VOI-aware Monte-Carlo Sampling in Trees

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

December 10, 2011

Abstract

Upper bounds for the VOI are provided for pure exploration in the Multi-armed Bandit Problem. Sampling policies based on the upper bounds are suggested. Empirical evaluation of the policies and comparison to the UCB1 and UCT policies is provided on random problem instances as well as on the Go game.

1 Introduction and Definitions

Taking a sequence of samples in order to minimize the regret of a decision based on the samples is abstracted by the *Multi-armed Bandit Problem.* In the Multi-armed Bandit problem 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.

The simple regret of a sampling policy for the Multi-armed Bandit Problem is the expected difference between the best expected reward $\max_i \mathbb{E}[X_i]$ and the expected reward $\mathbb{E}[X_j]$ of the arm with the best sample mean $\overline{X}_j = \max_i \overline{X}_i$:

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

where $\Delta_j = \max_i \mathbb{E}[X_i] - \mathbb{E}[X_j]$. Strategies that minimize the simple regret are called pure exploration strategies [3]. Principles of rational metareasoning [11] suggest that at each step the arm with the great value of information (VOI) must be pulled, and the sampling must be stopped and a decision must be made when no arm has positive VOI.

To estimate the VOI of pulling an arm, either a certain distribution of the rewards should be assumed (and updated based on observed rewards), or a distribution-independent bound on the VOI can be used as the VOI estimate. In this paper, we use *concentration inequalities* to derive distribution-independent bounds on the VOI.

2 Related Work

Efficient algorithms for Multi-Armed Bandits based on distribution-independent bounds, in particular UCB1, are introduced in [1]. The UCT algorithm, an extension of UCB1 to Monte-Carlo Tree Search is described in [8], and a successful application of UCT to playing the Go game is discussed in [5].

Pure exploration in Multi-armed bandits is explored in [3]. 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. [10] investigate stopping criteria for sampling based on the empirical Bernstein inequality; however, the stopping criteria are based on error probabilities rather than on value of information measures, and do not directly address the objective of regret minimization.

The principles of bounded rationality appear in [7]. [11] 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 single sample in a given search state is often zero, so a different approximating assumption must be used instead.

3 Some Concentration Inequalities

Let $X_1, \ldots, X_n$ be i.i.d. random variables with values from $[0, 1]$, $X = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

Hoeffding inequality [6]:

$$\Pr(X - \mathbb{E}[X] \geq a) \leq \exp(-2a^2 n) \tag{2}$$

Empirical Bernstein inequality (derived in Appendix 9):

$$\begin{aligned} \Pr(X - \mathbb{E}[X] \geq a) &\leq 2\exp\left(-\frac{a^2 n}{\frac{14}{3}\frac{n}{n-1}a + 2\overline{\sigma}_n^2}\right) \\ &\leq 2\exp\left(-\frac{a^2 n}{10a + 2\overline{\sigma}_n^2}\right) \end{aligned} \tag{3}$$

where sample variance $\overline{\sigma}_n^2$ is

$$\overline{\sigma}_n^2 = \frac{1}{n(n-1)} \sum_{1 \leq i < j \leq n} (X_i - X_j)^2 \tag{4}$$

Bounds (2, 3) are symmetrical around the mean. Bound (3) is tighter than (2) for small a and $\overline{\sigma}_n^2$.

4 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 an arm without pulling any arm at all. 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 non-positive, resulting in premature termination of the search. However, Λ_i can be estimated as the intrinsic value of perfect information Λ_i^p about the mean reward of the ith arm. Two cases are possible:

- the arm α with the highest sample mean is pulled, and the true mean of the arm is lower than the sample mean of the second-best arm β;
- another arm is pulled, and the true mean of the arm is higher than the sample mean of the best arm α.

Λ_i^p can be bounded from above as the probability that a different arm is selected, multiplied by the maximum possible increase in the reward:

Theorem 1. *The intrinsic value of perfect information Λ_i^p about the ith arm is bounded from above as*

$$\Lambda_i^p \leq \begin{cases} \overline{X}_\beta \Pr(\mathbb{E}[X_i] \leq \overline{X}_\beta) & \text{if } i = \alpha \\ (1 - \overline{X}_\alpha) \Pr(\mathbb{E}[X_i] \geq \overline{X}_\alpha) & \text{otherwise} \end{cases} \tag{5}$$

