FALL 2026 BAIS 7900

Homework 2

Instructor: *Mengxiao Zhang* Due: November 15, 2026

Total points: 50, plus 10 bonus points. Upload the PDF generated by LaTeX. For the implementation questions, submit runnable code with your PDF. Unless otherwise stated, all logarithms are natural, $K \geq 2$, $T > 2$, and losses lie in $[0,1]$. Denote by $[K] \triangleq \{1, \dots, K\}$ the arm set and by $\Delta_K \triangleq \{p \in [0,1]^K : \sum_{i=1}^K p_i = 1\}$ the probability simplex.

1. FTRL with Log-Barrier Regularizer **(15 pts)**

In Lecture 8, we established that Tsallis-INF achieves logarithmic regret in stochastic environments and minimax regret in adversarial environments. In this question, you will show that FTRL with the log-barrier regularizer also adapts to both environments, with an additional dependence on K and $\log T$. Denote by $\psi(p) \triangleq -\sum_{i=1}^K \log p_i$ the log-barrier regularizer on the simplex Δ_K. Consider the following algorithm.

Algorithm 1: FTRL with the Log-Barrier Regularizer

Input: arms $[K]$, horizon T, positive learning rates η_t chosen before drawing I_t
Initialize $\widehat{L}_{0,i} = 0$ for every $i \in [K]$;
for $t = 1, \dots, T$ **do**
  choose $p_t = \operatorname{argmin}_{p \in \Delta_K} \{\left\langle p, \widehat{L}_{t-1} \right\rangle + \psi(p)/\eta_t\}$;
  draw $I_t \sim p_t$ and observe ℓ_{t,I_t};
  set $\widehat{\ell}_{t,i} = \ell_{t,I_t} \mathbf{1}\{I_t = i\}/p_{t,i}$ for every $i \in [K]$;
  set $\widehat{L}_{t,i} = \widehat{L}_{t-1,i} + \widehat{\ell}_{t,i}$ for every $i \in [K]$;

For parts (a) and (b), use a fixed learning rate $\eta_t = \eta > 0$ and a fixed oblivious loss sequence in $[0,1]^K$. Denote by $L_T^\star \triangleq \min_{j \in [K]} \sum_{t=1}^T \ell_{t,j}$ the best arm's cumulative loss and by $\overline{\mathrm{Reg}}_T \triangleq \mathbb{E}[\sum_{t=1}^T \ell_{t,I_t}] - L_T^\star$ the pseudo-regret. You may use the following specialization of the FTRL decomposition and local-curvature argument from Lecture 7: for every comparator $u \in \Delta_K$,

$$\sum_{t=1}^T \left\langle p_t - u, \widehat{\ell}_t \right\rangle \leq \frac{\psi(u) - \psi(p_1)}{\eta} + \eta \sum_{t=1}^T \sum_{i=1}^K p_{t,i}^2 \widehat{\ell}_{t,i}^2. \tag{1}$$

(a) **(4 pts)** Choose a comparator approaching a best arm and use equation (1) to prove that

$$\overline{\mathrm{Reg}}_T \leq 1 + \frac{K \log T}{\eta} + \eta \mathbb{E}\left[\sum_{t=1}^T \ell_{t,I_t}^2\right].$$

Choose η to obtain the adversarial guarantee $\overline{\mathrm{Reg}}_T \leq \mathcal{O}(\sqrt{KT \log T})$.

(b) **(3 pts)** For $0 < \eta \leq 1/2$, prove that $\overline{\mathrm{Reg}}_T \leq 2 + 2K \log T/\eta + 2\eta L_T^\star$. Give an explicit choice of η, assuming that $L_T^\star$ is known, which guarantees that

$$\overline{\mathrm{Reg}}_T \leq \mathcal{O}\left(\sqrt{K L_T^\star \log T} + K \log T\right).$$

(c) **(4 pts)** Denote by $\mathbb{E}_t$ conditional expectation given the past and the current loss vector, before drawing I_t. Prove that $\mathbb{E}_t[(\widehat{\ell}_{t,i} - \ell_{t,I_t})^2] \le (1 - p_{t,i})/p_{t,i}$ for every $i \in [K]$. Denote by $V_t \triangleq \sum_{i=1}^{K} p_{t,i}^2 (\widehat{\ell}_{t,i} - \ell_{t,I_t})^2$ the squared local norm of the estimates after a common shift. Deduce that $\mathbb{E}_t[V_t] \le 2(1 - p_{t,j})$ for every $j \in [K]$.

