Improved Regret Bounds of (Multinomial) Logistic Bandits via Regret-to-Confidence-Set Conversion

Junghyun Lee[1], Se-Young Yun[1], and Kwang-Sung Jun[2]

[1] Kim Jaechul Graduate School of AI, KAIST, [2] Department of Computer Science, University of Arizona

{jh_lee00, yunseyoung}@kaist.ac.kr, kjun@cs.arizona.edu

Contributions

- We propose **regret-to-confidence-set conversion (R2CS)**, a new framework for converting *achievable* online learning regret bound to a confidence sequence, *without ever running the algorithm*!
- We apply R2CS to obtain the *tightest confidence set* for (multinomial) logistic losses, leading to the state-of-the-art regret guarantees for (multinomial) logistic bandits!
- Our confidence set is also numerically tight, leading to the best numerical regret by a large margin.

Logistic Bandits

Problem Setting

For $t \in [T]$:

1. The learner observes a potentially infinite (contextual) arm-set $\mathcal{X}_t \subset \mathbb{R}^d$
2. The learner chooses $\boldsymbol{x}_t \in \mathcal{X}_t$ according to some policy
3. Receive a *binary* reward $r_t | \boldsymbol{x}_t \sim \mathrm{Ber}(\mu(\langle \boldsymbol{x}_t, \boldsymbol{\theta}_\star \rangle))$,
   - $\boldsymbol{\theta}_\star \in \mathbb{R}^d$ is unknown
   - $\mu(z) = (1 + e^{-z})^{-1}$ is the logistic function

Goal. Minimize:

$$\mathrm{Reg}^B(T) := \sum_{t=1}^{T} \left\{ \mu(\langle \boldsymbol{x}_{t,\star}, \boldsymbol{\theta}_\star \rangle) - \mu(\langle \boldsymbol{x}_t, \boldsymbol{\theta}_\star \rangle) \right\},$$

where $\boldsymbol{x}_{t,\star} := \arg\max_{\boldsymbol{x} \in \mathcal{X}_t} \mu(\langle \boldsymbol{x}, \boldsymbol{\theta}_\star \rangle)$.

Applications. Discrete-valued rewards in interactive machine learning (e.g., clicks in news recommendations; Li et al. [2010])

Standard assumptions [Abeille et al., 2021]:

→ Assumption 1. $\mathcal{X}_t \subseteq \mathcal{B}^d(1)$ for all $t \geq 1$.

→ Assumption 2. $\boldsymbol{\theta}_\star \in \mathcal{B}^d(S)$ with known $S > 0$.

We define the following problem-dependent quantities:

$$\kappa_\star(T) := \left(\frac{1}{T} \sum_{t=1}^{T} \dot{\mu}(\boldsymbol{x}_{t,\star}^\intercal \boldsymbol{\theta}_\star) \right)^{-1}, \quad \kappa_{\mathcal{X}}(T) := \max_{t \in [T]} \max_{\boldsymbol{x} \in \mathcal{X}_t} \frac{1}{\dot{\mu}(\boldsymbol{x}^\intercal \boldsymbol{\theta}_\star)},$$

$$\text{and} \quad \kappa(T) := \max_{t \in [T]} \max_{\boldsymbol{x} \in \mathcal{X}_t} \max_{\boldsymbol{\theta} \in \mathcal{B}^d(S)} \frac{1}{\dot{\mu}(\boldsymbol{x}^\intercal \boldsymbol{\theta})}.$$

These can scale *exponentially in* S!

Prior Regret Guarantees

Regret lower bound:

Theorem 2 of Abeille et al. [2021]. Let $\mathcal{X}_t = \mathcal{S}^d(1)$. Then, for any problem instance $\boldsymbol{\theta}_\star$ and $T \geq d^2 \kappa_\star(\boldsymbol{\theta}_\star)$, there exists ϵ_T such that:

$$\min_{\pi:\text{policy}} \max_{\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_2 \leq \epsilon_T} \mathbb{E}[\mathrm{Reg}^B_{\boldsymbol{\theta},\pi}(T)] \geq \Omega\left(d \sqrt{\frac{T}{\kappa_\star(\boldsymbol{\theta}_\star)}} \right).$$

Regret upper bounds:

- `OFULog` [Abeille et al., 2021]: *Non-convex* confidence-set based UCB algorithm

$$dS^{\frac{3}{2}} \sqrt{\frac{T}{\kappa_\star(T)}} + \min\left\{ d^2 S^3 \kappa_{\mathcal{X}}(T), R_{\mathcal{X}}(T) \right\}.$$