Proof. For the case $i \neq \alpha$, the probability that the perfect information about the ith arm changes the final choice is $\Pr(\mathbb{E}[X_i] \geq \overline{X}_\alpha)$. $\mathbb{E}[X_i] \leq 1$ by definition, therefore the maximum increase in the expected reward is $(1 - \overline{X}_\alpha)$. Thus the intrinsic value of perfect information is at most $\Pr(\mathbb{E}[X_i] \geq \overline{X}_\alpha)(1 - \overline{X}_\alpha)$. Proof for the case $i = \alpha$ is similar. □

The search time is finite, and in a simple case the *search budget* is specified as the maximum number of samples. An estimate based on the perfect intrinsic VOI does not take in consideration the remaining number of samples. Given two arms with the same intrinsic perfect VOI, the VOI estimate of the arm pulled fewer times so far should be higher.

Definition 1. *The **blinkered estimate** of intrinsic VOI information of the ith arm is the intrinsic VOI of pulling the ith arm for the remaining budget.*

Theorem 2. *Denote the current number of samples of the ith arm as n_i. The blinkered estimate Λ_i^b of intrinsic value of information of pulling the ith arm for the remaining budget of N samples is bounded from above as*

$$\Lambda_i^b \leq \begin{cases} \overline{X}_\beta \frac{N}{N+n_i} \Pr(\overline{X}'_i \leq \overline{X}_\beta) \leq N \frac{\overline{X}_\beta}{n_i} \Pr(\overline{X}'_i \leq \overline{X}_\beta) & \text{if } i = \alpha \\ (1 - \overline{X}_\alpha) \frac{N}{N+n_i} \Pr(\overline{X}'_i \geq \overline{X}_\alpha) \leq N \frac{(1-\overline{X}_\alpha)}{n_i} \Pr(\overline{X}'_i \geq \overline{X}_\alpha) & \text{otherwise} \end{cases} \tag{6}$$

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

Proof. The proof is similar to the proof of Theorem 1, with $\overline{X}'_i$ substituted instead of $\mathbb{E}[X_i]$, and, if $i \neq \alpha$, an upper bound on X'_i is $X_\alpha + (1 - X_\alpha)\frac{N}{N+n_i}$; if $i = \alpha$, a lower bound on X'_i is $X_\beta - X_\beta \frac{N}{N+n_i}$. □

The probabilities in equations (5, 6) can be bounded from above using concentration inequalities. In particular, Lemma 1 (proved in Appendix 8) is based on the Hoeffding inequality (2):

Lemma 1. *The probabilities in equations (5, 6) are bounded from above as*

$$\begin{aligned} \Pr(\mathbb{E}[X_i] \leq \overline{X}_\beta | i = \alpha) &\leq \exp(-2(\overline{X}_i - \overline{X}_\beta)^2 n_i) \\ \Pr(\mathbb{E}[X_i] \geq \overline{X}_\alpha | i \neq \alpha) &\leq \exp(-2(\overline{X}_\alpha - \overline{X}_i)^2 n_i) \end{aligned} \tag{7}$$

$$\Pr(\overline{X}'_i \leq \overline{X}_\beta | i = \alpha) \leq 2 \exp\left(-2 \left(\frac{1 + \frac{n_i}{N}}{1 + \sqrt{\frac{n_i}{N}}} (\overline{X}_i - \overline{X}_\beta) \right)^2 n_i \right) \leq 2 \exp\left(-1.37(\overline{X}_i - \overline{X}_\beta)^2 n_i\right)$$

$$\Pr(\overline{X}'_i \geq \overline{X}_\alpha | i \neq \alpha) \leq 2 \exp\left(-2 \left(\frac{1 + \frac{n_i}{N}}{1 + \sqrt{\frac{n_i}{N}}} (\overline{X}_\alpha - \overline{X}_i) \right)^2 n_i \right) \leq 2 \exp\left(-1.37(\overline{X}_\alpha - \overline{X}_i)^2 n_i\right) \tag{8}$$

An upper bound on the intrinsic value of perfect information is obtained by substituting (7) into (5):

$$\Lambda_i^p \leq \hat{\Lambda}_i^p = \begin{cases} \overline{X}_\beta \cdot \exp\left(-2(\overline{X}_i - \overline{X}_\beta)^2 n_i\right) & \text{if } i = \alpha \\ (1 - \overline{X}_\alpha) \cdot \exp\left(-2(\overline{X}_\alpha - \overline{X}_i)^2 n_i\right) & \text{otherwise} \end{cases} \tag{9}$$

