L13 – Week 7
Introduction to Multi-armed Bandits (part 2)

CS 295 Optimization for Machine Learning

Ioannis Panageas

Recap of framework (stochastic)

Setting. *We are given K arms and time window T (known). At each time step* $t = 1 \ldots T$.

- *Player chooses arm a_t.*
- *Observes reward $r_t \in [0, 1]$ for the chosen arm.*

- The algorithm observes only the reward for the selected action, and nothing else.
- The reward for each action is IID. For each arm $a \in [K]$, there is a distribution D_a over reals, called the reward distribution (unknown). Every time this action is chosen, the reward is sampled independently from this distribution.

Goal: Minimize the regret

$$\boxed{R(T) = \mu^* T - \sum_{t=1}^{T} \mu(a_t) \text{ or } \mathbb{E}[R(T)].}$$

Upper Confidence Bounds

Definition (Confidence bounds)**.** *We define upper/lower confidence bounds for every arm a and round t*

$$UCB_t(a) = \hat{\mu}_t(a) + r_t(a), \;\; LCB_t(a) = \hat{\mu}_t(a) - r_t(a),$$

where $\hat{\mu}_t(a)$ is the average reward of arm a so far, $r_t(a) = \sqrt{\frac{2\log T}{n_t(a)}}$ (confidence radius) and $n_t(a)$ is the number of samples from arm a in round $1, ..., t$,

Upper Confidence Bounds

Definition (Confidence bounds). *We define upper/lower confidence bounds for every arm a and round t*

$$UCB_t(a) = \hat{\mu}_t(a) + r_t(a), \;\; LCB_t(a) = \hat{\mu}_t(a) - r_t(a),$$

where $\hat{\mu}_t(a)$ is the average reward of arm a so far, $r_t(a) = \sqrt{\frac{2\log T}{n_t(a)}}$ (confidence radius) and $n_t(a)$ is the number of samples from arm a in round $1, \ldots, t$,

Definition (UCB). *Consider the following algorithm:*

1. Try each arm once.
2. In each round t, pick $\arg\max_a UCB_t(a)$.

Analysis of UCB

Remarks:

- An arm a has the largest UCB_t for two reasons: The empirical reward is large (hence it is likely a has high reward) or confidence radius is large, thus the arm has not been explored much.

Either reason makes this arm worth choosing!

Analysis of UCB

Remarks:

- An arm a has the largest UCB_t for two reasons: The empirical reward is large (hence it is likely a has high reward) or confidence radius is large, thus the arm has not been explored much.

Either reason makes this arm worth choosing!

Theorem (Regret). *UCB algorithm achieves regret*

$$\mathbb{E}[R(T)] \text{ to be } O(\sqrt{KT \log T}).$$

Theorem (Regret v2). *UCB algorithm achieves regret*

$$\mathbb{E}[R(T)] \leq O(\log T) \left(\sum_{a:\mu(a)<\mu(a^*)} \frac{1}{\mu(a^*) - \mu(a)} \right).$$

Analysis of UCB

Let us define the "clean" event (we condition on that)

$$\mathcal{E} = \{\forall j, a \; |\hat{\mu}_j(a) - \mu(a)| \leq r_j(a)\}.$$

Let a^* be an optimal arm and assume that we chose arm a_t at time t then:

$$\mu(a_t) + 2r_t(a_t) \geq \hat{\mu}_t(a_t) + r_t(a_t) \text{ clean event}$$

Analysis of UCB

Let us define the "clean" event (we condition on that)

$$\mathcal{E} = \{\forall j, a \ |\hat{\mu}_j(a) - \mu(a)| \leq r_j(a)\}.$$

Let a^* be an optimal arm and assume that we chose arm a_t at time t then:

$$\mu(a_t) + 2r_t(a_t) \geq \hat{\mu}_t(a_t) + r_t(a_t) \text{ clean event}$$

$$= \mathrm{UCB}_t(a_t)$$

Analysis of UCB

Let us define the "clean" event (we condition on that)

$$\mathcal{E} = \{\forall j, a \ |\hat{\mu}_j(a) - \mu(a)| \leq r_j(a)\}.$$

Let a^* be an optimal arm and assume that we chose arm a_t at time t then:

$$\begin{aligned}
\mu(a_t) + 2r_t(a_t) &\geq \hat{\mu}_t(a_t) + r_t(a_t) \text{ clean event} \\
&= \mathrm{UCB}_t(a_t) \\
&\geq \mathrm{UCB}_t(a^*) \text{ since we chose } a_t
\end{aligned}$$

