Instance-Wise Minimax-Optimal Algorithms for Logistic Bandits

Marc Abeille[1], Louis Faury[1,2], Clément Calauzènes[1]
[1]Criteo AI Lab, [2]LTCI Télécom Paris

Motivation

Toward non-linear reward model.

- Parametric bandit results mostly concern the linear setting,
- Non-linearity often arises in real-world application,
- Impact of non-linearity on the exploration-exploitation tradeoff is poorly understood.

The logistic bandit setting.

- Non-linear reward signal,
- Compact and minimal setting,
- Widely used for practical applications.

We characterize the impact of non-linearity for Logistic Bandit:

→ First problem-dependent lower-bound,

→ Minimax-optimal algorithm.

The Logistic Bandit problem

The reward model.

- $\mathcal{X} \subset \mathbb{R}^d$ is the arm set,
- $r(x) \in \{0, 1\}$ is the reward associated with arm $x \in \mathcal{X}$,
- $\theta_\star \in \mathbb{R}^d$ *unknown* parameter.

[Binary reward]
$r(x) \sim \texttt{Bernoulli}\left(\mu(x^\top \theta_\star)\right)$

[Non-linear link function]
$\mu(z) = \left(1 + \exp(-z)\right)^{-1}$

The learning problem.

At each step $t \leq T$:

- Choose a arm $x_t \in \mathcal{X}$,
- Receive $r(x_t)$,

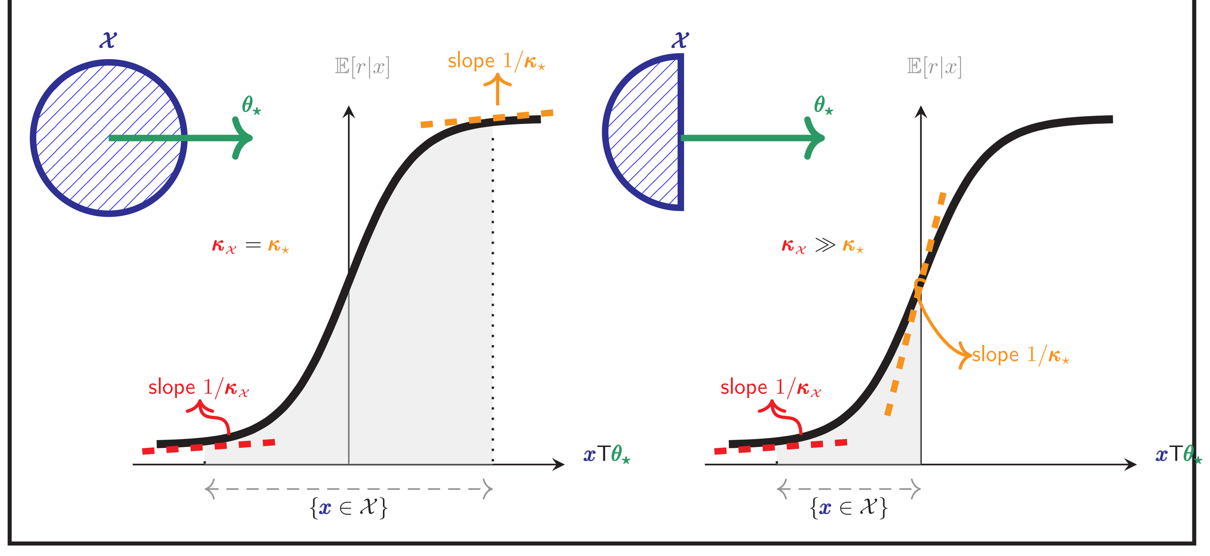
Objective: Minimize Regret

$$R_{\theta_\star}(T) = \sum_{t=1}^{T} \left[\max_{x \in \mathcal{X}} \mu(x^\top \theta_\star) - \mu(x_t^\top \theta_\star)\right] .$$

Quantifying non-linearity We consider two important *problem-dependent* constants:

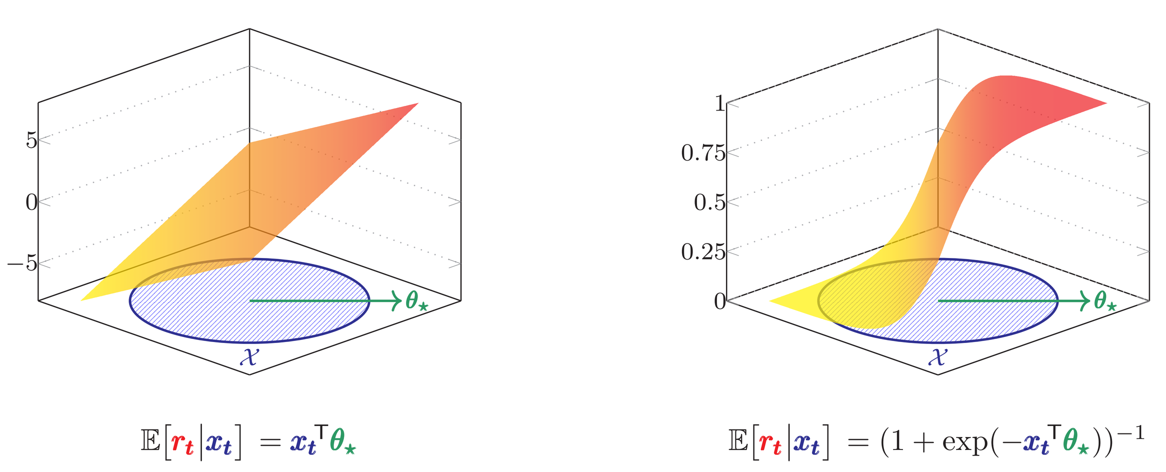
$$\kappa_\star := 1/\dot{\mu}(\max_{x \in \mathcal{X}} x^\top \theta_\star)$$
$$\kappa_{\mathcal{X}} := 1/\min_{x \in \mathcal{X}} \dot{\mu}(x^\top \theta_\star)$$

- $\kappa_\star$: "distance to linearity" around the optimal action,
- $\kappa_{\mathcal{X}}$: worst-case "distance to linearity" over the decision set.



Non-linearity: blessing or curse ?

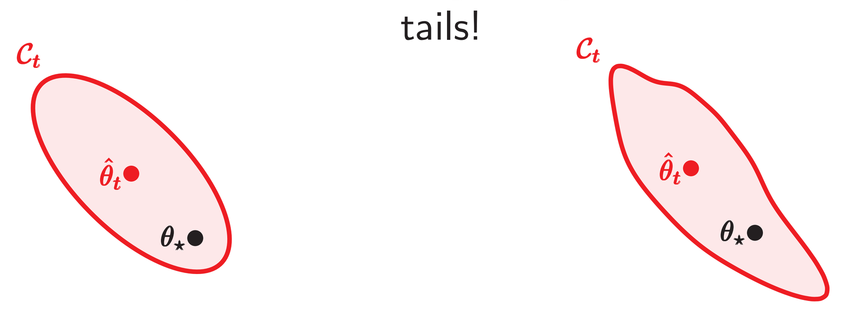
From LB to LogB



Impact on the learning.

Different richness of information associated with sampling an arm:

LB Same everywhere, **LogB** High in the center, low in the tails!



✓ Despite non-linearity → available conf. set. $\mathcal{C}_t$ for **LogB**

[Faury et al, Improved Optimistic Algorithms for Logistic Bandits, ICML'20]

✗ Some regions are *harder* to learn that other → the conf. set. $\mathcal{C}_t$ is *not* an ellipsoid!

Impact on the predicted performance

✓ **LogB** deviation in parameters → little to no deviation in performance *in the tails*

$$\|\theta - \theta_\star\| = \delta \quad \Rightarrow \quad \mu(x^\top\theta) \simeq \mu(x^\top\theta_\star)$$

Open question: does *easy* prediction cancel out *hard* learning?