An upper bound on the blinkered VOI estimate with ordering independent of N is obtained by substituting (8) into (6).

$$\Lambda_i^b \leq \hat{\Lambda}_i^b = \begin{cases} N\frac{\overline{X}_\beta}{n_i} \cdot 2\exp\left(-1.37(\overline{X}_i - \overline{X}_\beta)^2 n_i\right) & \text{if } i = \alpha \\ N\frac{(1-\overline{X}_\alpha)}{n_i} \cdot 2\exp\left(-1.37(\overline{X}_\alpha - \overline{X}_i)^2 n_i\right) & \text{otherwise} \end{cases} \tag{10}$$

The bound is generally tighter when N is small relative to n_i, and the estimate $\hat{\Lambda}_i^b$ can be viewed as a relaxation of the myopic VOI estimate to longer sequences of actions, such that the VOI estimate is positive for some of the actions.

Better bounds can be obtained through tighter estimates on the probabilities, for example, based on the empirical Bernstein inequality (3) or through a more careful application of the Hoeffding inequality (Appendix 10).

5 VOI-based Sampling Control

5.1 Selection Criterion

Following the principles of rational metareasoning, an arm with the highest VOI should be pulled at each step. The upper bounds established in Theorems 1, 2 can be used as VOI estimates. Blinkered VOI estimates (6, 10) can be viewed as approximations of VOI estimates for myopic or receding horizon policies. As illustrated by the empirical evaluation (Section 6), estimates based on upper bounds on the VOI result in rational sampling policies, and exhibit performance comparable to the performance of some state-of-the-art heuristic algorithms.

5.2 Termination Condition

The simplest termination condition for a sampling policy is the budget—a fixed number of samples performed before a decision is made. When a sample has a cost commensurable with the value of information of a measurement, an upper bound on the intrinsic VOI can be used to stop the sampling if the intrinsic VOI of any action is less than the cost of sampling C:

$$\text{stop if } \max_i \Lambda_i \leq C \tag{11}$$

Blinkered VOI estimates (6, 10) include the remaining budget N as a factor, but given the cost of a single sample c, the cost of the remaining samples accounted for in estimating the intrinsic VOI is $C = cN$. N can be dropped on both sides of the inequality, and a viable stopping condition is

$$\begin{aligned} \tfrac{1}{N}\Lambda_\alpha^b \leq \tfrac{\overline{X}_\beta}{n_\alpha}\Pr(\overline{X}'_\alpha \leq \overline{X}_\beta) \quad &\leq \quad c \\ \text{and} \\ \tfrac{1}{N}\max_i \Lambda_i^b \leq \max_i \tfrac{(1-\overline{X}_\alpha)}{n_i}\Pr(\overline{X}'_i \geq \overline{X}_\alpha) \quad &\leq \quad c \quad \forall i : i \neq \alpha \end{aligned}$$

The empirical evaluation (Section 6) confirms viability of this stopping condition and illustrates the influence of the sample cost c on the performance of the sampling policy.

5.3 Sample Redistribution in Trees

Monte-Carlo tree search [4] selects at each search step an action that appears to be the best based on outcomes of *search rollouts*. Two different criteria are employed in action selection during a rollout:

- at the first step, the simple regret of selecting an action must be minimized;

- starting with the second step of a rollout, the expected reward of the rollout must be made as close as possible to the optimum reward, so that the value of the action at the first step of the rollout is evaluated correctly. Thus, starting with the second step, the cumulative regret must be minimized.

A natural approach would be to apply an algorithm that minimizes the simple regret in Multi-armed bandits at the first step, and the cumulative regret during the rest of the rollout. However, this approach assumes that the information obtained from rollouts in the current state is discarded after an action is selected. In practice, most successful Monte-Carlo tree search algorithms re-use rollouts originating from an earlier search state and passing through the current search state; thus, the value of information of a rollout is determined not just by the influence on the choice of the action at the current state, but also by the influence on the choice at future search states, provided the rollout passes through the search states which will be visited.

One way to account for this re-use would be to incorporate the 'future' value of information into a VOI estimate. However, this approach appears to be complicated. Alternatively, one can behave myopically in search tree depth:

1. estimate VOI as though the information is discarded after each step;
2. stop early if the VOI is below a certain threshold (see Section 5.2);
3. save the unused sample budget for search in future state, such that if the nominal budget is N, and the unused budget in the last state is N_u, the search budget in the next state will be $N + N_u$.

