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 $B = [b_{ij}]_{2\times 2}$ 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 :

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

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

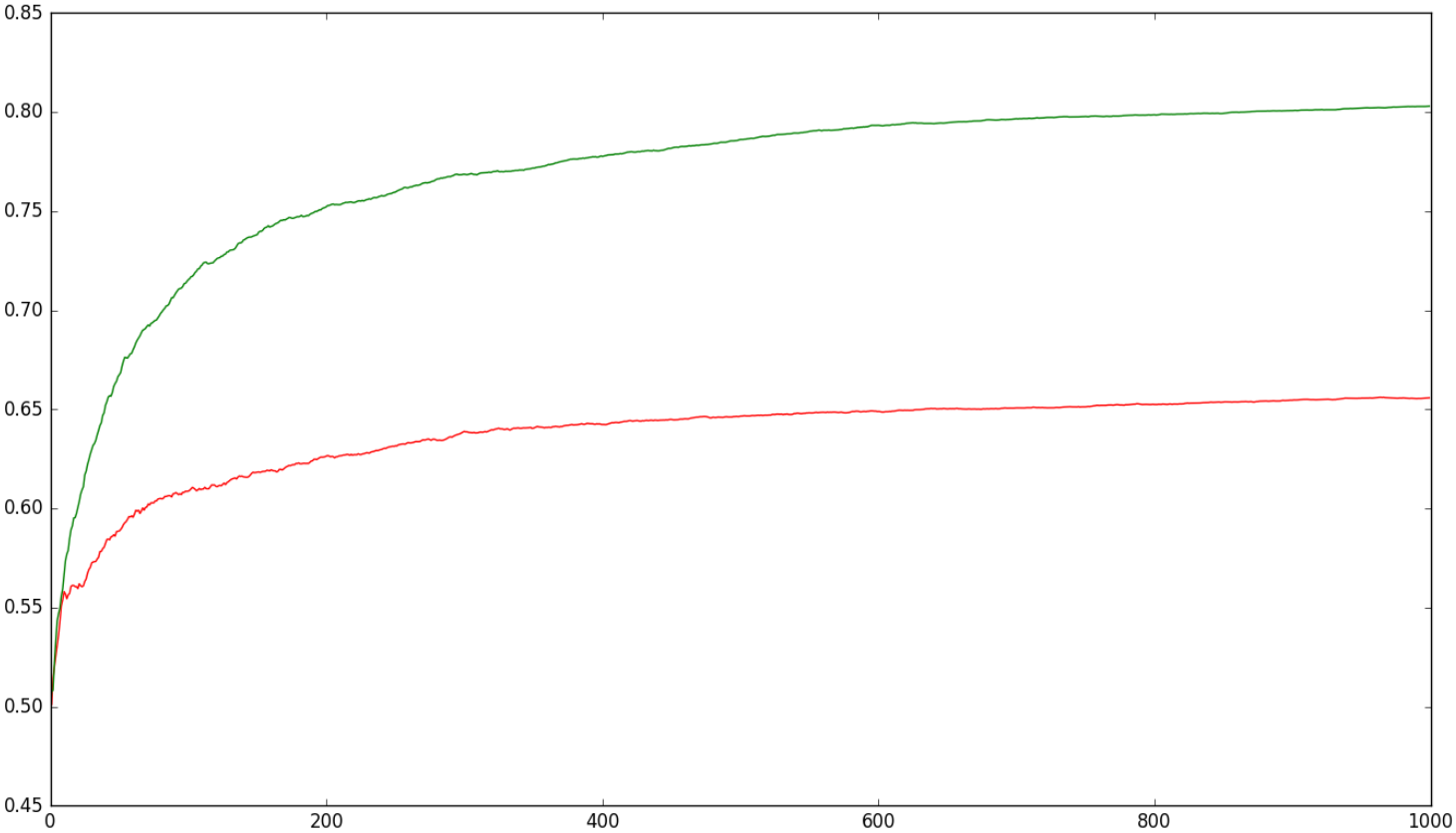
(where I_e is the indicator function of event e)

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.

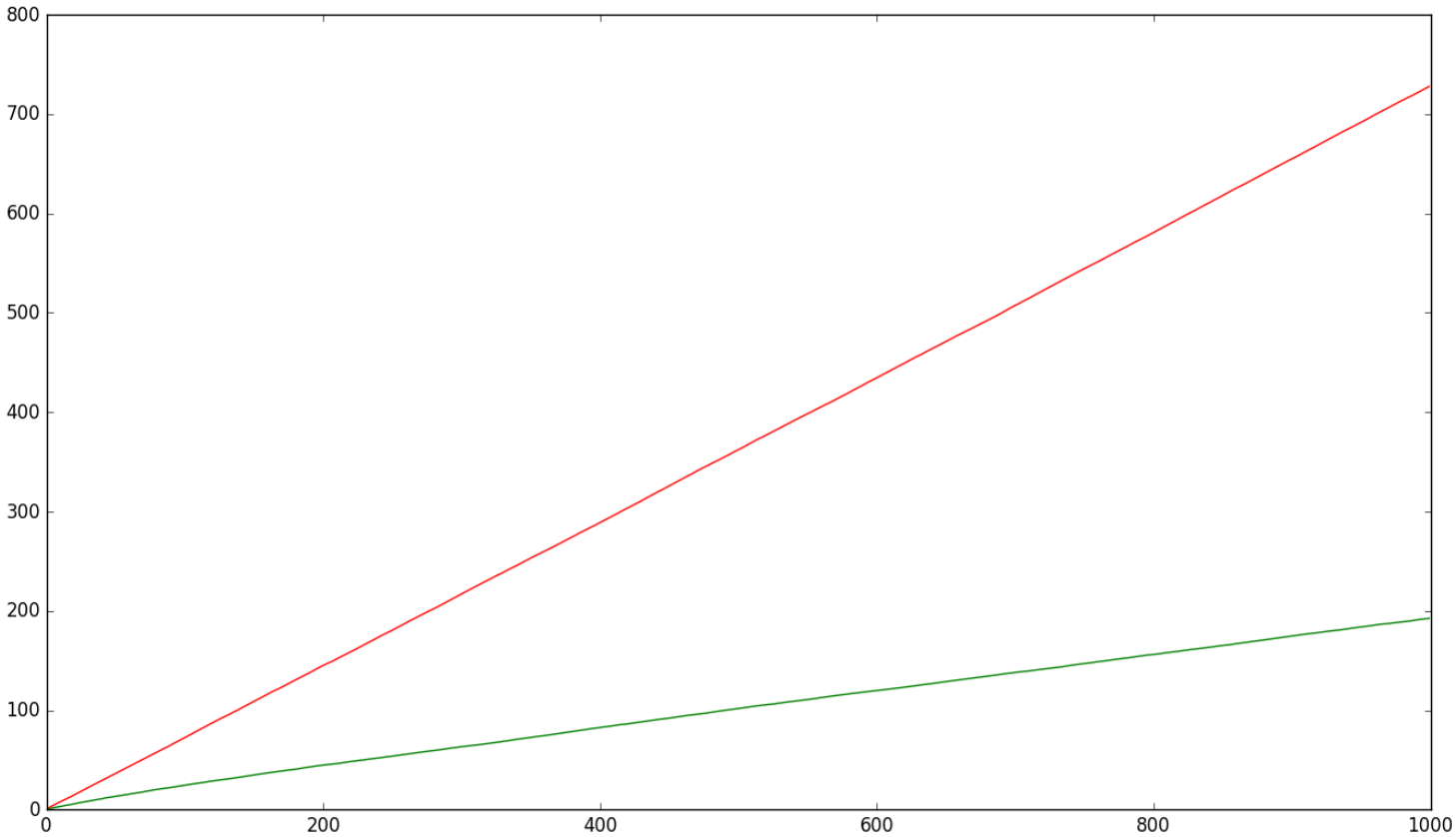
Optimal vs Greedy policy

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 :

