Weighted Linear Bandits for Non-Stationary Environments

Yoan Russac[1], Claire Vernade[2] and Olivier Cappé[1]



[1] CNRS, Inria, ENS, Université PSL [2] Deepmind

Roadmap

The Non-Stationary Linear Model

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 $A_{t,k} \in \mathbb{R}^d$ (with $\|A_{t,k}\|_2 \leq L$)

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

Optimality Criteria

Dynamic Regret Minimization

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

$$\iff \min \ \underbrace{\mathbb{E}\left(\sum_{t=1}^{T} \max_{a\in\mathcal{A}_t}\langle a - A_t, \theta_t^\star\rangle\right)}_{\text{dynamic regret}}$$

Difference to Specific Cases

1. When $\mathcal{A}_t \to I_d = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix}$
   - The model reduces to the (non-stationary) multiarmed bandit model
   - If $\theta_t^\star = \theta^\star$, there is a single best action $a^\star$
   - It is only necessary to control the deviations of $\hat{\theta}_t$ in the principal directions

2. If $\mathcal{A}_t \to I_d \otimes A_t = \begin{pmatrix} A_t & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & A_t \end{pmatrix}$, with $(A_t)_{t \geq 1}$ i.i.d.
   - ϵ-greedy exploration (may be) efficient

Non-Stationarity and Bandits

Two different approaches are commonly used to deal with non-stationary bandits

- Detecting changes in the distribution of the arms
- Building methods that are (somewhat) robust to variations of the environment

Their performance depends on the assumptions made on the sequence of environment parameters $(\theta_t^\star)_{t \geq 1}$

- In abruptly changing environments, changepoint detection methods are more efficient
- But they may fail in slowly-changing environments
- We expect robust policies to perform well in both environments

Our Approach

We only focus on robust policies

With that in mind, the non-stationarity in the $\theta_t^\star$ parameter is measured with the variation budget

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

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

Roadmap

Some references

- Garivier *et al.*(2011), *On upper-confidence bound policies for switching bandit problems*, COLT

Introduce sliding window and exponential discounting algorithms, analyzing them in the abrupt changes setting and providing a $O(T^{1/2})$ lower bound

- Besbes *et al.*(2014), *Stochastic multi-armed-bandit problem with non-stationary rewards*, NeurIPS

Consider the variation budget, prouve a $O(T^{2/3})$ lower bound and analyze an epoch-based variant of Exp3

- Wu *et al.*(2018), *Learning contextual bandits in a non-stationary environment*, ACM SIGIR

Introduce an algorithm (called dLinUCB) based on change detection for the linear bandit

- Cheung *et al.*(2019), *Learning to optimize under non-stationarity*, AISTATS

Adapt the sliding-window algorithm to the linear bandit

Garivier *et al.* paper

Sliding-Window UCB algorithm

At time t the `SW-UCB` policy selects the action that maximizes

$$A_t = \mathop{\arg\max}_{i \in \{1,\dots K\}} \frac{\sum_{s=t-\tau+1}^{t} X_s \mathbb{1}(I_s = i)}{\sum_{s=t-\tau+1}^{t} \mathbb{1}(I_s = i)} + \sqrt{\frac{\xi \log(\min(t,\tau))}{\sum_{s=t-\tau+1}^{t} \mathbb{1}(I_s = i))}}$$

Discounted UCB algorithm

At time t the `D-UCB` policy selects the action that maximizes

$$A_t = \mathop{\arg\max}_{i \in \{1,\dots K\}} \frac{\sum_{s=1}^{t} \gamma^{t-s} X_s \mathbb{1}(I_s = i)}{\sum_{s=1}^{t} \gamma^{t-s} \mathbb{1}(I_s = i)} + 2\sqrt{\frac{\xi \log((1-\gamma^{-t})/(1-\gamma))}{\sum_{s=1}^{t} \gamma^{t-s} \mathbb{1}(I_s = i)}}$$

with $\gamma < 1$

Roadmap

Assumptions

At each round $t \geq 1$ the learner

- Receives a finite set of arbitrary feasible actions $\mathcal{A}_t \subset \mathbb{R}^d$
- Selects an $\mathcal{F}_t = \sigma(X_1, A_1, \ldots, X_{t-1}, A_{t-1})$–measurable action $A_t \in \mathcal{A}_t$