In this approach, the cost of a sample in the current state is the VOI of increasing the budget of a future state by one sample. It is unclear whether this cost can be accurately estimated, but supposedly a fixed value for a given problem type and algorithm implementation would work. Indeed, the empirical evaluation (Section 6.2) confirms that stopping and sample redistribution based on a learned fixed cost substantially improve the performance of the VOI-based sampling policy in game tree search.

6 Empirical Evaluation

The experiments compare the UCT algorithm [8] with modified versions of UCT in which the samples at the first step are selected according to their VOI estimates. Three blinkered VOI estimates were used for the comparison:

VCT: based on Hoeffding inequality (2);

ECT: based on Hoeffding inequality with midpoint (see Appendix 10);

BCT: based on empirical Bernstein inequality (3).

For simplicity, the bounds on probability for unlimited number of samples were used in the estimate formulas. The estimates where computed as follows:

$$
\begin{aligned}
\Lambda_i^{VCT} &= \begin{cases} \frac{\overline{X}_\beta}{n_i} \cdot \exp\left(-2(\overline{X}_i - \overline{X}_\beta)^2 n_i\right) & \text{if } i = \alpha \\ \frac{(1-\overline{X}_\alpha)}{n_i} \cdot \exp\left(-2(\overline{X}_\alpha - \overline{X}_i)^2 n_i\right) & \text{otherwise} \end{cases} \\
\Lambda_i^{ECT} &= \min\left(\Lambda_i^{VCT}, \min_y \begin{cases} \frac{\overline{X}_\beta - y}{n_i} \cdot \exp\left(-2(\overline{X}_i - \overline{X}_\beta)^2 n_i\right) + \frac{\overline{X}_\beta}{n_i} \cdot \exp\left(-2(\overline{X}_i - y)^2 n_i\right) & \text{if } i = \alpha \\ \frac{y - \overline{X}_\alpha}{n_i} \cdot \exp\left(-2(\overline{X}_\alpha - \overline{X}_i)^2 n_i\right) + \frac{1-\overline{X}_\alpha}{n_i} \cdot \exp\left(-2(y - \overline{X}_i)^2 n_i\right) & \text{otherwise} \end{cases}\right) \\
\Lambda_i^{BCT} &= \min\left(\Lambda_i^{VCT}, \begin{cases} \frac{\overline{X}_\beta}{n_i} \cdot 2\exp\left(-\frac{(\overline{X}_i - \overline{X}_\beta)^2 n_i}{\frac{14}{3}\frac{n_i}{n_i-1}(\overline{X}_i - \overline{X}_\beta) + 2\overline{X}_i(1-\overline{X}_i)}\right) & \text{if } i = \alpha \\ \frac{(1-\overline{X}_\alpha)}{n_i} \cdot 2\exp\left(-\frac{(\overline{X}_\alpha - \overline{X}_i)^2 n_i}{\frac{14}{3}\frac{n_i}{n_i-1}(\overline{X}_\alpha - \overline{X}_i) + 2\overline{X}_i(1-\overline{X}_i)}\right) & \text{otherwise} \end{cases}\right)
\end{aligned}
\tag{12}
$$

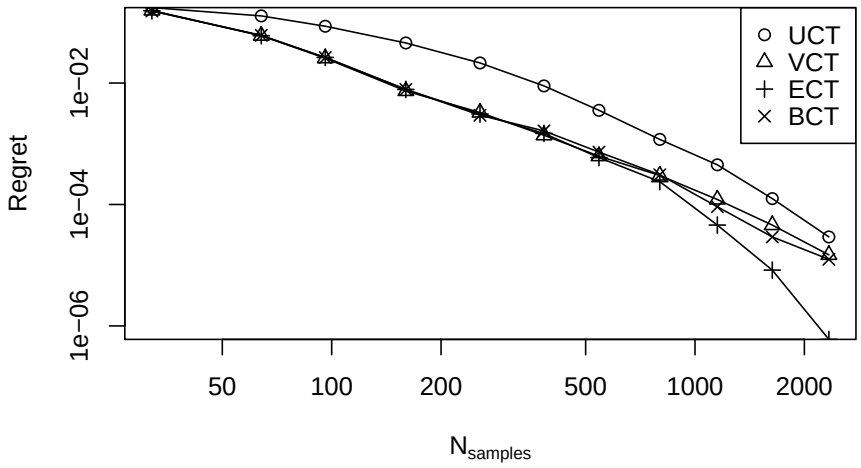

Figure 1: Random instances: regret vs. number of samples

6.1 Selecting The Best Arm

The sampling policies are first compared on random Multi-armed bandit problem instances, where the Monte-Carlo Tree Search algorithm is reduced to the sample selection policy at the first step: UCT is actually UCB1. In the problem of pure exploration in Multi-armed bandits [3], the smaller is the exploration factor, the higher is the expected simple regret; indeed, UCT exhibits the lowest simple regret in the experiments for the default exploration factor $Cp = \frac{\sqrt{2}}{2}$ [8].