- `OFULog-r` [Abeille et al., 2021]: Convex, loss-based confidence-set based UCB algorithm

$$dS^{\frac{5}{2}} \sqrt{\frac{T}{\kappa_\star(T)}} + \min\left\{ d^2 S^4 \kappa_{\mathcal{X}}(T), R_{\mathcal{X}}(T) \right\}.$$

- `ada-OFU-ECOLog` [Faury et al., 2022]: Online Newton step-based algorithm

$$dS \sqrt{\frac{T}{\kappa_\star(T)}} + d^2 S^6 \kappa(T).$$

Questions

- Can we construct tighter *convex, loss-based confidence set*, with improved dependency on S?
- Can this lead to a UCB algorithm that matches or beats `ada-OFU-ECOLog`?
- Does this lead to numerically meaningful performance?

Regret-to-Confidence-Set (R2CS)

R2CS starts by directly constructing a *likelihood loss-based* confidence set centered around the norm-constrained, unregularized maximum likelihood estimator (MLE), $\widehat{\boldsymbol{\theta}}_t$:

$$\widehat{\boldsymbol{\theta}}_t := \arg\min_{\|\boldsymbol{\theta}\|_2 \leq S} \left\{ \mathcal{L}_t(\boldsymbol{\theta}) \triangleq \sum_{s=1}^{t-1} \ell_s(\boldsymbol{\theta}) \right\}, \tag{1}$$

where ℓ_s is the logistic loss at time s, defined as

$$\ell_s(\boldsymbol{\theta}) := -r_s \log \mu(\langle \boldsymbol{x}_s, \boldsymbol{\theta} \rangle) - (1 - r_s) \log(1 - \mu(\langle \boldsymbol{x}_s, \boldsymbol{\theta} \rangle)).$$

Theorem 1. We have $\mathbb{P}[\forall t \geq 1, \ \boldsymbol{\theta}_\star \in \mathcal{C}_t(\delta)]$, where

$$\mathcal{C}_t(\delta) = \left\{ \boldsymbol{\theta} \in \mathcal{B}^d(S) : \mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\widehat{\boldsymbol{\theta}}_t) \leq \beta_{t-1}(\delta)^2 \right\}, \tag{2}$$

$$\beta_t(\delta) = \sqrt{10 d \log\left(\frac{St}{4d} + e \right) + 2((e-2) + S) \log \frac{1}{\delta}}. \tag{3}$$

This is a strict improvement over `OFULog-r`, which has the confidence radius of $\mathcal{O}_\delta\left(\sqrt{dS^3 \log t} \right)$.

Proof of R2CS for Logistic Losses

1. Decompose ℓ_s.

To use martingale concentrations, we begin by writing

$$r_s = \mu(\langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star \rangle) + \xi_s,$$

where ξ_s is a real-valued martingale difference noise.

The proof relies on the following two crucial lemmas:

Lemma 1. The following holds for any $\boldsymbol{\theta}$:

$$\ell_s(\boldsymbol{\theta}_\star) = \ell_s(\boldsymbol{\theta}) + \xi_s \langle \boldsymbol{x}_s, \boldsymbol{\theta} - \boldsymbol{\theta}_\star \rangle - \mathrm{KL}(\mu_s(\boldsymbol{\theta}_\star), \mu_s(\boldsymbol{\theta})).$$

Lemma 2. The following holds for any $\{\tilde{\boldsymbol{\theta}}_s\}$:

$$\mathcal{L}_{t+1}(\boldsymbol{\theta}_\star) - \mathcal{L}_{t+1}(\widehat{\boldsymbol{\theta}}_t) \leq \mathrm{Reg}^O(t) + \zeta_1(t) - \zeta_2(t), \tag{4}$$

where

$$\mathrm{Reg}^O(t) := \sum_{s=1}^{t} \left\{ \ell_s(\tilde{\boldsymbol{\theta}}_s) - \ell_s(\widehat{\boldsymbol{\theta}}_t) \right\},$$

$$\zeta_1(t) := \sum_{s=1}^{t} \xi_s \langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star - \tilde{\boldsymbol{\theta}}_s \rangle, \quad \zeta_2(t) := \sum_{s=1}^{t} \mathrm{KL}(\mu_s(\boldsymbol{\theta}_\star), \mu_s(\tilde{\boldsymbol{\theta}}_s)).$$

$\mathrm{Reg}^O(t)$ is the regret incurred by the online learning algorithm of our choice up to time t, $\zeta_1(t)$ is a sum of martingale difference sequences, and $\zeta_2(t)$ is a sum of KL's.

