Homework 3 Solutions

Instructor: Professor Agrawal
TA: Michael Hamilton
IEOR 8100-001

April 14, 2016

Problem 1. Suppose there are N arms, each arm i is associated with a a vector $x_i \in \mathbb{R}^d$. We considered the linear bandit problem of picking arm $I_t \in A_t$ at time t to minimize regret defined as:

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

In class we saw the LinUCB algorithm and showed it achieved regret:

$$R_{LinUCB}(T) = O(d\sqrt{T\log^2(T/\delta)}) \quad w.p.1-\delta \tag{2}$$

when $||W|| \leq \sqrt{d}, ||x_i|| \leq \sqrt{d}, |r_t| \leq 1$, which does not depend on the number of arms N. Chu et. al. study the same problem and achieve following regret that depends on N, the number of arms:

$$R_{Chu}(T) = O(\sqrt{dT\log^3(NT\log(T)/\delta)}) \quad w.p.1-\delta \tag{3}$$

Use these results show:

- (a) There are $K \leq T$ arms, each arm $i \in [K]$ is associated with x_i. At time t, given a subset A_t of the K arms, and need to pull an arm $I_t \in A_t$. Regret is defined as above, use the result of Chu et. al. to achieve regret $O(\sqrt{KdT\log^3(T\log(T)/\delta)})$ w.p $1-\delta$.
- (b) There are $K \leq T$ arms, and an unknown $w_i \in R^d$ for every arm i. At every time t, the decision maker observes a d-dimensional (context) vector x_t, and needs to pick an arm $I_t \in [K]$. Regret is now defined as:

$$\sum_{t=1}^{T}(\max_{i\in[K]} w_i^T x_t) - w_{I_t}^T x_t \tag{4}$$

  use the result of Chu et. al. to achieve regret $O(\sqrt{KdT\log^3(T\log(T)/\delta)})$ w.p $1-\delta$.

Solution a. The idea here will be embed the changing information at each time step as fixed information in a much larger dimension. To this end let

$$w = (w_1^T, w_2^T, \ldots, w_K^T) \in \mathbb{R}^{dK} \tag{5}$$

and let

$$x_i' = (0, \ldots, 0, x_i^T, 0, \ldots, 0) \in \mathbb{R}^{dK} \tag{6}$$

so that $< x_i', w >=< x_i, w_i >$ where $< \cdot, \cdot >$ is the dot product. Now applying the result of Chu et al. with $N = K \leq T$, and dimension dK instead of d gives the desired regret bound (with probability $1 - \delta$).

Solution b. Similar to part (a), define:

$$w = (w_1^T, w_2^T, \ldots, w_K^T) \in \mathbb{R}^{dK} \tag{7}$$

Now define A_t as a set of K arms, one corresponding to each of the vectors $y_1 = (x_t, 0, 0, 0 \ldots, 0)$, $y_2 = (0, x_t, 0, \ldots, 0)$, $y_3 = (0, 0, x_t, 0 \ldots, 0)$, ..., $y_K = (0, 0, 0, \ldots, 0, x_t)$. Now for i^{th} arm in A_t, $< y_i, w >=< x_t, w_i >$. Note that the universe of all arms is much larger than K, it has $K \times T$ arms, one for each $(i, x_t), i = 1, \ldots, K, t = 1, \ldots, T$. A_t contains just K of these.

Then the problem is equivalent to selecting $I_t \in A_t$, to minimize regret:

$$\sum_{t=1}^{T} (\max_{i \in A_t} < w^T, y_i >) - \sum_{t=1}^{T} < w, y_{I_t} > \tag{8}$$

Now the number of arms $= KT$ and the dimension is Kd, applying the result of Chu et al. again gives the desired regret.

Problem 2. 5000 thousands movies, $\{M_i\}_{i=1}^{5000}$ where each film has five numerical features, $M_i = (m_1, m_2, m_3, m_4, m_5)$. At each time step a user $U = (u_1, u_2, u_3, u_4, u_5)$ arrives. Assume there exists some $w \in \mathbb{R}^{10}$ such that:

$$Pr(U \text{ watches } M) = < w, (U, M) > \tag{9}$$

where again $< \cdot, \cdot >$ is the dot product.

Solution 2. We'll model this problem as a linear contextual bandit problem. Let the arrival of any given user be a time step and suppose there are T arrivals in total. At time t, consider the set of arms associated with U_t to be $A_t = \{(M_i, U_t), i \in [5000]\}$. Then regret is defined as:

$$R(T) = \sum_{i=1}^{T} \max_{x_{i^*} \in A_t} < w, x_{i^*} > - \sum_{i=1}^{T} < w, x_{I_T} > \tag{10}$$

Then the LinUCB algorithm (for example), with probability $1 - \delta$, achieves regret

$$R_{LinUCB}(T) \leq O(d\sqrt{(}T \log(T/\delta)) = O(\sqrt{(}T \log(T/\delta)) \tag{11}$$

where the equality follows from noting $d = 10$.

Problem 3. At every time t, play $x_t \in A$ and observe w_t and reward $< x_t, w_t >$. Online gradient descent achieved $DG\sqrt{T}$ regret for this problem, where $||w_t|| \leq G$ and $||x_t|| \leq D$. Consider the convex version of this problem: every time t, play $x_t \in A$ and observed concave function $f_t(\cdot)$ and reward $r_t = f_t(x_t)$. Regret is defined as:

$$R(T) = \left(\max_{x \in A} \sum_{i=1}^{T} f_t(x) \right) - \sum_{t=1}^{T} f_t(x_t) \tag{12}$$

Show any any algorithm for online linear optimization can be applied to the gradients of f_t to solve the online convex optimization problem while achieving the same regret bounds.

Solution . Since f_t is concave, we have:

$$f_t(x_t) - f_t(y) \leq \nabla f_t^T (x^* - x_t) \tag{13}$$

where x^* is the optimal "arm" for the online convex optimization problem. Then the claim immediately follows by defining $\{w_t = \nabla f_t(x^*)\}$, then letting $\{x_t\}$ be the sequence chosen by the arbitrary online linear optimization algorithm:

$$R_{convex}(T) = \left(\max_{x \in A} \sum_{i=1}^{T} f_t(x) \right) - \sum_{t=1}^{T} f_t(x_t) \leq \sum_{i=1}^{T} \nabla f_t(x^*)(x^* - x_t) = R_{lin}(T) \tag{14}$$

Problem 4. A media house wants to conduct a political survey using an online platform with students from A = 100 different institutes (where A is the index set). The media house has a budget of T = 1000 responses. At time t a school is chosen and a student drawn from that school responds with one of nine possible ideologies which is observed by the surveyor. Let $K_i \, i \in [9]$ be the number of students who respond a certain way by the end of study. Goal: $\min \max K_i$

Solution . This problem can be solved in a number of ways. Here's one approach: Think of each school as a distribution over the nine outcomes, and keep an empirical estimate for each school. We choose one of these distributions at each time t and observe reward vector $(0, \ldots, 1, \ldots, 0) \in \mathbb{R}$, a unit vector. Note we want the sum of rewards to be a uniform vector and also note a uniform vector minimizes the L_2 norm. Let r_t is the reward vector generated on picking school I_t at time t, then one way to formulate the problem is to choose I_ts online to minimize global convex function $|| \sum_{t=1}^{T} r_t||_2^2$. We can use the algorithm for maximizing a gloabl concave reward function with bandit vector feedback as seen in class.