CSCI 659 Homework 3

Spring 2026

Instructor: Haipeng Luo

This homework is due on **04/10, 11:59pm**. *See course website for more instructions on finishing and submitting your homework as well as the late policy. Total points: 70.*

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,

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

   (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

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

(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 $\alpha = \frac{\beta}{\beta-1}$.)

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

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

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 $a^\star \in [K]$:

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

as long as $\eta \le \frac{\beta}{1-\beta}\left(1 - e^{\frac{\beta-1}{\beta}}\right)$. 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.)

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):

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

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

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

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.)

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

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

(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.)

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 $p \in \Delta(K)$:

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

(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:

$$\mathbb{E}[\mathcal{R}_T] \le 1 + \frac{K \ln T}{\eta} + \mathbb{E}\left[\frac{\eta}{2}\sum_{t=1}^{T}\sum_{a=1}^{K} p_t(a)^2 \widehat{\ell}_t(a)^2\right] \tag{7}$$

$$= 1 + \frac{K \ln T}{\eta} + \mathbb{E}\left[\frac{\eta}{2}\sum_{t=1}^{T} \ell_t(a_t)^2\right]. \tag{8}$$

(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 $\eta = \min\{1, \sqrt{\frac{K\ln T}{L^\star}}\}$.

(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 $a^\star \in [K]$.

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

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

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.)

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 $\mathbb{E}[\mathcal{R}_{\mathcal{I}}] \leq B\sqrt{|\mathcal{I}|}$ 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 $1/2$.

   (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 $\mathcal{I}_{\mathcal{A}}$.

   (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 $\mathbb{E}[\mathcal{R}_{\mathcal{I}_{\mathcal{A}}}] \leq B\sqrt{|\mathcal{I}_{\mathcal{A}}|} = \mathcal{O}(T^{1/4})$.

5. (**Hedge for Online RL**) In this exercise, you will analyze yet another algorithm for online RL that enjoys $\mathcal{O}(\sqrt{T})$ regret (albeit being computationally inefficient, unlike the two algorithms discussed in Lecture 10). The idea is to simply run Hedge over the set of all deterministic policies $\Pi = \{\pi : X \to A\}$ (that is, mappings from a state to an action), with an appropriate loss estimator. Specifically, at time t, the algorithm samples $\pi_t \in \Pi$ from a distribution P_t over Π that follows the Hedge update:

$$P_t(\pi) \propto \exp\left(-\eta \sum_{s<t} \widehat{\ell}_s(\pi)\right), \quad \forall \pi \in \Pi$$

for some policy loss estimator $\widehat{\ell}_s(\pi)$. After that, the algorithm executes policy π_t and uses the observations from this episode to construct a new estimator $\widehat{\ell}_t(\pi)$ for each policy π. As in Lecture 10, for simplicity you can assume that the transition of the MDP is known.

(a) (6pts) Similar to the Exp4 algorithm for contextual bandits (Lecture 8), we need a good policy loss estimator that is unbiased and at the same time enjoys a low variance. More specifically, propose one such estimator and prove that for any policy π:

- (unbiasedness) $\mathbb{E}_t\left[\widehat{\ell}_t(\pi)\right] = \langle q^\pi, \ell_t \rangle$;
- (low variance) $\mathbb{E}_t\left[\widehat{\ell}_t(\pi)^2\right] \le H \sum_{(x,a)\in X\times A} \frac{q^\pi(x,a)}{q_t(x,a)}$

where $\mathbb{E}_t[\cdot]$ denotes the conditional expectation given everything before episode t, ℓ_t is the true loss function for episode t, q^π is the occupancy measure induced by π, and q_t is a shorthand for $\mathbb{E}_t[q^{\pi_t}] = \sum_{\pi\in\Pi} P_t(\pi) q^\pi$.

(b) (4pts) Continuing from the last question, prove that such a loss estimator ensures the following regret bound: for any $\pi \in \Pi$,

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle q_t - q^\pi, \ell_t \rangle\right] \le \frac{|X| \ln |A|}{\eta} + \eta H |X||A| T,$$

which, after picking the optimal η, is of order $\widetilde{\mathcal{O}}(|X|\sqrt{H|A|T})$.