Analysis of UCB

Let us define the "clean" event (we condition on that)

$$\mathcal{E} = \{\forall j, a \; |\hat{\mu}_j(a) - \mu(a)| \leq r_j(a)\}.$$

Let a^* be an optimal arm and assume that we chose arm a_t at time t then:

$$\begin{aligned}
\mu(a_t) + 2r_t(a_t) &\geq \hat{\mu}_t(a_t) + r_t(a_t) \text{ clean event} \\
&= \mathrm{UCB}_t(a_t) \\
&\geq \mathrm{UCB}_t(a^*) \text{ since we chose } a_t \\
&= \mu_t(a^*) + r_t(a^*) \geq \mu(a^*) \text{ clean event.}
\end{aligned}$$

Analysis of UCB

Let us define the "clean" event (we condition on that)

$$\mathcal{E} = \{\forall j, a \; |\hat{\mu}_j(a) - \mu(a)| \leq r_j(a)\}.$$

Let a^* be an optimal arm and assume that we chose arm a_t at time t then:

$$\begin{aligned} \mu(a_t) + 2r_t(a_t) &\geq \hat{\mu}_t(a_t) + r_t(a_t) \text{ clean event} \\ &= \mathrm{UCB}_t(a_t) \\ &\geq \mathrm{UCB}_t(a^*) \text{ since we chose } a_t \\ &= \mu_t(a^*) + r_t(a^*) \geq \mu(a^*) \text{ clean event.} \end{aligned}$$

Hence it holds

$$\boxed{2r_t(a_t) \geq \mu(a^*) - \mu(a_t) = \Delta(a_t).}$$

Analysis of UCB

Hence it holds

$$\boxed{2r_t(a_t) \geq \mu(a^*) - \mu(a_t) = \Delta(a_t).}$$

For each arm a consider the last time τ that a was pulled, then we get $n_T(a) = n_\tau(a)$ and $r_T(a) = r_\tau(a)$. We conclude that

Analysis of UCB

Hence it holds

$$\boxed{2r_t(a_t) \geq \mu(a^*) - \mu(a_t) = \Delta(a_t).}$$

For each arm a consider the last time τ that a was pulled, then we get $n_T(a) = n_\tau(a)$ and $r_T(a) = r_\tau(a)$. We conclude that

$$\boxed{2\sqrt{\frac{2\log T}{n_T(a)}} = 2r_T(a) \geq \mu(a^*) - \mu(a) = \Delta(a).}$$

Analysis of UCB

Hence it holds

$$\boxed{2r_t(a_t) \geq \mu(a^*) - \mu(a_t) = \Delta(a_t).}$$

For each arm a consider the last time τ that a was pulled, then we get $n_T(a) = n_\tau(a)$ and $r_T(a) = r_\tau(a)$. We conclude that

$$\boxed{2\sqrt{\frac{2\log T}{n_T(a)}} = 2r_T(a) \geq \mu(a^*) - \mu(a) = \Delta(a).}$$

The contribution of arm a to the total regret is

$$\Delta(a) \times n_T(a) \leq 2\sqrt{2n_T(a)\log T}.$$

Analysis of UCB

Hence it holds

$$\boxed{2r_t(a_t) \geq \mu(a^*) - \mu(a_t) = \Delta(a_t).}$$

For each arm a consider the last time τ that a was pulled, then we get $n_T(a) = n_\tau(a)$ and $r_T(a) = r_\tau(a)$. We conclude that

$$\boxed{2\sqrt{\frac{2\log T}{n_T(a)}} = 2r_T(a) \geq \mu(a^*) - \mu(a) = \Delta(a).}$$

