DYNAMICAL LINEAR BANDITS

POLITECNICO MILANO 1863

MARCO MUSSI, ALBERTO MARIA METELLI AND MARCELLO RESTELLI

{marco.mussi, albertomaria.metelli, marcello.restelli}@polimi.it

ICML International Conference On Machine Learning

MOTIVATION

- In real-world scenarios, actions leads both to **instantaneous** and **delayed** effects
- Delayed effects can be modeled by means of a **hidden** state
- The hidden state **evolves** depending on the previous hidden state and current **actions**

CONTRIBUTIONS

- We define **Dynamical Linear Bandits** to represent sequential problems with a hidden state evolving with a **linear unknown dynamics**
  - We show how to represent the problem with a unique **Markov Parameter**
  - We show that the optimal policy is a **constant** action
  - We provide a **lower bound**, also proving that we cannot avoid knowing an upper bound on the **spectral radius** of the transition matrix
- We propose **`DynLin-UCB`**, an **optimistic regret minimization** algorithm and we provide:
  - a **regret analysis** resulting in $\widetilde{\mathcal{O}}(\sqrt{T})$ expected regret
  - an extensive **numerical validation**

SETTING - OVERVIEW

DYNAMICAL LINEAR BANDITS (DLB)

$$\underbrace{y_t}_{\text{Reward}} = \underbrace{\langle \boldsymbol{\omega}, \mathbf{x}_t \rangle}_{\text{Delayed reward}} + \underbrace{\langle \boldsymbol{\theta}, \mathbf{u}_t \rangle}_{\text{Immediate reward}} + \underbrace{\eta_t}_{\text{Reward noise}}$$

$$\underbrace{\mathbf{x}_{t+1}}_{\text{New state}} = \underbrace{\mathbf{A}\mathbf{x}_t}_{\text{State contribution}} + \underbrace{\mathbf{B}\mathbf{u}_t}_{\text{Action contribution}} + \underbrace{\boldsymbol{\epsilon}_t}_{\text{State noise}}$$

- The state $\mathbf{x}_t \in \mathbb{R}^n$ is **not observable**
- The action $\mathbf{u}_t$ can be chosen in action space $\mathcal{U} \subseteq \mathbb{R}^d$
- $\boldsymbol{\omega}$, $\boldsymbol{\theta}$, $\mathbf{A}$, and $\mathbf{B}$ are **unknown**

ASSUMPTIONS

- Spectral radius: $\rho(\mathbf{A}) < 1$
- $\Phi(\mathbf{A}) = \sup_{\tau \geqslant 0} \|\mathbf{A}^\tau\|_2 / \rho(\mathbf{A})^\tau < \infty$

} STABILITY

- $\|\cdot\|_2$ of $\boldsymbol{\theta}, \boldsymbol{\omega}, \mathbf{B}, \mathbf{u}, \mathbf{x}$ bounded
- $\sup_{\mathbf{u}, \mathbf{u}' \in \mathcal{U}} \langle \boldsymbol{\theta}, \mathbf{u} - \mathbf{u}' \rangle \leqslant 1$

} BOUNDEDNESS

- η_t and $\boldsymbol{\epsilon}_t$ are σ^2-subgaussian } SUBGAUSSIANITY

SETTING - OPTIMAL POLICY AND LOWER BOUND

CUMULATIVE MARKOV PARAMETER

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

OPTIMAL POLICY

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

EXPECTED AVERAGE REWARD

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

ONLINE EXPECTED REGRET LOWER BOUND

$$\mathbb{E}\, R(\boldsymbol{\pi}, T) = \mathbb{E}\left[\sum_{t=1}^{T} J^* - y_t\right] \geqslant \Omega\left(\frac{d\sqrt{T}}{(1-\overline{\rho})^{1/2}}\right)$$

ALGORITHM - DYNAMICAL LINEAR UPPER CONFIDENCE BOUND (`DYNLIN-UCB`)

`DynLin-UCB` is an **optimistic regret minimization** algorithm that operates in **epochs**

- Played action is retrieved using an **optimistic** index
- **Epochs** are of increasing length H_m
  - Knowledge of an **upper bound on the spectral radius** $1 > \overline{\rho} \geqslant \rho(\mathbf{A})$
- The selected action is **persisted** for H_m times
- The **regression** estimate Markov parameters $\widehat{\mathbf{h}}_t$ is performed only using the **last sample**
  - i.e., when the hidden state is approximately **steady**

Algorithm 1 `DynLin-UCB`

```
Initialize V_0 = λI_d, b_0 = 0_d, ĥ_0 = 0_d,
    m ← 1, t ← 1
while t < T do
    Compute u*_t maximizing UCB_t(u)
    Define H_m = ⌈log m / log(1/ρ̄)⌉
    for j ∈ {1, ..., H_m} do
        Play u*_t = u*_{t-1} and Observe y_t
        t ← t + 1
    end
    Update V_t = V_{t-1} + u_t u_t^T
           b_t = b_{t-1} + u_t y_t
    Compute ĥ_t = V_t^{-1} b_t
    m ← m + 1
end
```

OPTIMISTIC ACTIONS CHOICE

$$\mathbf{u}_t^* = \arg\max_{\mathbf{u} \in \mathcal{U}} \mathrm{UCB}_t(\mathbf{u}) := \langle \widehat{\mathbf{h}}_{t-1}, \mathbf{u} \rangle + \beta_{t-1} \|\mathbf{u}\|_{\mathbf{V}_{t-1}^{-1}}$$

BOUND FOR `DYNLIN-UCB`

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

where $\overline{c}_1$, $\overline{c}_2$, and $\overline{\sigma}^2$ are constants, and $\lambda > 0$ is a regularization parameter

