Dynamical Linear Bandits

Marco Mussi[1], Alberto Maria Metelli[1] and Marcello Restelli[1]

[1] Politecnico di Milano

40th International Conference on Machine Learning, Honolulu, HI.

July 2023

In the customary **Multi-Armed Bandit** framework, there is **no notion of state**:

- The effect of the actions lasts for **one time-step** only
- There is **no chance** to model **action-dependent** phenomenas over time

- We consider a problem in which the **effect** of an action **persists over time**
- The effect of previous actions is modeled thanks to an **hidden state** evolving as a **linear** effect of the **actions**

- The state $\mathbf{x}_t \in \mathbb{R}^n$ is **not observable**
- The action $\mathbf{u}_t$ can be chosen in action space $\mathcal{U} \subseteq \mathbb{R}^d$
- At every time step we see a noisy realization of the reward y_t:

$$\underbrace{y_t}_{\text{Reward at time } t} = \underbrace{\langle \boldsymbol{\omega}, \mathbf{x}_t \rangle}_{\text{Current State Contribution}} + \underbrace{\langle \boldsymbol{\theta}, \mathbf{u}_t \rangle}_{\text{Action Contribution}} + \underbrace{\eta_t}_{\text{Subgaussian Random Noise}}$$

$$\underbrace{\mathbf{x}_{t+1}}_{\text{New State Vector}} = \underbrace{\mathbf{A}\mathbf{x}_t}_{\text{Previous State Contribution}} + \underbrace{\mathbf{B}\mathbf{u}_t}_{\text{Action Contribution}} + \underbrace{\boldsymbol{\epsilon}_t}_{\text{Subgaussian Random Noise}}$$

- $\boldsymbol{\omega}$, $\boldsymbol{\theta}$, $\mathbf{A}$, and $\mathbf{B}$ are **unknown**

- The goal is to minimize the **cumulative policy regret**:

$$\mathbb{E}\, R(\underline{\boldsymbol{\pi}}, T) = \mathbb{E}\left[\sum_{t=1}^{T} J^* - y_t\right]$$

where J^* is the value of J corresponding to the optimal policy ($J^* = \sup_{\underline{\boldsymbol{\pi}}} J(\underline{\boldsymbol{\pi}})$),

and:

$$J(\underline{\boldsymbol{\pi}}) := \liminf_{H \to +\infty} \mathbb{E}\left[\frac{1}{H}\sum_{t=1}^{H} y_t\right]$$

- (**Stability**) Spectral radius: $\rho(\mathbf{A}) < 1$
- (**Boundedness**) $\|\cdot\|_2$ of $\boldsymbol{\theta}$, $\boldsymbol{\omega}$, $\mathbf{B}$, $\mathbf{u}$, $\mathbf{x}$ bounded

  $\sup_{\mathbf{u},\mathbf{u}' \in \mathcal{U}} \langle \boldsymbol{\theta}, \mathbf{u} - \mathbf{u}' \rangle \leqslant 1$

Theorem (Optimal Policy)

Under **Stability** *and* **Boundedness** *Assumptions, an optimal policy* $\underline{\boldsymbol{\pi}}^*$ *maximizing the (infinite-horizon) expected average reward* $J(\underline{\boldsymbol{\pi}})$*, for every round* $t \in \mathbb{N}$ *and history* $H_{t-1} \in \mathcal{H}_{t-1}$ *is given by* $\boldsymbol{\pi}_t^*(H_{t-1}) = \mathbf{u}^*$*, defined as:*

$$\mathbf{u}^* \in \arg\max_{\mathbf{u} \in \mathcal{U}} J(\mathbf{u}) = \langle \mathbf{h}, \mathbf{u} \rangle,$$

where:

$$\mathbf{h} = \boldsymbol{\theta} + \mathbf{B}^T (\mathbf{I} - \mathbf{A})^{-T} \boldsymbol{\omega}.$$

Theorem (Lower Bound)

