Confidence Set Analysis in Preference Feedbacks

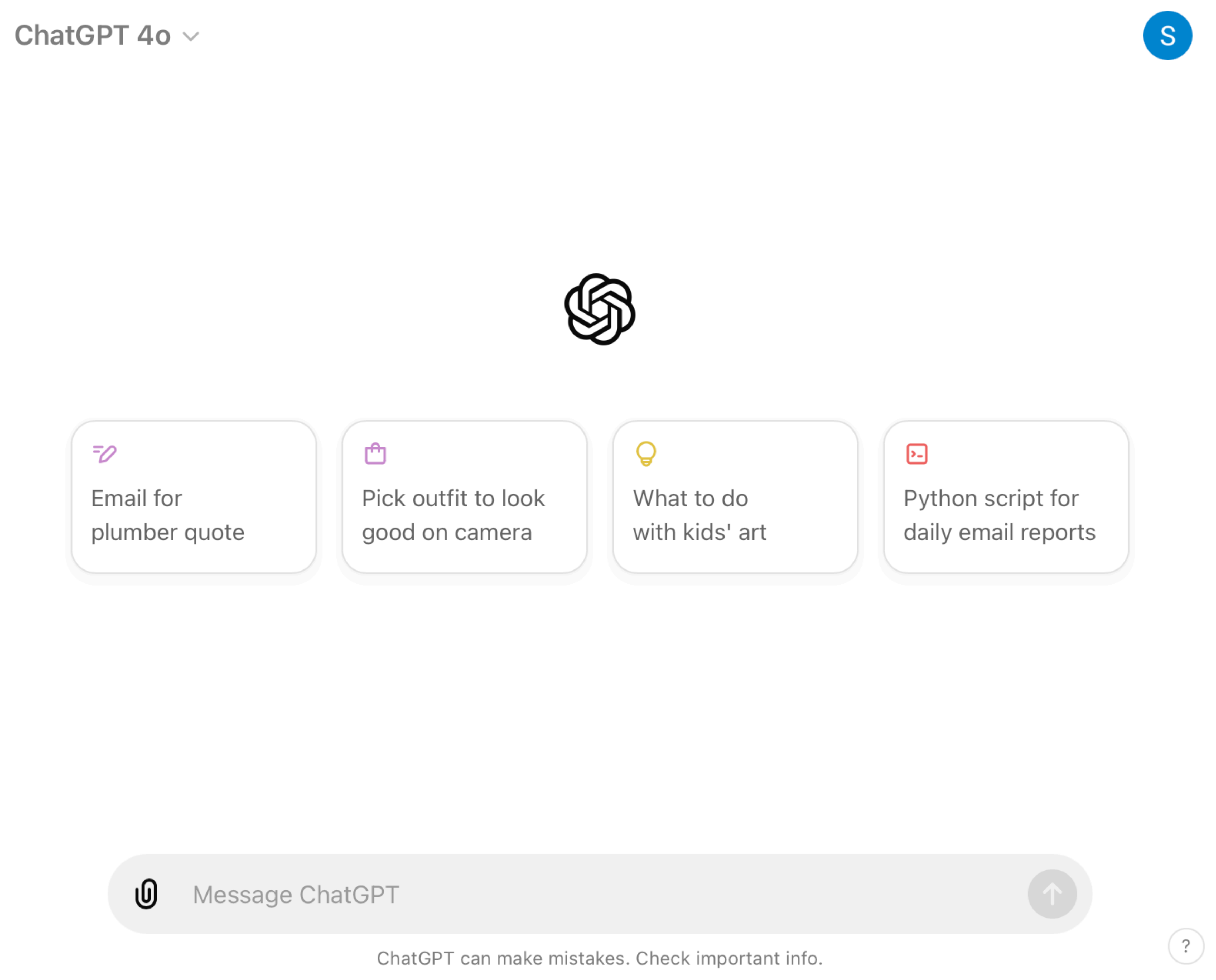
Se-Young Yun (KAIST AI)

