

# SELF-CONCORDANT ANALYSIS OF GENERALIZED LINEAR BANDITS WITH FORGETTING

Yoan Russac<sup>\*,1,2</sup>, Louis Faury<sup>\*,3</sup>, Olivier Cappé<sup>1,2</sup>, Aurélien Garivier<sup>1,4</sup>

<sup>1</sup>Inria, CNRS, <sup>2</sup>ENS, Université PSL, <sup>3</sup>Criteo AI Lab, LTCI Télécom Paris, <sup>4</sup>UMPA, ENS Lyon, \* Equal Contribution

## Motivations

- **Non-stationary environments:** ubiquitous in real-world applications.
  - **Generalized Linear Models:** broader rewards models of considerable practical relevance (**binary, categorical**).
- **Extension of forgetting strategies** designed for linear bandits to Generalized Linear Models.

## Preliminaries

At time  $t$ , **time-dependent finite set of arbitrary actions**  $\mathcal{A}_t = \{A_{t,1}, \dots, A_{t,K_t}\}$ , where  $A_{t,k} \in \mathbb{R}^d$ . After selection of  $a_t \in \mathcal{A}_t$  observation of a reward following:

$$\mathbb{E}[r_{t+1}|a_t] = \mu(a_t^\top \theta_t^*), \quad \text{with } \mu \text{ the inverse link function,}$$

**Dynamic Regret:**

$$R_T = \sum_{t=1}^T \max_{a \in \mathcal{A}_t} \mu(a^\top \theta_t^*) - \mu(a_t^\top \theta_t^*)$$

**Maximum likelihood estimator:** Solution of the convex program:

$$\hat{\theta}_t = \arg \min_{\theta \in \mathbb{R}^d} - \sum_{s=1}^{t-1} w_{s,t} \log \mathbb{P}_\theta(r_{s+1}|a_s) + \frac{\lambda}{2} \|\theta\|_2^2 \quad (1)$$

**Forgetting policies:** if  $w_{s,t} = \gamma^{t-1-s}$  **discounted policy** and if  $w_{s,t} = \mathbb{1}(t-s \leq \tau)$  **sliding window policy**.

## Assumptions

- Bounded actions and parameters:  $\forall t \geq 1, \forall a \in \mathcal{A}_t, \|a\|_2 \leq 1, \|\theta_t^*\|_2 \leq S$ .
- Bounded rewards:  $\forall t \geq 1, 0 \leq r_t \leq m$ .
- **Non-Stationarity:**  $\theta_t^*$  can change in an arbitrary fashion **up to  $\Gamma_T$  times**.
- **Self-Concordance:**

$$|\ddot{\mu}| \leq \dot{\mu}$$

- For the inverse link function:

$$c_\mu := \inf_{\theta: \|\theta\|_2 \leq S, a: \|a\|_2 \leq 1} \dot{\mu}(a^\top \theta) > 0 \quad \triangle 1/c_\mu \text{ can be exponentially large in } S!$$

## Challenges and Approach

- 1)  $c_\mu$  limitation of the practical interest of Generalized Linear Bandits algorithms. → Reducing **dependency in the  $c_\mu$  in non-stationary environments** → Extension of a Bernstein-like inequality of [1] to **weighted self-normalized martingales**.
- 2) **MLE not necessarily bounded**, existing algorithms require a complicated projection step or a prohibitively long burn-in phase. → **Finer characterization of the MLE** using self-concordance assumption. **Algorithm relying solely on this estimator without any projection**

## Concentration Result

To solve 1), **switching from a global analysis** featuring  $V_t = \sum_{s=1}^t w_{s,t}^2 a_s a_s^\top + \lambda I_d$  **to a local analysis** through  $H_t(\theta) = \sum_{s=1}^t w_{s,t}^2 \dot{\mu}(a_s^\top \theta) a_s a_s^\top + \lambda I_d$ .

→ How to handle the weights with a local analysis?

