Homework 3

Instructor: Haipeng Luo

1. (**Improved Blackbox MAB**) In Lecture 12 we discussed how to turn an arbitrary expert algorithm into a multi-armed bandit algorithm in a blackbox way, albeit with suboptimal regret $\mathcal{O}(T^{3/4})$ (ignoring dependence on K). In this exercise you need to prove that the regret can be improved if the expert algorithm enjoys a "small-loss" bound. Specifically, suppose we have a K-expert algorithm that at time t predicts $\widehat{p}_t$ and takes loss vector $c_t \in [0,1]^K$ as input, so that for any $a \in [K]$,

$$\sum_{t=1}^{T} \langle \widehat{p}_t, c_t \rangle - \sum_{t=1}^{T} c_t(a) \leq \mathcal{O}\left(\sqrt{\left(\sum_{t=1}^{T} c_t(a)\right) \ln K}\right).$$

Construct a multi-armed bandit algorithm using such expert algorithm so that its expected regret is bounded by $\mathcal{O}(T^{2/3}(K \ln K)^{1/3})$, ignoring other terms that have smaller dependence on T. (Hint: some uniform exploration is needed as discussed in Lecture 12).

2. (**Losses v.s Gains**) We have been doing loss-based learning throughout the semester. Sometimes it is more natural to do gain/reward-based learning. Take multi-armed bandit as an example. A gain-based setting would be: for each time $t = 1, \ldots, T$,

(1) the learner picks an action $a_t \in [K]$ while simultaneously the environment decides the gain vector $g_t \in [0,1]^K$,

(2) the learner receives and observes (only) the gain $g_t(a_t)$.

As usual we assume the environment is oblivious and the expected regret now becomes

$$\mathbb{E}\left[\mathcal{R}_T\right] = \max_{a \in [K]} \sum_{t=1}^{T} g_t(a) - \mathbb{E}\left[\sum_{t=1}^{T} g_t(a_t)\right].$$

(a) A direct generalization of Exp3 from the loss-based setting to the gain-based setting would be to pick a_t randomly according to a distribution p_t such that $p_t(a) \propto \exp(\eta \sum_{\tau=1}^{t-1} \widehat{g}_\tau(a))$, where $\widehat{g}_t$ is the importance weighted estimator at time t, that is, $\widehat{g}_t(a) = \frac{g_t(a)}{p_t(a)} \mathbf{1}\{a = a_t\}$. State informally why this algorithm should not work (Hint: think about how the weights change after picking an action).

(b) One can fix the above proposal by simply mixing a small amount of uniform exploration. Specifically, let $\widehat{p}_t$ be such that $\widehat{p}_t(a) \propto \exp(\eta \sum_{\tau=1}^{t-1} \widehat{g}_\tau(a))$. The algorithm selects a_t according to $p_t = (1-\alpha)\widehat{p}_t + \frac{\alpha}{K}\mathbf{1}$ where $\mathbf{1}$ is the all-one vector and construct estimator $\widehat{g}_t(a) = \frac{g_t(a)}{p_t(a)} \mathbf{1}\{a = a_t\}$. Prove that with $\alpha = \eta K$ and $\eta = \min\{\sqrt{\frac{\ln K}{TK}}, \frac{1}{2K}\}$, we have

$$\mathbb{E}\left[\mathcal{R}_T\right] \leq \mathcal{O}\left(\sqrt{TK \ln K} + K \ln K\right).$$

(Hint: redo the Hedge analysis in the gain-based setting, see where the proof breaks and why the uniform mixing helps. You will need to use the inequality $e^y \leq 1 + y + y^2$ for all $y \leq 1$.)

3. Prove that for *any* $\Delta \in [0,1]$, UCB has the following pseudo-regret bound:

$$\bar{\mathcal{R}}_T \leq \Delta T + \sum_{a : \Delta_a > \Delta} \left(\frac{16 \ln T}{\Delta_a} + 2\Delta_a\right).$$

where $\Delta_a = \mu(a) - \mu(a^\star)$ is the suboptimal gap (see Lecture 14). Based on this observation, further prove that UCB has worst-case regret bound $\bar{\mathcal{R}}_T = \mathcal{O}(\sqrt{TK \ln T} + K \ln T)$.

4. (**Small-loss Bounds for MAB**) While getting small-loss bounds is relatively easy in the full information setting, there are so far only two known techniques for getting small-loss bounds for the multi-armed bandit problem. You will analyze both of them in this exercise.

