CSCI 659 Homework 3

Fall 2022

Instructor: Haipeng Luo

This homework is due on **11/27, 11:59pm**. *See course website for more instructions on finishing and submitting your homework as well as the late policy. Total points: 60.*

1. (**Improved Analysis of FTRL for Bandits**) Consider the FTRL algorithm

$$p_t = \operatorname*{argmin}_{p \in \Delta(K)} \left\langle p, \sum_{s<t} \widehat{\ell}_s \right\rangle + \frac{1}{\eta}\psi(p) \tag{1}$$

where $\eta > 0$ is a learning rate, ψ is the Tsallis entropy $\psi(p) = \frac{1-\sum_{a=1}^K p(a)^\beta}{1-\beta}$ with a parameter $\beta \in (0,1)$, and $\widehat{\ell}_1, \ldots, \widehat{\ell}_T$ are arbitrary loss vectors. In Theorem 3 of Lecture 6, we prove a local-norm bound for this algorithm by showing the key step

$$\left\langle p_t - p_{t+1}, \widehat{\ell}_t \right\rangle - \frac{1}{\eta} D_\psi(p_{t+1}, p_t) \le \frac{\eta}{2} \|\widehat{\ell}_t\|^2_{\nabla^{-2}\psi(p_t)} = \frac{\eta}{2\beta} \sum_{a=1}^K p_t(a)^{2-\beta} \widehat{\ell}_t(a)^2 \tag{2}$$

as long as $\widehat{\ell}_t(a) \ge 0$. In this exercise, you need to prove the same statement (up to a constant of 2) under the weaker condition:

$$\eta p_t(a)^{1-\beta} \widehat{\ell}_t(a) \ge \frac{\beta}{1-\beta}\left(e^{\frac{\beta-1}{\beta}} - 1\right), \quad \forall t \in [T], a \in [K] \tag{3}$$

(it is weaker because the right-hand side is a negative number). Note that when $\beta \to 1$, this reduces to the condition $\eta \widehat{\ell}_t(a) \ge -1$ that we have seen for Hedge/Exp3. (While technical, this exercise will be helpful for Problems 2 and 3.)

(a) (3pts) The first step is still to bound $\langle p_t - p_{t+1}, \widehat{\ell}_t \rangle - \frac{1}{\eta} D_\psi(p_{t+1}, p_t)$ by $\langle p_t - q_t, \widehat{\ell}_t \rangle - \frac{1}{\eta} D_\psi(q_t, p_t)$ where $q_t = \operatorname{argmax}_{q \in \mathbb{R}_+^K} \langle p_t - q, \widehat{\ell}_t \rangle - \frac{1}{\eta} D_\psi(q, p_t)$. Prove that under condition (3), we have

$$\nabla\psi(q_t) = \nabla\psi(p_t) - \eta \widehat{\ell}_t, \tag{4}$$

or equivalently for all a,

$$\frac{1}{q_t(a)^{1-\beta}} = \frac{1}{p_t(a)^{1-\beta}} + \frac{1-\beta}{\beta} \eta \widehat{\ell}_t(a). \tag{5}$$

Proof. The gradient of the concave function $f(q) = \langle p_t - q, \widehat{\ell}_t \rangle - \frac{1}{\eta} D_\psi(q, p_t)$ is $-\widehat{\ell}_t - \frac{1}{\eta}(\nabla\psi(q) - \nabla\psi(p_t))$. If we can find a point $q^\star \in \mathbb{R}_+^K$ such that $\nabla f(q^\star) = \mathbf{0}$, then it must be the maximizer q_t, in which case both Eq. (4) and Eq. (5) are simple rewriting of $\nabla f(q_t) = \mathbf{0}$. Indeed, such $q^\star$ exists since the right-hand side of Eq. (5) is

$$\frac{1}{p_t(a)^{1-\beta}} \left(1 + \frac{1-\beta}{\beta} \eta p_t(a)^{1-\beta} \widehat{\ell}_t(a)\right) \ge \frac{e^{\frac{\beta-1}{\beta}}}{p_t(a)^{1-\beta}} > 0,$$

where the first inequality uses condition (3). $\square$

(b) (4pts) Use Eq. (4) to prove

$$\left\langle p_t - q_t, \widehat{\ell}_t \right\rangle - \frac{1}{\eta} D_\psi(q_t, p_t) = \frac{1}{\eta} D_\psi(p_t, q_t),$$

and use Eq. (5) to further prove

$$D_\psi(p_t, q_t) = \sum_{a=1}^{K} \left(q_t(a)^\beta - p_t(a)^\beta + \eta p_t(a) \widehat{\ell}_t(a) \right).$$

