Online Optimization and Learning (CS245)

Name: ID: Email:

Rules:

1. Deadline: **2022/04/25/00:00:00**.
   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: Online Mirror Descent for Adversarial Bandits

We have discussed Online Mirror Descent for Adversarial Bandits as follows.

Online Mirror Descent for Adversarial Bandits

Initialization: $x_1 = [1/K, ..., 1/K]$ and learning rate η.
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:** $x_{t+1} = \arg\min_{x \in \mathcal{K}} \langle x, -\hat{r}_t \rangle + \frac{1}{\eta} B_\psi(x; x_t)$

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 $\psi(x) = -\sum_{i=1}^K \sqrt{x_i}$.

- Please provide the regret analysis of the algorithm with a proper fixed learning rate η.
- Please provide the regret analysis of the algorithm with a proper adaptive learning rate η_t.
- Can you compare the regret of the algorithms with that of EXP3?

Solution:

Let's focus on adaptive learning rate (bonus)

We use loss $l_t(\hat{l}_t)$ to denote $-r_t(-\hat{r}_t)$ and suppose $r_t \in [0, 1]$ in the following proof. First, we will try to prove it using Lemma 3 in Lecture 4, which gives that for any $x \in \mathcal{K}$:

$$\langle x_t - x, \hat{l}_t \rangle \leqslant \frac{1}{\eta_t}(B(x; x_t) - B(x; x_{t+1})) + \frac{\eta_t}{2} \min \left\{ ||\hat{l}_t||^2_{(\nabla\psi^2(z_t))^{-1}}, ||\hat{l}_t||^2_{(\nabla\psi^2(z_t^{'}))^{-1}} \right\},$$

where z_t is between x_t and x_{t+1}; $z_t^{'}$ is between x_t and $x_{t+1}^{'}$ with $x_{t+1}^{'} = \arg\min \langle x, l_t \rangle + \frac{1}{\eta} B_\psi(x; x_t)$. Due to the formulation of $\psi(\cdot)$, we can observe that the Hessian of this regularizer is still diagonal and

$$\nabla\psi^2(z_t)_{i,i} = \frac{1}{4x_t^{3/2}(i)}$$

Then we can get:

$$\begin{aligned}\sum_{t=1}^T \langle x_t - x, \hat{l}_t \rangle &\leqslant \sum_{t=1}^T \frac{1}{\eta_t}(B(x; x_t) - B(x; x_{t+1})) + 2\sum_{t=1}^T \eta_t \sum_{i=1}^K \hat{l}_t^2(i) x_t^{3/2}(i) \\ &\leqslant \frac{B(x; x_1)}{\eta_1} + \sum_{t=1}^{T-1} (\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}) B(x; x_{t+1}) + 2\sum_{t=1}^T \eta_t \sum_{i=1}^K \hat{l}_t^2(i) x_t^{3/2}(i),\end{aligned}$$

Note:
With the inequality above, we can address Question 1 by letting $\eta_t = \eta$ (fixed learning rate). However in the case of time-varying learning rate, we are in trouble because $B_\psi(x; y) = \sum_{i=1}^K -\sqrt{x} + \sqrt{y} + \frac{x-y}{2\sqrt{y}} = \sum_{i=1}^K -\sqrt{x} + \frac{1}{2}\sqrt{y} + \frac{x}{2\sqrt{y}}$, where we are not able to provide an upper bound on $B_\psi(x; x_{t+1})$ since $x_{t+1,i}$ could be extremely small.

With adaptive learning rates, we need to introduce the following lemma, which is an "adaptive" version of Lemma 3 in Lecture 4.

