Extreme Value Monte Carlo Tree Search (Extended Abstract)

Masataro Asai[*1], **Stephen Wissow**[*2]

[1]MIT-IBM Watson AI Lab
[2]University of New Hampshire
masataro.asai@ibm.com, sjw@cs.unh.edu

Abstract

Monte-Carlo Tree Search (MCTS) combined with Multi-Armed Bandit (MAB) has had limited success in domain-independent classical planning until recently. Previous work (Wissow and Asai 2023) showed that UCB1, designed for bounded rewards, does not perform well when applied to the cost-to-go estimates of classical planning, which are unbounded in $\mathbb{R}$, then improved the performance by using a Gaussian reward MAB instead. We further sharpen our understanding of ideal bandits for planning tasks by resolving three issues: First, Gaussian MABs under-specify the support of cost-to-go estimates as $[-\infty, \infty]$. Second, Full-Bellman backup that backpropagates max/min of samples lacks theoretical justifications. Third, removing dead-ends lacks justifications in Monte-Carlo backup. We use *Extreme Value Theory Type 2* to resolve them at once, propose two bandits (UCB1-Uniform/Power), and apply them to MCTS for classical planning. We formally prove their regret bounds and empirically demonstrate their performance in classical planning.

1 Introduction

A recent breakthrough in Monte-Carlo Tree Search (MCTS) combined with Multi-Armed Bandit (MAB) applied to classical planning demonstrated that a better theoretical understanding of bandit-based algorithms can significantly improve search performance. Wissow and Asai (2023) showed why UCB1 bandit (Auer, Cesa-Bianchi, and Fischer 2002) does not perform well in classical planning: It assumes a reward distribution with a known, fixed, finite support such as $[0, 1]$ that is shared by all arms, i.e., that the cost-to-go estimates / heuristic functions would satisfy this. They then proposed UCB1-Normal2 bandit that assumes Gaussian rewards, which has the support $\mathbb{R} = [-\infty, \infty]$ that is impossible to violate.

We continue this trend to improve the understanding of heuristic search from the MAB standpoint. We aim to resolve three theoretical issues that remain in the previous work: The **first** is the under-specification that assumes cost-to-go estimates to be in $[-\infty, \infty]$, which should instead be $[0, \infty]$. In relaxation heuristics, this can be tighter, e.g., $h^{\max} \in [0, h^+]$ and $h^{\mathrm{FF}} \in [h^+, \infty]$. The **second** is the correct statistical characterization of *extrema* (maximum/minimum): Schulte and Keller (2014) noted that the use of averages in UCT is "rather odd" for optimization tasks such as planning and tried to address it without bandit-theoretic justification (Full Bellman backup that backpropagates the smallest mean among the arms.) The **third** is the dead-end removal. When a heuristic returns ∞ at a dead-end, a single dead-end node in a tree makes the evaluation of the root node ∞, because the average is ∞ if samples contain an ∞. Schulte and Keller (2014) removed these dead-end nodes from the tree, but it lacks statistical justification.

We introduce Extreme Value Theory (Fisher and Tippett 1928; Balkema and De Haan 1974, EVT) as the statistical foundation for understanding general optimization (minimization/maximization) tasks that resolves all issues above. EVTs are designed to model the statistics of *extrema* (minimum/maximum) of distributions using the *Extremal Limit Theorem*, unlike most statistical literature that models the *average* behavior based on the *Central Limit Theorem* (Laplace 1812, CLT). Among branches of EVTs, we identified *EVT Type 2* as our primary tool for designing new algorithms, which leads to Generalized Pareto (GP) distribution, which plays the same role as Gaussian distribution in CLT.

Based on this framework, we propose two novel MAB algorithms, UCB1-Power and UCB1-Uniform, for heuristic search applied to classical planning. Each of our novel bandits models a special case of GP distribution to avoid the numerical difficulty of estimating its parameters. We evaluate their performance over existing bandit-based MCTS, traditional GBFS and state-of-the-art diversified search algorithm called Softmin-Type(h) (Kuroiwa and Beck 2022). On 772 IPC instances under the same evaluation budget of 10^4 nodes using h^{FF} heuristics, GUCT-Power solved 55, 15.8, and 20 more instances than GBFS, GUCT-Normal2, and Softmin-Type(h), and GUCT-Uniform solved 56.8, 13, and 18.8 more instances, respectively.