Proof. Both are by direct calculations:

$$\begin{aligned}
&\left\langle p_t - q_t, \widehat{\ell}_t \right\rangle - \frac{1}{\eta} D_\psi(q_t, p_t) \\
&= \frac{1}{\eta} \left\langle p_t - q_t, \nabla\psi(p_t) - \nabla\psi(q_t) \right\rangle - \frac{1}{\eta} D_\psi(q_t, p_t) && \text{(Eq. (4))} \\
&= \frac{1}{\eta} \left\langle p_t - q_t, \nabla\psi(p_t) - \nabla\psi(q_t) \right\rangle - \frac{1}{\eta} \left(\psi(q_t) - \psi(p_t) - \left\langle \nabla\psi(p_t), q_t - p_t \right\rangle\right) \\
& && \text{(definition of Bregman divergence)} \\
&= \frac{1}{\eta} \left(\psi(p_t) - \psi(q_t) - \left\langle \nabla\psi(q_t), p_t - q_t \right\rangle\right) = \frac{1}{\eta} D_\psi(p_t, q_t),
\end{aligned}$$

and

$$\begin{aligned}
&D_\psi(p_t, q_t) \\
&= \psi(p_t) - \psi(q_t) - \left\langle \nabla\psi(q_t), p_t - q_t \right\rangle \\
&= \frac{1}{1-\beta} \sum_{a=1}^{K} \left(q_t(a)^\beta - p_t(a)^\beta + \beta q_t(a)^{\beta-1} (p_t(a) - q_t(a)) \right) \\
&= \frac{1}{1-\beta} \sum_{a=1}^{K} \left((1-\beta) q_t(a)^\beta - p_t(a)^\beta + \beta q_t(a)^{\beta-1} p_t(a) \right) \\
&= \frac{1}{1-\beta} \sum_{a=1}^{K} \left((1-\beta) q_t(a)^\beta - p_t(a)^\beta + \beta \left(\frac{1}{p_t(a)^{1-\beta}} + \frac{1-\beta}{\beta} \eta \widehat{\ell}_t(a) \right) p_t(a) \right) \\
& && \text{(Eq. (5))} \\
&= \sum_{a=1}^{K} \left(q_t(a)^\beta - p_t(a)^\beta + \eta p_t(a) \widehat{\ell}_t(a) \right).
\end{aligned}$$

□

(c) (4pts) Use Eq. (5) and the fact $(1+x)^\alpha \le 1 + \alpha x + \alpha(\alpha-1)x^2$ for any $\alpha < 0$ and $x \ge e^{1/\alpha} - 1$ to prove that the following holds under condition (3):

$$q_t(a)^\beta - p_t(a)^\beta + \eta p_t(a) \widehat{\ell}_t(a) \le \frac{\eta^2}{\beta} p_t(a)^{2-\beta} \widehat{\ell}_t(a)^2.$$

(Hint: you will need to apply the fact with $\alpha = \frac{\beta}{\beta-1}$.)

Proof. Using Eq. (5), we rewrite $q_t(a)^\beta$ as

$$q_t(a)^\beta = p_t(a)^\beta \left(1 + \frac{1-\beta}{\beta} \eta p_t(a)^{1-\beta} \widehat{\ell}_t(a) \right)^{\frac{\beta}{\beta-1}}.$$

Then we apply the provided inequality with $\alpha = \frac{\beta}{\beta-1} < 0$ and $x = \frac{1-\beta}{\beta} p_t(a)^{1-\beta} \eta \widehat{\ell}_t(a)$, which is at least $e^{1/\alpha} - 1$ under condition (3). This shows

$$\begin{aligned}
q_t(a)^\beta &\le p_t(a)^\beta \left(1 - \eta p_t(a)^{1-\beta} \widehat{\ell}_t(a) + \frac{\eta^2}{\beta} p_t(a)^{2-2\beta} \widehat{\ell}_t(a)^2 \right) \\
&= p_t(a)^\beta - \eta p_t(a) \widehat{\ell}_t(a) + \frac{\eta^2}{\beta} p_t(a)^{2-\beta} \widehat{\ell}_t(a)^2,
\end{aligned}$$

