ECAI 2012
Luc De Raedt et al. (Eds.)

doi:10.3233/978-1-61499-098-7-929

VOI-aware MCTS

David Tolpin and **Solomon Eyal Shimony** [1]

Abstract. UCT, a state-of-the art algorithm for Monte Carlo tree search (MCTS) in games and Markov decision processes, is based on UCB1, a sampling policy for the Multi-armed Bandit problem (MAB) that minimizes the cumulative regret. However, search differs from MAB in that in MCTS it is usually only the final "arm pull" (the actual move selection) that collects a reward, rather than all "arm pulls". In this paper, an MCTS sampling policy based on Value of Information (VOI) estimates of rollouts is suggested. Empirical evaluation of the policy and comparison to UCB1 and UCT is performed on random MAB instances as well as on Computer Go.

1 Introduction

MCTS, and especially UCT [9] appears in numerous search applications, such as [4]. Although these methods are shown to be successful empirically, most authors appear to be using UCT "because it has been shown to be successful in the past", and "because it does a good job of trading off exploration and exploitation". While the latter statement may be correct for the Multi-armed Bandit problem and for the UCB1 algorithm [1], we argue that a simple reconsideration from basic principles can result in schemes that outperform UCT.

The core issue is that in MCTS for adversarial search and search in "games against nature" the goal is typically to find the best first action of a good (or even optimal) policy, which is closer to minimizing the simple regret, rather than the cumulative regret minimized by UCB1. However, the simple and the cumulative regret cannot be minimized simultaneously; moreover, [3] shows that in many cases the smaller the cumulative regret, the greater the simple regret.

We begin with background definitions and related work. VOI estimates for arm pulls in MAB are presented, and a VOI-aware sampling policy is suggested, both for the simple regret in MAB and for MCTS. Finally, the performance of the proposed sampling policy is evaluated on sets of Bernoulli arms and on Computer GO, showing the improved performance.

2 Background and Related Work

Monte-Carlo tree search was initially suggested as a scheme for finding approximately optimal policies for Markov Decision Processes (MDP). MCTS explores an MDP by performing *rollouts*—trajectories from the current state to a state in which a termination condition is satisfied (either the goal or a cutoff state).

Taking a sequence of samples in order to minimize the regret of a decision based on the samples is captured by the Multi-armed Bandit problem (MAB) [11]. In MAB, we have a set of K arms. Each arm can be pulled multiple times. When the ith arm is pulled, a random reward X_i from an unknown stationary distribution is encountered. In the *cumulative setting*, all encountered rewards are collected. UCB1 [1] was shown to be near-optimal in this respect. UCT, an extension of UCB1 to MCTS is described in [9], and shown to outperform many state of the art search algorithms in both MDP and adversarial search [5, 4]. In the *simple regret setting*, the agent gets to collect only the reward of the last pull.

Definition 1. *The **simple regret** of a sampling policy for MAB is the expected difference between the best expected reward μ_* and the expected reward μ_j of the empirically best arm $\overline{X}_j = \max_i \overline{X}_i$:*

$$\mathbb{E}r = \sum_{j=1}^{K} \Delta_j \Pr(\overline{X}_j = \max_i \overline{X}_i) \tag{1}$$

where $\Delta_j = \mu_* - \mu_j$.

Strategies that minimize the simple regret are called pure exploration strategies [3].

A different scheme for control of sampling can use the principles of bounded rationality [8] and rational metareasoning [10, 6]. In search, one maintains a current best action α, and finds the expected gain from finding another action β to be better than the current best.

3 Upper Bounds on Value of Information

The intrinsic VOI Λ_i of pulling an arm is the expected decrease in the regret compared to selecting the best arm without pulling any arm at all. Two cases are possible:

- the arm α with the highest sample mean $\overline{X}_\alpha$ is pulled, and $\overline{X}_\alpha$ becomes lower than $\overline{X}_\beta$ of the second-best arm β;
- another arm i is pulled, and $\overline{X}_i$ becomes higher than $\overline{X}_\alpha$.

