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 $O(T^{3/4})$ (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 $\hat{p}_t$ and takes loss vector $c_t \in [0,1]^K$ as input, so that for any $a \in [K]$, 

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

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

2. (Losses vs. 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

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

(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 $p_t(a) \propto \exp(\eta \sum_{\tau=1}^{t-1} \hat{g}_\tau(a))$, where $\hat{g}_\tau$ is the importance weighted estimator at time $\tau$, that is, $\hat{g}_\tau(a) = \frac{\hat{n}_\tau(a)}{\hat{n}_\tau} \{a = a_\tau\}$. 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 $\tilde{p}_t$ be such that $\tilde{p}_t(a) \propto \exp(\eta \sum_{\tau=1}^{t-1} \hat{g}_\tau(a))$. The algorithm selects $a_t$ according to $p_t = (1 - \alpha)\tilde{p}_t + \frac{\alpha}{K} \mathbf{1}$ where $\mathbf{1}$ is the all-one vector and construct estimator $\hat{g}_t(a) = \frac{\hat{n}_t(a)}{\tilde{n}_t(a)} \mathbf{1}\{a = a_t\}$. Prove that with $\alpha = \eta K$ and $\eta = \min\{\sqrt{\frac{\ln K}{TK}}, \frac{1}{2K}\}$, we have

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

(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:

$$ \bar{R}_T \leq \Delta T + \sum_{a: \Delta_a > \Delta} \left( \frac{16 \ln T}{\Delta_a} + 2\Delta_a \right). $$
where $\Delta_a = \mu(a) - \mu(a^*)$ is the suboptimal gap (see Lecture 14). Based on this observation, further prove that UCB has worst-case regret bound $\widehat{R}_T = O(\sqrt{TK \ln T} + K \ln T)$.

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:

$$\nabla \psi(p_{t+1}) = \nabla \psi(p_t) - \eta \hat{e}_t$$

$$p_{t+1} = \arg\min_{p \in \Delta(K)} D_\psi(p, p_{t+1})$$

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

$$\frac{1}{\eta} \sum_{t=1}^T \left< p_t - q, \hat{e}_t \right> \leq \frac{D_\psi(q, p_1)}{\eta} + \frac{1}{\eta} \sum_{t=1}^T D_\psi(p_t, p_{t+1}).$$

(i) Prove that $D_\psi(p_t, p_{t+1}) \leq \eta^2 \sum_{a=1}^K p_t(a)\hat{e}_t(a)^2$. (Hint: use the inequality $x - x^2 \leq \ln(1 + x)$ for all $x \geq 0$.)

(ii) Further show that $D_\psi(p_t, p_{t+1}) \leq \eta^2 \ell_t(a_t)$. 

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

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

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

$$p_t(a) \propto 1\{\hat{p}_t(a) \geq \gamma\} \hat{p}_t(a).$$

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

$$\sum_{t=1}^T \left< \hat{p}_t, \hat{e}_t \right> - \sum_{t=1}^T \hat{e}(a^*) \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^T \sum_{a=1}^K (\hat{p}_t(a)\hat{e}_t(a))^2$$

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

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

(ii) Prove that $\left< \hat{p}_t, \hat{e}_t \right> \geq (1 - K\gamma)\ell_t(a_t)$ and $\hat{p}_t(a)\hat{e}_t(a)^2 \leq \ell_t(a)$. 

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

$$\sum_{t=1}^T \hat{e}_t(a) \leq \sum_{t=1}^T \ell_t(a') + \frac{1}{\gamma} + \frac{1}{\eta} \ln \frac{1}{\gamma}.$$
(iv) Combine everything to show that for any $a^*$,

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

Conclude the final small-loss regret bound by choosing specific $\eta$ and $\gamma$:

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