Proof sketch. Lemma 1 follows from the first-order Taylor expansion with integral remainder and careful terms rearranging. Lemma 2 then follows immediately.

2. Use state-of-the-art online regret for $\mathrm{Reg}^O(t)$.

Theorem 3 of Foster et al. [2018]. There is an online logistic regression algorithm with the following regret:

$$\mathrm{Reg}^O(t) \leq 10 d \log\left(\frac{St}{2d} + e \right). \tag{5}$$

We get $d \log S$ instead of dS, *for free*!

3. Use time-uniform Freedman to bound $\zeta_1(t)$.

Consequence of Lemma 3. For any $\eta \in \left[0, \frac{1}{2S}\right]$, the following holds w.p. at least $1 - \delta$: for all $t \geq 1$,

$$\zeta_1(t) \leq (e-2)\eta \sum_{s=1}^{t} \dot{\mu}(\boldsymbol{x}_s^\intercal \boldsymbol{\theta}_\star) \langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star - \tilde{\boldsymbol{\theta}}_s \rangle^2 + \frac{1}{\eta} \log \frac{1}{\delta}. \tag{6}$$

4. Use information-geometry to bound $\zeta_2(t)$.

Lemma 4. $\mathrm{KL}(\mu(z_2), \mu(z_1)) = D_m(z_1, z_2)$, where D_m is the Bregman divergence generated by $m(z) = \log(1 + e^z)$.

Combined with the self-concordant analysis [Abeille et al., 2021, Lemma 8], we obtain the following:

$$-\zeta_2(t) \leq -\frac{1}{2 + 2S} \sum_{s=1}^{t} \dot{\mu}(\boldsymbol{x}_s^\intercal \boldsymbol{\theta}_\star) \langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star - \tilde{\boldsymbol{\theta}}_s \rangle^2. \tag{7}$$

5. Combine everything.

Set $\eta = \frac{1}{2(e-2) + 2S}$, and plug Eqn. (5), (6), and (7) into Eqn. (4).

`OFULog+`

`OFULog+` is of the following form:

1. Obtain $\widehat{\boldsymbol{\theta}}_t$ (Eqn. (1)) and $\mathcal{C}_t(\delta)$ (Theorem 1)
2. Solve $(\boldsymbol{x}_t, \boldsymbol{\theta}_t) = \arg\max_{\boldsymbol{x} \in \mathcal{X}_t, \boldsymbol{\theta} \in \mathcal{C}_t(\delta)} \mu(\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle)$
3. Play $\boldsymbol{x}_t$, then observe/receive a reward $r_t \in \{0, 1\}$.

We then have the following *state-of-the-art* regret bound:

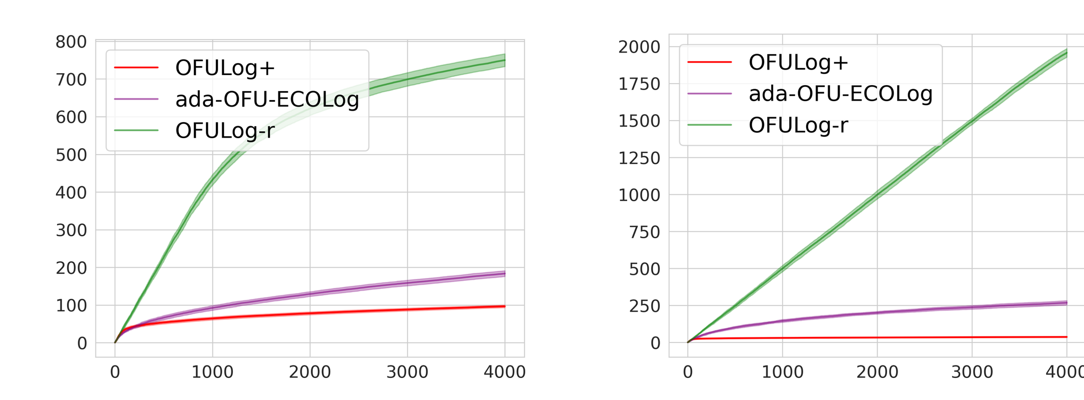
Theorem 3. `OFULog+` attains the following regret bound with probability at least $1 - \delta$:

$$\mathrm{Reg}^B(T) \lesssim_\delta dS \sqrt{\frac{T}{\kappa_\star(T)}} + \min\left\{ d^2 S^2 \kappa_{\mathcal{X}}(T), R_{\mathcal{X}}(T) \right\},$$

where $R_{\mathcal{X}}(T)$ is a term relating to the arm set geometry [Abeille et al., 2021, Section 4].