**Theorem 1.**

Let  $\tilde{H}_t = \sum_{s=1}^{t-1} w_s^2 \dot{\mu}(a_s^\top \theta_s^*) a_s a_s^\top + \lambda_{t-1} I_d$ ,  $\epsilon_{s+1} = r_{s+1} - \mu(a_s^\top \theta_s^*)$  and  $S_t = \sum_{s=1}^{t-1} w_s \epsilon_{s+1} a_s$ , then for any  $\delta \in (0, 1)$ ,

$$P \left( \|S_t\|_{\tilde{H}_t^{-1}} \leq \mathcal{O} \left( \sqrt{d \log \left( \frac{t}{\delta} \right)} \right) \right) \geq 1 - \delta.$$

→ High probability upper-bound independent of  $c_\mu$  thanks to the local analysis!

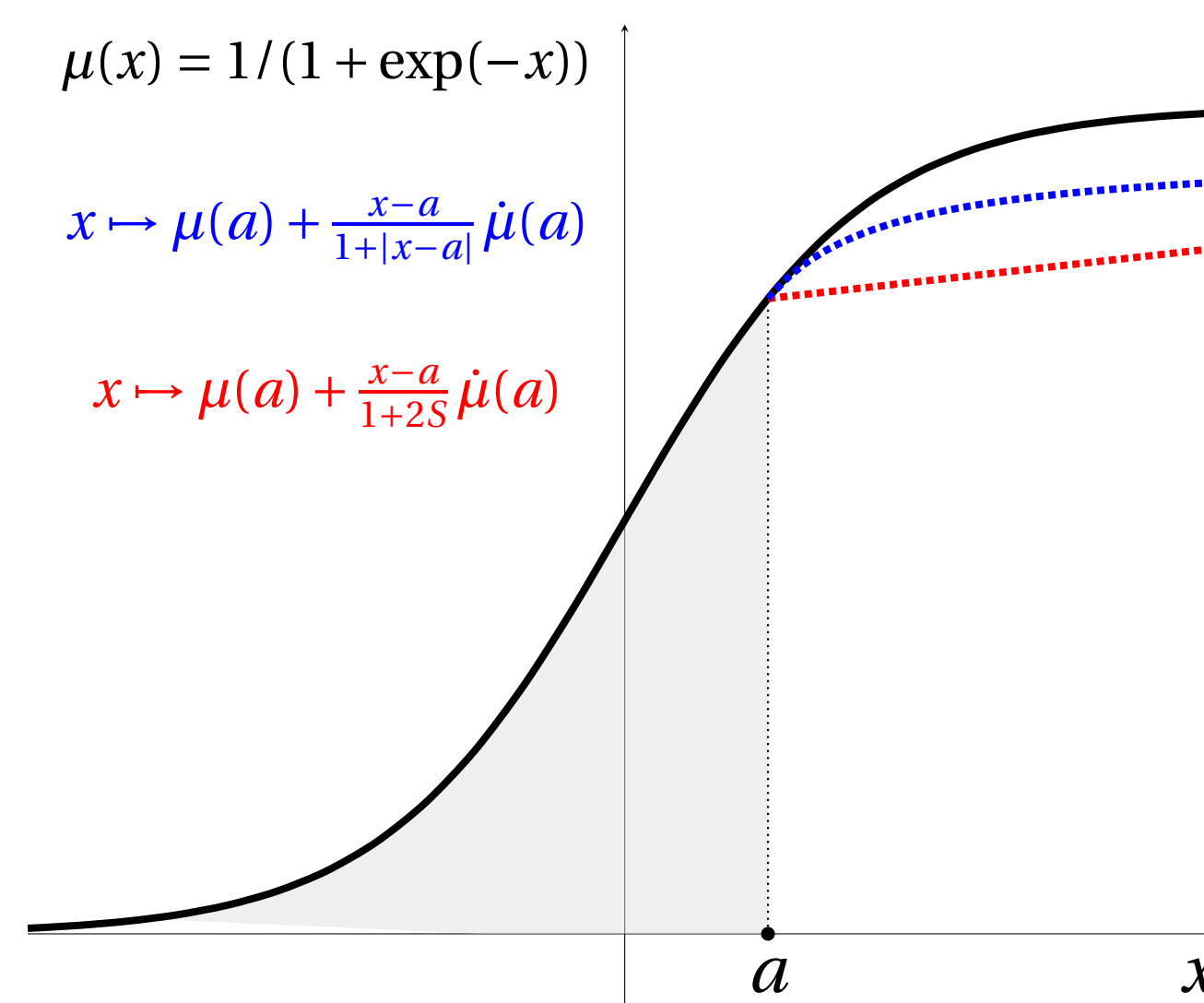
## Self-Concordance and MLE

Using a Taylor expansion and the **self-concordance assumption**, the authors in [1] uses:

$$\forall x, \quad \mu(x^\top \theta_t) \geq \mu(x^\top \theta^*) + \frac{|x^\top (\theta^* - \theta_t)|}{1 + 2S} \dot{\mu}(x^\top \theta^*)$$

Here, **tighter bound** to solve 2),

$$\forall x, \quad \mu(x^\top \hat{\theta}_t) \geq \mu(x^\top \theta^*) + \frac{|x^\top (\theta^* - \hat{\theta}_t)|}{1 + |x^\top (\theta^* - \hat{\theta}_t)|} \dot{\mu}(x^\top \theta^*)$$



## Comparison with Existing Works

Algorithm	Setting	Projection	Regret Upper Bound
GLM-UCB [2]	Stationary GLM	Non-convex	$\tilde{\mathcal{O}} \left( c_\mu^{-1} \cdot d \cdot \sqrt{T} \right)$
LogUCB1 [1]	Stationary Logistic	Non-convex	$\tilde{\mathcal{O}} \left( c_\mu^{-1/2} \cdot d \cdot \sqrt{T} \right)$
D-GLUCB [3]	Non-Stationary GLM	Non-convex	$\tilde{\mathcal{O}} \left( c_\mu^{-1} \cdot d^{2/3} \cdot \Gamma_T^{1/3} \cdot T^{2/3} \right)$
SC-D-GLUCB	Non-Stationary GLM + Gap Assumption	No projection	$\tilde{\mathcal{O}} \left( c_\mu^{-1/2} \cdot d \cdot \sqrt{\Gamma_T T} \right)$
SC-D-GLUCB	Non-Stationary GLM	No projection	$\tilde{\mathcal{O}} \left( c_\mu^{-1/3} \cdot d^{2/3} \cdot \Gamma_T^{1/3} \cdot T^{2/3} \right)$

Tab. 1: Comparison of regret guarantees for different algorithms in the GLM setting

## Regret Upper Bound

**Theorem 2.** Setting  $\gamma = 1 - (c_\mu^{1/2} \Gamma_T / (dT))^{2/3}$  and  $\lambda = d \log(T)$  leads to,

$$R_T = \mathcal{O} \left( c_\mu^{-1/3} d^{2/3} \Gamma_T^{1/3} T^{2/3} \right)$$

Adding an assumption on the gap, i.e. assuming that for all  $t$  and all suboptimal  $a \in \mathcal{A}_t$ ,  $\mu(a_t^\top \theta_t^*) - \mu(a^\top \theta^*) > \Delta$  and setting  $\gamma = 1 - \sqrt{\frac{c_\mu \Gamma_T}{d^2 T}}$  leads to,

$$R_T = \mathcal{O} \left( \Delta^{-1} c_\mu^{-1/2} d \sqrt{\Gamma_T T} \right)$$

## Algorithm

**Algorithm 1:** SC-SW-GLUCB

**Input:** Probability  $\delta$ , dimension  $d$ , regularization  $\lambda$ , upper bound for parameters  $S$ , sliding window length  $\tau$ .

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

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

Receive  $\mathcal{A}_t$ , compute  $\hat{\theta}_t$  according to Eq. (1) and  $\beta_t$  according to Eq. (??)