which proves the statement. □

(d) (3pts) Combining the three steps above, we have shown

$$\left\langle p_t - p_{t+1}, \widehat{\ell}_t \right\rangle - \frac{1}{\eta} D_\psi(p_{t+1}, p_t) \leq \frac{\eta}{\beta} \sum_{a=1}^{K} p_t(a)^{2-\beta} \widehat{\ell}_t(a)^2,$$

only two times worse compared to Eq. (2), but under the weaker condition (3). One benefit of this result is that it also implies the following: in MAB, when running Algorithm (1) with $\widehat{\ell}_1, \ldots, \widehat{\ell}_T$ being the inverse importance weighted loss estimators for $\ell_1, \ldots, \ell_T \in [0,1]^K$, we have for any arbitrary $a^\star \in [K]$:

$$\left\langle p_t - p_{t+1}, \widehat{\ell}_t \right\rangle - \frac{1}{\eta} D_\psi(p_{t+1}, p_t) \leq \frac{\eta}{\beta} \sum_{a=1}^{K} p_t(a)^{2-\beta} \left(\widehat{\ell}_t(a) - \ell_t(a^\star) \right)^2,$$

as long as $\eta \leq \frac{\beta}{1-\beta}\left(1 - e^{\frac{\beta-1}{\beta}}\right)$. Explain why this is true. (Hint: recall the cheating predictor trick discussed in Lecture 3 and consider running FTRL (1) on a different but equivalent loss sequence.)

Proof. Note that Eq. (1) is equivalent to running FTRL with an imaginary loss sequence $\widehat{\ell}_1 - \ell_1(a^\star)\mathbf{1}, \ldots, \widehat{\ell}_T - \ell_T(a^\star)\mathbf{1}$, that is,

$$p_t = \operatorname*{argmin}_{p \in \Delta(K)} \left\langle p, \sum_{s<t} (\widehat{\ell}_s - \ell_s(a^\star)\mathbf{1}) \right\rangle + \frac{1}{\eta}\psi(p),$$

where $\mathbf{1}$ is the all-one vector, since for any $p \in \Delta(K)$, $\langle p, \ell_s(a^\star)\mathbf{1}\rangle = \ell_s(a^\star)$ is a constant and does not affect the optimization problem. Moreover, by the condition on η and the facts $\widehat{\ell}_t(a) \geq 0$ and $\ell_t(a^\star) \in [0,1]$, we have

$$\eta p_t(a)^{1-\beta}(\widehat{\ell}_t(a) - \ell_t(a^\star)) \geq -\eta \ell_t(a^\star) \geq -\eta \geq \frac{\beta}{1-\beta}\left(e^{\frac{\beta-1}{\beta}} - 1\right),$$

and thus condition (3) holds for the imaginary loss sequence and the claimed bound holds. $\square$

2. (**Best-of-Both-Worlds for Tsallis Entropy**) In this exercise, you need prove that FTRL with Tsallis entropy ($\beta = 1/2$) and a time-varying learning rate, that is,

$$p_t = \operatorname*{argmin}_{p \in \Delta(K)} \left\langle p, \sum_{s<t} \widehat{\ell}_s \right\rangle + \frac{1}{\eta_t}\psi(p)$$

where $\psi(p) = -2\sum_{a=1}^{K} \sqrt{p(a)}$, $\eta_t = \frac{1}{2\sqrt{t}}$, and $\widehat{\ell}_1, \ldots, \widehat{\ell}_T$ are the inverse importance weighted loss estimators, satisfies Eq. (3) of Lecture 7, which further implies that it satisfies the strong best-of-both-worlds property according to Theorem 3 therein.

(a) (3pts) Let $\Phi_t^{\eta} = \min_{p \in \Delta(K)} \left\langle p, \sum_{s \le t} \widehat{\ell}_s \right\rangle + \frac{1}{\eta}\psi(p)$ and p'_{t+1} be the minimizer in the definition of $\Phi_t^{\eta_t}$. Prove the following two inequalities (hint: use Lemma 2 of Lecture 2 for the first one):

$$\Phi_{t-1}^{\eta_t} - \Phi_t^{\eta_t} \le -\left\langle p'_{t+1}, \widehat{\ell}_t \right\rangle - \frac{1}{\eta_t} D_\psi(p'_{t+1}, p_t)$$
$$\Phi_t^{\eta_t} - \Phi_t^{\eta_{t+1}} \le \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t+1}}\right)\psi(p_{t+1}).$$

