Multi-armed Bandits with externalities

VVN

2019

The model for two arms

Reward structure :

- We have a set of two arms and a corresponding set of two user types (each with populations $Z_1(t)$ and $Z_2(t)$).
- If the user arriving at time t is of type $Y(t)$ and is shown the arm $X(t)$, then the reward $R(t)$ is chosen from a Bernoulli distribution with mean $b_{X_t Y_t}$.
- Therefore we have a matrix $M = [b_{ij}]_{2\times 2}$ of the mean rewards where in each row i ,the element b_{ii} has the highest value.

The model for two arms

Updating the population : After reward $R(t)$ is obtained, the $Z_i(t)$'s are updated as follows -

- $Z_{X_t}(t+1) = Z_{X_t}(t) + R(t)$
- $Z_{-X_t}(t+1) = Z_{-X_t}(t) + (1 - R(t))$

where Z_{-X_t} is the population of the arm that was not recommended.
Notation : We use z_i to denote the proportion of the respective balls. For the policies, we use p to denote probability of choosing arm 1 given a user of type 1 appears and q the probability of choosing arm 2 given a user type 2 appears.

Solution via ODE for a probabilistic policy

Differential equation :

$$\frac{dz}{dt} = \frac{c + az}{A + t}$$

where $z = z_1(t)$,$A = z_1(0) + z_2(0)$(initial number of balls) and c, a are given by :

$$c = (1 - q)b_{10} + q(1 - b_{11})$$

$$a = -(1 + c - pb_{00} - (1 - p)(1 - b_{01}))$$

Solution via ODE for a probabilistic policy

- **Solution to the differential equation :**

$$z(t) = -\frac{c}{a} + (z(0) + \frac{c}{a})(1 + \frac{t}{A})^a$$

Here a is always negative so the proportion of balls of type 1 approaches $(-c/a)$ asymptotically. We can now maximize this term to find the optimum policy (p, q).

- The optimum policy (that maximizes long term population of type 1) can be found by minimizing δ such that $-c/a = 1 - \delta$. Therefore:

$$\delta = \frac{1 - pb_{00} - (1-p)(1-b_{01})}{1 + (1-q)b_{10} + q(1-b_{11}) - pb_{00} - (1-p)(1-b_{01})}$$

is to be minimized w.r.t (p, q).

Solution via ODE for a probabilistic policy

Optimum policy : The optimum policy that minimizes the δ mentioned above turns out to be :

$$p = I_{\{b_{00} > 1 - b_{01}\}}$$

$$q = I_{\{b_{10} < 1 - b_{11}\}}$$

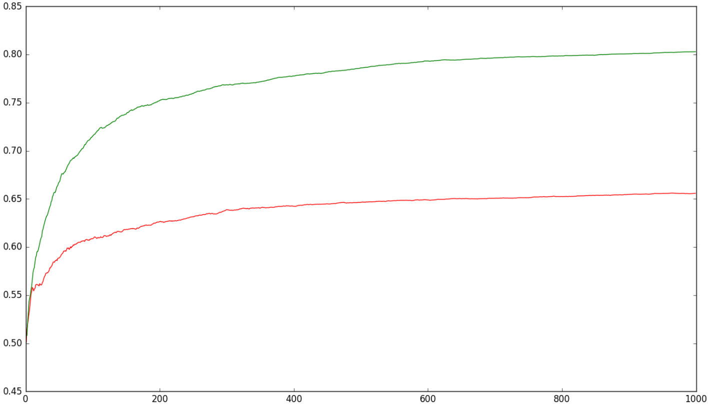
(where I_e is the indicator function of event e)

Optimal vs Greedy policy

- The policy described in the previous slide optimizes the proportion of users of a given type.
- If we decide, instead, to optimize the reward accumulated over time, we may get a different policy.
- The greedy policy (where we offer each user type the arm they prefer) gives better mean reward (given the user type) than the optimal policy in the previous slide.
- However, there is a tradeoff between the two policies because the greedy policy has a sub optimal proportion of more rewarding users.

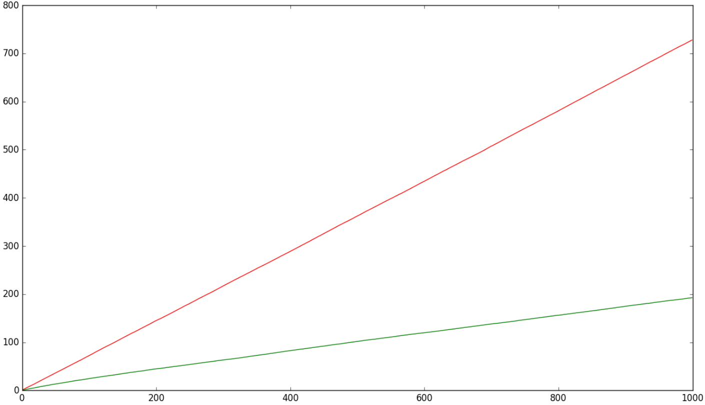
Optimal vs Greedy policy

Figure 1: Proportion of type 0 users vs time (Red : Greedy , Green : Optimal)



Optimal vs Greedy policy

Figure 2: Cumulative reward vs time (Red : Greedy, Green : Optimal)



Optimal vs Greedy policy

Case 1 : We first consider the case : $b_{00} < 1 - b_{01}$ and $b_{11} > 1 - b_{10}$ (In which case the optimal policy is to show the arm of the opposite type).