The contribution of arm a to the total regret is

$$\Delta(a) \times n_T(a) \leq 2\sqrt{2n_T(a)\log T}.$$

Hence the regret is bounded by

$$\boxed{2\sqrt{2\log T}\sum_a \sqrt{n_T(a)}.}$$

Analysis of UCB

Hence the regret is bounded by

$$\boxed{2\sqrt{2\log T}\sum_a \sqrt{n_T(a)}.}$$

Finally observe that $\sqrt{x}$ is a concave function hence, by Jensen's inequality we get

$$\frac{1}{K}\sum_a \sqrt{n_T(a)} \leq \sqrt{\frac{1}{K}\sum_a n_T(a)} \leq \sqrt{\frac{T}{K}}.$$

Analysis of UCB

Hence the regret is bounded by

$$\boxed{2\sqrt{2\log T}\sum_a \sqrt{n_T(a)}.}$$

Finally observe that $\sqrt{x}$ is a concave function hence, by Jensen's inequality we get

$$\frac{1}{K}\sum_a \sqrt{n_T(a)} \le \sqrt{\frac{1}{K}\sum_a n_T(a)} \le \sqrt{\frac{T}{K}}.$$

We conclude that the regret is bounded by

$$\boxed{O(\sqrt{TK\log T}).}$$

Analysis of UCB

Recall that we showed

$$\boxed{2\sqrt{\frac{2\log T}{n_T(a)}} = 2r_T(a) \geq \mu(a^*) - \mu(a) = \Delta(a).}$$

This implies that

$$n_T(a) \leq \frac{8\log T}{\Delta(a)^2}$$

Analysis of UCB

Recall that we showed

$$\boxed{2\sqrt{\frac{2\log T}{n_T(a)}} = 2r_T(a) \geq \mu(a^*) - \mu(a) = \Delta(a).}$$

This implies that

$$n_T(a) \leq \frac{8\log T}{\Delta(a)^2}$$

The contribution of arm a to the total regret is

$$\Delta(a) \times n_T(a) \leq \frac{8\log T}{\Delta(a)}.$$

Hence the regret is bounded by

$$\boxed{O(\log T)\sum_a \frac{1}{\Delta(a)}.}$$

Framework (adversarial bandits)

Setting. *We are given K arms and time window T (known). At each time step* $t = 1...T$.

- *Player chooses arm* a_t.
- *Adversary picks cost* $c_t(a)$ *for each arm* a.
- *Player observes cost* $c_t(a_t) \in [0, 1]$ *for the chosen arm.*

- The player observes only the cost for the selected action, and nothing else.

Goal: Minimize the regret

Framework (adversarial bandits)

Setting. *We are given K arms and time window T (known). At each time step* $t = 1 \ldots T$.

- *Player chooses arm a_t.*
- *Adversary picks cost $c_t(a)$ for each arm a.*
- *Player observes cost $c_t(a_t) \in [0,1]$ for the chosen arm.*

- The player observes only the cost for the selected action, and nothing else.

Goal: Minimize the regret

$$\boxed{R(T) = \sum_{t \in [T]} c_t(a_t) - \min_a \sum_{t \in [T]} c_t(a) \text{ or } \mathbb{E}[R(T)].}$$

MWU (recap)

Algorithm (MWUA). *We define the following algorithm:*

1. Initialize $w_i^0 = 1$ for all $i \in [n]$.
2. **For** t=1 … T **do**
3. **Choose** action i with probability proportional to w_i^{t-1}.
4. **For** each action i **do**
5. $w_i^t = (1-\epsilon)^{c_i^t} w_i^{t-1}$.
6. **End For**
7. **End For**

Remarks:

- We choose i with probability $\mathrm{p}_\mathrm{i}^\mathrm{t} = \frac{\mathrm{w}_i^{t-1}}{\sum_j w_j^{t-1}}$.
- $\mathrm{c}_\mathrm{i}^\mathrm{t}$ is the cost of action i at time t chosen by the adversary.

Can we use this for adversarial bandits? Reduction

Exp3 Algorithm

Algorithm (Exp3). *We define the following algorithm:*

1. Initialize $w_i^0 = 1$ for all $i \in [n]$.
2. **For** t=1 … T **do**
3. **Choose** action i with probability proportional to w_i^{t-1}.
4. **Only for** the chosen action (say i) **do**
5. $w_i^t = (1-\epsilon)^{c_i^t / p_i^t} w_i^{t-1}$.
6. **End For**
7. **End For**

Remarks:

- We choose i with probability $p_i^t = \frac{w_i^{t-1}}{\sum_j w_j^{t-1}}$.
- c_i^t is the cost of action i at time t chosen by the adversary.
- Essentially, we assume that all the actions got cost zero except the chosen action that got cost $\hat{c}_i^t \coloneqq c_i^t / p_i^t$.

