Online and Reinforcement Learning (2025)
Home Assignment 6

Davide Marchi 777881

Contents

1 PPO

2 Offline Evaluation of Bandit Algorithms

3 Part 1

Evaluating algorithms for online learning with limited feedback in real-life scenarios is challenging. While the most direct approach is to implement an algorithm and measure its performance live, this is often not feasible due to:

- **Potential risks, costs, and time delays.** Deploying a potentially suboptimal algorithm can lead to high financial or reputational costs. Moreover, once the algorithm is running, it takes time to collect sufficient data for analysis, which delays insights and can be inefficient if the algorithm underperforms.
- **Difficulty of controlled experimentation.** Once data is collected based on a particular algorithm's actions, it is nearly impossible to "replay" the same conditions to test different algorithms under identical circumstances. This lack of controlled repetition makes fair comparisons difficult.

4 Part 2

(a) Modification of UCB1 for Importance-Weighted Losses (Uniform Sampling)

To handle partial feedback when arms are chosen uniformly at random (with probability $1/K$ for each arm), we replace the usual empirical loss estimates in UCB1 with importance-weighted estimates. Concretely, whenever an arm i is chosen, its observed loss is scaled by the factor $\frac{1}{p_i(t)} = K$, ensuring an unbiased estimator. The main changes from standard UCB1 are thus:

- **Empirical Loss Update:** Instead of adding the raw observed loss $\ell_{i,t}$, we add $K \cdot \ell_{i,t}$ to the running total for arm i.
- **Confidence Bounds:** The increased variance (due to multiplying by K) is accounted for in the confidence term, typically by a constant factor in front of the usual $\sqrt{\frac{\ln t}{N_i(t)}}$ bound.

Pseudo-Code of the Modified Algorithm:

Initialize: For each arm $i \in \{1, \ldots, K\}$,

$$\widehat{L}_i(0) = 0, \quad N_i(0) = 0.$$

For $t = 1$ **to** T:

1. Select arm
$$A_t = \arg\min_{i \in \{1,\dots,K\}} \left(\widehat{L}_i(t-1) + c\sqrt{\frac{\ln(t-1)}{N_i(t-1)}} \right),$$
where $c > 0$ is a constant.

2. Observe the loss $\ell_{A_t,t}$ for the chosen arm.

3. Update counts:
$$N_{A_t}(t) = N_{A_t}(t-1) + 1, \quad N_j(t) = N_j(t-1) \quad \text{for } j \neq A_t.$$

4. Update the empirical loss estimate for arm A_t via importance weighting:
$$\widehat{L}_{A_t}(t) = \frac{N_{A_t}(t-1)\,\widehat{L}_{A_t}(t-1) + K\,\ell_{A_t,t}}{N_{A_t}(t)},$$
and set $\widehat{L}_j(t) = \widehat{L}_j(t-1)$ for $j \neq A_t$.

Detailed Regret Analysis.

Setup: Assume K arms with losses $\ell_{i,t} \in [0,1]$ and true mean μ_i for arm i. Under uniform sampling, when an arm is chosen its loss is scaled by K, so the empirical mean for arm i after n pulls is:
$$\widehat{\mu}_i(n) = \frac{1}{n}\sum_{s=1}^{n} K\,\ell_{i,t_s}.$$
Let $\mu^* = \min_i \mu_i$ and define the pseudo-regret as:
$$R_T = \sum_{t=1}^{T} \mathbb{E}[\ell_{A_t,t}] - T\,\mu^*.$$

Concentration Bound: Since each observed loss is in $[0,K]$, a concentration inequality (e.g., Hoeffding's) implies that with high probability,
$$|\widehat{\mu}_i(n) - \mu_i| \leq \sqrt{\frac{\alpha K^2 \ln T}{n}},$$
for some constant $\alpha > 0$.

