Non-linearity in Parametric Bandits: the Logistic Bandit case

joint work with **M. Abeille[1], C. Calauzènes[1] and O. Fercoq[2]**

[1] Criteo AI Lab [2] LTCI TelecomParis

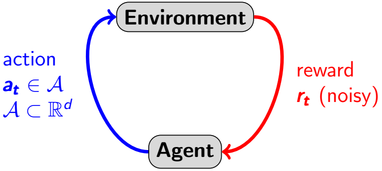
Laboratoire Traitement et Communication de l'Information

Presentation Outline

- **Goal.** Study **non-linearity** in sequential decision making problem.
  - Logistic Bandit: theoretical qualities.
    - $\rightsquigarrow$ Simple extension of the Linear Bandit.
    - $\rightsquigarrow$ Isolates the effect of non-linearity.
  - Logistic Bandit: practical relevance.
    - $\rightsquigarrow$ Model real-life problems with **binary** feedback.
    - $\rightsquigarrow$ news recommandation, clinical trials, ..
- **Logistic Bandit: high-level contributions.**
  - Improved algorithms with enhanced performances.
  - New theoretical insights: non-linearity makes the problem **easier**.

Warm-Up: Linear Bandits

- Repeated game with **structured** feedback.



noise

$$r_t = a_t^\top \theta_\star + \eta_t$$

unknown

- **Motivation.** Generalizes the classical Multi-Arm Bandit setting
  - encode similarities between actions.
  - handle infinite number of actions.
  - handle contextual information x_t: $f(\phi(a_t, x_t)^\top \theta_\star)$.

Warm-up: Linear Bandit (ctn'd)

- **Goal.** Minimize cumulative **regret**; with $\boldsymbol{a}_\star = \text{argmax}_{a \in \mathcal{A}}\, a^\mathsf{T} \boldsymbol{\theta}_\star$:

$$\text{Regret}_{\theta_\star}(T) := T \boldsymbol{a}_\star{}^\mathsf{T} \boldsymbol{\theta}_\star - \sum_{t=1}^{T} \boldsymbol{a}_t{}^\mathsf{T} \boldsymbol{\theta}_\star \ .$$

  ⇝ objective: $\text{Regret}_{\theta_\star}(T) \leq T^\alpha$ for $\alpha < 1$.

  ⇝ balance exploitation and exploration.

Warm-up: Linear Bandit (ctn'd)

- **Goal.** Minimize cumulative **regret**; with $\boldsymbol{a_\star} = \text{argmax}_{a\in\mathcal{A}}\, a^\mathsf{T}\boldsymbol{\theta_\star}$:

$$\text{Regret}_{\theta_\star}(T) := T\boldsymbol{a_\star}^\mathsf{T}\boldsymbol{\theta_\star} - \sum_{t=1}^{T}\boldsymbol{a_t}^\mathsf{T}\boldsymbol{\theta_\star}\ .$$

$\rightsquigarrow$ objective: $\text{Regret}_{\theta_\star}(T) \leq T^\alpha$ for $\alpha < 1$.

$\rightsquigarrow$ balance exploitation and exploration.

- **Solved**: **minimax-optimal** and efficient algorithms.

$$\text{Regret}(T) = \widetilde{\mathcal{O}}(d\sqrt{T})\ ,$$

where $\widetilde{\mathcal{O}}$ hides only logarithmic dependencies.

Warm-up: Linear Bandit (ctn'd)

- **Exploration/exploitation** trade-off vs. **optimism** in face of uncertainty.
  - **Learning** is performed via ordinary least-squares:

$$\hat{\theta}_t := \mathsf{V}_t^{-1}\left(\sum_{s=1}^{t-1} r_s a_s\right) \quad \text{where} \quad \mathsf{V}_t^{-1} = \sum_{s=1}^{t} a_s a_s^\mathsf{T} + \lambda \mathsf{I}_d \ .$$

Warm-up: Linear Bandit (ctn'd)

- **Exploration/exploitation** trade-off vs. **optimism** in face of uncertainty.
  - **Learning** is performed via ordinary least-squares:

$$\hat{\theta}_t := \mathrm{V}_t^{-1}\Big(\sum_{s=1}^{t-1} r_s a_s\Big) \quad \text{where} \quad \mathrm{V}_t^{-1} = \sum_{s=1}^{t} a_s a_s^\mathsf{T} + \lambda \mathrm{I}_d \ .$$