Other assumptions

- Sub-Gaussian Random Noise 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

Least Squares Estimator

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

Weighted Least Squares Estimator

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

Scale-Invariance Property

The weighted least squares estimator is given by

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

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

The Case of Exponential weights

Exponential Discount (Time-Dependent Weights)

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

Time-Independent Weights

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

$\hookrightarrow$ are equivalent, due to scale-invariance

Concentration Result

Theorem 1

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

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

where

$$V_t = \sum_{s=1}^{t} w_s A_s A_s^\top + \lambda_t I_d,$$

$$\widetilde{V}_t = \sum_{s=1}^{t} w_s^2 A_s A_s^\top + \mu_t I_d$$

On the Control of Deviations in the $V_t\widetilde{V}_t^{-1}V_t$ Norm

For the unweighted least squares estimator, the [Abbasi-Yadkori *et al.*, 2001] deviation bound features the $\|\hat{\theta}_t - \theta^\star\|_{V_t}$ norm

Here, the $V_t\widetilde{V}_t^{-1}V_t$ norm comes form the observation that

- The variance terms are related to w_s^2 which are featured in $\widetilde{V}_t$
- The weighted least squares estimator (and the matrix V_t) is defined with w_s

Remark: When $w_t = 1$, taking $\lambda_t = \mu_t$ yields $V_t\widetilde{V}_t^{-1}V_t = V_t$ and the usual concentration inequality

On the Role of μ_t

The sequence of parameters $(\mu_t)_{t\geq 1}$ is instrumental (results from the use of the Method of Mixtures) and could theoretically be chosen completely independently from λ_t and w_t

But taking μ_t proportional to λ_t^2, ensures that

- $V_t \widetilde{V}_t^{-1} V_t$ becomes scale-invariant
- $\lambda_t / \sqrt{\mu_t}$ becomes scale-invariant
- $\sum_{s=1}^{t} w_s^2 / \mu_t$ becomes scale-invariant

$\hookrightarrow$ Scale-invariant concentration inequality !

On the Use of Time-Dependent Regularization Parameters

- Using time-dependent regularization parameter λ_t, is required to avoid vanishing regularization
- In the sense that $d \log\left(1 + \frac{L^2 \sum_{s=1}^{t} w_s^2}{d\mu_t}\right)$ should not dominate the radius of the confidence region as t increases

In the setting with exponentially increasing weights $(w_s = \gamma^{-s})$

$$\lambda_t \propto w_t \quad \mu_t \propto \lambda_t^2$$

Roadmap

Concentration in the Non-Stationary Case

Moving back to the non-stationary environment $X_s = A_s^\top \theta_s^\star + \eta_s$ and assuming that $w_s = \gamma^{-s}$, $\lambda_s = \lambda\gamma^{-s}$

Let $\bar{\theta}_t = V_{t-1}^{-1}\left(\sum_{s=1}^{t-1}\gamma^{-s}A_sA_s^\top\theta_s^\star + \gamma^{t-1}\theta_t^\star\right)$ denote a "noiseless" proxy value for $\theta_t^\star$

Concentration in the Non-Stationary Case

Moving back to the non-stationary environment $X_s = A_s^\top \theta_s^\star + \eta_s$ and assuming that $w_s = \gamma^{-s}$, $\lambda_s = \lambda\gamma^{-s}$

Let $\bar{\theta}_t = V_{t-1}^{-1}\left(\sum_{s=1}^{t-1} \gamma^{-s} A_s A_s^\top \theta_s^\star + \gamma^{t-1}\theta_t^\star\right)$ denote a "noiseless" proxy value for $\theta_t^\star$

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*

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

Then, $\forall\delta > 0$,

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

D-LinUCB Algorithm (1)

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 $\mathsf{UCB}(a) = a^\top\hat{\theta} + \beta_{t-1}\sqrt{a^\top V^{-1}\widetilde{V}V^{-1}a}$

