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

# Homework 2 solutions

Instructor: Shipra Agrawal, TA: Michael Levi Hamilton

- 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 $\epsilon$-greedy algorithm for stochastic $N$-armed bandit we discussed in class: "at time $t$, with $\epsilon$ 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 $N/\epsilon$ 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 $T$.

*Proof.* Divide the N arms into two groups, $N_1$ and $N_2$ s.t. $\forall i \in N_1, s.t \Delta_i \leq \epsilon$ otherwise put the arm in $N_2$. Then:

$$\mathbb{E}[R(T)] \leq \sum_{t=1}^{T} \max_{i \in N_1} \Delta_i Pr(I_t \in N_1 - \{i^*\}) + \sum_{t=1}^{T} Pr(I_t \in N_2)\Delta_{I_t} \tag{2}$$

$$\leq \epsilon T + \sum_{t=1}^{T} Pr(I_t \in N_2) \tag{3}$$

Where we've noted that $\Delta_i \leq 1$ for all arms. Thus we can focus only on arms in group 2. The analysis will consider time $S$ by which we'll have explored each suboptimal arm enough times. Let $\Delta = \min_{i \in N_2} \Delta_i$ and let $n_{i,t}$ be the number of pulls of arm $i$ at $t$. Then define $S = \min\{t \mid n_{i,t} \geq 4 \log T/\Delta^2 \ \forall i \in N_2\}$. And note since each arm is randomly pulled with probability $\epsilon/N$,

$$\mathbb{E}[S] \leq \frac{N \log T}{\epsilon \Delta^2} \tag{4}$$

.

Now let's bound the probability of making a mistake for arms in $N_2$ after time $S$. Suppose at time $t$ the algorithm pulls a suboptimal arm. Then for $i \in N_2$

$$\Delta_i \leq \mu_{i^*} - \mu_{I_t} \leq (\mu_{i^*} - \hat{\mu_{i^*}}) + (\hat{\mu_{i^*}} - \hat{\mu_{I_t}}) + (\hat{\mu_{I_t}} - \mu_{I_t}) \tag{5}$$

$$\leq 2 \max_{i \in N_2} |\mu_i - \hat{\mu_i}| \tag{6}$$

where we've noted that the middle term in (3) is negative. Using this fact, the probability of pulling a suboptimal arm at $t$ is bounded by:

$$\mathbb{P}(\hat{\mu_{i,t}} \geq \hat{\mu_{i^*,t}}) \leq \mathbb{P}(\max_{i \in N_2} |\mu_i - \hat{\mu_i}| \geq \Delta/2) \leq N \max_{i \in N_2} \mathbb{P}(|\mu_i - \hat{\mu_i}| \geq \Delta/2)$$

where the second inequality follows from union bound. Finally note by Chernoff bounds, for all times $t \geq S$, since we've pulled each arm $4\frac{\log T}{\Delta^2}$ times, the probability $\mathbb{P}(|\mu_i - \hat{\mu_i}| \geq \Delta/2)$, is $O(\frac{1}{T^2})$. Putting it all together we see:

$$\begin{aligned}
\mathbb{E}[R(T)] &\leq \epsilon T + \sum_{t=1}^{T} Pr(I_t \in N_2) \\
&= \epsilon T + \mathbb{E}[\sum_{t=1}^{S} \mathbb{1}(I_t \in N_2)] + \mathbb{E}[\sum_{t=S}^{T} \mathbb{1}(I_t \in N_2)] \\
&\leq \epsilon T + \mathbb{E}[S] + NTO(\frac{1}{T^2}) \leq \epsilon T + \frac{N \log T}{\epsilon \Delta^2} + O(1)
\end{aligned}$$

Setting $\epsilon = \Delta = \left(\frac{N \log T}{T}\right)^{1/4}$ gives the result.

□

**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 $\epsilon$-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.

**Solution .**

Let $R_j$ be the regret in epoch $j$, we'll bound $R_j$ by:

$$\mathbb{E}[R_j] \leq O(N^{1/3} l_j^{2/3} \log {l_j}^{1/3}) \tag{7}$$

Before proving this note that for any $T$, we'll have $O(\log T)$ epochs and thus:

$$\mathbb{E}[R(T)] = \sum_{j}^{\log T} \mathbb{E}[R_j] \leq \sum_{j}^{\log T} N^{1/3} l_j^{2/3} \log {l_j}^{1/3} \leq O(N^{1/3} T^{2/3} \log T^{4/3}) \tag{8}$$

thus bounding the regret in each epoch gives our desired result.