  - **Planning** by resorting to confidence sets:

$$\boldsymbol{\theta}_\star \in \mathcal{C}_t(\delta) = \left\{\theta,\ \left\|\theta - \hat{\theta}_t\right\|_{\mathrm{V}_t}^2 \leq d\log(t/\delta).\right\} \quad \text{with proba. at least } 1-\delta$$

and enforcing optimism:

the hard part

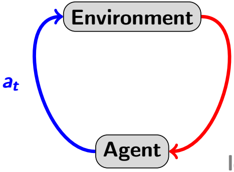
$$\text{play } a_{t+1} \in \operatorname{argmax}_{a\in\mathcal{A}} \max_{\theta\in\mathcal{C}_t(\delta)} a^\mathsf{T}\theta \ .$$

The Logistic Bandit

- **Motivations.** The Linear Bandit setting has (many) limitations;
  - **Theoretical**: towards rich reward models.
    - $\rightsquigarrow$ The real world is fundamentally non-linear.
    - $\rightsquigarrow$ Does the same principle work?
    - $\rightsquigarrow$ Will it be optimal?
  - **Practical**:
    - $\rightsquigarrow$ The Linear Bandit covers only continuous rewards.
    - $\rightsquigarrow$ What about binary rewards (click, sale, success)?

The Logistic Bandit

- Repeated game with structured **binary** feedback.



reward $\in \{0, 1\}$

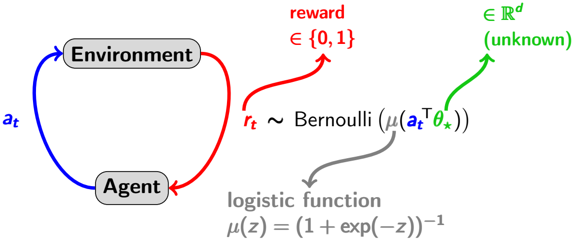
$\in \mathbb{R}^d$ (unknown)

$$r_t \sim \text{Bernoulli}\left(\mu({a_t}^\top \theta_\star)\right)$$

logistic function
$\mu(z) = (1 + \exp(-z))^{-1}$

The Logistic Bandit

- Repeated game with structured **binary** feedback.

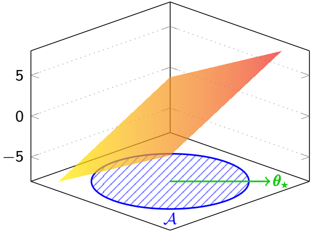


- **Regret.** The agent tries to minimize its cumulative pseudo-regret:

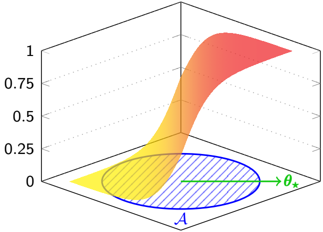
$$\text{Regret}_{\theta_\star}(T) := T\mu({a_\star}^\mathsf{T}\theta_\star) - \sum_{t=1}^{T}\mu({a_t}^\mathsf{T}\theta_\star)\ .$$

The Learning Problem (ctn'd)

- **Reward model.** Minimalist non-linear extension from the linear bandit.



$$\mathbb{E}[r_t | a_t] = a_t^\mathsf{T}\theta_\star$$



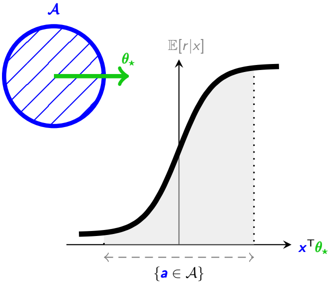
$$\mathbb{E}[r_t | a_t] = (1 + \exp(-a_t^\mathsf{T}\theta_\star))^{-1}$$

- **Exploration-exploitation.** Same recipe:
  - Learning: maximum likelihood (**logistic regression**).
  - Planning: Optimism through confidence sets.

- **Additional challenge.** Non-linearity: information vs. regret.

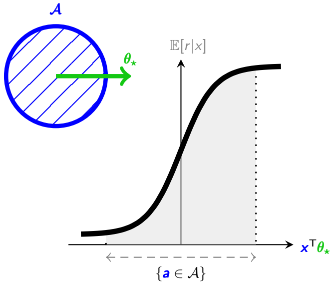
Quantifying non-linearity

- **Level of non-linearity = conditioning.**
  - How **flat** are the tails.