2 Extreme Value Theory Type 2

CLT states that the average of i.i.d. RVs converges in distribution to a Gaussian distribution. Extremal Limit Theorem Type 1 (Fisher and Tippett 1928) similarly states that the maximum of i.i.d. RVs converges in distribution to an *Extreme Value Distribution* (EVD). It is used for predicting the *block maxima*, such as the monthly maximum water level.

[*]These authors contributed equally.

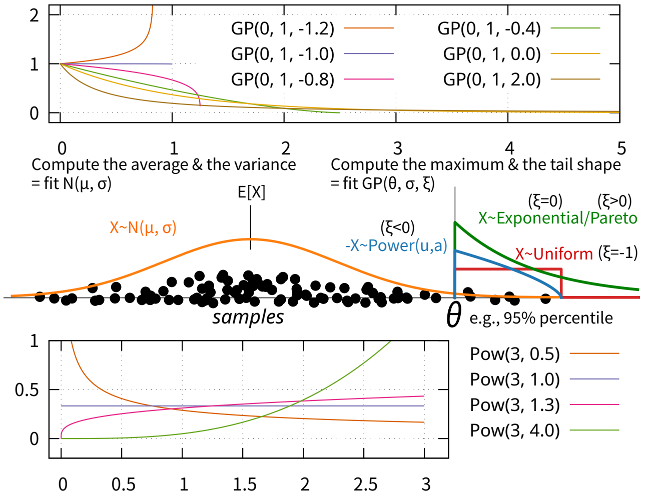

Figure 1: (top) Generalized pareto distribution $\mathrm{GP}(0, 1, \xi)$. (mid) Computing the average and the variance is seen as fitting $\mathcal{N}(\mu, \sigma)$; Computing the maximum and the shape of the tail distribution is seen as fitting $\mathrm{GP}(\mu, \sigma, \xi)$ with $\xi < 0$. (bottom) Power distribution $\mathrm{Pow}(3, a)$.

Extremal Limit Theorem Type 2 (Balkema and De Haan 1974) states that the excesses of i.i.d. RVs over a sufficiently high threshold θ converge in distribution to a Generalized Pareto (GP) distribution. It is used in *Peaks-Over-Threshold* analyses that predict exceedances over the safety limit.

$$\mathrm{GP}(\mathrm{x} \mid \theta, \sigma, \xi) = \begin{cases} \frac{1}{\sigma}\left(1 + \xi \frac{\mathrm{x}-\theta}{\sigma}\right)^{-\frac{\xi+1}{\xi}} & (\xi \neq 0) \\ \frac{1}{\sigma} \exp\left(-\frac{\mathrm{x}-\theta}{\sigma}\right) & (\xi = 0) \end{cases} \quad (\mathrm{x} > \theta)$$

θ, σ and ξ are called the location, the scale, and the shape parameter. It has a support $\mathrm{x} \in [\theta, \theta - \frac{\sigma}{\xi}]$ when $\xi < 0$ (a short-tailed distribution), otherwise $\mathrm{x} \in [\theta, \infty]$ (a heavy-tailed distribution). Fig. 1 shows a conceptual illustration of Peaks-Over-Threshold EVT. Given i.i.d. samples $x_1, \ldots, x_N$, extract a subset which exceeds a certain sufficiently high threshold θ, such as the top 5% element, and fit the parameters σ, ξ of $\mathrm{GP}(\theta, \sigma, \xi)$ on this subset. Then, the future exceeding data also follows $\mathrm{GP}(\theta, \sigma, \xi)$.

The short-tailed GP perfectly matches our requirements. Consider the maximization scenario, where the heuristic value is negated into a reward $-h^{\mathrm{FF}} \in [-\infty, -h^{+}]$. A short-tailed GP gives us an upper support $\theta - \frac{\sigma}{\xi}$, which is obtained by fitting σ and ξ to the data and works as an estimate of $-h^{+}$. GP also justifies discarding dead-ends ($-h^{\mathrm{FF}} = -\infty$) because GP is conditioned by $\mathrm{x} > \theta$. We use $-\theta = h(I) + 1$ for the initial state I.