ChatGPT

AI virtual assistant

ChatGPT

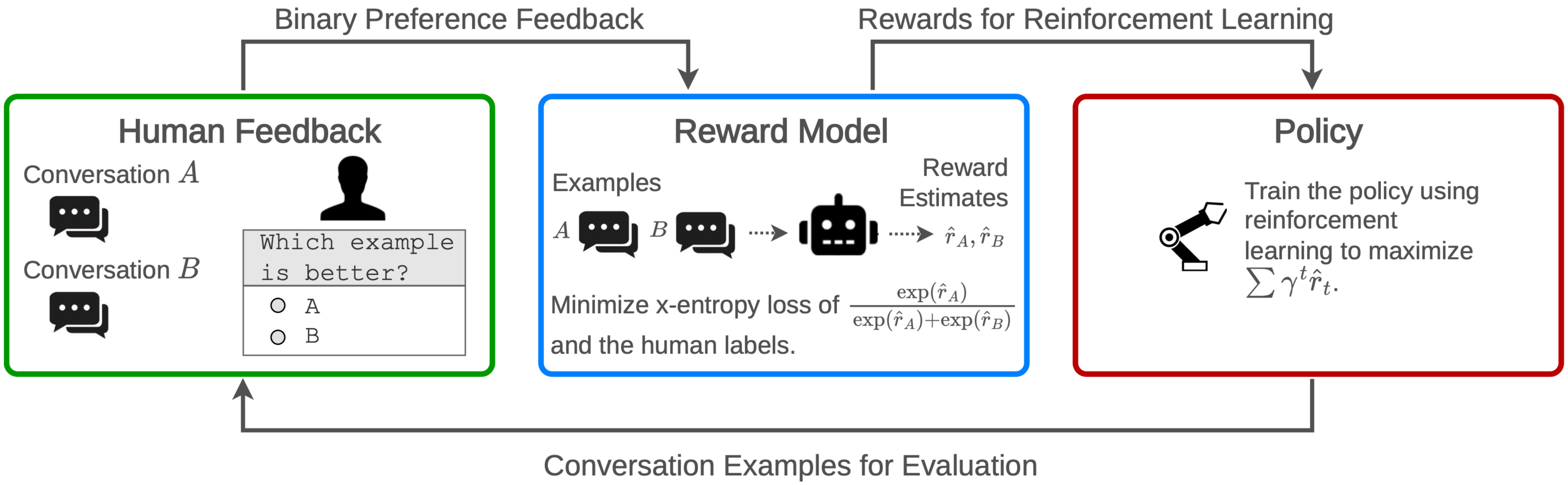
- The most successful AI service
- Strives to generate answers that align with users' intentions
- Preference Alignment from Human Feedback!!!



