*Proceedings of the 2019 Winter Simulation Conference*
*N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, eds.*

# MONTE CARLO TREE SEARCH WITH OPTIMAL COMPUTING BUDGET ALLOCATION

Yunchuan Li

Department of Electrical and Computer Engineering
Institute for Systems Research
University of Maryland
College Park, MD 20742, USA

## ABSTRACT

We analyze a tree search problem with an underlying Markov decision process, in which the goal is to identify the best action at the root that achieves the highest cumulative reward. We present a new tree policy that optimally allocates a limited computing budget to maximize a lower bound on the probability of correctly selecting the best action at each node. Compared to the widely used Upper Confidence Bound (UCB) type of tree policies, the new tree policy presents a more balanced approach to manage the exploration and exploitation trade-off when the sampling budget is limited. Furthermore, UCB assumes that the support of reward distribution is known, whereas our algorithm relaxes this assumption.

## 1 INTRODUCTION

We consider a reinforcement learning problem where an agent interacts with an underlying environment. A Markov Decision Process (MDP) with finite horizon is used to model the environment. In each move, the agent will take an action, receive a reward and land in a new state. The reward is usually random, and its distribution depends on both the state of the agent and the action taken. The distribution of the next state is also determined by the agent's current state and action. Our goal is to determine the optimal sequence of actions that leads to the highest expected reward. The optimality of the decision policy will be evaluated by the probability of correctly selecting the best action in the first stage of the underlying MDP.

If the distributions and the dynamics of the environment are known, the optimal set of actions can be computed through dynamic programming. Under more general settings where the agent does not have perfect information regarding the environment, Chang et al. (2005) proposed an adaptive algorithm based on a Multi-Armed Bandit (MAB) model and regret analysis with Upper Confidence Bound (UCB) (Auer et al. 2002). Kocsis and Szepesvári (2006) and Coulom (2007) applied UCB to tree search. In their method, the objective is to minimize overall regret, which is equivalent to maximizing cumulative reward throughout the sampling process.

## 2 MOTIVATION

Most bandit-based MCTS algorithms are designed to minimize the regret (or maximize the cumulative reward of the agent), whereas in many situations, the goal of the agent may be to efficiently determine the optimal set of actions within a limited sampling budget. To the best of our knowledge, there is limited effort in the literature that aims at addressing the latter problem.

Therefore, we are motivated to establish a tree policy that intelligently balances exploration and exploitation. In addition, many MCTS algorithms are only for MIN-MAX game trees, we are also interested in developing new tree policy that can be applied to more general types of tree search problems.

Algorithms that focus on minimizing regret tend to discourage exploration. This tendency can be seen in two ways. Suppose at some point an action was performed and received a small reward. To minimize regret, the algorithm would be discouraged from taking this action again. However, the small reward could likely be due to the randomness in the reward distribution. Mathematically, (Lai and Robbins 1985) showed that the number of times the optimal action is taken is exponentially more than sub-optimal ones. This leads to one motivation for our research: is there a selection policy that explores sub-optimal actions more so to make sure we really find the true optimal action?

Apart from the lack of exploration, most MCTS algorithms assume that the support of the reward distribution is known (typically assumed to be $[0,1]$). However, a general tree search problem may likely have an unknown range of rewards. In such case, assuming a range can lead to very poor performance. Therefore, another motivation of our research is to relax the known reward support assumption.

To tackle the challenge in balancing exploration and exploitation with a limited sampling budget for a tree policy, we model the tree selection problem at each stage as a statistical *Ranking & Selection* (R&S) problem and propose a new tree policy for MCTS based on an adaptive algorithm from the R&S community. Similar to the MAB problem, R&S assumes that we are given a set of alternatives with unknown reward distributions, and the goal is to select the alternative with the highest mean reward. Specifically, we will develop an MCTS tree policy based on the Optimal Computing Budget Allocation (OCBA) framework (Chen et al. 2000). The objective of the proposed OCBA tree policy is to maximize the Approximate Probability of Correct Selection (APCS), which is a lower bound on the probability of correctly selecting the optimal action at each node. Intuitively, the objective function of the new OCBA tree selection policy will lead to an optimal balance between exploration and exploitation with a limited sampling budget, and thus help address the drawbacks of existing work that either pursues pure exploration or exponentially discourage exploration (Kocsis and Szepesvári 2006). Our new OCBA tree policy also removes the known support assumption for reward distribution, because the new OCBA policy determines the sampling allocation based on the posterior distribution of each action, which is updated adaptively according to samples.

## REFERENCES

Auer, P., N. Cesa-Bianchi, and P. Fischer. 2002. "Finite-Time Analysis of the Multiarmed Bandit Problem". *Machine Learning* 47(2-3):235–256.

Chang, H. S., M. C. Fu, J. Hu, and S. I. Marcus. 2005. "An Adaptive Sampling Algorithm for Solving Markov Decision Processes". *Operations Research* 53(1):126–139.

Chen, C.-H., J. Lin, E. Yücesan, and S. E. Chick. 2000. "Simulation Budget Allocation for Further Enhancing the Efficiency of Ordinal Optimization". *Discrete Event Dynamic Systems* 10(3):251–270.

Coulom, R. 2007. "Efficient Selectivity and Backup Operators in Monte-Carlo Tree Search". In *Proceedings of the 5th International Conference on Computers and Games*, CG'06, 72–83. Berlin, Heidelberg: Springer-Verlag.

Kocsis, L., and C. Szepesvári. 2006. "Bandit based Monte-Carlo planning". In *Proceedings of the 17th European Conference on Machine Learning*, ECML'06, 282–293. Berlin, Heidelberg: Springer-Verlag.

Lai, T. L., and H. Robbins. 1985. "Asymptotically Efficient Adaptive Allocation Rules". *Advances in Applied Mathematics* 6(1):4–22.