1 Multi-armed Bandits

A multi-armed bandit problem [1] is a sequential decision problem over a set of possible actions ("arms"). At each time step, the player pulls one of the arms and recieves a pre-allocated and observable reward. The goal is to maximize the rewards obtained over a sequence of allocations and actions. The term multi-armed bandit comes from the sequential decision problem of playing multiple slot machines at once (the slot machines being the multi-armed bandit), and repeatedly choosing which arm to pull next. The player must balance the exploitation of arms that did well in the past and the exploration of arms that are currently underestimated but which may give higher payoffs in the future. When playing board games using reinforcement learning, each decision is made by simulating and evaluating as many possible game rollouts from the current state as time and resources allow. Algorithms for bandits (more specifically, for a tree-based version of the bandit problem) can be used to explore more efficiently the huge tree built by simulating game rollouts by focusing on the most promising subtrees (i.e. where the sampled return was greatest). A crucial algorithm in the literature, the UCT algorithm for hierarchical bandits of Kocsis and Szepesvari [2006] [2], which can be seen as an extenion of the UCB bandit algorithm, is directly applicable to these tree-based searches. UCT has demonstrable improvements over prior methods such as ϵ-greedy, where the optimal action is selected with probability 1-ϵ and a uniformly random action with the remaining probability ϵ. These improvements stem from analyzing the regret of a player who pulls the arms according to some strategy. We can make comparisons between this strategy's performance with that of an optimal strategy that, for any n step horizon, consistently plays the best arm. Stated more simply, we analyze the regret of a player who does not always play optimally. This regret is typically formulated as follows:

$$R_n \overset{\text{def}}{=} \max_{i=1,\ldots,K} \sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{I_t,t} \tag{1}$$

Where we have K ≥ 2 arms and sequences of rewards $X_{i,1}, X_{i,2}...$ associated with each arm $i = 1, ..., K$. Where at each time step $t = 1, ..., N$ the player selects an arm I_t and recieves the pre-allocated reward $X_{I_t,t}$.

The successes of UCB, and in turn the UCT algorithm, come from bounding this regret defined above. The UCT algorithm effectively moderates the tradeoff between exploitation and exploration by introducing a bias term which quantifies our uncertainty on the current estimates of the action values.

Though multi-armed bandits have been studied in a plethora of environments we are mainly interested in the stochastic and Markovian settings. These settings apply most directly to tree search and thus planning over actions in complex games. This tree search is a sequential decision problem over states in a finite MDP. Like with multi-armed bandits, the assumptions are that the reward and transition distributions are unknown, and we want to act in the MDP so as to maximize the rewards. Another model, with many applications, is that of sleeping bandits. There, it is assumed that the available actions vary over time. [3]

Another interesting result is that of Kearns et al. who showed that regardless of the size of the

state-space, fixed size trees suffice to find an action at the initial state whose value is within some error ϵ of the best action. [4]

References

[1] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc.*, vol. 58, no. 5, pp. 527–535, 09 1952. [Online]. Available: https://projecteuclid.org:443/euclid.bams/1183517370

[2] L. Kocsis and C. Szepesvári, "Bandit based monte-carlo planning," in *Proceedings of the 17th European Conference on Machine Learning*, ser. ECML'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 282–293. [Online]. Available: http://dx.doi.org/10.1007/11871842_29

[3] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012. [Online]. Available: http://dx.doi.org/10.1561/2200000024

[4] M. Kearns, Y. Mansour, and A. Y. Ng, "A sparse sampling algorithm for near-optimal planning in large markov decision processes," *Machine Learning*, vol. 49, no. 2, pp. 193–208, Nov 2002. [Online]. Available: https://doi.org/10.1023/A:1017932429737

[5] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2, pp. 235–256, May 2002. [Online]. Available: https://doi.org/10.1023/A:1013689704352

[6] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, Jan. 2003. [Online]. Available: https://doi.org/10.1137/S0097539701398375

[7] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," *J. Mach. Learn. Res.*, vol. 11, pp. 1563–1600, Aug. 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1756006.1859902