Contents

Chapter 1

Exploration Techniques

1.1 Undirected Exploration

Description, citing epsilon-greedy policies, Softmax policies and Boltzmann policies for MABs and RL

Algorithm:

- Soft Q-learning [42]

1.2 Counter-based exploration and OFU

Description.
Algorithms:

- UCB1
- HOO algorithm from X-Armed Bandits [22]
- GPUCB [91]
- PixelCNN algorithm from Count-based exploration with neural density models [75]. This paper builds upon Unifying Count-Based Exploration and Intrinsic Motivation [14].

1.3 Value Difference and Recency-based exploration

Brief description, citing:

- Value-Difference based Exploration: AdaptiveControl between epsilon-Greedy and Softmax [99]
- Efficient Exploration In Reinforcement Learning [98] refers to recency-based exploraiton.

1.4 Intrinsic Motivation

Description, including: Unifying Count-Based Exploration and Intrinsic Motivation [14], talks about the connection between intrinsic motivation and counter-based exploration

Algorithms:

- Vime: Variational information maximizing exploration [45]
- Diversity-Inducing Policy Gradient[63], similarly to us, uses an exploration bonus based on diversity between distributions

1.5 Thompson Sampling

To conclude with: Why is Posterior Sampling Better than Optimism for Reinforcement Learning? [73]

Bibliography

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

[2] Agrawal, R. The continuum-armed bandit problem. *SIAM Journal on Control and Optimization 33*, 6 (1995), 1926–1951.

[3] Agrawal, R. Sample mean based index policies by o (log n) regret for the multi-armed bandit problem. *Advances in Applied Probability 27*, 4 (1995), 1054–1078.

[4] Amari, S.-I. Natural gradient works efficiently in learning. *Neural computation 10*, 2 (1998), 251–276.

[5] Amari, S.-I., and Douglas, S. C. Why natural gradient? In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)* (1998), vol. 2, IEEE, pp. 1213–1216.

[6] Antos, A., Szepesvári, C., and Munos, R. Fitted q-iteration in continuous action-space mdps. In *Advances in neural information processing systems* (2008), pp. 9–16.

[7] Atan, O., Tekin, C., and Schaar, M. Global multi-armed bandits with hölder continuity. In *Artificial Intelligence and Statistics* (2015), pp. 28–36.

[8] Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research 3*, Nov (2002), 397–422.

[9] Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning 47*, 2-3 (2002), 235–256.

[10] Auer, P., Ortner, R., and Szepesvári, C. Improved rates for the stochastic continuum-armed bandit problem. In *International Conference on Computational Learning Theory* (2007), Springer, pp. 454–468.

[11] Bach, F. R., and Blei, D. M., Eds. *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015* (2015), vol. 37 of *JMLR Workshop and Conference Proceedings*, JMLR.org.

[12] Baird III, L. C. Advantage updating. Tech. rep., WRIGHT LAB WRIGHT-PATTERSON AFB OH, 1993.

[13] Baxter, J., and Bartlett, P. L. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research 15* (2001), 319–350.

[14] Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems* (2016), pp. 1471–1479.

[15] Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., Eds. *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada* (2018).

[16] Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence.* Oxford university press, 2013.

[17] Brafman, R. I., and Tennenholtz, M. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research 3*, Oct (2002), 213–231.

[18] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym, 2016.

[19] Bubeck, S., Cesa-Bianchi, N., et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning 5*, 1 (2012), 1–122.

[20] Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. Bandits with heavy tail. *IEEE Transactions on Information Theory 59*, 11 (2013), 7711–7717.

[21] Bubeck, S., and Munos, R. Open loop optimistic planning. In *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010* (2010), pp. 477–489.

[22] Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. X-armed bandits. *Journal of Machine Learning Research 12*, May (2011), 1655–1695.

[23] Bubeck, S., Stoltz, G., Szepesvári, C., and Munos, R. Online optimization in x-armed bandits. In *Advances in Neural Information Processing Systems* (2009), pp. 201–208.