Estimating the parameters of GP is known to be difficult. Thus we focus on its two subclasses: Uniform distribution $U(l, u)$ with an unknown support $[l, u]$, and Power distribution (Dallas 1976) $\mathrm{Pow}(u, a)$ with an unknown support $[0, u]$ and an unknown shape a. The price we pay is one degree of freedom in $\mathrm{GP}(\theta, \sigma, \xi)$: $U(l, u)$ has a fixed shape $\xi = -1$, and $\mathrm{Pow}(u, a)$ has a fixed lower bound 0. Note that GP models the maximum but Power models the minimum.

$$\mathrm{Pow}(\mathrm{x}|u, a) = \tfrac{a\mathrm{x}^{a-1}}{u^{a}}. \quad (0 < x < u,\ 0 < a)$$
$$U(\mathrm{x}|l, u) = \tfrac{1}{u-l}. \quad (l < x < u)$$

The expected values $\mathbb{E}[\mathrm{x}]$ are $\frac{ua}{a+1}$ and $\frac{l+u}{2}$, respectively. We can safely assume that states with smaller heuristic values (closer to the goal) are hard to find and rare during the search. Therefore, we assume $a \geq 1$ for Pow.

We introduce the Maximum Likelihood Estimators for Uniform and Power, then propose bandits that use these estimates, which are then used by MCTS for action selection.

Theorem 1. *Given i.i.d.* $x_1, \ldots, x_N \sim \mathrm{Pow}(\mathrm{x}|u, a)$*, the MLEs are* $\hat{u} = \max_i x_i$ *and* $\hat{a} = \left(\log \hat{u} - \frac{1}{N}\sum_i \log x_i\right)^{-1}$.

Theorem 2. *Given i.i.d.* $x_1, \ldots, x_N \sim U(\mathrm{x}|l, u)$*, the MLEs are* $\hat{u} = \max_i x_i$ *and* $\hat{l} = \min_i x_i$.

Backpropagation for these estimates from the leaves to the root uses existing backups. For $\hat{l}$ and $\hat{u}$ we use Full-Bellman backup (use the minimum/maximum among the children). For $\hat{a}$, we apply Monte-Carlo backup to the logarithms of heuristic values, then compute $\hat{a}$ combining $\hat{u}$ and the backed-up value. We propose two MABs that use them:

Theorem 3 (Main results)**.** *When* t_i*-th reward* r_{it_i} *of arm* i *follows* $U(l, u)$ *and* $\mathrm{Pow}(u, a)$ *with* $a \geq 1$*, we respectively define* LCB1-Uniform *and* LCB1-Power *as follows.*

$$\begin{aligned} \text{LCB1-Uniform}_i &= \tfrac{\hat{u}_i + \hat{l}_i}{2} - (\hat{u}_i - \hat{l}_i)\sqrt{6t_i \log T} \\ \text{LCB1-Power}_i &= \tfrac{\hat{u}_i \hat{a}_i}{\hat{a}_i + 1} - \hat{u}_i\sqrt{6t_i \log T} \end{aligned}$$

Let $\alpha \in [0, 1]$ *be an unknown problem-dependent constant and* u_i*,* l_i*,* a_i *be unknown ground-truth parameters of distributions of arm* i*. The cumulative regret is polynomially bounded as follows, where* $\beta = (2 - \alpha)^{1/a_i}$.

$$\tfrac{24(u_i - l_i)^2(1-\alpha)^2 \log T}{\Delta_i^2} + 1 + 2C + \tfrac{(1-\alpha)T(T+1)(2T+1)}{3}$$
$$\tfrac{6u_i^2(3-\beta)^2(\beta-1)^2 \log T}{\Delta_i^2} + 1 + 2C + \tfrac{(1-\alpha)T(T+1)(2T+1)}{3}$$

References

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

Balkema, A. A.; and De Haan, L. 1974. Residual Life Time at Great Age. *Annals of Probability*, 2(5): 792–804.

Dallas, A. 1976. Characterizing the Pareto and Power Distributions. *Annals of the Institute of Statistical Mathematics*, 28(1): 491–497.

Fisher, R. A.; and Tippett, L. H. C. 1928. Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(2): 180–190.

Kuroiwa, R.; and Beck, J. C. 2022. Biased Exploration for Satisficing Heuristic Search. In *Proc. of ICAPS*.

Laplace, P.-S. 1812. *Théorie analytique des probabilités*.

Schulte, T.; and Keller, T. 2014. Balancing Exploration and Exploitation in Classical Planning. In *Proc. of SOCS*.

Wissow, S.; and Asai, M. 2023. Scale-Adaptive Balancing of Exploration and Exploitation in Classical Planning. In *Proc. of HSDIP*.