Figure 1 shows experiment 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..2048, with multiplicative step of 2. The experiment for each number of samples was repeated 20000 times. UCT is worse than any of the VOI-aware sampling policies, for example, for 384 samples the simple regret of UCT is ≈ 40 times greater than of VCT, and ≈ 100 times greater than of ECT. On average, it takes 2.5 times as many samples for UCT to reach the regret of VCT. Different numbers of arms and distributions of means give similar results.

6.2 Playing Go Against UCT

One search domain in which Monte-Carlo tree search based on the UCT sampling policy has been particularly successful is playing the Go game [5]. This series of experiments compares the winning rate of UCT against VOI-aware policies (VCT, ECT, BCT) in Go and investigates the effect of early stopping and sample redistribution.

For the experiments, a modified version of Pachi [2], a state of the art Go program, was used:

- The UCT engine of Pachi was extended with VOI-aware sampling policies at the first step.
- The stopping condition for the VOI-aware policies was modified and based solely on the sample cost, specified as a constant parameter.
- The time-allocation mode based on the fixed number of samples was modified such that
  1. the same number of samples is available to the agent at each step, independently of the number of pre-simulated games;
  2. if samples were unused at the current step, they become available at the next step.

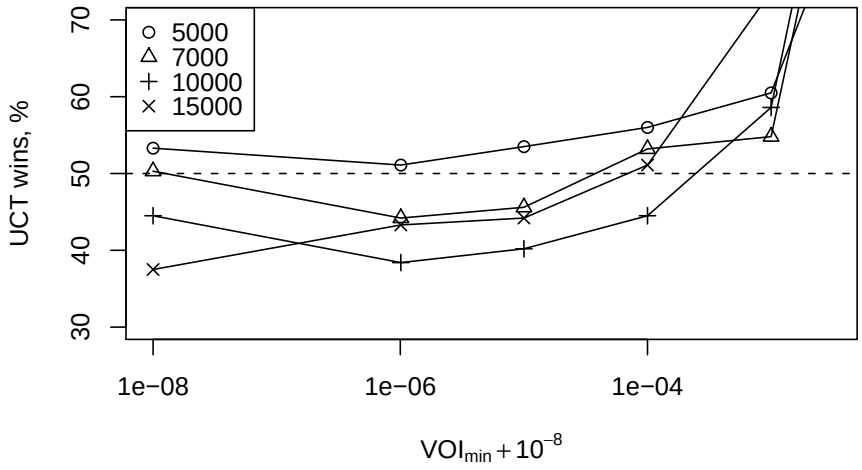

Figure 2: Go: winning rate — UCT against VCT

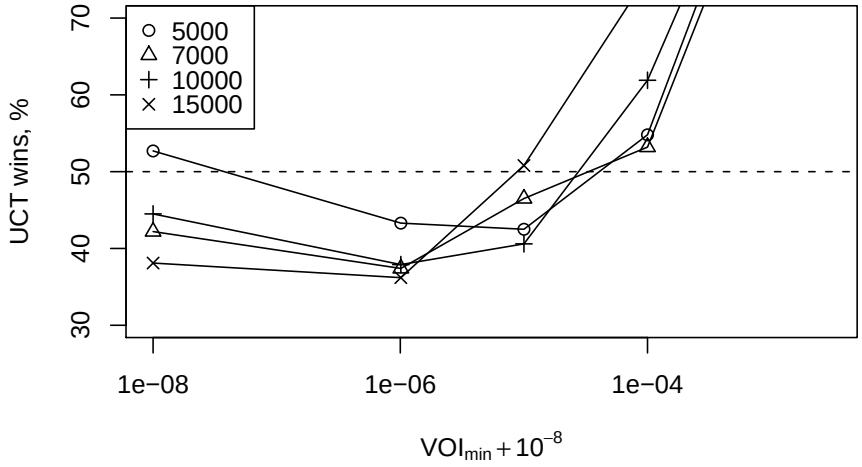

Figure 3: Go: winning rate — UCT against ECT

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 Monte-Carlo sampling phenomena which are the subject of the experiment.