[24] Cesa-Bianchi, N., and Lugosi, G. Combinatorial bandits. *Journal of Computer and System Sciences 78*, 5 (2012), 1404–1422.

[25] Chapelle, O., and Li, L. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems* (2011), pp. 2249–2257.

[26] Chen, W., Wang, Y., Yuan, Y., and Wang, Q. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *The Journal of Machine Learning Research 17*, 1 (2016), 1746–1778.

[27] Chentanez, N., Barto, A. G., and Singh, S. P. Intrinsically motivated reinforcement learning. In *Advances in neural information processing systems* (2005), pp. 1281–1288.

[28] Choshen, L., Fox, L., and Loewenstein, Y. Dora the explorer: Directed outreaching reinforcement action-selection. *arXiv preprint arXiv:1804.04012* (2018).

[29] Chowdhury, S. R., and Gopalan, A. Online learning in kernelized markov decision processes. *CoRR abs/1805.08052* (2018).

[30] Cochran, W. G. *Sampling techniques.* John Wiley & Sons, 2007.

[31] Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 442–450.

[32] Dann, C., and Brunskill, E. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems* (2015), pp. 2818–2826.

[33] Dann, C., Lattimore, T., and Brunskill, E. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems* (2017), pp. 5713–5723.

[34] Degris, T., White, M., and Sutton, R. S. Off-policy actor-critic. *arXiv preprint arXiv:1205.4839* (2012).

[35] Deisenroth, M. P., Neumann, G., Peters, J., et al. A survey on policy search for robotics. *Foundations and Trends® in Robotics 2*, 1–2 (2013), 1–142.

[36] Dorato, P., Abdallah, C. T., Cerone, V., and Jacobson, D. H. *Linear-quadratic control: an introduction.* Prentice Hall Englewood Cliffs, NJ, 1995.

[37] Dy, J. G., and Krause, A., Eds. *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018* (2018), vol. 80 of *JMLR Workshop and Conference Proceedings*, JMLR.org.

[38] Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., Legg, S., and Kavukcuoglu, K. IMPALA: scalable distributed deep-rl with importance weighted actor-learner architectures. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018* (2018), pp. 1406–1415.

[39] Gil, M., Alajaji, F., and Linder, T. Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences 249* (2013), 124–131.

[40] Glynn, P. W. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM 33*, 10 (1990), 75–84.

[41] Grondman, I., Busoniu, L., Lopes, G. A., and Babuska, R. A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42*, 6 (2012), 1291–1307.

[42] Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017* (2017), pp. 1352–1361.

[43] Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018* (2018), pp. 1856–1865.

[44] Hershey, J. R., and Olsen, P. A. Approximating the kullback leibler divergence between gaussian mixture models. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on* (2007), vol. 4, IEEE, pp. IV–317.

[45] Houthooft, R., Chen, X., Duan, Y., Schulman, J., De Turck, F., and Abbeel, P. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems* (2016), pp. 1109–1117.

[46] Ionides, E. L. Truncated importance sampling. *Journal of Computational and Graphical Statistics 17*, 2 (2008), 295–311.

[47] Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research 11*, Apr (2010), 1563–1600.

[48] Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems* (2018), pp. 4868–4878.

[49] Kalai, A. T., and Mohri, M., Eds. *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010* (2010), Omnipress.

[50] Kallus, N. Instrument-armed bandits. In *Algorithmic Learning Theory* (2018), pp. 529–546.

[51] Kaufmann, E., Korda, N., and Munos, R. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory* (2012), Springer, pp. 199–213.

[52] Kearns, M., and Singh, S. Near-optimal reinforcement learning in polynomial time. *Machine learning 49*, 2-3 (2002), 209–232.

[53] Kleinberg, R., Slivkins, A., and Upfal, E. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing* (2008), ACM, pp. 681–690.

[54] Kleinberg, R., Slivkins, A., and Upfal, E. Bandits and experts in metric spaces. *arXiv preprint arXiv:1312.1277* (2013).

[55] Kleinberg, R. D. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems* (2005), pp. 697–704.