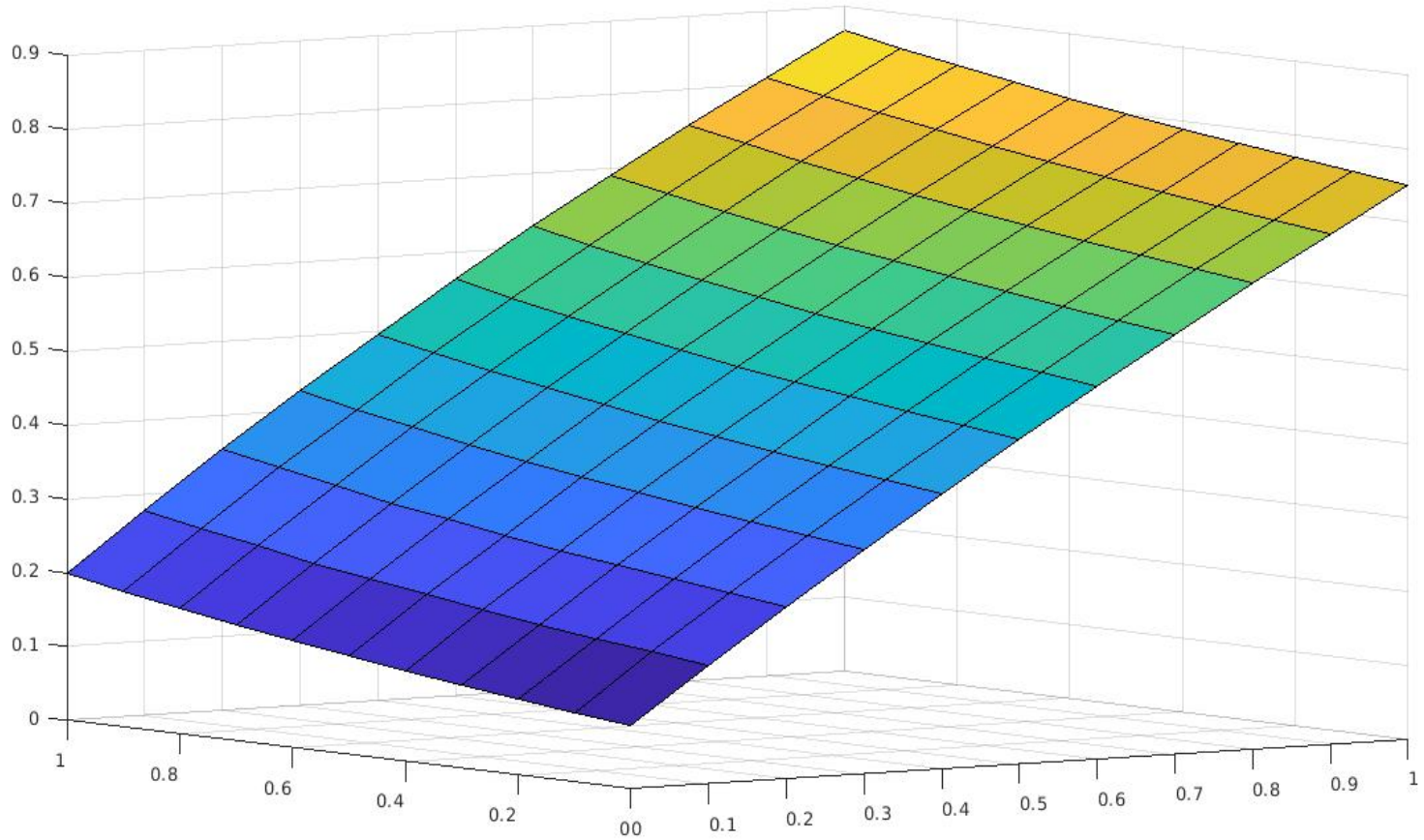
$$R = z(T)(pb_{00} + (1-p)(1-b_{01})) + (1-z(T))(qb_{11} + (1-q)(1-b_{10}))$$

- 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.

Notation

$$r_1 = pb_{00} + (1-p)b_{01}$$

$$r_2 = qb_{11} + (1-q)b_{10}$$

$$d_1 = p(1-b_{00}) + (1-p)b_{01}$$

$$d_2 = q(1-b_{11}) + (1-q)b_{10}$$

$$R_T(p,q) = T\left(\frac{r_1 d_1 + r_2 d_2}{d_1 + d_2}\right) + (r_1 - r_2)\left(z_0 - \frac{d_2}{d_1 + d_2}\right)\left(\sum_{t=1}^{T}\left(1 + \frac{t}{A}\right)^{-(d_1+d_2)}\right)$$

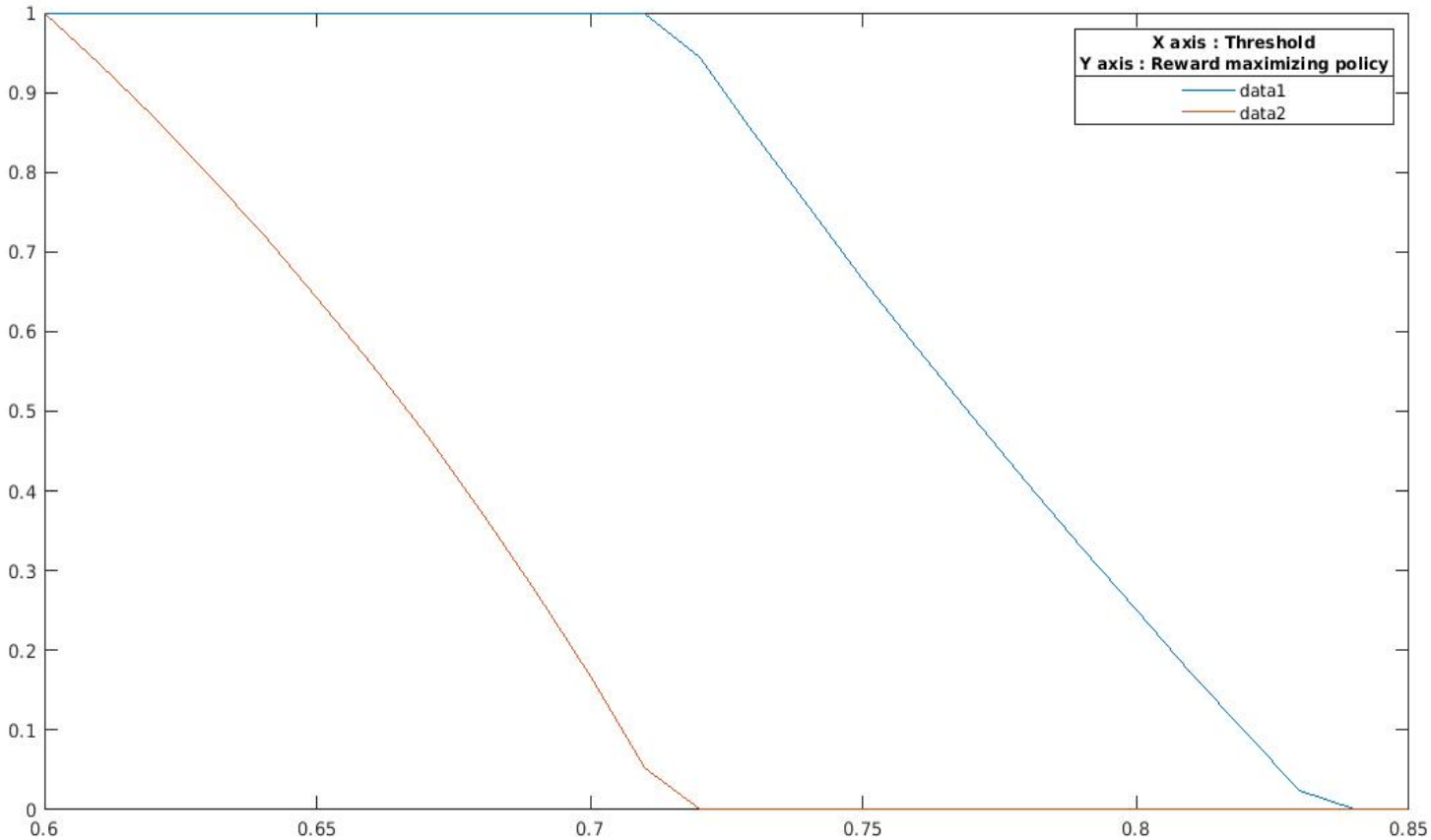
$$z_T = \frac{d_2}{d_1 + d_2} + \left(z_0 - \frac{d_2}{d_1 + d_2}\right)\left(1 + \frac{T}{A}\right)^{-(d_1+d_2)}$$

P1 and P2

- **Problem P1 :**
  Maximize $R_T(p, q)$
  given constraints $z_T(p, q) > z^*$ and $1 > p, q > 0$.
- **Problem P2 :**
  Maximize $z_T(p, q)$
  given constraints $R_T(p, q) > R^*$ and $1 > p, q > 0$.