The engines were compared on the 9x9 board, for 5000, 7000, 1000, and 15000 samples (game simulations) per ply, each experiment was repeated 1000 times. Figures 2–4 show the winning rate of UCT against VCT, ECT, BCT, correspondingly, vs. the stopping threshold (if the maximum VOI of a sample is below the threshold, the simulation is stopped, and a move is chosen). Each curve in the figures corresponds to a certain number of samples per ply. The lower the curve, the better is the corresponding policy compared to UCT.

As the results show, without sample redistribution UCT is as good as or better than the VOI-aware policies for lower numbers of samples per ply (5000, 7000), and the advantage of VOI-aware policies for larger number of samples is rather modest (e.g., ECT wins in 55% of games for 10000 samples per ply). For the stopping threshold of 10^{-6}, however, the VOI-aware policies are almost always better than UCT (only VCT for 5000 samples wins in less than 50% of games), with stronger policies (ECT, VCT) reaching the winning rate of 66%.

In agreement with intuition (Section 5.3), VOI-based stopping and sample redistribution is most

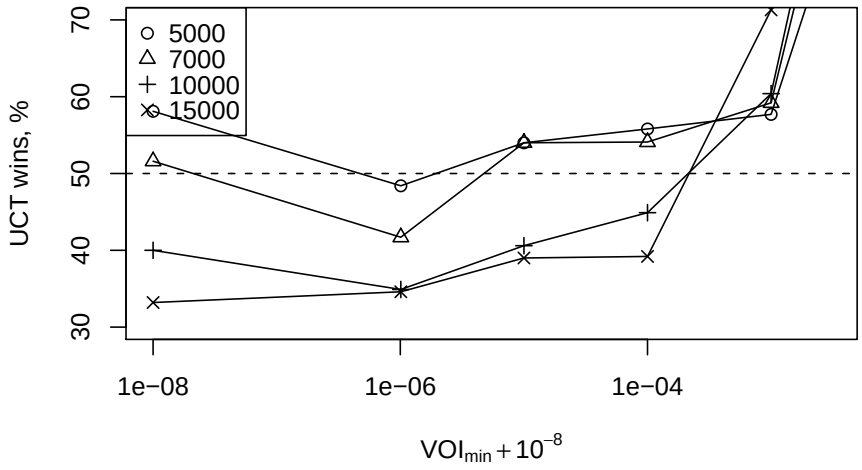

Figure 4: Go: winning rate — UCT against BCT

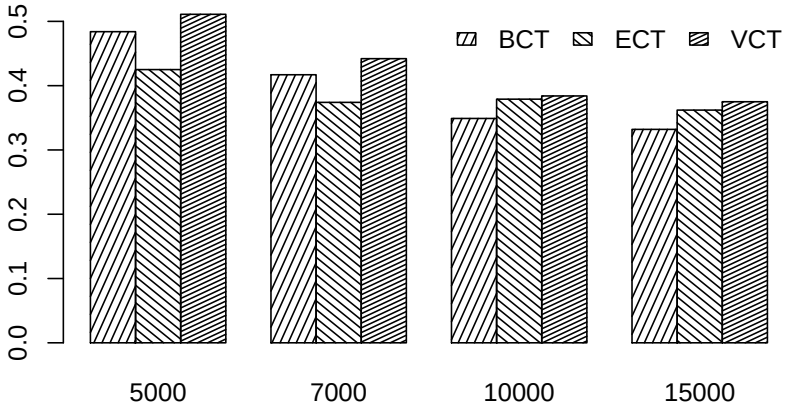

Figure 5: Go: best winning rate comparison

influential for medium numbers of samples per ply. When the maximum number of samples is too low, early stopping would result in poorly selected moves. On the other hand, when the maximum number of samples is sufficiently high, the VOI of increasing the maximum number of samples in a future state is low.

Figure 5 presents relative performance of different VOI-aware policies depending on the number of samples per ply by comparing the best achieved winning rates against UCT (again, the lower the bar, the better is the policy). Both ECT (based on the Hoeffding inequality with midpoint) and BCT (based on the empirical Bernstein inequality) are better than VCT. The difference between ECT and VCT is most prominent for smaller number of samples, and the winning rate of ECT increases relatively slowly with the number of samples. In contrast, BCT is worse than ECT and only slightly better than VCT for smaller number of samples, but the winning rate rapidly increases with the number of samples, and for 10000 and 15000 samples per ply BCT is the best of the policies with the winning rate reaching 67%.