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

  Play action A_t and **receive reward** X_t

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

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

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

D-LinUCB Algorithm (2)

Thanks to the scale-invariance property, for numerical stability of the implementation, we consider time-dependent weights

$$w_{t,s} = \gamma^{t-s} \quad \text{for} \quad 1 \leq s \leq t$$

The weighted least squares estimator is solution of

$$\hat{\theta}_t = \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t} \gamma^{t-s}(X_s - \langle A_s, \theta \rangle)^2 + \lambda/2 \|\theta\|_2^2$$

$\hookrightarrow$ this form is numerically stable and can be implemented recursively (but we revert to the standard form for the analysis)

D-LinUCB Algorithm (3)

And as usual, we consider optimistic arm selection in the sense that

$$A_t = \arg\max_{a \in \mathcal{A}_t} \max_{\theta} \langle a, \theta \rangle \quad \text{s.t.} \quad \underbrace{\|\theta - \hat{\theta}_{t-1}\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}} \leq \beta_{t-1}}_{\theta \in \mathcal{C}_t}$$

which is equivalent to

$$A_t = \arg\max_{a \in \mathcal{A}_t} \langle a, \hat{\theta}_{t-1} \rangle + \beta_{t-1} \|a\|_{V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}}$$

Theoretical Analysis

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*

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

Regret Decomposition

Let $\theta_t = \arg\max_{\theta \in \mathcal{C}_t} \langle A_t, \theta \rangle$ and $A_t^\star = \arg\max_{a \in \mathcal{A}_t} \langle a, \theta_t^\star \rangle$

We have $\langle A_t^\star, \bar{\theta}_t \rangle \leq \langle A_t, \theta_t \rangle$

Thus,

$$
\begin{aligned}
r_t &= \langle \max_{a \in \mathcal{A}_t} a - A_t, \theta_t^\star \rangle = \langle A_t^\star - A_t, \theta_t^\star \rangle \\
&= \langle A_t^\star - A_t, \bar{\theta}_t \rangle + \langle A_t^\star - A_t, \theta_t^\star - \bar{\theta}_t \rangle \\
&\leq \langle A_t, \bar{\theta}_t - \theta_t \rangle + \langle A_t^\star - A_t, \theta_t^\star - \bar{\theta}_t \rangle \\
&\leq \|A_t\|_{V_{t-1}^{-1} \widetilde{V}_{t-1} V_{t-1}^{t-1}} \|\bar{\theta}_t - \theta_t\|_{V_{t-1} \widetilde{V}_{t-1}^{-1} V_{t-1}} + \|A_t^\star - A_t\|_2 \|\theta_t^\star - \bar{\theta}_t\|_2 \quad \text{(C-S)} \\
&\leq \|A_t\|_{V_{t-1}^{-1} \widetilde{V}_{t-1} V_{t-1}^{-1}} \underbrace{\|\bar{\theta}_t - \theta_t\|_{V_{t-1} \widetilde{V}_{t-1}^{-1} V_{t-1}}}_{\substack{\text{Deviation term} \\ \leq 2\beta_{t-1} \text{ with prob. } 1-\delta}} + 2L \underbrace{\|\theta_t^\star - \bar{\theta}_t\|_2}_{\text{Bias term}}
\end{aligned}
$$

Controlling the Bias (1)