Solution for problem P1

Figure 4: Optimal Policy solution of problem P1 vs values of z^*



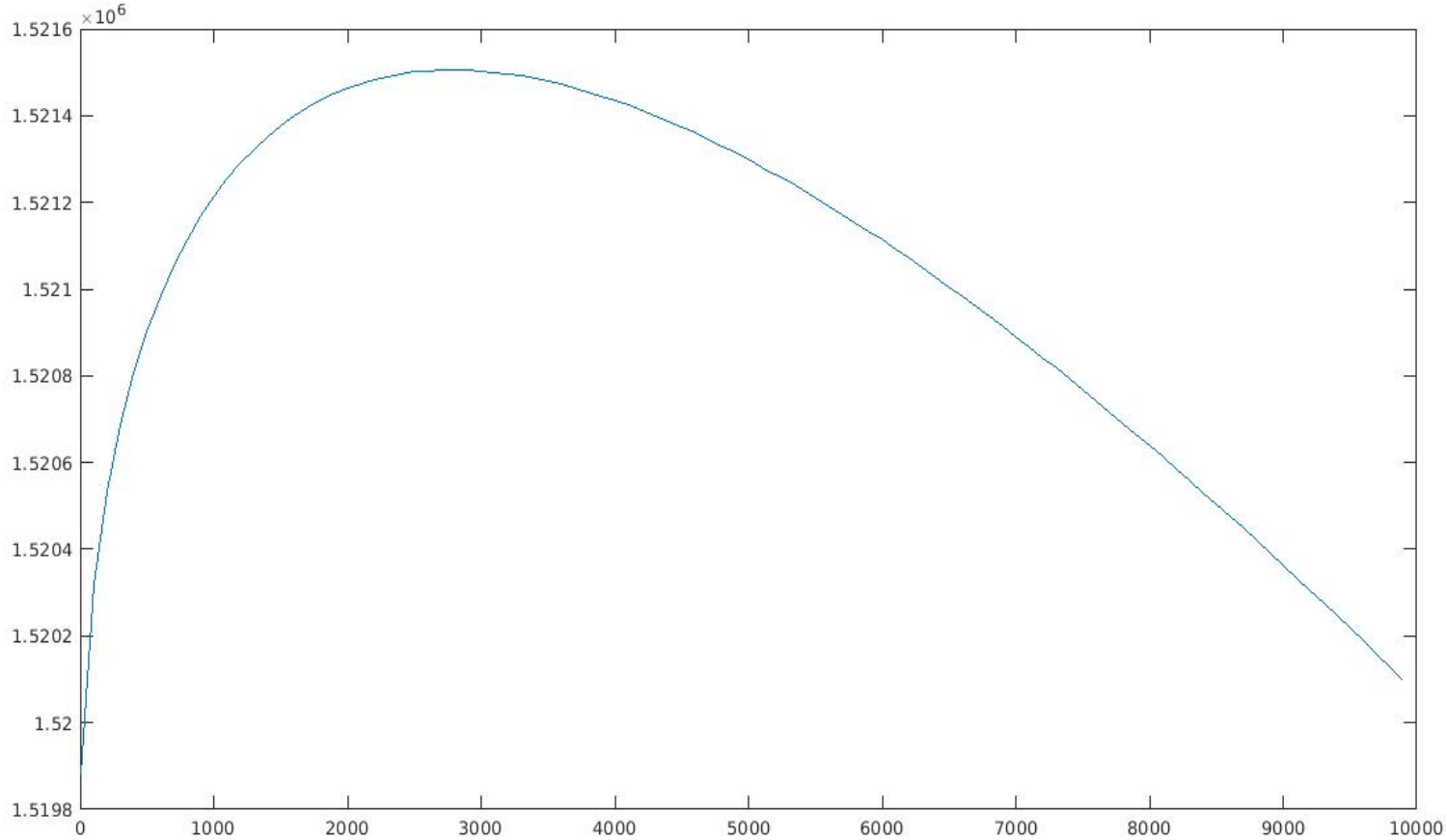
A Mixed Policy

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.

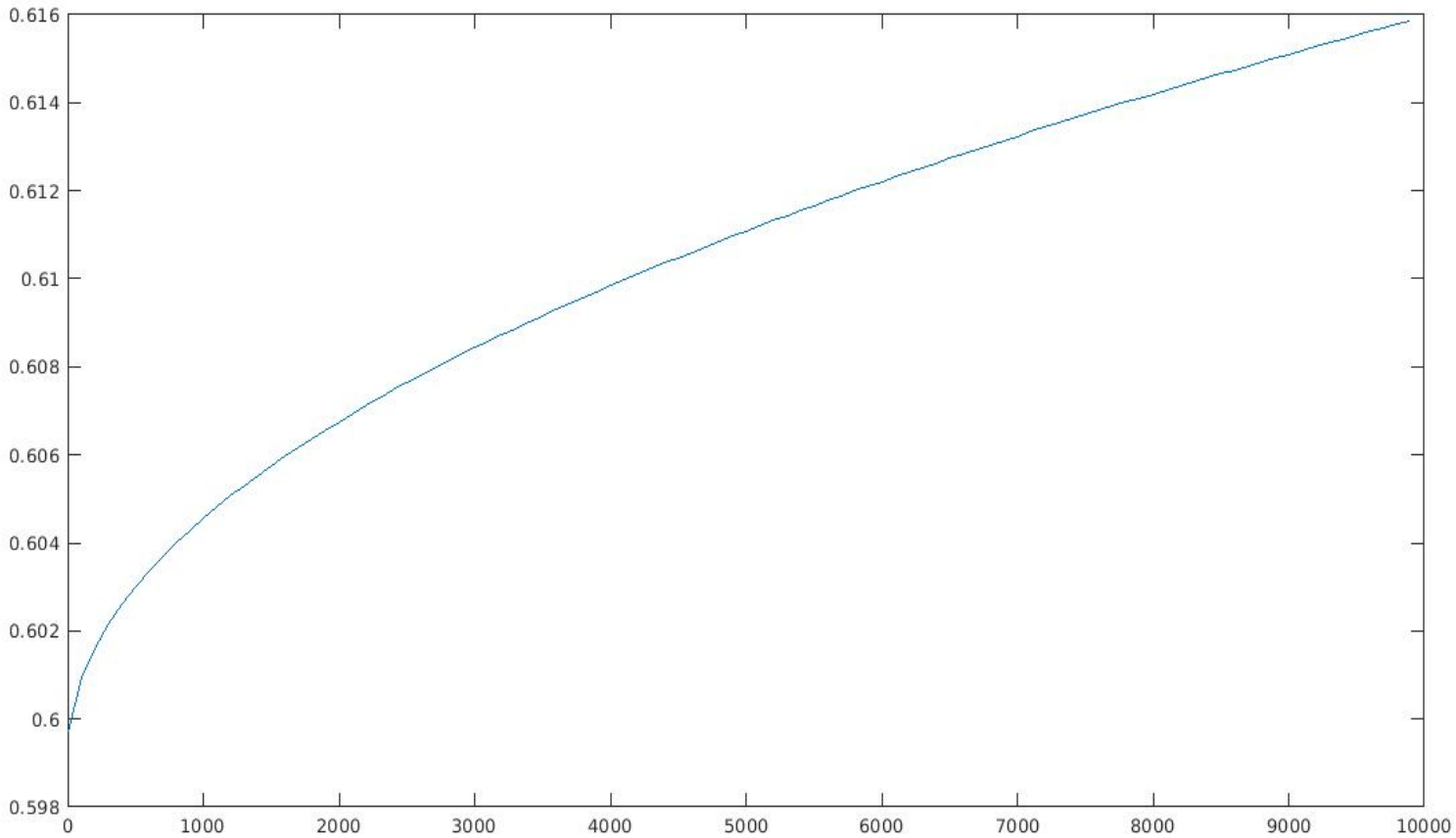
Mixed Policy Results for varying T_0

Figure 5: Cumulative reward at deadline Vs T_0 (deadline $T = 2 \times 10^6$)



Mixed Policy Results for varying T_0

Figure 6: 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. **Reason :** If $T_0^1 > T_0^2$ then

$$(1 + \frac{T_0^1}{A})^a > (1 + \frac{T_0^2}{A})^a$$

- 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.
- Manipulating the population first and then exploiting to get reward is the best policy to maximize cumulative reward up to some deadline T. **Reason :** As A increases, the time t' required for the population to reach some threshold under the optimal policy increases. This increases the time spent not being greedy, which in turn decreases the reward.

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.