(d) **(4 pts)** We now choose the learning rates using the observed quantities V_t. Set $A = K \log T$, initialize $S_0 = 0$, and use

$$\eta_t = \min\left\{\frac{1}{2}, \sqrt{\frac{A}{A + S_{t-1}}}\right\}.$$

Set $S_t = S_{t-1} + V_t$ after observing the loss. You may use without proof that this version of Algorithm 1 guarantees that

$$\mathbb{E}\left[\sum_{t=1}^{T} (\ell_{t,I_t} - \ell_{t,j})\right] \le 4A + 4\sqrt{A\,\mathbb{E}[S_T]}, \tag{2}$$

for every fixed comparator $j \in [K]$. First use part (c) to obtain $\mathcal{O}(\sqrt{KT \log T} + K \log T)$ regret against every fixed arm in an adversarial environment. Next consider independent and identically distributed loss vectors, and denote by $\mu_i \triangleq \mathbb{E}[\ell_{t,i}]$ the mean loss of arm i. Suppose that there is a unique best arm $i^\star$. Denote by $\Delta_i \triangleq \mu_i - \mu_{i^\star}$ the gap of arm $i \neq i^\star$ and by $\Delta_{\min} \triangleq \min_{i \neq i^\star} \Delta_i$ the smallest gap. Show that

$$\overline{\mathrm{Reg}}_T(i^\star) \triangleq \mathbb{E}\left[\sum_{t=1}^{T} (\ell_{t,I_t} - \ell_{t,i^\star})\right] = \sum_{t=1}^{T} \sum_{\substack{i \in [K]: \\ i \neq i^\star}} \Delta_i \mathbb{E}[p_{t,i}],$$

and deduce that

$$\overline{\mathrm{Reg}}_T(i^\star) \le \mathcal{O}\left(\frac{K \log T}{\Delta_{\min}}\right).$$

2. High-Probability Guarantee and Adaptive Adversary **(25 pts)**

In Lecture 6, we established an expected-regret guarantee for EXP3 against an oblivious adversary, which fixes the loss sequence before play. This question asks whether regret also concentrates around that guarantee. We then modify EXP3 to obtain a high-probability bound against an adaptive adversary, which may choose ℓ_t using the previous history but must choose it before seeing the current action I_t. Throughout this question, denote by $\mathrm{Reg}_T \triangleq \sum_{t=1}^{T} \ell_{t,I_t} - \min_{j \in [K]} \sum_{t=1}^{T} \ell_{t,j}$ the realized regret.

Algorithm 2: EXP3

Input: arms $[K]$, horizon T, learning rate $\eta > 0$
Initialize $w_{1,i} = 1$ for every $i \in [K]$;
for $t = 1, \dots, T$ **do**
  set $p_{t,i} = w_{t,i} / \sum_{j=1}^{K} w_{t,j}$ for every $i \in [K]$;
  draw $I_t \sim p_t$ and observe ℓ_{t,I_t};
  set $\widehat{\ell}_{t,i} = \ell_{t,I_t} \mathbf{1}\{I_t = i\}/p_{t,i}$ for every $i \in [K]$;
  set $w_{t+1,i} = w_{t,i} \exp(-\eta \widehat{\ell}_{t,i})$ for every $i \in [K]$;

- **EXP3 can have linear regret with constant probability. (15 pts)**
  Fix $\rho \in (0,1)$, an even horizon T, and $\eta \in [T^{-\rho}, 1]$. Consider two arms with losses $(\ell_{t,1}, \ell_{t,2}) = (0, \alpha)$ for $t \le T/2$ and $(1,0)$ for $t > T/2$, where $\alpha \in [1/4, 1/2]$. Show that, for every sufficiently large T, some choice of α makes Algorithm 2 satisfy $\Pr[\mathrm{Reg}_T \ge T/4] \ge 1/130$. Complete the following intermediate claims.

  (a) **(2 pts)** Denote by $N_2(t) \triangleq \sum_{u=1}^{t} \mathbf{1}\{I_u = 2\}$ the number of selections of arm 2. Set $q_0(\alpha) = 1/2$ and define

  $$q_{s+1}(\alpha) \triangleq \frac{q_s(\alpha)\exp(-\eta\alpha/q_s(\alpha))}{1 - q_s(\alpha) + q_s(\alpha)\exp(-\eta\alpha/q_s(\alpha))}.$$

  Prove that $p_{t,2} = q_{N_2(t-1)}(\alpha)$ for $t \le T/2$.

