OReL Assignment 1

bkx591

February 2024

1 Follow the Lead (FTL) algorithm fo i.i.d full information games (25 points)

I seek to derive a bound for the pseudo regret of FTL in i.i.d. full information games with K possible actions, and outcomes bounded in the $[0,1]$ interval. The FTL algorithm can be formalized as follows: At each round t, select the action a_t such that $a_t = \arg\max_a \hat{\mu}_{t-1}(a)$ (in the setting of rewards and not loss).

Let us define the following terms:

- $\mu(a)$: Expected reward of action a.
- $\hat{\mu}_t(a)$: Empirical estimate of the reward of action a at round t.
- a^*: Optimal action with the highest expected reward.
- $\Delta(a)$: Sub-optimality gap, $\Delta(a) = \mu(a^*) - \mu(a)$.
- $N_T(a)$: Number of times action a was played up to round T.

Given $K = 2$, FTL plays a suboptimal arm, $a \neq a^*$, at rounds t, for which $\hat{\mu}_{t-1}(a) \geq \max_{a'} \hat{\mu}_{t-t}(a')$ with probability

$$\mathbb{P}(\hat{\mu}_{t-1}(a) \geq \hat{\mu}_{t-1}(a^*))$$

Intuitively this probability becomes smaller and smaller as t increases, since the empirical mean of any action gets closer to the actual mean, so that if Δ is large, the probability of choosing a suboptimal action becomes smaller even faster per the definition of FTL.

Let us analyze how often this happens by looking at the stochastic nature of the rewards, where even the suboptimal action can have a higher empirical mean due to random fluctuations in the short term. Let us look at the expected number of times the suboptimal action is played which can be written as $\mathbb{E}[N_T(a)]$, where $N_T(a)$ is the number of times action a is played up to round T. We have that $N_T(a) = \sum_{t=1}^{T} \mathbb{1}(A_t = a)$ such that

$$\mathbb{E}[N_T(a)] \leq \sum_{t=1}^{T} \mathbb{P}(\hat{\mu}_{t-1}(a) \geq \hat{\mu}_{t-1}(a^*)) \tag{1}$$

This smells like Hoeffding, so let's try to bound the probability of the event where the empirical mean of action a is at least as large as the empirical mean of the optimal action a^* at time $t-1$ and call that event E_t. We need to account for the possibility of either action's ($K = 2$) empirical mean being misleading. Following the analysis approach from the slides on Exploration-Exploitation trade-off, we consider two events of the empirical mean being on the "wrong" side of the separation line. Formally we define two events A_t^* and A_t in which cases, the event E_t would occur.

$$\begin{aligned} A_t^* &= \{\hat{\mu}_{t-1}(a^*) \leq \mu_{t-1}(a^*) - \frac{\Delta}{2}\} \\ A_t &= \{\hat{\mu}_{t-1}(a) \geq \mu_{t-1}(a) + \frac{\Delta}{2}\} \end{aligned} \tag{2}$$

Now let's do two times Hoeffding and get the following

$$\mathbb{P}(A_t^*) = \mathbb{P}(\hat{\mu}_{t-1}(a^*) - \mu_{t-1}(a^*) \geq \frac{\Delta}{2}) \leq e^{-2t(\frac{\Delta}{2})^2} \tag{3}$$

$$\mathbb{P}(A_t) = \mathbb{P}(\hat{\mu}_{t-1}(a) - \mu_{t-1}(a) \geq \frac{\Delta}{2}) \leq e^{-2t(\frac{\Delta}{2})^2} \tag{4}$$

$$\begin{aligned} \mathbb{P}(E_t) &= \mathbb{P}(\hat{\mu}_{t-1}(a) \geq \hat{\mu}_{t-1}(a^*)) && (5) \\ &= \mathbb{P}(A_t^* \cup A_t) && (6) \\ &\leq \mathbb{P}(A_t^*) + \mathbb{P}(A_t) && (7) \\ &= 2e^{-2t(\frac{1}{2}\Delta)^2} && (8) \end{aligned}$$