[56] Kupcsik, A. G., Deisenroth, M. P., Peters, J., and Neumann, G. Data-efficient generalization of robot skills with contextual policy search. In *Twenty-Seventh AAAI Conference on Artificial Intelligence* (2013).

[57] Lai, T. L., and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics 6*, 1 (1985), 4–22.

[58] Lakshmanan, K., Ortner, R., and Ryabko, D. Improved regret bounds for undiscounted continuous reinforcement learning. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015* (2015), pp. 524–532.

[59] Lange, S., Gabel, T., and Riedmiller, M. Batch reinforcement learning. In *Reinforcement learning*. Springer, 2012, pp. 45–73.

[60] Lattimore, T., and Hutter, M. Near-optimal pac bounds for discounted mdps. *Theoretical Computer Science 558* (2014), 125–143.

[61] Lattimore, T., and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press (preprint), 2019.

[62] Magureanu, S., Combes, R., and Proutiere, A. Lipschitz bandits: Regret lower bounds and optimal algorithms. *arXiv preprint arXiv:1405.4758* (2014).

[63] Masood, M. A., and Doshi-Velez, F. Diversity-inducing policy gradient: Using mmd to find a set of policies that are diverse in terms of state-visitation.

[64] McCluskey, L., Williams, B. C., Silva, J. R., and Bonet, B., Eds. *Proceedings of the Twenty-Second International Conference on Automated Planning and Scheduling, ICAPS 2012, Atibaia, São Paulo, Brazil, June 25-19, 2012* (2012), AAAI.

[65] Medina, A. M., and Yang, S. No-regret algorithms for heavy-tailed linear bandits. In *Proceedings of The 33rd International Conference on*

Machine Learning (New York, New York, USA, 20–22 Jun 2016), M. F. Balcan and K. Q. Weinberger, Eds., vol. 48 of *Proceedings of Machine Learning Research*, PMLR, pp. 1642–1650.

[66] Mersereau, A. J., Rusmevichientong, P., and Tsitsiklis, J. N. A structured multiarmed bandit problem and the greedy policy. *IEEE Transactions on Automatic Control 54*, 12 (2009), 2787–2802.

[67] Metelli, A. M., Papini, M., Faccio, F., and Restelli, M. Policy optimization via importance sampling. In *Advances in Neural Information Processing Systems* (2018), pp. 5447–5459.

[68] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016* (2016), pp. 1928–1937.

[69] Moré, J. J., and Thuente, D. J. Line search algorithms with guaranteed sufficient decrease. *ACM Transactions on Mathematical Software (TOMS) 20*, 3 (1994), 286–307.

[70] Ok, J., Proutière, A., and Tranos, D. Exploration in structured reinforcement learning. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.* (2018), pp. 8888–8896.

[71] Ortner, R., and Ryabko, D. Online regret bounds for undiscounted continuous reinforcement learning. In *Proceedings of the 25th International Conference on Neural Information Processing Systems-Volume 2* (2012), Curran Associates Inc., pp. 1763–1771.

[72] Osband, I., Russo, D., and Roy, B. V. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.* (2013), pp. 3003–3011.

[73] Osband, I., and Van Roy, B. Why is posterior sampling better than optimism for reinforcement learning? In *Proceedings of the 34th*

International Conference on Machine Learning-Volume 70 (2017), JMLR. org, pp. 2701–2710.

[74] Osband, I., Van Roy, B., and Wen, Z. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning* (2016), pp. 2377–2386.

[75] Ostrovski, G., Bellemare, M. G., van den Oord, A., and Munos, R. Count-based exploration with neural density models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (2017), JMLR. org, pp. 2721–2730.

[76] Owen, A., and Zhou, Y. Safe and Effective Importance Sampling. *Journal of the American Statistical Association* (Mar. 2000), 135–143.

[77] Owen, A. B. *Monte Carlo theory, methods and examples.* 2013.

[78] Owen, A. B. *Monte Carlo theory, methods and examples.* 2013.

[79] Pandey, S., Chakrabarti, D., and Agarwal, D. Multi-armed bandit problems with dependent arms. In *Proceedings of the 24th international conference on Machine learning* (2007), ACM, pp. 721–728.