UCB1 Index: Define the index for arm i at round t as:
$$I_i(t) = \widehat{\mu}_i(N_i(t)) + c\sqrt{\frac{\ln t}{N_i(t)}},$$

where c is chosen to cover the K-dependent variance. Standard analysis shows that the number of suboptimal pulls for any arm i (with gap $\Delta_i = \mu_i - \mu^*$) is bounded by

$$N_i(T) \leq \frac{C \ln T}{\Delta_i^2},$$

with C including factors from the importance weighting. Thus, the regret is bounded by

$$R_T \leq \sum_{i \neq i^*} \Delta_i \, N_i(T) = O\Big(\sqrt{K \, T \, \ln T}\Big)$$

in the adversarial or worst-case scenario.

Conclusion (a): The modified UCB1, with importance-weighted updates, yields a pseudo-regret of order

$$O\Big(\sqrt{K \, T \, \ln T}\Big),$$

up to constants depending on c and the loss distribution.

(b) Why the Modified UCB1 Cannot Exploit the Small Variance

In the modified UCB1, each observed loss is multiplied by K (since $p_i(t) = 1/K$). Even if the original loss distribution has small variance, the effective variance is inflated by the factor K. Since the algorithm builds confidence intervals based on worst-case deviations, the benefit of a small underlying variance is lost. In other words, the importance-weighted update makes the algorithm behave conservatively—as if the variance were higher—thus preventing exploitation of any small-variance structure.

(c) Modifying EXP3 for Importance-Weighted Losses (Uniform Sampling)

We consider a logging policy that selects arms uniformly at random (with probability $1/K$ each). The EXP3 algorithm is modified so that when arm A_t is selected and its loss $\ell_{A_t,t}$ is observed, an unbiased estimator for the loss of each arm is formed by

$$\tilde{\ell}_{i,t} = \begin{cases} K \, \ell_{A_t,t} & \text{if } i = A_t, \\ 0 & \text{otherwise.} \end{cases}$$

Then, the exponential weights update is applied:

$$w_i(t+1) = w_i(t) \, \exp\big(-\eta \, \tilde{\ell}_{i,t}\big),$$

with initial weights $w_i(1) = 1$ for all i.

Pseudo-Code for Modified EXP3:

1. **Initialization:** Set $w_i(1) = 1$ for each $i \in \{1, \ldots, K\}$.
2. **For each round** $t = 1, 2, \ldots, T$:
   (a) **(Logging policy)** An arm A_t is chosen uniformly at random (i.e., with probability $1/K$).
   (b) **(Observe loss)** Observe the loss $\ell_{A_t,t}$ for the chosen arm.
   (c) **(Form importance-weighted loss)** For each arm i, define
   $$\tilde{\ell}_{i,t} = \begin{cases} K\,\ell_{A_t,t} & \text{if } i = A_t, \\ 0 & \text{otherwise.} \end{cases}$$
   (d) **(Update weights)** Update
   $$w_i(t+1) = w_i(t)\,\exp\bigl(-\eta\,\tilde{\ell}_{i,t}\bigr),$$
   where $\eta > 0$ is the learning rate.

Detailed Regret Analysis.

Setup: Let $L_i = \sum_{t=1}^{T} \ell_{i,t}$ be the cumulative loss of arm i and $L^* = \min_i L_i$ be the loss of the best arm. The (pseudo-)regret is:

$$R_T = \sum_{t=1}^{T} \ell_{A_t,t} - L^*.$$

We aim to show that $\mathbb{E}[R_T] = O\Bigl(\sqrt{K\,T\,\ln K}\Bigr)$.

