MCTS Based on Simple Regret

David Tolpin, Solomon Eyal Shimony
Department of Computer Science,
Ben-Gurion University of the Negev, Beer Sheva, Israel
{tolpin,shimony}@cs.bgu.ac.il

Abstract

UCT, a state-of-the art algorithm for Monte Carlo tree search (MCTS) in games and Markov decision processes, is based on UCB, a sampling policy for the Multi-armed Bandit problem (MAB) that asymptotically 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". Therefore, it makes more sense to minimize the simple regret, as opposed to the cumulative regret. We begin by introducing policies for multi-armed bandits with lower finite-time and asymptotic simple regret than UCB, using it to develop a two-stage scheme (SR+CR) for MCTS which outperforms UCT empirically.

We then observe that optimizing the sampling process is itself a meta-reasoning problem, a solution of which can use value of information (VOI) techniques. Although the theory of VOI for search exists, applying it to MCTS is non-trivial, as typical myopic assumptions fail. Lacking a working VOI theory for MCTS, we nevertheless propose a sampling scheme that is "aware" of VOI, achieving an algorithm that in empirical evaluation outperforms both UCT and the other proposed algorithms.

Introduction

Monte-Carlo tree sampling, and especially a version based on the UCT formula (Kocsis and Szepesvári 2006) appears in numerous search applications, such as (Eyerich, Keller, and Helmert 2010). Although these methods are shown to be successful empirically, most authors appear to be using the UCT formula "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 and for the UCB method (Auer, Cesa-Bianchi, and Fischer 2002), we argue that it is inappropriate for search. The problem is not that UCT does not work; rather, we argue that a simple reconsideration from basic principles can result in schemes that outperform UCT.

The core issue is that in search (especially for adversarial search and for "games against nature" - optimizing behavior under uncertainty) the goal is typically to either find a good (or optimal) strategy, or even to find the best first action of such a policy. Once such an action is discovered, it is not beneficial to further sample that action, "exploitation" is thus meaningless for search problems. Finding a good first action is closer to the pure exploration variant, as seen in the selection problem (Bubeck, Munos, and Stoltz 2011; Tolpin and Shimony 2010). In the selection problem, it is much better to minimize the *simple* regret. However, the simple and the cumulative regret cannot be minimized simultaneously; moreover, (Bubeck, Munos, and Stoltz 2011) shows that in many cases the smaller the cumulative regret, the greater the simple regret.

We begin with introduction of several sampling schemes with better bounds for the simple regret on sets. We then extend the results to sampling in trees by combining the proposed sampling schemes on the first step of a rollout with UCT during the rest of the rollout. Finally, we empirically evaluate the performance of the proposed sampling schemes on sets of Bernoulli arms, in randomly generated trees, and on the sailing domain.

Related Work

Efficient algorithms for Multi-Armed Bandits based on distribution-independent bounds, in particular UCB1, are introduced in (Auer, Cesa-Bianchi, and Fischer 2002). The UCT algorithm, an extension of UCB1 to Monte-Carlo Tree Search is described in (Kocsis and Szepesvári 2006). Pure exploration in Multi-armed bandits is explored in (Bubeck, Munos, and Stoltz 2011). On the one hand, the paper proves certain upper and lower bounds for UCB1 and uniform sampling, showing that an upper bound on the simple regret is exponential in the number of samples for uniform sampling, while only polynomial for UCB1. On the other hand, empirical performance of UCB1 appears to be better than of uniform sampling.

The principles of bounded rationality appear in (Horvitz 1987). (Russell and Wefald 1991) provided a formal description of rational metareasoning and case studies of applications in several problem domains. One obstacle to straightforward application of the principles of rational metareasoning to Monte-Carlo sampling is the metagreedy assumption, according to which samples must be selected as though at most one sample can be taken before an action is chosen. In Monte-Carlo sampling, the value of information of any sin-

gle sample in a given search state is often zero, so a different approximating assumption must be used instead.

Sampling Based on Simple Regret

In the Multi-armed Bandit problem (Vermorel and Mohri 2005) 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 returned. The reward is bounded between 0 and 1.

