Intermediate Observations in Factored-Reward Bandits

Simone Drago[1] Marco Mussi[1] Marcello Restelli[1] Alberto Maria Metelli[1]

[1]Politecnico di Milano

Adaptive and Learning Agents Workshop @ AAMAS 2024, Auckland, New Zealand

In the customary **Multi-Armed Bandit** framework, we consider a problem where:

- We have K **arms**, each representing an action
- The actions are **independent**
- There is **no structure** in the reward

However, in several cases, we may have:

- A structure in the **actions** and/or in the **reward** model
- Access to **intermediate effects** which may help the **learning process**

We consider the scenario in which we want to **sell a product online**:

- We have to choose a **price-budget pair**:
  - the **price** we set determines the users' propensity to buy (the so-called **conversion rate**)
  - the **advertising budget** we invest influences the number of potential customers that will be exposed (i.e., the number of **impressions**)
- We have access to **intermediate observations**:
  - the conversion rate, which depends on the price
  - the expected number of impressions, which depends on the budget
- Our objective is to maximize the revenue (i.e., **reward**) that is a function of the **product** between **intermediate observations**

- We can solve this problem using **standard Multi-Armed Bandit** techniques considering **price-budget couples** as actions
- However, if we look just at the reward and disregard this **factored** structure, the learning problem will:
  - present an **unnecessarily large action space**, including all the possible combinations of action components
  - suffer a **possibly amplified effect of the noise** in the reward due to the product of the noisy intermediate observations

- At every round $t \in [\![T]\!]$, we choose an **action vector**:

$$\mathbf{a}(t) = (a_1(t), \ldots, a_d(t)) \in \mathcal{A} := [\![k_1]\!] \times \cdots \times [\![k_d]\!]$$

  - $\forall i \in [\![d]\!]$ we have k_i options
  - d is the action vector dimension

- We observe a vector of d **intermediate observations** $\mathbf{x}(t) = (x_1(t), \ldots, x_d(t))$ and receive as reward the **product of the observations** $r(t) = \prod_{i \in [\![d]\!]} x_i(t)$

- The i^{th} component $x_i(t)$ of the intermediate observation vector $\mathbf{x}(t)$ is the effect of the i^{th} action component $a_i(t)$ in the action vector: $x_i(t) = \mu_{i,a_i(t)} + \epsilon_i(t)$
  - $\mu_{i,a_i(t)} \in [0, 1]$ is the **expected observation** of the i^{th} component $a_i(t)$
  - $\epsilon_i(t)$ is σ^2-subgaussian noise

- An **optimal action vector** is:

$$\mathbf{a}^* = (a_1^*, \ldots, a_d^*) \in \operatorname*{arg\,max}_{\mathbf{a}=(a_1,\ldots,a_d)\in\mathcal{A}} \prod_{i\in[\![d]\!]} \mu_{i,a_i}$$

and we abbreviate $\mu_i^* = \mu_{i,a_i^*}, \forall i \in [\![d]\!]$

- We define the **suboptimality gaps** related to:
  - the i^{th} **action component** $\Delta_{i,a_i} := \mu_i^* - \mu_{i,a_i}$ for $a_i \in [\![k_i]\!]$
  - the **action vector** $\mathbf{a} = (a_1, \ldots, a_d) \in \mathcal{A}$ as $\Delta_{\mathbf{a}} := \prod_{i\in[\![d]\!]} \mu_i^* - \prod_{i\in[\![d]\!]} \mu_{i,a_i}$

- The **goal** of an algorithm $\mathfrak{A}$ is to minimize the **expected cumulative regret**:

$$\mathbb{E}[R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}})] := \mathbb{E}\left[T \prod_{i\in[\![d]\!]} \mu_i^* - \sum_{t\in[\![T]\!]} \prod_{i\in[\![d]\!]} \mu_{i,a_i(t)}\right] = \mathbb{E}\left[\sum_{t\in[\![T]\!]} \Delta_{\mathbf{a}(t)}\right]$$

Theorem (Worst-Case Lower Bound)

