Weighted Linear Bandits for non-stationary environments

Yoan Russac, Claire Vernade*, Olivier Cappé

CNRS, Inria, ENS, Université PSL, Deepmind*

Motivations

- Extending the analysis provided in Garivier and Moulines [2011] to the contextual case
- Propose an analysis valid in both **abruptly** and **slowly changing environments**
- Building policies robust to changepoints in the distribution instead of detecting them

Non-Stationary Linear Bandits Setting

At time t, the learner has access to a **time-dependent finite set of arbitrary actions** $\mathcal{A}_t = \{A_{t,1}, \dots, A_{t,K_t}\}$, where $A_{t,k} \in \mathbb{R}^d$.

They can only be probed one at a time, i.e., the learner

- Chooses an action $A_t \in \mathcal{A}_t$
- and observes only the noisy linear reward

$$X_t = A_t^\top \theta_t^\star + \eta_t$$

where η_t is a σ-subgaussian random noise

Specificity of the model

- **Non-Stationarity** $\theta_t^\star$ depends on t
- **Unstructured action set**

The dynamic regret is defined as

$$\mathbb{E}[R(T)] = \mathbb{E}\left(\sum_{t=1}^{T} \max_{a \in \mathcal{A}_t} \langle a - A_t, \theta_t^\star \rangle\right)$$

A key quantity for quantifying the non-stationarity is the **variation budget** defined as

$$\sum_{s=1}^{T-1} \|\theta_s^\star - \theta_{s+1}^\star\|_2 \leq B_T$$

A large variation budget can be either due to large scarce changes of $\theta_t^\star$ or frequent but small deviations

Assumptions

- η_t is, conditionally on the past, σ-subgaussian
- Bounded actions: $\forall t \geq 1, \forall a \in \mathcal{A}_t, \|a\|_2 \leq L$
- Bounded parameters: $\forall t \geq 1, \|\theta_t^\star\|_2 \leq S$
- $\forall t \geq 1, \forall a \in \mathcal{A}_t, |\langle a, \theta_t^\star \rangle| \leq 1$

Weighted least squares estimator

The usual least squares estimator is defined by,

$$\hat{\theta}_t^{OLS} = \arg\min_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t} (X_s - A_s^\top \theta)^2 + \frac{\lambda}{2}\|\theta\|_2^2,$$

whereas, the weighted least squares estimator is defined by

$$\hat{\theta}_t = \arg\min_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t} w_s (X_s - A_s^\top \theta)^2 + \frac{\lambda_t}{2}\|\theta\|_2^2.$$

It has the following closed form solution

$$\hat{\theta}_t = \left(\sum_{s=1}^{t} w_s A_s A_s^\top + \lambda_t I_d\right)^{-1} \sum_{s=1}^{t} w_s A_s X_s.$$

$\hookrightarrow$ $\hat{\theta}_t$ is unchanged if all the weights $(w_s)_{s \leq t}$ and the regularization parameter λ_t are multiplied by a same constant α

Concentration Result

In a stationary environment the following concentration result holds when using a weighted least squares estimator,

Theorem 1. Assuming that $\theta_t^\star = \theta^\star$, for any sequences of actions $(A_t)_{t \geq 1}$ (predictable based on past actions and rewards) and positive weights $(w_t)_{t \geq 1}$ and for all $\delta > 0$, with probability higher than $1 - \delta$,

$$\mathbb{P}\left(\forall t, \|\hat{\theta}_t - \theta^\star\|_{V_t \widetilde{V}_t^{-1} V_t} \leq \frac{\lambda_t}{\sqrt{\mu_t}} S + \sigma \sqrt{2\log(\tfrac{1}{\delta}) + d \log\left(1 + \frac{L^2 \sum_{s=1}^{t} w_s^2}{d \mu_t}\right)}\right)$$

where,

$$V_t = \sum_{s=1}^{t} w_s A_s A_s^\top + \lambda_t I_d \quad \text{and} \quad \widetilde{V}_t = \sum_{s=1}^{t} w_s^2 A_s A_s^\top + \mu_t I_d.$$

Extension for Non-Stationarity

We use particular weights of the form $w_s = \gamma^{-s}$ and particular regularization terms $\lambda_t = \lambda \gamma^{-t}$ and $\mu_t = \lambda \gamma^{-2t}$, where $0 < \gamma < 1$. A noiseless proxy value for $\theta_t^\star$ is defined as follow

$$\bar{\theta}_t = V_{t-1}^{-1}\left(\sum_{s=1}^{t-1} \gamma^{-s} A_s A_s^\top \theta_s^\star + \gamma^{t-1} \theta_t^\star\right)$$

