recursive Estimation. Trans. of Mathematical Monographs* **47**. American Math. Society, Providence, RI.

[102] NGUYEN, N. and YIN, G. (2023). Stochastic approximation with discontinuous dynamics, differential inclusions, and applications. *Ann. Appl. Probab.* **33** 780–823.

[103] PAGARE, T., BORKAR, V. S. and AVRACHENKOV, K. (2023). Full gradient deep reinforcement learning for average-reward criterion. In *5th Conference on Learning for Dynamics and Control* 235–247. PMLR.

[104] PELLETIER, M. (1998). On the almost sure asymptotic behaviour of stochastic algorithms. *Stochastic Process. Appl.* **78** 217–244.

[105] PELLETIER, M. (1999). An almost sure central limit theorem for stochastic approximation algorithms. *J. Multivariate Anal.* **71** 76–93.

[106] PEMANTLE, R. (1990). Nonconvergence to unstable points in urn models and stochastic approximations. *Ann. Probab.* **18** 698–712.

[107] PEZESHKI-ESFAHANI, H. and HEUNIS, A. J. (1997). Strong diffusion approximations for recursive stochastic algorithms. *IEEE Trans. Inf. Theory* **43** 512–523.

[108] POWELL, W. B. (2017). *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Wiley, Hoboken, NJ.

[109] POWELL, W. B. (2017). *Reinforcement Learning and Stochastic Optimization: A Unified Framework for Sequential Decisions*. Wiley, Hoboken, NJ.

[110] PRABUCHANDRAN, K. J., BHATNAGAR, S. and BORKAR, V. S. (2016). Actor-critic algorithms with online feature adaptation. *ACM Trans. Model. Comput. Simul.* **26** 1–26.

[111] PRASHANTH, L. A. and FU, M. C. (2022). Risk-sensitive reinforcement learning via policy gradient search. *Found. Trends Mach. Learn.* **15** 537–693.

[112] QU, G. and WIERMAN, A. (2020). Finite-time analysis of asynchronous stochastic approximation and Q-learning. In *Proceedings of the 33rd Annual Conference on Learning Theory* 3185–3205.

[113] RAJ, A., ZHU, L., GURBUZBALABAN, M. and SIMSEKLI, U. (2023). Algorithmic stability of heavy-tailed SGD with general loss functions. In *Proceedings of* 40*th International Conference on Machine Learning*, 28578–28597. PMLR, Honolulu, HI.

[114] RAMASWAMY, A. and BHATNAGAR, S. (2018). Stability of stochastic approximations with "controlled Markov" noise and temporal difference learning. *IEEE Trans. Automat. Control* **64** 2614–2620.

[115] RASHID, T., SAMVELYAN, M., DE WITT, C. S., FARQUHAR, G., FOERSTER, J. and WHITESON, S. (2020). Monotonic value function factorisation for deep multi-agent reinforcement learning. *J. Mach. Learn. Res.* **21** 1–51.

[116] ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Statist.* **22** 400–407.

[117] ROBBINS, H. and SIEGMUND, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics* (J. S. Rustagi, ed.) 233–257. Academic Press, San Diego.

[118] ROY, A., BORKAR, V. S., KARANDIKAR, A. and CHAPORKAR, P. (2021). Online reinforcement learning of optimal threshold policies for Markov decision processes. *IEEE Trans. Automat. Control* **67** 3722–3729.

[119] RUMMERY, G. A. and NIRANJAN, M. (1994). On-line Q-learning using connectionist systems. Report CUED/F-INFENG/TR 166, Cambridge Univ. Engineering Department.

[120] SACKS, J. (1958). Asymptotic distribution of stochastic approximation procedures. *Ann. Math. Statist.* **29** 373–405.

[121] SARYKALIN, S., SERRAINO, G. and URYASEV, S. (2008). Value-at-risk vs. conditional value-at-risk in risk management and optimization'. In *State-of-the-art decision-making tools in the information-intensive age*. Tutorials in Operations Research, INFORMS 270-294.

[122] SCHULMAN, J., LEVINE, S., ABBEEL, P., JORDAN, M. and MORITZ, P. (2015). Trust region policy optimization. In *Proceedings of the International Conference on Machine Learning* 1889–1897. PMLR.

[123] SCHULMAN, J., WOLSKI, F., DHARIWAL, P., RADFORD, A. and KLIMOV, O. (2017). Proximal policy optimization algorithms. arXiv preprint. Available at arXiv:1707.06347.

[124] SCHULTZ, W., DAYAN, P. and MONTAGUE, P. R. (1997). A neural substrate of prediction and reward. *Science* **275** 1593–1599.

[125] SUTTON, R. S. and BARTO, A. G. (2018). *Reinforcement Learning: An Introduction*, 2nd ed. MIT Press, Cambridge, MA.

[126] THOPPE, G. and BORKAR, V. S. (2019). A concentration bound for stochastic approximation via Alekseev's formula. *Stoch. Syst.* **9** 1–26.

