Online and Reinforcement Learning (2025)
Home Assignment 7

Davide Marchi 777881

Contents

1 Offline Evaluation of Bandit Algorithms - the Practical Part

1.1 Part 1

Modified UCB1 for Offline Evaluation

In this subsection, we present a modified version of the UCB1 algorithm for offline evaluation under a uniform logging policy. Given a log of T rounds in which, at each round t, the logging policy selects an arm A_t uniformly at random and observes a loss $\ell_{A_t,t} \in [0,1]$, we replay UCB1. Updates occur only when the arm chosen by UCB1 matches the logged arm; in such cases, the observed loss is scaled by the importance weight K to yield an unbiased estimate.

Offline UCB1 with Importance Weighting:

1. **Input:** Number of arms K, total rounds T, logged data $\{(A_t, \ell_{A_t,t})\}_{t=1}^{T}$, constant $c > 0$.
2. **Initialize:** For each arm $i = 1, \ldots, K$, set $L_i \leftarrow 0$ and $N_i \leftarrow 0$.
3. **For** $t = 1$ to T:
   (a) For $i = 1$ to K:
   - If $N_i = 0$, set $UCB_i \leftarrow +\infty$.
   - Otherwise, set
   $$UCB_i \leftarrow \frac{L_i}{N_i} + c\sqrt{\frac{\ln(t)}{N_i}}.$$
   (b) Let $i_t \leftarrow \arg\min_i UCB_i$.
   (c) **If** $A_t = i_t$, then update
   $$L_{A_t} \leftarrow L_{A_t} + K\,\ell_{A_t,t}, \quad N_{A_t} \leftarrow N_{A_t} + 1.$$
4. **Output:** The pairs $\{(L_i,\, N_i)\}_{i=1}^{K}$.

Modified EXP3 for Offline Evaluation (Anytime Version)

Here, we describe a modified version of the EXP3 algorithm adapted for offline evaluation in the anytime setting. With a uniform logging policy, the logged data is used to update the cumulative importance-weighted loss only when the logging policy's chosen arm coincides with the arm that the EXP3 algorithm would have drawn. In the anytime version, the learning rate is set to vary with time as

$$\eta_t = \sqrt{\frac{2\ln(K)}{K\,t}}.$$

Offline EXP3 with Importance Weighting:

1. **Input:** Number of arms K, total rounds T, logged data $\{(A_t, \ell_{A_t,t})\}_{t=1}^{T}$.
2. **Initialize:** For each $i = 1, \ldots, K$, set $w_i \leftarrow 1$ and $\tilde{L}_i \leftarrow 0$.
3. **For** $t = 1$ to T:
   (a) Set
   $$\eta_t \leftarrow \sqrt{\frac{2\,\ln(K)}{K\,t}}.$$
   (b) Compute
   $$W \leftarrow \sum_{i=1}^{K} \exp\bigl(-\eta_t\,\tilde{L}_i\bigr).$$
   (c) For each arm i, set
   $$p_t(i) \leftarrow \frac{\exp(-\eta_t\,\tilde{L}_i)}{W}.$$
   (d) **If** A_t is the arm drawn according to p_t, then update
   $$\tilde{L}_{A_t} \leftarrow \tilde{L}_{A_t} + K\,\ell_{A_t,t}.$$
4. **Output:** The final cumulative losses $\{\tilde{L}_i\}_{i=1}^{K}$ and the distribution $\{p_T(i)\}_{i=1}^{K}$.