Proof. Since $\Phi_{t-1}^{\eta_t}$ and $\Phi_t^{\eta_t}$ are defined with the same learning rate, and p_t and p'_{t+1} are respectively the minimizer in their definition, we can apply Lemma 2 of Lecture 2 to obtain

$$\begin{aligned}\Phi_{t-1}^{\eta_t} - \Phi_t^{\eta_t} &\le \left\langle p'_{t+1}, \sum_{s<t} \widehat{\ell}_s - \sum_{s \le t} \widehat{\ell}_s \right\rangle - D_{\frac{1}{\eta_t}\psi}(p'_{t+1}, p_t) \\ &= -\left\langle p'_{t+1}, \widehat{\ell}_t \right\rangle - \frac{1}{\eta_t} D_\psi(p'_{t+1}, p_t),\end{aligned}$$

proving the first statement. The second statement is by definition:

$$\Phi_t^{\eta_t} - \Phi_t^{\eta_{t+1}} \le \left\langle p_{t+1}, \sum_{s \le t} \widehat{\ell}_s \right\rangle + \frac{1}{\eta_t}\psi(p_{t+1}) - \Phi_t^{\eta_{t+1}} = \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t+1}}\right)\psi(p_{t+1}).$$

□

(b) (4pts) Use the previous results to prove that for any distribution $p \in \Delta(K)$,

$$\begin{aligned}\sum_{t=1}^{T} \left\langle p_t - p, \widehat{\ell}_t \right\rangle \le & \underbrace{\sum_{t=1}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)(\psi(p) - \psi(p_t))}_{\text{penalty term}} \\ &+ \underbrace{\sum_{t=1}^{T} \left(\left\langle p_t - p'_{t+1}, \widehat{\ell}_t \right\rangle - \frac{1}{\eta_t} D_\psi(p'_{t+1}, p_t)\right)}_{\text{stability\&negative term}},\end{aligned}$$

where we define $1/\eta_0 = 0$ for convenience. (Note that when η_t stays the same for all $t \ge 1$, this bound exactly recovers Lemma 3 of Lecture 2.)

Proof. Using the first result from the last question, we have

$$\begin{aligned}\sum_{t=1}^{T} \left\langle p_t, \widehat{\ell}_t \right\rangle &\le \sum_{t=1}^{T} \left(\Phi_t^{\eta_t} - \Phi_{t-1}^{\eta_t}\right) + \text{stability\&negative term} \\ &= \Phi_T^{\eta_T} - \Phi_0^{\eta_1} + \sum_{t=2}^{T} \left(\Phi_{t-1}^{\eta_{t-1}} - \Phi_{t-1}^{\eta_t}\right) + \text{stability\&negative term}.\end{aligned}$$

Further using the second result from the last question, we continue with

$$\sum_{t=1}^{T}\left\langle p_t,\widehat{\ell}_t\right\rangle \le \Phi_T^{\eta_T}-\Phi_0^{\eta_1}-\sum_{t=2}^{T}\left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}\right)\psi(p_t)+\text{stability\&negative term}.$$

Finally, noting $\Phi_0^{\eta_1}=\frac{1}{\eta_1}\psi(p_1)$ and

$$\Phi_T^{\eta_T}\le\left\langle p,\sum_{t=1}^{T}\widehat{\ell}_t\right\rangle+\frac{1}{\eta_T}\psi(p)=\left\langle p,\sum_{t=1}^{T}\widehat{\ell}_t\right\rangle+\sum_{t=1}^{T}\left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}\right)\psi(p)$$

competes the proof after rearranging. □

(c) (3pts) Prove that for any action $a^\star\in[K]$, the per-round penalty term satisfies

$$\left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}\right)(\psi(p)-\psi(p_t))\le 4\sum_{a\neq a^\star}\sqrt{\frac{p_t(a)}{t}}.$$

Proof. Note that $\psi(p)$ is maximized when p concentrates on one action and thus

$$\psi(p)-\psi(p_t)\le 2\left(\sum_{a=1}^{K}\sqrt{p_t(a)}-1\right)\le 2\sum_{a\neq a^\star}\sqrt{p_t(a)}.$$

On the other than, using the specific value of the learning rate $\eta_t=\frac{1}{2\sqrt{t}}$, we have

$$\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}=2(\sqrt{t}-\sqrt{t-1})=\frac{2}{\sqrt{t}+\sqrt{t-1}}\le\frac{2}{\sqrt{t}}.$$