[127] THORNDIKE, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychol. Rev.: Monogr. Suppl.* **2**.

[128] TSITSIKLIS, J. N. (1994). Asynchronous stochastic approximation and Q-learning. *Mach. Learn.* **16** 185–202.

[129] TSITSIKLIS, J. N. (2007). NP-hardness of checking the unichain condition in average cost MDPs. *Oper. Res. Lett.* **35** 319–323.

[130] TSITSIKLIS, J. N. and VAN ROY, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Trans. Automat. Control* **42** 674–690.

[131] TSITSIKLIS, J. N. and VAN ROY, B. (1999). Average cost temporal-difference learning. *Automatica* **35** 1799–1808.

[132] VIDYASAGAR, M. (2023). Convergence of stochastic approximation via martingale and converse Lyapunov methods. *Math. Control Signals Systems* **35** 351–374.

[133] WAN, Y., NAIK, A. and SUTTON, RṠ. (2021). Learning and planning in average-reward Markov decision processes. In *Proceedings of the 38th International Conference on Machine Learning* 10653–10662.

[134] WATKINS, C. J. C. H. (1989). Learning from delayed rewards. Ph.D. thesis, King's College, Univ. Cambridge.

[135] WATKINS, C. J. C. H. and DAYAN, P. (1992). Q-learning. *Mach. Learn.* **8** 279–292.

[136] WHITTLE, P. (1988). Restless bandits: Activity allocation in a changing world. *J. Appl. Probab.* **25** 287–298.

[137] WIENER, N. (2019). *Cybernetics, or Control and Communication in the Animal and the Machine*. MIT Press, Cambridge, MA.

[138] WILLIAMS, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.* **8** 229–256.

[139] YU, H. and BERTSEKAS, D. P. (2009). Convergence results for some temporal difference methods based on least squares. *IEEE Trans. Automat. Control* **54** 1515–1531.

Statistical Science
2025, Vol. 40, No. 4, 671–691
https://doi.org/10.1214/25-STS1017

Replicable Bandits for Digital Health Interventions

Kelly W. Zhang, Nowell Closser, Anna L. Trella and Susan A. Murphy

Abstract. Adaptive treatment assignment algorithms, such as bandit algorithms, are increasingly used in digital health intervention clinical trials. Frequently, the data collected from these trials is used to conduct causal inference and related data analyses to decide how to refine the intervention and whether to roll out the intervention more broadly. This work studies inference for estimands that depend on the adaptive algorithm itself; a simple example is the mean reward under the adaptive algorithm. Specifically, we investigate the replicability of statistical analyses concerning such estimands when using data from trials deploying adaptive treatment assignment algorithms. We demonstrate that many standard statistical estimators can be inconsistent and fail to be replicable across repetitions of the clinical trial, even as the sample size grows large. We show that this nonreplicability is intimately related to properties of the adaptive algorithm itself. We introduce a formal definition of a "replicable bandit algorithm" and prove that under such algorithms, a wide variety of common statistical estimators are guaranteed to be consistent and asymptotically normal. We present both theoretical results and simulation studies based on a mobile health oral health self-care intervention. Our findings underscore the importance of designing adaptive algorithms with replicability in mind, especially for settings like digital health, where deployment decisions rely heavily on replicated evidence. We conclude by discussing open questions on the connections between algorithm design, statistical inference, and experimental replicability.

Key words and phrases: bandit algorithms, digital health, replicability, adaptive treatment assignment.

REFERENCES

[1] ALBERS, N., NEERINCX, M. A. and BRINKMAN, W.-P. (2022). Addressing people's current and future states in a reinforcement learning algorithm for persuading to quit smoking and to be physically active. *PLoS ONE* **17** e0277295.

[2] AMAGAI, S., PILA, S., KAAT, A. J., NOWINSKI, C. J. and GERSHON, R. C. (2022). Challenges in participant engagement and retention using mobile health apps: Literature review. *J. Med. Internet Res*. **24** e35120.

[3] ANAND, K. and ZEEVI, A. (2021). A closer look at the worst-case behavior of multi-armed bandit algorithms. *Adv. Neural Inf. Process. Syst*. **34** 8807–8819.

[4] ATHEY, S., BERGSTROM, K., HADAD, V., JAMISON, J. C., ÖZLER, B., PARISOTTO, L. and DOHBIT SAMA, J. (2023). Can personalized digital counseling improve consumer search for modern contraceptive methods? *Sci. Adv.* **9** eadg4420.

