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

Junghyun Lee (KAIST AI), Se-Young Yun (KAIST AI), Kwang-Sung Jun (Dept. of CS, Univ. of Arizona)

KAIST AI
Kim Jaechul Graduate School

Computer Science

Optimization and Statistical Inference LAB

OptiML
Optimization & Machine Learning

The Plan for Today

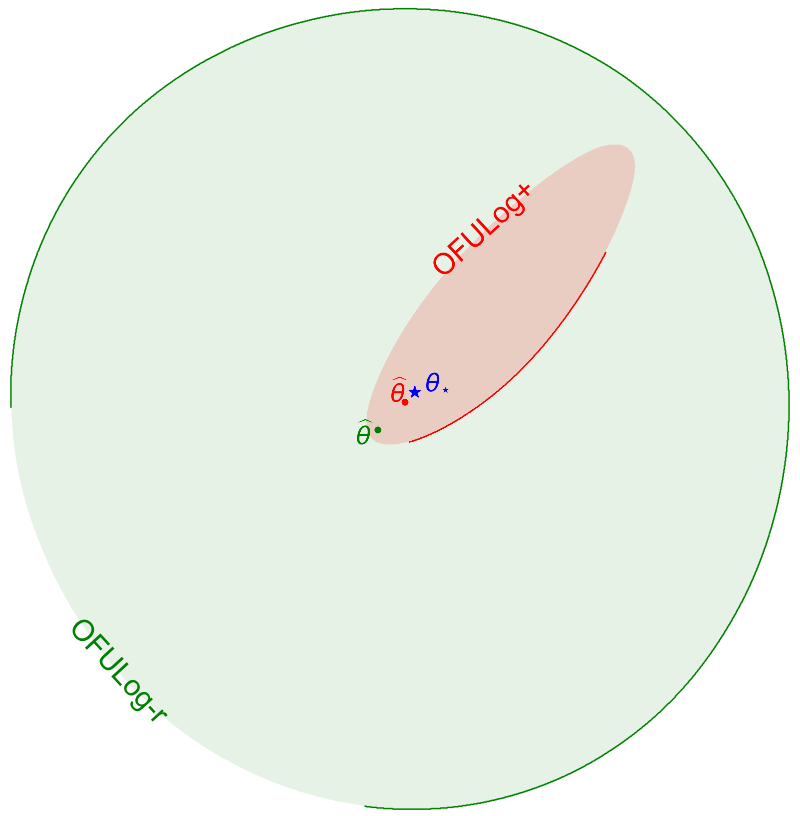
- Logistic Bandits 101
- New confidence set for logistic bandits via (online) regret-to-confidence-set (**O2CS**)
- Improved Regrets for Logistic Bandits
- Conclusion and Future Works

The Plan for Today

- Logistic Bandits 101
- New confidence set for logistic bandits via (online) regret-to-confidence-set (**O2CS**)
- Improved Regrets for Logistic Bandits
- Conclusion and Future Works

~~~

We propose a framework in which one can construct a confidence set using an *achievable* online learning regret bound (*without ever running the alg*), and apply it to improve the regret bounds of (multinomial) logistic bandits.

~~~

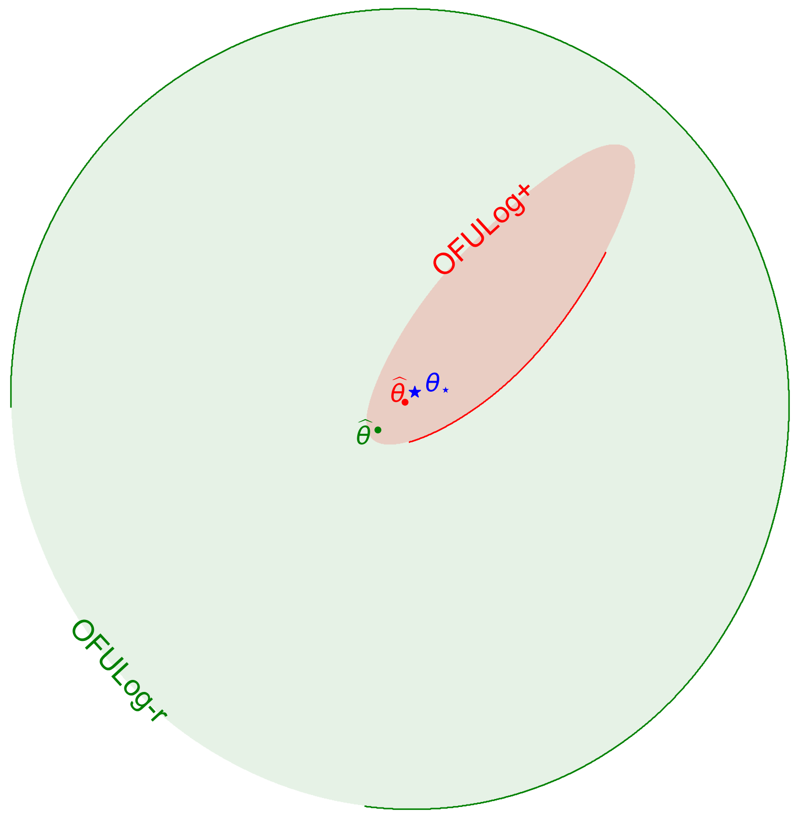
The Plan for Today



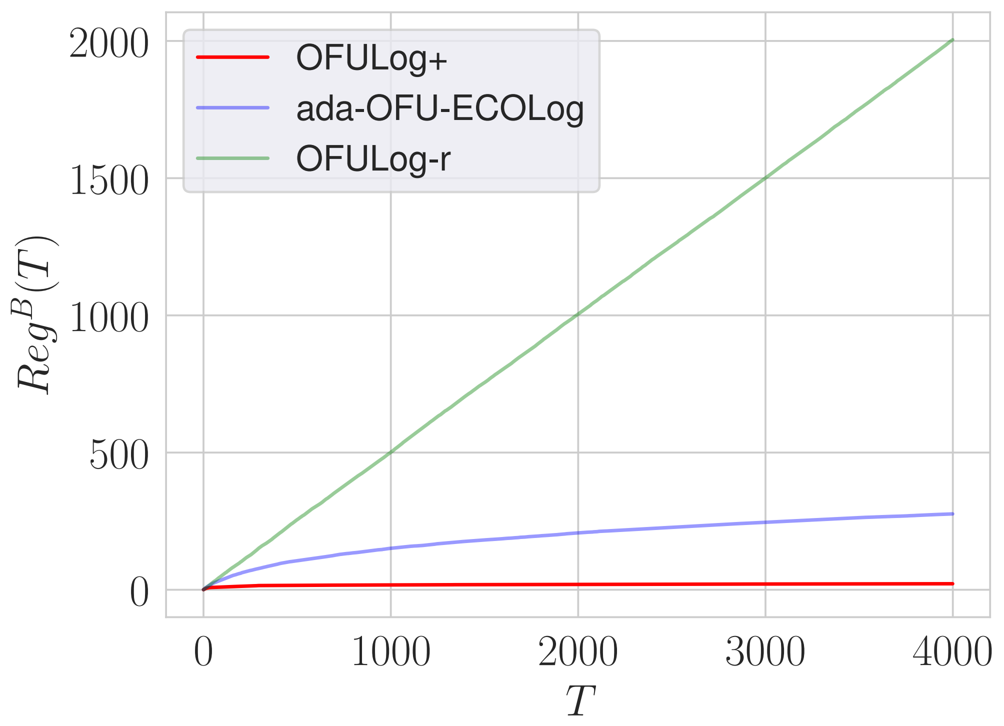
- Logistic Bandits 101
- New confidence set for logistic bandits via (online) regret-to-confidence-set (**O2CS**)
- Improved Regrets for Logistic Bandits
- Conclusion and Future Works

~~~

We propose a framework in which one can construct a confidence set using an *achievable* online learning regret bound (*without ever running the alg*), and apply it to improve the regret bounds of (multinomial) logistic bandits.