[80] Peters, J., and Schaal, S. Reinforcement learning of motor skills with policy gradients. *Neural networks 21*, 4 (2008), 682–697.

[81] Precup, D., and Teh, Y. W., Eds. *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017* (2017), vol. 70 of *Proceedings of Machine Learning Research*, PMLR.

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

[83] Rényi, A. On measures of entropy and information. Tech. rep., Hungarian Academy of Sciences Budapest Hungary, 1961.

[84] Robbins, H. Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*. Springer, 1985, pp. 169–177.

[85] Rummery, G. A., and Niranjan, M. *On-line Q-learning using connectionist systems*, vol. 37. University of Cambridge, Department of Engineering Cambridge, England, 1994.

[86] Saritaç, A. Ö., and Tekin, C. Combinatorial multi-armed bandit problem with probabilistically triggered arms: A case with bounded regret. In *Signal and Information Processing (GlobalSIP), 2017 IEEE Global Conference on* (2017), IEEE, pp. 111–115.

[87] Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International Conference on Machine Learning* (2015), pp. 1889–1897.

[88] Sehnke, F., Osendorfer, C., Rückstiess, T., Graves, A., Peters, J., and Schmidhuber, J. Policy gradients with parameter-based exploration for control. In *International Conference on Artificial Neural Networks* (2008), Springer, pp. 387–396.

[89] Sehnke, F., Osendorfer, C., Rückstiess, T., Graves, A., Peters, J., and Schmidhuber, J. Parameter-exploring policy gradients. *Neural Networks 23*, 4 (2010), 551–559.

[90] Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *ICML* (2014).

[91] Srinivas, N., Krause, A., Kakade, S., and Seeger, M. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (Haifa, Israel, June 2010), J. Fürnkranz and T. Joachims, Eds., Omnipress, pp. 1015–1022.

[92] Strehl, A. L., Li, L., and Littman, M. L. Reinforcement learning in finite mdps: Pac analysis. *Journal of Machine Learning Research 10*, Nov (2009), 2413–2444.

[93] Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine learning 3*, 1 (1988), 9–44.

[94] Sutton, R. S. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin 2*, 4 (1991), 160–163.

[95] Sutton, R. S., and Barto, A. G. *Reinforcement learning: An introduction.* MIT press, 2018.

[96] Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems* (2000), pp. 1057–1063.

[97] Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika 25*, 3/4 (1933), 285–294.

[98] Thrun, S. B. Efficient exploration in reinforcement learning.

[99] Tokic, M., and Palm, G. Value-difference based exploration: adaptive control between epsilon-greedy and softmax. In *Annual Conference on Artificial Intelligence* (2011), Springer, pp. 335–346.

[100] Vakili, S., Liu, K., and Zhao, Q. Deterministic sequencing of exploration and exploitation for multi-armed bandit problems. *arXiv preprint arXiv:1106.6104* (2011).

[101] Veach, E., and Guibas, L. J. Optimally combining sampling techniques for Monte Carlo rendering. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques - SIGGRAPH '95* (1995), ACM Press, pp. 419–428.

[102] Wang, Z., Zhou, R., and Shen, C. Regional multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics* (2018), pp. 510–518.

[103] Watkins, C. J. C. H. *Learning from delayed rewards.* PhD thesis, King's College, Cambridge, 1989.

[104] Weinstein, A., and Littman, M. L. Bandit-based planning and learning in continuous-action markov decision processes. In *Proceedings of the Twenty-Second International Conference on Automated Planning and Scheduling, ICAPS 2012, Atibaia, São Paulo, Brazil, June 25-19, 2012* (2012).

[105] Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning 8*, 3-4 (1992), 229–256.

[106] Yu, X., Shao, H., Lyu, M. R., and King, I. Pure exploration of multi-armed bandits with heavy-tailed payoffs.

[107] Zhao, T., Hachiya, H., Niu, G., and Sugiyama, M. Analysis and improvement of policy gradient estimation. In *Advances in Neural Information Processing Systems* (2011), pp. 262–270.