IEOR 8100-001: Learning and Optimization for Sequential Decision Making

Homework 2

Instructor: Shipra Agrawal **Due on: Feb 24, 2016**

- All problems carry equal points.
- You may discuss the problems on this assignment with others, but you must write your own solutions.
- Even if you cannot solve a problem completely, make sure to provide your partial/suboptimal solution to get partial credit.
- Turn in the solutions in class or email them to ieor8100-001-spring2016@columbia.edu before class on due date.

Problem 1. Recall the ϵ-greedy algorithm for stochastic N-armed bandit we discussed in class: "at time t, with ϵ probability, play any arm uniformly at random , and with $1-\epsilon$ probability, take the greedy choice". Prove that with $\epsilon = \left(\frac{N}{T}\right)^{1/4}$, this algorithm achieves following bound on problem-independent expected regret in time T:

$$E[R(T)] \leq O((N \log T)^{1/4} T^{3/4}) \tag{1}$$

You may assume $N \leq T$. Bounds within $\log(NT)$ factors of the stated bound will earn full credit.
Hint: In every N/ϵ steps, "every arm" is played once (approximately). After how many steps, arm i and arm 1 will have had sufficient number of plays, so that if $\Delta_i \geq \epsilon$, then arm i will not be played with high probability?

Note that this algorithm requires advance knowledge of the time horizon T that the algorithm will be used for. The regret may not be bounded as above if the algorithm is run for a longer horizon. Compare this to UCB and Thompson Sampling which bound regret $R(T)$ by $O(\sqrt{NT\log(T)})$ for all T.

Problem 2. Modify the algorithm in Problem 1 in the following way. Work in epochs of doubling length, i.e., epoch j is of length $\ell_j = 2^{j-1}$, starts at time step $t = 2^{j-1}+1$ and ends at $t = 2^j$. In epoch j, run ϵ-greedy with $\epsilon = \left(\frac{N}{\ell_j}\right)^{1/3}$. Prove that this algorithm achieves the regret bound $O(N^{1/3}T^{2/3}(\log T)^{4/3})$ for *any* time horizon T.

You may assume $N \leq T$. Bounds within $\log(NT)$ factors of the stated bound will earn full credit.
Hint: Observe that in the beginning of epoch j, for every arm, there will be close to $\frac{\epsilon \ell_j}{N}$ samples available from previous epochs.

Problem 3. Consider following arm-elimination algorithm for stochastic N-armed bandit problem: Proceed in rounds, in every round play once every arm that has not been eliminated yet. So, in the first round, you will play every arm once. At the beginning of every round, for every arm i, check if $\text{UCB}_{i,t-1} \leq \text{LCB}_{j,t-1}$ for some j not eliminated yet. If yes, eliminate arm i.

This algorithm uses slightly modified definitions of UCB (and LCB) from those discussed in class:

$$UCB_{i,t} = \hat{\mu}_{i,t} + \sqrt{\frac{\log(T)}{n_{i,t}}}, LCB_{i,t} = \hat{\mu}_{i,t} - \sqrt{\frac{\log(T)}{n_{i,t}}}$$

Note that above definitions use knowledge of time horizon T. In class, we used $\log(t)$ instead of $\log(T)$ in the exploration term above. Show that this algorithm will achieve expected regret bound of $O(\sum_{i \neq I^*} \frac{\log(T)}{\Delta_i})$. You may assume that there is a unique optimal arm, and that $N \leq T$.

Again observe that this algorithm requires advance knowledge of T, and cannot provide small regret for *any* time horizon. In fact for an algorithm to work for infinite time horizon, it must never completely eliminate an arm. ***Hint:*** Prove the following claims (a) The best arm I^* will not be eliminated in any round with high probability, (b) A suboptimal arm i will be eliminated after at most $\log(T)/\Delta_i^2$ rounds. Regret bound should directly follow from these claims.

Problem 4. Suppose that the reward r_t generated on playing arm I_t at time t is observed after a delay of γ, i.e. at time step $t+\gamma$. You may assume unique optimal arm and advance knowledge of T. Regret in time T is defined as before:

$$\mathcal{R}(T) = T\mu^* - \sum_{t=1}^{T} \mu_{I_t}$$

1. Show that a natural modification of UCB algorithm achieves $O(\sum_{i \neq I^*} \frac{\log(T)}{\Delta_i} + \sum_{i \neq I^*} \Delta_i \gamma)$ bound on expected regret in time T for this problem. (Or, provide an alternate algorithm that achieves this regret bound.)
2. Provide an algorithm that achieves an improved regret bound of $O(\sum_{i \neq I^*} \frac{\log(T)}{\Delta_i} + \sum_i \Delta_i + \gamma \log \gamma)$, assuming $\gamma \leq N \leq T$.
   Hint: Use the algorithm in Problem 3 with required modifications.

Problem 5. [Additional feedback] Suppose that we are given an undirected graph G, where N nodes of the graph correspond to the Narms. Two arms i, j are neighbors iff there is an edge between i and j. On playing arm $I_t = i$ at time t, you receive reward r_t generated i.i.d. from (unknown) distribution ν_i with mean μ_i. Additionally, for all neighbors j of i in graph G, you get to observe $X_{t,j}$ generated i.i.d. from (unknown) distribution ν_j. Regret is defined as before: $\mathcal{R}(T) = T\mu^* - \sum_t \mu_{I_t}$. The aim is to use these extra observations to learn faster and achieve a better regret bound. Provide an algorithm so that if the graph G is such that it can be covered by $k \leq N$ cliques, then the worst case regret bound is improved from $O(\sqrt{NT\log(T)})$ to $O(\sqrt{kT\log(T)\log(N)})$.
You may assume $N \leq T$, unique optimal arm, and if required, advance knowledge of T. Bounds within $\log(NT)$ factors of the stated bound will earn full credit.
Hint: In the algorithm given in Problem 3 for example, after every round, every arm in clique C has $|C|$ new samples (instead of 1 in the regular problem). This directly reduces the number of rounds that arm i needs. Though you have to be careful about the fact that some arms from this clique may have been eliminated, which decreases the effective $|C|$ over time.

(As an example application of this problem setup, think about arms as people in social network, and pulling an arm means making an offer to a person in the network. Reward is receiving a like or a click from the person who the offer was given. You may also observe the feeling for the offer (likes/shares/clicks) from his/her friends.)