Exp3 Algorithm

Algorithm (Exp3). *We define the following algorithm:*

1. Initialize $w_i^0 = 1$ for all $i \in [n]$.
2. **For** t=1 ... T **do**
3. **Choose** action i with probability proportional to w_i^{t-1}.
4. **Only for** the chosen action (say i) **do**
5. $w_i^t = (1-\epsilon)^{c_i^t/p_i^t} w_i^{t-1}$.
6. **End For**
7. **End For**

Remarks:

- We choose i with probability $\mathrm{p_i^t} = \frac{w_i^{t-1}}{\sum_j w_j^{t-1}}$.
- $\mathrm{c_i^t}$ is the cost of action i at time t chosen by the adversary.
- Essentially, we assume that all the actions got cost zero except the chosen action that got cost $\hat{c}_i^t := \mathrm{c_i^t}/p_i^t$.

What is the cost of every action? Each a r.v that is an unbiased estimator!

Formally we ensure that $\mathbb{E}[\hat{c}_i^t | p^t] = c_i^t$ for all i.

Exp3 Algorithm

Algorithm (Exp3). *We define the following algorithm:*

1. Initialize $w_i^0 = 1$ for all $i \in [n]$.
2. **For** t=1 … T **do**
3. **Choose** action i with probability proportional to w_i^{t-1}.
4. **Only for** the chosen action (say i) **do**
5. $w_i^t = (1-\epsilon)^{c_i^t / p_i^t} w_i^{t-1}$.
6. **End For**
7. **End For**

Remarks:

- We choose i with probability $p_i^t = \frac{w_i^{t-1}}{\sum_j w_j^{t-1}}$.
- c_i^t is the cost of action i at time t chosen by the adversary.
- Essentially, we assume that all the actions got cost zero except the chosen action that got cost $\hat{c}_i^t := c_i^t / p_i^t$.

What is the cost of every action? Each a r.v that is an unbiased estimator!

Formally we ensure that $\mathbb{E}[\hat{c}_i^t | p^t] = c_i^t$ for all i.

We will choose $\epsilon = \sqrt{\frac{2 \log K}{TK}}$ and we will get regret $O(\sqrt{TK \log K})$.

Analysis of Exp3

Recall that for the analysis of MWU we defined a potential function Φ_t (sum of weights).

We set $\Phi_t = -\frac{1}{\epsilon} \log \sum_i e^{-\epsilon \sum_{\tau=1}^{t-1} \hat{c}_i^{\tau}}$.

Set $L_i^t = \sum_{\tau=1}^{t} \hat{c}_i^{\tau}$. Observe that

Analysis of Exp3

Recall that for the analysis of MWU we defined a potential function Φ_t (sum of weights).

We set $\Phi_t = -\frac{1}{\epsilon}\log \sum_i e^{-\epsilon \sum_{\tau=1}^{t-1} \hat{c}_i^\tau}$.

Set $L_i^t = \sum_{\tau=1}^{t} \hat{c}_i^\tau$. Observe that

$$\Phi_{t+1} - \Phi_t = -\frac{1}{\epsilon}\log \frac{\sum_i e^{-\epsilon L_i^t}}{\sum_i e^{-\epsilon L_i^{t-1}}}$$

Analysis of Exp3

Recall that for the analysis of MWU we defined a potential function Φ_t (sum of weights).

We set $\Phi_t = -\frac{1}{\epsilon} \log \sum_i e^{-\epsilon \sum_{\tau=1}^{t-1} \hat{c}_i^\tau}$.

Set $L_i^t = \sum_{\tau=1}^{t} \hat{c}_i^\tau$. Observe that

$$\Phi_{t+1} - \Phi_t = -\frac{1}{\epsilon} \log \frac{\sum_i e^{-\epsilon L_i^t}}{\sum_i e^{-\epsilon L_i^{t-1}}}$$

$$= -\frac{1}{\epsilon} \log \frac{\sum_i e^{-\epsilon L_i^{t-1}} e^{-\epsilon \hat{c}_i^t}}{\sum_i e^{-\epsilon L_i^{t-1}}}$$

Analysis of Exp3

Recall that for the analysis of MWU we defined a potential function Φ_t (sum of weights).