[5] ATKINSON, A. C. (2004). Adaptive biased-coin designs for clinical trials with several treatments. *Discuss. Math. Probab. Stat*. **24** 85–108. MR2118925

[6] ATKINSON, A. C. and BISWAS, A. (2017). Optimal response and covariate-adaptive biased-coin designs for clinical trials with continuous multivariate or longitudinal responses. *Comput. Statist. Data Anal*. **113** 297–310. MR3662409 https://doi.org/10.1016/j.csda.2016.05.022

[7] BATTALIO, S. L., CONROY, D. E., DEMPSEY, W., LIAO, P., MENICTAS, M., MURPHY, S., NAHUM-SHANI, I., QIAN, T., KUMAR, S. et al. (2021). Sense2stop: A micro-randomized trial using wearable sensors to optimize a just-in-time-adaptive stress management intervention for smoking relapse prevention. *Contemp. Clin. Trials* **109** 106534.

[8] BELTZER, M. L., AMEKO, M. K., DANIEL, K. E., DAROS, A. R., BOUKHECHBA, M., BARNES, L. E. and

Kelly W. Zhang is an Assistant Professor at Imperial College London, London, United Kingdom (e-mail: kelly.zhang@imperial.ac.uk). Nowell Closser is a Ph.D. student, Department of Statistics, Harvard University, Cambridge, Massachusetts, USA (e-mail: nowellclosser@g.harvard.edu). Anna L. Trella is a Ph.D. student at Harvard University, Cambridge, Massachusetts 02138, USA (e-mail: annatrella@g.harvard.edu). Susan Murphy is Professor, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, USA (e-mail: samurphy@g.harvard.edu).

TEACHMAN, B. A. (2022). Building an emotion regulation recommender algorithm for socially anxious individuals using contextual bandits. *Br. J. Clin. Psychol.* **61** 51–72.

[9] BIBAUT, A., DIMAKOPOULOU, M., KALLUS, N., CHAMBAZ, A. and DER VAN LAAN, M. (2021). Post-contextual-bandit inference. *Adv. Neural Inf. Process. Syst.* **34** 28548–28559.

[10] BIBAUT, A. and KALLUS, N. (2025). Demystifying inference after adaptive experiments. *Annu. Rev. Stat. Appl.* **12** 407–423. MR4879684

[11] BORUVKA, A., ALMIRALL, D., WITKIEWITZ, K. and MURPHY, S. A. (2018). Assessing time-varying causal effect moderation in mobile health. *J. Amer. Statist. Assoc.* **113** 1112–1121. MR3862343 https://doi.org/10.1080/01621459.2017.1305274

[12] CHAKRABORTY, B., MURPHY, S. and STRECHER, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Stat. Methods Med. Res.* **19** 317–343. MR2757118 https://doi.org/10.1177/0962280209105013

[13] CHOW, S.-C. and CHANG, M. (2008). Adaptive design methods in clinical trials–a review. *Orphanet J. Rare Dis.* **3** 1–13.

[14] CHOW, S.-C. and CHANG, M. (2017). Adaptive clinical trial design. *Quant. Methods IV/AIDS Res.* 41–62.

[15] CHOW, S.-C., CHANG, M. and PONG, A. (2005). Statistical consideration of adaptive methods in clinical development. *J. Biopharm. Statist.* **15** 575–591. MR2190570 https://doi.org/10.1081/BIP-200062277

[16] COHEN, J. (2013). Statistical power analysis for the behavioral sciences. Routledge.

[17] DESHPANDE, Y., MACKEY, L., SYRGKANIS, V. and TADDY, M. Accurate inference for adaptive linear models. In *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.). *Proceedings of Machine Learning Research* **80** 1194–1203. PMLR, Stockholmsmässan.

[18] DONALD, B. A. (2012). Adaptive clinical trials in oncology. *Nat. Rev. Clin. Oncol.* **9** 199–207.

[19] EATON, E., HUSSING, M., KEARNS, M. and SORRELL, J. (2024). Replicable reinforcement learning. *Adv. Neural Inf. Process. Syst.* **36**.

[20] ERRAQABI, A., LAZARIC, A., VALKO, M., BRUNSKILL, E. and LIU, Y.-E. (2017). Trading off rewards and errors in multi-armed bandits. In *Artificial Intelligence and Statistics* 709–717. PMLR.

[21] ESFANDIARI, H., KALAVASIS, A., KARBASI, A., KRAUSE, A., MIRROKNI, V. and VELEGKAS, G. (2022). Replicable bandits. arXiv preprint. Available at arXiv:2210.01898.

[22] EYSENBACH, B. and LEVINE, S. (2021). Maximum entropy rl (provably) solves some robust rl problems. arXiv preprint. Available at arXiv:2103.06257.

[23] FAN, L. and GLYNN, P. W. (2021). The fragility of optimized bandit algorithms. arXiv preprint. Available at arXiv:2109.13595.

[24] FARO, J. M., CHEN, J., FLAHIVE, J., NAGAWA, C. S., ORVEK, E. A., HOUSTON, T. K., ALLISON, J. J., PERSON, S. D., SMITH, B. M. et al. (2023). Effect of a machine learning recommender system and viral peer marketing intervention on smoking cessation: A randomized clinical trial. *JAMA Netw. Open* **6** e2250665–e2250665.