Related Work and Contributions

Related work.

[Filippi et al., NIPS'10] $R_{\theta_\star}(T) \;=\!\lesssim\; \kappa_{\mathcal{X}} d\sqrt{T}$

[Faury et al., ICML'20] $R_{\theta_\star}(T) \lesssim d\sqrt{T} + \kappa_{\mathcal{X}}$

[Dong et al., COLT'19] In the worst case, $R_{\theta_\star}(T)$ must increase with $\kappa_{\mathcal{X}}$

Contributions.

Theorem 1. (Regret Upper Bound) The regret of **OFU-Log** satisfies with high-probability:

$$R(T) \lesssim d\sqrt{\frac{T}{\kappa_\star}} + (\kappa_{\mathcal{X}}).$$

Theorem 2. (Local Lower Bound) Let $\mathcal{X} = \mathcal{S}_d(0, 1)$, for any $\theta_\star$ and T large enough, it exists $\epsilon > 0$ small enough s.t.

$$\min_\pi \max_{\|\theta - \theta_\star\| \leq \epsilon} \mathbb{E}\left[R_\theta^\pi(T)\right] = \Omega\left(d\sqrt{\frac{T}{\kappa_\star}}\right).$$

Optimistic Algorithm OFULog

```
for t = {0, ..., T} do
  Set λ_t ← d log(t).
  (Learning) Solve θ̂_t = arg min_θ L_t(θ).
  (Planning) Solve (x_t, θ_t) ∈ arg max_{X, C_t(δ)} μ(x^T θ).
  Play x_t and observe reward r_{t+1}.
end for
```

where $\mathcal{L}_t(\theta)$ and $\mathcal{C}_t(\delta)$ are the log-likelihood function and confidence set associated with the learning problem.

Ideas Behind the Lower Bound

Objective and approach

- we shoot for a *problem-dependent* lower-bound
- standard approach consider worst-case over *all possible instance*
- inspired by *[Simchowitz et al., ICML'20]* → *local* lower-bound
- consider worst-case over all nearby alternative around a given *problem instance*.

ideas

- we consider a given instance parametrized by $\theta_\star$,
- let π denote a policy that outputs a sequence of arms, and $R^\pi_{\theta_\star}(T)$ the induced expected regret.

Small regret ↔ low exploration

$$R^\pi_{\theta_\star}(T) \propto 1/\kappa_\star \sum_{t=1}^{T} \|x_t - x_\star(\theta_\star)\|^2, \quad x_\star(\theta_\star) = \arg\max_{x \in \mathcal{X}} \mu(x^\top\theta_\star)$$

- $R^\pi_{\theta_\star}(T)$ small ↔ $x_t \simeq x_\star(\theta_\star)$,
- directions orthogonal to $x_\star(\theta_\star)$ are poorly explored!
- *Larger* $\kappa_\star$ → *smaller* impact when deviating from $x_\star(\theta_\star)$!

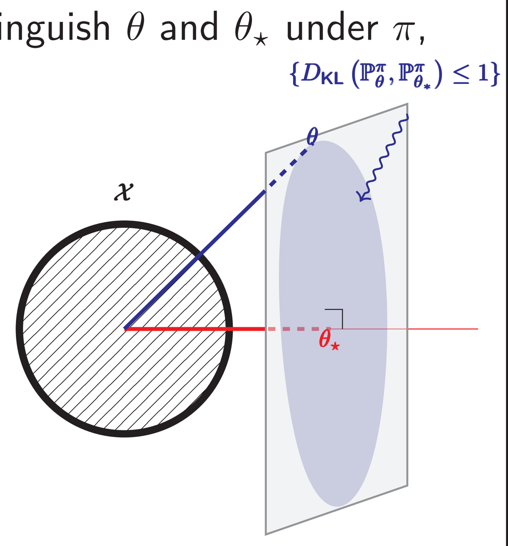
Low exploration ↔ large set of plausible alternative