Theorem 2. Let $\mathcal{C}_t = \{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_{t-1}\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}} \leq \beta_{t-1}\}$ denote the confidence ellipsoid with

$$\beta_t = \sqrt{\lambda} S + \sigma \sqrt{2\log(1/\delta) + d \log\left(1 + \frac{L^2(1 - \gamma^{2t})}{\lambda d (1 - \gamma^2)}\right)}$$

Then, $\forall \delta > 0$,

$$\mathbb{P}\left(\forall t \geq 1, \bar{\theta}_t \in \mathcal{C}_t\right) \geq 1 - \delta$$

High probability upper bound on the regret of `D-LinUCB`

Theorem 3. Assuming that $\sum_{s=1}^{T-1} \|\theta_s^\star - \theta_{s+1}^\star\|_2 \leq B_T$, the regret of the `D-LinUCB` algorithm may be bounded for all $\gamma \in (0,1)$ and integer $D \geq 1$, with probability at least $1 - \delta$, by

$$R_T \leq 2LDB_T + \frac{4L^3 S}{\lambda} \frac{\gamma^D}{1-\gamma} T + 2\sqrt{2}\beta_T \sqrt{dT} \sqrt{T \log(1/\gamma) + \log\left(1 + \frac{L^2}{d\lambda(1-\gamma)}\right)}$$

Algorithm

Algorithm 1: `D-LinUCB`

Input: Probability δ, subgaussianity constant σ, dimension d, regularization λ, upper bound for actions L, upper bound for parameters S, discount factor γ.

Initialization: $b = 0_{\mathbb{R}^d}$, $V = \lambda I_d$, $\widetilde{V} = \lambda I_d$, $\hat{\theta} = 0_{\mathbb{R}^d}$

for $t \geq 1$ **do**

  Receive $\mathcal{A}_t$, compute $\beta_{t-1} =$
  $\sqrt{\lambda} S + \sigma \sqrt{2 \log\left(\frac{1}{\delta}\right) + d \log\left(1 + \frac{L^2(1-\gamma^{2(t-1)})}{\lambda d (1-\gamma^2)}\right)}$

  for $a \in \mathcal{A}_t$ **do**

    Compute $\mathrm{UCB}(a) = a^\top \hat{\theta} + \beta_{t-1} \|a\|_{V^{-1}\widetilde{V}V^{-1}}$

  $A_t = \arg\max_a(\mathrm{UCB}(a))$

  Play action A_t and **receive reward** X_t

  Updating phase:

  $V = \gamma V + A_t A_t^\top + (1-\gamma)\lambda I_d,$

  $\widetilde{V} = \gamma^2 \widetilde{V} + A_t A_t^\top + (1-\gamma^2)\lambda I_d$

  $b = \gamma b + X_t A_t, \quad \hat{\theta} = V^{-1} b$

Experiments

1. Synthetic data: $K = 20$, $d = 2$, $L = 1$, $S = 1$ and $\theta^\star$ is evolving over the experiment
2. Synthetic data based on a real world dataset.

We compare `D-LinUCB` with 1/ `SW-UCB` of Cheung et al. [2019], 2/ `dLinUCB` the changepoint detection method from Wu et al. [2018].