~~~

The Plan for Today



- Logistic Bandits 101
- New confidence set for logistic bandits via (online) regret-to-confidence-set (**O2CS**)
- Improved Regrets for Logistic Bandits
- Conclusion and Future Works



~~~

We propose a framework in which one can construct a confidence set using an *achievable* online learning regret bound (*without ever running the alg*), and apply it to improve the regret bounds of (multinomial) logistic bandits.

~~~

The Plan for Today

- Logistic Bandits 101
- New confidence set for logistic bandits via (online) regret-to-confidence-set (**O2CS**)
- Improved Regrets for Logistic Bandits
- Conclusion and Future Works

~~~

We propose a framework in which one can construct a confidence set using an *achievable* online learning regret bound (*without ever running the alg*), and apply it to improve the regret bounds of (multinomial) logistic bandits.

~~~

Logistic Bandits 101

Motivation

- Useful in modeling exploration-exploitation dilemma with *binary/discrete-valued* rewards and items' feature vectors
  - e.g., news recommendation ('click', 'no click'), online ad placement ('click', 'show me later', 'never show again', 'no click')
- Naive reduction to linear bandits is quite suboptimal[Li et al., WWW'10; ICMLW'11]!

The Web Conference 2023 - Seoul Test of Time Award
(presented at The Web Conference 2023 in Austin)

Winners: **Wei Chu**, **Lihong Li**, **John Langford** and **Robert Schapire**
for their paper "A Contextual-Bandit Approach to Personalized News Article Recommendation".

Logistic Bandits 101

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 $x_t \in \mathcal{X}_t$ according to some policy
3. Receive a *binary* reward $r_t \sim \mathrm{Ber}(\mu(\langle x_t, \theta_\star \rangle))$
    - $\theta_\star$ is unknown to the learner
    - $\mu(z) := (1 + e^{-z})^{-1}$ is the logistic function, $\dot{\mu}(z) = \mu(z)(1 - \mu(z))$ is its first derivative

Goal:

$$\text{Minimize } \mathrm{Reg}^B(T) := \sum_{t=1}^{T} \left\{ \mu(\langle x_{t,\star}, \theta_\star \rangle) - \mu(\langle x_t, \theta_\star \rangle) \right\}, \text{ where } x_{t,\star} := \mathrm{argmax}_{x \in \mathcal{X}_t} \langle x, \theta_\star \rangle.$$

Logistic Bandits 101

Assumptions

Assumption 1. $\bigcup_{t=1}^{\infty} \mathcal{X}_t \subseteq \mathbf{B}^d(1)$

Assumption 2. $\theta_\star \in \mathbf{B}^d(S)$ => today's main quantity of interest!

We consider the following quantities describing the difficulty of the problem:

$$\kappa_\star(T) := \left(\frac{1}{T} \sum_{t=1}^{T} \dot{\mu}(\langle x_{t,\star}, \theta_\star \rangle) \right)^{-1}, \quad \kappa_{\mathcal{X}}(T) := \max_{t \in [T]} \max_{x \in \mathcal{X}_t} \frac{1}{\dot{\mu}(\langle x, \theta_\star \rangle)}.$$

They can scale ***exponentially in*** S [Faury et al., ICML'20]

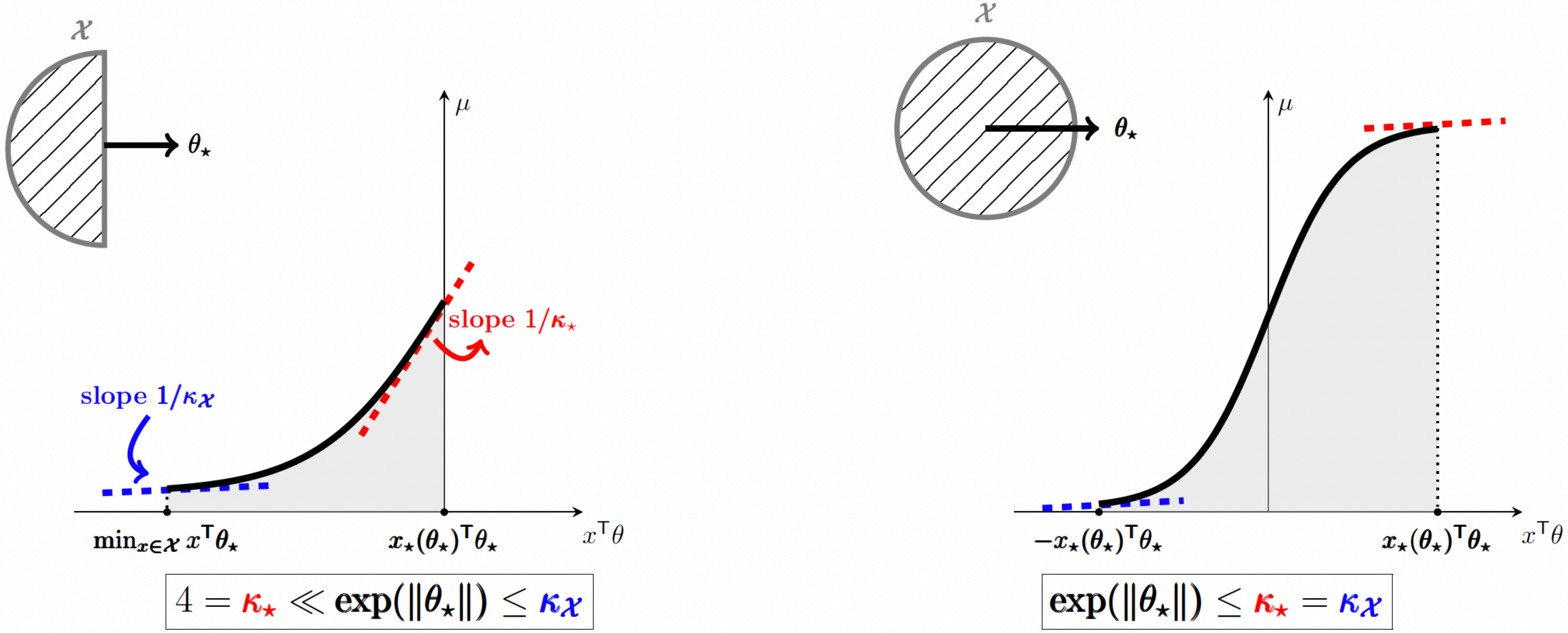
Logistic Bandits 101

$d\sqrt{T/\kappa_\star(T)}$ is minimax optimal (taken from slides of L. Faury on his website)

Theorem 2. **[Local Lower-Bound; Abeille et al., AISTATS'21]** Let $\mathcal{X}_t = \mathbf{S}^d(1)$ and . Then, for any problem instance $\theta_\star$ and for $T \geq d^2\kappa_\star(\theta_\star)$, there exists $\epsilon_T > 0$ such that:

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

- More nonlinear (flatter tail), the easier!
- Transient regret (small t):
  - Exploration of "detrimental" arms
- **Permanent regret (large t):**
  - Sub-linear regret, as the estimate is sufficiently close to $\theta_\star$
  - Linear bandit with local slope around $\theta_\star$, $\dot{\mu}(\langle x_\star, \theta_\star\rangle) \sim \dfrac{1}{\kappa_\star(T)}$



(a) Assymetric arm-set.

