The UCB Algorithm

Paper *Finite-time Analysis of the Multiarmed Bandit Problem*
by Auer, Cesa-Bianchi, Fischer, Machine Learning 27, 2002

Presented by Markus Enzenberger.
Go Seminar, University of Alberta.

March 14, 2007

Multiarmed Bandit Problem

- Example for the exploration vs exploitation dilemma
- K independent gambling machines (armed bandits)
- Each machine has an unknown stationary probability distribution for generating the reward
- Observed rewards when playing machine i: $X_{i,1}, X_{i,2}, \ldots$
- Policy A chooses next machine based on previous sequence of plays and rewards

Regret

Regret of policy A

$$\mu^* n - \mu_j \sum_{j=1}^{K} \mathbb{E}[T_j(n)]$$

μ_i	expectation of machine i
μ^*	expectation of optimal machine
T_j	number of times machine j was played

The regret is the expected loss after n plays due to the fact that the policy does not always play the optimal machine.

Lai and Robbins (1985)

Policy for a class of reward distributions (including: normal, Bernoulli, Poisson) with regret asymptotically bounded by logarithm of n:

$$\mathbb{E}[T_j(n)] \leq \frac{\ln(n)}{D(p_j||p^*)} \qquad n \to \infty$$

$D(p_j||p^*)$ Kullback-Leibler divergence between reward densities

- This is the best possible regret
- Policy computes upper confidence index for each machine
- Needs entire sequence of rewards for each machine

In this Paper

- Show policies with logarithmic regret uniformly over time
- Policies are simple and efficient
- Notation: $\Delta_i := \mu^* - \mu_i$

Theorem 1 (UCB1)

Policy with finite-time regret logarithmically bounded for arbitrary sets of reward distributions with bounded support

Deterministic policy: UCB1.
Initialization: Play each machine once.
Loop:

– Play machine j that maximizes $\bar{x}_j + \sqrt{\frac{2 \ln n}{n_j}}$, where $\bar{x}_j$ is the average reward obtained from machine j, n_j is the number of times machine j has been played so far, and n is the overall number of plays done so far.

Theorem 1. *For all $K > 1$, if policy* UCB1 *is run on K machines having arbitrary reward distributions $P_1, \ldots, P_K$ with support in $[0, 1]$, then its expected regret after any number n of plays is at most*

$$\left[8 \sum_{i:\mu_i<\mu^*} \left(\frac{\ln n}{\Delta_i}\right)\right] + \left(1 + \frac{\pi^2}{3}\right)\left(\sum_{j=1}^{K} \Delta_j\right)$$

where $\mu_1, \ldots, \mu_K$ are the expected values of $P_1, \ldots, P_K$.

- $\mathbb{E}[T_j(n)] \leq \frac{8}{\Delta_j^2} \ln(n)$ worse than Lai and Robbins
- $D(p_j||p^*) \geq 2\Delta_j^2$ with best possible constant 2
  $\longrightarrow$ UCB2 brings main constant arbitrarily close to $\frac{1}{2\Delta_j^2}$

Theorem 2 (UCB2)

More complicated version of UCB1 with better constants for bound on regret.

Deterministic policy: UCB2.
Parameters: $0 < \alpha < 1$.
Initialization: Set $r_j = 0$ for $j = 1, \ldots, K$. Play each machine once.
Loop:

1. Select machine j maximizing $\bar{x}_j + a_{n,r_j}$, where $\bar{x}_j$ is the average reward obtained from machine j, a_{n,r_j} is defined in (3), and n is the overall number of plays done so far.
2. Play machine j exactly $\tau(r_j + 1) - \tau(r_j)$ times.
3. Set $r_j \leftarrow r_j + 1$.