- We quantify the *similarity* between instances θ, $\theta_\star$ under policy π by the *discrepancy*

$$D_{\text{KL}}\left(\mathbb{P}^\pi_\theta, \mathbb{P}^\pi_{\theta_\star}\right)$$

large $D_{\text{KL}}\left(\mathbb{P}^\pi_\theta, \mathbb{P}^\pi_{\theta_\star}\right)$ → *easy* to distinguish θ and $\theta_\star$ under π,
small $D_{\text{KL}}\left(\mathbb{P}^\pi_\theta, \mathbb{P}^\pi_{\theta_\star}\right)$ → *hard* to distinguish θ and $\theta_\star$ under π,

$$D_{\text{KL}}\left(\mathbb{P}^\pi_\theta, \mathbb{P}^\pi_{\theta_\star}\right) \propto \sqrt{\frac{T}{\kappa_\star}}\|\theta - \theta_\star\|^2$$



- *large* $\kappa_\star$ degrades the richness of acquired information,

→ $D_{\text{KL}}\left(\mathbb{P}^\pi_\theta, \mathbb{P}^\pi_{\theta_\star}\right)$ decreases with $\kappa_\star$,

Tension and trade-off

- Policy π cannot perform well on two *distinct* instances,
- but may not yield *similar* information.

Trade-off

- Let π perform well for $\theta_\star$,
- consider an alternative instance θ such that $\|\theta - \theta_\star\|^2 \approx \sqrt{\frac{\kappa_\star}{T}}$,
- the regret of π for the instance θ must be large:

$$R^\pi_\theta(T) \approx 1/\kappa_\star \sum_{t=1}^{T} \|x_t - x_\star(\theta)\|^2 \approx 1/\kappa_\star \sum_{t=1}^{T} \|x_\star(\theta_\star) - x_\star(\theta)\|^2$$
$$\approx T\|\theta_\star - \theta\|^2/\kappa_\star \approx \sqrt{T/\kappa_\star}.$$

Ideas Behind the Upper Bound

Permanent and transitory regimes

- Regret decomposition:

$$R_{\theta_\star}(T) = \underbrace{R^{\text{perm}}(T)}_{\tilde{\mathcal{O}}(\sqrt{T})} + \underbrace{R^{\text{trans}}(T)}_{\tilde{\mathcal{O}}(1)}$$

Permanent regime: intuition.

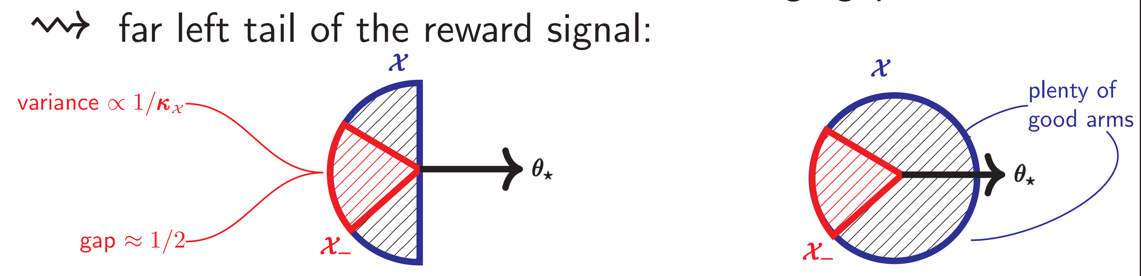
- Sublinear regret ⇒ play mostly around the best arm $x_\star$.
  ⇝ Almost a linear bandit with slope $1/\kappa_\star$.
- A finer analysis is coherent with this conceptual argument:

$$R^{\text{perm}}(T) \leq d\sqrt{\sum_{t=1}^{T} \dot{\mu}(x_t^\top\theta_\star)} \approx d\sqrt{T/\kappa_\star}.$$

- Formal proof: thanks to self-concordance property.