- **Important quantities.** The level of non-linearity is problem-dependent.

Quantifying non-linearity

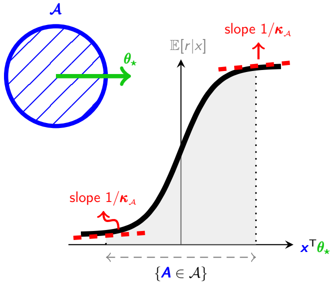


- **Level of non-linearity = conditioning.**
  - How **flat** are the tails.

- **Important quantities.** The level of non-linearity is problem-dependent.
  - Historically characterized by a constant $\kappa_{\mathcal{A}}$:

$$\kappa_{\mathcal{A}} := \frac{1}{\min_{a\in\mathcal{A}} \dot{\mu}(a^\top \theta_\star)} \,.$$

Quantifying non-linearity



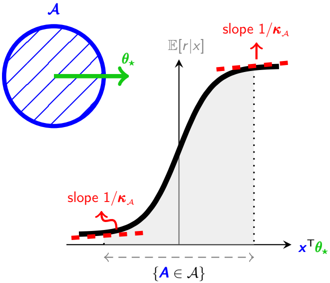
- **Level of non-linearity = conditioning.**
  - How **flat** are the tails.

- **Important quantities.** The level of non-linearity is problem-dependent.
  - Historically characterized by a constant $\kappa_{\mathcal{A}}$:

$$\kappa_{\mathcal{A}} := \frac{1}{\min_{a\in\mathcal{A}} \dot{\mu}(a^\top\theta_\star)} \; .$$

  - The more non-linear the reward, the **bigger**.

Quantifying non-linearity



- **Level of non-linearity = conditioning.**
  - How **flat** are the tails.

- **Important quantities.** The level of non-linearity is problem-dependent.
  - Historically characterized by a constant $\kappa_\mathcal{A}$:

$$\kappa_\mathcal{A} := \frac{1}{\min_{a\in\mathcal{A}} \dot{\mu}(a^\top\theta_\star)} .$$

  - The more non-linear the reward, the **bigger**.
  - Typically $\kappa_\mathcal{A} \geq \exp(\|\theta_\star\|)$! In practical case; $\kappa_\mathcal{A} \sim 10^3$.

Previous approaches

- A lot of existing work on the logistic bandit: [Filippi et al. 2010; Li et al 2017; Kveton et al. 2019; Dong et al. 2019];
- All rely on a global **linearization** approach.

Previous approaches

- A lot of existing work on the logistic bandit: [Filippi et al. 2010; Li et al 2017; Kveton et al. 2019; Dong et al. 2019];
- All rely on a global **linearization** approach.
- Information vs. regret : worst of both world!
  - Confidence set (at algorithmic design time):

$$\boldsymbol{\theta}_\star \in \mathcal{C}_t(\delta) = \left\{\theta,\ \left\|\theta - \hat{\theta}_t\right\|^2_{\mathbf{V}_t} \leq \boldsymbol{\kappa}_{\mathcal{A}} d \log(t/\delta)\right\} \quad \text{with proba. at least } 1-\delta$$

Previous approaches

- A lot of existing work on the logistic bandit: [Filippi et al. 2010; Li et al 2017; Kveton et al. 2019; Dong et al. 2019];

- All rely on a global **linearization** approach.

- Information vs. regret : worst of both world!
  - Confidence set (at algorithmic design time):

$$\boldsymbol{\theta}_\star \in \mathcal{C}_t(\delta) = \left\{\theta,\ \left\lVert\theta - \hat{\theta}_t\right\rVert^2_{\mathbf{V}_t} \leq \boldsymbol{\kappa}_{\mathcal{A}}\, d\log(t/\delta)\right\} \qquad \text{with proba. at least } 1-\delta$$

  - Prediction error (at analysis time):

$$\mu(\boldsymbol{a}^\mathsf{T}\boldsymbol{\theta}_\star) - \mu(\boldsymbol{a}^\mathsf{T}\theta) \leq \boldsymbol{a}^\mathsf{T}(\boldsymbol{\theta}_\star - \theta)/4\ .$$

Previous approaches (ctn'd)

- Global linearization $\Rightarrow$ disappointing results!
  - Poor regret guarantees.

$$\text{Regret}_{\theta_\star}(T) = \widetilde{\mathcal{O}}\left(\boldsymbol{\kappa}_{\mathcal{A}} d\sqrt{T}\right) .$$

  - Because the algorithms are **over-explorative**.
  - Disappointing story about the effects of non-linearity.

