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

Homework 1: multi-armed bandits with IID rewards

This homework would not be collected or graded, and would not affect your grade. However, We encourage you to try it to solidify your understanding of the course material. Also, you’d be exposed to “instantaneous regret” and “doubling trick”, important concepts only briefly mentioned in lecture.

Please feel free to refer to the the book draft, and to discuss solutions with others. All problems except (2b) can be solved by a fairly basic application of concepts covered in class. Problem (2b) requires a careful argument; try it if you feel adventurous!

In problems 1-4, clearly identify the ”clean event”. It should be possible to use the same “clean event” for all/most problems.

Please alert us if you see bugs or typos!

Notation. We will use some notation from the class. T is the time horizon, K is the number of arms, a_t is the arm chosen at time t, μ^* is the expected reward of the best arm. For each arm a, $\mu(a)$ is the expected reward, and $\Delta(a) = \mu^* - \mu(a)$ is the “badness”.

Problem 1: rewards from a small interval. Consider a version of the problem in which all the realized rewards are in the interval $[\frac{1}{2}, \frac{1}{2} + \epsilon]$ for some $\epsilon \in (0, \frac{1}{2})$. Define versions of UCB1 and Successive Elimination attain improved regret bounds (both logarithmic and root-T) that depend on the ϵ.

Hint: Use a more efficient version of Hoeffding Inequality from the “probability and concentration recap” slides from the first lecture. No need to repeat all steps from the analysis in class as long as you understand which steps in the analysis are changed.

Problem 2: instantaneous regret. Recall: *instantaneous regret* at time t is defined as $\Delta(a_t)$.

(a) Prove that Successive Elimination achieves “instance-independent” regret bound of the form

$$\mathbb{E}[\Delta(a_t)] \leq \frac{\text{polylog}(T)}{\sqrt{t/K}} \quad \text{for each round } t \in [T]. \tag{1}$$

(b) Let us argue that UCB1 does not achieve the regret bound in (1). More precisely, let us consider a version of UCB1 with $UCB_t(a) = \bar{\mu}_t(a) + 2 \cdot r_t(a)$, where $\bar{\mu}_t(a)$ and $r_t(a)$ are as defined in class. (It is easy to see that the analysis from the class carries over to this version.) Focus on two arms, and prove that this algorithm cannot achieve a regret bound of the form

$$\mathbb{E}[\Delta(a_t)] \leq \frac{\text{polylog}(T)}{t^\gamma},\ \gamma > 0 \quad \text{for each round } t \in [T]. \tag{2}$$

Hint: Fix reward function μ. Focus on the clean event. If (2) holds, then the bad arm cannot be played after some time T_0. Consider the last time the bad arm is played, call it $t_0 \leq T_0$. Derive a lower bound on the UCB of the best arm at t_0 (stronger lower bound than the one proved in class). Consider what this lower implies for the UCB of the bad arm at time t_0. Observe that eventually, after some number of plays of the best arm, the bad arm will be chosen again, assuming a large enough time horizon T. Derive a contradiction with (2).

Take-away: for “bandits with predictions”, the simple solution of predicting the last-played arm to be the best arm does not always work, even for a good algorithm such as UCB1.

(c) Derive a regret bound for Explore-first with N steps of exploration, namely: an "instance-independent" upper bound on the instantaneous regret. (There are two cases: $t \leq N$ and $t > N$, the first case being trivial.)

Problem 3: bandits with predictions. Recall that in "bandits with predictions", after T rounds the algorithm outputs an arm y_T: a prediction for the best arm. We focus on the instantaneous regret $\Delta(y_T)$ for the predicted arm.

(a) Take any bandit algorithm with an instance-independent regret bound $E[R(T)] \leq f(T)$, and construct an algorithm for "bandits with predictions" such that $\mathbb{E}[\Delta(y_T)] \leq f(T)/T$.