Definition 1. *The **simple regret** of a sampling policy for the Multi-armed Bandit Problem 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 (Bubeck, Munos, and Stoltz 2011). Such strategies are used to select the best arm, and, by extension, the best action in MCTS. MCTS is used to solve Markov Decision Processes (MDP) approximately. An MDP is defined by the set of states S, the set of actions A, the transition table $T(s, a, s')$, the reward table $R(s, a, s')$, the initial state s and the goal state t: (S, A, T, R, s, t) (Russell and Norvig 2003). MCTS explores an MDP by *rollouts*—trajectories from the current state to a state in which a termination condition is satisfied (either the goal state, or a cutoff state for which the reward is evaluated approximately). The UCB algorithm (that attempts to minimize the cumulative regret) (Auer, Cesa-Bianchi, and Fischer 2002) had been extended into the tree search sampling scheme known as UCT (Kocsis and Szepesvári 2006).

UCT performance can be improved by combining UCB with a sampling scheme that minimizes the simple regret of selecting an action at the current root node. Indeed, the algorithm must select an action with the minimum regret *given the assumption that after performing the selected action the algorithm performs optimally*, which corresponds to maximizing the value of partial information (Russell and Norvig 2003). Therefore, an improved allocation scheme would

- maximize the value of partial information by sampling actions to minimize the **simple** regret of the selection at the current root node, and
- as the goal of sampling in deeper tree nodes is estimating the value of a node, rather than selection, it makes sense to minimize the **cumulative** regret of the rollouts from the second step onwards.

Ultimately, an "optimal" sampling in the meta-reasoning sense should be used at the first step in the above scheme. Nevertheless, this task is daunting for the following reasons:

- Defining the cost of a sample is not easy, and even if we simply use time-cost as an approximation, we get an intractable meta-reasoning problem.
- Applying the standard myopic and subtree independence assumptions, we run into serious problems. Even in the standard selection problem (Tolpin and Shimony 2010), we get a non-concave utility function and premature stopping of the algorithm. This is due to the fact that the value of information of a single measurement (analogous to a sample in MCTS) is frequently less than its time-cost, even though this is not true for multiple measurements. When applying the selection problem to MCTS, the situation is exacerbated. The utility of an action is usually bounded, and thus in many cases a single sample may be insufficient to change the current best action, *regardless* of its outcome. As a result, we frequently get a *zero* "myopic" value of information for a single sample.

As the ultimate goal is extremely difficult to achieve, and even harder to analyze, we introduce simple schemes more amenable to analysis, and compare them to UCB (on sets) and UCT (in trees).

Sampling on Sets

Definition 2. *Scheme $\mathbf{UCB}(\alpha)$ repeatedly pulls arm i that maximizes upper confidence bound b_i on the reward:*

$$b_i = \overline{X}_i + \sqrt{\frac{\alpha \ln n}{n_i}} \tag{2}$$

where $\overline{X}_i$ is the average reward obtained from arm i, n_i is the number of times arm i was pulled, and n is the total number of pulls so far.

The best known upper bound on the simple regret of UCB(α) is polynomially decreasing in the number of samples (see (Bubeck, Munos, and Stoltz 2011), Theorems 2,3).

An upper bound on the simple regret of uniform sampling is exponentially decreasing in the number of samples (see (Bubeck, Munos, and Stoltz 2011), Proposition 1). However, empirically UCB(α) yields a lower simple regret than uniform sampling.

We introduce here two sampling schemes with superpolynomially decreasing upper bounds on the simple regret. The bounds suggest that these schemes achieve a lower simple regret than uniform sampling; indeed, this is confirmed by experiments.

We first consider ε-greedy sampling as a straightforward generalization of uniform sampling:

Definition 3. *The ε-**greedy** sampling scheme pulls the empirically best arm with probability ε and any other arm with probability $\frac{1-\varepsilon}{K-1}$.*

This sampling scheme exhibits an exponentially decreasing simple regret:

Theorem 1. *For any $\pi > 0$ and $\gamma > 0$ there exists N such that for any number of samples $n > N$ the simple regret of the ε-greedy sampling scheme is bounded from above as*

$$\mathbb{E}r_{\varepsilon\text{-}greedy} \leq (1+\gamma) \sum_{i=1}^{K} \Delta_i \exp(\cdot) \tag{3}$$

with probability at least $1 - \pi$. the bound is minimized for $\varepsilon = \frac{1}{2}$.