  (b) **(5 pts)** Prove that, for sufficiently large T, there exist $\alpha \in [1/4, 1/2]$ and an integer $s \ge 1$ such that $q_s(\alpha) = 1/(8T)$ and $\sum_{u=0}^{s-1} 1/q_u(\alpha) \le T/8$.

  (c) **(3 pts)** For this choice, prove that $\Pr[N_2(T/2) \ge s+1] \ge 1/65$.

  (d) **(4 pts)** Prove that $\Pr[\mathrm{Reg}_T \ge T/4] \ge (1 - T\exp(-\eta T)/2)/65$ and deduce the claimed constant-probability lower bound.

  (e) **(1 pt)** Why is this consistent with EXP3's expected regret bound $\mathcal{O}(\sqrt{KT\log K})$ under its usual learning rate?

- **A high-probability guarantee for EXP3-IX. (10 pts)**
  We now replace the importance-weighted estimates by estimates with *implicit exploration*.

Algorithm 3: EXP3-IX with Fixed Parameters

Input: arms $[K]$, horizon T, learning rate $\eta > 0$, implicit exploration parameter $\gamma > 0$

Initialize $w_{1,i} = 1$ for every $i \in [K]$;

for $t = 1, \dots, T$ **do**

- set $p_{t,i} = w_{t,i} / \sum_{j=1}^{K} w_{t,j}$ for every $i \in [K]$;
- draw $I_t \sim p_t$ and observe ℓ_{t,I_t};
- set $\widehat{\ell}_{t,i} = \ell_{t,I_t}\mathbf{1}\{I_t = i\}/(p_{t,i} + \gamma)$ for every $i \in [K]$;
- set $w_{t+1,i} = w_{t,i}\exp(-\eta\widehat{\ell}_{t,i})$ for every $i \in [K]$;

Denote by $\mathbb{E}_t$ conditional expectation given the past and the current loss vector, before drawing I_t.

(a) **(2 pts)** Suppose that $a_{t,i} \in [0, 2\gamma]$ is determined by the history before round t. Prove that

$$\mathbb{E}_t\left[\exp\left(\sum_{i=1}^{K} a_{t,i}(\widehat{\ell}_{t,i} - \ell_{t,i})\right)\right] \le 1.$$

(b) **(2 pts)** For every $\delta \in (0,1)$, deduce that

$$\Pr\left[\sum_{t=1}^{T}\sum_{i=1}^{K} a_{t,i}(\widehat{\ell}_{t,i} - \ell_{t,i}) > \log\frac{1}{\delta}\right] \le \delta.$$

In particular, prove that, with probability at least $1-\delta$,

$$\sum_{t=1}^{T}(\widehat{\ell}_{t,j}-\ell_{t,j}) \le \frac{\log(K/\delta)}{2\gamma},$$

simultaneously for all $j \in [K]$.

(c) **(1 pt)** Prove that $\left\langle p_t, \widehat{\ell}_t \right\rangle = \ell_{t,I_t} - \gamma \sum_{i=1}^{K} \widehat{\ell}_{t,i}$ and $\sum_{i=1}^{K} p_{t,i}\widehat{\ell}_{t,i}^2 \le \sum_{i=1}^{K} \widehat{\ell}_{t,i}$.

(d) **(2 pts)** Use the exponential-weights inequality from Lecture 6 to prove that

$$\sum_{t=1}^{T}(\ell_{t,I_t}-\ell_{t,j}) \le \frac{\log K}{\eta} + \sum_{t=1}^{T}(\widehat{\ell}_{t,j}-\ell_{t,j}) + (\eta+\gamma)\sum_{t=1}^{T}\sum_{i=1}^{K}\widehat{\ell}_{t,i},$$

for every $j \in [K]$.

(e) **(3 pts)** Assume that $\eta \le \gamma$. Use parts (a)–(d) to prove that, with probability at least $1-\delta$,

$$\mathrm{Reg}_T \triangleq \sum_{t=1}^{T}\ell_{t,I_t} - \min_{i\in[K]}\sum_{t=1}^{T}\ell_{t,i} \le \frac{\log K}{\eta} + \frac{\log(2K/\delta)}{2\gamma} + (\eta+\gamma)KT + \log\frac{2}{\delta}.$$

Give explicit, time-independent choices of η and γ which yield $\mathcal{O}(\sqrt{KT\log(K/\delta)} + \log(1/\delta))$ regret with probability at least $1-\delta$. Explain why the proof applies to the adaptive adversary defined above.

