Gil Landau
12/31/18
ESE 545

Project 4 Report

Part 1:

We can view this is as a non-stochastic multi-armed bandit problem. Each arm is a type of ad delivered, when a user appears. For each round (column) *t*, we pick a type of ad (row) *i* and assign it with a reward $r_{i,t}$ and our goal is to maximize the reward of $\sum_{t=1}^{T} r_{i,t}$, which corresponds to the total reward of selecting an ad that a user will click. We want to minimize our total regret, defined as our reward subtracted from the reward of the arm with the most success: $R_T = \max_{i1..k} \sum_{t=1}^{T} r_{i,t} - \sum_{t=1}^{T} r'_{i,t}$. Our ideal goal is for $\frac{R_T}{T}$, the average regret, to approach 0.

Part 2:

You can view the code for this part in the section labelled "Part 2." This is an implementation of the EXP3 algorithm, using partial feedback. I have altered it to maximize reward, rather than minimize loss. I define the probability distribution as $p_{i,t} = \frac{e^{\eta(\text{total reward of i at t})}}{\sum_{i=1}^{n} e^{\eta(\text{total reward or i at t})}}$. I compute the total reward of a chosen arm at t, A_t, that yield a reward, X_t, as

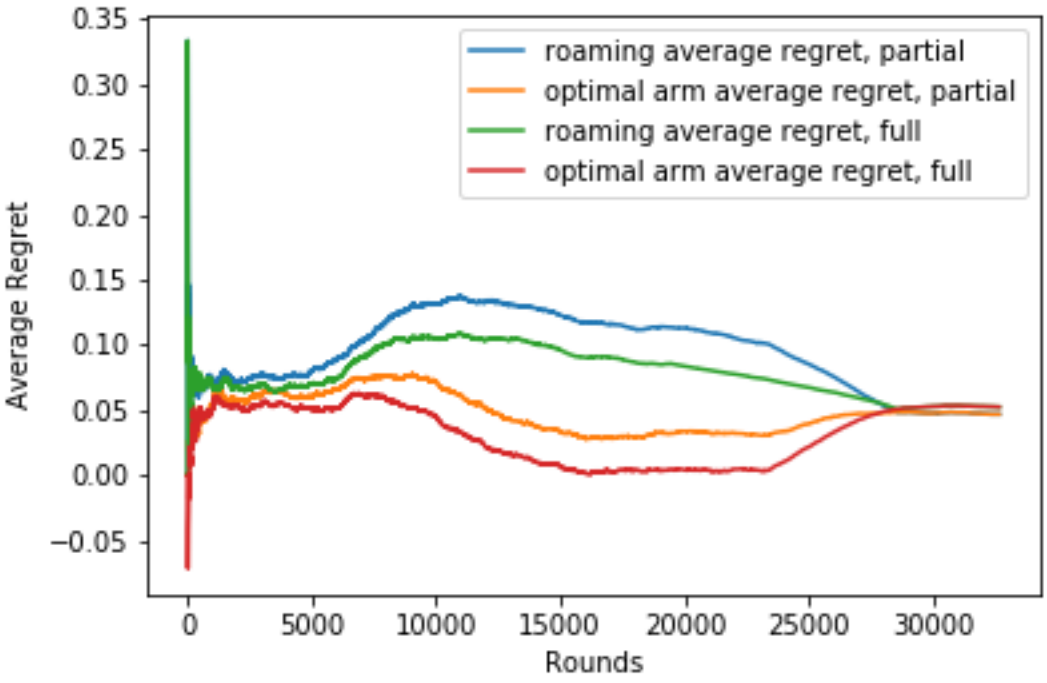
(total reward of i at t) = (total reward of i at t) + 1 - $\frac{(1-X_t)1(A_t = i)}{p_{i,t}}$, so I punish arms that yield failures. I choose η from various readings, that discuss more about learning rate selection as a function of the number of arms and the number of rounds (see here for the textbook: Bandit Algorithms - Tor Lattimore and Csaba Szepesv´ari). I previously tried various rates of η, ranging from small constants, such as .000085, and variable constants, such as $\frac{1}{t}$, which yielded poor results.

Part 3:

You can view the code for this part in the section labelled "Part 3." Here I use a full-feedback version of EXP3. That means instead of creating an estimation of the rewards for the arm that I "pull," I use all the rewards for round t and update the total rewards for all arms. So

(total reward of i at time t) = (total reward of i at time t) + $X_{i,t}$.

Here you can see the graph of the average regret for both partial-feedback and full-feedback EXP3 algorithms. I define two different types of averages: the "optimal arm" average and the "roaming" average.



The first is the "optimal arm" average. With full hindsight, I determine the arm that yielded the most rewards. I then track its rewards over time t, such that $\text{rbest}_t = \text{rbest}_{t-1} + X_{\text{best},t}$. This is the "optimal" reward that my algorithm subtracts from in order to determine regret ($R_t = \text{rbest}_t - \sum_{a=1}^{t} r'_a$), where r' is the reward of my algorithm at time a.

The second is the “roaming” average, which is a bit harsher (although eventually both averages converge, due to obvious reasons). I define this as $\max_{i1..k} \sum_{a=1}^{t} r_{i,a}$. For a time t, it is the “optimal arm.” When t=T, the “roaming” average is the same as the “optimal arm” average.

What I found is that full-feedback does yield a deeper decrease in regret, getting as low as .00105 for optimal arm and .0508 for roaming, respectively. This makes sense, since I am using the actual reward information, rather than estimating it. Ultimately, however, regret does creep up for all four averages and all four average converge roughly to the same point. The lowest regret comes from the partial feedback EXP3, at .0468506. Full feedback converges at .05239306.