Proof outline: TODO □

ε-greedy is based solely on sampling the empirically best arm with a higher probability then the rest of the arms, and ignores information about empirical means of other arms. On the other hand, UCB distributes samples in accordance with empirical means, but, in order to minimize cumulative regret, chooses the empirically best arm too often. Intuitively, a better scheme for simple regret minimization would distribute samples in a way similar to UCB, but would sample the best arm less. This can be achieved by replacing $\log(\cdot)$ in Equation 2 with a faster growing sublinear function, for example, $\sqrt{\cdot}$.

Definition 4. *Scheme* $\mathbf{UCB}_{\sqrt{\cdot}}(\alpha)$ *repeatedly pulls arm* i *that maximizes* b_i*:*

$$b_i = \overline{X}_i + \sqrt{\frac{\alpha\sqrt{n}}{n_i}} \tag{4}$$

where, as before, $\overline{X}_i$ *is the average reward obtained from arm* i*,* n_i *is the number of times arm* i *was pulled, and* n *is the total number of pulls so far.*

This scheme also exhibits a superpolynomially decreasing simple regret:

Theorem 2. *For any* $\pi > 0$ *and* $\gamma > 0$ *there exists* N *such that for any number of samples* $n > N$ *the simple regret of the UCB*$_{\sqrt{\cdot}}(\alpha)$ *sampling scheme is bounded from above as*

$$\mathbb{E}r_{ucb\sqrt{\cdot}} \leq (1+\gamma)\sum_{i=1}^{K}\Delta_i \exp\left(-2\alpha\sqrt{n}\right) \tag{5}$$

with probability at least $1-\pi$*.*

Proof outline: TODO □

Sampling in Trees

UCT (Kocsis and Szepesvári 2006) is a generalization of UCB for MCTS. UCT applies UCB at each step of a rollout. We suggest an improvement on UCT, which combines different sampling schemes on the first step and during the rest of a rollout:

Definition 5. *The* ***SR+CR MCTS sampling scheme*** *selects an action at the current root node according to a scheme suitable for minimizing the simple regret (****SR****), such as* $\frac{1}{2}$*-greedy or UCB*$_{\sqrt{\cdot}}$*, and then selects actions according to UCB, minimizing the cumulative regret (****CR****).*

Such two-stage sampling scheme outperforms UCT, and, additionally, is significantly less sensitive to the tuning of the exploration factor α of UCT, since the conflict between the need for a larger value of α on the first step (simple regret) and a smaller value for the rest of the rollout (cumulative regret) (Bubeck, Munos, and Stoltz 2011) is resolved. In fact, a sampling scheme that uses UCB at all steps but a larger value of α for the first step than for the rest of the steps, outperforms UCT. The pseudocode of the two-stage rollout is in Algorithm 1.

Algorithm 1 Two-stage Monte-Carlo tree search sampling

```
procedure ROLLOUT(node, depth=1)
    if ISLEAF(node, depth) then
        return 0
    else
        if depth=1 then
            action ← FIRSTACTION(node)
        else
            action ← NEXTACTION(node)
        end if
        next-node ← NEXTSTATE(node, action)
        reward ← REWARD(node, action, next-node)
                 + ROLLOUT(next-node, depth+1)
        UPDATESTATS(node, action, reward)
    end if
end procedure
```