3. Implementation I: Best of Both Worlds **(5 pts)**

In Lecture 8, we established a best-of-both-worlds guarantee for Tsallis-INF. You will compare its empirical performance with EXP3 and UCB in a stochastic environment and an environment whose loss distributions change over time.

Use $K = 8$, $T = 50000$, and $\Delta = 1/8$. Generate Bernoulli losses independently across arms and rounds, with the following mean vectors.

- *Stochastic environment.* Use the fixed means $\mu_1 = (1-\Delta)/2$ and $\mu_i = (1+\Delta)/2$ for every $i \neq 1$.
- *Changing environment.* Starting with phase $m = 0$, alternate between the mean vectors $(0, \Delta, \ldots, \Delta)$ in even phases and $(1-\Delta, 1, \ldots, 1)$ in odd phases. Phase m lasts $\lceil 100(1.6)^m \rceil$ rounds; truncate the last phase at T. Fix the phase schedule before play. Arm 1 has the smallest mean loss on every round, with gap Δ to every other arm.

(a) **(2 pts)** Implement the following three algorithms, using the same parameters in both environments.

- *Tsallis-INF.* Use the Lecture 8 regularizer $\psi_{1/2}(p) = 2(1 - \sum_{i=1}^{K}\sqrt{p_i})$, learning rate $\eta_t = 1/(2\sqrt{t})$, and importance-weighted estimates.
- *EXP3.* Use Algorithm 2 with $\eta = \sqrt{\log K/(KT)}$.
- *UCB.* Pull arms $1, \ldots, K$ once in this order. Subsequently choose an arm minimizing $\widehat{\mu}_{t-1,i} - \sqrt{2\log t / N_i(t-1)}$, where $N_i(t-1)$ is its number of previous pulls and $\widehat{\mu}_{t-1,i}$ is its empirical mean observed loss. Break ties by the smallest arm index.

(b) **(2 pts)** Run 100 independent repetitions, using seeds $0, \ldots, 99$. For each seed and environment, generate one full loss matrix and use it for all three algorithms, revealing only the selected loss to each algorithm. Use separate random streams for environment generation and each algorithm's action draws. For each run, record $G_t \triangleq \Delta \sum_{u=1}^{t} \mathbf{1}\{I_u \neq 1\}$, whose expectation is the pseudo-regret against arm 1. Plot the mean G_t with pointwise 95% normal-approximation Monte Carlo confidence intervals in each environment, and report its mean and standard error at $t \in \{1000, 3000, 10000, 30000, 50000\}$.

(c) **(1 pt)** Explain the differences between the algorithms in the two environments and connect your observations to their guarantees. In particular, discuss the effect of changing mean losses on UCB and the stochastic performance of EXP3.

Submit runnable code, the commands needed to reproduce the experiments, the figures, and a short discussion. Do not tune parameters separately by environment.

4. Implementation II: EXP3 and EXP3-IX **(5 pts)** You will see how implicit exploration improves the concentration of regret in the two-arm environment from the first part of Question 2. Use $T = 10000$ and, for each $\alpha \in \{0.28, 0.285, 0.29, 0.295, 0.30\}$, set $(\ell_{t,1}, \ell_{t,2}) = (0, \alpha)$ for $t \leq T/2$ and $(1, 0)$ afterward. Run Algorithm 2 and Algorithm 3 with the same fixed learning rate $\eta = \sqrt{\log 2/T}$, and use $\gamma = \eta$ for EXP3-IX.

(a) **(2 pts)** Implement both algorithms. For each α, perform 500 independent runs with seeds $0, \ldots, 499$, using separate action-draw streams for the two algorithms. Record the realized regret $\mathrm{Reg}_t \triangleq \sum_{u=1}^{t} \ell_{u,I_u} - \min_{j\in[2]} \sum_{u=1}^{t} \ell_{u,j}$ after every round. At the end, the benchmark is $\alpha T/2$, so $\mathrm{Reg}_T = \sum_{t=1}^{T} \ell_{t,I_t} - \alpha T/2$.

(b) **(3 pts)** For each α, plot the mean regret curve across the 500 runs, with a shaded band of one sample standard deviation above and below the mean. Also give side-by-side boxplots of the terminal regret distributions on a shared vertical scale. Report the mean, sample variance, 95th and 99th percentiles, and the fraction of runs with $\mathrm{Reg}_T \geq T/4$. Include all runs in these summaries.

Submit runnable code, reproduction commands, the plots, and the summary table. Keep the parameters fixed across the five values of α.