(a) The first approach is Online Mirror Descent:

$$\begin{aligned} \nabla\psi(p'_{t+1}) &= \nabla\psi(p_t) - \eta\widehat{\ell}_t \\ p_{t+1} &= \operatorname*{argmin}_{p\in\Delta(K)} D_\psi(p, p'_{t+1}) \end{aligned}$$

with a special regularizer $\psi(p) = -\sum_{a=1}^K \ln p(a)$ ($\widehat{\ell}_t$ is the usual importance weighted estimator and D_ψ is the Bregman divergence). Recall that we have shown the following general regret bound for OMD: $\forall q \in \Delta(K)$,

$$\sum_{t=1}^T \left\langle p_t - q, \widehat{\ell}_t \right\rangle \le \frac{D_\psi(q, p_1)}{\eta} + \frac{1}{\eta}\sum_{t=1}^T D_\psi(p_t, p'_{t+1}).$$

(i) Prove that $D_\psi(p_t, p'_{t+1}) \le \eta^2 \sum_{a=1}^K p_t(a)^2 \widehat{\ell}_t(a)^2$. (Hint: use the inequality $x - x^2 \le \ln(1+x)$ for all $x \ge 0$.)

(ii) Further show that $D_\psi(p_t, p'_{t+1}) \le \eta^2 \ell_t(a_t)$.

(iii) Conclude the following small-loss regret bound for any action $a^\star$ by picking specific q and η:

$$\mathbb{E}\left[\sum_{t=1}^T \ell_t(a_t) - \sum_{t=1}^T \ell_t(a^\star)\right] = \mathcal{O}\left(\sqrt{\left(\sum_{t=1}^T \ell_t(a^\star)\right) K \ln T} + K \ln T\right).$$

(b) The second approach is based on Exp3 with a weight clipping trick. Specifically, at time t first compute $\widehat{p}_t(a) \propto \exp\left(-\eta \sum_{\tau=1}^{t-1} \widehat{\ell}_\tau(a)\right)$ as Exp3. Next zero out the weights that are smaller than some threshold $\gamma \in (0, 1)$ and renormalize, that is, compute

$$p_t(a) \propto \mathbf{1}\{\widehat{p}_t(a) \ge \gamma\}\widehat{p}_t(a).$$

Finally sample $a_t \sim p_t$ and construct estimator $\widehat{\ell}_t(a) = \frac{\ell_t(a)}{p_t(a)}\mathbf{1}\{a = a_t\}$. Recall that by the Hedge analysis we have for any $a^\star$,

$$\sum_{t=1}^T \left\langle \widehat{p}_t, \widehat{\ell}_t \right\rangle - \sum_{t=1}^T \widehat{\ell}(a^\star) \le \frac{\ln K}{\eta} + \eta \sum_{t=1}^T \sum_{a=1}^K \widehat{p}_t(a)\widehat{\ell}_t(a)^2$$

(i) Prove that $\mathbb{E}_{a_t\sim p_t}[\widehat{\ell}_t(a)] \le \ell_t(a)$ for any a (which means that the estimator is no longer unbiased) and also

$$1 - K\gamma \le \frac{\widehat{p}_t(a_t)}{p_t(a_t)} \le 1.$$

(ii) Prove that $\left\langle \widehat{p}_t, \widehat{\ell}_t \right\rangle \ge (1 - K\gamma)\ell_t(a_t)$ and $\widehat{p}_t(a)\widehat{\ell}_t(a)^2 \le \widehat{\ell}_t(a)$.

(iii) Prove that for any two actions $a, a' \in [K]$, their cumulative estimated losses are close in the following sense:

$$\sum_{t=1}^T \widehat{\ell}_t(a) \le \sum_{t=1}^T \widehat{\ell}_t(a') + \frac{1}{\gamma} + \frac{1}{\eta}\ln\frac{1}{\gamma}.$$

(Hint: let $T_a = \max\{t : \widehat{\ell}_t(a) \ne 0\}$ be the last time when $\widehat{\ell}_t(a)$ is non-zero, and first see how close $\sum_{t=1}^{T_a} \widehat{\ell}_t(a)$ and $\sum_{t=1}^{T_a} \widehat{\ell}_t(a')$ are.)

(iv) Combine everything to show that for any $a^\star$,

$$(1 - K\gamma) \sum_{t=1}^{T} \left(\ell_t(a_t) - \widehat{\ell}_t(a^\star) \right) \leq \frac{\ln K}{\eta} + (\gamma + \eta) K \sum_{t=1}^{T} \widehat{\ell}_t(a^\star) + \frac{\eta K}{\gamma} + K \ln \frac{1}{\gamma}.$$

Conclude the final small-loss regret bound by choosing specific η and γ:

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t) - \sum_{t=1}^{T} \ell_t(a^\star) \right] = \mathcal{O}\left(\sqrt{\left(\sum_{t=1}^{T} \ell_t(a^\star) \right) K \ln K} + K \ln \left(K \sum_{t=1}^{T} \ell_t(a^\star) \right) \right).$$