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

1. The Model
2. Related work
3. Concentration Result
4. Application to Non-Stationary Linear Bandits
5. Empirical Performances
The Non-Stationary Linear Model

At time $t$, the learner has access to a time-dependent finite set of arbitrary actions $A_t = \{A_{t,1}, \ldots, A_{t,K_t}\}$, 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 $A_t \in A_t$
- and observes only the noisy linear reward $X_t = A_t^\top \theta_t^* + \eta_t$
  where $\eta_t$ is a $\sigma$-subgaussian random noise

Specificity of the model

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

Dynamic Regret Minimization

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

dynamic regret
The Model

Difference to Specific Cases

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

2. If $A_t \rightarrow I_d \otimes A_t = \begin{pmatrix} A_t & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & A_t \end{pmatrix}$, with $(A_t)_{t \geq 1}$ i.i.d.
   - $\epsilon$-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 \((\theta^*_t)_{t \geq 1}\)

- 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 $\theta_t^*$ parameter is measured with the variation budget

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

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

1 The Model

2 Related work

3 Concentration Result

4 Application to Non-Stationary Linear Bandits

5 Empirical Performances
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, prove 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

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

**Discounted UCB algorithm**

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

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

with $\gamma < 1$
Roadmap

1. The Model
2. Related work
3. Concentration Result
4. Application to Non-Stationary Linear Bandits
5. Empirical Performances
Assumptions

At each round $t \geq 1$ the learner

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

Other assumptions

- **Sub-Gaussian Random Noise** $\eta_t$ is, conditionally on the past, $\sigma$-subgaussian
- **Bounded Actions** $\forall t \geq 1, \forall a \in A_t, \|a\|_2 \leq L$
- **Bounded Parameters** $\forall t \geq 1, \|\theta^*_t\|_2 \leq S$
- $\forall t \geq 1, \forall a \in A_t, |\langle a, \theta^*_t \rangle| \leq 1$
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

\[ \hat{\theta}_t = \arg \min_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t} w_s (X_s - A_s^\top \theta)^2 + \frac{\lambda_t}{2} \| \theta \|_2^2 \]
The weighted least squares estimator is given by

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

$\hat{\theta}_t$ is unchanged if all the weights $w_s$ and the regularization parameter $\lambda_t$ are multiplied by a same constant $\alpha$
The Case of Exponential weights

**Exponential Discount (Time-Dependent Weights)**

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

**Time-Independent Weights**

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

are equivalent, due to scale-invariance
Theorem 1

Assuming that $\theta_t^* = \theta^*$, 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 $1 - \delta$,

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

where

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

$$
\hat{V}_t = \sum_{s=1}^{t} w_s^2 A_s A_s^\top + \mu_t I_d
$$
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 $\|\hat{\theta}_t - \theta^*\|_{V_t}$ norm.

Here, the $V_t \widetilde{V}_t^{-1} V_t$ norm comes from 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 $\mu_t$

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 $\lambda_t$ and $w_t$

But taking $\mu_t$ proportional to $\lambda_t^2$, ensures that

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

$\rightarrow$ Scale-invariant concentration inequality!
On the Use of Time-Dependent Regularization Parameters

- Using time-dependent regularization parameter $\lambda_t$, 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 ($w_s = \gamma^{-s}$)

\[
\begin{align*}
\lambda_t &\propto w_t \\
\mu_t &\propto \lambda_t^2
\end{align*}
\]
Roadmap

1. The Model
2. Related work
3. Concentration Result
4. Application to Non-Stationary Linear Bandits
5. Empirical Performances
Concentration in the Non-Stationary Case

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

Let $\bar{\theta}_t = V_{t-1}^{-1} \left( \sum_{s=1}^{t-1} \gamma^{-s} A_s A_s^\top \theta^*_s + \gamma^{t-1} \theta^*_t \right)$ denote a “noiseless” proxy value for $\theta^*_t$. 
Concentration in the Non-Stationary Case

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

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

**Theorem 2**

Let $C_t = \{ \theta \in \mathbb{R}^d : \| \theta - \hat{\theta}_{t-1} \|_{V_{t-1} \tilde{V}_{t-1} V_{t-1}} \leq \beta_{t-1} \}$ 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, $\forall \delta > 0$,

$$
\mathbb{P} \left( \forall t \geq 1, \bar{\theta}_t \in C_t \right) \geq 1 - \delta
$$
D-LinUCB Algorithm (1)

**Algorithm 1: D-LinUCB**

**Input:** Probability $\delta$, subgaussianity constant $\sigma$, dimension $d$, regularization $\lambda$, upper bound for actions $L$, upper bound for parameters $S$, discount factor $\gamma$.

