Contents

Chapter 1

Preliminaries

This is an introduction...

1.1 Markov Decision Processes

A Markov Decision Process (MDP) is a way to model the interaction between an agent, which is the learner and the decision maker, and an environment. When the agent performs an action a, the environment responds presenting a new state s to the agent and rewarding the agent with a certain scalar signal called *immediate reward*. Importantly, the environment dynamics of a MDP are *stationary*: they do not depend upon time. Moreover, the current state s_{h+1} at time step h must depend only on the previous state s_h and action a_h. This property, called Markovian Property, entails that the agent is somehow pushed to forget the states and actions of the past. However, there is another mechanism of MDPs which encourages the agent to take into account the long term consequences of its choices. In fact, the objective of the agent is to maximize over time the *cumulative reward* or *return*, which is the cumulated sum of the immediate rewards obtained after each action undertaken by the agent. Evidently, sometimes it is more profitable to sacrifice immediate reward in order to reach a higher cumulative reward in the long term. This mechanism pushes the agent to take into account the future in its current decisions.

Hence, MDPs are a powerful tool capable of modelling many challenges that we encounter in life, science and engineering. Take, for example, a hungry newborn infant in his mother's arms. Through various attempts, the infant moves around its arms, head and mouth looking for food. In this ways, he

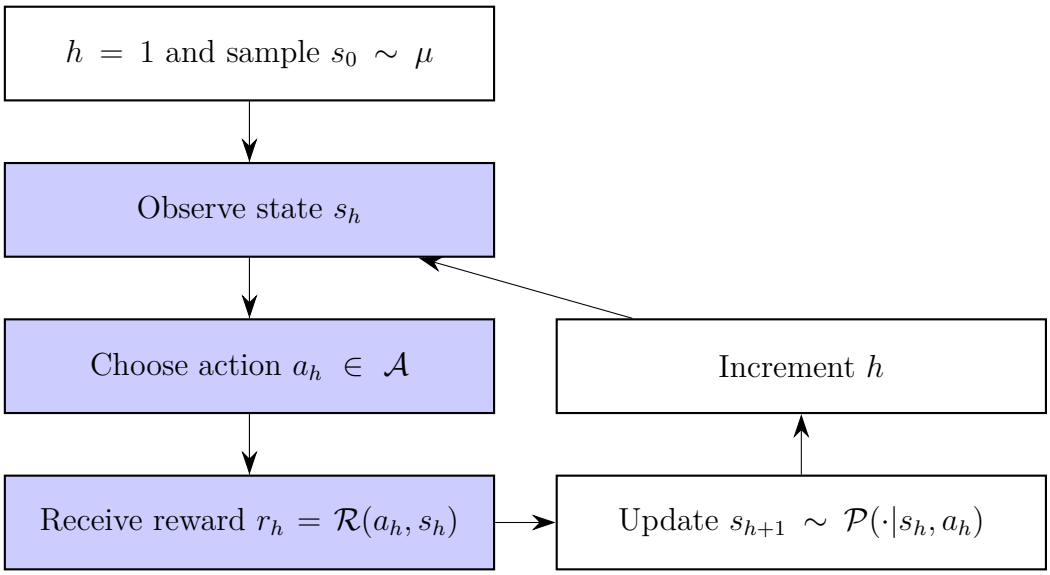

Figure 1.1: Interaction protocol for Markov decision processes

changes its state until the moment he eventually receives a positive reward, the mother's milk. Little by little, the baby will learn the precise sequence of actions and states that rewards him with milk.

Formally, a discrete-time continuous MDP [64, 73] is a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \mu \rangle$, where:

- $\mathcal{S} \in \mathbb{R}^{d_\mathcal{S}}$ is the $d_\mathcal{S}$-dimensional continuous *state space*, i.e., the set of all possible observable states of the environment;
- $\mathcal{A} \in \mathbb{R}^{d_\mathcal{A}}$ is the $d_\mathcal{A}$-dimensional continuous *action space*, i.e., the set of all the possible actions that the agent can perform. Sometimes, not all the actions are performable in all states. In such cases, we can define the set $\mathcal{A}(s)$ for all $s \in \mathcal{S}$, s.a $\mathcal{A} = \bigcup_{s \in \mathcal{S}} \mathcal{A}(s)$;
- $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is a function called *transition model* such that, for every $s \in \mathcal{S}$ and $a \in \mathcal{A}$, it assigns a probability measure $\mathcal{P}(\cdot|s, a)$ over $\mathcal{S}$. Its corresponding probability density function is $P(\cdot|s, a)$;
- $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to [-R_{\max}, R_{\max}]$ is a bounded *reward* function such that, every time the agent chooses action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$, it receives a reward $r = \mathcal{R}(a, s)$. $R_{\max}$ is the maximum absolute reward that the agent can receive;
- $\gamma \in (0, 1]$ is a discount factor to apply to future rewards;

- $\mu \in \Delta(\mathcal{S})$ is the initial state distribution, such that the initial state is drawn as $s_0 \sim \mu$.

In the more general case, the state space $\mathcal{S}$ and the action space $\mathcal{A}$, which are the sensor and actuator possibilities of the agent, respectively, can be both continuous and discrete, finite or infinite. In what follows, we will focus on the continuous case because it is the most relevant to our work.

The time is typically modelled as a discrete sequence of decision steps represented by the natural numbers: $\mathcal{H} = \{0, 1, \dots, H\}$, where $H \in \mathbb{N}$ is the *horizon* of a given task, which can be either infinite $H = \infty$ or finite. The agent-environment interaction ends when either the horizon is reached or a *terminal state* is reached. Tasks that always end in a finite amount of steps (an *episode*) are called *episodic*. A *trajectory* is the sequence of states and actions $\tau = \{s_0, a_0, s_1, a_1, \dots, s_{H-1}, a_{H-1}\}$ up to the last time step of the episode.

The core of a MDP agent is the *policy* π, a mapping from perceived states of the environment to possible actions. A policy can also take as input the past history of set and actions, but in our work we only consider *memoryless policies*. Such mappings can be deterministic $\pi : \mathcal{S} \to \mathcal{A}$ or stochastic $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, such that the current action is drawn as $a_h \sim \pi(\cdot|s_h)$. Note that the memoryless property of policies does not imply that the agent cannot take into account what it has seen before when making an action choice. In fact, the agent has multiple ways to leverage its experience and learn through time, as we will see later on.

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. 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.

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

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

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

[6] 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.

[7] 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.

[8] 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.

[9] 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).

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

[11] 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.

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

[13] 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.

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

[15] 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.

[16] 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.

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

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

[19] 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.

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

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

[22] 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.

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

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

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

[36] 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.

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

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

[39] 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.

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

[41] 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.

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

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

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

[45] 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.

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

[47] Lattimore, T., and Szepesvári, C. Bandit algorithms.

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

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

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

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

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

[61] 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.

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

[63] 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.

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

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

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

[67] 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.

[68] 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.

[69] 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.

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

[71] 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.

[72] 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.

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

[74] 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.

[75] 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.

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

[77] 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.

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

[79] 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).

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