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 4, 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 is provided on random problems 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 μ_* and the expected reward μ_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 $\Delta_j = \mu_* - \mu_j$. Strategies that minimize the simple regret are called pure exploration strategies [1]. Principles of rational metareasoning [4] 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 Some Concentration Inequalities

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

Hoeffding's inequality [2]:

$$\Pr(X - \mathbb{E}[X] \ge a) \le \exp(-2na^2) \tag{2}$$

Empirical Bernstein's inequality [3]: [1]

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

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

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

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

3 Upper Bounds on Value of Information

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

$$\Lambda_i^p \le \begin{cases} \Pr(\mathbb{E}[X_i] \le \overline{X}_\beta)\overline{X}_\beta & \textit{if } i = \alpha \\ \Pr(\mathbb{E}[X_i] \ge \overline{X}_\alpha)(1 - \overline{X}_\alpha) & \textit{otherwise} \end{cases} \qquad (5)$$

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

$$\Lambda_i^b \le \begin{cases} \Pr(\overline{X}_i' \le \overline{X}_\beta)\overline{X}_\beta\frac{N}{N+n_i} < \Pr(\overline{X}_i' \le \overline{X}_\beta)\overline{X}_\beta\frac{N}{n_i} & \textit{if } i = \alpha \\ \Pr(\overline{X}_i' \ge \overline{X}_\alpha)(1 - \overline{X}_\alpha)\frac{N}{N+n_i} < \Pr(\overline{X}_i' \ge \overline{X}_\alpha)(1 - \overline{X}_\alpha)\frac{N}{n_i} & \textit{otherwise} \end{cases} \qquad (6)$$

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

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

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

$$\begin{aligned}
\Pr(\mathbb{E}[X_i] \le \overline{X}_\beta | i = \alpha) &\le \exp(-2(\overline{X}_i - \overline{X}_\beta)^2 n_i) \\
\Pr(\mathbb{E}[X_i] \ge \overline{X}_\alpha | i \ne \alpha) &\le \exp(-2(\overline{X}_\alpha - \overline{X}_i)^2 n_i) \\
\Pr(X_i' \le \overline{X}_\beta | i = \alpha) &\le 2\exp\left(-\frac{(\overline{X}_i - \overline{X}_\beta)^2 n_i}{2}\right) \\
\Pr(X_i' \ge \overline{X}_\alpha | i \ne \alpha) &\le 2\exp\left(-\frac{(\overline{X}_\alpha - \overline{X}_i)^2 n_i}{2}\right) \qquad (7)
\end{aligned}$$

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 B).