(b) Symmetric arm-set (unit-ball).

Logistic Bandits 101

State-of-the-Arts, so-far

- **OFULog** [Abeille et al., AISTATS'21]. *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., AISTATS'21]. Convex relaxation of OFULog ~ loss-based confidence set

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

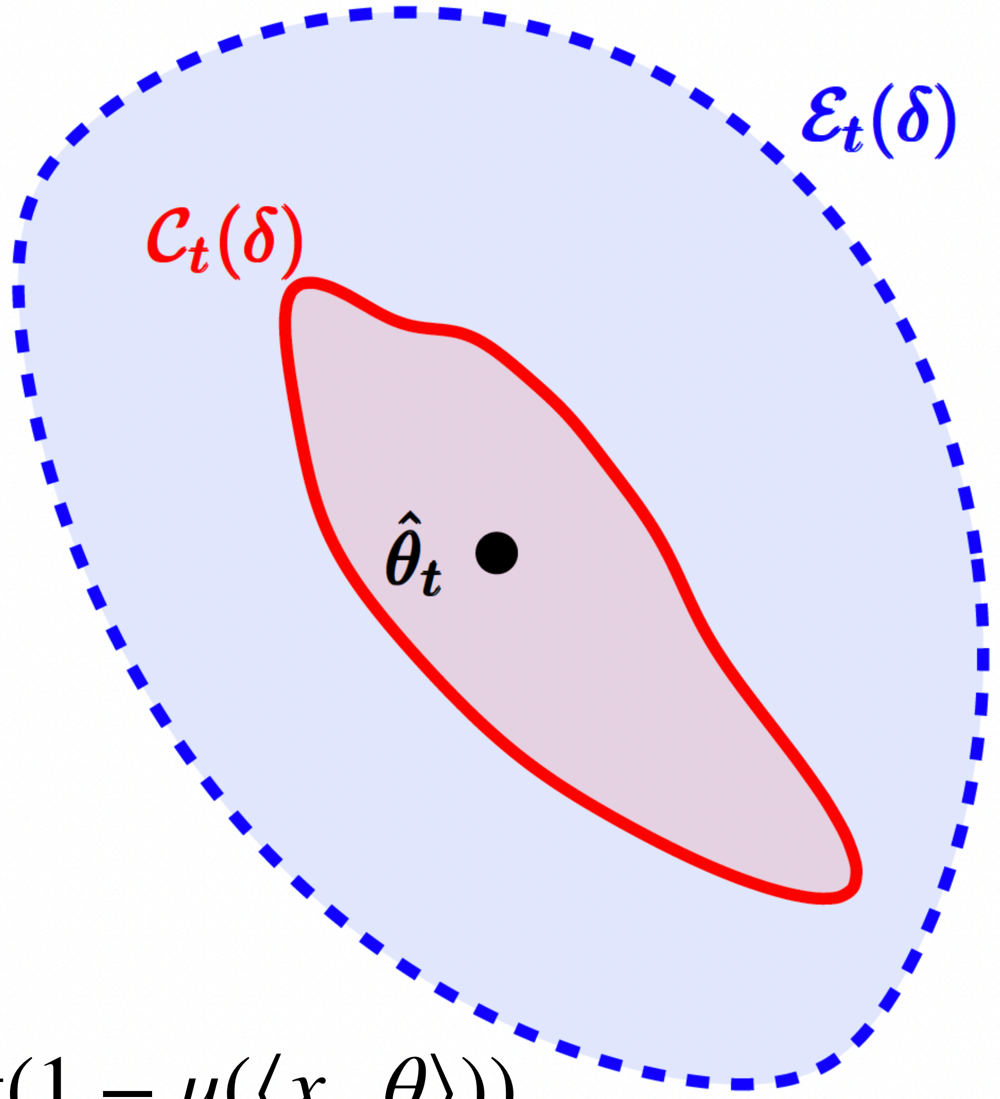
- **ada-OFU-ECOLog** [Faury et al., AISTATS'22]. Online Newton step [Hazan et al., 2007]-based algorithm

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

Can we construct tighter (improved dependency in S) *loss-based confidence set*?? Can we make UCB great again (i.e., UCB-type algorithm that matches or beats ada-OFU-ECOLog)?

Logistic Bandits 101

More details in OFULog(-r)