$\rightsquigarrow$ The more non-linear the problem, the larger the regret!

Previous approaches (ctn'd)

- Global linearization $\Rightarrow$ disappointing results!
  - Poor regret guarantees.

$$\text{Regret}_{\theta_\star}(T) = \widetilde{\mathcal{O}}\left(\boldsymbol{\kappa}_{\mathcal{A}} d\sqrt{T}\right) .$$

  - Because the algorithms are **over-explorative**.
  - Disappointing story about the effects of non-linearity.
  - $\rightsquigarrow$ The more non-linear the problem, the larger the regret!

- Our goal is to improve this with:
  - Enhanced confidence sets for $\boldsymbol{\theta}_\star$.
  - Improvement treatment of the **local** behavior of the reward signal.

Improved confidence set

- Let $\mathrm{H}_t(\theta) = \sum_{s=1}^{t-1} \dot{\mu}(a_s^\mathsf{T}\theta) a_s a_s^\mathsf{T} + \lambda \mathrm{I}_d$; then

$$\boxed{\theta_\star \in \mathcal{E}_t(\delta) := \left\{\theta,\ \left\|\theta - \hat{\theta}_t\right\|_{\mathrm{H}_t(\theta)}^2 \leq d \log(t/\delta)\right\} \text{ with proba } \geq 1-\delta\ .}$$

- Based on a **new concentration inequality** for self-normalized process.

- Smaller than $\mathcal{C}_t(\delta)$ by **at least** $\sqrt{\boldsymbol{\kappa}_{\mathcal{A}}}$.

- Undergoes convex relaxation for tractability.

Improved confidence set

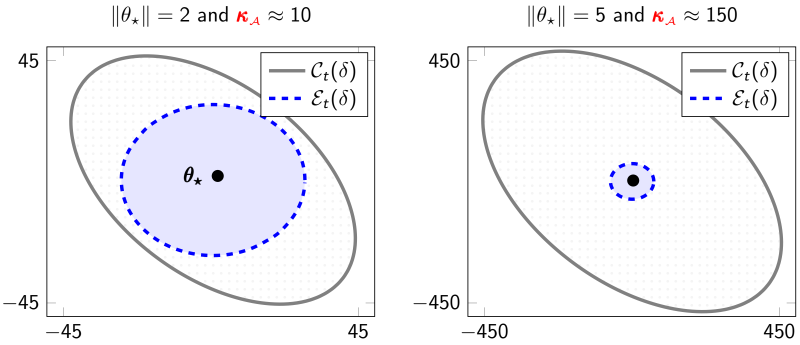


Figure: Visualization of two-dimensional Logistic bandit confidence sets.

- Smaller confidence set $\Rightarrow$ less explorative algorithm, better performance.

Improved algorithm and analysis

- We use the same recipe for enforcing **optimism**:

$$\text{play } a_t = \text{argmax}_{a \in \mathcal{A}} \max_{\theta \in \mathcal{E}_t(\delta)} \mu(a^\mathsf{T}\theta) \,.$$

Improved algorithm and analysis

- We use the same recipe for enforcing **optimism**:

$$\text{play } a_t = \text{argmax}_{a\in\mathcal{A}} \max_{\theta\in\mathcal{E}_t(\delta)} \mu(a^\mathsf{T}\theta) \; .$$

- We introduce a new analysis for the Logistic Bandit:

Improved algorithm and analysis

- We use the same recipe for enforcing **optimism**:

$$\text{play } a_t = \text{argmax}_{a\in\mathcal{A}} \max_{\theta\in\mathcal{E}_t(\delta)} \mu(a^\mathsf{T}\theta) \; .$$

- We introduce a new analysis for the Logistic Bandit:
  - Leverage the **self-concordance** property of the logistic function.
  - Allows for a **local** treatment of the non-linearity.
  - Strikes the right balance between information and regret.