RLHF

Reinforcement Learning from Human Feedback

RLHF: a key ingredient of recent success of LLMs

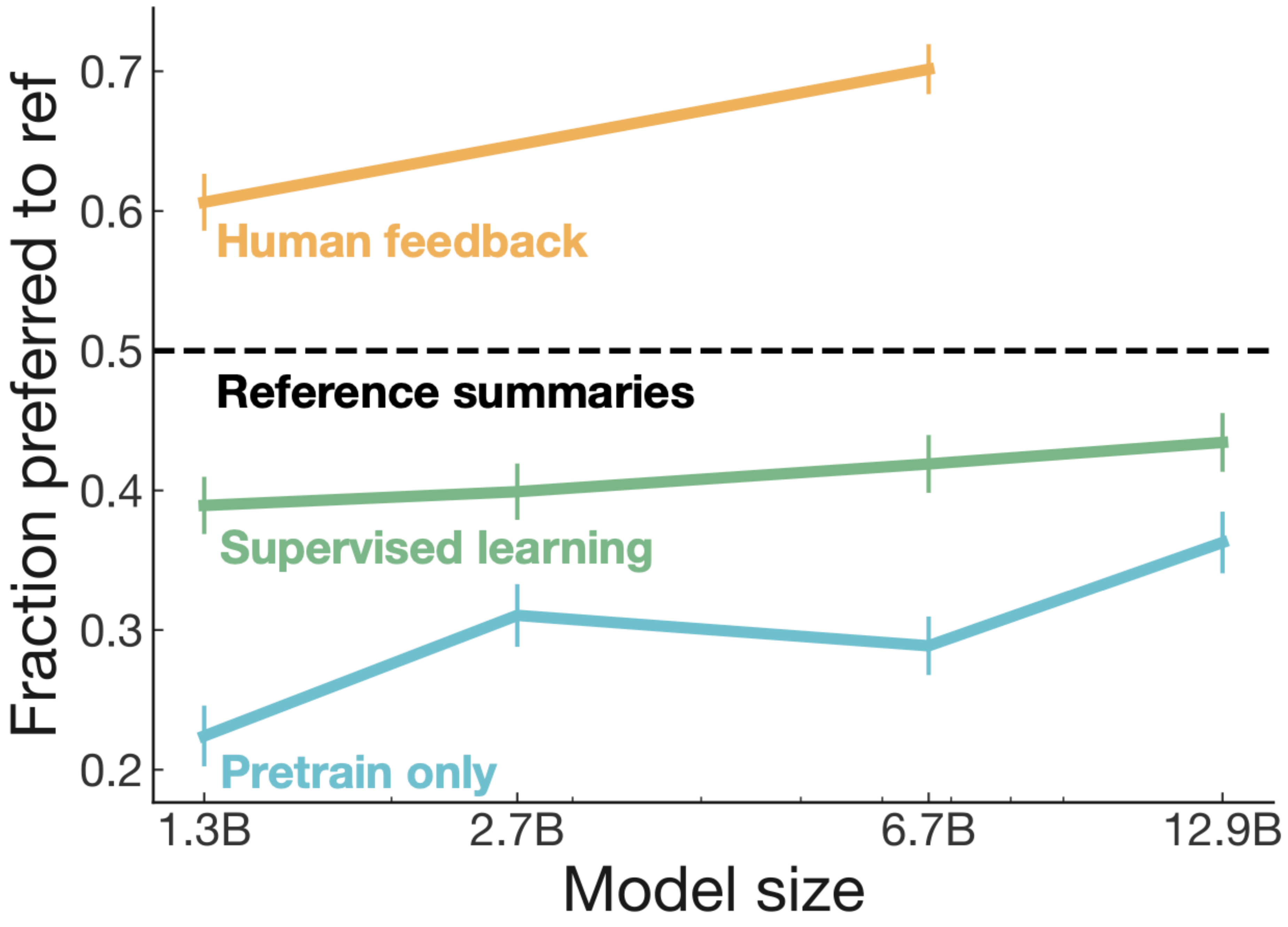


Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback

RLHF's Efficiency

RLHF significantly outperforms baselines

Ex> English Summarization Task



Stiennon et al "Learning to summarize from human feedback"

Bradley-Terry Model

Probability model for pairwise comparisons

Bradley-Terry Model: a probability model for the outcome of pairwise comparisons

$$\mathbb{P}(i > j) = \frac{e^{r_i}}{e^{r_i} + e^{r_j}} = \frac{1}{1 + e^{-(r_i - r_j)}}$$

- The probability that item i wins against item j is represented using reward scores r_i and r_j
- RLHF learns reward scores using the Bradley-Terry model

$$\arg\min_{r \in \mathcal{R}} - \sum_{t=1}^{T} \log \frac{e^{r_{w_t}}}{e^{r_{w_t}} + e^{r_{l_t}}}$$

  - w_t, l_t are the winner index and the loser index at the t-th comparison
- **Questions: Uncertainty? Confidence? Reward Modes? Other Probability Models?**

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

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

KAIST AI
Kim Jaechul Graduate School

Optimization and Statistical Inference LAB

OptiML
Optimization & Machine Learning

Computer Science

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

Linear Contextual Bandit

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 $r_t \sim \text{Ber}(\langle x_t, \theta_\star \rangle)$
   - $\theta_\star$ is unknown to the learner

