Stochastic Search and Optimisation
Multi-armed Bandits: Heuristics

There are a variety of heuristic approaches to the Multi-armed Bandit problem. Most of them implicitly make use of the Gittins Index Theorem, in that they use indices based on the performance of individual bandits.

In what follows let $X(t)$ be the return from the t-th decision, where

$$X(t) \sim \text{Bernoulli}(\theta_{i(t)}).$$

Uniform Bandit Algorithm

Even-Dar, Mannor & Mansour (2002). PAC bounds for multi-armed bandit and Markov decision processes, *Computational Learning Theory*.

The Uniform Bandit Algorithm says pull each arm w times, then choose the bandit with the best average award.

Theorem Suppose that we have a n bandits with mean returns θ_i, $i = 1, \ldots, n$. Then for any $\delta, \epsilon > 0$ we can find a w such that

$$\mathbb{P}(\theta^* - \max_i \hat{\theta}_i > \epsilon) < \delta.$$

Moreover we can take $w = \epsilon^{-2} \log(n/\delta)$.

Here $\theta^* = \max_i \theta_i$ and $\hat{\theta}_i$ is the sample average of the w returns from bandit i.

In the machine learning literature they say that the Uniform Bandit Algorithm is ‘Probably Approximately Correct’ (PAC). The proof is an application of Chernoff’s bound.

Chernoff Bound

Let X_i be i.i.d. random variables with mean μ, absolutely bounded by K, then

$$\mathbb{P}\left(\left|\mu - \frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq \epsilon\right) \leq \exp(-(\epsilon/K)^2 n).$$

Cumulative Regret

Cumulative regret is used to measure the performance of a heuristic algorithm compared to the optimum.

Let $\mathbf{i} = (i(1), \ldots, i(N))$ be our sequence of decisions, then the *expected cumulative regret* at time N is

$$\mathcal{R}(N) = N\theta^* - \sum_{i=1}^{N} \mathbb{E}\theta_{i(t)}.$$

Here we are thinking of each $i(t)$ as a random variable, depending on $i(1), \ldots, i(t-1)$ and outcomes $X_{i(1)}, \ldots, X_{i(t-1)}$.

Lower Bound

Lai & Robbins (1985) Asymptotically efficient adaptive allocation rules, *Advances in Applied Mathematics*.

Let $\Delta_i = \theta^* - \theta_i$ then

$$\liminf_{N\to\infty} \frac{\mathcal{R}(N)}{\log N} \geq \sum_{i:\Delta_i>0} \frac{\Delta_i}{\theta_i \log \frac{\theta_i}{\theta^*} + (1-\theta_i)\log \frac{1-\theta_i}{1-\theta^*}}$$

That is, the best decision rule will have expected cumulative regret at least $O(\log N)$.

There are a number of decision rules that achieve this asymptotic rate.

ϵ-greedy Algorithm

Auer, Cesa-Bianchi & Fisher (2002) Finite-time analysis of the multiarmed bandit problem, *Machine Learning*.

For a sequence $\{\epsilon_t\}$ the ϵ-greedy algorithm says that at time t with probability $1 - \epsilon_t$ choose the bandit with the highest $\hat{\theta}_i$, and with probability ϵ_t choose a bandit uniformly at random.

Theorem Put $\Delta = \min_{\Delta_i > 0} \Delta_i$ and

$$\epsilon_t = \min\left\{\frac{6n}{\Delta^2 t}, 1\right\}.$$

Then for some constant c

$$\mathcal{R}(N) \leq \left(c \sum_{i:\Delta_i > 0} \frac{\Delta_i}{\Delta^2}\right) \log N.$$

Upper Confidence Bound Algorithm (UCB)

Auer, Cesa-Bianchi & Fisher (2002) Finite-time analysis of the multiarmed bandit problem, *Machine Learning*.

Let $T_i(t)$ be the number of times bandit i has been played up to and including time t, and let $\hat{\theta}_i(t)$ be the sample average of the returns from bandit i. The the UCB1 strategy is

$$i(t) = \arg\max_i \left(\hat{\theta}_i(t-1) + \sqrt{\frac{2\log t}{T_i(t-1)}} \right)$$

Theorem For some c

$$\mathcal{R}(N) \leq \sum_{i:\Delta_i > 0} \min\left\{ \frac{c}{\Delta_i} \log N, N\Delta_i \right\}$$

The proof is an application of Hoeffding's Inequality.

Hoeffding's Inequality

Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample of random variables taking values in $[0, 1]$. Then for any $\epsilon > 0$,

$$\mathbb{P}\left(\mathbb{E}X_1 \leq \bar{X} + \sqrt{\frac{-\log \epsilon}{2n}}\right) \geq 1 - \epsilon.$$

Bayesian

Agrawal & Goyal (2012) Analysis of Thompson sampling for the multi-armed bandit problem

Recall that the state of (information about) bandit i at time t is modeled as $\Theta_i(t) \sim \text{beta}(\alpha_i(t), \beta_i(t))$.[1]

Thompson sampling uses a *random* index for bandit i at time t, distributed as $\Theta_i(t)$. That is, we generate independent beta random variables for each bandit, and choose the bandit with the largest one.

Theorem For Thompson sampling the cumulative regret can be bounded by

$$\mathcal{R}(N) \leq O\left(\left(\sum_{i:\Delta_i>0} \Delta_i^{-2}\right)^2 \log N\right).$$

[1] So $T_i(t) = \alpha_i(t) + \beta_i(t) - 2$.

Finite sample performance

`mab_regret.r`