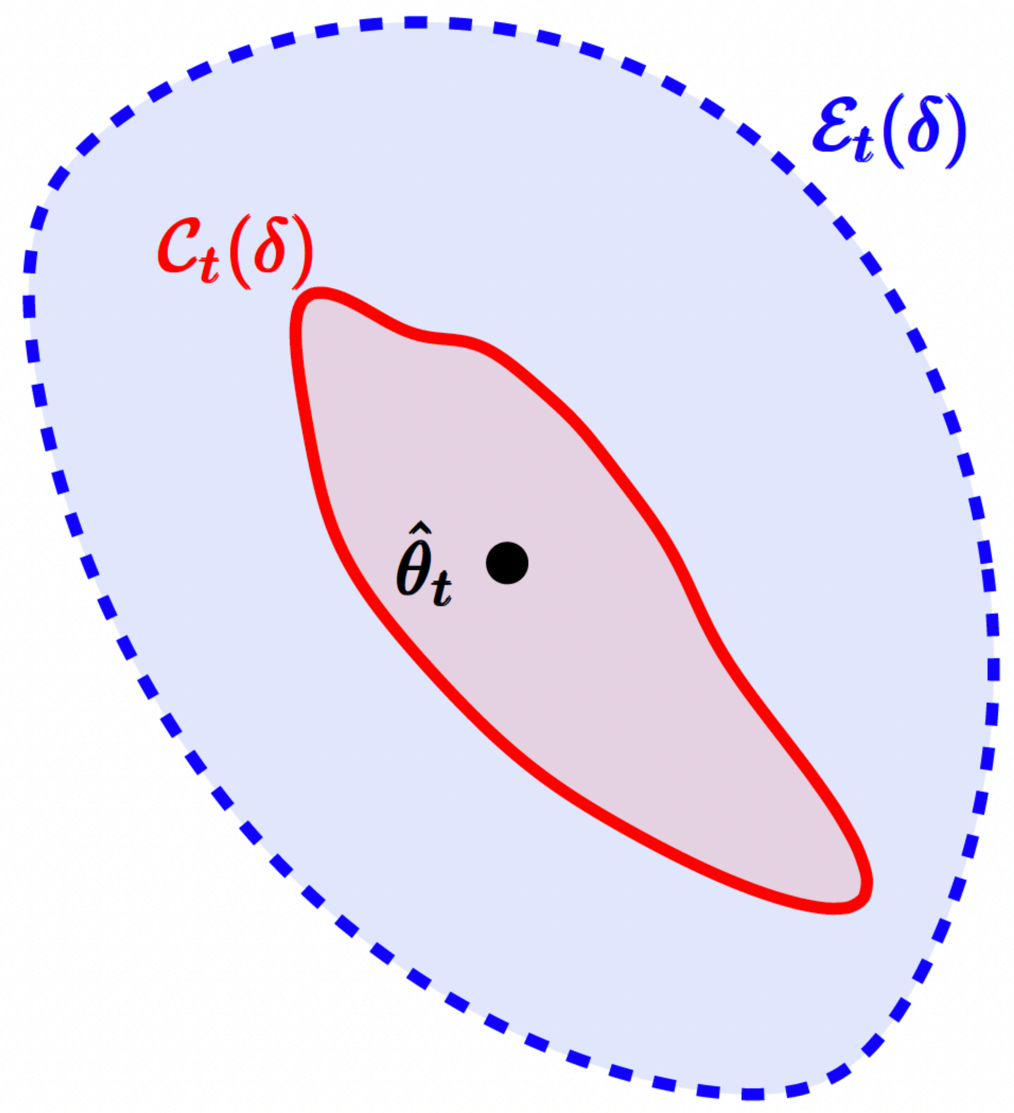


- OFULog and OFULog-r are of the following form:
  1. Solve $\widehat{\theta}_t = \text{argmin}_{\theta \in \mathbb{R}^d} \left[\mathscr{L}_t(\theta) \triangleq \sum_{s=1}^{t-1} \ell_s(\theta) + \lambda_t \|\theta\|_2^2 \right]$, where $\ell_s(\theta) := -r_s \log \mu(\langle x_s, \theta \rangle) - (1 - r_s) \log(1 - \mu(\langle x_s, \theta \rangle))$
  2. Obtain a confidence-set $C_t(\delta) \subseteq \mathbb{B}^d(S)$ satisfying $\mathbb{P}\left[\forall t \geq 1,\ \theta_\star \in C_t(\delta)\right] \geq 1 - \delta$.
  3. Solve $(x_t, \theta_t) = \text{argmax}_{x \in \mathcal{X}_t, \theta \in C_t(\delta)} \mu(\langle x, \theta \rangle)$, play x_t and observe/receive a reward r_t

- **OFULog** [Abeille et al., AISTATS'21]: $C_t(\delta) := \left\{ \theta \in \mathbb{B}^d(S) : \left\| \nabla \mathscr{L}_t(\theta) - \nabla \mathscr{L}_t(\widehat{\theta}_t) \right\|_{\mathbf{H}_t^{-1}(\theta)} \leq \mathcal{O}\left(\sqrt{dS \log t}\right) \right\}$
- **OFULog-r** [Abeille et al., AISTATS'21]: $C_t(\delta) := \left\{ \theta \in \mathbb{B}^d(S) : \mathscr{L}_t(\theta) - \mathscr{L}_t(\widehat{\theta}_t) \leq \mathcal{O}\left(\sqrt{dS^3 \log t}\right) \right\} (= E_t(\delta))$

The *multiplicative S*'s comes from rather naive applications of self-concordant ($|\ddot{\mu}| \leq \dot{\mu}$) analyses [Bach, 2010]

Logistic Bandits 101



$C_t(\delta)$: Gradient-based Confidence set [Abeille et al., AISTATS'21]

- **OFULog** **[Abeille et al., AISTATS'21]:**

$$C_t(\delta) := \left\{ \theta \in \mathbb{B}^d(S) : \left\| \nabla \mathcal{L}_t(\theta) - \nabla \mathcal{L}_t(\widehat{\theta}_t) \right\|_{\mathbf{H}_t^{-1}(\theta)} \leq \mathcal{O}\left(\sqrt{dS \log t}\right) \right\}$$

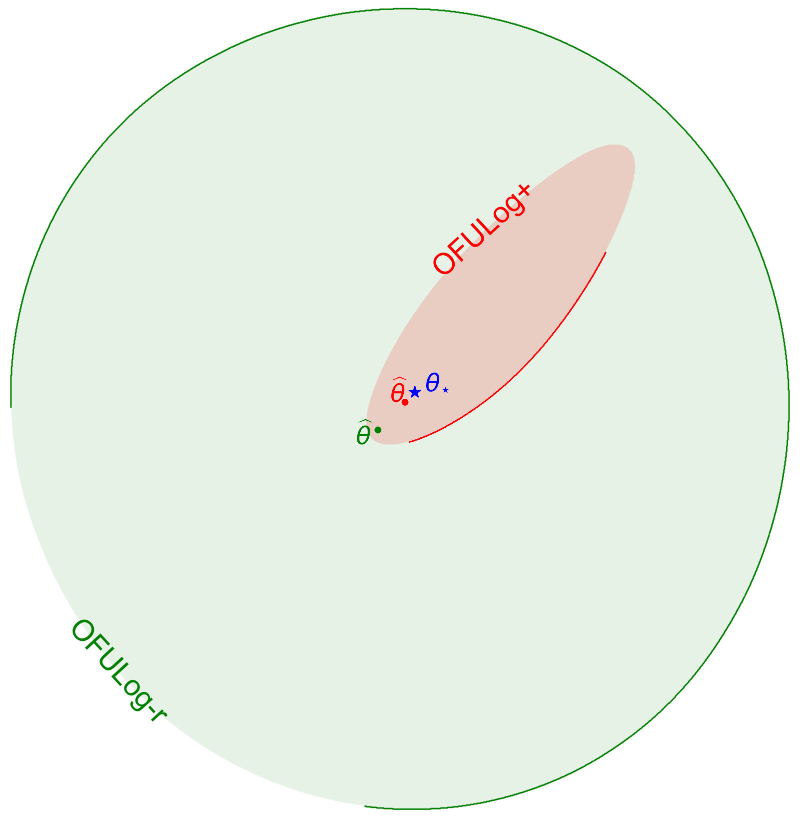
- Gradient of the logistic loss:

$$\nabla \mathcal{L}_t(\theta_\star) = \underbrace{\sum_{s=1}^{t-1} \left(\mu(\langle x_s, \theta_\star \rangle) - r_s\right) x_s}_{\text{sum of martingale differences}} + 2\lambda_t \theta_\star$$

  - If $\widehat{\theta}_t$ is a good estimator, then the gradient at $\theta_\star$ should be near zero!
  - We can quantify "pointwise confidence" with the inverse of Hessian (covariance) $H(\theta_\star) = \sum_{s=1}^{t-1} \dot{\mu}(\langle x_s, \theta_\star \rangle) x_s x_s^\top + \lambda I$

- In order to obtain a confidence set, we require the local metric $\|\cdot\|_{\mathbf{H}_t^{-1}(\theta)}$ that depends on the choice of θ
- Relaxing $C_t(\delta)$ to the loss-based set $E_t(\delta)$ gives a *convex* confidence set, but is not tight in S

The Plan for Today



- Logistic Bandits 101
- New confidence set for logistic bandits via (online) regret-to-confidence-set (**O2CS**)
- Improved Regrets for Logistic Bandits
- Conclusion and Future Works

~~~

We propose a framework in which one can construct a confidence set using an *achievable* online learning regret bound (*without ever running the alg*), and apply it to improve the regret bounds of (multinomial) logistic bandits.

