Linear Bandits with Memory

G. Clerici
(UniMi)

P. Laforgue
(UniMi)

N. Cesa-Bianchi
(UniMi, PoliMi)

Motivations

Non-stationary effects (e.g. satiation)

are key in music recommendation

Motivations

We address problems arising from music recommendation,
where songs are usually characterized by features (e.g., music genres)

We propose a new bandit model
to investigate **non-stationary** effects (e.g. satiation)
in a linear setting where **actions are d-dimensional vectors**
(e.g. dimensions corresponds to different music genres)

Linear Bandits[1]

- Let $\mathcal{A} \subset \mathbb{R}^d$ be set of actions. The **reward** of action $a_t \in \mathcal{A}$ is now defined as
$$X_t = \langle a_t, \theta^* \rangle + \eta_t$$
where $\theta^* \in \mathbb{R}^d$ is the unknown parameter.
- The regret is defined as
$$R_T = \mathbb{E}\left[\hat{R}_T\right] = \mathbb{E}\left[\sum_{t=1}^{T} \max_{a\in\mathcal{A}} \langle \theta^*, a \rangle - \sum_{t=1}^{T} X_t\right]$$
- In the context of music recommender systems, the interdependencies between actions raise new challenges in terms of satiation!

[1] Abbasi-Yadkori, Yasin; Pàl, Dávid; Szepesvàri, Csaba. Improved algorithms for linear stochastic bandits. Advances in neural information processing systems, 2011, 24.

Linear Bandits with Memory (LBM) (1/2)

Let $\mathcal{A} \subset \mathbb{R}^d$ be set of actions. The **reward** of action $a_t \in \mathcal{A}$ is now defined as

$$X_t = \langle a_t, \, A(a_{t-m}, \ldots, a_{t-1}) \; \theta^* \rangle + \eta_t$$

where $\theta^* \in \mathbb{R}^d$ is the unknown parameter and

$$A(a_1, \ldots, a_m) = \left(A_0 + \textstyle\sum_{s=1}^m a_s a_s^\top \right)^\gamma \qquad \textbf{Memory matrix}$$

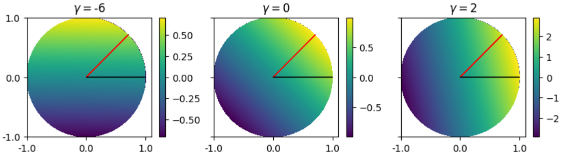
where m and γ are known.

Linear Bandits with Memory (LBM) (2/2)

Memory matrix: $A(a_1, \ldots, a_m) = \left(A_0 + \sum_{s=1}^{m} a_s a_s^\top\right)^\gamma$,

m is the window size, γ controls the nature of the non-stationarity:

- $\gamma = 0$ standard linear bandits
- $\gamma < 0$ rotting rested bandits[2]
- $\gamma > 0$ rising rested bandits[3]



$m = 1$
θ^*
$\mathbf{a_{t-1}}$

[2] Seznec, Julien, et al. Rotting bandits are no harder than stochastic ones. In: The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, 2019.
[3] Metelli, Alberto Maria, et al. Stochastic rising bandits. In: International Conference on Machine Learning. PMLR, 2022.

LBM: definition of regret

- We define the regret as:

$$R_T = \sum_{t=1}^{T} r_t^* - \mathbb{E}\Big[\sum_{t=1}^{T} X_t\Big]$$

where $r_t^* = \langle a_t^*, A(a_{t-m}^*, \ldots, a_{t-1}^*)\theta^* \rangle$ and $(a_t^*)_{t \geq 1}$ is the optimal sequence of actions (OPT), i.e.

$$a_1^*, \ldots, a_T^* = \underset{a_1, \ldots, a_T \in \mathcal{A}}{\operatorname{argmax}} \sum_{t=1}^{T} \langle a_t, A(a_{t-m}, \ldots, a_{t-1})\theta^* \rangle.$$

LBM: approximation and estimation

- We consider cyclic policies of blocks of size $m + L$
- Approximation error:

$$\mathrm{OPT} - \sum_{t=1}^{T} \tilde{r} \leq \frac{2mR}{m+L} T$$

- Estimation: adaptation of the `OFUL`[4] algorithm

[4] Abbasi-Yadkori, Yasin; Pàl, Dávid; Szepesvàri, Csaba. Improved algorithms for linear stochastic bandits. Advances in neural information processing systems, 2011, 24.