For any policy $\underline{\boldsymbol{\pi}}$ (even stochastic), there exists a DLB fulfilling **Stability** *and* **Boundedness** *Assumptions, such that for sufficiently large $T \geqslant \mathcal{O}\left(\frac{d^2}{1-\rho(\mathbf{A})}\right)$, policy $\underline{\boldsymbol{\pi}}$ suffers an expected regret lower bounded by:*

$$\mathbb{E}R(\underline{\boldsymbol{\pi}}, T) \geqslant \Omega\left(\frac{d\sqrt{T}}{(1-\rho(\mathbf{A}))^{\frac{1}{2}}}\right).$$

Algorithm 1: `DynLin-UCB`.

Input : Regularization parameter $\lambda > 0$,
Exploration coefficients $(\beta_{t-1})_{t \in [\![T]\!]}$,
Spectral radius upper bound $\overline{\rho} < 1$

Initialize $t \leftarrow 1, \quad \mathbf{V}_0 = \lambda \mathbf{I}_d, \quad \mathbf{b}_0 = \mathbf{0}_d, \quad \widehat{\mathbf{h}}_0 = \mathbf{0}_d$,

Define $M = \min\{M' \in \mathbb{N} : \sum_{m=1}^{M'} 1 + \lfloor \frac{\log m}{\log(1/\overline{\rho})} \rfloor > T\} - 1$

for $m \in [\![M]\!]$ **do**
- Compute $\mathbf{u}_t \in \arg\max_{\mathbf{u} \in \mathcal{U}} \mathsf{UCB}_t(\mathbf{u})$ where $\mathsf{UCB}_t(\mathbf{u}) := \langle \widehat{\mathbf{h}}_{t-1}, \mathbf{u} \rangle + \beta_{t-1} \|\mathbf{u}\|_{\mathbf{V}_{t-1}^{-1}}$
- Play arm $\mathbf{u}_t$ and observe y_t
- Define $H_m = \lfloor \frac{\log m}{\log(1/\overline{\rho})} \rfloor$
- **for** $j \in [\![H_m]\!]$ **do**
  - Update $\mathbf{V}_t = \mathbf{V}_{t-1}, \quad \mathbf{b}_t = \mathbf{b}_{t-1}$
  - $t \leftarrow t + 1$
  - Play arm $\mathbf{u}_t = \mathbf{u}_{t-1}$ and observe y_t
- **end**
- Update $\mathbf{V}_t = \mathbf{V}_{t-1} + \mathbf{u}_t \mathbf{u}_t^{\intercal}, \mathbf{b}_t = \mathbf{b}_{t-1} + \mathbf{u}_t y_t$
- Compute $\widehat{\mathbf{h}}_t = \mathbf{V}_t^{-1} \mathbf{b}_t$
- $t \leftarrow t + 1$

end

Theorem (Policy Regret Upper Bound)

Under **Stability** *and* **Boundedness** *Assumptions, selecting:*

$$\beta_t := \frac{\overline{c}_1}{\sqrt{\lambda}} \log(e(t+1)) + \overline{c}_2\sqrt{\lambda} + \sqrt{2\overline{\sigma}^2 \left(\log\left(\frac{1}{\delta}\right) + \frac{d}{2} \log\left(1 + \frac{tU^2}{d\lambda}\right)\right)},$$

and $\delta = 1/T$, `DynLin-UCB` *suffers an expected regret bounded as (highlighting the dependencies on* T, $\overline{\rho}$, d, *and* σ *only):*

$$\mathbb{E}\, R(\underline{\boldsymbol{\pi}}^{\texttt{DynLin-UCB}}, T) \leqslant \mathcal{O}\left(\frac{d\sigma\sqrt{T}(\log T)^{\frac{3}{2}}}{1-\overline{\rho}} + \frac{\sqrt{dT}(\log T)^2}{(1-\overline{\rho})^{\frac{3}{2}}} + \frac{1}{(1-\rho(\mathbf{A}))^2} \right).$$

Thank You for your Attention!