Dynamical Linear Bandits for Long-Lasting Vanishing Rewards

Marco Mussi, Alberto Maria Metelli and Marcello Restelli
{marco.mussi, albertomaria.metelli, marcello.restelli}@polimi.it
Politecnico Milano 1863

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 stare 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 that the optimal policy is a **constant** action
- We propose `DynLin-UCB`, an **anytime optimistic** algorithm and we provide:
  - a **regret analysis** resulting in $\widetilde{\mathcal{O}}(\sqrt{T})$ expected regret
  - a **numerical validation** in comparison with bandit baselines

Setting

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

Cumulative Markov Parameter

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

Dynamical Linear Upper Confidence Bound (DynLin-UCB)

Expected Average Reward

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

Optimal Policy

- Play the **constant** action

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

Algorithm

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

- Played action is retrieved using an **optimistic** index
- **Epochs** are of increasing length H_m (anytime algorithm)
  - 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 $\mathbf{V}_0 = \lambda \mathbf{I}_d$, $\mathbf{b}_0 = \mathbf{0}_d$, $\widehat{\mathbf{h}}_0 = \mathbf{0}_d$, $m \leftarrow 1, t \leftarrow 1$
while $t < T$ **do**
  Compute $\mathbf{u}_t^*$ maximizing $\mathrm{UCB}_t(\mathbf{u})$
  Define $H_m = \lceil \log m / \log(1/\overline{\rho}) \rceil$
  for $j \in \{1, \dots, H_m\}$ **do**
    Play $\mathbf{u}_t^* = \mathbf{u}_{t-1}^*$
    Observe y_t
    $t \leftarrow t + 1$
  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 $\widehat{\mathbf{h}}_t = \mathbf{V}_t^{-1} \mathbf{b}_t$
  $m \leftarrow m + 1$
end

Regret Analysis

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 Regret Bound

$$\mathbb{E}\, R(\texttt{DynLin-UCB}, T) = \mathbb{E}\left[\sum_{t=1}^{T} J^* - y_t\right] \leqslant \widetilde{\mathcal{O}}\left(\frac{(1 + \|\mathbf{A}\|_F) d \sqrt{T}}{(1 - \overline{\rho})^{3/2}}\right)$$

where $\|\cdot\|_F$ is the Frobenius norm

Similar Settings

Dynamical Linear Bandits can be seen as:

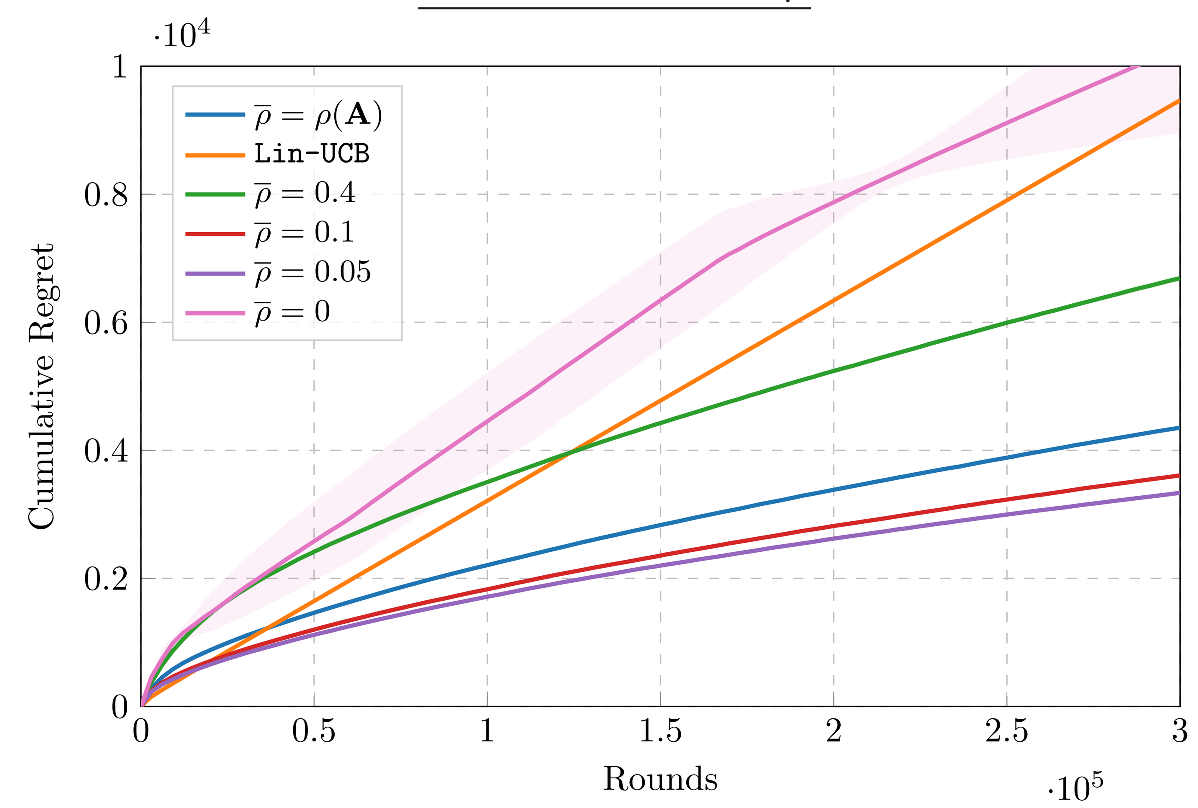
- **Partially Observable Markov Decision Processes** [Littman et al., 1995], in which the state $\mathbf{x}_t$ is non-observable, and the learner has access to the noisy observation y_t
- Multiple Input Single Output discrete–time **Linear Time–Invariant System** [Kalman, 1963]
- Non–contextual **Linear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