Regret guarantees

- Enhanced regret guarantee; denote $a_\star$ the **best** action:

$$\mathrm{Regret}_{\theta_\star}(T) = \widetilde{\mathcal{O}}\left(d\sqrt{\dot{\mu}(a_\star^\mathsf{T}\theta_\star)T}\right) \ .$$

Regret guarantees

- Enhanced regret guarantee; denote $\boldsymbol{a_\star}$ the **best** action:

$$\text{Regret}_{\theta_\star}(T) = \widetilde{\mathcal{O}}\left(d\sqrt{\dot{\mu}(a_\star^\mathsf{T}\theta_\star)T}\right) .$$

- Illustration for the unit ball arm-set: $\dot{\mu}(a_\star^\mathsf{T}\theta_\star) = 1/\boldsymbol{\kappa}_\mathcal{A} \approx \exp(-\|\boldsymbol{\theta_\star}\|)$.
  - The regret is:

$$\boxed{\text{Regret}_{\theta_\star}(T) = \widetilde{\mathcal{O}}\left(d\sqrt{T/\boldsymbol{\kappa}_\mathcal{A}}\right) .}$$

  - Improvement by $\boldsymbol{\kappa}_\mathcal{A}{}^{3/2} \approx \exp(3\|\boldsymbol{\theta_\star}\|/2)$!

Regret guarantees

- Enhanced regret guarantee; denote $a_\star$ the **best** action:

$$\text{Regret}_{\theta_\star}(T) = \widetilde{\mathcal{O}}\left(d\sqrt{\dot{\mu}(a_\star^\mathsf{T}\theta_\star)T}\right) .$$

- Illustration for the unit ball arm-set: $\dot{\mu}(a_\star^\mathsf{T}\theta_\star) = 1/\kappa_{\mathcal{A}} \approx \exp(-\|\theta_\star\|)$.
  - The regret is:

$$\boxed{\text{Regret}_{\theta_\star}(T) = \widetilde{\mathcal{O}}\left(d\sqrt{T/\kappa_{\mathcal{A}}}\right) .}$$

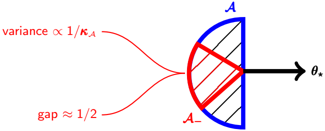
  - Improvement by ${\kappa_{\mathcal{A}}}^{3/2} \approx \exp(3\|\theta_\star\|/2)$!

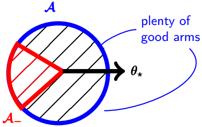
- This rate is **minimax-optimal** w.r.t d, T and $\kappa_{\mathcal{A}}$.

Effects of non-linearity

- Non-linearity seems to be beneficial!
  - The larger $\boldsymbol{\kappa}_{\mathcal{A}}$, the smaller the regret!
- Not entirely true; non-linearity can impact a **transitory phase**.
  - Second-order term of the regret.
  - Happens before highly rewarding areas of $\mathcal{A}$ are identified.

$$\text{Regret}_{\theta_\star}(T) = \widetilde{\mathcal{O}}\left(d\sqrt{\dot{\mu}(a_\star^\mathsf{T}\theta_\star)T} + R^{\text{transitory}}(T)\right) .$$

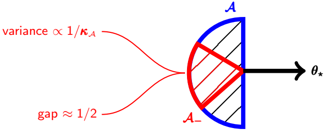


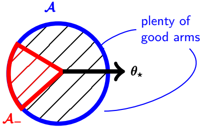


Effects of non-linearity

- Non-linearity seems to be beneficial!
  - The larger $\boldsymbol{\kappa}_{\mathcal{A}}$, the smaller the regret!
- Not entirely true; non-linearity can impact a **transitory phase**.
  - Second-order term of the regret.
  - Happens before highly rewarding areas of $\mathcal{A}$ are identified.

$$\text{Regret}_{\theta_\star}(T) = \widetilde{\mathcal{O}}\left(d\sqrt{\dot{\mu}(a_\star^\mathsf{T}\theta_\star)T} + R^{\text{transitory}}(T)\right) .$$



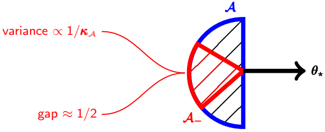


$\boxed{R^{\text{transitory}}(T) = \tilde{\mathcal{O}}(\boldsymbol{\kappa}_{\mathcal{A}})}$