The *myopic* VOI estimate is of limited applicability to Monte-Carlo sampling, since the effect of a single sample is small, and the myopic VOI estimate will often be zero. However, for the common case of a fixed budget of samples per node, Λ_i can be estimated as the intrinsic VOI Λ_i^b of pulling the ith arm for the rest of the budget. Let us denote the current number of samples of the ith arm by n_i, and the remaining number of samples by N:

Theorem 1. *Λ_i^b is bounded from above as*

$$\Lambda_\alpha^b \le \frac{N\overline{X}_\beta}{N+n_\alpha} \Pr(\overline{X}'_\alpha \le \overline{X}_\beta) \le \frac{N\overline{X}_\beta}{n_\alpha} \Pr(\overline{X}'_\alpha \le \overline{X}_\beta) \tag{2}$$
$$\Lambda_{i|i\neq\alpha}^b \le \frac{N(1-\overline{X}_\alpha)}{N+n_i} \Pr(\overline{X}'_i \ge \overline{X}_\alpha) \le \frac{N(1-\overline{X}_\alpha)}{n_i} \Pr(\overline{X}'_i \ge \overline{X}_\alpha)$$

where $\overline{X}'_i$ is the sample mean of the ith arm after $n_i + N$ samples.

The probabilities can be bounded from above using the Hoeffding inequality [7]:

[1] Ben-Gurion University of the Negev, Israel, email: {tolpin,shimony}@cs.bgu.ac.il

Theorem 2. *The probabilities in equations (2) are bounded from above as*

$$\Pr(\overline{X}'_\alpha \le \overline{X}_\beta) \le 2\exp\left(-\varphi(n_\alpha)(\overline{X}_\alpha - \overline{X}_\beta)^2 n_\alpha\right)$$
$$\Pr(\overline{X}'_{i|i\neq\alpha} \ge \overline{X}_\beta) \le 2\exp\left(-\varphi(n_i)(\overline{X}_\alpha - \overline{X}_i)^2 n_i\right) \quad (3)$$

where $\varphi(n) = 2(\frac{1+n/N}{1+\sqrt{n/N}})^2 > 1.37$.

Corollary 1. *An upper bound on the VOI estimate* Λ_i^b *is obtained by substituting (3) into (2).*

$$\Lambda_\alpha^b \le \hat{\Lambda}_\alpha^b = \frac{2N\overline{X}_\beta}{n_\alpha}\exp\left(-1.37(\overline{X}_\alpha - \overline{X}_\beta)^2 n_\alpha\right)$$
$$\Lambda_{i|i\neq\alpha}^b \le \hat{\Lambda}_i^b = \frac{2N(1-\overline{X}_\alpha)}{n_i}\exp\left(-1.37(\overline{X}_\alpha - \overline{X}_i)^2 n_i\right) \quad (4)$$

4 VOI-based Sample Allocation

Following the principles of rational metareasoning, for pure exploration in Multi-armed Bandits an arm with the highest VOI should be pulled at each step. The upper bounds established in Corollary 1 can be used as VOI estimates. In MCTS, pure exploration takes place at the first step of a rollout, where an action with the highest utility must be chosen. MCTS differs from pure exploration in Multi-armed Bandits in that the distributions of the rewards are not stationary. However, VOI estimates computed as for stationary distributions work well in practice. As illustrated by the empirical evaluation (Section 5), estimates based on upper bounds on the VOI result in a rational sampling policy exceeding the performance of some state-of-the-art heuristic algorithms.

5 Empirical Evaluation

5.1 Selecting The Best Arm

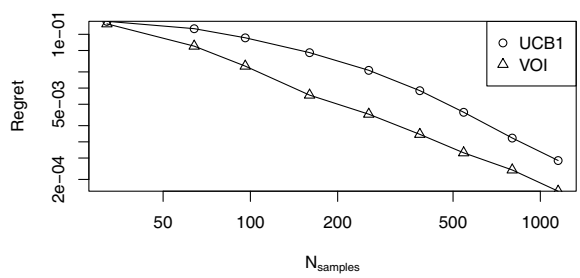

Figure 1. Random instances: regret vs. number of samples

The sampling policies are first compared on random Multi-armed bandit problem instances. Figure 1 shows results for randomly-generated Multi-armed bandits with 32 Bernoulli arms, with the mean rewards of the arms distributed uniformly in the range $[0, 1]$, for a range of sample budgets $32..1024$, with multiplicative step of 2. The experiment for each number of samples was repeated 10000 times. UCB1 is always considerably worse than the VOI-aware sampling policy.