Solution via Voter model

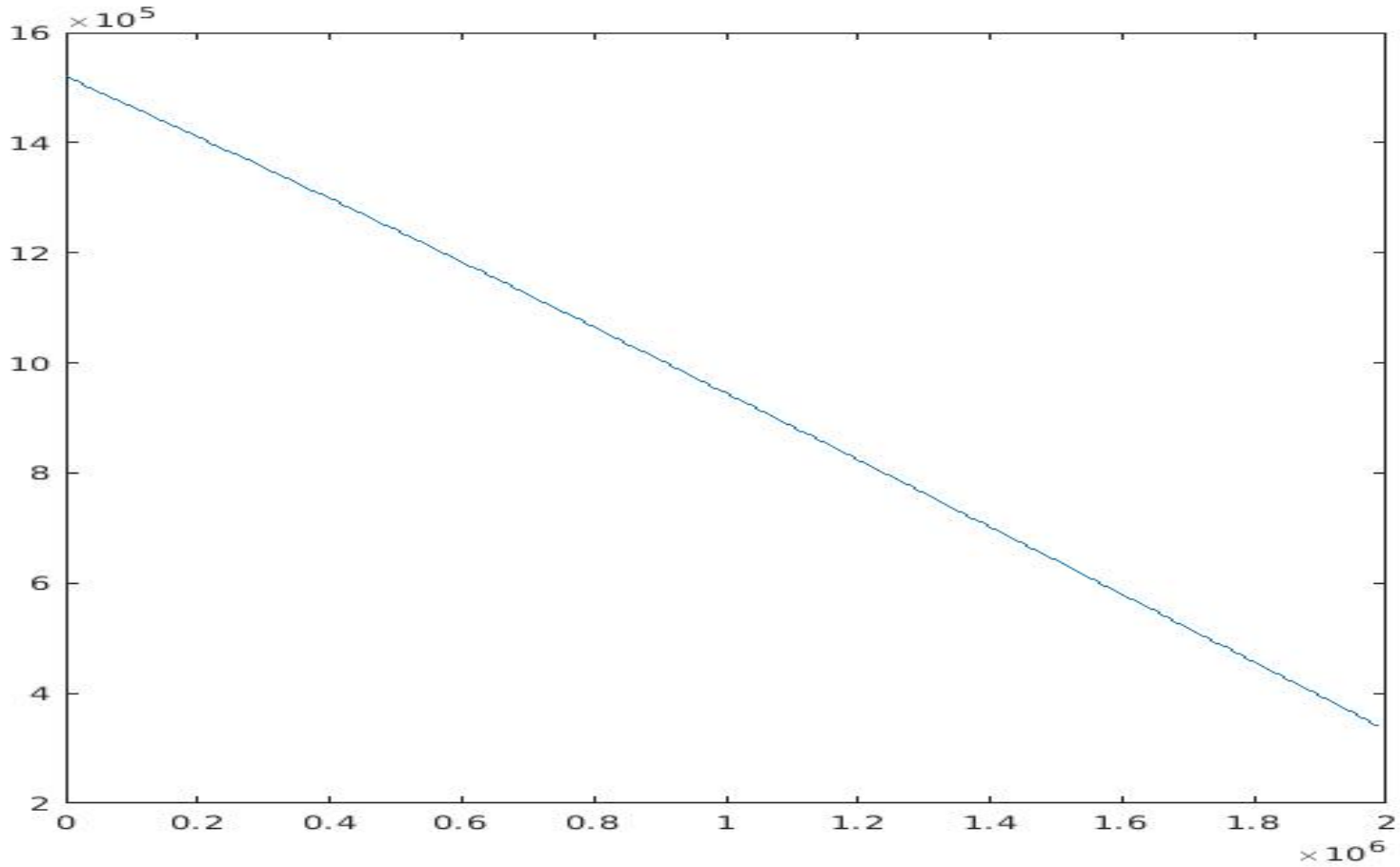
- **Model :** The randomly chosen voter changes preference when it obtains reward $= 1$ on the arm of opposite preference or reward $= 0$ on the arm of its own preference.
- The proportion of preference 0 users as a function of time for a fixed policy becomes :

$$z(t) = \frac{d_2}{d_1 + d_2} + (z_0 - \frac{d_2}{d_1 + d_2})e^{-t\frac{d_1+d_2}{A}}$$

  where c and a are defined as before.
- The optimal policy for this case turns out to be the same as the one obtained previously.

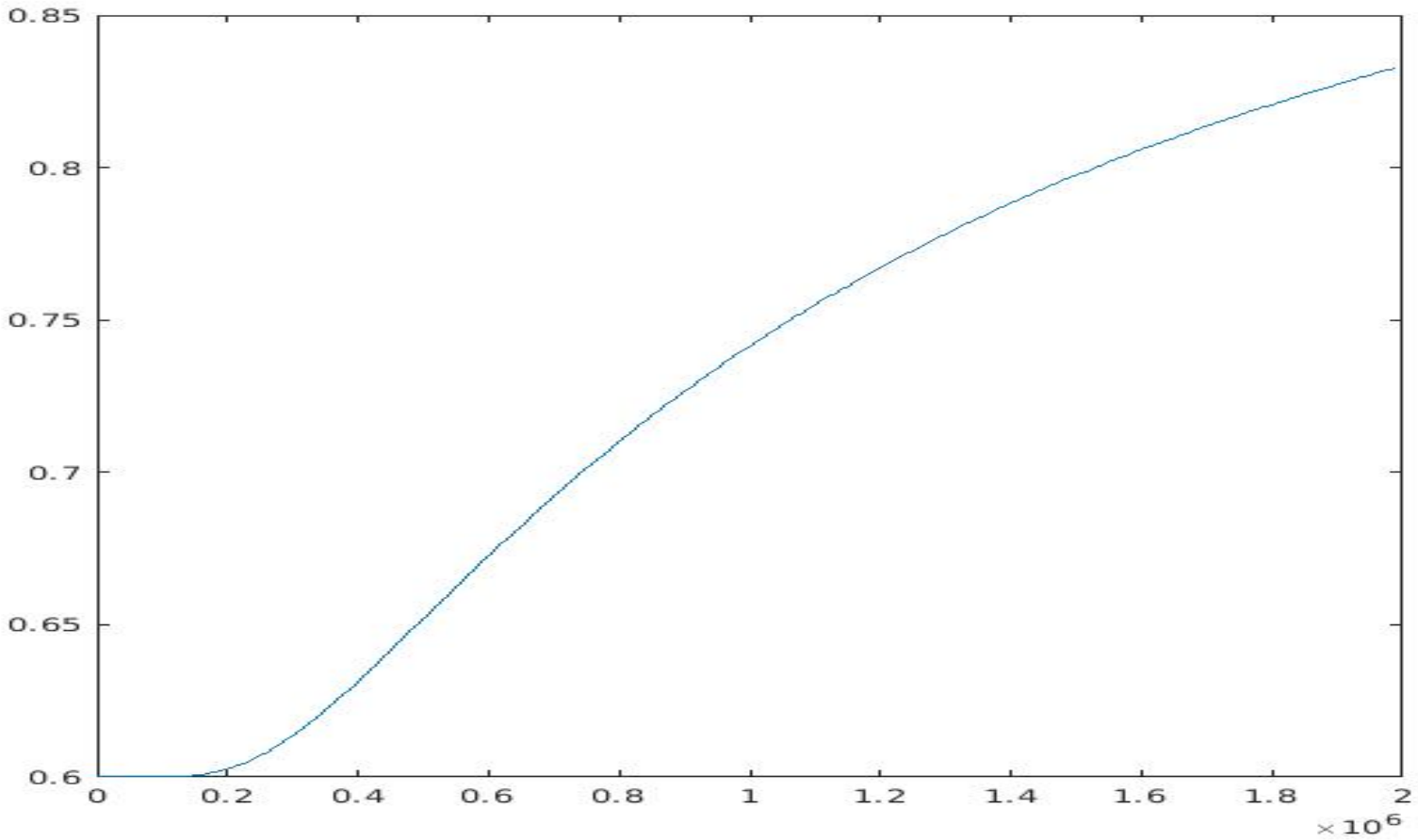
Mixed Policy Results for varying T_0

Figure 7: Cumulative reward at deadline Vs T_0 (deadline $T = 2 \times 10^6$)



Mixed Policy Results for varying T_0

Figure 8: Proportion of type 0 at deadline Vs T_0 (deadline $T = 2 \times 10^6$)