[25] FIGUEROA, C. A., AGUILERA, A., CHAKRABORTY, B., MODIRI, A., AGGARWAL, J., DELIU, N., SARKAR, U., WILLIAMS, J. J. and LYLES, C. R. (2021). Adaptive learning algorithms to optimize mobile applications for behavioral health: Guidelines for design decisions. *J. Amer. Med. Inform. Assoc.* **28** 1225–1234.

[26] FORMAN, E. M., GOLDSTEIN, S. P., CROCHIERE, R. J., BUTRYN, M. L., JUARASCIO, A. S., ZHANG, F. and FOSTER, G. D. (2019). Randomized controlled trial of ontrack, a just-in-time adaptive intervention designed to enhance weight loss. *Transl. Behav. Med.* **9** 989–1001.

[27] FREIDLING, T., ZHAO, Q. and GAO, Z. (2024). Selective randomization inference for adaptive experiments. arXiv preprint. Available at arXiv:2405.07026.

[28] GHOSH, S., GUO, Y., HUNG, P.-Y., COUGHLIN, L., BONAR, E., NAHUM-SHANI, I., WALTON, M. and MURPHY, S. (2024). rebandit: Random effects based online rl algorithm for reducing cannabis use. arXiv preprint. Available at arXiv:2402.17739.

[29] GOLDSTEIN, S. P., GRAHAM THOMAS, J., FOSTER, G. D., TURNER-MCGRIEVY, G., BUTRYN, M. L., HERBERT, J. D., MARTIN, G. J. and FORMAN, E. M. (2020). Refining an algorithm-powered just-in-time adaptive weight control intervention: A randomized controlled trial evaluating model performance and behavioral outcomes. *Health Inform. J.* **26** 2315–2331.

[30] GOODMAN, S. N., FANELLI, D. and IOANNIDIS, J. P. A. (2016). What does research reproducibility mean? *Sci. Transl. Med.* **8**. 341ps12–341ps12.

[31] GREENHILL, S., RANA, S., GUPTA, S., VELLANKI, P. and VENKATESH, S. (2020). Bayesian optimization for adaptive experimental design: A review. *IEEE Access* **8** 13937–13948.

[32] HADAD, V., HIRSHBERG, D. A., ZHAN, R., WAGER, S. and ATHEY, S. (2021). Confidence intervals for policy evaluation in adaptive experiments. *Proc. Natl. Acad. Sci. USA* **118** Paper No. e2014602118. MR4294058 https://doi.org/10.1073/pnas.2014602118

[33] HIRANO, K. and PORTER, J. R. (2012). Impossibility results for nondifferentiable functionals. *Econometrica* **80** 1769–1790. MR2977437 https://doi.org/10.3982/ECTA8681

[34] HIRANO, K. and PORTER, J. R. (2023). Asymptotic representations for sequential decisions, adaptive experiments, and batched bandits. arXiv preprint. Available at arXiv:2302.03117.

[35] HOEY, J., POUPART, P., VON BERTOLDI, A., CRAIG, T., BOUTILIER, C. and MIHAILIDIS, A. (2010). Automated handwashing assistance for persons with dementia using video and a partially observable Markov decision process. *Comput. Vis. Image Underst.* **114** 503–519.

[36] HOWARD, S. R., RAMDAS, A., MCAULIFFE, J. and SEKHON, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *Ann. Statist.* **49** 1055–1080. MR4255119 https://doi.org/10.1214/20-aos1991

[37] HU, F. and ROSENBERGER, W. F. (2006). *The Theory of Response-Adaptive Randomization in Clinical Trials*. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR2245329 https://doi.org/10.1002/047005588X

[38] HUDGENS, M. G. and HALLORAN, M. E. (2008). Toward causal inference with interference. *J. Amer. Statist. Assoc.* **103** 832–842. MR2435472 https://doi.org/10.1198/016214508000000292

[39] IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. MR3309951 https://doi.org/10.1017/CBO9781139025751

[40] IMPAGLIAZZO, R., LEI, R., PITASSI, T. and SORRELL, J. (2022). Reproducibility in learning. In *STOC '22—Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing* 818–831. ACM, New York. MR4490043

[41] JEFFRIES, N. and GELLER, N. L. (2015). Longitudinal clinical trials with adaptive choice of follow-up time. *Biometrics* **71** 469–477. MR3366251 https://doi.org/10.1111/biom.12287

[42] JENNISON, C. and TURNBULL, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. CRC Press/CRC, Boca Raton, FL. MR1710781

[43] KARBASI, A., VELEGKAS, G., YANG, L. and ZHOU, F. (2024). Replicability in reinforcement learning. *Adv. Neural Inf. Process. Syst.* **36**.

[44] KHAMARU, K., DESHPANDE, Y., LATTIMORE, T., MACKEY, L. and WAINWRIGHT, M. J. (2021). Near-optimal inference in adaptive linear regression. arXiv preprint. Available at arXiv:2107.02266.