5.2 Playing Go Against UCT

The policies were also compared on Computer Go, a search domain in which UCT-based MCTS has been particularly successful [5]. A modified version of Pachi [2], a state of the art Go program, was used for the experiments. The UCT engine was extended with a VOI-aware sampling policy, and a time allocation mode ensuring that both the original UCT policy and the VOI-aware policy use the same average number of samples per node was added. (While the UCT engine is not the most powerful engine of Pachi, it is still a strong player; on the other hand, additional features of more advanced engines would obstruct the MCTS phenomena which are the subject of the experiment.) The engines were compared on the 9x9 board, for 5000, 7000, 10000, and 15000 samples per ply, each experiment was repeated 1000 times. Figure 2 shows the winning rate of VOI against UCT vs. the number of samples. For most numbers of samples per node, VOI outperforms UCT.

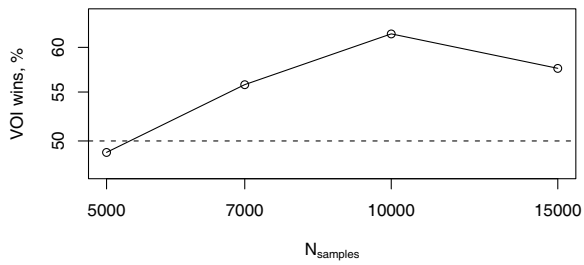

Figure 2. Go: winning rate — VOI against UCT

6 Summary and Future Work

This work suggested a Monte-Carlo sampling policy in which sample selection is based on upper bounds on the value of information. Empirical evaluation showed that this policy outperforms heuristic algorithms for pure exploration in MAB, as well as for MCTS.

MCTS still remains a largely unexplored field of application of VOI-aware algorithms. More elaborate VOI estimates, taking into consideration re-use of samples in future search states should be considered. The policy introduced in the paper differs from the UCT algorithm only at the first step, where the VOI-aware decisions are made. Consistent application of principles of rational metareasoning at all steps of a rollout may further improve the sampling.

REFERENCES

[1] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer, 'Finite-time analysis of the Multiarmed bandit problem', *Mach. Learn.*, **47**, 235–256, (May 2002).
[2] Petr Braudiš and Jean Loup Gailly, 'Pachi: State of the art open source Go program', in *ACG 13*, (2011).
[3] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz, 'Pure exploration in finitely-armed and continuous-armed bandits', *Theor. Comput. Sci.*, **412**(19), 1832–1852, (2011).
[4] Patrick Eyerich, Thomas Keller, and Malte Helmert, 'High-quality policies for the canadian travelers problem', in *In Proc. AAAI 2010*, pp. 51–58, (2010).
[5] Sylvain Gelly and Yizao Wang, 'Exploration exploitation in Go: UCT for Monte-Carlo Go', *Computer*, (2006).
[6] Nicholas Hay and Stuart J. Russell, 'Metareasoning for Monte Carlo tree search', Technical Report UCB/EECS-2011-119, EECS Department, University of California, Berkeley, (Nov 2011).
[7] Wassily Hoeffding, 'Probability inequalities for sums of bounded random variables', *Journal of the American Statistical Association*, **58**(301), pp. 13–30, (1963).
[8] Eric J. Horvitz, 'Reasoning about beliefs and actions under computational resource constraints', in *Proceedings of the 1987 Workshop on Uncertainty in Artificial Intelligence*, pp. 429–444, (1987).
[9] Levente Kocsis and Csaba Szepesvári, 'Bandit based Monte-Carlo planning', in *ECML*, pp. 282–293, (2006).
[10] Stuart Russell and Eric Wefald, *Do the right thing: studies in limited rationality*, MIT Press, Cambridge, MA, USA, 1991.
[11] Joannès Vermorel and Mehryar Mohri, 'Multi-armed bandit algorithms and empirical evaluation', in *ECML*, pp. 437–448, (2005).