VOI-aware Sampling

A further improvement can be achieved by computing or estimating the value of information (VOI) of the rollouts and choosing a rollout that maximizes the VOI. VOI of a rollout can be computed when the sample distribution of an action is known up to the parameters, such as the normal distribution with an unknown mean and/or variance. Alternatively, the value of information can be estimated from the set of samples, and the need to assume a particular shape of the distribution can be lifted. In one realization of the latter approach the VOI of performing an action is estimated as an *upper bound on the value of perfect information about an arm, divided by the number of pulls of the arm.*

$$\begin{aligned} VOI_\alpha &\approx \frac{\overline{X}_\beta}{n_\alpha+1}\exp\left(-2(\overline{X}_\alpha-\overline{X}_\beta)^2 n_\alpha\right) \qquad (6)\\ VOI_i &\approx \frac{1-\overline{X}_\alpha}{n_i+1}\exp\left(-2(\overline{X}_\alpha-\overline{X}_i)^2 n_i\right),\ i\neq\alpha \\ \text{where } \alpha &= \arg\max_i \overline{X}_i \\ \beta &= \arg\max_{i,\, i\neq\alpha} \overline{X}_i \end{aligned}$$

Early experiments with this approach demonstrated a significantly lower simple regret on a wide range of problems.

Empirical Evaluation

The results were empirically verified on Multi-armed Bandit instances, on search trees, and on the sailing domain, as defined in (Kocsis and Szepesvári 2006). In most cases, the experiments showed a lower average simple regret for $\frac{1}{2}$-greedy an UCB$_{\sqrt{\cdot}}$ than for UCB on sets, and for the SR+CR scheme than for UCT in trees.

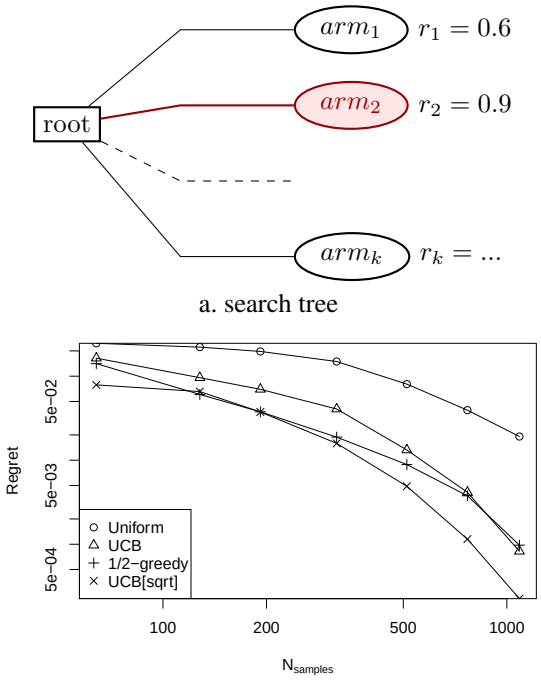
Simple regret in multi-armed bandits

Figure 1 presents a comparison of MCTS sampling schemes on Multi-armed bandits. Figure 1.a shows the search tree corresponding to a problem instance. Each arm returns a random reward drawn from a Bernoulli distribution. The search selects an arm and compares the expected reward, unknown

to the algorithm during the sampling, to the expected reward of the best arm. For such problem instances, the MCTS sampling schemes coincide with the selection algorithms employed at the first step of a rollout.

Figure 1.b shows the regret vs. the number of samples, averaged over 10^4 experiments for randomly generated instances of 32 arms.

For smaller numbers of samples, $\frac{1}{2}$-greedy achieves the best performance; for larger numbers of samples, UCB$_{\sqrt{\cdot}}$ outperforms $\frac{1}{2}$-greedy. A combination of $\frac{1}{2}$-greedy and UCB$_{\sqrt{\cdot}}$ dominates UCB over the whole range.



