Jointly Efficient and Optimal Algorithms for Logistic Bandits

Louis Faury[1], Marc Abeille[1], Kwang-Sung Jun[2], Clément Calauzènes[1]

[1]Criteo AI Lab, [2]University of Arizona

Motivation

Logistic Bandits. *Structured* and *binary* bandit games;

- neat study of non-linearity in parametric bandits,
- highly relevant in practice (sequential decision-making under binary feedback).

Context.

- fruitful stream of research on the impact of *non-linearity* in Logistic Bandits,
- led to the development of *statistically* efficient (minimax-optimal) algorithms,
- however (prohibitively) computationally *inefficient*.

$\rightsquigarrow$ although well understood from a learning-theoretic standpoint, we are still missing fast algorithms for Logistic Bandits.

Can we achieve computational efficiency without sacrificing statistical tightness?

Logistic Bandits

Reward model

- $\mathcal{A} \subset \mathbb{R}^d$ is the arm set,
- $r_{t+1} = r(a_t) \in \{0,1\}$ is the associated reward,
- $\theta_\star \in \mathbb{R}^d$ *unknown* parameter.

[Binary reward]

$r(a) \sim \texttt{Bernoulli}(\mu(a^\mathsf{T}\theta_\star))$

[Non-linear link function]

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

Learning problem

At each step $t \leq T$:

- choose arm $a_t \in \mathcal{A}$,
- receive reward r_{t+1}.

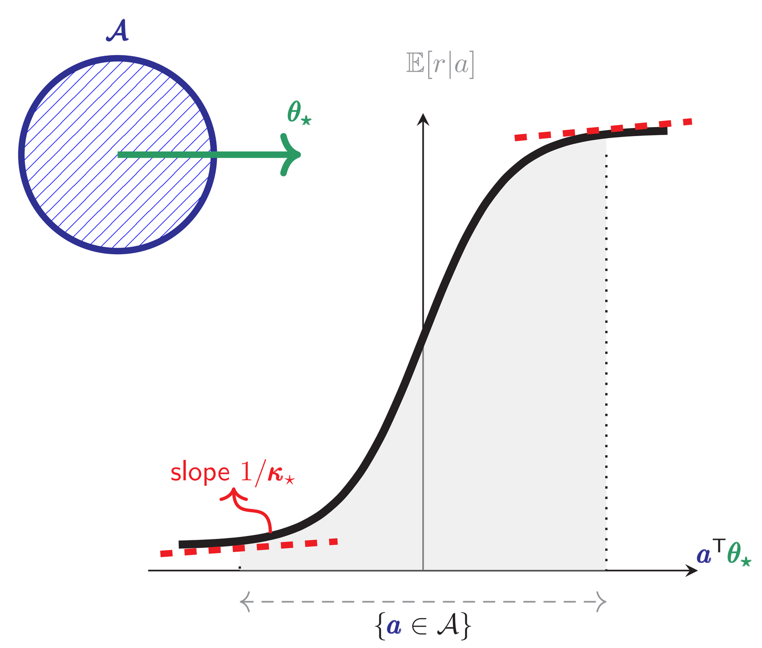
Objective is to minimize regret: $\text{Regret}(T) = \sum_{t=1}^T \left[\max_{a\in\mathcal{A}} \mu(a^\mathsf{T}\theta_\star) - \mu(a_t^\mathsf{T}\theta_\star)\right]$.

Level of non-linearity

Measured by a *problem-dependent* constant:

$$\kappa_\star := 1/\dot{\mu}(\max_{a\in\mathcal{A}} a^\mathsf{T}\theta_\star)$$

- (inverse of) minimal variance,
- "distance to linearity" over the decision set,
- typically large as $\kappa_\star \approx \exp(\|\theta_\star\|)$.



Minimax-Optimality

Minimax-Optimal algorithms w.r.t T, d and $\kappa_\star$:

[Abeille et al., AISTATS'21] $\text{Regret}(T) \lesssim d\sqrt{T/\kappa_\star}\ (+\kappa_\star)$

$\rightsquigarrow$ exponential improvement over previous approaches.

$\rightsquigarrow$ the more non-linear the problem, the easier!

$\rightsquigarrow$ matching lower-bound.

Optimistic algorithm based on confidence region $\mathcal{C}_t(\delta)$; play $a_t = \arg\max_{\mathcal{A}} \max_{\theta\in\mathcal{C}_t(\delta)} a^\mathsf{T}\theta$.

Main teachings from those optimal algorithms:

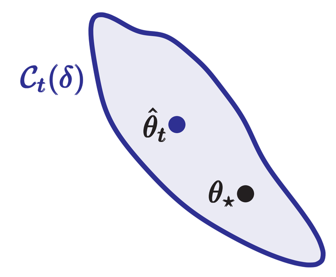
- the learning procedure must leverage the *local* behavior of non-linearity,
- the learner might suffer an inevitably large regret during an initial *transitory* regime,
- in the *permanent* regime, the local behavior around $\theta_\star$ dominates.