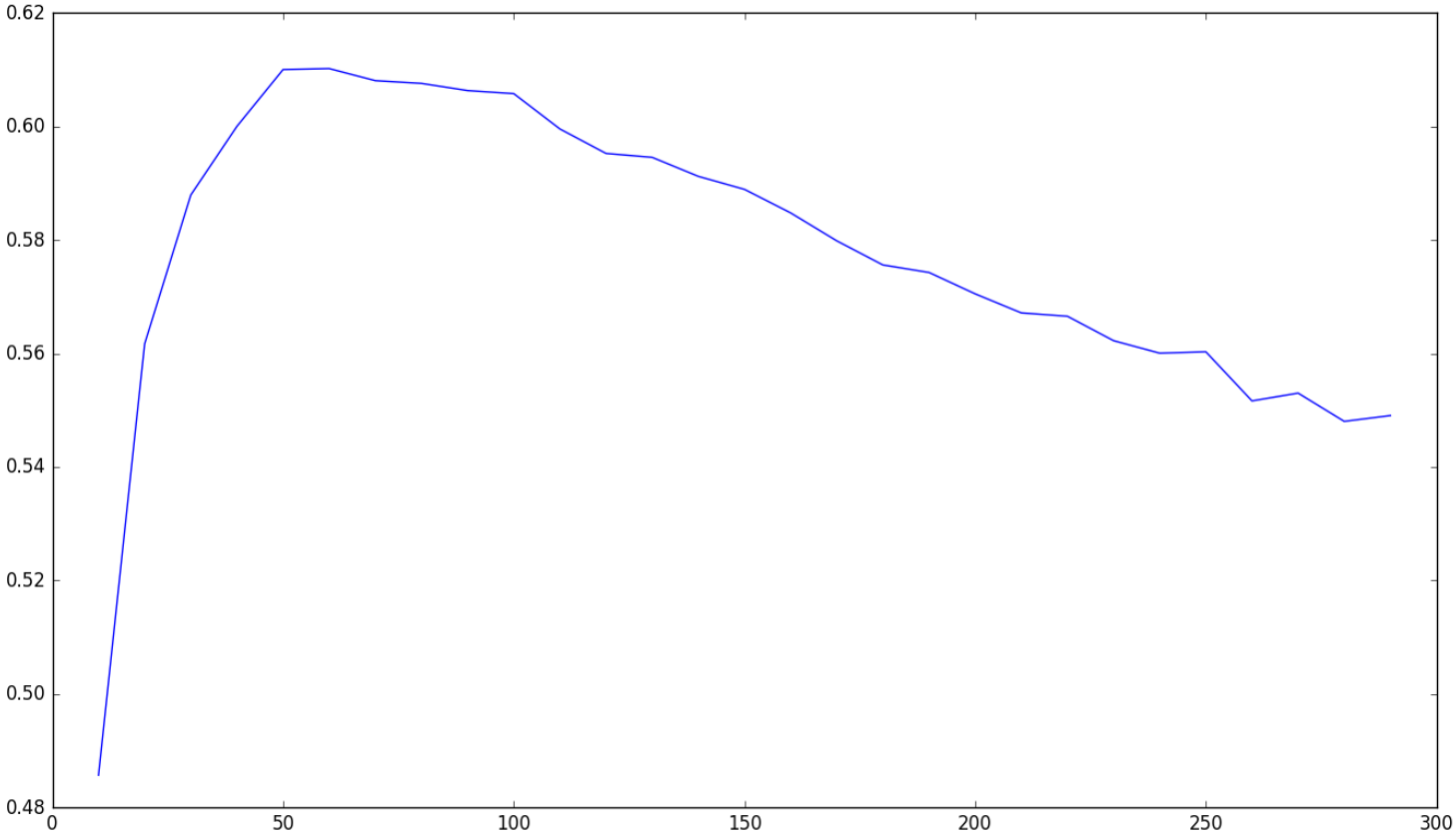
Unknown matrix

Now we move on to the case where the matrix B is unknown and we wish to maximize the type 0 preference user population at a deadline T.

- The general problem we seek to tackle is :
  Problem : Given a deadline T, find a policy that maximizes the population of type 0 users at $t = T$ (or alternatively, minimize the expected cumulative regret between it and the optimal policy).
- We first restrict ourselves to the following policy :
  Policy 1 : Show arms uniformly at random and keep updating the estimate of matrix B till a time T_{thresh}. After this, show arms according to the optimal policy for the estimated matrix $\hat{B}$.

Unknown matrix

Figure 9: Proportion of type 0 users at deadline Vs T_{thresh} (deadline $T = 500$)



Unknown matrix

- The plot in the previous slide was for a case when the optimal policy is supposed to be $p = q = 1$.
- A similar plot is obtained for other values of matrix B, demonstrating the tradeoff between estimation of B and exploitation in all cases.
- The optimal T_{thresh} changes with the matrix B, which is why we need a policy that does estimation simultaneously and incorporates some measure of certainty of the estimate while recommending arms.

Analyzing the Explore-Then-Commit policy

Notation :

- Z_t =number of type 0 balls in urn at time t
- z_t =proportion of type 0 balls in the urn at time t

We define regret at time T as :

$$Regret(T) = E[\sum_{t=1}^{T}(\Delta Z_t^{opt} - \Delta Z_t^{pol})]$$

where the ΔZ_t^{opt} is the number of type 0 balls added at time t *given that the proportion of type 0 balls in the urn is* z_t^{pol}. We wish to minimize this regret.

Analyzing the Explore-Then-Commit policy

The above definition gives us the following expression for regret at time T.

$$Regret(T) = R_{explore} + R_{exploit}$$

where,

$$R_{explore} = 0.5(\sum_{1}^{2m} z_t)|b_{00} + b_{01} - 1| + 0.5(2m - \sum_{1}^{2m} z_t)|b_{11} + b_{10} - 1|$$

and

$$R_{exploit} = p'(\sum_{2m}^{T} z_t)|b_{00} + b_{01} - 1| + q'(T - 2m - \sum_{2m}^{T} z_t)|b_{11} + b_{10} - 1|$$

Here p' and q' are the probabilities of "bad" events happening. That is, if say $b_{00} + b_{01} > 1$ then $p' = Prob(\hat{b_{00}} + \hat{b_{01}} < 1)$.

Analyzing the Explore-Then-Commit policy

Let $|b_{00} + b_{01} - 1| = \Delta_0$ and $|b_{11} + b_{10} - 1| = \Delta_1$. Also let $S_i^j = \sum_{t=i}^{j} z_t$.Then :

$$Regret(T) = [0.5S_1^{2m}\Delta_0 + 0.5(2m - S_1^{2m})\Delta_1] + [p' S_{2m}^{T}\Delta_0 + q'(T - 2m - S_{2m}^{T})\Delta_1$$

If we want an upper bound on the regret, then we need an upper bound on p', q'. This demands a lower bound on the number of times a particular element of the matrix B was sampled.

Analyzing the Explore-Then-Commit policy

Some observations that may be useful :

- In the exploration phase, $U_0 = max(z_0, \frac{1-b_{00}+b_{01}}{1-b_{00}-b_{11}+b_{01}+b_{10}}) \geq z_t$ and $z_t \geq min(z_0, \frac{1-b_{00}+b_{01}}{1-b_{00}-b_{11}+b_{01}+b_{10}}) = L_0$
- By the above observation, $2mU_0 \geq \sum_1^{2m} z_t \geq 2mL_0$.
- Similarly we can find bounds (in terms of the asymptotic proportion values) for the sum term in the exploitation phase. Let those bounds be $(T-2m)U_1, (T-2m)L_1$.

Analyzing the Explore-Then-Commit policy

Using the Hoeffding bound and the above observations, we get :

$$p' \leq e^{-mL_0\Delta_0^2}$$

$$q' \leq e^{-m(1-U_0)\Delta_1^2}$$

The upper bound on regret becomes :

$$[mU_0+(T-2m)U_1e^{-mL_0\Delta_0^2}]\Delta_0+[m(1-L_0)+(T-2m)(1-L_1)e^{-m(1-U_0)\Delta_1^2}]\Delta_1$$

Analyzing the Explore-Then-Commit policy

Deriving a bounds on regret for a special case : Consider the case where we have $b_{00} = b_{11}$ and $b_{01} = b_{10}$. For this case we get the following two bounds on the regret as defined previously :

- **Gap-dependent bound :**

$$R_T \leq m + e^{\frac{-m\Delta^2}{2}}(T - 2m)$$

- **Gap-independent bound :**

$$R_T \leq \mathcal{O}(T^{2/3}(log(T))^{1/3})$$

ETC algorithm results using the Hoeffding bounds

- We now plot the analytical expressions for the population trajectory for the ETC algorithm.
- The expression used to plot this is :

$$z(t) = z_\infty + (z_{explore} - z_\infty)(1 + t/A)^{-(d_1+d_2)}$$

where d_1 and d_2 depend on the policy (p, q).