Combining these two facts proves the statement. □

(d) (6pts) For the per-round stability&negative term, since $\eta_t=\frac{1}{2\sqrt{t}}\le\frac{1}{2}\le 1-\frac{1}{e}=\frac{\beta}{1-\beta}\left(1-e^{\frac{\beta-1}{\beta}}\right)$ (recall $\beta=1/2$), we can apply the results from Problem 1(d), which says: for any $a^\star\in[K]$,

$$\left\langle p_t-p'_{t+1},\widehat{\ell}_t\right\rangle-\frac{1}{\eta_t}D_\psi(p'_{t+1},p_t)\le 2\eta_t\sum_{a=1}^{K}p_t(a)^{\frac{3}{2}}\left(\widehat{\ell}_t(a)-\ell_t(a^\star)\right)^2.$$

Prove $\mathbb{E}_t\left[\sum_{a=1}^{K}p_t(a)^{\frac{3}{2}}\left(\widehat{\ell}_t(a)-\ell_t(a^\star)\right)^2\right]\le 3\sum_{a\neq a^\star}\sqrt{p_t(a)}$ where $\mathbb{E}_t$ is the conditional expectation given everything before round t. (Therefore, combining all steps, we have shown Eq. (3) of Lecture 7 for this algorithm.)

Proof. We proceed as follows:

$$\begin{aligned}
&\mathbb{E}_t\left[\sum_{a=1}^{K}p_t(a)^{\frac{3}{2}}\left(\widehat{\ell}_t(a)-\ell_t(a^\star)\right)^2\right]\\
&=\mathbb{E}_t\left[\sum_{a=1}^{K}p_t(a)^{\frac{3}{2}}\left(\widehat{\ell}_t(a)^2-2\widehat{\ell}_t(a)\ell_t(a^\star)+\ell_t(a^\star)^2\right)\right]\\
&=\sum_{a=1}^{K}p_t(a)^{\frac{3}{2}}\left(\frac{\ell_t(a)^2}{p_t(a)}-2\ell_t(a)\ell_t(a^\star)+\ell_t(a^\star)^2\right) &&\text{(Lemmas 1 and 2 of Lecture 6)}\\
&=\sum_{a=1}^{K}\left(\sqrt{p_t(a)}\ell_t(a)^2-p_t(a)^{\frac{3}{2}}\ell_t(a)\ell_t(a^\star)\right)+\sum_{a=1}^{K}p_t(a)^{\frac{3}{2}}\left(\ell_t(a^\star)^2-\ell_t(a)\ell_t(a^\star)\right).
\end{aligned}$$

For the first summation, the terms corresponding to $a\neq a^\star$ are together bounded by $\sum_{a\neq a^\star}\sqrt{p_t(a)}$ (by ignoring the second negative term); the term corresponding to $a=a^\star$ is

$$\sqrt{p_t(a^\star)}(1-p_t(a^\star))\ell_t(a^\star)^2\le 1-p_t(a^\star)=\sum_{a\neq a^\star}p_t(a)\le\sum_{a\neq a^\star}\sqrt{p_t(a)}.$$

Similarly, for the second summation, the terms corresponding to $a \neq a^\star$ are together bounded by $\sum_{a \neq a^\star} \sqrt{p_t(a)}$ (again by ignoring the second negative term); the term corresponding to $a = a^\star$ is simply 0. Combining all these bounds finishes the proof. $\square$

3. (**Log-Barrier Regularizer**) Consider running the following FTRL algorithm for MAB with an oblivious adversary:

$$p_t = \operatorname*{argmin}_{p\in\Delta(K)} \left\langle p, \sum_{s<t} \widehat{\ell}_s \right\rangle + \frac{1}{\eta}\psi(p)$$

where $\eta > 0$ is a fixed learning rate, $\psi(p) = -\sum_{a=1}^{K} \ln p(a)$ is the *log-barrier* regularizer, and $\widehat{\ell}_1, \ldots, \widehat{\ell}_T$ are the inverse importance weighted loss estimators. By the same machinery introduced in Lecture 6, it can be shown that this algorithm ensures for any $p \in \Delta(K)$:

$$\begin{aligned}\sum_{t=1}^{T} \left\langle p_t - p, \widehat{\ell}_t \right\rangle &\le \frac{\psi(p)-\psi(p_1)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \left\| \widehat{\ell}_t \right\|^2_{\nabla^{-2}\psi(p_t)} \\ &= \frac{\psi(p)-\psi(p_1)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{a=1}^{K} p_t(a)^2 \widehat{\ell}_t(a)^2. \qquad (6)\end{aligned}$$

(You do not need to prove this fact, but are encouraged to verify it yourself.)