To bound the regret in each epoch we'll use a similar argument as in (Q1). We'll slightly modify the $\{\epsilon_j\}$ so as to ensure that in each epoch we have the necessary number of samples. Define the sequence of $\epsilon' s$ to be:

$$\epsilon_j = \left(\frac{4N \log(l_j)}{l_j}\right)^{1/3} \tag{9}$$

Fix the epoch, the analysis of $\mathbb{E}[R_j]$ will proceed as follows:

- Note $\sum_{i=1}^{j-2} 2^i = 2^{j-1} - 1 = \ell_j - 1$, so the size of our current epoch $j$ is approximately the size of all previous epochs combined.
- We'll split the arms, letting $N_1 = \{i \mid \Delta_i \leq \epsilon_j\}$ and $N_2 = \{i \mid \Delta_i \geq \epsilon_j\}$ .
- By end of previous epoch, i.e., by time $t = 2^{j-1}$, each arm $i \in N_2$ will have been pulled $n_{i,t}$, times where

$$\begin{aligned} E[n_{i,t}] &\geq (\ell_j - 1)\frac{\epsilon_{j-1}}{N} \\ &= \left(\frac{N}{\ell_{j-1}}\log(\ell_{j-1})\right)^{1/3}\frac{\ell_j - 1}{N} \\ &\geq \left(\frac{\ell_{j-1}}{N}\right)^{2/3}\log(\ell_{j-1})^{1/3} \\ &\geq \frac{1}{2}\left(\frac{\ell_j}{N}\right)^{2/3}\log(\ell_{j-1}\log(\ell_j)^{1/3} \\ &= \frac{\log(\ell_j)}{\epsilon_j^2} \end{aligned}$$

Therefore, by Chernoff bounds, in epoch $j$ with probability $1 - O(\frac{1}{\ell_j^2})$, we can distinguish arms with $\Delta_i \geq \epsilon_j$.

Putting this together, with probability $1 - O(\frac{1}{T^2})$, we only pull arms in $N_1$ in epoch $j$, giving

$$E[R_j] \leq \epsilon_j l_j + \frac{1}{\ell_j^2} = O(\log(\ell_j)^{1/3} N^{1/3} {\ell_j}^{2/3})$$

which as noted before, gives the result.

**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.

**Solution .**
We'll prove the two claims as given in the hint.

**Lemma 1.** *1 Best arm $i^*$ will not be eliminated with high probability*

*Proof.* For arm $i^*$ to be eliminated at round t we must have some arm j such that $LCB_{j,t} \geq UCB_{i^*,t}$. By Chernoff bounds both:

$$UCB_{i^*,t} \geq \mu_{i*} \tag{10}$$

and

$$\mu_j \geq LCB_{j,t} \tag{11}$$

hold with probability $\geq 1 - \frac{2}{T^2}$. Which implies the probability of being eliminated in any round is $O(1/T^2)$ and the probability of ever being eliminated is $O(1/T)$. □

**Lemma 2.** *2 Suboptimal arm i will be eliminated after $\frac{\log T}{\Delta_i^2}$ rounds.*

*Proof.* Note in round $k$, the probability of arm $j$ being eliminated is lower bounded by the probability that arm $i^*$ eliminates it. That is $\mathbb{P}$(j eliminated) $\geq \mathbb{P}\left(\hat{\mu_{i^*,t}} - \hat{\mu_{j,t}} \geq 2\sqrt{\frac{\log T}{k}}\right)$. Letting $k$ be $\frac{8 \log T}{\Delta_i^2}$ and recalling $\mathbb{P}(|\hat{\mu_{i,k}} - \mu_i| \geq \Delta_i/2) \leq \frac{2}{T^2}$, implies:

$$LCB_{i^*,t} = \hat{\mu_{i*,t}} - \Delta_j/2 \geq \mu_{i^*,t} - \Delta_j/2 \geq \mu_j + \Delta_j/2 \geq UCB_{j,t} \tag{12}$$

with probability $O(1 - 1/T^2)$ □

Let $B$ be the event that optimal arm $i^*$ ever gets eliminated. Then by conditioning and using the above lemma's:

$$E[R(T)] \leq E[R(T)|B]\mathbb{P}(B) + \sum_{j \neq i^*} \Delta_j E[n_{j,T}] \leq 1 + O\left(\sum_{i \neq i^*} \frac{\log T}{\Delta_i}\right) \tag{13}$$

**Problem 4.** Suppose that the reward $r_t$ generated on playing arm $I_t$ at time $t$ is observed after a delay of $\gamma$, 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.*

**Solution 4.1.** Consider the UCB algorithm with delayed update rule:

$$UCB_{i,t} = \hat{\mu_{i,t-\gamma}} + \sqrt{\frac{\log T}{n_{i,t-\gamma}}} \tag{14}$$

where we first start the algorithm with $\gamma$ pulls of each arm. The regret incurred after the first $N\gamma$ pulls is then less than or equal to the regret of a standard UCB algorithm on $T - N\gamma$ pulls. Thus:

$$E[R_\gamma(T)] \leq \sum_{j \neq i^*} \Delta_j \gamma + E[R_{UCB}(T - N\gamma)] \tag{15}$$

$$= O(\sum_{j \neq i^*} \frac{\log(T - N\gamma)}{\Delta_j} + \Delta_j \gamma) \leq O(\sum_{j \neq i^*} \frac{\log T}{\Delta_j} + \Delta_j \gamma) \tag{16}$$