- Also, we use the following fact

$$\mathcal{E}(z(t)) = z_{right} * P(right) + z_{wrong} * P(wrong)$$

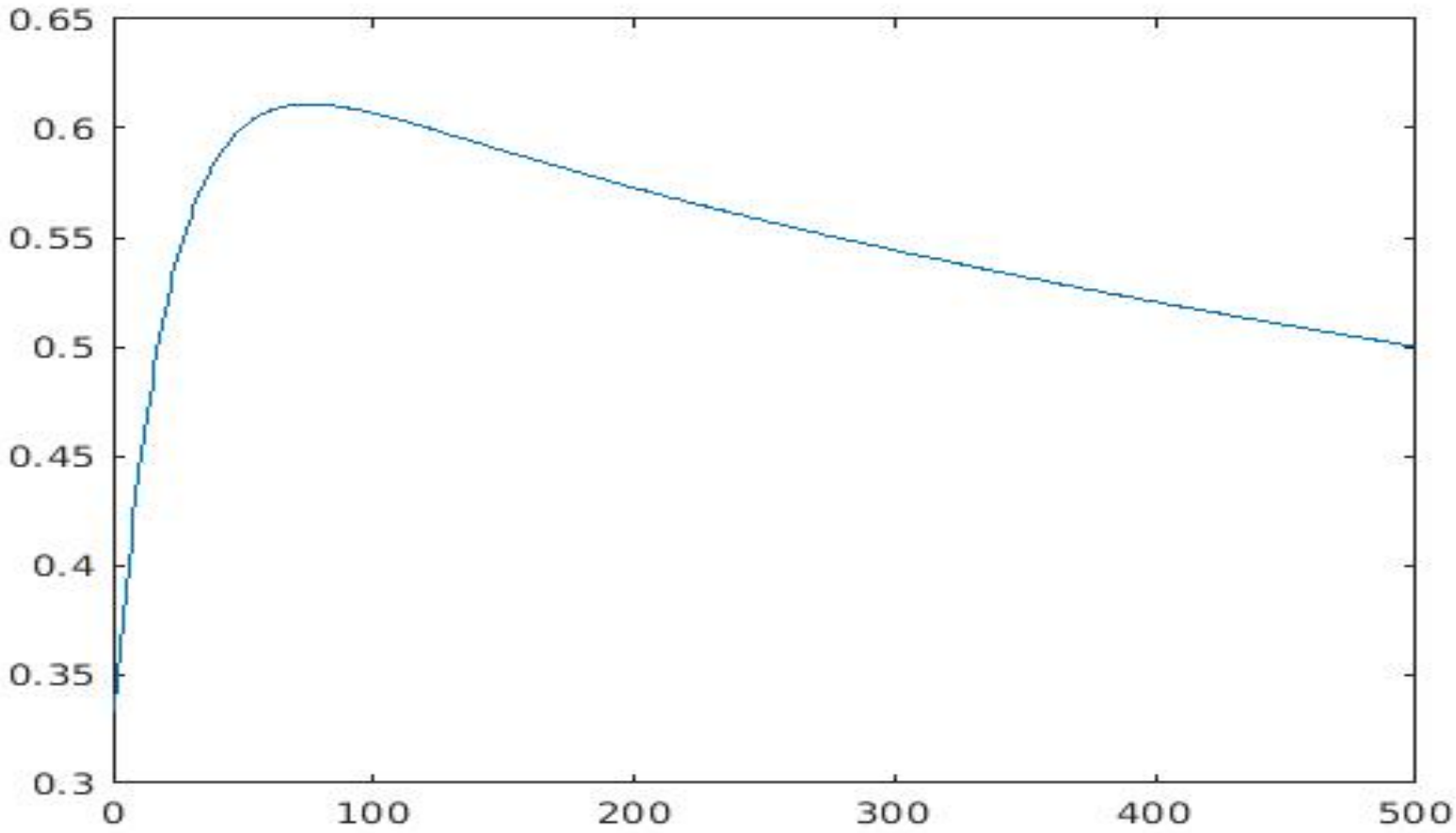
where the events "wrong" and "right" are the events that our estimates at the end of the exploration phase are wrong and right respectively.

- We use the following bound on $P(wrong)$:

$$P(wrong) \leq e^{-kT_{thresh}\Delta^2}$$

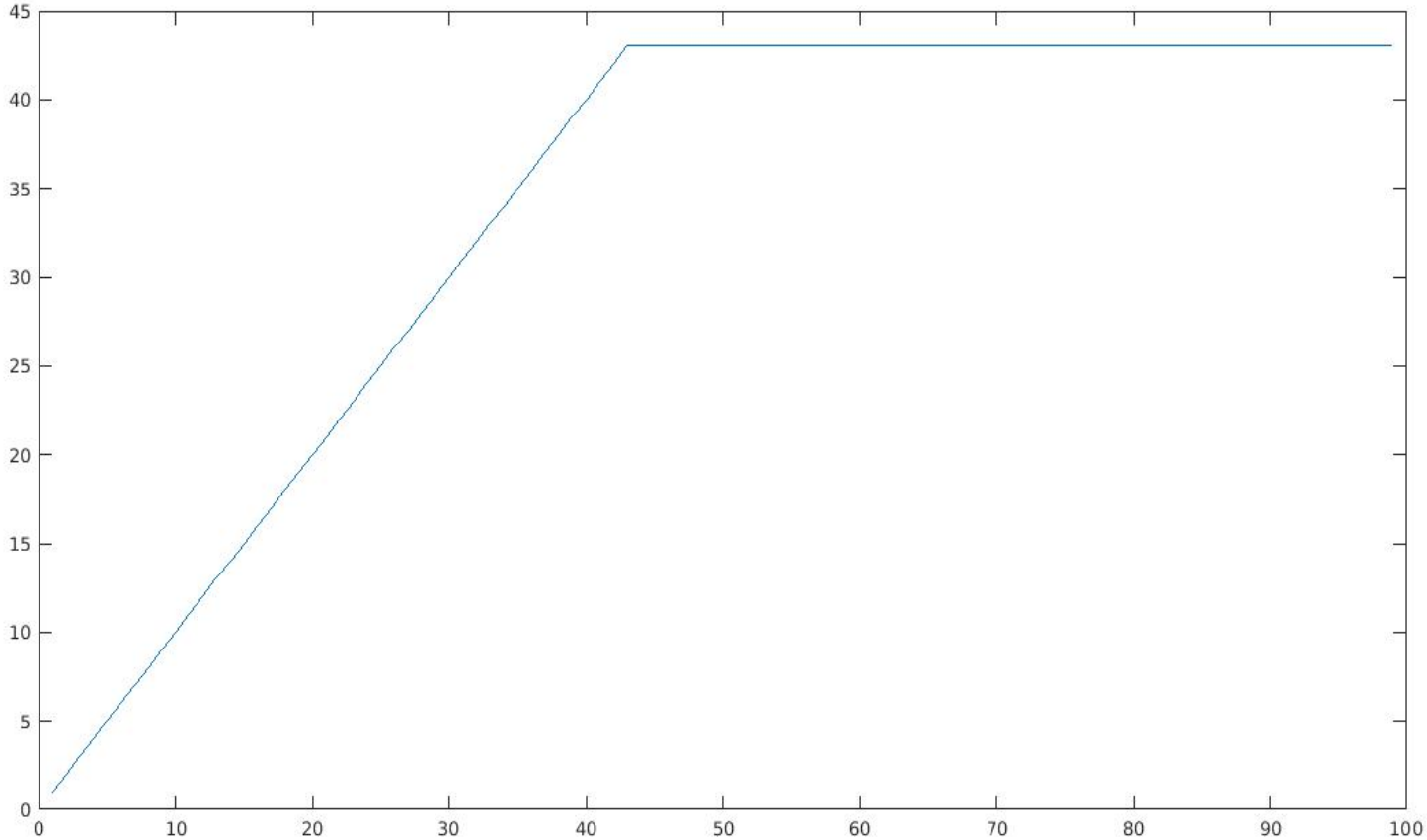
ETC algorithm results using the Hoeffding bounds

Figure 10: Proportion vs Value of threshold for ETC policy with deadline = 500



ETC algorithm results using the Hoeffding bounds

Figure 11: Optimal threshold vs deadline



Need for a better policy

- Since the matrix B is unknown and the optimal threshold $2m$ in our Explore-Then-Commit policy depends heavily on the values in B (through Δ_i), we need a policy that keeps track of our confidence in the estimates of the elements of the matrix B.
- Therefore we try out two policies that use concepts similar to those used in the UCB and Thompson sampling algorithms.

UCB-like algorithm

This algorithm follows the following steps. In each time step, do :

- Keep an estimate matrix of the matrix B as $\hat{B}$.
- Define :

$$UCB_{ij} = \hat{b}_{ij} + \sqrt{\frac{klog(t)}{T_{ij}(t)}}$$

$$LCB_{ij} = \hat{b}_{ij} - \sqrt{\frac{klog(t)}{T_{ij}(t)}}$$

  where $T_{ij}(t)$ is the number of times b_{ij} was sampled till time t.
- If $UCB_{00} + UCB_{01} - 1 < 0$ then set $p = 0$. If $LCB_{00} + LCB_{01} - 1 > 0$ then set $p = 0$. If neither are true, set $p = 0.5$.
- Do a similar procedure for setting the value of q.