7 Summary and Future Work

This work suggested Monte-Carlo sampling policies in which sample selection and termination are based on upper bounds on the value of information. A simplified model of accounting for future value of information of a sample, based on early termination and sample redistribution, was proposed for Monte-Carlo sampling in trees. Empirical evaluation showed that these policies outperform heuristic algorithms for simple regret minimization in Multi-armed bandits, as well as for tree search.

Monte-Carlo tree search still remains a largely unexplored field of application of VOI-aware algorithms. More elaborated VOI estimates, taking into consideration re-use of samples in future search states should be considered. The policies introduced in the paper differ 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 policies.

8 Proof of Lemma 1

. Equations (7) is a direct application of the Hoeffding inequality (2).

Equations (8) follow from the observation that if $i \neq \alpha$, $\overline{X}_i' > \overline{X}_\alpha$ if and only if the mean $\overline{X}_i^+$ of N samples from $n_i + 1$ to N is at least $\overline{X}_\alpha + (\overline{X}_\alpha - \overline{X}_i)\frac{n_i}{N}$.

For any δ, the probability that $\overline{X}_i'$ is greater than $\overline{X}_\alpha$ is less than the probability that $\mathbb{E}[X_i] \geq \overline{X}_i + \delta$ *or* $\overline{X}_i^+ \geq \mathbb{E}[X_i] + \overline{X}_\alpha - \overline{X}_i - \delta + (\overline{X}_\alpha - \overline{X}_i)\frac{n_i}{N}$, thus, by the union bound, less than the sum of the probabilities:

$$\Pr(\overline{X}_i' \geq \overline{X}_\alpha) \leq \Pr(\mathbb{E}[X_i] - \overline{X}_i \geq \delta) + \Pr\left(\overline{X}_i^+ - \mathbb{E}[X_i] \geq \overline{X}_\alpha - \overline{X}_i - \delta + (\overline{X}_\alpha - \overline{X}_i)\frac{n_i}{N}\right) \tag{13}$$

Bounding the probabilities on the right-hand side using the Hoeffding inequality, obtain:

$$\Pr(\overline{X}_i' \geq \overline{X}_\alpha) \leq \exp(-2\delta^2 n_i) + \exp\left(-2\left((\overline{X}_\alpha - \overline{X}_i)\left(1 + \frac{n_i}{N}\right) - \delta\right)^2 N\right) \tag{14}$$

Find δ for which the two terms on the right-hand side of (14) are equal:

$$\exp(-\delta^2 n) = \exp(-2\left((\overline{X}_\alpha - \overline{X}_i)(1 + \frac{n_i}{N}) - \delta\right)^2 N) \tag{15}$$

Solve (15) for δ: $\delta = \frac{1+\frac{n_i}{N}}{1+\sqrt{\frac{n_i}{N}}}(\overline{X}_\alpha - \overline{X}_i) \geq (\sqrt{2}-1)(\overline{X}_\alpha - \overline{X}_i)$. Substitute δ into (14) and obtain

$$\begin{aligned}
\Pr(\overline{X}_i' \geq \overline{X}_\alpha) &\leq 2\exp\left(-2\left(\frac{1+\frac{n_i}{N}}{1+\sqrt{\frac{n_i}{N}}}(\overline{X}_\alpha - \overline{X}_i)\right)^2 n_i\right) \\
&\leq 2\exp(-2(\sqrt{2}-1)^2(\overline{X}_\alpha - \overline{X}_i)^2 n_i) \leq 2\exp(-1.37(\overline{X}_\alpha - \overline{X}_i)^2 n_i)
\end{aligned} \tag{16}$$

Derivation for the case $i = \alpha$ is similar. □

9 Empirical Bernstein Inequality

Theorem 4 in [9] states that

$$\Pr\left(\mathbb{E}[X] - \overline{X}_n \geq \sqrt{\frac{2\overline{\sigma}_n^2 \ln 2/\delta}{n}} + \frac{7\ln 2/\delta}{3(n-1)}\right) \leq \delta \tag{17}$$

Therefore

$$\Pr\left(\mathbb{E}[X] - \overline{X}_n \geq \sqrt{\left(\frac{7\ln 2/\delta}{3(n-1)}\right)^2 + \frac{2\overline{\sigma}_n^2 \ln 2/\delta}{n}} + \frac{7\ln 2/\delta}{3(n-1)}\right) \leq \delta. \tag{18}$$