LBM: approximation and estimation

- We consider cyclic policies of blocks of size $m + L$
- Approximation error:

$$\text{OPT} - \sum_{t=1}^{T} \tilde{r} \leq \frac{2mR}{m+L} T$$

- Estimation: adaptation of the `OFUL`[5] algorithm
- Theorem: our algorithm `OFUL-memory (O3M)` achieves

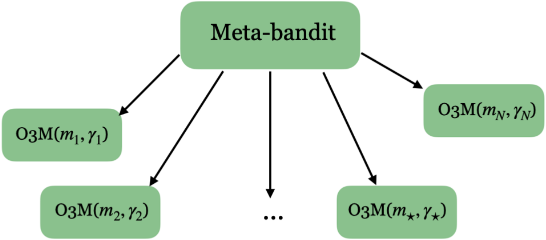
$$R_T = \mathcal{O}\left(\underbrace{\frac{2mR}{m+L} T}_{\text{approx.}} + \underbrace{dL(m+1)^{\gamma^+} \sqrt{T}}_{\text{estimation}} \right) = \tilde{\mathcal{O}}\left(\sqrt{d}\ (m+1)^{\frac{1}{2}+\gamma^+}\ T^{3/4} \right)$$

[5] Abbasi-Yadkori, Yasin; Pàl, Dávid; Szepesvàri, Csaba. Improved algorithms for linear stochastic bandits. Advances in neural information processing systems, 2011, 24.

LBM: approximation, estimation, and solution

If m and γ are unknown?

We propose a **meta-bandit algorithm for model selection**, `Bandit Combiner`[6] on `OFUL-memory (O3M)`



$$N \leq d\sqrt{m^\star} \text{ number of instances, } T \geq (m^\star + 1)^{2\gamma_\star^+}/m_\star d^4, \text{ and}$$
$$(m^\star, \gamma^\star) \in S = \{(m_1, \gamma_1), \ldots, (m_N, \gamma_N)\}$$

[6] Cutkosky, Ashok; Das, Abhimanyu; Purohit, Manish. Upper confidence bounds for combining stochastic bandits. arXiv preprint arXiv:2012.13115, 2020.

LBM: what happens if m and γ are unknown? (2/2)

Theorem: the regret of Bandit Combiner on OFUL-memory is upper bounded by

$$\tilde{\mathcal{O}}\Big(M\, d\, (m_\star + 1)^{1+\frac{3}{2}\gamma_\star^+}\; T^{3/4}\Big)$$

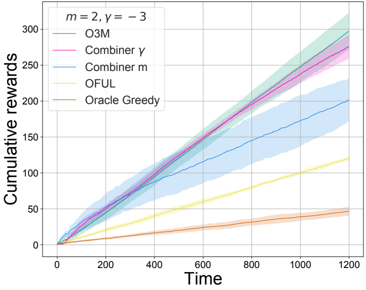
where

- d dimension
- $m_\star$ and $\gamma_\star$ are the true parameters of the LBM problem
- $M = (\max_j m_j)/(\min_j m_j)$
- T time horizon

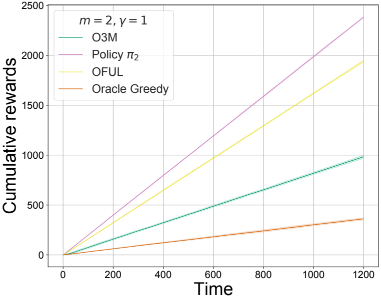
LBM: experiments

We tested `OFUL-memory (O3M)` and `Bandit Combiner on O3M` against natural benchmarks: a rotting ($\gamma < 0$) and on a rising ($\gamma > 0$) instance.

Rotting instance



Rising instance



Thank you

Thank you for your attention!

→ **GitHub**: GiuliaClerici/Linear-Bandits-with-Memory