Thompson-like algorithm

This algorithm follows the following steps. In each time step, do :

- Keep two matrices $A_{2\times 2}(t)$ and $B_{2\times 2}(t)$ (initialised to all ones at $t = 0$).
- Sample a matrix of values $B_{sample}(t)$ such that :

$$B^{ij}_{sample} \sim \beta(A_{ij}(t), B_{ij}(t))$$

- Set $p = I_{B^{00}_{sample}+B^{01}_{sample}-1>0}$ and $q = I_{B^{11}_{sample}+B^{10}_{sample}-1<0}$.
- If reward in time slot t is R_t, then update the A and B matrices as :

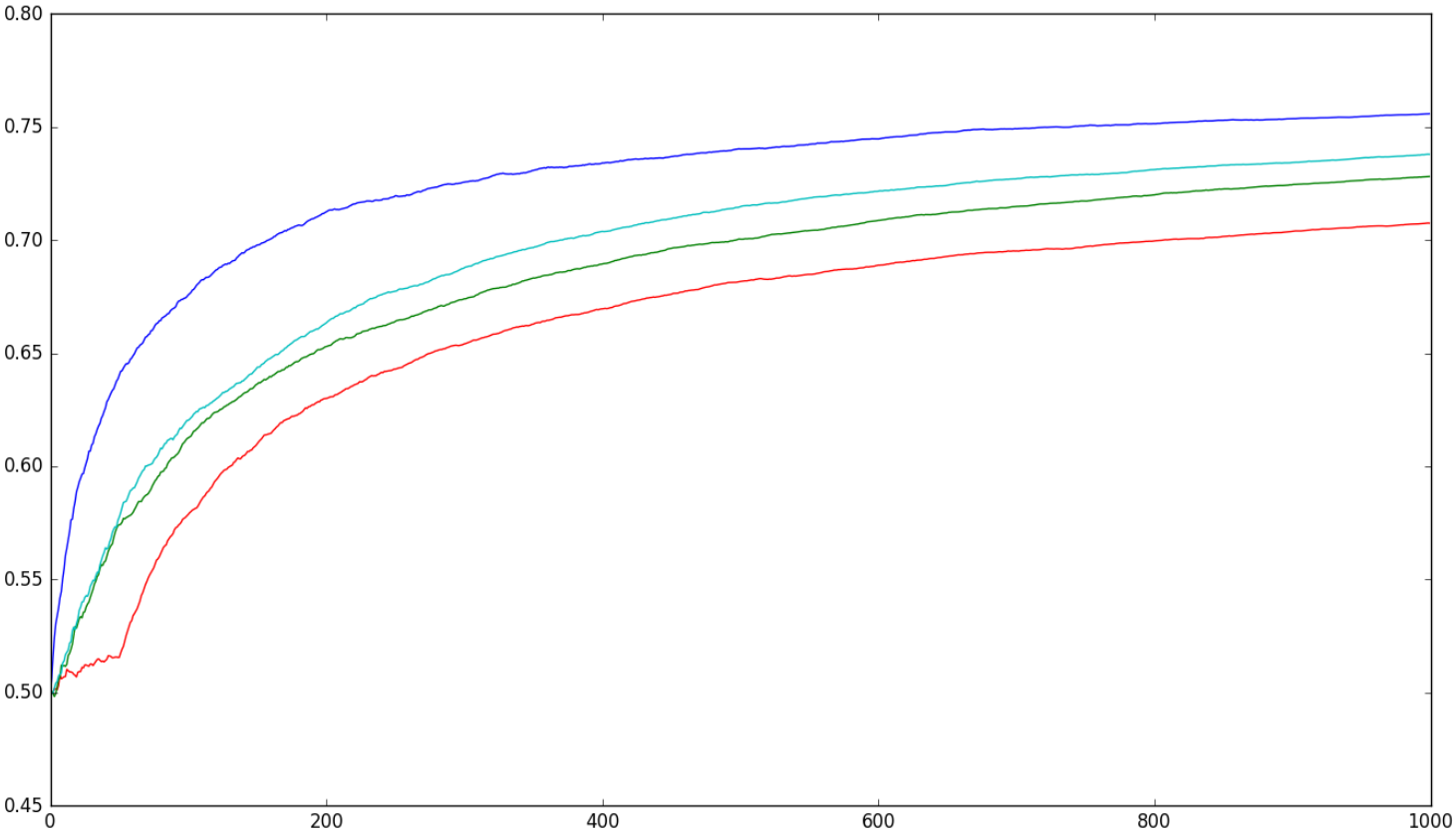
$$A_{ij}(t+1) = A_{ij}(t) + R_t$$

$$B_{ij}(t+1) = A_{ij}(t) + 1 - R_t$$

where the arm preference i and a arm recommended j in the time slot t.

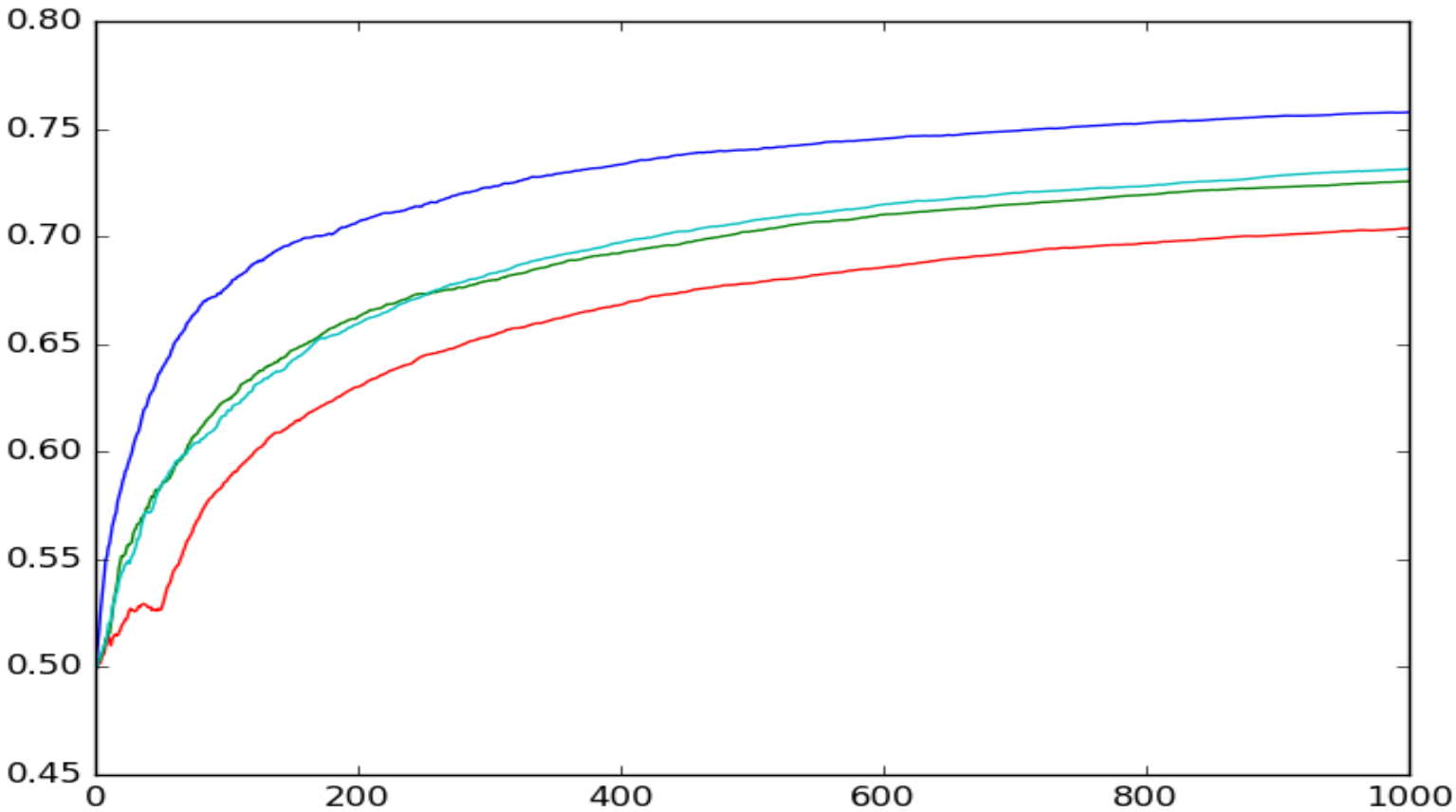
Results on UCB and Thompson

Figure 12: Proportion of type 0 users for various policies Vs Time (Blue : Optimal, Red : ETC, Green : UCB, Cyan : Thompson)