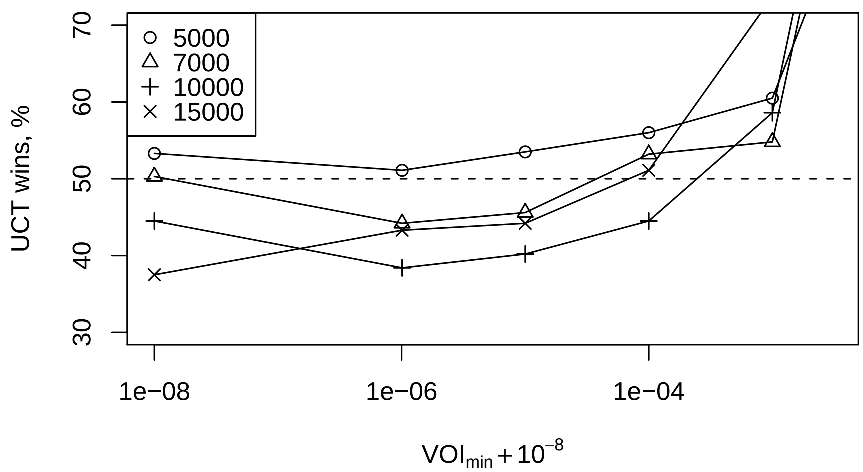

Figure 1: Winning rate: UCT against VCT

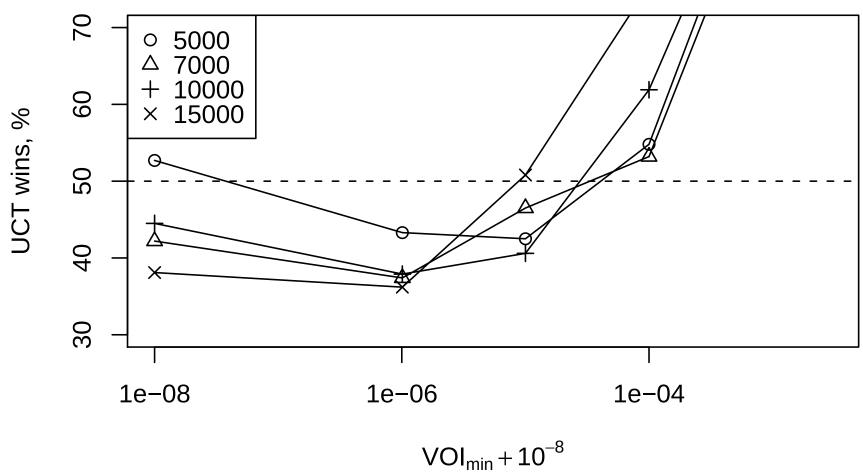

Figure 2: Winning rate: UCT against ECT

4 Empirical Evaluation

5 VOI-based Sampling Control

5.1 Selection Criterion

Following the principles of rational metareasoning, an arm with the highest upper bound $\hat{\Lambda}$ on the perfect value of information should be pulled at each step. This way, arms known to have a low VOI would be pulled less frequently.

5.2 Termination Condition

The upper bounds (**??**, **??**) decrease exponentially with the number of pulls n. When the upper bound of the VOI for all arms becomes lower than a threshold λ, which can be chosen based on resource constraints, the sampling should be stopped, and an arm should be chosen.