(a) (4pts) Let $a^\star$ be the fixed optimal action in hindsight. To derive the expected regret bound of this algorithm using Eq. (6), you will find that we cannot simply pick $p = e_{a^\star}$ (the distribution that concentrates on action $a^\star$), since $\psi(p) = +\infty$ in this case. Instead, pick a p that is close to $a^\star$ and prove the following two statements:

$$\begin{aligned}\mathbb{E}[\mathcal{R}_T] &\le 1 + \frac{K \ln T}{\eta} + \mathbb{E}\left[\frac{\eta}{2}\sum_{t=1}^{T}\sum_{a=1}^{K} p_t(a)^2 \widehat{\ell}_t(a)^2\right] \qquad (7)\\ &= 1 + \frac{K \ln T}{\eta} + \mathbb{E}\left[\frac{\eta}{2}\sum_{t=1}^{T} \ell_t(a_t)^2\right]. \qquad (8)\end{aligned}$$

Proof. Let $p = (1 - \frac{1}{KT})e_{a^\star} + \frac{1}{KT}\mathbf{1}$, so that

$$\psi(p) - \psi(p_1) = \sum_{a=1}^{K} \ln \frac{p_1(a)}{p(a)} = \sum_{a=1}^{K} \ln \frac{1}{Kp(a)} \le K \ln T,$$

and

$$\begin{aligned}\mathbb{E}[\mathcal{R}_T] = \mathbb{E}\left[\sum_{t=1}^{T} \langle p_t - e_{a^\star}, \ell_t\rangle\right] &\le \mathbb{E}\left[\sum_{t=1}^{T} \langle p_t - p, \ell_t\rangle\right] + \frac{1}{KT}\sum_{t=1}^{T}\langle \mathbf{1}, \ell_t\rangle \\ &\le \mathbb{E}\left[\sum_{t=1}^{T} \left\langle p_t - p, \widehat{\ell}_t\right\rangle\right] + 1.\end{aligned}$$

Combining these with Eq. (6) proves Eq. (7). Eq. (8) is simply by the definition of the loss estimator: $\sum_{a=1}^{K} p_t(a)^2 \widehat{\ell}_t(a)^2 = \sum_{a=1}^{K} p_t(a)^2 \frac{\ell_t(a)^2}{p_t(a)^2}\mathbf{1}\{a = a_t\} = \ell_t(a_t)^2$. □

(b) (3pts) With the optimal η, Eq. (8) shows that the regret of this algorithm is $\mathcal{O}(\sqrt{TK \ln T})$, slightly worse than Exp3 or FTRL with Tsallis entropy. However, one benefit of this algorithm is that it actually ensures a small-loss bound $\widetilde{\mathcal{O}}(\sqrt{L^\star K} + K)$ where $L^\star = \sum_{t=1}^{T} \ell_t(a^\star)$ is the total loss of the optimal action. To see this, manipulate Eq. (8) to prove

$$\mathbb{E}[\mathcal{R}_T] \le 2 + \frac{2K \ln T}{\eta} + \eta L^\star,$$

as long as $\eta \le 1$, which then leads to the claimed small-loss bound if $\eta = \min\{1, \sqrt{\frac{K \ln T}{L^\star}}\}$.

Proof. First, bound $\ell_t(a_t)^2$ by $\ell_t(a_t)$ in Eq. (8). Then, use the definition of regret $\mathbb{E}[\mathcal{R}_T] = \mathbb{E}[\sum_{t=1}^{T} \ell_t(a_t)] - L^\star$ and rearrange Eq. (8) to arrive at:

$$\left(1 - \frac{\eta}{2}\right)\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t)\right] - L^\star \le 1 + \frac{K \ln T}{\eta},$$

which is equivalent to

$$\left(1-\frac{\eta}{2}\right)\mathbb{E}[\mathcal{R}_T] \le 1 + \frac{K\ln T}{\eta} + \frac{\eta}{2}L^\star.$$

Dividing both sides by $1-\frac{\eta}{2}$ and lower bounding it by $1/2$ (due to the condition $\eta \le 1$) finishes the proof. □

(c) By the same reasoning as in Problem 1(d), one can also improve Eq. (7) to

$$\mathbb{E}[\mathcal{R}_T] \le 1 + \frac{K\ln T}{\eta} + \mathbb{E}\left[\eta\sum_{t=1}^{T}\sum_{a=1}^{K} p_t(a)^2(\widehat{\ell}_t(a)-\ell_t(a_t))^2\right],$$

which, together with a doubling trick on tuning η, leads to

$$\mathbb{E}[\mathcal{R}_T] \le B\sqrt{(K\ln T)\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K} p_t(a)^2(\widehat{\ell}_t(a)-\ell_t(a_t))^2\right]} \tag{9}$$

for some constant $B>0$.

