Online Optimization and Learning (CS245)

Name: ID: Email:

Rules:

1. Deadline: **2025/4/14/23:59:59**.
   The grade of the late submission subjects to the decaying policy $(75\%, 50\%, 25\%)$.
2. Please do latex your homework and no handwriting is accepted.
3. Submit your homework to TA(guohq@shanghaitech.edu.cn), including your PDF and Code, with filename "name+id+CS245HW2.zip".
4. Plagiarism is not allowed. You will fail this homework if any plagiarism is detected.

Problem 1: Explore-then-Exploit in Bandits

Explore-then-Exploit is a simple and efficient algorithm for K-armed bandit problems.

Explore-then-Exploit
Initialization: Time horizon T, exploration times N. **Exploration:** In the first phase, the choice of arms does not depend on the observed rewards, and each arm is played for N times. **Exploitation:** In all remaining rounds, the algorithm selects the arm with the highest empirical mean reward based on the exploration phase.

Consider the stochastic multi-armed bandit setting:

- After N times explorations, for any arm a,derive an upper bound on the expected estimation error $|\mu_a - \bar{\mu}_a|$, where μ_a is the true mean reward and $\bar{\mu}_a$ is its empirical mean estimate.
- Choose proper N, and prove an **upper bound** on the regret for the above algorithm.
- Specialize the analysis to the two-armed case and establish a **lower bound** on the regret of the algorithm.

Solution:

Problem 2: Online Mirror Descent for Adversarial/Stochastic Bandits

We have discussed Online Mirror Descent is possible to achieve good performance for Adversarial/Stochastic Bandits.

Online Mirror Descent for Adversarial/Stochastic Bandits

Initialization: $x_1 = [1/K, ..., 1/K]$ and learning rate η_t.
For each round $t = 1, \cdots, T$:

- **Learner:** Sample an arm i from x_t.
- **Environment:** Observe the reward of arm i : $r_t(i)$.
- **Estimator:** $\hat{r}_t(i) = r_t(i)/x_t(i)$ and 0 otherwise.
- **Update:** $x_{t+1} = \arg\min_{x \in \mathcal{K}} \langle x, -\hat{r}_t \rangle + \frac{1}{\eta_t} B_\Psi(x; x_t)$.

If the regularizer $\Psi(x)$ is the negative entropy function, B_ψ is the KL divergence and the algorithm is the classical EXP3 algorithm.
Now we consider a different regularizer $\psi(x) = -\sum_{i=1}^{K} \sqrt{x_i}$.

- For adversarial bandits, please provide the regret analysis of the algorithm with a proper adaptive learning rate η_t and compare it with the regret of EXP3.
- For stochastic bandits, please try to provide a possible problem dependent regret analysis of the algorithm with a proper adaptive learning rate η_t.

Solution:

Problem 3: Bandit Algorithms

Consider the following protocol of Bandits problem.

Learning in Bandits

Initialization: K arms.
For each round $t = 1, \cdots, T$:

- **Learner:** Choose an arm $i \in [K]$.
- **Environment:** Observe the loss of picked arm $\ell_{t,i}$.

In this problem, we provide an environment with $K = 32$ arms and $T = 5000$ rounds, where each round you will receive a **loss** of your picked arm (note to be consistent with Homework 1, the environment returns the loss instead of reward).

Let's apply the following algorithms:

- Explore-then-exploit Algorithm in Problem 1.
- UCB Algorithm: A classical algorithm for stochastic bandits.
- Thompson Sampling Algorithm: A classical algorithm for stochastic bandits.
- EXP3 Algorithm: A classical algorithm for adversarial bandits.
- Online Mirror Descent with a log-barrier regularizer.
- Online Mirror Descent with $\Psi(x) = -\sum_i \sqrt{x_i}$ in Problem 2.

Like in Homework 1, you are supposed to choose the proper learning rates and plot the trajectories of algorithms.

Please read the code sample and implement the algorithms with Python 3.

Note after you submit the code, we will also test your algorithm in other environments.