Abruptly changing environment

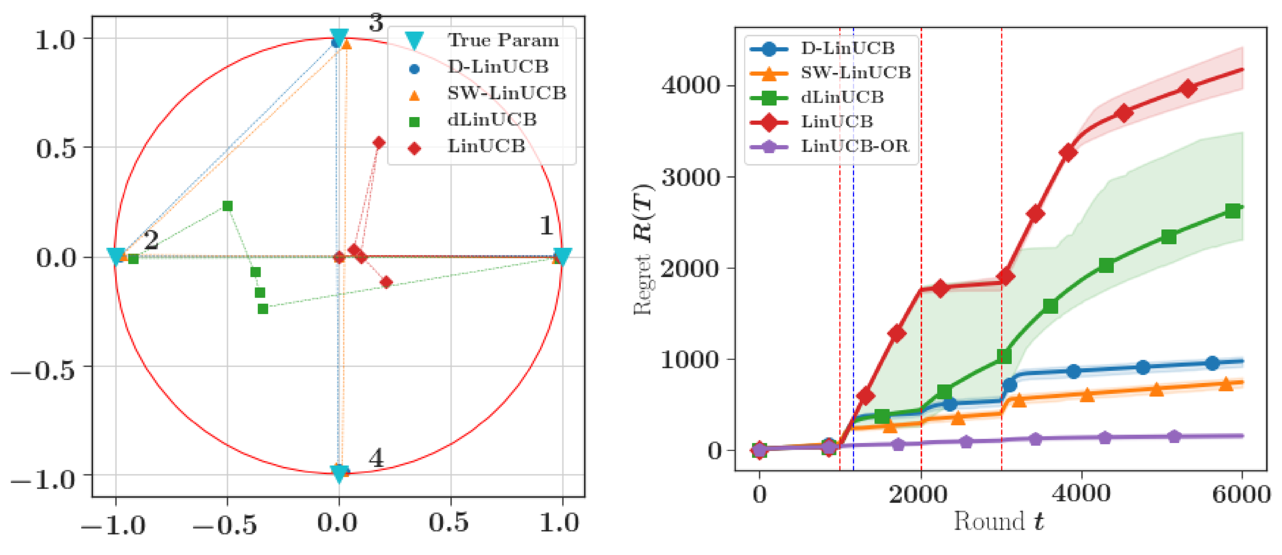

Fig. 1: (*left*) estimated parameters, (*right*) accumulated regret, both averaged on 100 runs

Slowly changing environment

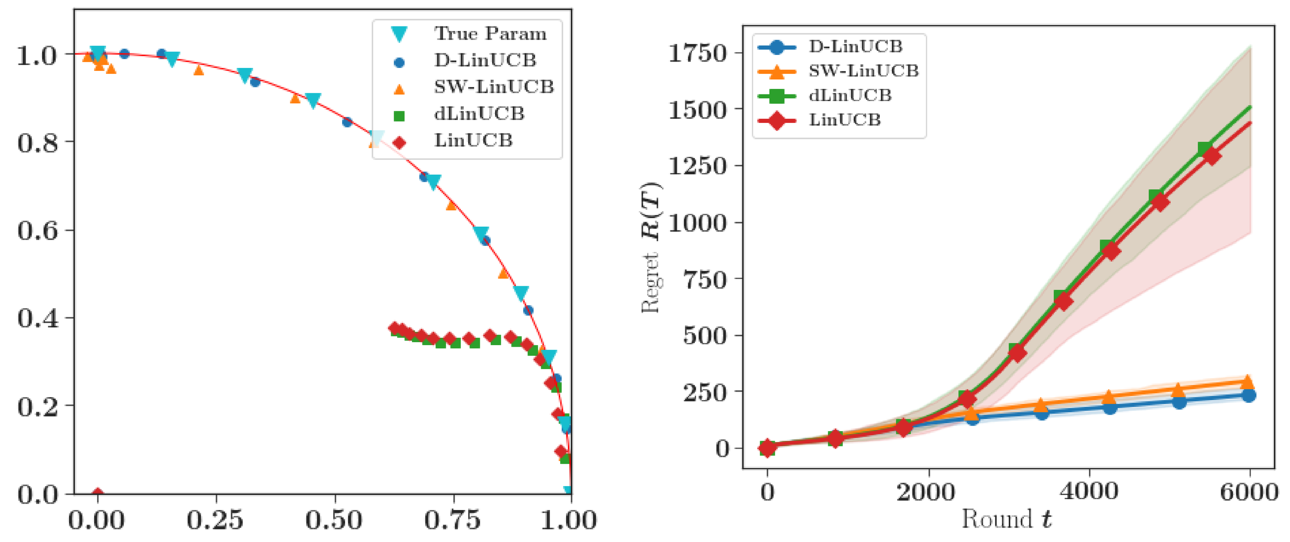

Fig. 2: (*left*) estimated parameters, (*right*) accumulated regret, both averaged on 100 runs

Asymptotic Upper Bound

Theorem 4. By choosing $\gamma = 1 - (B_T/(dT))^{2/3}$, the regret of the `D-LinUCB` algorithm is asymptotically upper bounded with high probability by $O(d^{2/3} B_T^{1/3} T^{2/3})$ when $T \to \infty$.

Conclusions and Remarks

- We assume that the variation budget is known all along this work. In Cheung et al. [2019], a first solution is presented to relax such hypothesis.
- Providing an algorithm with an upper bound matching the lower bound presented in Besbes et al. [2014] up to logarithmic terms.

References

O. Besbes, Y. Gur, and A. Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in neural information processing systems*, pages 199–207, 2014.

W. C. Cheung, D. Simchi-Levi, and R. Zhu. Learning to optimize under non-stationarity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1079–1087, 2019.

A. Garivier and E. Moulines. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pages 174–188. Springer, 2011.

Q. Wu, N. Iyer, and H. Wang. Learning contextual bandits in a non-stationary environment. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 495–504, New York, NY, USA, 2018. ACM.