Improved Optimistic Algorithms for Logistic Bandits

Louis Faury[1,2], Marc Abeille[1]
Clément Calauzènes[1], Olivier Fercoq[2]

[1] Criteo AI Labs

[2] LTCI Telecom Paris

Scope

Logistic Bandit.

- sequential decision making model.
- **powerful** extension to the Linear Bandit.
- **binary** reward, ubiquitous in applications of contextual bandits.

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

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

action $\in \mathcal{X}_t$

logistic function

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

Repeated game. At each round t:

1. Environment reveals $\mathcal{X}_t \in \mathbb{R}^d$ arbitrary arm-set (possibly infinite).
2. Player plays arm $x_t \in \mathcal{X}_t$
3. Player receives the reward $r(x_t)$.

Learning problem. Minimize cumulative pseudo-regret up to round T:

$$R(T) = \sum_{t=1}^{T} \Big[\underbrace{\text{argmax}_{x \in \mathcal{X}_t} \mu(\theta_\star{}^\top x)}_{\text{max reward in hindsight}} - \mu(\theta_\star{}^\top x_t) \Big]$$

Topic of this talk. We study a problem-dependent **constant** κ

- κ measures the **non-linearity** of the reward signal.
- κ can be **very large**, especially in real-life problems.

Why. Troublesome dependencies of existing algorithms

- exploration bonus $\propto \kappa$
- as a result: $\text{Regret}(T) = \tilde{\mathcal{O}}(\kappa d \sqrt{T})$.

Raise two major drawbacks

- practical: poor empirical performances.
- gap between linear and non-linear bandits.

Contributions

Novel algorithm. LogUCB2 for which we prove:

$$\text{Regret}(T) = \tilde{\mathcal{O}}(d\sqrt{T} + \boldsymbol{\kappa})$$

- reduced dependency in $\boldsymbol{\kappa}$.
- solves an **open question** since [Filippi et al. 2010].

Novel analysis with improved treatment of the reward's non-linearity.

How. Old and new:

- self-concordance property of the logistic loss.
- **new tail-inequality** for self-normalized vectorial martingales.
- **information-preserving** projections.

Optimistic algorithms

Exploration/exploitation trade-off via **optimism** (OFU).

- for **generalized linear bandits** [Filippi et al. 2010, Li et al. 2017]
- includes the logistic bandit

$$\text{play } x_t = \text{argmax}_{x \in \mathcal{X}_t} \underbrace{\mu(\hat{\theta}_t^\top x)}_{\text{exploitation}} + \underbrace{\textbf{bonus}(x)}_{\text{exploration}}$$

Exploration bonus: mitigate some defects in the prediction

- designed by upper-bounding the **prediction error**:

$$\textbf{bonus}(x) \geq \mu(\theta_\star^\top x) - \mu(\hat{\theta}_t^\top x)$$

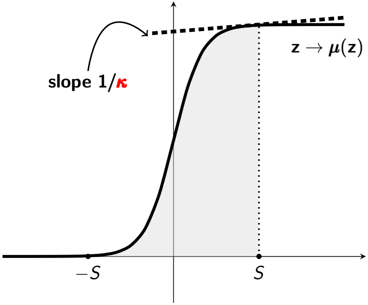
- The tighter the bonus, the better the algorithm
- For GLM-UCB [Filippi et al. 2010]:

$$\textbf{bonus}(x) \propto \boldsymbol{\kappa}$$

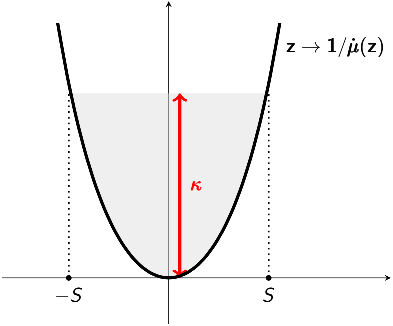
A key quantity (1/2)

Non-linear reward signal: $\boldsymbol{\kappa}$ as a distance from the Linear Bandit setting

$$\boxed{\kappa = \max_{\|x\|_2 \leq 1, \|\theta\|_2 \leq S} 1/\dot{\mu}(\theta^\top x)} \text{ when } \|\theta_\star\| \leq S.$$