~~~

Regret-to-Confidence-Set Conversion (R2CS)

Main Theorem - Improved Confidence Set for Logistic Loss

- Let us consider *norm-constrained, **unregularized*** **MLE**:

$$\widehat{\theta}_t := \mathrm{argmin}_{\theta \in \mathbb{B}^d(S)} \left[\mathscr{L}_t(\theta) := \sum_{s=1}^{t-1} \ell_s(\theta) \right], \text{ where } \ell_s(\theta) := -r_s \log \mu(\langle x_s, \theta \rangle) - (1 - r_s) \log(1 - \mu(\langle x_s, \theta \rangle))$$

Theorem 1. **[Lee et al., AISTATS'24]** We have $\mathbb{P}\left[\forall t \geq 1,\ \theta_\star \in C_t(\delta)\right] \geq 1 - \delta$, where

$$C_t(\delta) := \left\{ \theta \in \mathbb{B}^d(S) : \mathscr{L}_t(\theta) - \mathscr{L}_t(\widehat{\theta}_t) \leq \beta_t(\delta)^2 \right\},$$

$$\beta_t(\delta) := \sqrt{10d \log\left(\frac{St}{4d} + e\right) + 2((e-2) + S)\log\frac{1}{\delta}} = \mathcal{O}(\sqrt{(d+S)\log t})$$

Strict improvement over prior (loss-based & convex) confidence-set radius of $\mathcal{O}\left(\sqrt{dS^3 \log t}\right)$

Regret-to-Confidence-Set Conversion (R2CS)

Proof Sketch of Theorem 1

Decomposing the logistic loss with any online learning algorithm $\tilde{\theta}_s$:

$$\mathcal{L}_t(\theta_\star) - \mathcal{L}_t(\hat{\theta}_t) = \sum_{s=1}^{t-1} \ell_s(\theta_\star) - \ell_s(\hat{\theta}_t) = \underbrace{\sum_{s=1}^{t-1} \left(\ell_s(\tilde{\theta}_s) - \ell_s(\hat{\theta}_t) \right)}_{\color{blue}{\mathrm{Reg}^O(t)}} + \underbrace{\sum_{s=1}^{t-1} \left(\ell_s(\theta_\star) - \ell_s(\tilde{\theta}_s) \right)}_{\color{red}{\zeta(t) = \zeta_1(t) - \zeta_2(t)}}$$

where $\color{red}{\zeta_1(t)} := \sum_{s=1}^{t-1} \xi_s \langle x_s, \tilde{\theta}_s - \theta_\star \rangle, \quad \color{red}{\zeta_2(t)} := \sum_{s=1}^{t-1} \mathrm{KL}(\mu_s(\langle x_s, \theta_\star \rangle), \mu_s(\langle x_s, \tilde{\theta}_s \rangle))$

- $\color{blue}{\mathrm{Reg}^O(t)}$ is the online regret up to time t, and $\color{red}{\zeta(t)}$ is the superiority of the online learning algorithm in terms of loss compared to $\theta_\star$ which is expected very small (independent to t) with high probability since $\theta_\star$ is the problem instance parameter.
- $\hat{\theta}_t$ is the optimal parameter for the entire batch til time t, while $\tilde{\theta}_s$ is the online prediction.

Regret-to-Confidence-Set Conversion (R2CS)

Proof Sketch of Theorem 1

1. Decomposing the logistic loss such that the $\beta_t(\delta)^2$ is expressed as a sum of $\mathrm{Reg}^O(t)$, **regret of *any* online learning algorithm of our choice**, $\zeta_1(t)$, a sum of martingales, and $-\zeta_2(t)$, a (negative) sum of KL-divergences.
2. For $\mathrm{Reg}^O(t)$, we utilize the state-of-the-art online regret of Foster et al., (COLT'18), which reduces the usual dS to $d \log S$, *without ever running the algorithm*.
3. For $\zeta_1(t)$, we utilize a novel anytime variant of the Freedman's concentration inequality [Freedman, 1975] for martingales.
4. For $-\zeta_2(t)$, we utilize the Bregman geometrical interpretation of the KL-divergence, along with self-concordant results.

Regret-to-Confidence-Set Conversion (R2CS)

Proof of Theorem 1

1. Decomposing the logistic loss such that the $\beta_t(\delta)^2$ is expressed as a sum of $\mathrm{Reg}^O(t)$, regret of *any* online learning algorithm of our choice, $\zeta_1(t)$, a sum of martingales, and $-\zeta_2(t)$, a (negative) sum of KL-divergences.

Note that $r_s = \mu(\langle x_s, \theta_\star \rangle) + \xi_s$ for some martingale difference noise ξ_s.

Lemma 1 & 2. **[Lee et al., AISTATS'24]** For the logistic loss ℓ_s and *any* sequence of parameters $\{\tilde{\theta}_s\}$ (e.g., "outputted" from some online algorithm), the following holds:

$$\sum_{s=1}^{t} \ell_s(\theta_\star) - \ell_s(\widehat{\theta}_t) \leq \mathrm{Reg}^O(t) + \zeta_1(t) - \zeta_2(t) .$$

The proof utilizes second-order Taylor expansion of ℓ_s with *integral remainder*!

Regret-to-Confidence-Set Conversion (R2CS)

Proof of Theorem 1

2. For $\text{Reg}^O(t)$, we utilize the state-of-the-art online regret of Foster et al., (COLT'18), which reduces the usual dS to $d \log S$, *without ever running the algorithm*.

Theorem 3. **[Foster et al., COLT'18]** There exists an (improper learning) algorithm for online logistic regression with the following regret:

$$\text{Reg}^O(t) \leq 10d \log \left(\frac{St}{4d} + e \right).$$

Note how we get $d \log S$ instead of dS!! Even better, we get this *without ever running the algorithm*, which in this case, is quite expensive!

Regret-to-Confidence-Set Conversion (R2CS)

Proof of Theorem 1

3. For $\zeta_1(t)$, we utilize a novel anytime variant of the Freedman's concentration inequality [Freedman, 1975] for martingales.