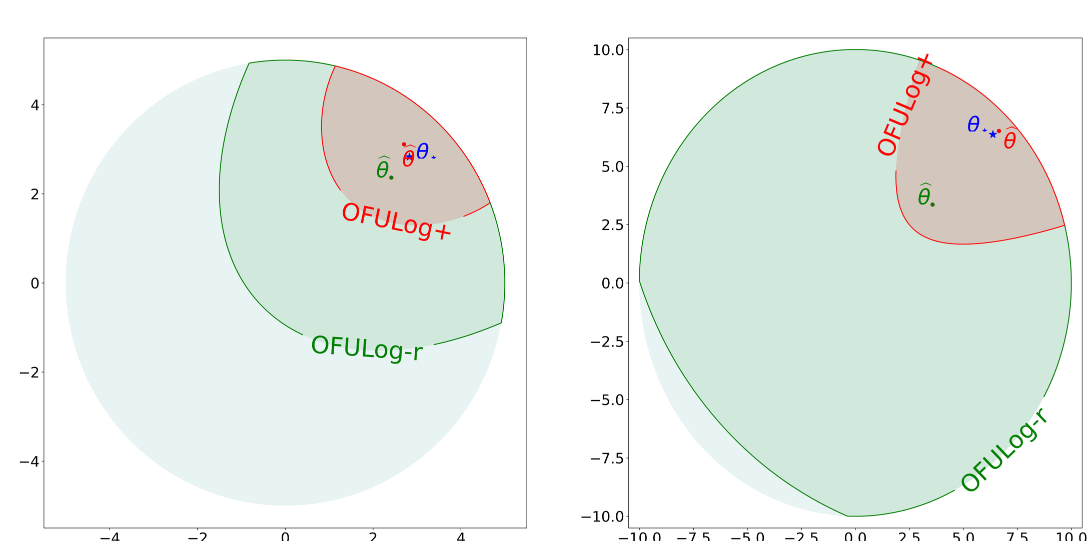
Proof novelties. Time-uniform Freedman (Lemma 3) and elliptical potential count lemma [Gales et al., 2022, Lemma 7].

Experiments



(a) $S = 5$ (b) $S = 10$

Figure 1:Numerical regrets.



(a) $S = 5$ (b) $S = 10$

Figure 2:Confidence sets at $t = 4000$ from a single run.

Multinomial Logistic (MNL) Bandits

Via R2CS, we attain the *state-of-the-art* regret bound for MNL bandits over prior arts [Amani and Thrampoulidis, 2021, Zhang and Sugiyama, 2023]:

Theorem 5. `MNL-UCB+` attains the following regret bound with probability at least $1 - \delta$:

$$\mathrm{Reg}^B(T) \lesssim_\delta d\sqrt{KS} \min\left\{ \kappa(T) T, \sqrt{ST} + dKS\kappa(T) \right\}.$$

Open Problems

- poly(S)-free regret for (multinomial) logistic bandits?
- Extension to GLM bandits?

References

M. Abeille et al. Instance-Wise Minimax-Optimal Algorithms for Logistic Bandits. In *AISTATS*, 2021.

S. Amani and C. Thrampoulidis. UCB-based Algorithms for Multinomial Logistic Regression Bandits. In *NeurIPS*, 2021.

L. Faury et al. Jointly Efficient and Optimal Algorithms for Logistic Bandits. In *AISTATS*, 2022.

D. J. Foster et al. Logistic Regression: The Importance of Being Improper. In *COLT*, 2018.

S. B. Gales et al. Norm-Agnostic Linear Bandits. In *AISTATS*, 2022.

L. Li et al. A Contextual-Bandit Approach to Personalized News Article Recommendation. In *WWW*, 2010.

Y.-J. Zhang and M. Sugiyama. Online (Multinomial) Logistic Bandit: Improved Regret and Constant Computation Cost. In *NeurIPS*, 2023.

KAIST AI Kim Jaechul Graduate School

Computer Science

Optimization and Statistical Inference LAB

OptiML Optimization & Machine Learning