a. search tree

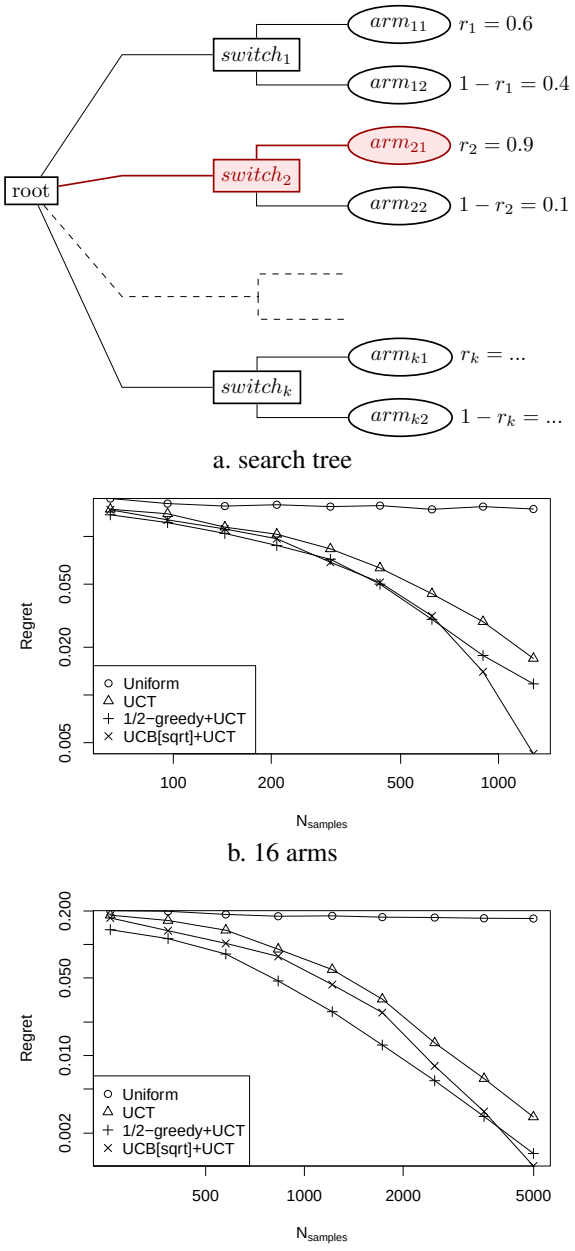
b. regret vs. number of samples

Figure 1: Simple regret in MAB

Monte Carlo tree search

The second set of experiments was performed on randomly generated trees crafted in such a way that uniform random sampling selects a direction at the root randomly. The degree of the root is a parameter of the tree generator. The degree of all nodes at distance 1 from the root is 2, and all nodes at distance 2 from the roots are leaves. The average reward of two children of each node at distance 1 is 0.5. Thus, a uniform sampling scheme results in the same average reward for all edges at the root, and an adaptive sampling scheme, such as UCT, has to be used.

Figure shows a sketch of the search tree (Figure .a) and the dependency of the regret vs. the number of samples for trees with root degree 16 (Figure .b) and 64 (Figure .c). The dependencies look differently from Multi-armed bandit instances, but the algorithms exhibit a similar relative performance: either $\frac{1}{2}$-greedy+UCT or UCB$_{\sqrt{\cdot}}$+UCT gives the lowest regret, UCB$_{\sqrt{\cdot}}$+UCT dominates UCT everywhere except for small numbers of instances. The advantage of both $\frac{1}{2}$-greedy+UCT and UCB$_{\sqrt{\cdot}}$+UCT grows with the number of arms.