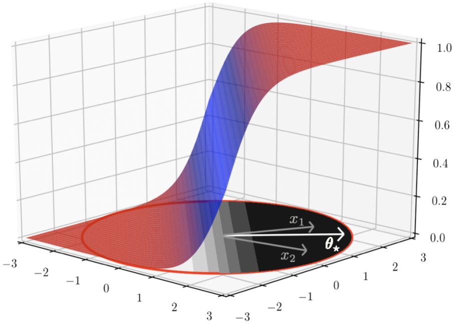
⇒ The more **non-linear** the reward, the bigger κ.



⇒ $\kappa \geq \exp(\|\theta_\star\|_2)$
exponential growth !

A key quantity (2/2)

κ characterizes the **hardness** of the **learning** problem.



- x_1 and x_2: almost always same reward ← small conditional variance.
- Typically:

$$\|\hat{\theta}_t - \theta_\star\|_2^2 \propto \kappa$$

where $\hat{\theta}_t$ is the **maximum likelihood** estimator

κ large $\Leftrightarrow$ estimating $\theta_\star$ is **hard**

GLM-UCB-like algorithms

- Bonus design: **linearization** and use of $\mathbf{V_t} = \sum_{s=1}^{t-1} x_s x_s^\top + \lambda \mathbf{I_d}$.

$$\overbrace{\mu(x^T\hat{\theta}_t) - \mu(x^T\theta_\star)}^{\text{prediction error}} \leq \boldsymbol{L} \|x\|_{\mathbf{V_t}^{-1}} \|\hat{\theta}_t - \theta_\star\|_{\mathbf{V_t}}$$

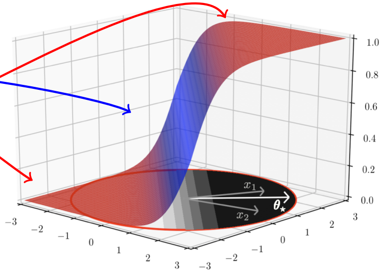
$$\Rightarrow \boxed{\text{bonus}(x) = \boldsymbol{L\kappa}\|x\|_{\mathbf{V_t}^{-1}}}$$

- Notice:

$\boldsymbol{L}$ = worst-case prediction-wise

$\boldsymbol{\kappa}$ = worst-case parameter-wise

$\Rightarrow \boldsymbol{L\kappa}$ = worst of both worlds!



Challenges

- Switch from a global (i.e $\mathbf{V_t}$) to a **local** analysis through:

$$\mathbf{H_t}(\theta) = \sum_{s=1}^{t-1} \dot{\mu}(x_s^\top \theta) x_s x_s^\top + \lambda \mathbf{I_d} \tag{1}$$

- Design a **local** bonus thanks to:

$$\boxed{\mu(x^T\hat{\theta}_t) - \mu(x^T\theta_\star) \lessapprox \dot{\mu}(x^T\hat{\theta}_t)\|x\|_{\mathbf{H_t^{-1}}(\hat{\theta}_t)}\|\hat{\theta}_t - \theta_\star\|_{\mathbf{H_t}(\hat{\theta}_t)}}$$

so **easy prediction** can cancel out **hard learning**.

- Challenges:
  - **Control** $\|\hat{\theta}_t - \theta_\star\|_{\mathbf{H_t}(\hat{\theta}_t)}$ to design a bonus **(challenge 1)**
  - **Prove** that the bonus vanishes quickly (sub-linear regret) **(challenge 2)**

both independently of $\boldsymbol{\kappa}$.