[45] KHAN, O., EL MISTIRI, M., RIVERA, D. E., MARTIN, C. A. and HEKLER, E. (2022). A Kalman filter-based hybrid model predictive control algorithm for mixed logical dynamical systems: Application to optimized interventions for physical activity. In 2022 *IEEE* 61*st Conference on Decision and Control* (*CDC*) 2586–2593. IEEE, New York.

[46] KOMIYAMA, J., ITO, S., YOSHIDA, Y. and KOSHINO, S. (2024). Replicability is asymptotically free in multi-armed bandits. arXiv preprint. Available at arXiv:2402.07391.

[47] KUMAR KRISHNAMURTHY, S., ZHAN, R., ATHEY, S. and BRUNSKILL, E. (2023). Proportional response: Contextual bandits for simple and cumulative regret minimization. *Adv. Neural Inf. Process. Syst.* **36** 30255–30266.

[48] LABER, E. B., LIZOTTE, D. J., QIAN, M., PELHAM, W. E. and MURPHY, S. A. (2014). Dynamic treatment regimes: Technical challenges and applications. *Electron. J. Stat.* **8** 1225–1272. MR3263118 https://doi.org/10.1214/14-EJS920

[49] LABER, E. B. and MURPHY, S. A. (2011). Adaptive confidence intervals for the test error in classification. *J. Amer. Statist. Assoc.* **106** 904–913. MR2894746 https://doi.org/10.1198/jasa.2010.tm10053

[50] LATTIMORE, T. and SZEPESVÁRI, C. (2020). *Bandit Algorithms*. Cambridge Univ. Press, Cambridge.

[51] LAUFFENBURGER, J. C., YOM-TOV, E., KELLER, P. A., MCDONNELL, M. E., BESSETTE, L. G., FONTANET, C. P., SEARS, E. S., KIM, E., HANKEN, K. et al. (2021). Reinforcement learning to improve non-adherence for diabetes treatments by optimising response and customising engagement (reinforce): Study protocol of a pragmatic randomised trial. *BMJ Open* **11** e052091.

[52] LEEB, H. and PÖTSCHER, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* **21** 21–59. MR2153856 https://doi.org/10.1017/S0266466605050036

[53] LIAO, P., GREENEWALD, K., KLASNJA, P. and MURPHY, S. (2020). Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **4** 1–22.

[54] LIAO, P., KLASNJA, P., TEWARI, A. and MURPHY, S. A. (2016). Sample size calculations for micro-randomized trials in mHealth. *Stat. Med.* **35** 1944–1971. MR3513494 https://doi.org/10.1002/sim.6847

[55] LIU, X., DELIU, N., CHAKRABORTY, T., BELL, L. and CHAKRABORTY, B. (2025). Thompson sampling for zero-inflated count outcomes with an application to the Drink Less mobile health study. *Ann. Appl. Stat.* **19** 1403–1425. MR4912973 https://doi.org/10.1214/25-aoas2030

[56] LUEDTKE, A. R. and VAN DER LAAN, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist.* **44** 713–742. MR3476615 https://doi.org/10.1214/15-AOS1384

[57] MAYSTRE, L., RUSSO, D. and ZHAO, Y. (2023). Optimizing audio recommendations for the long-term: a reinforcement learning perspective. arXiv preprint. Available at arXiv:2302.03561.

[58] MIHAILIDIS, A., BOGER, J. N., CRAIG, T. and HOEY, J. (2008). The coach prompting system to assist older adults with dementia through handwashing: An efficacy study. *BMC Geriatr.* **8** 1–18.

[59] MINTZ, Y., ASWANI, A., KAMINSKY, P., FLOWERS, E. and FUKUOKA, Y. (2020). Nonstationary bandits with habituation and recovery dynamics. *Oper. Res.* **68** 1493–1516. MR4166309 https://doi.org/10.1287/opre.2019.1918

[60] MINTZ, Y., ASWANI, A., KAMINSKY, P., FLOWERS, E. and FUKUOKA, Y. (2023). Behavioral analytics for myopic agents. *European J. Oper. Res.* **310** 793–811. MR4594035 https://doi.org/10.1016/j.ejor.2023.03.034

[61] MOODIE, E. E. M. and RICHARDSON, T. S. (2010). Estimating optimal dynamic regimes: Correcting bias under the null. *Scand. J. Stat.* **37** 126–146. MR2675943 https://doi.org/10.1111/j.1467-9469.2009.00661.x

[62] MUSSI, M., DRAGO, S., RESTELLI, M. and METELLI, A. M. (2025). Factored-reward bandits with intermediate observations: Regret minimization and best arm identification. *Artificial Intelligence* **347** Paper No. 104362. MR4919687 https://doi.org/10.1016/j.artint.2025.104362

[63] NAHUM-SHANI, I., GREER, Z. M., TRELLA, A. L., ZHANG, K. W., CARPENTER, S. M., RUENGER, D., ELASHOFF, D., MURPHY, S. A. and SHETTY, V. (2024). Optimizing an adaptive digital oral health intervention for promoting oral self-care behaviors: Micro-randomized trial protocol. *Contemp. Clin. Trials* **139** 107464.