- Say we try to find a policy for which the equilibrium reward attained per unit time is maximum. That is, we maximize :

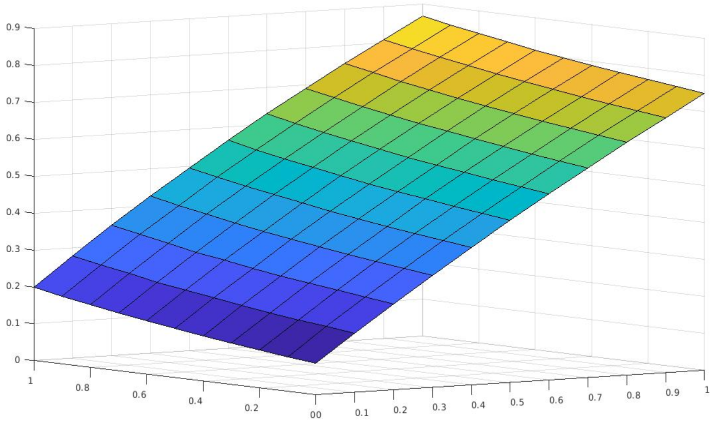
$$R = z(T)(pb_{00} + (1-p)(1-b_{01})) + (1-z(T))(qb_{11} + (1-q)(1-b_{10}))$$

- Optimizing this over (p, q), we get that the greedy policy almost always gives us the maximum reward (taking $T = 100, 1000, 10000$).
- The reward obtained vs (p, q) is plotted in the following slide for a particular Bernoulli matrix satisfying case 1.
- The greedy policy is sub-optimal only in cases where b_{10} is close* to b_{11} in value (in which case the policy $p = 1, q = 0$ becomes more rewarding, but only slightly*).

*This can be made more precise by explicitly differentiating R above and seeing the signs of the derivatives

Optimal vs Greedy policy

Figure 3: XY - plane : Values of (p,q), Z-axis : Value of equilibrium reward



Optimal vs Greedy policy

Case 2 and 3: In the case : $b_{00} > 1 - b_{01}$ and $b_{11} > 1 - b_{10}$ (for which the optimal policy is $p = 1, q = 0$) or for the case $b_{00} < 1 - b_{01}$ and $b_{11} < 1 - b_{10}$(for which the optimal policy is $p = 0, q = 1$), we obtain the same results as the previous case.
Case 4 : In the case : $b_{00} > 1 - b_{01}$ and $b_{11} < 1 - b_{10}$, the optimal and the greedy policy coincide and hence the optimal policy also gives us the maximum reward.

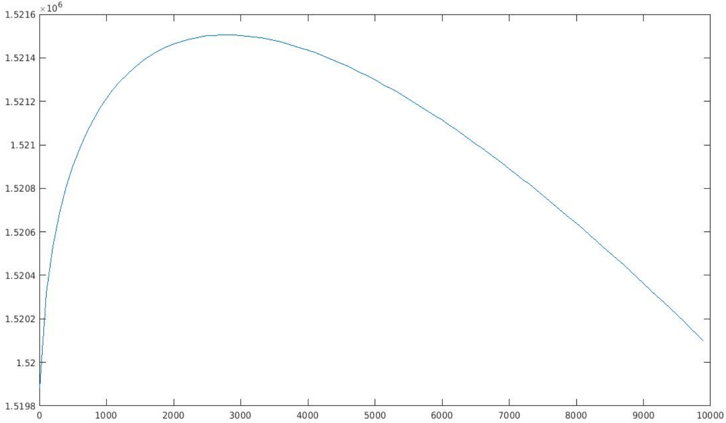
A Mixed Policy

An alternative policy might be to play the optimal policy for some time T_0 and then use the greedy policy till some deadline T.
In this case, we can put either of the following constraints as our aim, and optimize over varying T_0:

- **P1 :** Maximize the reward accrued at deadline given that the population of type 0 must be greater than a threshold.
- **P2 :** Maximize population at deadline given reward greater than a threshold.

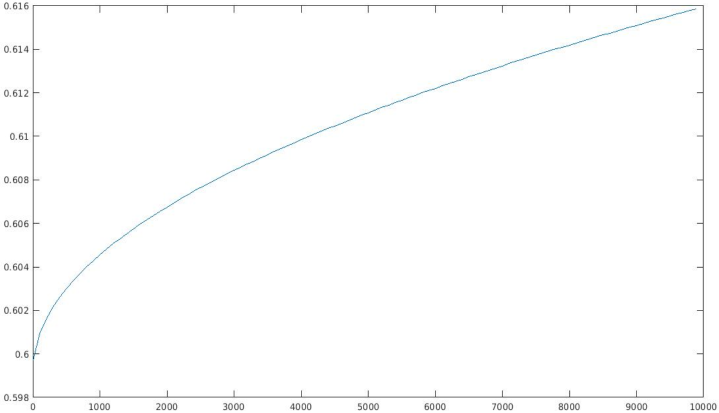
Mixed Policy Results for varying T_0

Figure 4: Cumulative reward at deadline Vs T_0 (deadline $T = 2 \times 10^6$)



Mixed Policy Results for varying T_0

Figure 5: Proportion of type 0 at deadline Vs T_0 (deadline $T = 2 \times 10^6$)



Mixed Policy Results for varying T_0

Further observations :

- The graph of proportion of type 0 users vs T_0 is always strictly increasing.
- The graph of cumulative reward vs T_0 is strictly decreasing after reaching the maxima at atmost one point.
- For small values of deadline T, the reward is strictly decreasing for all values of T_0.

Mixed Policy Results for varying T_0

Finding the optimum T_0 :

- Given the deadline, plot the two graphs in the preceding slides using the deadline and reward matrix.
- To solve P1, find the value T_0^* after which the value of the population exceeds the threshold in Fig 5. Then choose $T_0^{opt} > T_0^*$ at which the graph in Fig 4 attains maxima.
- To solve P2, follow similar procedure, except that the threshold is now in Fig 4.

Link to all codes

Github :
https://github.com/vivien98/MultiArmedBandit-simulations