Step 1: Master Inequality. Define the potential $\Phi_t = \ln W(t)$ where $W(t) = \sum_{i=1}^{K} w_i(t)$. By the exponential weights update and a standard inequality we have:

$$\Phi_{t+1} - \Phi_t \le -\eta \sum_{i=1}^{K} p_t(i)\,\tilde{\ell}_{i,t} + \frac{\eta^2}{2} \sum_{i=1}^{K} p_t(i)\,\tilde{\ell}_{i,t}^2,$$

with $p_t(i) = \frac{w_i(t)}{W(t)}$. Summing over $t = 1$ to T,

$$\Phi_{T+1} - \Phi_1 \le -\eta \sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i)\,\tilde{\ell}_{i,t} + \frac{\eta^2}{2} \sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i)\,\tilde{\ell}_{i,t}^2.$$

Since $W(1) = K$ (so $\Phi_1 = \ln K$), we next lower-bound Φ_{T+1}.

Step 2: Lower Bound on Φ_{T+1}. For any fixed arm i,

$$w_i(T+1) = \exp\Bigl(-\eta \sum_{t=1}^{T} \tilde{\ell}_{i,t}\Bigr).$$

Thus,

$$W(T+1) \geq w_{i^*}(T+1) = \exp\Big(-\eta \sum_{t=1}^{T} \tilde{\ell}_{i^*,t}\Big),$$

where i^* is the best arm. Taking logs,

$$\Phi_{T+1} \geq -\eta \sum_{t=1}^{T} \tilde{\ell}_{i^*,t}.$$

Step 3: Combine the Bounds. We have:

$$-\eta \sum_{t=1}^{T} \tilde{\ell}_{i^*,t} - \ln K \leq -\eta \sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i)\, \tilde{\ell}_{i,t} + \frac{\eta^2}{2} \sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i)\, \tilde{\ell}_{i,t}^2.$$

Rearrange to obtain:

$$\sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i)\, \tilde{\ell}_{i,t} - \sum_{t=1}^{T} \tilde{\ell}_{i^*,t} \leq \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{K} p_t(i)\, \tilde{\ell}_{i,t}^2.$$

Step 4: Relate to Actual Losses. Since only the chosen arm contributes,

$$\sum_{i=1}^{K} p_t(i)\, \tilde{\ell}_{i,t} = p_t(A_t)\, (K\, \ell_{A_t,t}).$$

Under uniform sampling, in expectation $p_t(A_t) = \frac{1}{K}$; hence,

$$\mathbb{E}\Big[\sum_{i=1}^{K} p_t(i)\, \tilde{\ell}_{i,t}\Big] = \ell_{A_t,t}.$$

Similarly, one shows that

$$\sum_{i=1}^{K} p_t(i)\, \tilde{\ell}_{i,t}^2 \leq K\, \ell_{A_t,t},$$

since $\ell_{A_t,t} \in [0,1]$.

Taking expectations and summing over t,

$$\mathbb{E}\Big[\sum_{t=1}^{T} \ell_{A_t,t}\Big] - \sum_{t=1}^{T} \ell_{i^*,t} \leq \frac{\ln K}{\eta} + \frac{\eta K}{2}\, \mathbb{E}\Big[\sum_{t=1}^{T} \ell_{A_t,t}\Big].$$

Since $\ell_{A_t,t} \leq 1$ implies $\mathbb{E}[\sum_{t=1}^{T} \ell_{A_t,t}] \leq T$, we have

$$\mathbb{E}[R_T] \leq \frac{\ln K}{\eta} + \frac{\eta K T}{2}.$$

Choosing

$$\eta = \sqrt{\frac{2 \ln K}{K T}},$$

balances the two terms and gives

$$\mathbb{E}[R_T] \leq 2\sqrt{2}\,\sqrt{K\,T\,\ln K}.$$

Thus, the expected regret is

$$\mathbb{E}[R_T] = O\Big(\sqrt{K\,T\,\ln K}\Big).$$

(d) Anytime Modification of EXP3

For the anytime version of EXP3 (which does not assume knowledge of a fixed horizon T), we replace the constant learning rate with one that depends on the current round t. A common choice is:

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

With this time-varying learning rate, the expected regret bound remains of the same order as the fixed-horizon version:

$$O\Big(\sqrt{K\,T\,\ln K}\Big),$$

though with a slightly larger constant factor than if T were known in advance.