a. search tree

b. 16 arms

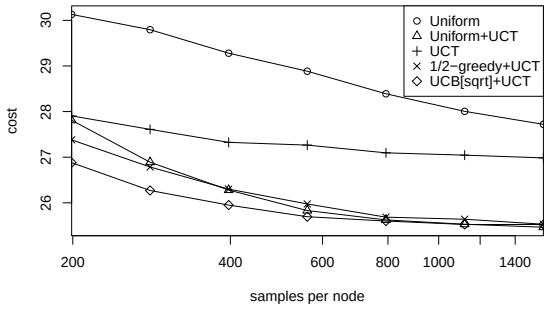
c. 64 arms

Figure 2: MCTS: a path to the best arm

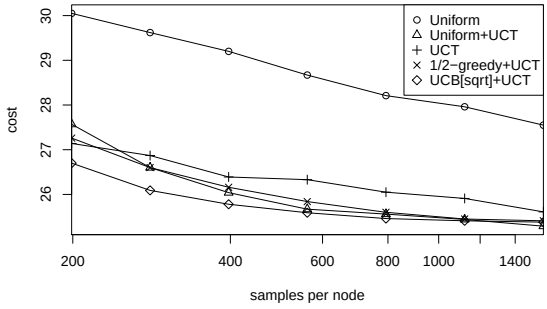
The sailing domain

Figures 3–6 show results of experiments on the sailing domain. Figure 3 shows the regret vs. the number of samples, computed for a range of values of α. Figure 3.a shows the median cost, and Figure 3.b — the minimum costs. UCT is always worse than either $\frac{1}{2}$-greedy+UCT or UCB$_{\sqrt{\cdot}}$+UCT, and is sensitive to the value of α: the median cost is much higher than the minimum cost for UCT.

For both $\frac{1}{2}$-greedy+UCT and UCB$_{\sqrt{\cdot}}$+UCT, the difference is significantly less prominent.



a. median cost



b. minimum cost

Figure 3: The sailing domain, 6×6 lake, cost vs. samples

Figure 4 shows the regret vs. the exploration factor for different numbers of samples. UCB$_{\sqrt{\cdot}}$+UCT is always better than UCT, and $\frac{1}{2}$-greedy+UCT is better than UCT expect for a small range of values of the exploration factor α.

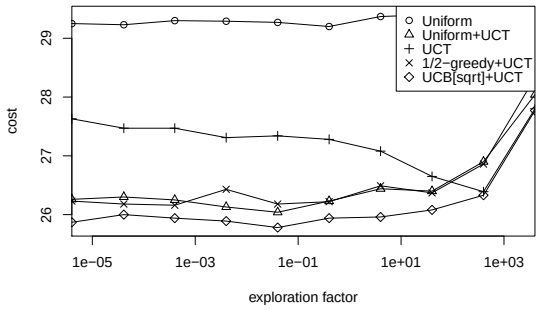
Figure 5 shows the cost vs. the exploration factor for lakes of different sizes. The relative difference between the sampling schemes becomes more prominent when the lake size increases.

Figure 6 compares UCB$_{\sqrt{\cdot}}$+UCT with UCT with a different exploration factor at the root (UCB+UCT). α for the rest of the steps was chosen to ensure the best performance from earlier experiments on the domain. Both algorithms exhibit similar dependency, but as the number of samples grows, UCB$_{\sqrt{\cdot}}$+UCT achieves smaller average regrets and is less sensitive to the choice of the value for α at the first step.

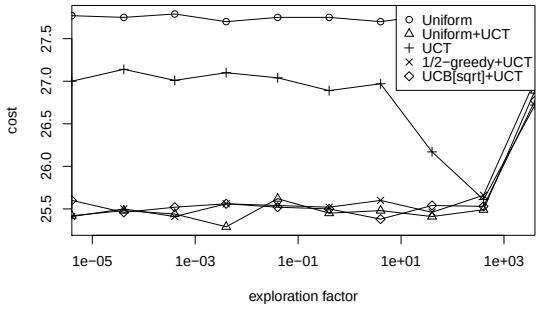
Summary and Future Work

The MCTS SR+CR scheme presented in the paper differs from UCT at the first step of the rollout, when the 'simple' selection regret is minimized instead of the cumulative regret. Both the theoretical analysis and the empirical evaluation provide evidence for better general performance of the proposed scheme.

The improvement is inspired by the notion of value of information (VOI), but VOI is used implicitly in the analysis of the algorithm, rather than computed or learned explicitly in order to plan the rollouts. Preliminary results on VOI-aware sampling suggest that a further improvement in the