Note: Surprisingly, taking $y_T = a_t$ does not work in general, see problem 2(b). Taking y_T to be the arm with a maximal reward does not seem to work, either: more precisely, we are not aware of a way to complete the proof. But there is a simple solution ...

Take-away: We can easily obtain $\mathbb{E}[\Delta(y_T)] = O(\sqrt{K \log(T)/T}$ from standard algorithms such as UCB1 and Successive Elimination. However, as parts (bc) show, one can do much better!

(b) Consider Successive Elimination with $y_T = a_T$. Prove that (with a slightly modified definition of the confidence radius) this algorithm can achieve

$$\mathbb{E}[\Delta(y_T)] \leq T^{-\gamma} \quad \text{if } T > T_{\mu,\gamma},$$

where $T_{\mu,\gamma}$ depends only on the mean rewards $(\mu(a) : a \in \mathcal{A})$ and the γ. This holds for an arbitrarily large constant γ, with only a multiplicative-constant increase in regret.

Hint: Put the γ inside the confidence radius, so as to make the "failure probability" sufficiently low. Argue that in the clean event, when T is large enough, only the best arm remains.

(c) Consider a very simple algorithm which alternates the arms in a round-robin fashion, and chooses an arm with the largest average reward for the prediction y_T. Prove that such algorithm achieves:

$$\mathbb{E}[\Delta(y_T)] \leq e^{-\Omega(T)} \quad \text{if } T > T_{\mu},$$

where T_μ depends only on the mean rewards $(\mu(a) : a \in \mathcal{A})$.

Hint: Consider Hoeffding Inequality with an arbitrary constant α in the confidence radius. Pick α as a function of the time horizon T so that the failure probability is as small as needed. Let Δ be the smallest non-zero $\Delta(a)$ among all arms, and pick T large enough so that the confidence radius in Hoeffding Inequality becomes smaller than, say, $\Delta/2$. In the definition of the clean event, no need to take a union bound over all rounds $t < T$.

Problem 4: doubling trick. Take any bandit algorithm $\mathcal{A}$ for fixed time horizon T. Convert it to an algorithm $\mathcal{A}_\infty$ which runs forever, in phases $i = 1, 2, 3, \ldots$ of 2^i rounds each. In each phase i algorithm $\mathcal{A}$ is restarted and run with time horizon 2^i.

(a) State and prove a theorem which converts an instance-independent upper bound on regret for $\mathcal{A}$ into similar bound for $\mathcal{A}_\infty$ (so that this theorem applies to both UCB1 and Explore-first).

(b) Do the same for $\log(T)$ instance-dependent upper bounds on regret. (Then regret increases by a $\log(T)$ factor.)

Note: Consider a regret bound of the form $C \cdot f(T)$, where $f(\cdot)$ does not depend on mean rewards μ and C does not depend on T. Such regret bound is called *instance-independent* if C does not depend on μ, and *instance-dependent* otherwise.

Problem 5: lower bound for non-adaptive exploration. Consider an algorithm such that:

- in the first N rounds ("exploration phase") the choice of arms does not depend on the observed rewards, for some N that is fixed before the algorithm starts;
- in all remaining rounds, the algorithm only uses rewards observed during the exploration phase.

Focus on the case of two arms, and prove that any such algorithm must have regret $\mathbb{E}[R(T)] \geq \Omega(T^{2/3})$ in the worst case.

Hint: Regret is a sum of regret from exploration, and regret from exploitation. For "regret from exploration", we can use two instances: $(\mu_1, \mu_2) = (1, 0)$ and $(\mu_1, \mu_2) = (0, 1)$, *i.e.,* one arm is very good and another arm is very bad. For "regret from exploitation" we can invoke the impossibility result for "bandits with predictions" (Corollary 3.11 in Chapter 3).

Take-away: Regret bound for Explore-First cannot be substantially improved. Further, allowing Explore-first to pick different arms in exploitation does not help.