Theorem 2 (UCB2)

$$a_{n,r} = \sqrt{\frac{(1+\alpha)\ln(en/\tau(r))}{2\tau(r)}}$$

where

$$\tau(r) = \lceil (1+\alpha)^r \rceil.$$

Theorem 2. *For all $K > 1$, if policy* UCB2 *is run with input $0 < \alpha < 1$ on K machines having arbitrary reward distributions $P_1, \ldots, P_K$ with support in $[0, 1]$, then its expected regret after any number*

$$n \geq \max_{i:\mu_i<\mu^*} \frac{1}{2\Delta_i^2}$$

of plays is at most

$$\sum_{i:\mu_i<\mu^*} \left(\frac{(1+\alpha)(1+4\alpha)\ln\left(2e\Delta_i^2 n\right)}{2\Delta_i} + \frac{c_\alpha}{\Delta_i} \right) \tag{4}$$

where $\mu_1, \ldots, \mu_K$ are the expected values of $P_1, \ldots, P_K$.

- First term brings constant arbitrarily close to $\frac{1}{2\Delta_j^2}$ for small α
- $c_\alpha \to \infty$ as $\alpha \to 0$
- Let $\alpha = \alpha_n$ slowly decrease

Theorem 3 (ϵ_n-GREEDY)

Similar result for ϵ-greedy heuristic.
(ϵ needs to go to 0; constant ϵ has linear regret)

Randomized policy: ε_n-GREEDY.
Parameters: $c > 0$ and $0 < d < 1$.
Initialization: Define the sequence $\varepsilon_n \in (0, 1]$, $n = 1, 2, \ldots$, by

$$\varepsilon_n \stackrel{\text{def}}{=} \min\left\{1, \frac{cK}{d^2 n}\right\}$$

Loop: For each $n = 1, 2, \ldots$

- Let i_n be the machine with the highest current average reward.
- With probability $1 - \varepsilon_n$ play i_n and with probability ε_n play a random arm.

Theorem 3. *For all $K > 1$ and for all reward distributions $P_1, \ldots, P_K$ with support in $[0, 1]$, if policy ε_n-*GREEDY *is run with input parameter*

$$0 < d \leq \min_{i:\mu_i < \mu^*} \Delta_i,$$

*then the probability that after any number $n \geq cK/d$ of plays ε_n-*GREEDY *chooses a suboptimal machine j is at most*

$$\frac{c}{d^2 n} + 2\left(\frac{c}{d^2} \ln \frac{(n-1)d^2 e^{1/2}}{cK}\right)\left(\frac{cK}{(n-1)d^2 e^{1/2}}\right)^{c/(5d^2)} + \frac{4e}{d^2}\left(\frac{cK}{(n-1)d^2 e^{1/2}}\right)^{c/2}.$$

- For c large enough, the bound is of order $c/(d^2 n) + o(1/n)$ $\rightarrow$ logarithmic bound on regret
- Bound on instantaneous regret
- Need to know lower bound d on expectation between best and second-best machine

Theorem 4 (UCB1-NORMAL)

Indexed based policy with logarithmically bounded finite-time regret for normally distributed reward distributions with unknown mean and variance.

Theorem 4 (UCB1-NORMAL)

Deterministic policy: UCB1-NORMAL.
Loop: For each $n = 1, 2, \ldots$

- If there is a machine which has been played less then $\lceil 8 \log n \rceil$ times then play this machine.
- Otherwise play machine j that maximizes

$$\bar{x}_j + \sqrt{16 \cdot \frac{q_j - n_j \bar{x}_j^2}{n_j - 1} \cdot \frac{\ln(n-1)}{n_j}}$$

  where $\bar{x}_j$ is the average reward obtained from machine j, q_j is the sum of squared rewards obtained from machine j, and n_j is the number of times machine j has been played so far.
- Update $\bar{x}_j$ and q_j with the obtained reward x_j.

Theorem 4. *For all $K > 1$, if policy* UCB1-NORMAL *is run on K machines having normal reward distributions $P_1, \ldots, P_K$, then its expected regret after any number n of plays is at most*

$$256(\log n)\left(\sum_{i:\mu_i<\mu^*} \frac{\sigma_i^2}{\Delta_i}\right) + \left(1 + \frac{\pi^2}{2} + 8\log n\right)\left(\sum_{j=1}^{K} \Delta_j\right)$$

where $\mu_1, \ldots, \mu_K$ and $\sigma_1^2, \ldots, \sigma_K^2$ are the means and variances of the distributions $P_1, \ldots, P_K$.