We set $\Phi_t = -\frac{1}{\epsilon} \log \sum_i e^{-\epsilon \sum_{\tau=1}^{t-1} \hat{c}_i^\tau}$.

Set $L_i^t = \sum_{\tau=1}^{t} \hat{c}_i^\tau$. Observe that

$$\begin{aligned}
\Phi_{t+1} - \Phi_t &= -\frac{1}{\epsilon} \log \frac{\sum_i e^{-\epsilon L_i^t}}{\sum_i e^{-\epsilon L_i^{t-1}}} \\
&= -\frac{1}{\epsilon} \log \frac{\sum_i e^{-\epsilon L_i^{t-1}} e^{-\epsilon \hat{c}_i^t}}{\sum_i e^{-\epsilon L_i^{t-1}}} \\
&= -\frac{1}{\epsilon} \log \mathbb{E}_{i \sim p^t} [e^{-\epsilon \hat{c}_i^t}]
\end{aligned}$$

Analysis of Exp3

Recall that for the analysis of MWU we defined a potential function Φ_t (sum of weights).

We set $\Phi_t = -\frac{1}{\epsilon} \log \sum_i e^{-\epsilon \sum_{\tau=1}^{t-1} \hat{c}_i^\tau}$.

Set $L_i^t = \sum_{\tau=1}^{t} \hat{c}_i^\tau$. Observe that

$$
\begin{aligned}
\Phi_{t+1} - \Phi_t &= -\frac{1}{\epsilon} \log \frac{\sum_i e^{-\epsilon L_i^t}}{\sum_i e^{-\epsilon L_i^{t-1}}} \\
&= -\frac{1}{\epsilon} \log \frac{\sum_i e^{-\epsilon L_i^{t-1}} e^{-\epsilon \hat{c}_i^t}}{\sum_i e^{-\epsilon L_i^{t-1}}} \\
&= -\frac{1}{\epsilon} \log \mathbb{E}_{i \sim p^t} [e^{-\epsilon \hat{c}_i^t}] \\
&\geq -\frac{1}{\epsilon} \log \mathbb{E}_{i \sim p^t} [1 - \epsilon \hat{c}_i^t + \frac{1}{2} \epsilon^2 \hat{c}_i^{t\,2}]
\end{aligned}
$$

Since $e^{-x} \leq 1 - x + \frac{1}{2} x^2$.

Analysis of Exp3

$$\Phi_{t+1} - \Phi_t \geq -\frac{1}{\epsilon} \log \mathbb{E}_{i \sim p^t}[1 - \epsilon \hat{c}_i^t + \frac{1}{2}\epsilon^2 \hat{c}_i^{t\ 2}]$$

$$= -\frac{1}{\epsilon} \log \left(1 - \mathbb{E}_{i \sim p^t}[\epsilon \hat{c}_i^t - \frac{1}{2}\epsilon^2 \hat{c}_i^{t\ 2}]\right)$$

Analysis of Exp3

$$\Phi_{t+1} - \Phi_t \geq -\frac{1}{\epsilon} \log \mathbb{E}_{i \sim p^t}[1 - \epsilon \hat{c}_i^t + \frac{1}{2}\epsilon^2 \hat{c}_i^{t\ 2}]$$

$$= -\frac{1}{\epsilon} \log \left(1 - \mathbb{E}_{i \sim p^t}[\epsilon \hat{c}_i^t - \frac{1}{2}\epsilon^2 \hat{c}_i^{t\ 2}]\right)$$

$$\geq \frac{1}{\epsilon} \mathbb{E}_{i \sim p^t}[\epsilon \hat{c}_i^t - \frac{1}{2}\epsilon^2 \hat{c}_i^{t\ 2}]$$