Challenge 1: a novel tail-inequality

1. Let $\{x_t\}_{t=1}^{\infty}$ a $\mathcal{F}_t$-adapted **stochastic process** in $\mathcal{B}_2(d)$
2. Let $\{\varepsilon_t\}_{t=2}^{\infty}$ a $\mathcal{F}_t$-adapted **martingale difference sequence** s.t:

$$|\varepsilon_t| \leq 1, \qquad \boldsymbol{\sigma_t^2} := \mathbb{E}[\varepsilon_{t+1}|\mathcal{F}_t] < +\infty$$

Let $\lambda > 0$ and for any $t \geq 1$ define:

$$S_t := \sum_{s=1}^{t-1} \varepsilon_{s+1} x_s \qquad \mathbf{H}_t := \sum_{s=1}^{t-1} \boldsymbol{\sigma_s^2} x_s x_s^T + \lambda \mathbf{I}_d$$

Theorem (informal)

With probability at least $1-\delta$:

$$\forall t \geq 1, \|S_t\|_{\mathbf{H}_t^{-1}} = \mathcal{O}\left(\sqrt{d \log(t/\delta)}\right)$$

Bernstein-equivalent of the tail-inequality for the Linear Bandit [Theorem 1, Abbasi-Yadkori. 2011]

Challenge 1: improved deviation-bounds

Application to the Logistic Bandit. In the logistic model:

Proposition (Deviation-bound, informal)

$$\forall t \geq 1, \quad \left\|\hat{\theta}_t - \theta_\star\right\|_{\mathbf{H}_t(\theta_\star)} \leq (1+2S)\sqrt{d\log(t)} \qquad w.h.p$$

Improvement over past results. Using the **linearization** strategy and the Linear Bandit tail-inequality:

$$\forall t \geq 1, \quad \left\|\hat{\theta}_t - \theta_\star\right\|_{\mathbf{V_t}} \leq \boldsymbol{\kappa}\sqrt{d\log(t)} \qquad \text{w.h.p}$$

⇒ from global to **local**
⇒ independent of $\boldsymbol{\kappa}$

} **challenge 1: ✔**

Challenge 2

- With these results we can design the **local** bonus:

$$\text{bonus}(x, \hat{\theta}_t) = \dot{\mu}(\hat{\theta}_t^\top x)\|x\|_{\mathbf{H_t}^{-1}(\hat{\theta}_t)}\beta_t(\delta) + \underbrace{C\boldsymbol{\kappa}\|x\|^2_{\mathbf{V_t}^{-1}}}_{\text{second order term}}$$

with $\beta_t \sim \sqrt{d \log(t)}$ and play:

$$\boxed{x_t = \text{argmax}_{x \in \mathcal{X}_t} \left[\mu(x^\top \hat{\theta}_t) + \text{bonus}(x, \hat{\theta}_t)\right]}$$

- To finish the analysis, we need to bound:

$$\sum_{t=1}^{T} \text{bonus}(x_t, \hat{\theta}_t) \leq \beta_T(\delta) \underbrace{\sum_{t=1}^{T} \dot{\mu}(\hat{\theta}_t^\top x_t)\|x\|_{\mathbf{H_t}^{-1}(\hat{\theta}_t)}}_{\text{leading regret term}} + C\boldsymbol{\kappa} \underbrace{\sum_{t=1}^{T} \|x_t\|^2_{\mathbf{V_t}^{-1}}}_{\log(T)}$$

$$\overset{?}{\leq} \sqrt{T} \qquad \Leftarrow \textbf{the bonus vanishes}$$

Challenge 2: admissible log-odds

Decreasing bonus $\Leftrightarrow$ increasing **information/knowledge**.

Why it is not obvious.

- How is information measured? At round t:
  - In MAB, for arm x:

$$\#\{x_t = x,\, s \leq t\}$$

  - In Linear Bandit:

$$\|x\|_{\mathbf{V}_t}$$

increasing

  - In Logistic Bandits, for arm x:

$$\|x\|_{\mathbf{H}_t(\hat{\theta}_t)} \quad \text{??}$$

$$\left(\mathbf{H}_t(\hat{\theta}_t) = \sum_{s=1}^{t-1} \dot{\mu}(x_s^\top \hat{\theta}_t) x_s x_s^\top + \lambda \mathbf{I_d}\right)$$

What it means.

- Updating $\hat{\theta}_t$ can **degrade** past information
- $\Rightarrow$ no reason the bonus should vanish!

Challenge 2: admissible log-odds (ctn'd)

Solution (informal).