For every algorithm $\mathfrak{A}$*, there exists an FRB* $\underline{\boldsymbol{\nu}}$ *such that for* $T \geq \mathcal{O}\left(d^2\right)$*,* $\mathfrak{A}$ *suffers an expected cumulative regret of at least:*

$$\mathbb{E}\left[R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}})\right] \geq \frac{\sigma}{4\sqrt{2}} \sum_{i \in [\![d]\!]} \sqrt{k_i T}.$$

In particular, if $k_i =: k$ *for every* $i \in [\![d]\!]$*, we have:*

$$\mathbb{E}\left[R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}})\right] \geq \Omega(\sigma d \sqrt{kT}).$$

Theorem (Instance-Dependent Lower Bound)

For every consistent algorithm $\mathfrak{A}$ *and FRB* $\underline{\boldsymbol{\nu}}$ *with unique optimal arm* $\mathbf{a}^* \in \mathcal{A}$ *it holds:*

$$\liminf_{T \to +\infty} \frac{\mathbb{E}\left[R_T(\mathfrak{A}, \underline{\boldsymbol{\nu}})\right]}{\log T} \geq \underline{C}(\underline{\boldsymbol{\nu}}) = \min_{(L_{\mathbf{a}})_{\mathbf{a} \in \mathcal{A} \setminus \{\mathbf{a}^*\}}} \sum_{\mathbf{a} \in \mathcal{A} \setminus \{\mathbf{a}^*\}} L_{\mathbf{a}} \Delta_{\mathbf{a}}$$

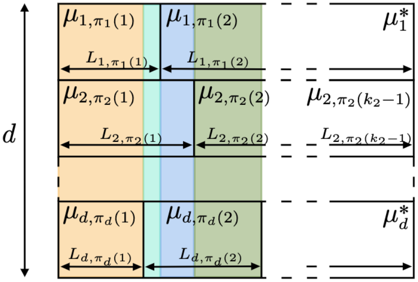
$$\text{s.t.} \quad L_{i,j} = \textstyle\sum_{\mathbf{a} \in \mathcal{A} \setminus \{\mathbf{a}^*\}, a_i = j} L_{\mathbf{a}}, \quad \forall i \in [\![d]\!],\ j \in [\![k_i]\!] \setminus \{a_i^*\}$$

$$L_{i,j} \geq \frac{2\sigma^2}{\Delta_{i,j}^2}, \quad \forall i \in [\![d]\!], j \in [\![k_i]\!] \setminus \{a_i^*\}$$

$$L_{\mathbf{a}} \geq 0, \quad \forall \mathbf{a} \in \mathcal{A} \setminus \{\mathbf{a}^*\}.$$

- We consider $L_{i,j} = \mathbb{E}[N_{i,j}] / \log T$ to handle the asymptotic nature of the bound

- To solve the optimization problem, we have to search for the **best way** to arrange the pulls
- We can make use of **rearrangement inequality** for integrals to find the best solution (Luttinger and Friedberg, 1976)



- We present Factored Upper Confidence Bound (F-UCB)
- F-UCB performs a **UCB-like** exploration (Auer et al., 2002) **independently** for every dimension $i \in [\![d]\!]$
- Then, we study its **theoretical guarantees**

Algorithm: F−UCB.

Input : Exploration Parameter α, Subgaussian proxy σ, Action component size k_i, $\forall i \in [\![d]\!]$

```
Initialize $N_{i,a_i}(0) \leftarrow 0$, $\widehat{\mu}_{i,a_i}(0) \leftarrow 0$ $\forall a_i \in [\![k_i]\!]$, $i \in [\![d]\!]$
for $t \in [\![T]\!]$ do
    Select $\mathbf{a}(t) \in \underset{\mathbf{a}=(a_1,\dots a_d)^{\mathsf{T}} \in \mathcal{A}}{\arg\max} \prod_{i \in [\![d]\!]} \mathrm{UCB}_{i,a_i}(t)$ where $\mathrm{UCB}_{i,a_i}(t) = \widehat{\mu}_{i,a_i}(t-1) + \sigma\sqrt{\frac{\alpha \log t}{N_{i,a_i}(t-1)}}$
    Play $\mathbf{a}(t)$ and observe $\mathbf{x}(t) = (x_1(t), \dots, x_d(t))^{\mathsf{T}}$
    Update $\widehat{\mu}_{i,a_i(t)}(t)$ and $N_{i,a_i(t)}(t)$ for every $i \in [\![d]\!]$
end
```