[64] NAIR, Y. and JANSON, L. (2023). Randomization tests for adaptively collected data. arXiv preprint. Available at arXiv:2301.05365.

[65] NATIONAL ACADEMIES OF SCIENCES, POLICY, GLOBAL AFFAIRS, BOARD ON RESEARCH DATA, INFORMATION, DIVISION ON ENGINEERING, PHYSICAL SCIENCES, COMMITTEE ON APPLIED, THEORETICAL STATISTICS, BOARD ON MATHEMATICAL SCIENCES (2019). *Reproducibility and Replicability in Science*. National Academies Press.

[66] NEYMAN, J. (1992). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. In *Breakthroughs in Statistics*: *Methodology and Distribution* 123–150. Springer, Berlin.

[67] OFFER-WESTORT, M., COPPOCK, A. and GREEN, D. P. (2019). Adaptive experimental design: Prospects and applications in political science. Available at SSRN 3364402.

[68] PARK, J. and LEE, U. (2023). Understanding disengagement in just-in-time mobile health interventions. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **7** 1–27.

[69] PARK, J. J., THORLUND, K. and MILLS, E. J. (2018). Critical concepts in adaptive clinical trials. *Clin. Epidemiol.* 343–351.

[70] PARMIGIANI, G. (2023). Defining replicability of prediction rules. *Statist. Sci.* **38** 543–556. MR4665025 https://doi.org/10.1214/23-sts891

[71] PIETTE, J. D., NEWMAN, S., KREIN, S. L., MARINEC, N., CHEN, J., WILLIAMS, D. A., EDMOND, S. N., DRISCOLL, M., LACHAPPELLE, K. M. et al. (2022). Artificial intelligence (ai) to improve chronic pain care: Evidence of ai learning. *Intell.-Based Med.* 100064.

[72] QIAN, T., WALTON, A. E., COLLINS, L. M., KLASNJA, P., LANZA, S. T., NAHUM-SHANI, I., RABBI, M., RUSSELL, M. A., WALTON, M. A. et al. (2022). The microrandomized trial for developing digital interventions: Experimental design and data analysis considerations. *Psychol. Methods*.

[73] QIAN, T., YOO, H., KLASNJA, P., ALMIRALL, D. and MURPHY, S. A. (2021). Estimating time-varying causal excursion effects in mobile health with binary outcomes. *Biometrika* **108** 507–527. MR4298759 https://doi.org/10.1093/biomet/asaa070

[74] RABBI, M., PFAMMATTER, A., ZHANG, M., SPRING, B., CHOUDHURY, T. et al. (2015). Automated personalized feedback for physical activity and dietary behavior change with mobile phones: A randomized controlled trial on adults. *JMIR mHealth uHealth* **3** e4160.

[75] ROMANO, J. P. and SHAIKH, A. M. (2012). On the uniform asymptotic validity of subsampling and the bootstrap. *Ann. Statist.* **40** 2798–2822. MR3097960 https://doi.org/10.1214/12-AOS1051

[76] ROSENBERGER, W. F., VIDYASHANKAR, A. N. and AGARWAL, D. K. (2001). Covariate-adjusted response-adaptive designs for binary response. *J. Biopharm. Statist.* **11** 227–236.

[77] RUBIN, D. B. (2008). Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* **75** 591–593.

[78] RUSSO, D. J., VAN ROY, B., KAZEROUNI, A., OSBAND, I., WEN, Z. et al. (2018). A tutorial on Thompson sampling. *Found. Trends Mach. Learn.* **11** 1–96.

[79] SCHÄFER, H. (2006). Adaptive designs from the viewpoint of an academic biostatistician. *Biom. J.* **48** 586–590. MR2247041 https://doi.org/10.1002/bimj.200610246

[80] SHAIKH, H., MODIRI, A., WILLIAMS, J. J. and RAFFERTY, A. N. (2019). Balancing student success and inferring personalized effects in dynamic experiments. In *EDM*.

[81] SIMCHI-LEVI, D. and WANG, C. (2023). Multi-armed bandit experimental design: Online decision-making and adaptive inference. In *International Conference on Artificial Intelligence and Statistics* 3086–3097. PMLR.

[82] SIMCHI-LEVI, D., WANG, C. and ZHENG, Z. (2023). Nonstationary experimental design under linear trends. *Adv. Neural Inf. Process. Syst.* **36** 32102–32116.

[83] SUTTON, R. S. and BARTO, A. G. (2018). *Reinforcement Learning*: *An Introduction*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3889951

[84] THOMPSON, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25** 285–294.