- **Project** $\hat{\theta}_t$ to a set of information-preserving estimators.
- Set of **admissible log-odds**:

$$\mathcal{W}_t := \left\{\theta,\ \dot{\mu}(x_s^\top \theta) \geq \dot{\mu}(x_s^\top \hat{\theta}_s) \text{ for all } s \geq t-1\right\}$$

- Notice:

$$\begin{aligned}\hat{\theta}_t \in \mathcal{W}_t &\Rightarrow \dot{\mu}(x_s^\top \hat{\theta}_t) \geq \dot{\mu}(x_s^\top \hat{\theta}_s)\\ &\Rightarrow \mathbf{H_t}(\hat{\theta}_t) \succeq \sum_{s=1}^{t-1} \dot{\mu}(x_s^\top \hat{\theta}_s) x_s x_s^\top + \lambda \mathbf{I_d} := \mathbf{L_t}\\ &\Rightarrow \|x\|_{\mathbf{H_t}(\hat{\theta}_t)} \geq \|x\|_{\mathbf{L_t}} \quad \leftarrow \textbf{increasing!}\end{aligned}$$

- We can prove:

$$\boxed{\hat{\theta}_t \in \mathcal{W}_t \Rightarrow \sum_{t=1}^{T} \dot{\mu}(\hat{\theta}_t^\top x_t) \|x\|_{\mathbf{H_t}^{-1}(\hat{\theta}_t)} \leq d\sqrt{T} + C\boldsymbol{\kappa} \log T}$$

challenge 2: ✔

LogUCB-2 (wrap-up)

Algorithm 1 Log-UCB2

Input: regularization parameter λ
Initialize the set of admissible log-odds $\mathcal{W}_0 = \Theta$
for $t \geq 1$ **do**
$\tilde{\theta}_t = \text{argmin}_{\theta \in \mathcal{W}_t \cap \Theta} \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{H}_t^{-1}(\theta)}$ **← project $\hat{\theta}_t$ on $\mathcal{W}_t$**
Observe the contexts-action feature set $\mathcal{X}_t$.
Play $x_t = \text{argmax}_{x \in \mathcal{X}_t} \mu(x^\top \tilde{\theta}_t) + b_t(x)$.
Observe rewards r_{t+1}.
Compute log-odds $\ell_t = \sup_{\theta' \in \mathcal{C}_t(\delta)} x_t^\top \theta'$. **← minimum information**
Add the new constraint to the feasible set:

$$\mathcal{W}_{t+1} = \mathcal{W}_t \cap \{\theta : -\ell_t \leq \theta^\top x_t \leq \ell_t\}.$$

end for

LogUCB-2 (wrap-up)

Algorithm	Regret Upper Bound	Setting
GLM-UCB [Filippi et al. 2010]	$\mathcal{O}\left(\boldsymbol{\kappa} \cdot d \cdot T^{1/2} \cdot \log(T)^{3/2}\right)$	GLM
Thompson Sampling [Abeille et Lazaric. 2017]	$\mathcal{O}\left(\boldsymbol{\kappa} \cdot d^{3/2} \cdot T^{1/2} \log(T)\right)$	GLM
SupCB-GLM[1] [Li et al. 2017]	$\mathcal{O}\left(\boldsymbol{\kappa} \cdot (d \log K)^{1/2} \cdot T^{1/2} \log(T)\right)$	GLM, K actions
LogUCB1 **(this work)**	$\mathcal{O}\left(\boldsymbol{\kappa}^{\mathbf{1/2}} \cdot d \cdot T^{1/2} \log(T)\right)$	Logistic model
LogUCB2 **(this work)**	$\mathcal{O}\left(d \cdot T^{1/2} \log(T) + \boldsymbol{\kappa} \cdot d^2 \cdot \log(T)^2\right)$	Logistic model

Comparison of frequentist regret guarantees for the logistic bandit with respect to $\boldsymbol{\kappa}$, d and T.

Take-home messages

Critical dependence on $\boldsymbol{\kappa}$.

- Linearization strategies $\Rightarrow$ **prohibitive** practical performance

Tackled through a local analysis.

- new tail-inequality for self-normalized martingales
- self-concordance of log-loss

and information-preserving estimators.

- set of admissible log-odds.

Closing the gap with linear bandits

- $R_T = \tilde{\mathcal{O}}\left(d\sqrt{T} + \boldsymbol{\kappa}\right)$

Thank you!