Step 7 is by union bound. Now that we have the probability of such an event at time t, we need to sum it up for time T which gives us the sum

$$\sum_{t=1}^{T} \mathbb{P}(E_t) = \sum_{t=1}^{T} 2e^{-2t(\frac{\Delta}{2})^2} = 2\sum_{t=1}^{T} e^{-2t(\frac{\Delta}{2})^2} \tag{9}$$

which is the sum of a geometric series, mentioned in guideline 6. In our case we are interested in the sum as T goes to infinity, which gives us the simplified convergence of $\sum_{t=0}^{\infty} r^t = \frac{1}{1-r}$ with $r = e^{-2(\frac{\Delta}{2})^2}$. Notice that this step implies that in this full information i.i.d. setting, the regret does not grow with time. We now have the following

$$2\sum_{t=0}^{\infty} e^{-2t(\frac{\Delta}{2})^2} = \frac{2}{1 - e^{-2(\frac{\Delta}{2})^2}} \tag{10}$$

The pseudo regret $\bar{R}_T$ is the sum of the expected losses due to playing the sub-optimal action. Each time the sub-optimal action is selected, we incur a loss of Δ, so we have

$$\bar{R}_T \leq \sum_{a:\Delta>0} \Delta \frac{2}{1 - e^{-2(\frac{\Delta}{2})^2}} \tag{11}$$

2 Improved Parametrization of UCB1

Note: Pay attention to the cumulative regret values (y-axis) in the plots. Initially, the plots might look identical, but the regret value-range differs greatly.

Which values of Δ lead to higher regret?

We observe that lower values of Δ leads to higher regret, which can be explained by the nature of the exploration-exploitation trade-off in multi-armed bandit problems. A lower delta implies a smaller difference in the expected rewards between the best arm and the sub-optimal ones. Therefore, the algorithm spends more time exploring to confidently identify the best arm, which inherently leads to more regret accumulated during this exploration phase.

What can you say about the relative performance of the two parametrizations?

The modified UCB1 with improved confidence intervals is theoretically designed to provide a tighter confidence bound. This tighter bound means that the algorithm is expected to be more efficient in distinguishing between the arms' rewards, leading to quicker convergence to the optimal arm and thus lower cumulative regret. The empirical simulations comparing the two, shows how significant this improvement is in practice. If the simulations align with the theoretical expectations, you would observe that the modified UCB1 outperforms the standard UCB1, especially as the number of rounds, T, increases and the difference in delta becomes more critical for performance.

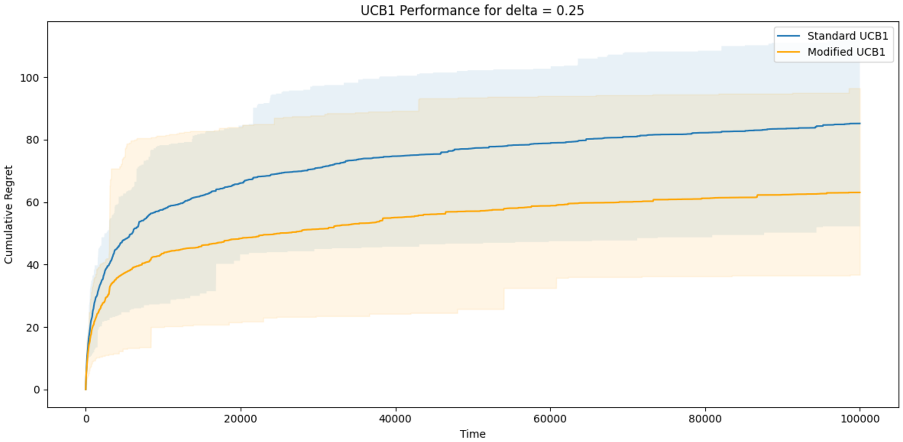

Figure 1: Empirical pseudo regret over time for $T = 100000$, $S = 20$ simulations, and $\Delta = \frac{1}{4}$, for the UCB1 algorithm (blue) and the modified version from the assignment (orange)

