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

# Homework 3

Instructor: Shipra Agrawal **Due on: Mar 30, 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.** We have discussed the following linear bandit problem in class: There are $N$ arms, each arm $i$ is associated with a known $d$-dimensional vector $x_i$. At time $t$, you are given a subset of arms $A_t$, you need to pick an arm $I_t \in A_t$. After picking $I_t$, you observe $r_t \in \mathbb{R}$, where $E[r_t|I_t = i] = w^T x_i$ for some unknown fixed $d$-dimensional vector $w$. We defined regret as

$$\sum_{t=1}^{T}(\max_{i\in A_t} w^T x_i) - \sum_{t=1}^{T} w^T x_{I_t},$$

and discussed a (LinUCB) algorithm for this problem which achieves $O(d\sqrt{T\log^2(T/\delta)})$ regret upper bound with probability $1-\delta$, assuming $\|w\| \leq \sqrt{d}, \|x_i\| \leq \sqrt{d}, |r_t| \leq 1$. This regret bound did not depend on $N$, which means number of arms $N$ can be very large or even infinite. A similar algorithm by Chu et al. 2011 (Complete reference below) can achieve $O(\sqrt{dT\log^3(NT\log(T)/\delta)})$ regret with probability $1-\delta$, which is useful when $N$ is large but not extremely large, for example when $N = O(T)$. Use these results to obtain useful regret bounds for following variations of this problem. State any further assumptions you make.

(a) There are $K \leq T$ arms, each arm $i = 1, \ldots, K$ is associated with a known $d$-dimensional vector $x_i$ and an unknown $w_i \in \mathbb{R}^d$. At time $t$, you are given a subset $A_t$ of the $K$ arms, and need to pull an arm $I_t \in A_t$. The expected reward observed on pulling arm $I_t = i$ is $x_i^T w_i$. Regret in time $T$ is defined as

$$\sum_{t=1}^{T}\left(\max_{i\in A_t} w_i^T x_i\right) - \sum_{t=1}^{T} w_{I_t}^T x_{I_t}$$

Use the linear bandit results stated above to obtain $O(\sqrt{KdT\log^3(T\log(T))})$ regret bounds for this problem.

(a) There are $K$ arms $K \leq T$, and an unknown $w_i \in \mathbb{R}^d$ for every arm $i$. At every time $t$, the decision maker observes a $d$-dimensional (context) vector $x_t$, and needs to to pick an arm $I_t \in \{1, \ldots, K\}$. The expected reward on pulling arm $I_t = i$ is $w_i^T x_t$. Regret is defined as:

$$\sum_{t=1}^{T}\left(\max_{i\in\{1,\ldots,K\}} w_i^T x_t\right) - \sum_{t=1}^{T} w_{I_t}^T x_t$$

Use the linear bandit results stated above to obtain $O(\sqrt{KdT\log^3(T\log(T))})$ regret bounds for this problem.

**Reference:** Contextual Bandits with Linear Payoff Functions. Wei Chu. Lihong Li. Lev Reyzin. Robert E. Schapire. AISTATS 2011. http://research-srv.microsoft.com/pubs/178848/camera-ready.pdf

**Problem 2.** iflix has a catalog of 50000 movies and hundreds of thousands of users. Every movie M is described by 5 numerical features (M1, M2, M3, M4, M5). Whenever some user U comes to the website, the user's profile can be observed, which is composed by 5 numerical features (U1, U2, U3, U4, U5). iflix wants to recommend a movie to the user whenever he/she comes to the website. Assume that the probability of user U watching a recommended movie M is a linear function of (M1, M2, M3, M4, M5, U1, U2, U3, U4, U5). Let us call the movie that user U is most likely to watch as "user's favorite movie". Over a course of a day many different users come to the website, iflix wants to recommend movies to maximize total number of movies watched, or to minimize expected regret, where regret is defined as the difference between "Total Number of movies watched" and "Total Number of movies that *would be watched* if every user were recommended their favorite movie". Model this problem as a bandit problem and state regret bounds. State any further assumptions you make.

**Problem 3.** We discussed the following adversarial linear full-information problem (also known as online learning or online linear optimization): every time $t$, play an $x_t \in A$ and observe $w_t$, where $x_t, w_t \in \mathbb{R}^d$, the reward is $x_t^T w_t$. We defined regret as difference between algorithm's reward and reward achieved on playing the best single $x \in A$ at all time steps: $\left(\max_{x \in A} \sum_t x^T w_t\right) - \sum_t x_t^T w_t$. Online gradient ascent algorithm achieved $DG\sqrt{T}$ regret bound for this problem, where $G$ is an upper bound on $\|w_t\|$ and $D$ is a bound on $\|x_t\|$. Now, consider the following convex version of this problem: every time $t$, play $x_t \in A$, and observe a concave function $f_t(\cdot)$, reward is $f_t(x_t)$. This is called online convex optimization. Regret is again defined as difference between algorithm's reward and reward achieved on playing the best single $x \in A$ at all time steps, i.e. regret in time $T$:

$$\left(\max_{x \in A} \sum_{t=1}^{T} f_t(x)\right) - \sum_{t=1}^{T} f_t(x_t).$$

Show that *any* algorithm for online linear optimization can be applied to the gradients of $f_t$ to solve online convex optimization to achieve the same regret bounds. (for example, gradient ascent algorithm would achieve $DG\sqrt{T}$ bounds with $G$ being an upper bound on the magnitude of gradient of $f_t$ and $D$ is a bound on $\|x_t\|$). If needed, you may assume that $f_t$ is a smooth function.

**Problem 4.** A media house wants to conduct a political survey using an online platform which has a random set of students from 100 different institutes. The media house has budget to pay 1000 survey responders. On the online platform the surveyor can see no information about a student other than the institute he/she belongs to. Assume that whenever the surveyor asks a student to take the survey, the student responds, and takes the survey truthfully. Also, there are many more than 1000 students from each institute on the platform. In the survey, the responder is asked to classify his/her political ideology into one of 9 predefined categories. Each institute has a different (fixed unknown) distribution of students over the 9 categories. Use one of the bandit formulations discussed in the course to design an exploration-exploitation strategy that the surveyor could use to conduct the survey, such that in the end the survey is filled by almost equal number of students of each political ideology, if possible. State any further assumptions made.