Lemma 3. [Lee et al., AISTATS'24] Let $\{X_s\}_{s=1}^t$ be a martingale difference sequence satisfying $\max_s |X_s| \leq R$ a.s., and let $\mathcal{F}_s := \sigma(X_1, \cdots, X_s)$. Then for any $\delta \in (0,1)$ and any $\eta \in [0, 1/R]$, the following holds:

$$\mathbb{P}\left[\forall t \geq 1,\ \sum_{s=1}^t X_s \leq (e-2)\eta \sum_{s=1}^t \mathbb{E}[X_s^2 | \mathcal{F}_{s-1}] + \frac{1}{\eta}\log\frac{1}{\delta}\right] \geq 1 - \delta.$$

The proof is based on Theorem 1 of Beygelzimer et al. (ICML'11) and the Ville's inequality [Ville, 1939]

With this, we have the following: for any choice of $\delta \in (0,1)$ and $\eta \in \left[0, \frac{1}{2S}\right]$,

$$\mathbb{P}\left[\forall t \geq 1,\ \zeta_1(t) \leq (e-2)\eta \sum_{s=1}^t \dot{\mu}(\langle x_s, \theta_\star \rangle)\langle x_s, \theta_\star - \tilde{\theta}_s \rangle^2 + \frac{1}{\eta}\log\frac{1}{\delta}\right] \geq 1 - \delta$$

Regret-to-Confidence-Set Conversion (R2CS)

Proof of Theorem 1

4. For $-\zeta_2(t)$, we utilize the Bregman geometrical interpretation of the KL-divergence, along with self-concordant results.

Observation. $D_m(z_1, z_2) := m(z_1) - m(z_2) - \nabla m(z_2)^\intercal (z_1 - z_2) = \int_{z_1}^{z_2} m''(z)(z_1 - z)dz$.

Lemma 4. [Lee et al., AISTATS'24] $\mathrm{KL}(\mu(z_1), \mu(z_1)) = D_m(z_1, z_2)$, where $m(z) := \log(1 + e^z)$.

Combining above with self-concordant analysis [Lemma 8 of Abeille et al., AISTATS'21], we have:

$$-\zeta_2(t) \leq -\frac{1}{2+2S} \sum_{s=1}^{t} \dot{\mu}(\langle x_s, \theta_\star \rangle) \langle x_s, \theta_\star - \tilde{\theta}_s \rangle^2$$

Regret-to-Confidence-Set Conversion (R2CS)

Proof of Theorem 1

Combining everything, we have: with probability at least $1-\delta$, for all $t \in [T]$,

$$\sum_{s=1}^{t} \ell_s(\theta_\star) - \ell_s(\widehat{\theta}_t)$$

$$\leq \mathrm{Reg}^O(t) + \zeta_1(t) - \zeta_2(t)$$

$$\leq 10d \log\left(\frac{St}{4d} + e\right) + (e-2)\eta \sum_{s=1}^{t} \dot{\mu}(\langle x_s, \theta_\star\rangle)\langle x_s, \theta_\star - \tilde{\theta}_s\rangle^2 + \frac{1}{\eta}\log\frac{1}{\delta} - \frac{1}{2+2S}\sum_{s=1}^{t} \dot{\mu}(\langle x_s, \theta_\star\rangle)\langle x_s, \theta_\star - \tilde{\theta}_s\rangle^2$$

$$\leq 10d \log\left(\frac{St}{4d} + e\right) + 2((e-2)+S)\log\frac{1}{\delta},$$

where we choose $\eta = \frac{1}{2(e-2)+2S)} < \frac{1}{2S}$ that satisfies $-\frac{1}{2+2S} + \frac{e-2}{2(e-2)+2S} < 0$.

Related Work: Online-to-Something Conversions

Online Learning -> Concentration of Measure

Online-to-confidence-set: Start from some online learning algorithm $\mathscr{A}$ with regret $\sum_{s=1}^{t} \ell_s(\theta_s) - \ell_s(\theta_\star) \leq B(t)$, then bound LHS to obtain a quadratic-type confidence set on $\theta_\star$ that depends on the outputs of $\mathscr{A}$ whose radius scales with $B(t)$ [**Abbasi-Yadkori et al., AISTATS'12**; Jun et al., NeurIPS'17]

Advantages of O2SC: "progress in constructing better algorithms for online prediction problems directly translates into tighter confidence sets" [Abbasi-Yadkori et al., AISTATS'12]; see Chapter 23.3 of Lattimore and Szepesvári (2020)

BUT, what if the online learning has a trade-off between computational complexity and regret??

e.g., online logistic regression: good regret & bad computational complexity [Foster et al., COLT'18] or worse regret & good computational complexity [Jézéquel et al., COLT'20]

Our algorithm does not run the online learning part!

Related Work: Online-to-Something Conversions

Online-to-PAC Conversion

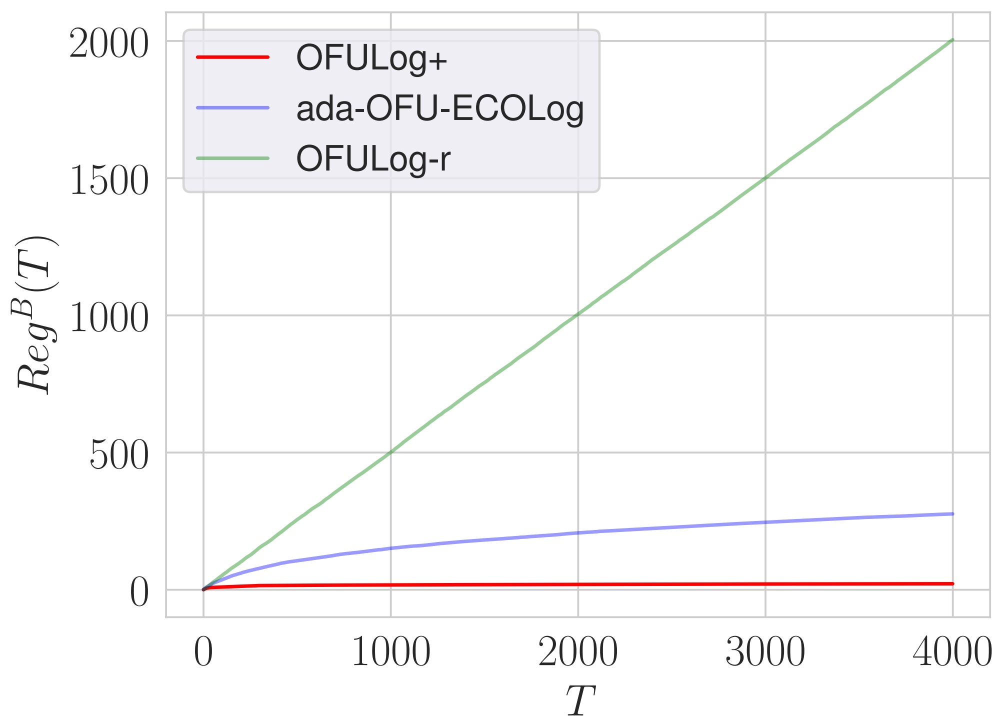
Recently, Lugosi & Neu (arXiv'23) introduced **online-to-PAC** conversion:

"… the *existence* of an online learning algorithm with bounded regret in this game implies a bound on the generalization error of the statistical learning algorithm up to a martingale concentration term that is independent of the complexity of the statistical learning method."

=> very similar spirit, but the goal is different from ours.

The Plan for Today

- Logistic Bandits 101
- New confidence set for logistic bandits via (online) regret-to-confidence-set (**O2CS**)
- Improved Regrets for Logistic Bandits
- Conclusion and Future Works



~~~

We propose a framework in which one can construct a confidence set using an *achievable* online learning regret bound (*without ever running the alg*), and apply it to improve the regret bounds of (multinomial) logistic bandits.

