IE5504 - Systems Modeling and Advanced Simulation
Group Project Plan

supervisor: Prof. Lee Loo Hay

Group Member:

Li Qingxuan, Li Yue, Liu Weizhi,
Wong Manyu, Yuan yuhe

LIU Weizhi *

Department of Industrial & Systems Engineering
National University of Singapore

September 15, 2014

*weizhiliu2009@gmail.com

1 Introduction

We intend to investigate the multi-armed bandit problem which can be formulated as below:
Maximize your total reward R with at most n trials given K slot machines (bandit) whose reward distribution is quite different.

This problem is very similiar to what OCBA are supposed to solve. Treat the slot machines as possible designs and the amount of tries as computing resources. **Consequently, we are very interesed in comparing the performance of OCBA and traditional methods to solve multi-armed bandit problem and figure out the reason why they might perform differently.**

To ease the problem without loss of generality for multi-armed bandit problem, we assume the reward of slot machine k obeys Bernoulli distribution p_k. Therefore, the conjugate distribution is Beta distribution.

2 Strategy

We propose two preliminary approaches to solve the problem described as section 1. More detailed algorithms will be developed further.

2.1 Bayesian Bandit

The process of Bayesian bandit algorithm is as follow:

1. Initialize the Bernoulli success rate for all bandits and generate a success/failure (1/0) sequence with length n for each bandit. (Pseudo Random)
2. Calculate the APCS (Approximate Probability of Correct Selection [1] or Stochastic Ordering) of all bandits based on their prior distribution.
3. Find the bandit B with highest APCS (if not unique, then uniformly randomly choose a bandit)
4. Observe the reward of bandit B from the given sequence
5. Update the posterior distribution of bandit B
6. Stop if total number of trials run out otherwise return to step 2

2.2 Optimal Computing Budget Allocation

The process of Optimal Computing Budget Allocation is as follow:

1. Initialize the Bernoulli success rate for all bandits and generate a success/failure (1/0) sequence with length n for each bandit.
2. For each bandit, initially try N_0 times and calculate their corresponding mean and variance
3. Find the bandit B with highest mean as the best bandit in the initial period.
4. Allocate the number of trials to different bandit according to the allocation rule [2] in order to maximize the probability of correct selection.

[1] please refer to **Stochastic Simulation Optimization - An Optimal Computing Budget Allocation**, P37
[2] please refer to the allocation rule at **Stochastic Simulation Optimization - An Optimal Computing Budget Allocation**, P46

5. Try the bandit for respective allocated nums and then calculated the mean and variance for each bandit.

6. Stop if total number of trials run out otherwise return to step 3

3 Evaluation

We adopt the Pseudo Regret as a performance measure for the algorithm to find the optimal bandit.

$$Pseudo\ Regret = \max_i \sum_{t=1}^{n} R_{i,t} - \sum_{t=1}^{n} R_{I_t,t}$$

where I_t is the bandit selected at time t.

4 Extension

If time is enough, some general scenarios could be considered

- Find more strategies to solve the multi-armed bandit problem and make comparisions.
- Integration of markov decision process to solve multi-armed bandit problem.
- Modify the reward distribution to more relastic distributions.