5. Bonus Question: First-Order Bound in Contextual Bandits **(10 bonus pts)**

In Lecture 9, we established that an online square-loss regression oracle with regret bound $B(T)$ gives contextual-bandit regret $\mathcal{O}(\sqrt{KTB(T)})$. In this problem, you will show that an online log-loss regression oracle gives a first-order guarantee, whose leading term depends on the loss of the optimal policy. We use the FastCB sampling rule of Dylan Foster and Akshay Krishnamurthy (2021).

Use the realizable setting from Lecture 9. Thus $\mathcal{F} \subseteq \{f : \mathcal{X} \times [K] \to [0, 1]\}$ contains an unknown $f^\star$ satisfying $\mathbb{E}[\ell_{t,i} \mid \mathcal{H}_{t-1}, x_t] = f^\star(x_t, i)$, where $\mathcal{H}_{t-1}$ is the previous history. Conditional on this information and x_t, the current action draw is independent of the current loss vector. Contexts may depend on the previous history. Denote by $\pi^\star(x) \in \operatorname{argmin}_{i\in[K]} f^\star(x, i)$ an optimal policy, with fixed tie-breaking, and by $L^\star \triangleq \mathbb{E}[\sum_{t=1}^{T} \ell_{t,\pi^\star(x_t)}]$ its expected cumulative loss.

Denote by $\ell_{\log}(u, y) \triangleq -y \log u - (1-y)\log(1-u)$ the log loss, using the usual boundary conventions. Suppose that an online regression oracle provides predictions $\widehat{y}_{t,i} \in (0,1)$ and, for every adaptively chosen sequence of selected instances and outcomes, guarantees that

$$\sum_{t=1}^{T} \ell_{\log}(\widehat{y}_{t,I_t}, \ell_{t,I_t}) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell_{\log}(f(x_t, I_t), \ell_{t,I_t}) \le \mathrm{Reg}_{\log}, \tag{3}$$

where $\mathrm{Reg}_{\log} > 0$ is a known upper bound. All K predictions on a round are queried from the same oracle state, and only the selected context-action pair is used to update the oracle.

Algorithm 4: FastCB with a Log-Loss Regression Oracle

Input: online log-loss regression oracle, fixed parameter $\gamma \ge 2K$
for $t = 1, \dots, T$ **do**
observe x_t and query $\widehat{y}_{t,i}$ for every $i \in [K]$ without updating the oracle;
choose $b_t \in \mathrm{argmin}_{i \in [K]} \widehat{y}_{t,i}$ using fixed tie-breaking;
set $p_{t,i} = \widehat{y}_{t,b_t} / [K\widehat{y}_{t,b_t} + \gamma(\widehat{y}_{t,i} - \widehat{y}_{t,b_t})]$ for every $i \ne b_t$;
set $p_{t,b_t} = 1 - \sum_{i \in [K]: i \ne b_t} p_{t,i}$;
draw $I_t \sim p_t$ and observe ℓ_{t,I_t};
update the oracle once using (x_t, I_t, ℓ_{t,I_t});

The proof reduces to the following one-round allocation inequality and the oracle guarantee.

(a) **(5 pts)** Prove the following lemma. Fix a prediction vector $y \in (0,1)^K$, a mean-loss vector $f \in [0,1]^K$, and $\gamma \ge 2K$. Construct p from y using the sampling rule in Algorithm 4, and choose $i^\star \in \mathrm{argmin}_{i \in [K]} f_i$. Then

$$\sum_{i=1}^{K} p_i (f_i - f_{i^\star}) \le \frac{8K}{\gamma} \sum_{i=1}^{K} p_i f_i + 4\gamma \sum_{i=1}^{K} p_i \frac{(y_i - f_i)^2}{y_i + f_i}.$$

(b) **(5 pts)** Using part (a) and equation (3), prove that for every $\gamma \ge 16K$,

$$\mathbb{E}\left[\sum_{t=1}^{T} (\ell_{t,I_t} - \ell_{t,\pi^\star(x_t)})\right] \le \frac{16KL^\star}{\gamma} + 16\gamma \mathrm{Reg}_{\log}.$$

You may use $(u-v)^2/(u+v) \le 2\,\mathrm{kl}(u,v)$, where $\mathrm{kl}(u,v) \triangleq u\log(u/v) + (1-u)\log((1-u)/(1-v))$ is binary relative entropy, with the usual boundary conventions. Assume that $L^\star$ is known. Choose γ to obtain the guarantee $\mathcal{O}(\sqrt{KL^\star \mathrm{Reg}_{\log}} + K\mathrm{Reg}_{\log})$.