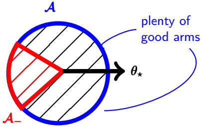
Effects of non-linearity

- Non-linearity seems to be beneficial!
  - The larger $\boldsymbol{\kappa}_{\mathcal{A}}$, the smaller the regret!
- Not entirely true; non-linearity can impact a **transitory phase**.
  - Second-order term of the regret.
  - Happens before highly rewarding areas of $\mathcal{A}$ are identified.

$$\text{Regret}_{\theta_\star}(T) = \widetilde{\mathcal{O}}\left(d\sqrt{\dot{\mu}(a_\star^\mathsf{T}\theta_\star)T} + R^{\text{transitory}}(T)\right) .$$



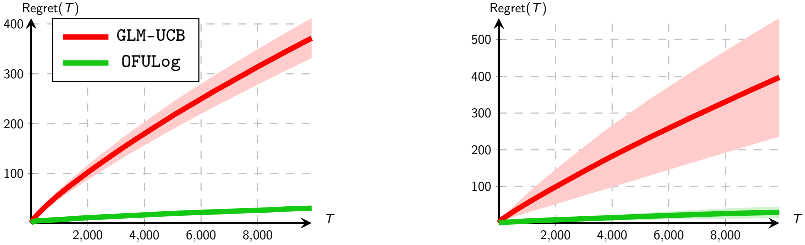
$\boxed{R^{\text{transitory}}(T) = \tilde{\mathcal{O}}(\boldsymbol{\kappa}_{\mathcal{A}})}$



$\boxed{R^{\text{transitory}}(T) = \tilde{\mathcal{O}}(1)}$

Empirical performances

- Compared with the GLM-UCB of [Filippi et al. 2010].



(a) $\boldsymbol{\kappa}_{\mathcal{A}} = 50$

(b) $\boldsymbol{\kappa}_{\mathcal{A}} = 400$

Empirical comparison of `GLM-UCB` and `OFULog` on two LogB toy experiments. The regret curves are averaged over 50 independent runs. Standard-deviation is reported in shaded colors around the averaged cumulative regret. The arm-set $\mathcal{A}$ is composed of 40 arms drawn uniformly at random in the 2-dimensional ball at the beginning of each run.

Empirical performances (ctn'd)

- Check the impact of non-linearity:

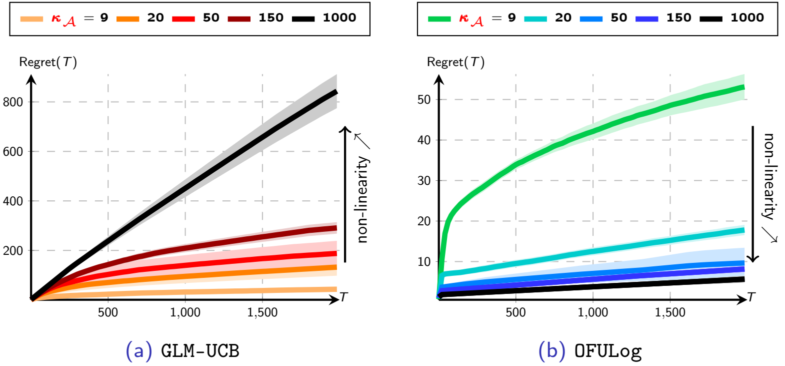


Figure: Comparing the effect of non-linearity on GLM-UCB and OFULog by varying the level of non-linearity in a Logistic Bandit setting.

Thank you!

Some references

- **Previous work (most relevant)**

  - Filippi et al. Parametric Bandits: the Generalized Linear case. *NeurIPS*, 2010.
  - Li et al. Provably Optimal Algorithms for Generalized Linear Bandits. *ICML*, 2017.
  - Dong et al. On the Performance of Thompson Sampling on Logistic Bandits. *COLT*, 2019.

- **Material for this talk was taken from**:

  - F., Abeille, Calauzènes and Fercoq. Improved Optimistic Algorithms for Logistic Bandits. *ICML*, 2020.
  - Abeille, F. and Calauzènes. Instance-Wise Minimax-Optimal Algorithms for Logistic Bandits. *AISTATS*, 2021.

- **Extension** to non-stationary settings.

  - Russac, F., Cappé, Garivier. Self-Concordant Analysis of Generalized Linear Bandits with Forgetting *AISTATS*, 2021.
  - F., Russac, Abeille and Calauzènes. Regret Bounds for Generalized Linear Bandits under Parameter Drift. *ALT*, 2021.