**Play action**  $a_t = \arg \max_{a \in \mathcal{A}_t} \mu(a^\top \hat{\theta}_t) + \frac{\beta_t^2}{\sqrt{c_\mu}} \|a\|_{V_t}^{-1}$

**Receive reward**  $r_{t+1}$

**Updating phase:**

**if**  $t < \tau$  **then**

$V_{t+1} \leftarrow a_t a_t^\top + V_t$

**else**

$V_{t+1} \leftarrow a_t a_t^\top - a_{t-\tau} a_{t-\tau}^\top + V_t$

## Experiments in Abruptly Changing Environments

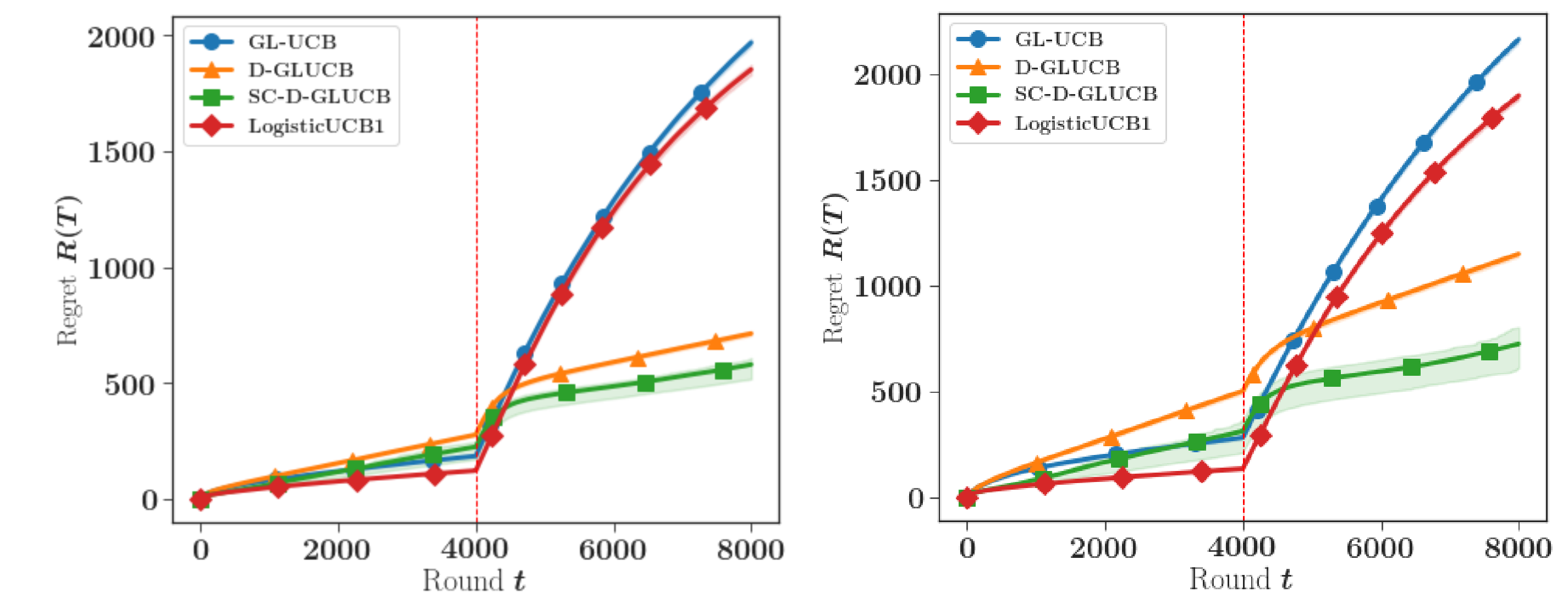


Fig. 1: Regret of the different algorithms in a 2D abruptly changing environment averaged on 200 independent experiments and the 25% associated quantiles. (left)  $c_\mu^{-1} = 400$ , (right)  $c_\mu^{-1} = 1000$

## References

- [1] L. Faury, M. Abeille, C. Calauzènes, and O. Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR, 2020.
- [2] S. Filippi, O. Cappé, A. Garivier, and C. Szepesvári. Parametric bandits: the generalized linear case. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems-Volume 1*, pages 586–594, 2010.
- [3] Y. Russac, O. Cappé, and A. Garivier. Algorithms for non-stationary generalized linear bandits. *arXiv preprint arXiv:2003.10113*, 2020.