Goal:

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

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:

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

Logistic Bandits 101

Preference Feedback

Preference Feedback:

- The agent selects a tuple (x, a, a′) to present to a human labeller
- Some papers consider a linear reward model $r_{\theta_\star}(x, a) = \langle \phi(x, a), \theta_\star \rangle$ where ϕ is a known feature map
- The preference feedback follows the Bernoulli response such that a is preferred over a' with probability $\mu(\langle \phi(x, a) - \phi(x, a'), \theta_\star \rangle)$

Goal:

- **How to find $\theta_\star$ accurately within a given labeling budget?**
- **How to define a good confidence range of $\theta_\star$?**

Logistic Bandits 101

Assumptions

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

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:

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

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

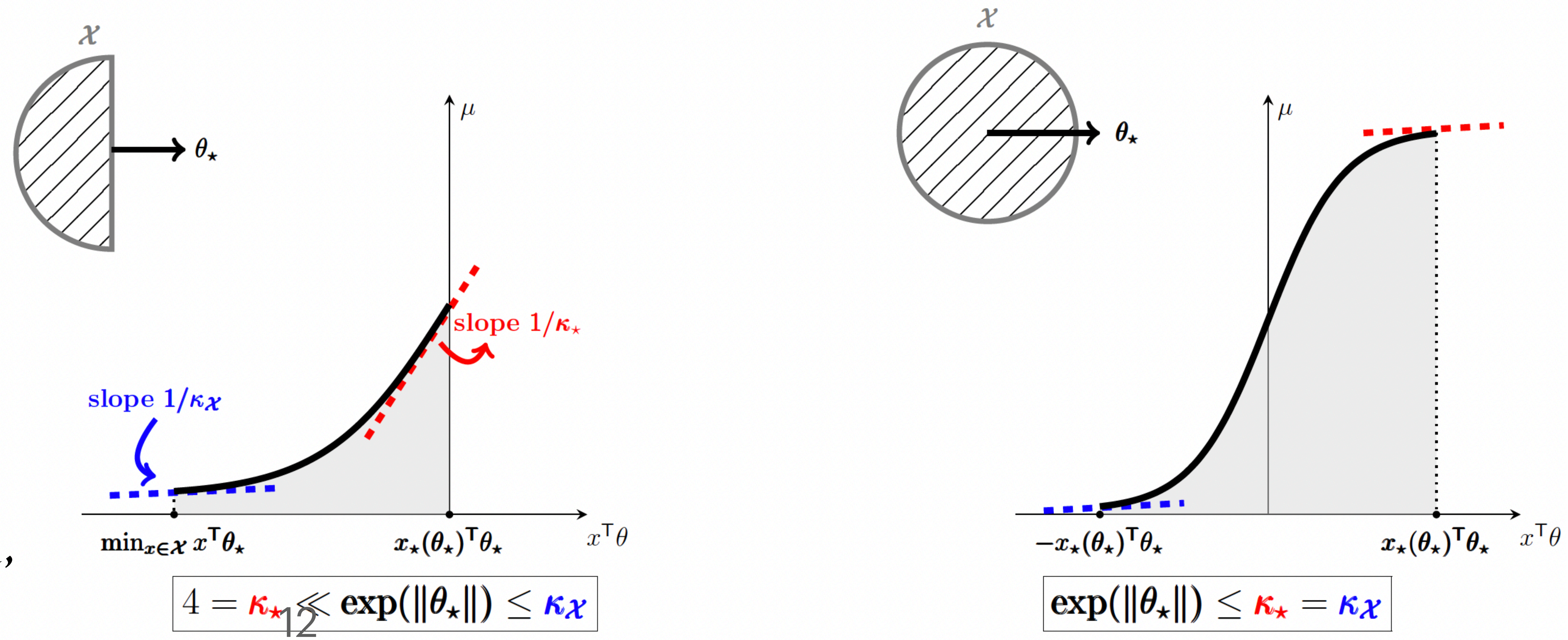
Logistic Bandits 101

$d\sqrt{T/\kappa_\star(T)}$ 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

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

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

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

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

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

Can we construct tighter (improved dependency in S) *loss-based confidence set*??