Lemma 1 *Let ψ be twice-differentiable and with the Hessian positive definite in the interior of their domains, and $\frac{1}{\eta_{t+1}}\psi(x) \geqslant \frac{1}{\eta_t}\psi(x)$, $\forall x \in \mathcal{K}$. Online mirrored descent algorithm achieves*

$$\sum_{t=1}^{T}\langle x_t - x, \hat{l}_t\rangle \leqslant \frac{1}{\eta_T}\psi(x) - \frac{1}{\eta_1}\psi(x_1) + \frac{1}{2}\sum_{t=1}^{T}\eta_t \min\{||\hat{l}_t||^2_{(\nabla\psi^2(z_t))^{-1}}, ||\hat{l}_t||^2_{(\nabla\psi^2(z_t^{'}))^{-1}}\},$$

where z_t is between x_t and x_{t+1}; $z_t^{'}$ is between x_t and $x_{t+1}^{'}$ with $x_{t+1}^{'} = \arg\min\langle x, \hat{l}_t\rangle + \frac{1}{\eta_t}B_\psi(x; x_t)$.

Remark 1 *This is **Lemma 7.13** in **A Modern Introduction to Online Learning**, you can find detailed proof in corresponding chapter. The proof is a bit similar to the proof of Lemma 3 in lecture4. It transforms the online mirror descent update to a FTRL update: $x_{t+1} = \arg\min_{x\in\mathcal{K}}\sum_{i=1}^{t}\langle \hat{l}_i, x\rangle + \frac{1}{\eta_t}\psi(x)$. Actually they are equivalent with linear losses, you can easily prove it through optimal condition.*

Set $\eta_t \propto 1/\sqrt{t}$, to satisfy $\frac{1}{\eta_{t+1}}\psi(x) \geqslant \frac{1}{\eta_t}\psi(x)$, we let $\hat{\psi}(x) = \sqrt{K} + \psi(x) \geqslant 0$ (that holds since $\mathcal{K}$ is a set of probability simplex). Then our online mirror descent update is equal to $x_{t+1} = \arg\min_{x\in\mathcal{K}}\langle x, \hat{l}_t\rangle + \frac{1}{\eta_t}B_{\hat{\psi}}(x; x_t)$.

From Lemma 1 above and Hessian of the modified regularizer, we have

$$\textit{Regret} \leqslant \frac{1}{\eta_T}\hat{\psi}(x^*) - \frac{1}{\eta_1}\hat{\psi}(x_1) + 2\mathbb{E}[\sum_{t=1}^{T}\eta_t\sum_{i=1}^{K}\hat{l}_t^2(i)x_t^{3/2}(i)],$$

Let's focus on the last term

$$\begin{aligned}
\mathbb{E}[\sum_{i=1}^{K}\hat{l}_t^2(i)x_t^{3/2}(i)] &= \mathbb{E}[\mathbb{E}[\sum_{i=1}^{K}\hat{l}_t^2(i)x_t^{3/2}(i)|A_1, \ldots, A_{t-1}]] \\
&= \mathbb{E}[\sum_{i=1}^{K}\frac{l_t^2(i)}{x_t^2(i)}x_t^{3/2}(i)x_t(i)] \\
&= \mathbb{E}[\sum_{i=1}^{K}l_t^2(i)\sqrt{x_t(i)}] \\
&\leqslant \mathbb{E}[\sqrt{\sum_{i=1}^{K}l_t^2(i)}\sqrt{\sum_{i=1}^{K}l_t^2(i)x_t(i)}] \\
&\leqslant \sqrt{K},
\end{aligned}$$

where the first equality comes from Adam's Law that $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$ and A_t denotes event that you have made decision in round t.

Combine all above inequalities and recall $\eta_t \propto 1/\sqrt{t}$, we have

$$\begin{aligned}
\textit{Regret} &\leqslant \sqrt{T}\hat{\psi}(x^*) + 2\sqrt{K}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \\
&\leqslant 3\sqrt{TK},
\end{aligned}$$

which gives $O(\sqrt{KT})$ regret bound.

Problem 2: 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 three algorithms:

- UCB Algorithm: A classical algorithm for stochastic bandits.
- EXP3 Algorithm: A classical algorithm for adversarial bandits.
- Online Mirror Descent with $\Psi(x) = -\sum_i \sqrt{x_i}$ in Problem 1.

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

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

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

Note:
In OMD with Tsallis Entropy, we don't have an explicit update for x_{t+1} like in EXP3. To implement this algorithm, you can use Newton's method approximation or solver package likes 'CVXPY' to get the answer of this convex problem.
It doesn't matter if you choose a bad learning rate, but the algorithm should be right.