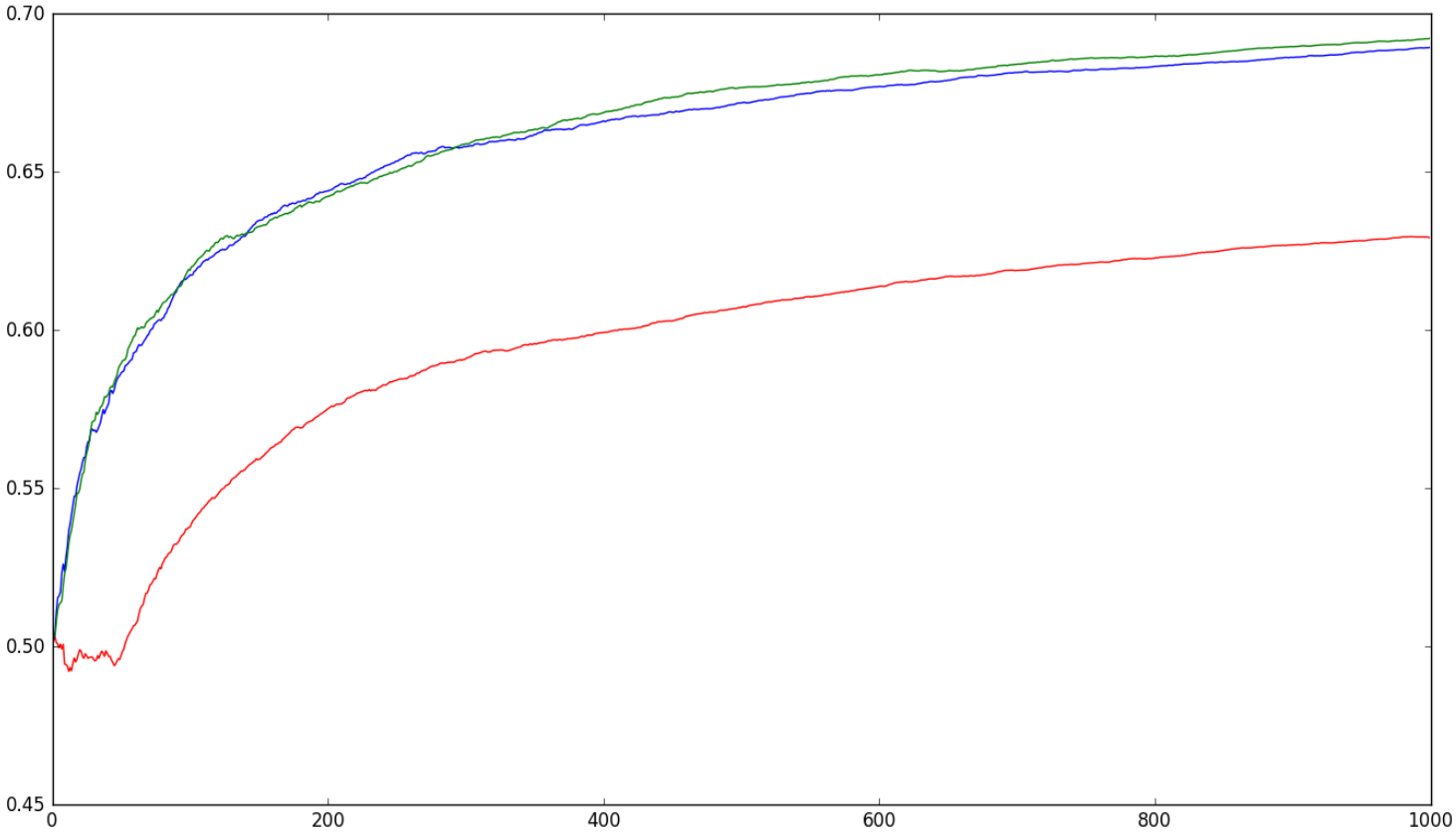
Results on UCB and Thompson

Figure 13: Proportion of type 0 users for various policies Vs Time (Blue : Optimal, Red : ETC, Green : UCB, Cyan : Thompson)



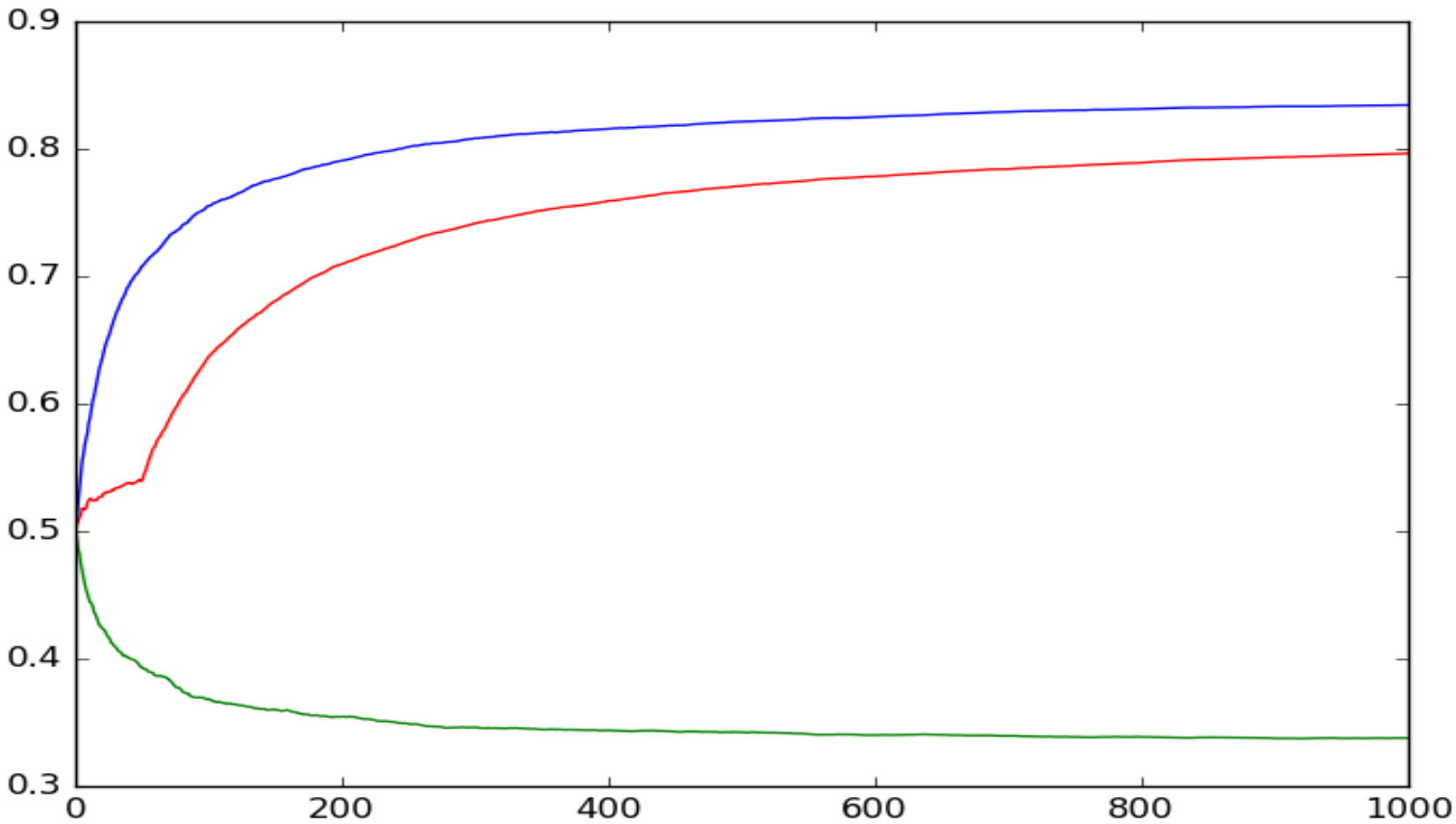
Results on UCB and Thompson

Figure 14: Proportion of type 0 users for various policies Vs Time(Blue : Optimal, Red : ETC, Green : Overly optimistic UCB)



Results on UCB and Thompson

Figure 15: Proportion of type 0 users for various policies Vs Time (Blue : Optimal, Red : ETC, Green : Overly optimistic UCB)



Observations

- Thompson sampling (appropriately tuned) performs better than UCB, which in turn performs better than ETC.
- For some values of matrix B, overly optimistic UCB (i.e. UCB with k $= 0$) is as good as the optimal policy. However, this policy gives us opposite results for other values of B.

Regret bounds on Thompson sampling

- We now state bounds on the regret for the Thompson sampling policy applied to our model. Assume that the initial proportion of populations is 0.5 and the optimal policy reaches the value z^* asymptotically. Here, $\Delta_0 = |b_{00} + b_{01} - 1|, \Delta_1 = |b_{10} + b_{11} - 1|$
- **Gap dependent bound :**

$$R_T^{Thomp} \leq \frac{3z^*}{8}(\frac{1}{\Delta_0} + \frac{1}{\Delta_1})\sum_{i=1}^{T}\frac{1}{i} \leq \frac{3z^*}{8}(\frac{1}{\Delta_0} + \frac{1}{\Delta_1})log(T)$$

- **Gap independent bound :**

$$R_T^{Thomp} \leq \mathcal{O}(log(T))$$

Note that R_T^{ETC} was $\leq \mathcal{O}(T^{2/3}(log(T))^{1/3})$

Analogous results for Voter Model (fixed total population)

- It turns out that regret expression for the voter model is just twice the regret expression for our previous urn model.
- The equation of the trajectory in the voter model case becomes :

$$z(t) = \frac{d_2}{d_1 + d_2} + (z_0 - \frac{d_2}{d_1 + d_2})e^{-t\frac{d_1+d_2}{A}}$$

- Therefore, we see that the expression for the asymptotic proportion of type 0 users is still the same (i.e. $\frac{d_2}{d_1+d_2}$). This means the optimal policy also turns out to be the same.
- Therefore all of our regret analysis results still hold for the fixed population case (but for a factor of 2).

Link to all codes

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