**Solution 4.2.** We'll use the rounds algorithm given in problem 3 with the UCB modification given above, but now not seeding the algorithm with $N\gamma$ pulls. Instead we'll seed the algorithm with one free pull. Since $\gamma \leq N$,

the UCB estimates are well defined. Let $i$ be the $\ell^{th}$ arm to be eliminated, and $L_i$ be the number of rounds until the arm $i$ is eliminated. If $i$ be the $\ell^{th}$ arm to be eliminated, then number of remaining arms when $i$ is eliminated is $N-\ell$, and we know that until arm $i$ was eliminated there were always atleast $N-\ell$ arms in play (actually more initially). Therefore, if $N-\ell \geq \gamma$, then we know there was always a gap of at least $N-\ell \geq \gamma$ time steps between two plays of arm $i$. Therefore, the delay did not really effect arm $i$'s observations (except for the last one round), and we have $L_i \leq \frac{\log T}{\Delta_i^2} + 1$. But, if $i$ is $\ell^{th}$ arm to be eliminated for $\ell = N-\gamma-1, N-\gamma-2, \ldots, 2$, then the number of remaining arms at time time $i$ was eliminated was $N-\ell < \gamma$, the gap between two plays of arm $i$ could be less than $\gamma$, and therefore, the delay did effect observations from arm $i$. In particular, we may not be able to use samples from its last $\frac{\gamma}{N-\ell}$ plays, giving $L_i = \frac{\log T}{\Delta_i^2} + 1 + \frac{\gamma}{N-\ell}$ for $\ell = N-\gamma-1, N-\gamma-2, \ldots, 2$. Using this observation,

$$E[R(T)] \leq 1 + \sum_{j \neq i^*} E[L_i]\Delta_i \tag{17}$$

$$= 1 + \sum_{i=1}^{N} \Delta_i (\frac{\log T}{\Delta_i^2} + 1) + \sum_{\ell=N-\gamma-1}^{1} ceil(\gamma/(N-\ell)) \tag{18}$$

$$\leq \sum_{i \neq i^*} \frac{\log T}{\Delta_i} + \Delta_i + \sum_{j=1}^{\gamma} \frac{\gamma}{j} \tag{19}$$

$$\leq \sum_{i \neq i^*} \frac{\log T}{\Delta_i} + \Delta_i + \gamma \log(\gamma) \tag{20}$$

**Problem 5.** [Additional feedback] Suppose that we are given an undirected graph $G$, where $N$ nodes of the graph correspond to the $N$arms. 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 $\nu_i$ with mean $\mu_i$. Additionally, for all neighbors $j$ of $i$ in graph $G$, you get to observe $X_{t,j}$ generated i.i.d. from (unknown) distribution $\nu_j$. 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.)

**Solution .** We propose a similar algorithm as in (Q3), proceeding in rounds and eliminating arms in the same fashion. The difference now is in a single round we can collect many observations for an individual arm. Clearly the same proofs for high probability elimination of suboptimal arms and high probability preservation of the optimal arm still hold.

Let $C_i,\ i \in [k]$ be the k cliques that cover G and denote the $j^{th}$ arm in the $i^{th}$ clique as the pair (i,j). Following problem 3, we'll need $O(\frac{\log T}{\Delta_{i,j}^2})$ observations to eliminate an arm with high probability. Fix an ordering in which arms in a clique i are eliminated, then the $(i,j)^{th}$ arm to be eliminated needs only $O(\frac{\log T}{\Delta_{i_j}^2}\frac{1}{|C_i|-j})$ rounds to be eliminated.

As done in earlier problems, let us bound regret only for arms in $N_1$, which is the arms with $\Delta_i \geq \Delta =$

$\sqrt{\frac{k}{T}\log T \log N}$. The total regret from remaining arms is bounded by $\sqrt{kT \log T \log N}$.

$$E[R(T)] = \sum_{i=1}^{N_1} \Delta_i E[n_{i,T}] \tag{21}$$

$$\leq \sum_{i=1}^{k} \sum_{j=1, j \in N_1}^{|C_i|} \frac{\log T}{j\Delta_{i,j}} \tag{22}$$

$$\leq \sum_{i=1}^{k} \sum_{j=1}^{|C_i|} \frac{\log T}{j\Delta} \tag{23}$$

$$= O\left(\Delta T + \frac{k \log N \log T}{\Delta}\right) \tag{24}$$

Where the first inequality follows from our analysis in (Q3) and the discussion above. The second inequality follows from noting that $\Delta_{i,j} \geq \Delta$ for arms in $N_1$. Substituting $\Delta = \sqrt{\frac{k}{T}\log T \log N}$ gives the final result.