[85] TRAGOS, E., O'REILLY-MORGAN, D., GERACI, J., SHI, B., SMYTH, B., DOHERTY, C., LAWLOR, A. and HURLEY, N. (2023). Keeping people active and healthy at home using a reinforcement learning-based fitness recommendation framework. In 32*nd International Joint Conference on Artificial Intelligence*, *IJCAI*-23 6237–6245.

[86] TRELLA, A. L., ZHANG, K. W., CARPENTER, S. M., ELASHOFF, D., GREER, Z. M., NAHUM-SHANI, I., RUENGER, D., SHETTY, V. and MURPHY, S. A. (2024). Oralytics reinforcement learning algorithm.

[87] TRELLA, A. L., ZHANG, K. W., NAHUM-SHANI, I., SHETTY, V., DOSHI-VELEZ, F. and MURPHY, S. A. (2022). Designing reinforcement learning algorithms for digital interventions: Pre-implementation guidelines. *Algorithms* **15**.

[88] TRELLA, A. L., ZHANG, K. W. NAHUM-SHANI, I., SHETTY, V., DOSHI-VELEZ, F. and MURPHY, S. A. (2023). Reward design for an online reinforcement learning algorithm supporting oral self-care. In *Thirty-Fifth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-23)*.

[89] TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data*. *Springer Series in Statistics*. Springer, New York. MR2233926

[90] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

[91] VAN DER VAART, A. W. and WELLNER, J. A. (1996). Weak convergence. In *Weak Convergence and Empirical Processes* 16–28. Springer, Berlin.

[92] WAUDBY-SMITH, I., WU, L., RAMDAS, A., KARAMPATZIAKIS, N. and MINEIRO, P. (2024). Anytime-valid off-policy inference for contextual bandits. *ACM/IMS J. Data Sci.* **1** Art. 10. MR4922618

[93] WEI, Y., ZHENG, P., DENG, H., WANG, X., LI, X. and FU, H. (2020). Design features for improving mobile health intervention user engagement: Systematic review and thematic analysis. *J. Med. Internet Res.* **22** e21687.

[94] WHITE, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48** 817–838. MR0575027 https://doi.org/10.2307/1912934

[95] WU, H., TAN, S., LI, W., GARRARD, M., OBENG, A., DIMMERY, D., SINGH, S., WANG, H., JIANG, D. et al. (2022). Interpretable personalized experimentation. In *Proceedings of the* 28*th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 4173–4183.

[96] XU, Z., ZHANG, K. W. and MURPHY, S. A. (2024). The fallacy of minimizing local regret in the sequential task setting. arXiv preprint. Available at arXiv:2403.10946.

[97] YAO, J., BRUNSKILL, E., PAN, W., MURPHY, S. and DOSHI-VELEZ, F. (2021). Power constrained bandits. In *Proceedings of the* 6*th Machine Learning for Healthcare Conference* (K. Jung, S. Yeung, M. Sendak, M. Sjoding and R. Ranganath, eds.) *Proceedings of Machine Learning Research* **149**, 209–259. PMLR.

[98] YIN, G. and SHEN, Y. (2005). Adaptive design and estimation in randomized clinical trials with correlated observations. *Biometrics* **61** 362–369. MR2140907 https://doi.org/10.1111/j.1541-0420.2005.00333.x

[99] YING, M., KHAMARU, K. and ZHANG, C.-H. (2024). Adaptive linear estimating equations. *Adv. Neural Inf. Process. Syst.* **36**.

[100] YOM-TOV, E., FERARU, G., KOZDOBA, M., MANNOR, S., TENNENHOLTZ, M. and HOCHBERG, I. (2017). Encouraging physical activity in patients with diabetes: Intervention using a reinforcement learning system. *J. Med. Internet Res.* **19** e338.

[101] YU, B. and KUMBIER, K. (2020). Veridical data science. *Proc. Natl. Acad. Sci. USA* **117** 3920–3929. MR4075122 https://doi.org/10.1073/pnas.1901326117

[102] ZHAN, R., HADAD, V., HIRSHBERG, D. A. and ATHEY, S. (2021). Off-policy evaluation via adaptive weighting with data from contextual bandits. In *Proceedings of the* 27*th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* 2125–2135.

[103] ZHANG, K. W, CLOSSER, N., TRELLA, A. L and MURPHY, S. A (2025). Supplement to "Replicable bandits for digital health interventions." https://doi.org/10.1214/25-STS1017SUPP

[104] ZHANG, K. W., JANSON, L. and MURPHY, S. A. (2020). Inference for batched bandits. *Adv. Neural Inf. Process. Syst.* **33** 9818–9829.

[105] ZHANG, K. W., JANSON, L. and MURPHY, S. A. (2021). Statistical inference with m-estimators on adaptively collected data. *Adv. Neural Inf. Process. Syst.* **34** 7460–7471.