(i) (6pts) Let $\mathbb{E}_t$ be the conditional expectation given everything before round t. Prove that for any action $a\in[K]$, we have $\mathbb{E}_t\left[(\widehat{\ell}_t(a)-\ell_t(a_t))^2\right] \le \frac{1-p_t(a)}{p_t(a)}$ and

$$\mathbb{E}_t\left[\sum_{a=1}^{K} p_t(a)^2(\widehat{\ell}_t(a)-\ell_t(a_t))^2\right] \le 2(1-p_t(a^\star))$$

for any action $a^\star\in[K]$.

Proof. By definitions, we have

$$\begin{aligned}(\widehat{\ell}_t(a)-\ell_t(a_t))^2 &= \left(\frac{\ell_t(a_t)\mathbf{1}\{a\neq a_t\}}{p_t(a_t)} - \ell_t(a_t)\right)^2 \\ &= \frac{(\mathbf{1}\{a\neq a_t\}-p_t(a_t))^2\ell_t(a_t)^2}{p_t(a_t)^2} \le \frac{(\mathbf{1}\{a\neq a_t\}-p_t(a_t))^2}{p_t(a_t)^2}\end{aligned}$$

and thus

$$\mathbb{E}_t\left[(\widehat{\ell}_t(a)-\ell_t(a_t))^2\right] \le p_t(a)\frac{(1-p_t(a))^2}{p_t(a)^2} + \sum_{b\neq a}^{K} p_t(b) = \frac{1-p_t(a)}{p_t(a)},$$

proving the first statement. The second statement holds consequently:

$$\begin{aligned}&\mathbb{E}_t\left[\sum_{a=1}^{K} p_t(a)^2(\widehat{\ell}_t(a)-\ell_t(a_t))^2\right] \le \sum_{a=1}^{K} p_t(a)^2\frac{1-p_t(a)}{p_t(a)} \\ &= \sum_{a=1}^{K} p_t(a)(1-p_t(a)) \le 1-p_t(a^\star) + \sum_{a\neq a^\star} p_t(a) = 2(1-p_t(a^\star)).\end{aligned}$$

□

(ii) (5pts) Consider the same condition stated in Theorem 3 of Lecture 7: the environment is such that

$$\mathbb{E}[\mathcal{R}_T] \ge \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a\neq a^\star} p_t(a)\Delta(a)\right] - C$$

for some action $a^\star$, gap measures $\Delta(a)>0$ for $a\neq a^\star$, and a constant $C>0$. Combine Eq. (9) and the result of the last question to prove that this algorithm satisfies

$$\mathbb{E}[\mathcal{R}_T] = \mathcal{O}\left(\frac{K\ln T}{\Delta_{\min}} + \sqrt{\frac{CK\ln T}{\Delta_{\min}}}\right),$$

where $\Delta_{\min} = \min_{a\neq a^\star}\Delta(a)$ (that is, a weaker best-of-both-worlds result). (Hint: read the proof of Theorem 5 in Lecture 3 again.)

Proof. Combining Eq. (9) and the last question, we have

$$
\begin{aligned}
\mathbb{E}[\mathcal{R}_T] &\le B\sqrt{(2K\ln T)\mathbb{E}\left[\sum_{t=1}^{T} 1 - p_t(a^\star)\right]} \\
&= B\sqrt{\frac{2K\ln T}{\Delta_{\min}}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a\neq a^\star} p_t(a)\Delta_{\min}\right]} \\
&\le B\sqrt{\frac{2K\ln T}{\Delta_{\min}}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a\neq a^\star} p_t(a)\Delta(a)\right]} \\
&\le B\sqrt{\frac{2K\ln T}{\Delta_{\min}}(\mathbb{E}\left[\mathcal{R}_T\right] + C)} && \text{(condition of the environment)} \\
&\le B\sqrt{\frac{2K\ln T}{\Delta_{\min}}}(\sqrt{\mathbb{E}\left[\mathcal{R}_T\right]} + \sqrt{C}).
\end{aligned}
$$