- Like UCB1, but since kind of distribution is known, sample variance is used to estimate variance of distribution
- Proof depends on bounds for tails of χ^2 and Student distribution, which were only verified numerically

Independence Assumptions

- Theorem 1–3 also hold for rewards that are not independent across machines: $X_{i,s}$ and $X_{j,t}$ might be dependent for any s, t and $i \neq j$
- The rewards of a single machine do not need to be independent and identically-distributed. Weaker assumption: $\mathbb{E}[X_{i,t} | X_{i,1}, \ldots, X_{i,t-1}] = \mu_i$ for all $1 \leq t \leq n$

UCB1-TUNED

Fined-tuned version of UCB taking the measured variance into account (no proven regret bounds)
Upper confidence bound on variance of machine j

$$V_j(s) \stackrel{\text{def}}{=} \left(\frac{1}{s} \sum_{\tau=1}^{s} X_{j,\tau}^2 \right) - \bar{X}_{j,s}^2 + \sqrt{\frac{2 \ln t}{s}}$$

Replace upper confidence bound in UCB1 by

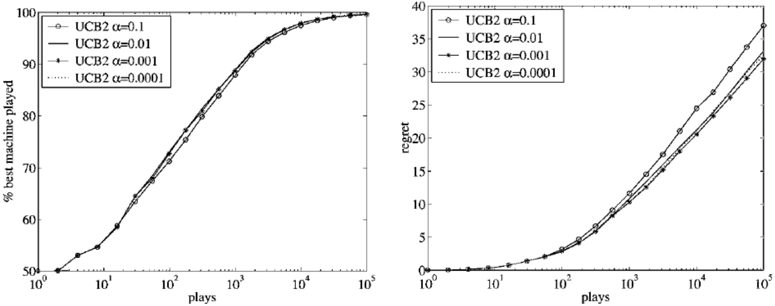
$$\sqrt{\frac{\ln n}{n_j} \min\{1/4, V_j(n_j)\}}$$

1/4 is upper bound on variance of a Bernoulli random variable

Distributions

	1	2	3	4	5	6	7	8	9	10
1	0.9	0.6								
2	0.9	0.8								
3	0.55	0.45								
11	0.9	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
12	0.9	0.8	0.8	0.8	0.7	0.7	0.7	0.6	0.6	0.6
13	0.9	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
14	0.55	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45

Best value for α



- Relatively insensitive, as long as α is small
- Use fixed $\alpha = 0.001$

Summary of Results

- An optimally tuned ϵ_n-GREEDY performs almost always best
- Performance of not well-tuned ϵ_n-GREEDY degrades rapidly
- In most cases UCB1-TUNED performs comparably to a well-tuned ϵ_n-GREEDY
- UCB1-TUNED not sensitive to the variances of the machines
- UCB2 performs similar to UCB1-TUNED, but always slightly worse

Comparison on Distribution 11

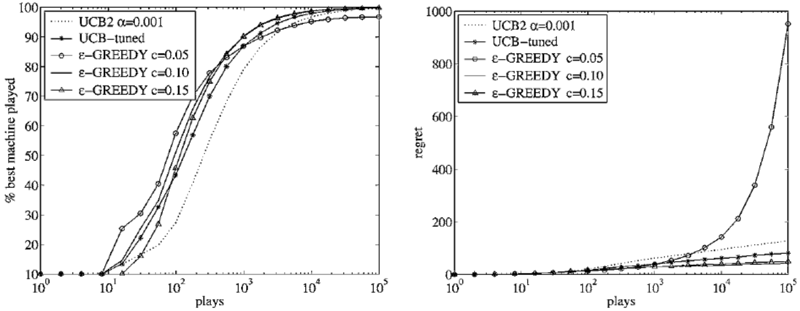


Figure 9. Comparison on distribution 11 (10 machines with parameters 0.9, 0.6, . . . , 0.6).

Conclusions

- Simple, efficient policies for the bandit problem on any set of reward distributions with known bounded support with uniform logarithmic regret
- Based on upper confidence bounds (with exception of ϵ_n-GREEDY)
- Robust with respect to the introduction of moderate dependencies

- Many extensions of this work are possible
- Generalize to non-stationary problems
- Based on Gittins allocation indices (needs preliminary knowledge or learning of the indices)