$a = \sqrt{\left(\frac{7\ln 2/\delta}{3(n-1)}\right)^2 + \frac{2\overline{\sigma}_n^2 \ln 2/\delta}{n}} + \frac{7\ln 2/\delta}{3(n-1)}$ is a root of square equation $a^2 - a\frac{14\ln 2/\delta}{3(n-1)} - \frac{2\overline{\sigma}_n^2 \ln 2/\delta}{n} = 0$ which, solved for $\delta \triangleq \Pr(\mathbb{E}[X] - \overline{X}_n \geq a)$, gives

$$\Pr(\mathbb{E}[X] - \overline{X}_n \geq a) \leq 2\exp\left(-\frac{na^2}{\frac{14}{3}\frac{n}{n-1}a + 2\overline{\sigma}_n^2}\right)$$

Other derivations, giving slightly different results, are possible.

10 Better Hoeffding-Based Bound on Value of Perfect Information

Bounds (9, 10) have the form

$$\Lambda_i \leq \hat{\Lambda}_i = \begin{cases} (\overline{X}_\beta - B)\exp\left(-\gamma(\overline{X}_i - \overline{X}_\beta)^2 n_i\right) & \text{if } i = \alpha \\ (A - \overline{X}_\alpha)\exp\left(-\gamma(\overline{X}_\alpha - \overline{X}_i)^2 n_i\right) & \text{otherwise} \end{cases} \tag{19}$$

where A is the upper bound on the posterior sample mean $\overline{X}'_i$ if $i \neq \alpha$, B is the lower bound on $\overline{X}'_i$ if $i = \alpha$. For $i = \alpha$, the bound can be improved by selecting a midpoint $B < y < \overline{X}_\beta$ and computing the bound as the sum of two parts:

- $\overline{X}_\beta - y$ multiplied by the probability that $\overline{X}'_i \leq \overline{X}_\beta$;
- $\overline{X}_\beta - B$ multiplied by the probability that $\overline{X}'_i \leq y$.

.

$$\Lambda_{i|i=\alpha} \leq \hat{\Lambda}_{i|i=\alpha} = (\overline{X}_\beta - y)\exp\left(-\gamma(\overline{X}_i - \overline{X}_\beta)^2 n_i\right) + (\overline{X}_\beta - B)\exp\left(-\gamma(\overline{X}_i - y)^2 n_i\right) \tag{20}$$

If y that minimizes $\hat{\Lambda}_{i|i=\alpha}$ exists, the minimum is achieved when $\frac{d\hat{\Lambda}_{i|i=\alpha}}{dy} = 0$, that is, when y is the root of the following equation:

$$2\gamma(\overline{X}_\beta - B)(\overline{X}_i - y)n_i = \exp\left(-\gamma\left(\frac{\overline{X}_i - \overline{X}_\beta}{\overline{X}_i - y}\right)^2 n_i\right) \tag{21}$$

The derivation for the case $i \neq \alpha$ is obtained by substitution $1 - A, 1 - \overline{X}_i, 1 - \overline{X}_\alpha, 1 - y$ instead of $B, \overline{X}_i, \overline{X}_\beta, y$ into (20):

$$\Lambda_{i|i\neq\alpha} \leq \hat{\Lambda}_{i|i\neq\alpha} = (y - \overline{X}_\alpha)\exp\left(-\gamma(\overline{X}_\alpha - \overline{X}_i)^2 n_i\right) + (A - \overline{X}_\alpha)\exp\left(-\gamma(y - \overline{X}_i)^2 n_i\right) \tag{22}$$

A closed-form solution for y cannot be obtained, but given $A, B, \overline{X}_\alpha, \overline{X}_\beta, n$, the value of y can be efficiently computed.

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] Guillaume Chaslot, Sander Bakkes, Istvan Szita, and Pieter Spronck. Monte-Carlo tree search: A new framework for game AI. In *AIIDE*, 2008.

[5] Sylvain Gelly and Yizao Wang. Exploration exploitation in Go: UCT for Monte-Carlo Go. *Computer*, 2006.

[6] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):pp. 13–30, 1963.

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

[8] Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *ECML*, pages 282–293, 2006.

[9] Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample-variance penalization. In *COLT*, 2009.

[10] Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical Bernstein stopping. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 672–679, New York, NY, USA, 2008. ACM.

[11] Stuart Russell and Eric Wefald. *Do the right thing: studies in limited rationality*. MIT Press, Cambridge, MA, USA, 1991.