Finally, using the fact $x \le b\sqrt{x} + c \Rightarrow x \le b^2 + 2c$ proves

$$
\mathbb{E}[\mathcal{R}_T] \le \frac{2B^2K\ln T}{\Delta_{\min}} + 2B\sqrt{\frac{2CK\ln T}{\Delta_{\min}}}.
$$

$\square$

4. (**Impossibility of Strongly Adaptive Algorithms**) In this exercise, you need to show that strongly adaptive algorithms are impossible for the adversarial MAB problem even with only two actions, that is, no algorithm can guarantee $\mathbb{E}[\mathcal{R}_\mathcal{I}] \le B\sqrt{|\mathcal{I}|}$ for all interval $\mathcal{I}$ simultaneously, where B is an absolute constant.

   (a) (4pts) We prove by contradiction. Suppose that such a strongly adaptive algorithm $\mathcal{A}$ exists. Consider running it in a 2-armed bandit problem where $\ell_t(1)$ is always $1/2$ and $\ell_t(2)$ is always 1 for all t. Prove that there must exist an interval $\mathcal{I}_\mathcal{A}$ of length $\frac{\sqrt{T}}{4B}$ (assumed to be an integer for simplicity), where the total expected number of times $\mathcal{A}$ selects action 2 is at most $1/2$.

   Proof. Clearly, in this environment the expected regret of the algorithm equals one half of the expected number of times action 2 is selected. Partition the entire T rounds evenly into $4B\sqrt{T}$ intervals, each of length $\frac{\sqrt{T}}{4B}$. If in every one of these intervals, the expected number of times $\mathcal{A}$ selects action 2 is larger than $1/2$, then the total number of times action 2 is selected is larger than $2B\sqrt{T}$, which is a contradiction to the fact that $\mathcal{A}$ guarantees $\mathbb{E}[\mathcal{R}_{[1,T]}] \le B\sqrt{T}$. This proves the claimed statement. □

   (b) (4pts) Continuing with the last question, use Markov's inequality (link) to show that with probability at least $1/2$, $\mathcal{A}$ never picks action 2 on interval $\mathcal{I}_\mathcal{A}$.

   Proof. Let n be the number of times $\mathcal{A}$ selects action 2 on interval $\mathcal{I}_\mathcal{A}$, which satisfies $\mathbb{E}[n] \le 1/2$ according to the last question. Directly applying Markov's inequality tells us $\Pr(n \ge 1) \le \mathbb{E}[n] \le \frac{1}{2}$, and thus $\Pr(n < 1) \ge \frac{1}{2}$. Noting that $n < 1$ is the same as $\mathcal{A}$ never selecting action 2 on interval $\mathcal{I}_\mathcal{A}$ finishes the proof. □

   (c) (4pts) Finally, consider a new environment that is different from the previous one only on interval $\mathcal{I}_\mathcal{A}$, where $\ell_t(2)$ is now always 0 (while $\ell_t(1)$ stays the same) for all $t \in \mathcal{I}_\mathcal{A}$. Prove that running the same algorithm $\mathcal{A}$ in this environment gives $\mathbb{E}[\mathcal{R}_{\mathcal{I}_\mathcal{A}}] = \Omega(\sqrt{T})$, a contradiction to the strongly adaptive property which says $\mathbb{E}[\mathcal{R}_{\mathcal{I}_\mathcal{A}}] \le B\sqrt{|\mathcal{I}_\mathcal{A}|} = \mathcal{O}(T^{1/4})$.

   Proof. Note that in this new environment, with probability at least $1/2$, $\mathcal{A}$ behaves exactly the same as in the previous environment, because it never picks action 2 on $\mathcal{I}_\mathcal{A}$ and thus never observes anything different from the previous environment. On the other hand, action 2 is the better action on $\mathcal{I}_\mathcal{A}$ with no loss at all, so not picking action 2 at all on $\mathcal{I}_\mathcal{A}$ incurs $\frac{1}{2}|\mathcal{I}_\mathcal{A}|$ regret. Therefore, $\mathbb{E}[\mathcal{R}_{\mathcal{I}_\mathcal{A}}] \ge \frac{1}{2} \cdot \frac{1}{2}|\mathcal{I}_\mathcal{A}| = \Omega(\sqrt{T})$, proving the statement. □