Analysis of Exp3

$$\begin{aligned}
\Phi_{t+1} - \Phi_t &\geq -\frac{1}{\epsilon} \log \mathbb{E}_{i\sim p^t}[1 - \epsilon \hat{c}_i^t + \frac{1}{2}\epsilon^2 \hat{c}_i^{t\,2}] \\
&= -\frac{1}{\epsilon} \log \left(1 - \mathbb{E}_{i\sim p^t}[\epsilon \hat{c}_i^t - \frac{1}{2}\epsilon^2 \hat{c}_i^{t\,2}]\right) \\
&\geq \frac{1}{\epsilon}\mathbb{E}_{i\sim p^t}[\epsilon \hat{c}_i^t - \frac{1}{2}\epsilon^2 \hat{c}_i^{t\,2}] \\
&= \sum_i p_i^t \hat{c}_i^t - \frac{1}{2}\epsilon \sum_i p_i^t \hat{c}_i^{t\,2}
\end{aligned}$$

Analysis of Exp3

$$\begin{aligned}\Phi_{t+1} - \Phi_t &\geq -\frac{1}{\epsilon} \log \mathbb{E}_{i\sim p^t}[1 - \epsilon \hat{c}_i^t + \frac{1}{2}\epsilon^2 \hat{c}_i^{t\,2}] \\ &= -\frac{1}{\epsilon} \log \left(1 - \mathbb{E}_{i\sim p^t}[\epsilon \hat{c}_i^t - \frac{1}{2}\epsilon^2 \hat{c}_i^{t\,2}]\right) \\ &\geq \frac{1}{\epsilon}\mathbb{E}_{i\sim p^t}[\epsilon \hat{c}_i^t - \frac{1}{2}\epsilon^2 \hat{c}_i^{t\,2}] \\ &= \sum_i p_i^t \hat{c}_i^t - \frac{1}{2}\epsilon \sum_i p_i^t \hat{c}_i^{t\,2}\end{aligned}$$

By taking expectation we get

$$\boxed{\mathbb{E}[\Phi_{t+1} - \Phi_t] \geq \sum_i p_i^t c_i^t - \frac{1}{2}\epsilon \sum_i c_i^{t\,2} \geq \sum_i p_i^t c_i^t - \frac{K\epsilon}{2}}$$

Analysis of Exp3

We conclude that (telescopic sum)

$$\boxed{\mathbb{E}[\Phi_T - \Phi_1] \geq \sum_{t=1}^{T} \sum_{i} p_i^t c_i^t - \frac{KT\epsilon}{2}}$$

Analysis of Exp3

We conclude that (telescopic sum)

$$\boxed{\mathbb{E}[\Phi_T - \Phi_1] \geq \sum_{t=1}^{T} \sum_i p_i^t c_i^t - \frac{KT\epsilon}{2}}$$

Finally

$$\mathbb{E}[\Phi_T - \Phi_1] \leq \mathbb{E}[L_{i^*}^T - (-\frac{1}{\epsilon} \log K)] = \sum_t c_{i^*}^t + \frac{1}{\epsilon} \log K.$$

Hence

$$\mathbb{E}[R(T)] = \sum_t \sum_i p_i^t c_i^t - \sum_t c_{i^*}^t \leq \frac{KT\epsilon}{2} + \frac{1}{\epsilon} \log K$$

Analysis of Exp3

We conclude that (telescopic sum)

$$\boxed{\mathbb{E}[\Phi_T - \Phi_1] \geq \sum_{t=1}^{T} \sum_i p_i^t c_i^t - \frac{KT\epsilon}{2}}$$

Finally

$$\mathbb{E}[\Phi_T - \Phi_1] \leq \mathbb{E}[L_{i^*}^T - (-\frac{1}{\epsilon}\log K)] = \sum_t c_{i^*}^t + \frac{1}{\epsilon}\log K.$$

Hence

$$\mathbb{E}[R(T)] = \sum_t \sum_i p_i^t c_i^t - \sum_t c_{i^*}^t \leq \frac{KT\epsilon}{2} + \frac{1}{\epsilon}\log K$$

We choose $\epsilon = \sqrt{\frac{2\log K}{TK}}$ and it follows that $\boxed{\mathbb{E}[R(T)] \text{ is } O(\sqrt{TK\log K}).}$

Conclusion

- Introduction to Multi-armed bandits.
  - UCB.
  - Exp3

- Next lecture we will talk about basics in Markov Decision Processes and Stochastic Games.