ONLINE EXPECTED REGRET UPPER BOUND

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

SIMILAR SETTINGS

Dynamical Linear Bandits can be seen as:

- **Partially Observable Markov Decision Processes** (Littman et al., 1995), in which the state $\mathbf{x}_t$ is non-observable, and the learner has access to the noisy observation y_t
- Multiple Input Single Output discrete–time **Linear Time–Invariant System** (Kalman, 1963)
- Non–contextual **Linear Bandit** (Abe and Long, 1999) when the hidden state does not affect the reward, i.e., when $\boldsymbol{\omega} = \mathbf{0}$

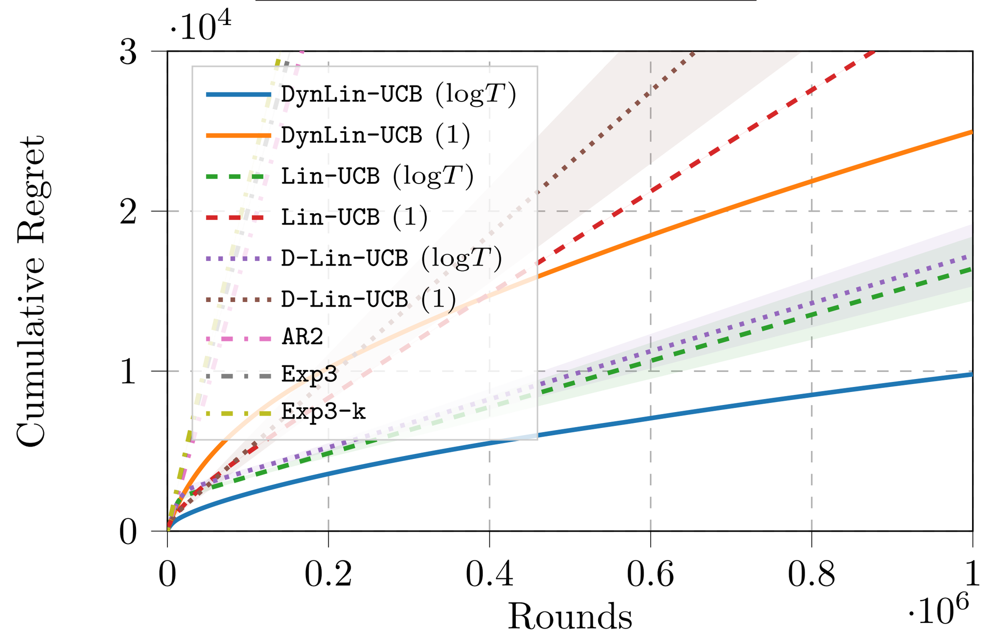
EXPERIMENTAL VALIDATION

EXPERIMENTAL SETTINGS

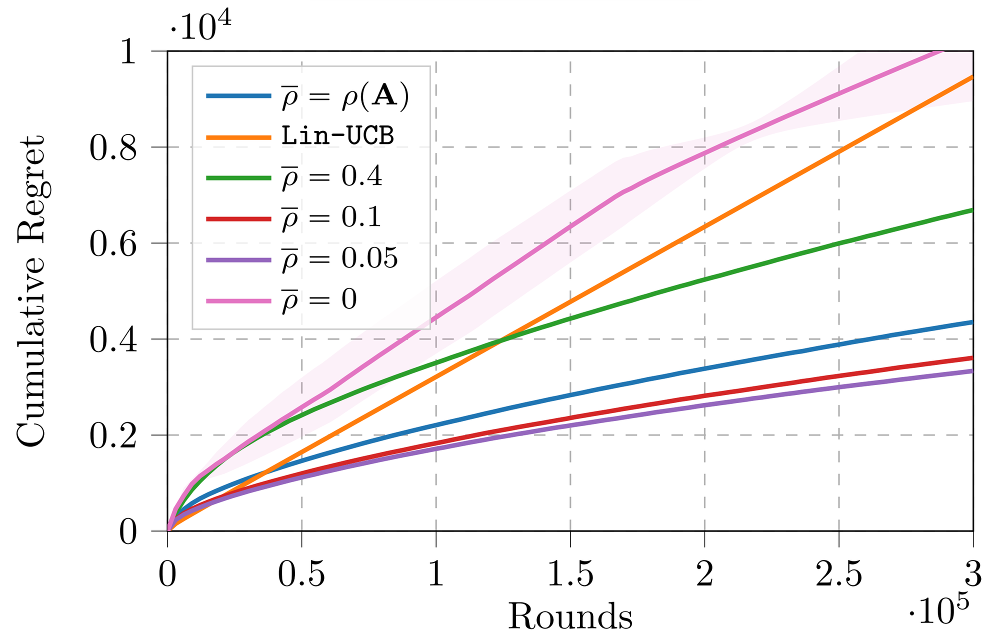
$\mathbf{A} = \mathrm{diag}(0.2, 0, 0.1)$ $\quad \mathbf{B} = \mathrm{diag}(0.25, 0, 0.1)$

$\boldsymbol{\theta} = (0, 0.5, 0.1)^\top$ $\quad \boldsymbol{\omega} = (1, 0, 0.1)^\top$ $\quad \sigma^2 = 10^{-2}$

CUMULATIVE REGRET



SENSITIVITY TO $\overline{\rho}$



REFERENCES

Naoki Abe and Philip M. Long. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pages 3–11, 1999.

Rudolf Emil Kalman. Mathematical description of linear dynamical systems. *Journal of the Society for Industrial and Applied Mathematics*, 1(2):152–192, 1963.

Michael L. Littman, Anthony R. Cassandra, and Leslie Pack Kaelbling. Learning policies for partially observable environments: Scaling up. In *ICML*, pages 362–370, 1995.