~~~

Improved Regret of Logistic Bandits

OFULog+

- Note that our algorithm is of the same form with OFULog-r, except we've only changed the confidence set radius, $\mathcal{O}\left(\sqrt{dS^3 \log t}\right)$ to $\mathcal{O}\left(\sqrt{(d+S)\log t}\right)$, which we call *OFULog+*

Theorem 3. **[Lee et al., AISTATS'24]** OFULog+ incurs the following regret bound w.p. at least $1-\delta$:

$$\mathrm{Reg}^B(T) \lesssim \underbrace{dS\sqrt{\frac{T}{\kappa_\star(T)}}}_{\text{permanent term}} + \underbrace{\min\left\{d^2S^2\kappa_{\mathcal{X}}(T), R_{\mathcal{X}}(T)\right\}}_{\text{transient term}}$$

(Refer to our paper for the missing definitions)

Improved Regret of Logistic Bandits

OFULog+ is the state-of-the-art, taking S into account

- **OFULog** [Abeille et al., AISTATS'21]. *Non-convex* confidence-set-based UCB algorithm

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

- **OFULog-r** [Abeille et al., AISTATS'21]. Convex relaxation of OFULog

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

- **ada-OFU-ECOLog** [Faury et al., AISTATS'22]. Online Newton step (ONS) [Hazan et al., 2007]-based algorithm

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

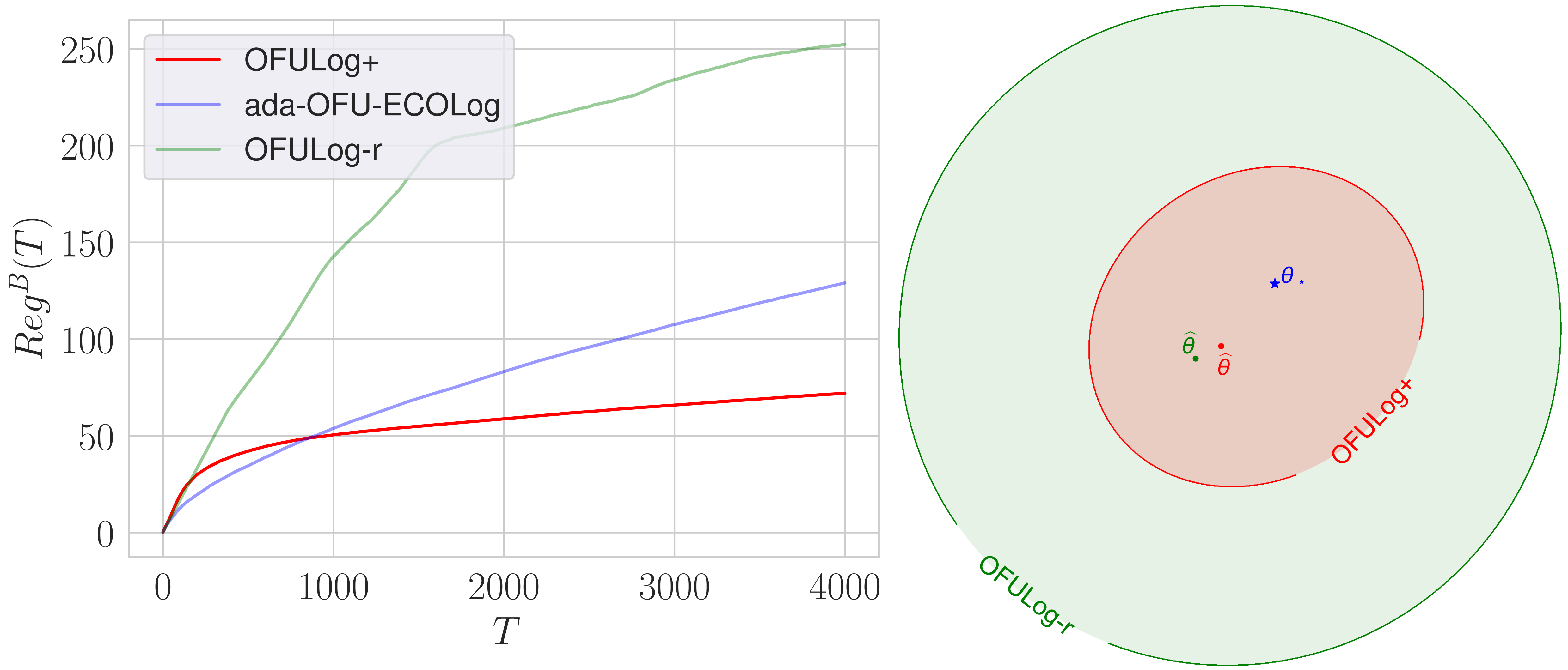
- **OFULog+** [Lee et al., AISTATS'24]. Tight loss-based confidence set

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

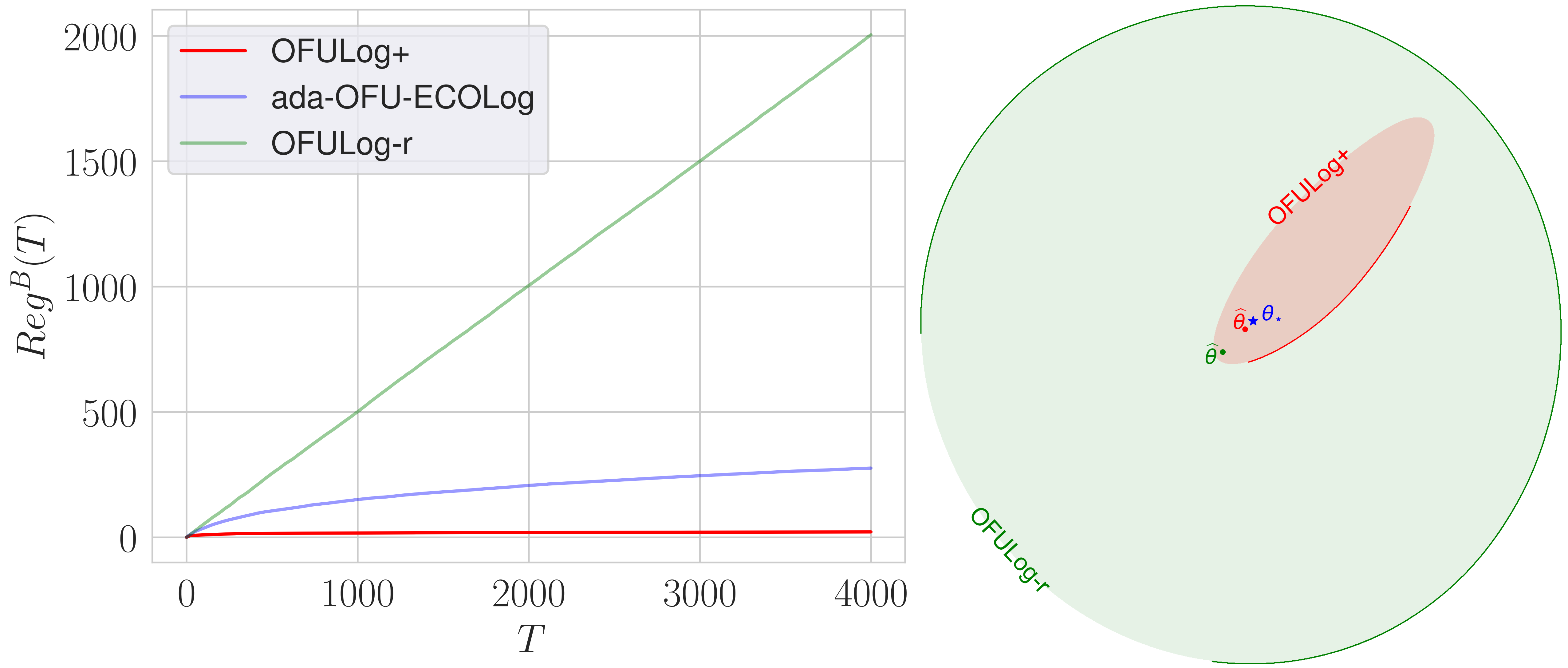
Improved Regret of Logistic Bandits

Experiments

- One may wonder, does shaving off dependencies on S really help in practice?
- Synthetic experiments show that this is indeed beneficial, by a large margin!!
  - (In fact, we believe that the current analysis can be made tighter, which may explain the large margin shown in the experiments)