Experimental Validation

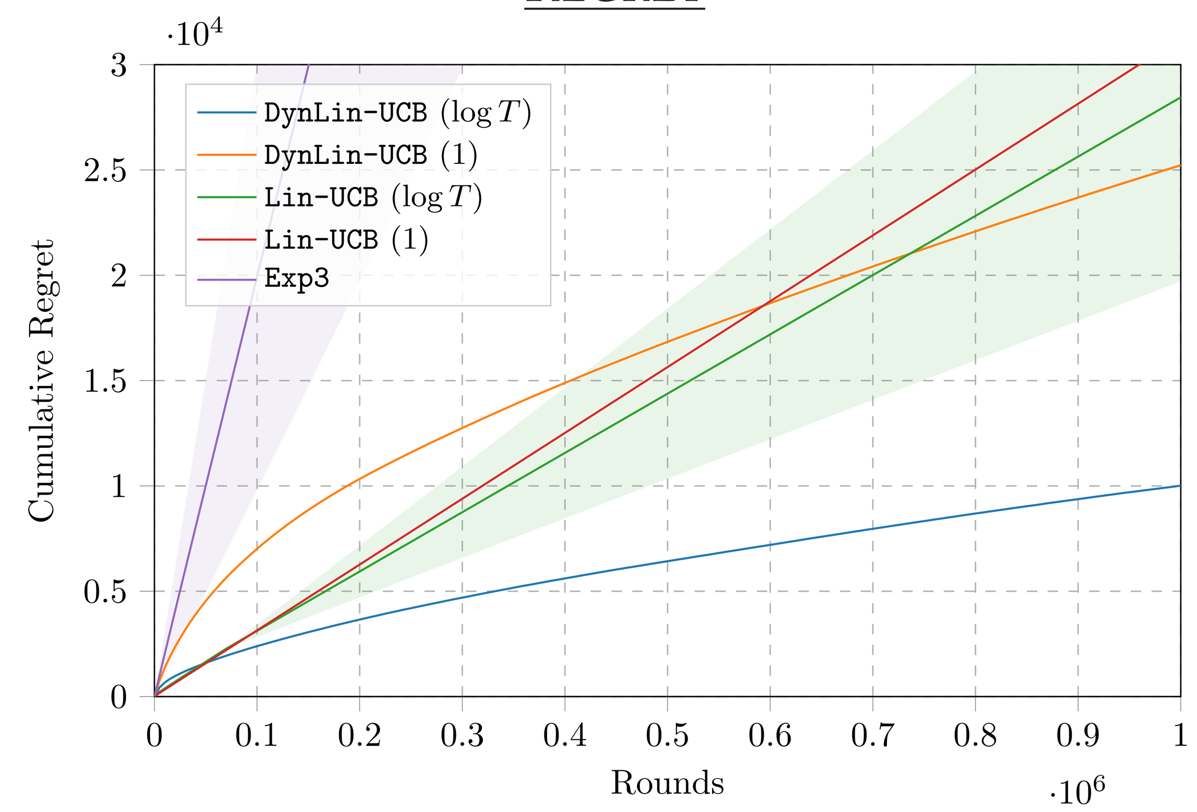
Experimental Settings

$\mathbf{A} = \mathrm{diag}((0.2, 0, 0.1)) \quad (\rho(\mathbf{A}) = 0.2)$
$\mathbf{B} = \mathrm{diag}((0.25, 0, 0.1))$
$\boldsymbol{\theta} = (0, 0.5, 0.1)^\intercal \qquad \boldsymbol{\omega} = (1, 0, 0.1)^\intercal$
$\eta \sim \mathcal{N}(0, 10^{-3}) \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, 10^{-3})$

Sensitivity to $\overline{\rho}$



Regret



References

Naoki Abe and Philip M. Long. Associative reinforcement learning using linear probabilistic concepts. In *International Conference on Machine Learning*, 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 *International Conference on Machine Learning*, pages 362–370, 1995.