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

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

Receive $A_t$, 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** $a \in A_t$ **do**

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

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

Play action $A_t$ and receive reward $X_t$

**Updating phase:**

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

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

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

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

\(\rightarrow\) this form is numerically stable and can be implemented recursively (but we revert to the standard form for the analysis)
And as usual, we consider optimistic arm selection in the sense that

\[ A_t = \arg \max_{a \in A_t} \max_{\theta} \langle a, \theta \rangle \quad \text{s.t.} \quad \|\theta - \hat{\theta}_{t-1}\|_{V_{t-1} \tilde{V}_{t-1} V_{t-1}} \leq \beta_{t-1} \]

\[ \hat{\theta}_{t-1} \in \mathcal{C}_t \]

which is equivalent to

\[ A_t = \arg \max_{a \in A_t} \langle a, \hat{\theta}_{t-1} \rangle + \beta_{t-1} \|a\|_{V_{t-1} \tilde{V}_{t-1} V_{t-1}} \]
Theoretical Analysis

**Theorem 3**

Assuming that $\sum_{s=1}^{T-1} \| \theta_s^* - \theta_{s+1}^* \|_2 \leq B_T$, 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

$$R_T \leq 2LDB_T + \frac{4L^3S}{\lambda} \frac{\gamma^D}{1 - \gamma} T + 2\sqrt{2\beta T} \sqrt{dT} \sqrt{T \log(1/\gamma) + \log \left( 1 + \frac{L^2}{d\lambda(1 - \gamma)} \right)}$$
Let $\theta_t = \arg \max_{\theta \in C_t} \langle A_t, \theta \rangle$ and $A_t^* = \arg \max_{a \in A_t} \langle a, \theta_t^* \rangle$

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

Thus,

$$r_t = \langle \max_{a \in A_t} a - A_t, \theta_t^* \rangle = \langle A_t^* - A_t, \theta_t^* \rangle$$

$$= \langle A_t^* - A_t, \bar{\theta}_t \rangle + \langle A_t^* - A_t, \theta_t^* - \bar{\theta}_t \rangle$$

$$\leq \langle A_t, \bar{\theta}_t - \theta_t \rangle + \langle A_t^* - A_t, \theta_t^* - \bar{\theta}_t \rangle$$

$$\leq \|A_t\| V_t^{-1} \tilde{V}_{t-1} V_{t-1}^{-1} \|\bar{\theta}_t - \theta_t\| V_{t-1} \tilde{V}_{t-1}^{-1} V_{t-1} + \|A_t^* - A_t\|_2 \|\theta_t^* - \bar{\theta}_t\|_2 \quad \text{(C-S)}$$

$$\leq \|A_t\| V_t^{-1} \tilde{V}_{t-1} V_{t-1} \|\bar{\theta}_t - \theta_t\| V_{t-1} \tilde{V}_{t-1}^{-1} V_{t-1} \quad \text{Deviation term}$$

$$+ 2L \|\theta_t^* - \bar{\theta}_t\|_2 \quad \text{Bias term}$$

$$\leq 2 \beta_{t-1} \text{ with prob. } 1 - \delta$$
Controlling the Bias (1)

Let $D > 0$,

\[
\|\theta_t^* - \bar{\theta}_t\|_2 = \|V_{t-1}^{-1} \sum_{s=1}^{t-1} \gamma^{-s} A_s A_s^\top (\theta_s^* - \theta_t^*)\|_2
\]

\[
\leq \| \sum_{s=t-D}^{t-1} V_{t-1}^{-1} \gamma^{-s} A_s A_s^\top (\theta_s^* - \theta_t^*)\|_2 + \| V_{t-1}^{-1} \sum_{s=1}^{t-D-1} \gamma^{-s} A_s A_s^\top (\theta_s^* - \theta_t^*)\|_2
\]

\[
\leq \| \sum_{s=t-D}^{t-1} V_{t-1}^{-1} \gamma^{-s} A_s A_s^\top \sum_{p=s}^{p} (\theta_p^* - \theta_{p+1}^*)\|_2 + \| \sum_{s=1}^{t-D-1} \gamma^{-s} A_s A_s^\top (\theta_s^* - \theta_t^*)\|_{V_{t-1}^{-2}}
\]

\[
\leq \| \sum_{p=t-D}^{t-1} V_{t-1}^{-1} \gamma^{-s} A_s A_s^\top \sum_{s=t-D}^{p} (\theta_p^* - \theta_{p+1}^*)\|_2 + \sum_{s=1}^{t-D-1} \gamma^{t-1-s} \| A_s A_s^\top (\theta_s^* - \theta_t^*)\|_2
\]

\[
\leq \sum_{p=t-D}^{t-1} \| V_{t-1}^{-1} \sum_{s=t-D}^{p} \gamma^{-s} A_s A_s^\top (\theta_p^* - \theta_{p+1}^*)\|_2 + \frac{2L^2 S}{\lambda} \sum_{s=1}^{t-D-1} \gamma^{t-1-s}
\]

\[
\leq \sum_{p=t-D}^{t-1} \lambda_{\text{max}} \left( V_{t-1}^{-1} \sum_{s=t-D}^{p} \gamma^{-s} A_s A_s^\top \right) \| \theta_p^* - \theta_{p+1}^* \|_2 + \frac{2L^2 S}{\lambda} \frac{\gamma^D}{1 - \gamma}.
\]
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

1. The Model
2. Related work
3. Concentration Result
4. Application to Non-Stationary Linear Bandits
5. Empirical Performances
Performance in Abruptly-Changing Environment

**Figure:** Performances of the algorithms in the abruptly-changing environment. The plot on the left corresponds 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.