Online Optimization and Learning (CS245)

Name: ID: Email:

Rules:

1. Deadline: **2025-03-29/23:59:59**.
   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+CS245HW1.zip".
4. Plagiarism is not allowed. You will fail this homework if any plagiarism is detected.

Problem 1: Adaptive Online Learning.

Online Mirror Desecent

Initialization: $x_1 \in \mathcal{K}$, learning rate η_t and regularizer $\psi(\cdot)$.
For $t = 1, \cdots, T$:

- **Learner:** Submit x_t.
- **Environment:** Observe the loss gradient l_t.
- **Update:** $x_{t+1} = \arg\min_{x \in \mathcal{K}} \ \langle l_t, x \rangle + \frac{1}{\eta_t} B_\psi(x; x_t)$.

Follow-The-Regularized-Leader

Initialization: $x_1 \in \mathcal{K}$, learning rate η_t and regularizer $R(\cdot)$.
For $t = 1, \cdots, T$:

- **Learner:** Submit x_t.
- **Environment:** Observe the loss function $f_t(x) = \langle l_t, x \rangle$
- **Update:** $x_{t+1} = \arg\min_{x \in \mathcal{K}} \ \sum_{s=1}^{t} f_s(x) + \frac{1}{\eta_t} R(x)$.

Here we introduce the Online Mirror Descent (OMD) and Follow-The-Regularized-Leader (FTRL) algorithms for linear loss function $f_t(x) = \langle l_t, x \rangle \geqslant 0$. Let's consider the following problem (please provide the detailed steps to justify your answers):

- Let regularizer $\psi(x) = R(x) = \frac{1}{2}\|x\|^2$, please design adaptive learning rate for η_t to prove that OMD and FTRL achieves sub-linear regret, where the regret is defined as usually

$$\mathcal{R}(T) = \sum_{t=1}^{T} f_t(x_t) - \min_{x \in \mathcal{K}} \sum_{t=1}^{T} f_t(x).$$

- FTRL and OMD both have regularization terms, can you explain the connections between these two algorithms?

Problem 2: Online Mirror Descent for LLM Preference Alignment.

Reinforcement learning from human feedback (RLHF) has effectively aligned large language models (LLMs) with human preferences. Given a prompt x_t sampled from the dataset according to a specific distribution, the LLM model samples a pair of responses (y_t^1, y_t^2) from the policy distribution $\pi_t \in \Pi$ (where Π represents the set of all possible response distributions). A human evaluator then selects the preferred response, and the LLM is fine-tuned based on this preference feedback.

Here, we consider a simplified formulation of the problem. After choosing the policy π_t, the LLM model has access to the reward function $r_t(y) = \mathbb{E}_{x_t}[\mathbb{P}(y \succ \pi_t)|x_t] = \mathbb{E}_{y' \sim \pi_t}[\mathbb{P}(y \succ y')]$ (which denotes the expected win rate of response y against the current policy π_t). Below, we present a Mirror Descent Algorithm designed to address this problem.

Online Mirror Descent

Initialization: $\pi_1 \in \Pi$, Number of iterations T, learning rate η, preference oracle $\mathbb{P}$.
For $t = 1, \cdots, T$:

- **Learner:** Submit π_t and construct response pairs.
- **Environment:** The preference oracle $\mathbb{P}$ outputs the reward function r_t.
- **Algorithm:** (Given the prediction M_{t+1},) Conduct the following (Optimistic) Online Mirror Descent to output the next-step policy π_{t+1}:

$$\pi_{t+1} = \arg\max_{\pi \in \Pi} \langle \pi, r_t \rangle - \frac{1}{\eta} \mathrm{KL}(\pi \| \pi_t).$$

Let's consider the following problem (please provide the detailed steps to justify your answers):

- Can you provide the closed-form updates for the above Online Mirror Descent algorithm?
- To judge the performance of the algorithm, let $\bar{\pi} := \frac{1}{T} \sum_{t=1}^{T} \pi_t$, $J(\pi_1, \pi_2) := \mathbb{E}_{x, y^1 \sim \pi_1,\ y^2 \sim \pi_2}[\mathbb{P}(y^1 \succ y^2)|x] = \mathbb{E}_{y^1 \sim \pi_1,\ y^2 \sim \pi_2}[\mathbb{P}(y^1 \succ y^2)]$, the corresponding DualGap is defined as

  $$\mathrm{DualGap}(\bar{\pi}) := \max_{\pi_1} J(\pi_1, \pi) - \min_{\pi_2} J(\pi, \pi_2).$$

  When $\mathrm{DualGap}(\bar{\pi}) = 0$, it indicates that the Nash equilibrium has been reached. Consequently, the objective is to minimize the DualGap, as it serves as a measure of how closely a policy approximates the Nash equilibrium. Please choose proper η and prove the DualGap of **Online Mirror Descent**.
- For this problem, please write up the **Optimistic Online Mirror Descent** algorithm, set proper learning rates, and prove the DualGap.

(Hint: Find the relationship between *Regret* and *Dual Gap*, and recall the symmetric nature of the game.)

Problem 3: (Adaptive) Gradient Descent Algorithms for Linear Regression.

In this problem, you will train a linear regression model to predict estate prices.

The dataset is available at: `https://archive.ics.uci.edu/dataset/477/real+estate+valuation+data+set`, where the first six fields are the features of the estate and the last field is the price. For this dataset, you need to:

- Split the data into train and test sets in a proper way and utilize a linear regression model to predict the evaluation level. You are asked to implement Gradient Descent Algorithm, Adaptive Gradient Descent Algorithm, Root Mean Square Propagation Algorithm, and Adam Algorithm in training your model. (You may need to preprocess the data before training to ensure convergence.)
- Please plot your training losses, test your models on your test dataset, and compare their performance.

Please implement algorithm code with Python 3, and make sure your code works.