Let $D > 0$,

$$
\begin{aligned}
\|\theta_t^\star - \bar{\theta}_t\|_2 &= \|V_{t-1}^{-1} \sum_{s=1}^{t-1} \gamma^{-s} A_s A_s^\top (\theta_s^\star - \theta_t^\star)\|_2 \\
&\leq \| \sum_{s=t-D}^{t-1} V_{t-1}^{-1} \gamma^{-s} A_s A_s^\top (\theta_s^\star - \theta_t^\star)\|_2 + \|V_{t-1}^{-1} \sum_{s=1}^{t-D-1} \gamma^{-s} A_s A_s^\top (\theta_s^\star - \theta_t^\star)\|_2 \\
&\leq \| \sum_{s=t-D}^{t-1} V_{t-1}^{-1} \gamma^{-s} A_s A_s^\top \sum_{p=s}^{t-1} (\theta_p^\star - \theta_{p+1}^\star)\|_2 + \| \sum_{s=1}^{t-D-1} \gamma^{-s} A_s A_s^\top (\theta_s^\star - \theta_t^\star)\|_{V_{t-1}^{-2}} \\
&\leq \| \sum_{p=t-D}^{t-1} V_{t-1}^{-1} \gamma^{-s} A_s A_s^\top \sum_{s=t-D}^{p} (\theta_p^\star - \theta_{p+1}^\star)\|_2 + \sum_{s=1}^{t-D-1} \frac{\gamma^{t-1-s}}{\lambda} \|A_s A_s^\top (\theta_s^\star - \theta_t^\star)\|_2 \\
&\leq \sum_{p=t-D}^{t-1} \|V_{t-1}^{-1} \sum_{s=t-D}^{p} \gamma^{-s} A_s A_s^\top (\theta_p^\star - \theta_{p+1}^\star)\|_2 + \frac{2L^2 S}{\lambda} \sum_{s=1}^{t-D-1} \gamma^{t-1-s} \\
&\leq \sum_{p=t-D}^{t-1} \lambda_{\max} \left(V_{t-1}^{-1} \sum_{s=t-D}^{p} \gamma^{-s} A_s A_s^\top \right) \|\theta_p^\star - \theta_{p+1}^\star\|_2 + \frac{2L^2 S}{\lambda} \frac{\gamma^D}{1-\gamma}.
\end{aligned}
$$

Controlling the Bias (2)

- It is essential to introduce the D term and to control the two terms differently
- The oldest terms $(s < t - D)$ have fewer importance and can be bounded roughly
- For the most recent terms $(t - D \leq s \leq t - 1)$, a more precise analysis is necessary

Optimal Asymptotic Regret

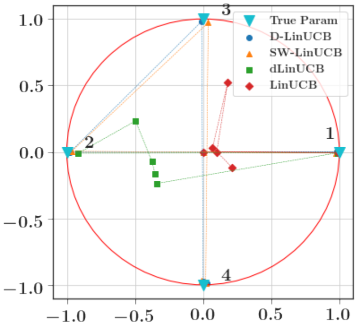
Theorem 4

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* $T \to \infty$.

*And $D = \log(T)/(1-\gamma)$

Roadmap

Performance in Abruptly-Changing Environment



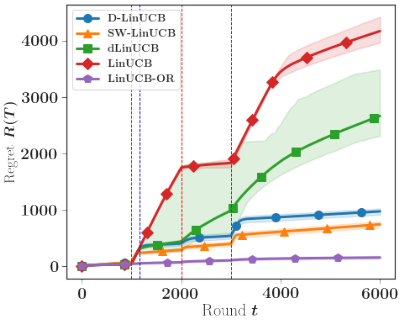


Figure: Performances of the algorithms in the abrutply-changing environment. The plot on the left correspond to the estimated parameter and the one on the right to the accumulated regret, averaged on $N = 100$ independent experiments

Performance in Slowly-Changing Environment

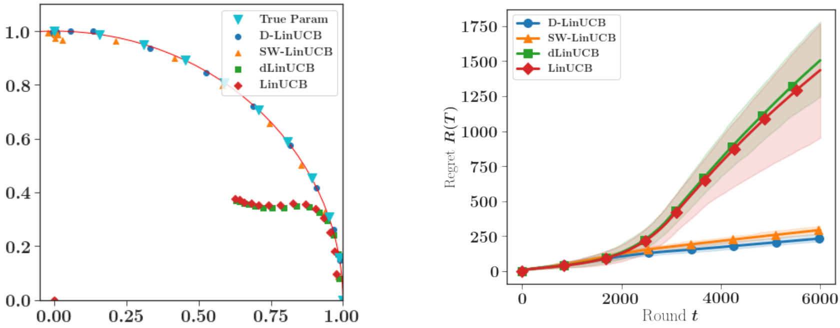


Figure: Performances of the algorithms in the slowly-varying environment. The plot on the left correspond to the estimated parameter and the one on the right to the accumulated regret, averaged on $N = 100$ independent experiments