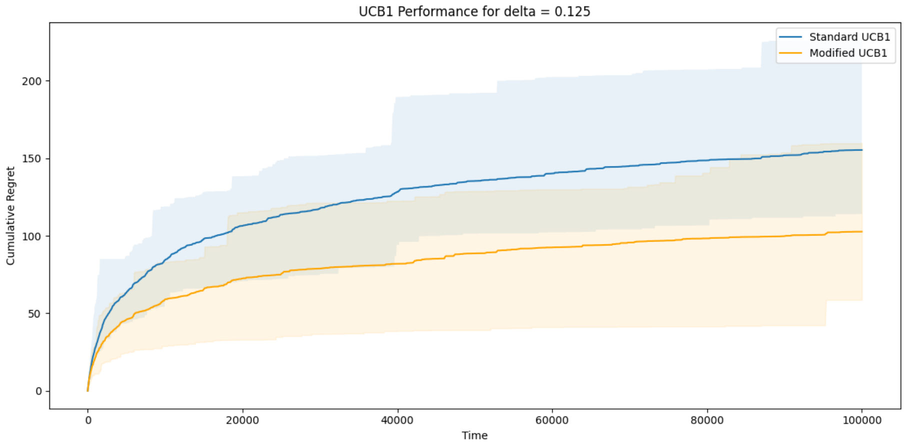

Figure 2: Empirical pseudo regret over time for $T = 100000$, $S = 20$ simulations, and $\Delta = \frac{1}{8}$, for the UCB1 algorithm (blue) and the modified version from the assignment (orange)

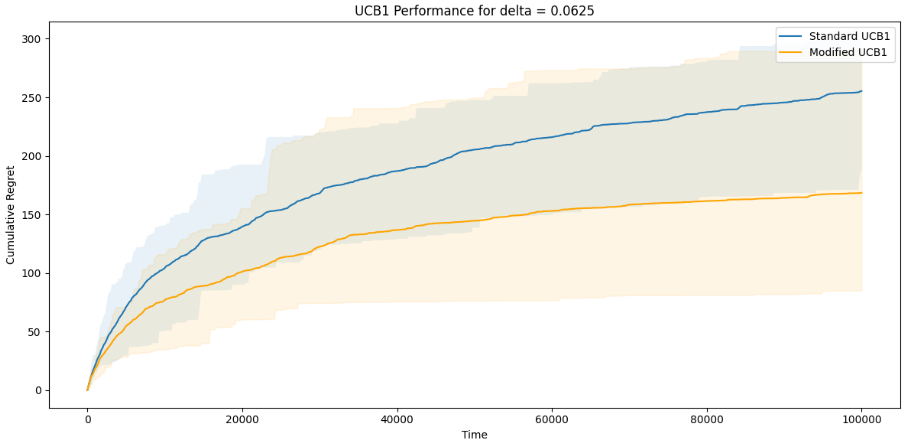

Figure 3: Empirical pseudo regret over time for $T = 100000$, $S = 20$ simulations, and $\Delta = \frac{1}{16}$, for the UCB1 algorithm (blue) and the modified version from the assignment (orange)

3 Example of Policies in RiverSwim

Considering the given RiverSwim MDPs, I classify them as follows:

(i) $\pi_a \in \prod^{SD}$ as it prescribes a deterministic action based only on the current state.

(ii) $\pi_b \in \prod^{SR}$ because it involves a random action (coin flip) regardless of the state.

(iii) $\pi_c \in \prod^{HD}$ since the action depends on the history (previous state's index).

(iv) $\pi_d \in \prod^{HR}$ as it combines randomness (coin flip) with the state information to decide the action.

(v) $\pi_e \in \prod^{SR}$ if we do not consider the external 'rainy' condition as part of the state. If we do, and the weather changes independently of the agent's state, it remains SR as the policy itself does not have a historical component.