Transitory regime and detrimental arms.

- *Detrimental arm* $\mathcal{X}_-$: low-information and large gap
  ⇝ far left tail of the reward signal:



- Transitory regime: how long before discarding detrimental arms:

$$R^{\text{trans}}{}_{\theta_\star}(T) \leq \min\left(\kappa_{\mathcal{X}}, \sum_{t=1}^{T} \mathbb{1}(x_t \in \mathcal{X}_-)\right)$$

- Fast if the proportion of detrimental arm is small:

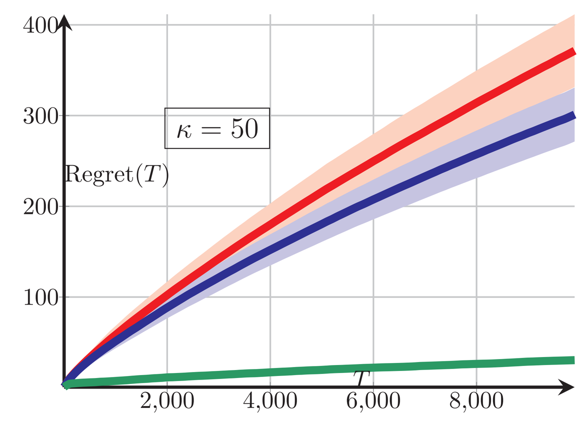
Proposition 1. (Transitory regret) With h.p :

$$R^{\text{trans}}{}_{\theta_\star}(T) \lesssim_T d^2 + dK \qquad \text{if } |\mathcal{X}_-| \leq K\,,$$
$$R^{\text{trans}}{}_{\theta_\star}(T) \lesssim_T d^3 \qquad \text{if } \mathcal{X} = \mathcal{B}_d(0, 1)\,.$$

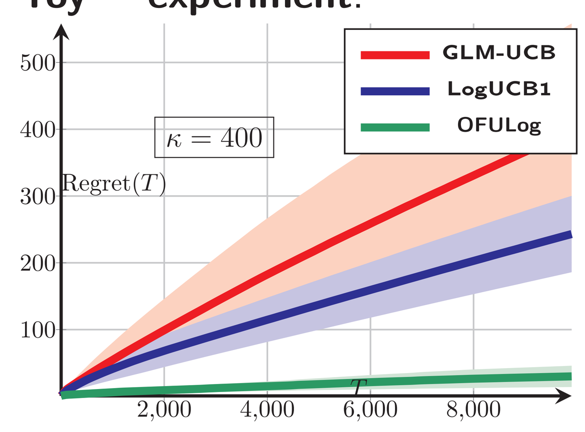
⇝ independent of $\kappa_{\mathcal{X}}$ for reasonable configurations.

Algorithmic design and experiments

- **Convex relaxation.** bla



- **Toy experiment.**



Conclusion

- Blah
- Blah
- Blah
- Blah

References

Y. Abbasi-Yadkori, Cs. Szepesvári. Regret Bounds for the Adaptive Control of Linear Quadratic Systems. In *Proceedings of COLT*, 2011.

S. Bittanti and M.C. Campi. Adaptive control of linear time invariant systems: the "bet on the best" principle. *Commun. Inf. Syst. Volume 6*, 2006.

M.K.S. Faradonbeh, A. Tewari, and G. Michailidis. Finite Time Analysis of Optimal Adaptive Policies for Linear-Quadratic Systems. *arXiv:1711.07230*.

Y. Ouyang, M. Gagrani, and R. Jain. Learning-based Control of Unknown Linear Systems with Thompson Sampling. *arXiv:1709.04047*.

I. Osband, and B. Van Roy. Model-based Reinforcement Learning and the Eluder Dimension *NIPS*, 2014.

M. Abeille, and A. Lazaric. Thompson Sampling for Linear-Quadratic Control Problems. *Proceedings of AISTATS*, 2017.