Modeling Extreme Values in Monte Carlo Tree Search

Abstract. Abstract

1 Introduction

Monte Carlo Tree Search (MCTS) [3, 5] is a simulation-based search technique that has received much attention in the search community due to its success in two-player turn-taking games, such as Go [4].

≪ *more about mcts ...* ≫

During a simulation, actions are selected using Upper Confidence Bound for Trees (UCT) [5]. This selection policy is based on a regret minimization method for bandit problems, Upper Confidence Bound (UCB) [1], which provides a good balance between exploration and exploitation.

Using the UCB selection policy, the strategy computed by MCTS provably converges to the optimal strategy as the number of simulations approaches infinity in two-player zero-sum games. In this setting, every state has a unique optimal value and the optimal strategy is one that chooses the action leading to a new state with the highest such value. The proof uses Hoeffding's inequality to show that the difference between the empirical mean of a set of sampled payoffs and the true mean of the underlying distribution is bounded and approaches zero as the number samples grow. Therefore, if the empirical mean of the samples can be shown to converge to the optimal value, the optimal strategy simply chooses the action with the highest value.

Suppose now that we have a sequential problem where the goal is not simply to maximize expected utility but to find (or avoid) strategies that admit extreme values. For example, there my be hard constraints such as avoiding any strategy that could leading to the death of a patient. As another example, the underlying problem may lack inherent stochasticity, which is only introduced by the sampling algorithm itself. In these situations, selection policies that model the statistical behavior of extreme values may be preferred.

In this work, we develop an MCTS-style simulation-based algorithm for sequential decision-making problems based on extreme value theory [2]. We show that, in certain situations, using extreme value theory to guide simulations can be preferred to bandit theory. We prove the consistency of MCTS in single-player games using this selection mechanism. Finally, we thoroughly compare the EVT-based MCTS versus bandit-based on several complex optimization problems as well a spectrum of randomly-generated sequential problems.

2 Background

$\ll$ *formalize basics of sequential decision-making problems ...* $\gg$

2.1 Bandit-Based Monte Carlo Tree Search

A stochastic bandit problem is one where an player is faced with K actions. When an action is taken, the player received a random payoff X_i which is generated from some fixed distribution. The player can repeat this process n times, receiving payoffs $X_{i,t}$ for choosing action i at time t. The goal is maximize the expected payoff $\mathbb{E}[\sum_{t=1}^{n} X_{I_t,t}]$ where $I_t \in \{1, \ldots, K\}$ is the (generally randomized) choice made by the player at time t.

The standard way to express the quality of a bandit algorithm is to measure its expected regret

$$R_n = \max_{1 \leq i \leq K} \left(\mathbb{E}\left[\sum_{t=1}^{n} X_{i,t}\right] - \mathbb{E}\left[\sum_{t=1}^{n} X_{I_t,t}\right] \right),$$

that is to quantify how much the player would have preferred to choose the single best action in hindsight.

Let $T_i(n) = \sum_{t=1}^{n} \mathbb{I}(I_t = i)$ be the number of times action i was chosen up to time n, and $\bar{X}_{i,n} = \sum_{t=1}^{n} \mathbb{I}(I_t = i) X_{i,t} / T_i(n)$ is the empirical mean of the value of action i after n plays, where the indicator function $\mathbb{I}(\text{cond}) = 1$ if cond is true, or 0 otherwise.

Algorithm UCB1 chooses

$$I_t = \operatorname*{argmax}_{i \in \{1,\ldots,K\}} \{\bar{X}_{i,n} + B_i(t)\},$$

where $B_i(t)$ is an exploration bias parameter that is generally a function of the observed sequence up to time t.

Hoeffding's inequality, as adapted from [5, Equations 3-4], states that

$$\Pr(|\bar{X}_{i,n} - \mu_i| \geq B_i(t)) \leq n^{-4}$$

and if $\lim_{t \to \infty} B_i(t) = 0$ then eventually the empirical mean will match true mean μ_i.

In Monte Carlo Tree Search, UCB is applied recursively, as if treating every state as its own bandit problem. The true value of action i is then the minimax value of the child state and the empirical mean $\bar{X}_{i,n}$ converges to this value in two-player zero-sum games with perfect information. The proof is non-trivial because in this setting, the payoffs are not simply random variables generated from some distribution but rather the reward received from an arbitrarily complex stochastic (non-stationary) process which is itself the another "recursive sub-bandit problem". Nonetheless, under certain conditions that the payoffs and process are well-behaved, convergence can still be guaranteed.

2.2 Extreme Value Theory

In subsection 2.1 we outlined how to compute the value of an action in UCT. One can relabel the time steps $1 \leq s \leq T_i(n)$ and consider, for some action i, the sequence of payoffs $\mathbf{X}_i = (X_1, X_2, \dots, X_{T_i(n)})$, guiding the simulations by analyzing their empirical means $f(\mathbf{X}_i) = \bar{X}_{i,n} = \sum_{s=1}^{T_i(n)} X_s / T_i$.

Extreme value theory concerns the statistical behavior of a different quantity

$$f(\mathbf{X}_i) = M_{i,n} = \max\{X_1, X_2, \dots, X_{T_i(n)}\}.$$

In contrast to the bandit setting, $M_{i,n}$ is not necessarily an estimator of the mean of the underlying payoff distribution, but rather models extreme values in of the payoff sequences.

Define $M^*_{i,n} = (M_{i,n} - b_n)/a_n$ for some constants a_n and b_n. By [2, Theorem 3.1.1], the cumulative distribution function

$$\Pr(M^*_{i,n} \leq z) \rightarrow G(z), \text{ where } G(z) = \exp\left\{ - \left[1 + \xi \left(\frac{z - \mu_i}{\sigma} \right) \right]^{-1/\xi} \right\}$$

is known as the **generalized extreme value** (GEV) family of distributions. Therefore, the foundation of extreme value theory is an analog of the central limit theorem for $M_{i,n}$. Furthermore the GEV distribution $G(z)$ can be inferred from observed data using standard techniques from statistical machine learning such as maximum likelihood estimation.

3 MCTS Based on Extreme Value Theory

In this section, we develop our MCTS algorithm based on extreme value theory. In essence, at each state we replace the bandit-inspired regret minimizers (UCB) with GEV prediction machines. Each one observes payoff data from each of its actions and build its own GEV distributions from observed payoff data, which in turn informs it how to select on successive simulations. As in standard MCTS, the problem is defined recursively, so that the payoffs generated are indeed obtained as function of many decision-making steps rather than a single step.

4 Experiments

5 Conclusion

References

1. Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time analysis of the multiarmed bandit problem. Machine Learning 47(2/3), 235–256 (2002)
2. Coles, S.: An Introduction to Statistical Modeling of Extreme Values. Springer (2001)

3. Coulom, R.: Efficient selectivity and backup operators in Monte-Carlo tree search. In: Proceedings of the 5th international conference on Computers and games. CG'06, vol. 4630, pp. 72–83 (2007)
4. Gelly, S., Kocsis, L., Schoenauer, M., Sebag, M., Silver, D., Szepesvári, C., Teytaud, O.: The grand challenge of computer Go: Monte Carlo tree search and extensions. Communications of the ACM 55(3), 106–113 (March 2012)
5. Kocsis, L., Szepesvári, C.: Bandit-based Monte Carlo planning. In: 15th European Conference on Machine Learning. LNCS, vol. 4212, pp. 282–293 (2006)