With a bit more work it is possible to derive a tight lower bound of $\Omega\left(\sqrt{KT\ln\frac{N}{K}}\right)$ (Chase et al., 2025). And by replacing EXP3-style approach with Tsallis-INF-style techniques it is possible to derive an algorithm with matching $O\left(\sqrt{KT\ln\frac{N}{K}}\right)$ regret upper bound (Kale, 2014).

7.7 Exercises

Exercise 7.1 (*Find an online learning problem from real life*)**.** Find two examples of real life problems that fit into the online learning framework (online, not reinforcement!). For each of the two examples explain what is the set of actions an algorithm can take, what are the losses (or rewards) and what is the range of the losses/rewards, whether the problem is stateless or contextual, and whether the problem is i.i.d. or adversarial, and with full information or bandit feedback.

Exercise 7.2 (*Follow The Leader (FTL) algorithm for i.i.d. full information games*)**.** Follow the leader (FTL) is a playing strategy that at round t plays the action that was most successful up to round t ("the leader"). 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 (you can work with rewards or losses, as you like). You can use the following guidelines (which assume a game with rewards):

1. You are allowed to solve the problem for $K=2$. (The guidelines are not limited to $K=2$.)

2. It may be helpful to write the algorithm down explicitly. For the analysis it does not matter how you decide to break ties.

3. Let $\mu(a)$ be expected reward of action a and let $\hat{\mu}_t(a)$ be empirical estimate of the reward of action a at round t (the average of rewards observed so far). Let a^* be an optimal action (there may be more than one optimal action, but then things only get better [convince yourself that this is true], so we can assume that there is a single a^*). Let $\Delta(a)=\mu(a^*)-\mu(a)$. FTL may play $a\neq a^*$ at rounds t for which $\hat{\mu}_{t-1}(a)\geq\max_{a'}\hat{\mu}_{t-1}(a')$ (in the case of two arms it means $\hat{\mu}_{t-1}(a)\geq\hat{\mu}_{t-1}(a^*)$). So you should analyze how often this may happen.

4. Note that the number of times an action a was played can be written as $N_T(a)=\sum_{t=1}^{T}\mathbb{1}(A_t=a)$, and that $\mathbb{E}\left[\mathbb{1}(A_t=a)\right]\leq\mathbb{P}(\hat{\mu}_{t-1}(a)\geq\max_{a'}\hat{\mu}_{t-1}(a'))\leq\mathbb{P}(\hat{\mu}_{t-1}(a)\geq\hat{\mu}_{t-1}(a^*))$, where $\mathbb{1}$ is the indicator function.

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

6. At some point in the proof you will need to sum up a geometric series. A geometric series is a series of a form $\sum_{t=0}^{\infty}r^t$, and for $r<1$ we have $\sum_{t=0}^{\infty}r^t=\frac{1}{1-r}$. In your case r will be an exponent $r=e^{\alpha}$ for some constant α.

7. At the end you should get a bound of a form $\bar{R}_T\leq\sum_{a:\Delta(a)>0}\frac{c}{1-\exp(-\Delta(a)^2/2)}\Delta(a)$, where c is a constant.

Important observations to make:

1. *Note that in the full information i.i.d. setting the regret does not grow with time!!! (Since the bound is independent of T.)*

2. *Note that even though you have used $\Delta(a)$ in the analysis of the algorithm, you do not need to know it in order to define the algorithm! I.e., you can run the algorithm even if you do not know $\Delta(a)$.*

Exercise 7.3 (*Decoupling exploration and exploitation in i.i.d. multiarmed bandits*)**.** Assume an i.i.d. multiarmed bandit game, where the observations are not coupled to the actions. Specifically, we assume that at each round of the game the player is allowed to observe the reward of a single arm, but it does not have to be the same arm that was played at that round (and if it is not the same arm, the player does not observe its own reward, but instead observes the reward of the alternative option).

Derive a playing strategy and a regret bound for this game. (You should solve the problem for a general K and you should get that the regret does not grow with time.)