4 Bounds on H-step Values

Considering a discounted MDP M with rewards supported on $[0, 1]$, The H-step value function of a policy π, for a given $H \in \mathbb{N}$ is defined as

$$U^{\pi,H}(s) = \mathbb{E}^{\pi}\left[\sum_{t=1}^{H} \gamma^{t-1} r_t | s_1 = s\right], \forall s \in S$$

(i) Given $\epsilon > 0$, we will determine values of H such that $|U^{\pi,H}(s) - V^{\pi}(s)| \leq \epsilon$ for all s and for all π. Recall that $V^{\pi}(s) = \mathbb{E}^{\pi}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t | s_1 = s\right]$ and let's then first define $U^{\pi,H}(s) - V^{\pi}(s)$:

$$\begin{aligned} |U^{\pi,H}(s) - V^{\pi}(s)| &= |\mathbb{E}^{\pi}\left[\sum_{t=H+1}^{\infty} \gamma^{t-1} r_t | s_1 = s\right]| && (12) \\ &= |\mathbb{E}^{\pi}\left[\gamma^H \sum_{t=H+1}^{\infty} \gamma^{t} r_t | s_1 = s\right]| && (13) \\ &\leq |\sum_{t=H+1}^{\infty} \gamma^{t-1}| && (14) \\ &= |\frac{\gamma^H}{1-\gamma}| && (15) \end{aligned}$$

Step 13 modifies the sum, such that we can move the first γ^H out of the sum. Step 14 and 15 follows from the infinite sum of a geometric series where the sum does not start at $t = 0$. Next up, we want this sum to be at most ϵ, so we solve for H:

$$\frac{\gamma^H}{1-\gamma} \leq \epsilon \quad (16)$$

$$\gamma^H \leq \epsilon(1-\gamma) \quad (17)$$

$$H \log(\gamma) \leq \log(\epsilon(1-\gamma)) \quad (18)$$

$$H \geq \frac{\log(\epsilon(1-\gamma))}{\log(\gamma)} \quad (19)$$

Since $0 < \gamma < 1$, we have that $\log(\gamma) < 0$, so we must have the following:

$$H \geq \lceil | \frac{\log(\epsilon(1-\gamma))}{\log(\gamma)} | \rceil \tag{20}$$

(ii) Interpreting the derived bounds gives us an idea of the practical significance of H in the context of a finite-horizon MDP. In many cases it is neither practical, necessary, nor possible to calculate a policy for an infinite horizon, so having a bound that guarantees that the sum of the expected discounted rewards from step $H + 1$ and onwards, is within an ϵ tolerance of the infinite horizon value. In other words, it allows us to reasonably ignore long term effects and limit how far into the future we need to plan.

5 Policy Evaluation in RiverSwim

Considering the 5-state RiverSwium MDP with $\gamma = 0.95$ and the policy π (as defined in the assignment), I've calculated the Monte Carlo simulation to V^{π} using $T = 200$, and $n = 100$, as well as calculated the exact value through direct computation using Bellman equation $V^{\pi} = (I - \gamma P^{\pi})^{-1} r^{\pi}$ and gotten the following values:

State	Estimated V^{π} (Monte Carlo)	Exact V^{π}
$s1$	1.82008716	1.60365251
$s2$	1.74905262	1.89408739
$s3$	2.60629086	3.15492297
$s4$	7.19907226	7.29485928
$s5$	8.76096827	8.77220123

Table 1: Comparison of estimated and exact state values in RiverSwim MDP

6 Modifications to RiverSwim

(i) Consider the state s_L in the RiverSwim MDP representing the right-hand side bank. We introduce a stochastic reward process $R(s_L, a)$ which is defined by the following probability distribution:

- With probability $p = 0.08$, the agent is fined $R_f = -5000$ DKK for being caught by a beach guard.
- With probability $1 - p = 0.92$, the agent collects an item worth $R_c = 1000$ DKK.

The expected reward for taking any action a in state s_L is then given by the expectation:

$$\mathbb{E}[R(s_L, a)] = p \cdot R_f + (1 - p) \cdot R_c \tag{21}$$

$$\mathbb{E}[R(s_L, a)] = 0.08 \cdot (-5000) + 0.92 \cdot 1000 \tag{22}$$

This expected value must be incorporated into the reward matrix of the MDP to reflect the new reward structure due to the beach guard's presence.

(ii) I don't know if it is even allowed, but I would maintain two policies depending on whether we have the item or not. In both policies I would insert a new state s_{L+1} that would be accessible from s_L with reward 0 s.t. $r(s_L, \text{right}) = 0$. Transitioning to s_{L+1} would engage the second policy, in which only left actions are probable/allowed. After swimming only left, a reward of 1 is given when transitioning to state s_1, after which the initial policy would come into effect again.