[106] ZHANG, K. W., JANSON, L. and MURPHY, S. A. (2022). Statistical inference after adaptive sampling for longitudinal data. arXiv preprint. Available at arXiv:2202.07098.

[107] ZHANG, L.-X., HU, F., CHEUNG, S. H. and CHAN, W. S. (2007). Asymptotic properties of covariate-adjusted response-adaptive designs. *Ann. Statist.* **35** 1166–1182. MR2341702 https://doi.org/10.1214/009053606000001424

[108] ZHOU, M., FUKUOKA, Y., MINTZ, Y., GOLDBERG, K., KAMINSKY, P., FLOWERS, E., ASWANI, A. et al. (2018). Evaluating machine learning–based automated personalized daily

step goals delivered through a mobile phone app: Randomized controlled trial. *JMIR mHealth uHealth* **6** e9117.

[109] ZHOU, M., MINTZ, Y., FUKUOKA, Y., GOLDBERG, K., FLOWERS, E., KAMINSKY, P., CASTILLEJO, A. and ASWANI, A. (2018). Personalizing mobile fitness apps using reinforcement learning. In *CEUR Workshop Proceedings* **2068**. NIH Public Access.

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

The purpose of the Institute is to foster the development and dissemination of the theory and applications of statistics and probability.

IMS OFFICERS

President: Kavita Ramanan, Division of Applied Mathematics, Brown University, Providence, RI 02912, USA

President-Elect: Richard J. Samworth, Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Cambridge, CB3 0WB, UK

Past President: Tony Cai, Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA 19104-6304, USA

Executive Secretary: Peter Hoff, Department of Statistical Science, Duke University, Durham, NC 27708-0251, USA

Treasurer: Jiashun Jin, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA

Program Secretary: Annie Qu, Department of Statistics, University of California, Irvine, Irvine, CA 92697-3425, USA

IMS EDITORS

The Annals of Statistics. *Editors*: Hans-Georg Müller, Department of Statistics, University of California, Davis, Davis, CA 95616, USA. Harrison Zhou, Department of Statistics and Data Science, Yale University, New Haven, CT 06520, USA

The Annals of Applied Statistics. *Editor-in-Chief*: Lexin Li, Department of Biostatistics and Epidemiology, University of California, Berkeley, Berkeley, CA 94720-7360, USA

The Annals of Probability. *Editors*: Paul Bourgade, Courant Institute of Mathematical Sciences, New York University, New York, NY 10012-1185, USA. Julien Dubedat, Department of Mathematics, Columbia University, New York, NY 10027, USA

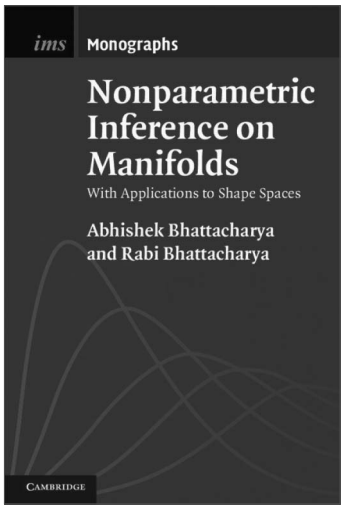
The Annals of Applied Probability. *Editors*: Jian Ding, School of Mathematical Sciences, Peking University, 100871, Bejing, China. Claudio Landim, IMPA, 22461-320, Rio de Janeiro, Brazil

Statistical Science. *Editor*: Moulinath Banerjee, Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA

The IMS Bulletin. *Editor*: Tati Howell, bulletin@imstat.org

The Institute of Mathematical Statistics presents

IMS MONOGRAPHS



Nonparametric Inference on Manifolds

With Applications to Shape Spaces

Abhishek Bhattacharya, Rabi Bhattacharya

This book introduces in a systematic manner a general nonparametric theory of statistics on manifolds, with emphasis on manifolds of shapes. The theory has important and varied applications in medical diagnostics, image analysis, and machine vision. An early chapter of examples establishes the effectiveness of the new methods and demonstrates how they outperform their parametric counterparts. Inference is developed for both intrinsic and extrinsic Fréchet means of probability distributions on manifolds, then applied to shape spaces defined as orbits of landmarks under a Lie group of transformations—in particular, similarity, reflection similarity, affine and projective transformations. In addition, nonparametric Bayesian theory is adapted and extended to manifolds for the purposes of density estimation, regression and classification. Ideal for statisticians who analyze manifold data and wish to develop their own methodology, this book is also of interest to probabilists, mathematicians, computer scientists and morphometricians with mathematical training.

IMS member? Claim your 40% discount: www.cambridge.org/ims

Hardback price US$51.00 (non-member price $85.00)

Cambridge University Press, in conjunction with the Institute of Mathematical Statistics, established the IMS Monographs and IMS Textbooks series of high-quality books. The Series Editors are Xiao-Li Meng, Susan Holmes, Ben Hambly, D. R. Cox and Alan Agresti.