COMS E6998.001 "Bandits & RL" (Fall'17) Instructors: Alekh Agarwal & Alex Slivkins

Homework 2: adversarial bandits

This homework would not be collected or graded, and would not affect your grade. We encourage you to try it to solidify your understanding of the course material.

Please feel free to refer to the the book draft, and to discuss solutions with others. All problems can be solved by a fairly basic application of concepts covered in class.

Preliminaries. We will use notation from the class. T is the time horizon, K is the number of arms. At each round t, a_t is the arm chosen by the algorithm, and $c_t(a)$ is the cost of arm a. The total cost of arm a is $\texttt{cost}(a) = \sum_{t=1}^{T} c_t(a)$. The total cost of an algorithm ALG is $\texttt{cost}(\texttt{ALG}) = \sum_{t=1}^{T} c_t(a_t)$. Regret is defined as $R(T) = \texttt{cost}(\texttt{ALG}) - \texttt{cost}^*$, where $\texttt{cost}^* = \min_{\text{arms } a} \texttt{cost}(a)$.

In all problems, assume that the costs are chosen by oblivious adversary.

Problem 1. Consider binary prediction with expert advice, with a perfect expert. Prove that any algorithm makes at least $\Omega(\min(T, \log K))$ mistakes in the worst case.

Take-away: The majority vote algorithm is worst-case-optimal for instances with a perfect expert.

Hint: For simplicity, let $K = 2^d$ and $T \geq d$, for some integer d. Construct a distribution over problem instances such that each algorithm makes $\Omega(d)$ mistakes in expectation. Recall that each expert e corresponds to a binary sequence $e \in \{0,1\}^T$, where e_t is the prediction for round t. Put experts in 1-1 correspondence with all possible binary sequences for the first d rounds. Pick the "perfect expert" u.a.r. among the experts.

Problem 2 (i.i.d. costs and hindsight regret). Consider online learning with experts, for the special case of i.i.d. costs.

(a) Prove that $\min_a \mathbb{E}[\texttt{cost}(a)] \leq \mathbb{E}[\min_a \texttt{cost}(a)] + O(\sqrt{T \log(KT)})$.

Take-away: All $\sqrt{T}$ upper regret bounds from Lecture 2 carry over to "hindsight regret".

Hint: Consider the "clean event": namely, that the event inside the Hoeffding inequality holds for the cost sequence of each arm.

(b) Prove that there is a problem instance with a deterministic adversary for which any algorithm suffers regret

$$\mathbb{E}[\texttt{cost}(\texttt{ALG}) - \min_{a \in [K]} \texttt{cost}(a)] \geq \Omega(\sqrt{T \log K}).$$

Hint: Assume all arms have 0-1 costs with mean $\frac{1}{2}$. Use the following standard fact:

$$\mathbb{E}[\min_a \texttt{cost}(a)] \leq \tfrac{T}{2} - \Omega(\sqrt{T \log K}). \tag{1}$$

(This is in fact a fact about random walks.)

Note: This example does not carry over to "foresight regret". Indeed, each arm has expected reward of $\frac{1}{2}$ in each round, so any algorithm trivially achieves 0 "foresight regret".

Take-away: The $O(T \log K)$ regret bound for `Hedge` is the best possible for hindsight regret. Further, $\log(T)$ upper regret bounds for foresight regret do not carry over to hindsight regret in full generality.

(c) Prove that algorithms UCB1 and Successive Elimination achieve the same logarithmic regret bound (Theorem 2.9 in the book) for hindsight regret, if the best-in-foresight arm a^* is unique.

Hint: Under the "clean event", $\texttt{cost}(a) < T \cdot \mu(a) + O(\sqrt{T \log T})$ for each arm $a \neq a^*$, where $\mu(a) = \mathbb{E}[c_t(a)]$ is the mean per-round cost. Therefore, a^* is also the best-in-hindsight arm, unless $\mu(a) - \mu(a^*) < O(\sqrt{T \log T})$ for some arm $a \neq a^*$ (in which case the claimed regret bound holds trivially).

Problem 3. Prove that any deterministic algorithm for the online learning problem with K experts and 0-1 costs suffers total cost T for some deterministic-oblivious adversary, even if $\texttt{cost}^* \leq T/K$.

Take-away: With a deterministic algorithm, cannot even extend the guarantee for WMA (Theorem 6.8 in the book) to the general case of online learning with experts, let alone have $o(T)$ regret.

Hint: Fix the algorithm. Construct the problem instance by induction on round t, so that the chosen arm has cost 1 and all other arms have cost 0.

Problem 4 (lower bound). Consider adversarial bandits with experts advice. For any given (K, N, T), construct a randomized problem instance for which any algorithm satisfies

$$\mathbb{E}[R(T)] \geq \Omega\left(\sqrt{KT \log(N)/\log(K)}\right). \tag{2}$$

Hint: Split the time interval $1..T$ into $M = \frac{\ln N}{\ln K}$ non-overlapping sub-intervals of duration T/M . For each sub-interval, construct the randomized problem instance from Chapter **??** (independently across the sub-intervals). Each expert recommends the same arm within any given sub-interval; the set of experts includes all experts of this form.

Problem 5 (slowly changing costs). Consider a randomized oblivious adversary such that the expected cost of each arm changes by at most ϵ from one round to another, for some fixed and known $\epsilon > 0$. Use algorithm `Exp4` to obtain dynamic regret

$$\mathbb{E}[R^*(T)] \leq O(T) \cdot (\epsilon K \log KT)^{1/3}. \tag{3}$$

Note: Regret bounds for dynamic regret are typically of the form $\mathbb{E}[R^*(T)] \leq C \cdot T$, where C is a "constant" determined by K and the parameter(s). The intuition here is that the algorithm pays a constant per-round "price" for keeping up with the changing costs. The goal here is to make C smaller, as a function of K and ϵ.

Hint: Recall the application of `Exp4` to n-shifting regret, denote it `Exp4`(n). Let $\texttt{OPT}_n = \min \texttt{cost}(\pi)$, where the min is over all n-shifting policies π, be the benchmark in n-shifting regret. Analyze the "discretization error": the difference between $\texttt{OPT}_n$ and $\texttt{OPT}^* = \sum_{t=1}^{T} \min_a c_t(a)$, the benchmark in dynamic regret. Namely: prove that $\texttt{OPT}_n - \texttt{OPT}^* \leq O(\epsilon T^2/n)$. Derive an upper bound on dynamic regret that is in terms of n. Optimize the choice of n.