Computational Efficiency

Learning. [Abeille et al., 2021] use *variance-sensitive* confidence sets:

$$\mathcal{C}_t(\delta) := \left\{\theta,\ \left\|\theta - \hat{\theta}_t\right\|^2_{\mathbf{H}_t(\theta)} \lesssim d\log(t)\right\}$$

$$\begin{cases} \hat{\theta}_t := \arg\min_\theta \sum_{s=1}^t \ell_s(\theta) \text{ the MLE estimator,} \\ \mathbf{H}_t(\theta) := \sum_{s=1}^t \dot{\mu}(a_s^\mathsf{T}\theta)a_s a_s^\mathsf{T} + \lambda\mathbf{I}_d\,. \end{cases}$$



Expensive to maintain;

- testing membership $\theta \in \mathcal{C}_t(\delta)$ costs $\Omega(t)$ operations!
- batch computation of the MLE,
- non-ellipsoidal confidence sets impacts the complexity of the *planning* mechanism.

Goal. New confidence sets for Logistic Bandits such that:

- *sufficient statistics* require $\mathcal{O}(1)$ operations,
- without sacrificing tightness.

ECOLog Procedure

Iterative Procedure. Given some convex set Θ, compute estimators $\{\theta_t\}_t$ of $\theta_\star$ through:

ECOLog

1. $\theta_t \leftarrow \arg\min_{\theta\in\Theta} \left\|\theta - \theta_{t-1}\right\|^2_{\mathbf{W}_{t-1}} + \eta\ell_t(\theta)$
2. $\mathbf{W}_t \leftarrow \mathbf{W}_{t-1} + \dot{\mu}(a_t^\mathsf{T}\theta_t)a_t a_t^\mathsf{T}$

where $\eta = \text{diam}(\Theta)$ is the *learning-rate* and $\ell_t(\cdot)$ is the immediate log-loss.

Main Idea. Inspired from Online Convex Optimization literature [Jézéquel et al., 2020];

- based on *local* quadratic lower-bounds for the logistic-loss:

$$\ell_t(\theta_\star) \geq \ell_t(\theta) + \nabla\ell_t(\theta)^\mathsf{T}(\theta_\star - \theta) + \dot{\mu}(a_t^\mathsf{T}\theta)(a_t^\mathsf{T}(\theta_\star - \theta))^2\,.$$

- decompose the total log-loss $\mathcal{L}_t(\theta) := \sum_{s=1}^t \ell_s(\theta_t)$ as:

$$\mathcal{L}_t(\theta) = \tilde{\mathcal{L}}_{t-1}(\theta) + l_t(\theta) = \eta^{-1}(\theta - \theta_{t-1})^\mathsf{T}\mathbf{W}_{t-1}{}^\mathsf{T}(\theta - \theta_{t-1}) + l_t(\theta)\,.$$

- minimize this *strongly convex* proxy to obtain θ_t at $\tilde{\mathcal{O}}(d^2)$ cost.

Confidence Regions. Emulates the confidence set $\mathcal{C}_t(\delta)$ through $\mathbf{W}_t$;

Confidence Set

Let $\delta \in (0,1]$. If $\theta_\star \in \Theta$ then:

$$\mathcal{E}_t(\delta) = \left\{\left\|\theta - \theta_t\right\|^2_{\mathbf{W}_t} \lesssim \exp(\eta)d\log(t/\delta)\right\}$$

is a confidence region for $\theta_\star$; $\mathbb{P}(\theta_\star \in \mathcal{E}_t(\delta)) \geq 1-\delta$.

$\rightsquigarrow$ *local* behavior captured through Θ, an *admissible* parameter-set.

Forced-Exploration phase allows to identify a *small* admissible set Θ;

- such that $\eta = \text{diam}(\theta) \approx 1$,
- for at most $\kappa_\star$ rounds (*e.g.* second-order regret term),
- hardcodes the transitory phase of [Abeille et al., 2021].

Algorithm and Regret Bound

Algorithm. Combines forced-exploration and ECOLog;

OFU-ECOLog

- Perform $\tau = \kappa_\star$ rounds of forced exploration to obtain Θ such that:
  $\theta_\star \in \Theta$ with high proba. and $\text{diam}(\Theta) \leq 1$.
- For $t \geq \tau$:
  1. play the optimistic arm $a_t = \arg\max_{a\in\mathcal{A}} \max_{\theta\in\mathcal{E}_t(\delta)} a^\mathsf{T}\theta$,
  2. receive r_{t+1}, compute $(\theta_{t+1}, \mathbf{W}_{t+1})$ by running ECOLog.