$S = 2,\ \kappa = 9$



$S = 10,\ \kappa = 22028$

The Plan for Today

- Logistic Bandits 101
- New confidence set for logistic bandits via (online) regret-to-confidence-set (**O2CS**)
- Improved Regrets for Logistic Bandits
- Conclusion and Future Works

~~~

We propose a framework in which one can construct a confidence set using an *achievable* online learning regret bound (*without ever running the alg*), and apply it to improve the regret bounds of (multinomial) logistic bandits.

~~~

Conclusion and Future Works

Conclusion

1. **Regret-to-confidence-set conversion (R2CS):** a new framework that converts an *achievable* online learning regret guarantee to a confidence set, without ever running the online algorithm explicitly.
2. We apply R2CS to obtain tightest confidence set for logistic losses, which then leads to the state-of-the-art regret guarantee of logistic bandits.
3. We empirically show that our new confidence-set based UCB algorithm attains the best performance.

Omitted from this presentation:

- Guarantees for multinomial logistic loss/bandits
- Extensive discussions on other related work
- and more

Conclusion and Future Works

Future Works

1. Extend our R2CS framework to various settings such as sparse logistic bandits [Oh et al., ICML'21], generalized linear bandits [Filippi et al., NIPS'10], norm-agnostic scenario [Gales et al., AISTATS'22], and multinomial logistic MDPs [Hwang & Oh, AAAI'23].
2. As logistic bandits can be seen as utility-based dueling bandits with top-1 feedback (Bradley-Terry model), apply our analysis to make the recent guarantees on RLHF [Wu & Sun, ICLR'24] tighter?
3. Any relation to Thompson sampling [Abeille & Lazaric, 2017]?
4. Any relation to universal inference [Wasserman et al., 2020] and sequential likelihod ratio confidence set [Emmenegger et al., NeurIPS'23]?
5. Any relation to Decision Estimation Coefficients [Foster et al., arXiv'21]?
6. …etc!!

Thank you for your attention!

(arXiv will be updated with camera-ready ver soon)

References

Logistic Bandits

J. Ville. Étude critique de la notion de collectif. *Monographies des Probabilités*. Paris: Gauthier-Villars, 1939.

D. A. Freedman. On Tail Probabilities for Martingales. *The Annals of Probability,* 3(1):100-118, 1975.

E. Hazan, Z. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning,* 69(2):169-192, 2007.

F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics,* 4(none):384-414, 2010.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A Contextual-Bandit Approach to Personalized News Article Recommendation. In *WWW 2010.*

L. Li, W. Chu, J. Langford, T. Moon, and X. Wang. An Unbiased Offline Evaluation of Contextual Bandit Algorithms with Generalized Linear Models. In *ICML 2011 Workshop on On-line Trading of Exploration and Exploitation.*

A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. Schapire. Contextual Bandit Algorithms with Supervised Learning Guarantees. In *AISTATS 2011.*

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved Algorithms for Linear Stochastic Bandits. In *NeurIPS 2011.*

D. J. Foster, S. Kale, H. Luo, M. Mohri, and K. Sridharan. Logistic Regression: The Importance of Being Improper. In *COLT 2018.*

R. Jézéquel, P. Gaillard, and A. Rudi. Efficient improper learning for online logistic regression. In *COLT 2020.*

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

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

L. Faury, M. Abeille, K.-S. Jun, and C. Calauzènes. Jointly Efficient and Optimal Algorithms for Logistic Bandits. In *AISTATS 2022.*

References

Related Work, Future Work

O. Dekel, C. Gentile, and K. Sridharan. Robust selective sampling from single and multiple teachers. In *COLT 2010.*

C. Gentil and F. Orabona. On Multilabel Classification and Ranking with Bandit Feedback. *Journal of Machine Learning Research,* 15(70):2451-2487, 2014.

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

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Online-to-Confidence-Set Conversions and Applications to Sparse Stochastic Bandits. In *AISTATS 2012*.

L. Zhang, T. Yang, R. Jin, Y. Xiao, and Z. Zhou. Online Stochastic Linear Optimization under One-bit Feedback. In *ICML 2016*.

M. Abeille and A. Lazaric. Linear Thompson camping revisited. *Electronic Journal of Statistics,* 11(2):5165-5197, 2017.

K.-S. Jun, A. Bhargava, R. Nowak, and R. Willett. Scalable Generalized Linear Bandits: Online Computation and Hashing. In *NeurIPS 2017.*

A. Rakhlin and K. Sridharan. On Equivalence of Martingale Tail Bounds and Deterministic Regret Inequalities. In *COLT 2017*.

K.-S. Jun and F. Orabona. Parameter-Free Online Convex Optimization with Sub-Exponential Noise. In *COLT 2019*.

D. J. Foster and A. Rakhlin. Beyond UCB: Optimal and Efficient Contextual Bandits with Regression Oracles. In *ICML 2020.*

L. Wasserman, A. Ramdas, and S. Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880-16890, 2020.

M. Oh, G. Iyengar, and A. Zeevi. Sparsity-Agnostic Lasso Bandit. In *ICML 2021*.

D. J. Foster, S. Kakade, J. Qian, and A. Rakhlin. The Statistical Complexity of Decision Making. In *arXiv preprint arXiv:2112.13487*.

S. B. Gales, S. Sethuraman, and K.-S. Jun. Norm-Agnostic Linear Bandits. In *AISTATS 2022*.

F. Orabona and K.-S. Jun. Tight Concentrations and Confidence Sequences from the Regret of Universal Portfolio. *IEEE Transactions on Information Theory*, 70(1):436-455, 2023.

T. Hwang and M. Oh. Model-Based Reinforcement Learning with Multinomial Logistic Function Approximation. In *AAAI 2023*.

G. Lugosi and G. Neu. Online-to-PAC Conversions: Generalization Bounds via Regret Analysis. In *arXiv preprint arXiv:2305.19674*.

N. Emmenegger, M. Mutný, and A. Krause. Likelihood Ratio Confidence Sets for Sequential Decision Making. In *NeurIPS 2023*.

R. Wu and W. Sun. Making RL with Preference-based Feedback Efficient via Randomization. In *ICLR 2024*.