The Multi-armed Bandit Problem

Ben Wiley and **Micah Brown**
{bewiley,msbrown}@davidson.edu
Davidson College
Davidson, NC 28035
U.S.A.

Abstract

This paper analyzes a well known division of reinforcement learning known as the multi-armed bandit problem. The multi-armed bandit problem attempts to determine the best policy when faced with multiple choices. We analyze two different policies for picking machines: upper confidence bound (UCB1) and $\mathcal{E}_n$ -GREEDY. [alg] was a better algorithm due to...

1 Introduction

When given multiple options in a situation with unknown outcomes, naturally one does not know whether to explore more options or stick to the best option that one knows. A multi-armed bandit problem enables us to study the exploration-exploitation trade-off in reinforcement learning. In this problem, an agent is faced with multiple "arms" (though often the term slot-machine is used), in which each arm has a reward distribution. The agent's goal is to determine a policy that informs it which arm to pick at any given state. Should it pick the arm that it knows to be the most successful to this point, or should it explore other options in the case that it can find a better alternative? That is the central question of the exploration-exploitation trade-off that we would like to answer.

This problem is interesting to Artificial Intelligence researchers due to the number of interesting practical uses that it has been applied to. For example, the multi-armed bandit problem has been used to model clinical trial treatments, ad placement, website optimization, and computer game-playing (Bubeck and Cesa-Bianchi 2012). Furthermore, the idea of reinforcement learning can summarize most of the field of Artificial Intelligence: place an agent in an unknown environment, with unknown rules, and have the agent determine what to do to thrive (Russell and Norvig 2003).

The multi-armed bandit problem has been analyzed extensively by (Auer, Cesa-Bianchi, and Fischer 2002), and this paper attempts to reproduce a subset of their work. The remainder of this paper includes background information on reinforcement learning and the multi-armed bandit problem, the experiments we performed, the results, and finally our conclusions.

2 Background

Describe any background information that the reader would need to know to understand your work. You do not have to explain algorithms or ideas that we have seen in class. Rather, use this section to describe techniques that you found elsewhere in the course of your research, that you have decided to bring to bear on the problem at hand. Don't go overboard here — if what you're doing is quite detailed, it's often more helpful to give a sketch of the big ideas of the approaches that you will be using. You can then say something like "the reader is referred to X for a more in-depth description of...", and include a citation.

Alternately, you may have designed a novel approach for the problem — your own algorithm or heuristic, say. A description of these would also be placed in this section (use subsections to better organize the content in this case).

3 Experiments

In this section, you should describe your experimental setup. What were the questions you were trying to answer? What was the experimental setup (number of trials, parameter settings, etc.)? What were you measuring? You should justify these choices when necessary. The accepted wisdom is that there should be enough detail in this section that I could reproduce your work *exactly* if I were so motivated.

4 Results

Present the results of your experiments. Simply presenting the data is insufficient! You need to analyze your results. What did you discover? What is interesting about your results? Were the results what you expected? Use appropriate visualizations. Prefer graphs and charts to tables as they are easier to read (though tables are often more compact, and can be a better choice if you're squeezed for space). **Always** include information that conveys the uncertainty in your measurements: mean statistics should be plotted with error bars, or reported in tables with a ± range. The

95%-confidence interval is a commonly reported statistic.

5 Conclusions

In this section, briefly summarize your paper — what problem did you start out to study, and what did you find? What is the key result / take-away message? It's also traditional to suggest one or two avenues for further work, but this is optional.

References

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* 47(2-3):235–256.

Bubeck, S., and Cesa-Bianchi, N. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *CoRR* abs/1204.5721.

Russell, S. J., and Norvig, P. 2003. *Artificial Intelligence: A Modern Approach*. Pearson Education.