Computational Cost. At each round at most $\mathcal{O}(d^2|\mathcal{A}|\log(t))$ operations, since:

- ECOLog can be solved to arbitrary precision ε at cost $d\log(1/\varepsilon)$,
- the confidence region is *ellipsoidal*; closed-form for the optimistic arm (exploration bonus).

Regret Bounds. Minimax-optimal rates;

Regret

The regret of OFU-ECOLog satisfies with high probability:

$$\text{Regret}(T) \lesssim d\sqrt{T/\kappa_\star} + \kappa_\star\,.$$

Adaptive Version. ada-OFU-ECOLOG dilutes the warm-up throughout the learning.

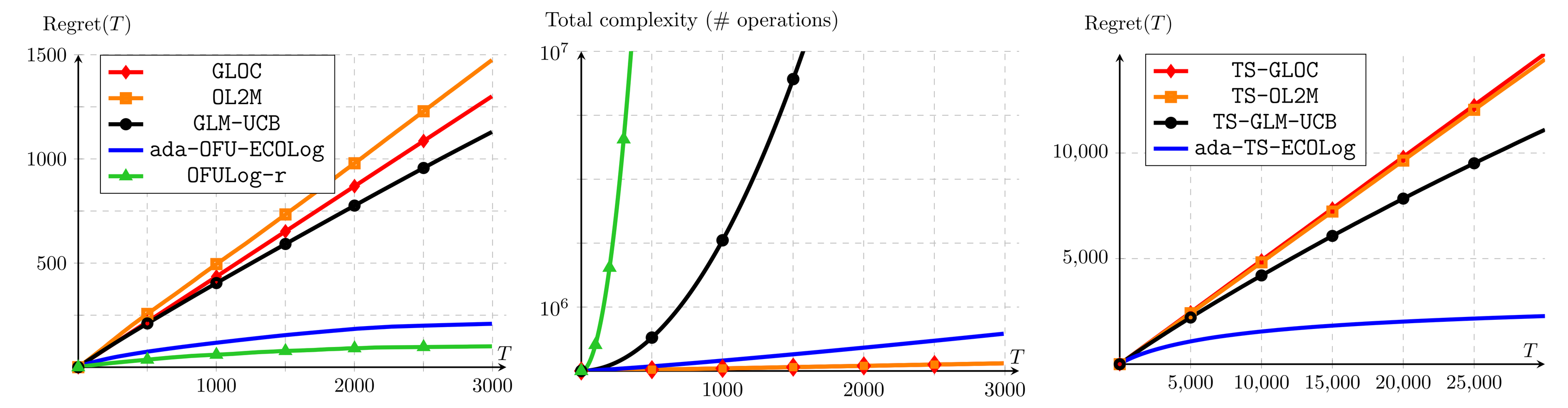
- based on a *data-dependent* width for the confidence regions,
- ✔ preserves statistical efficiency while ultimately removing the need for forced-exploration,
- ✔ coherent with [Abeille et al. 2021]: low-order $\kappa_\star$ dependencies can sometimes be avoided.

Conclusion

Joint *statistical* and *computational* efficiency;

Algorithm	Regret Bound	Cost Per-Round	Minimax	Efficient
GLM-UCB [Filippi et al. 2010]	$\tilde{\mathcal{O}}\left(\kappa_\star d\sqrt{T}\right)$	$\mathcal{O}\left(d^2\|\mathcal{A}\|T\right)$	✘	✘
GLOC, OL2M [Jun et al. 2017] [Zhang et al. 2016]	$\tilde{\mathcal{O}}\left(\kappa_\star d\sqrt{T}\right)$	$\mathcal{O}\left(d^2\|\mathcal{A}\|\right)$	✘	✔
OFULog-r [Abeille et al. 2021]	$\tilde{\Theta}\left(d\sqrt{T/\kappa_\star}\right)$	$\mathcal{O}\left(d^2\|\mathcal{A}\|T\right)$	✔	✘
(ada-)OFU-ECOLog (this paper)	$\tilde{\Theta}\left(d\sqrt{T/\kappa_\star}\right)$	$\tilde{\mathcal{O}}\left(d^2\|\mathcal{A}\|\right)$	✔	✔

Numerical simulations corroborates theoretical results;



$d=2,\ |\mathcal{A}|=20,\ \kappa=400$ $d=2,\ |\mathcal{A}|=20,\ \kappa=400$ $d=5,\ \mathcal{A}=\mathcal{B}_5,\ \kappa=400$.

References

S. Filippi, O. Cappé, A. Garivier and C. Szepesvári. Parametric Bandits: The Generalized Linear Case. *NeurIPS*, 2010.

L. Faury, M. Abeille, C. Calauzènes and O. Fercoq. Improved Optimistic Algorithms for Logistic Bandits. *ICML*, 2020.

M.Abeille, L.Faury and C.Calauzènes. Instance-Wise Minimax-Optimal Algorithms for Logistic Bandits. *AISTATS*, 2021.