Theorem (Worst-Case Upper Bound for F-UCB)

For any FRB $\underline{\boldsymbol{\nu}}$, F-UCB with $\alpha > 2$ suffers an expected regret bounded as:

$$\mathbb{E}\left[R_T(\texttt{F-UCB}, \underline{\boldsymbol{\nu}})\right] \leq 4\sigma \sum_{i \in [\![d]\!]} \sqrt{\alpha k_i T \log T} + g(\alpha) \sum_{i \in [\![d]\!]} k_i.$$

In particular, if $k_i =: k$, for every $i \in [\![d]\!]$, we have:

$$\mathbb{E}\left[R_T(\texttt{F-UCB}, \underline{\boldsymbol{\nu}})\right] \leq \widetilde{\mathcal{O}}(\sigma d \sqrt{kT}).$$

Theorem (Instance-Dependent Upper Bound for F-UCB)

For a given FRB $\underline{\boldsymbol{\nu}}$, F-UCB with $\alpha > 2$ suffers an expected regret bounded as:

$$\mathbb{E}\left[R_T(\texttt{F-UCB}, \underline{\boldsymbol{\nu}})\right] \le \overline{C}(\texttt{F-UCB}, \underline{\boldsymbol{\nu}}) = \max_{(N_{\mathbf{a}})_{\mathbf{a}\in\mathcal{A}}} \sum_{\mathbf{a}\in\mathcal{A}\setminus\{\mathbf{a}^*\}} N_{\mathbf{a}}\Delta_{\mathbf{a}}$$

$$\text{s.t.} \quad N_{i,j} = \textstyle\sum_{\mathbf{a}\in\mathcal{A}\setminus\{\mathbf{a}^*\}, a_i = j} N_{\mathbf{a}}, \quad \forall i \in [\![d]\!],\ j \in [\![k_i]\!] \setminus \{a_i^*\}$$

$$N_{i,j} \le \frac{4\alpha\sigma^2 \log T}{\Delta_{i,j}^2} + g(\alpha), \quad \forall i \in [\![d]\!],\ j \in [\![k_i]\!] \setminus \{a_i^*\}$$

$$\textstyle\sum_{\mathbf{a}\in\mathcal{A}} N_{\mathbf{a}} = T$$

$$N_{\mathbf{a}} \ge 0, \qquad \forall \mathbf{a} \in \mathcal{A}$$

Corollary (Explicit Instance-Dependent Upper Bound for F-UCB)

For a given FRB $\underline{\boldsymbol{\nu}}$, F-UCB with $\alpha > 2$ suffers an expected regret bounded by:

$$\mathbb{E}\left[R_T(\texttt{F-UCB}, \underline{\boldsymbol{\nu}})\right] \leq \overline{C}(\texttt{F-UCB}, \underline{\boldsymbol{\nu}})$$
$$\leq 4\alpha\sigma^2 \log T \sum_{i \in [\![d]\!]} \mu^*_{-i} \sum_{j \in [\![k_i]\!] \setminus \{a^*_i\}} \Delta_{i,j}^{-1} + g(\alpha) \sum_{i \in [\![d]\!]} k_i,$$

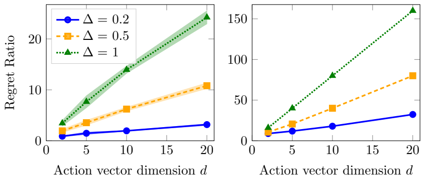
where $\mu^*_{-i} = \prod_{l \in [\![d]\!] \setminus \{i\}} \mu^*_l \leq 1$ for every $i \in [\![d]\!]$.