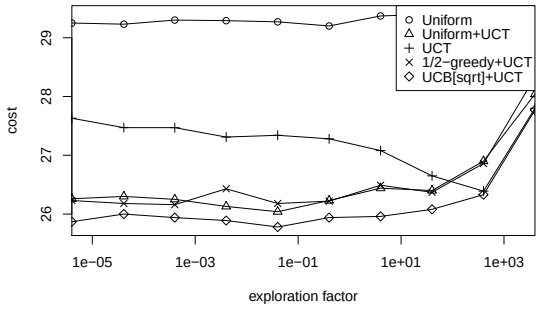


a. 397 rollouts

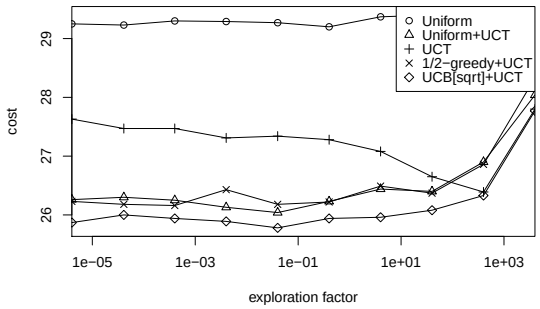


c. 1585 rollouts

Figure 4: The sailing domain, 6×6 lake, cost vs. factor



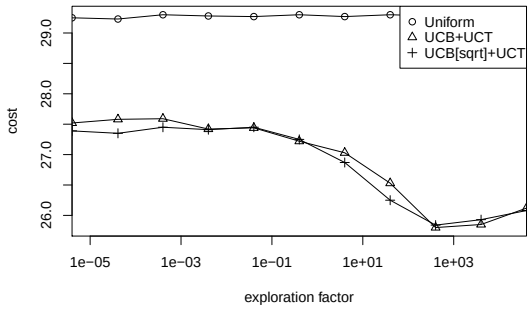
b. 6×6 lake



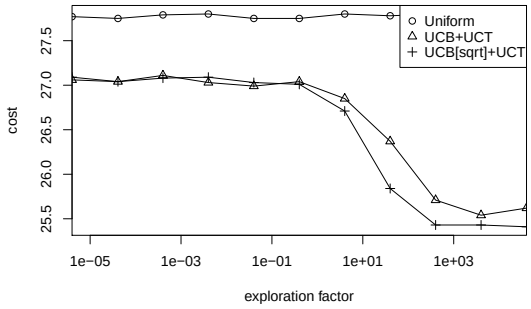
c. 10×10 lake

Figure 5: The sailing domain, 397 samples, cost vs. factor

sampling performance can be achieved. However, application of the theory of rational metareasoning to Monte Carlo Tree Search is an open problem (Hay and Russell 2011), and both a solid theoretical model and empirically efficient VOI



b. 397 rollouts



d. 1585 rollouts

Figure 6: The sailing domain, UCB vs. UCB$_{\sqrt{\cdot}}$, 6×6 lake

estimates still need to be developed.

Acknowledgments

The research is partially supported by Israel Science Foundation grant 305/09, by the Lynne and William Frankel Center for Computer Sciences, and by the Paul Ivanier Center for Robotics Research and Production Management.

References

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

Bubeck, S.; Munos, R.; and Stoltz, G. 2011. Pure exploration in finitely-armed and continuous-armed bandits. *Theor. Comput. Sci.* 412(19):1832–1852.

Eyerich, P.; Keller, T.; and Helmert, M. 2010. High-quality policies for the canadian travelers problem. In *In Proc. AAAI 2010*, 51–58.

Hay, N., and Russell, S. J. 2011. Metareasoning for monte carlo tree search. Technical Report UCB/EECS-2011-119, EECS Department, University of California, Berkeley.

Horvitz, E. J. 1987. Reasoning about beliefs and actions under computational resource constraints. In *Proceedings of the 1987 Workshop on Uncertainty in Artificial Intelligence*, 429–444.

Kocsis, L., and Szepesvári, C. 2006. Bandit based monte-carlo planning. In *ECML*, 282–293.

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

Russell, S., and Wefald, E. 1991. *Do the right thing: studies in limited rationality*. Cambridge, MA, USA: MIT Press.

Tolpin, D., and Shimony, S. E. 2010. Semi-myopic measurement selection for optimization under uncertainty. Technical Report 10-01, Lynne and William Frankel Center for Computer Science at Ben Gurion University of the Negev, Israel.

Vermorel, J., and Mohri, M. 2005. Multi-armed bandit algorithms and empirical evaluation. In *ECML*, 437–448.