[1] see Appendix A for derivation

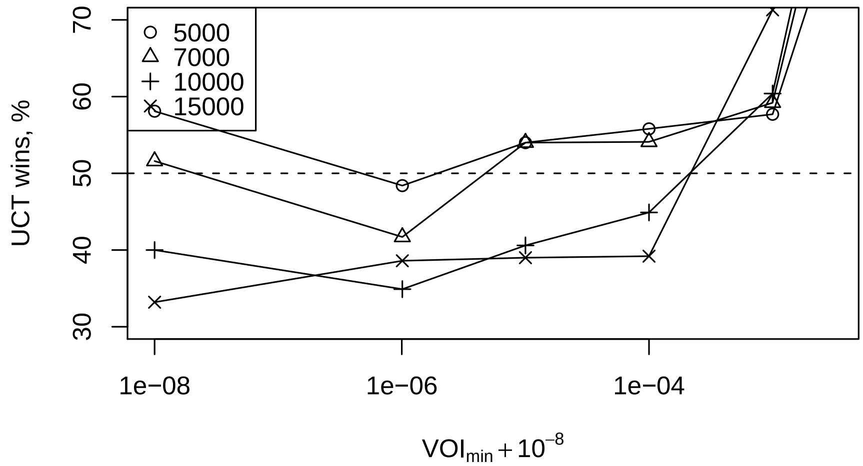

Figure 3: Winning rate: UCT against BCT

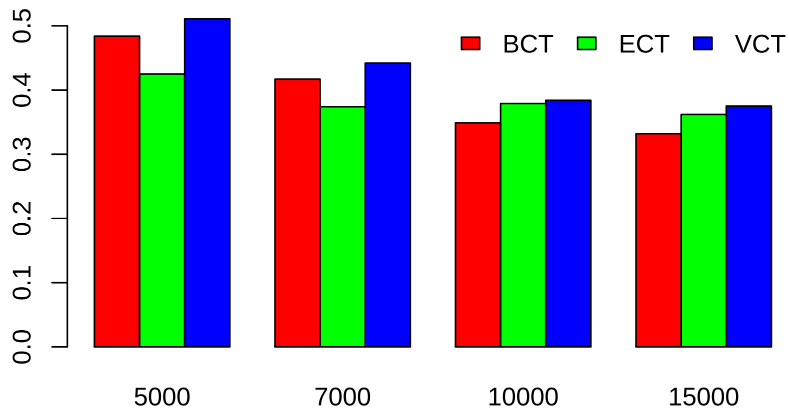

Figure 4: Best winning rate comparison

5.3 Sample Redistribution in Trees

UCT forwards samples to the next search stage when the winner at the current state is known with high confidence. VOI-based termination condition can be used to stop the sampling in the current state early and save the remaining samples for re-use in a later search state.

A Empirical Bernstein Inequality

Theorem 4 in [3] states that

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

Therefore

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

$a = \sqrt{\left(\frac{7\ln 2/\delta}{3(n-1)}\right)^2 + \frac{2\overline{\sigma}_n^2 \ln 2/\delta}{n}} + \frac{7\ln 2/\delta}{3(n-1)}$ 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 $\delta \triangleq \Pr(\mathbb{E}[X] - \overline{X}_n \ge a)$, gives

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

Other derivations, giving slightly different results, are possible.

B Better Hoeffding-Based Bound on Value of Perfect Information

The bound can be supposedly be improved by selecting a midpoint $0 < \gamma < \overline{X}_\beta$ and computing the bound as the sum of two parts:

- $\overline{X}_\beta - \gamma$ multiplied by the probability that $\mu_\alpha \le \overline{X}_\beta$;
- $\overline{X}_\beta$ multiplied by the probability that $\mu_\alpha \le \gamma$.

.

$$V_{\overline{X}=\overline{X}_\alpha} \le (\overline{X}_\beta - \gamma)\exp\left(-2n(\overline{X}_\alpha - \overline{X}_\beta)^2\right) + \overline{X}_\beta \exp\left(-2n(\overline{X}_\alpha - \gamma)^2\right)$$

The minimum of V^* is achieved when $\frac{dV^*}{d\gamma} = 0$, that is, when γ is the root of the following equation:

$$4\overline{X}_\beta n(\overline{X}_\alpha - \gamma) = \exp\left(-2n\left(\frac{\overline{X}_\alpha - \overline{X}_\beta}{\overline{X}_\alpha - \gamma}\right)^2\right)$$

If a root in the interval $0 \le \gamma \le \beta$ exists, then the number of samples is bounded as

$$n \le \frac{1}{4\overline{X}_\beta(\overline{X}_\alpha - \overline{X}_\beta)}$$

by observing that the right-hand side is at most 1 (a negative power), and the left-hand side is at least $4\overline{X}_\beta n(\overline{X}_\alpha - \overline{X}_\beta)$. So, the bound can supposedly be improved for smaller values of n. The improvement is more significant when the current best and second-best sample means are close.

The derivation for the other case (sampling an item that can be better than the current best) is obtained by substitution $1-\overline{X}, 1-\overline{X}_\alpha, 1-\gamma$ instead of $\overline{X}_\alpha, \overline{X}_\beta, \gamma$:

$$V_{\overline{X} \neq \overline{X}_\alpha} \leq (\gamma - \overline{X}_\alpha) \exp\left(-2n(\overline{X}_\alpha - \overline{X})^2\right) + (1 - \overline{X}_\alpha) \exp\left(-2n(\gamma - \overline{X})^2\right)$$

The anticipated influence of the improved VOI estimate would be that the selected item will be a less discovered one and further from the current best or second-best.

A closed-form solution for γ cannot be obtained, but given $\overline{X}_\alpha, \overline{X}_\beta, n$, the value of γ can be efficiently computed. It should be determined empirically whether the improved estimate has justified influence on the performance of the algorithm.

References

[1] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theor. Comput. Sci.*, 412(19):1832–1852, 2011.

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

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

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