Instance-dependent Optimality of F-UCB

- For $T \to +\infty$, we observe that:

$$\frac{\overline{C}(\texttt{F-UCB}, \underline{\boldsymbol{\nu}})}{\underline{C}(\underline{\boldsymbol{\nu}}) \log T} \leq \frac{2d\alpha\Delta}{1-(1-\Delta)^d} \overset{\Delta \to 1}{=} 2\alpha d$$

- F-UCB performs worse than the lower bound, with an **additional dependence on d**
- In the figure, we compare:
  - **(left)** the ratio between the regret obtained by running F-UCB and the instance-dependent lower bound
  - **(right)** the bound above



- **F-UCB** does not enjoy instance-depedent optimality due to the **lack of syncronization** over the components of the action vector
- To overcome this problem, we propose `F-Track`
- `F-Track` is an algorithm which computes and **tracks the lower bound** (Lattimore and Szepesvari, 2017)

Algorithm: F-Track.

Input : Warm-up sample size N_0, Threshold ϵ_T, Action component size k_i, $\forall i \in [\![d]\!]$,

$t \leftarrow 1$

1

while $\min_{i \in [\![d]\!]} \min_{j \in [\![k_i]\!]} N_{i,j}(t) < N_0$ **do**

  Pull action vector $\mathbf{a}(t)$ with $a_i(t) = (t-1) \mod k_i + 1$ for all $i \in [\![d]\!]$, $t \leftarrow t+1$

end

$T_{\text{warm-up}} \leftarrow t - 1$

2

Estimate the suboptimality gaps $\forall i \in [\![d]\!],\ j \in [\![k_i]\!]$: $\widehat{\Delta}_{i,j} := \max_{j' \in [\![k_i]\!]} \widehat{\mu}_{i,j'}(T_{\text{warm-up}}) - \widehat{\mu}_{i,j}(T_{\text{warm-up}})$

Compute the number of pulls $\widehat{N}_{i,j} = 2\sigma^2 f_T(1/T)\widehat{\Delta}_{i,j}^{-2}$ for every action component $i \in [\![d]\!]$ and $j \in [\![k_i]\!]$

Compute the number of pulls $\widehat{N}_{\mathbf{a}}$ for every action vector $\mathbf{a} \in \mathcal{A}$ by solving the LP of the ID Lower Bound

while $t \leqslant T$ *and* $\max_{i \in [\![d]\!], j \in [\![k_i]\!]} |\widehat{\mu}_{i,j}(T_{\text{warm-up}}) - \widehat{\mu}_{i,j}(t-1)| \leqslant 2\epsilon_T$ **do**

  Pull action vector $\mathbf{a}(t) \in \arg\min\{N_{\mathbf{a}}(t) : \mathbf{a} \in \mathcal{A} \text{ and } N_{\mathbf{a}}(t) \leqslant \widehat{N}_{\mathbf{a}}\}$, $t \leftarrow t+1$

end

3

Discard all data and play F-UCB until $t = T$

Theorem (Instance-Dependent Upper Bound for `F-Track`)

For any FRB $\underline{\boldsymbol{\nu}}$, `F-Track` run with:

$$f_T(\delta) := \left(1 + \frac{1}{\log T}\right)\left(c \log\log T + \log\left(\frac{1}{\delta}\right)\right),$$

$$N_0 = \left\lceil \sqrt{\log T} \right\rceil \quad \textit{and} \quad \epsilon_T = \sqrt{\frac{2\sigma^2 f_T(1/\log T)}{N_0}},$$

suffers an expected regret of:

$$\limsup_{T \to +\infty} \frac{\mathbb{E}\left[R_T(\texttt{F-Track}, \underline{\boldsymbol{\nu}})\right]}{\log T} = \underline{C}(\underline{\boldsymbol{\nu}}).$$

- We presented the **Factored-Reward Bandits**, where we perform a **set of actions**, whose effects can be **observed**, and the reward is the **product** of those effects
- We characterized the **statistical complexity** of the setting from both the **worst-case** and **instance-dependent** perspectives
- We presented **F-UCB**, and we characterized its **instance-dependent** and **worst-case** guarantees and we discuss its **instance-dependent limitations**
- To overcome the F-UCB's limitations, we presented **F-Track**, which shows asymptotical **instance-depependent** optimality

References

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256.

Lattimore, T. and Szepesvari, C. (2017). The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 728–737. PMLR.

Luttinger, J. and Friedberg, R. (1976). A new rearrangement inequality for multiple integrals. *Archive for Rational Mechanics and Analysis*, 61:45–64.