Logistic Bandits 101

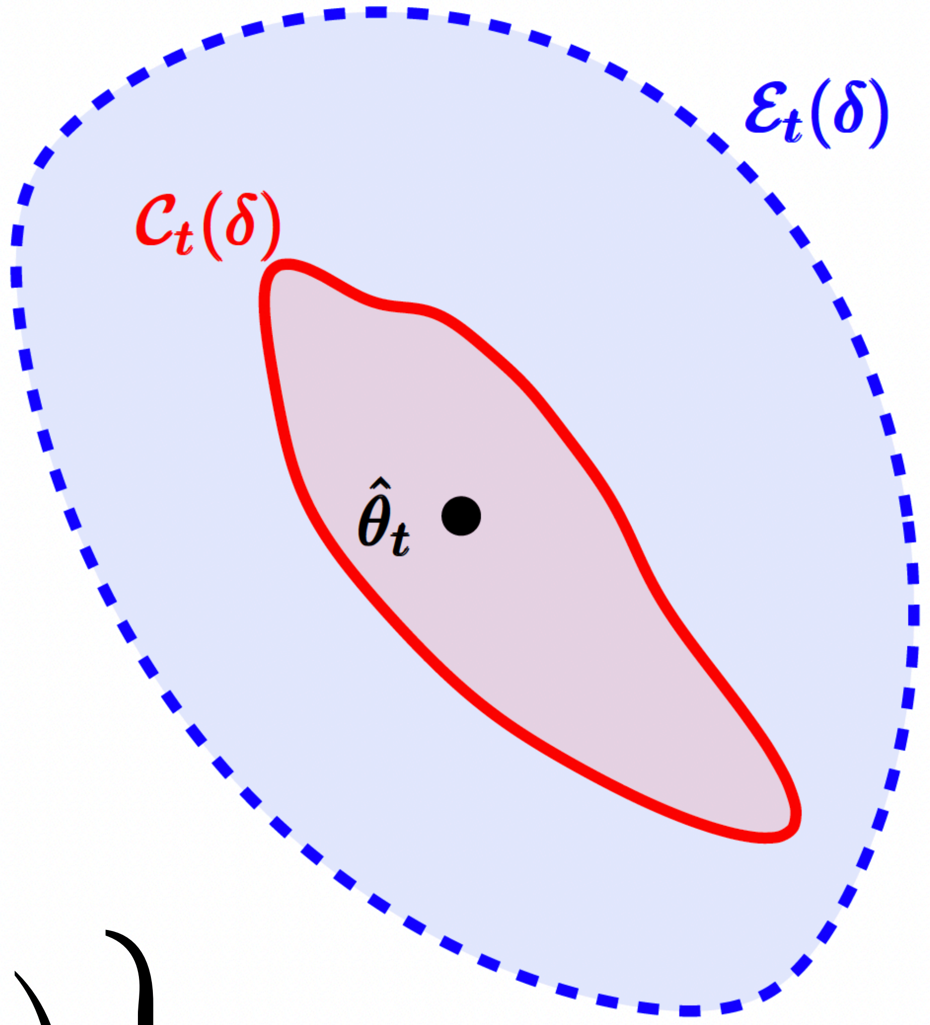
More details in OFULog(-r)

- OFULog and OFULog-r are of the following form:

1. Solve $\widehat{\theta}_t = \text{argmin}_{\theta \in \mathbb{R}^d} \left[\mathcal{L}_t(\theta) \triangleq \sum_{s=1}^{t-1} \ell_s(\theta) + \lambda_t \|\theta\|_2^2 \right]$, 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 a reward r_t

Logistic Bandits 101

More details in OFULog(-r)



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

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

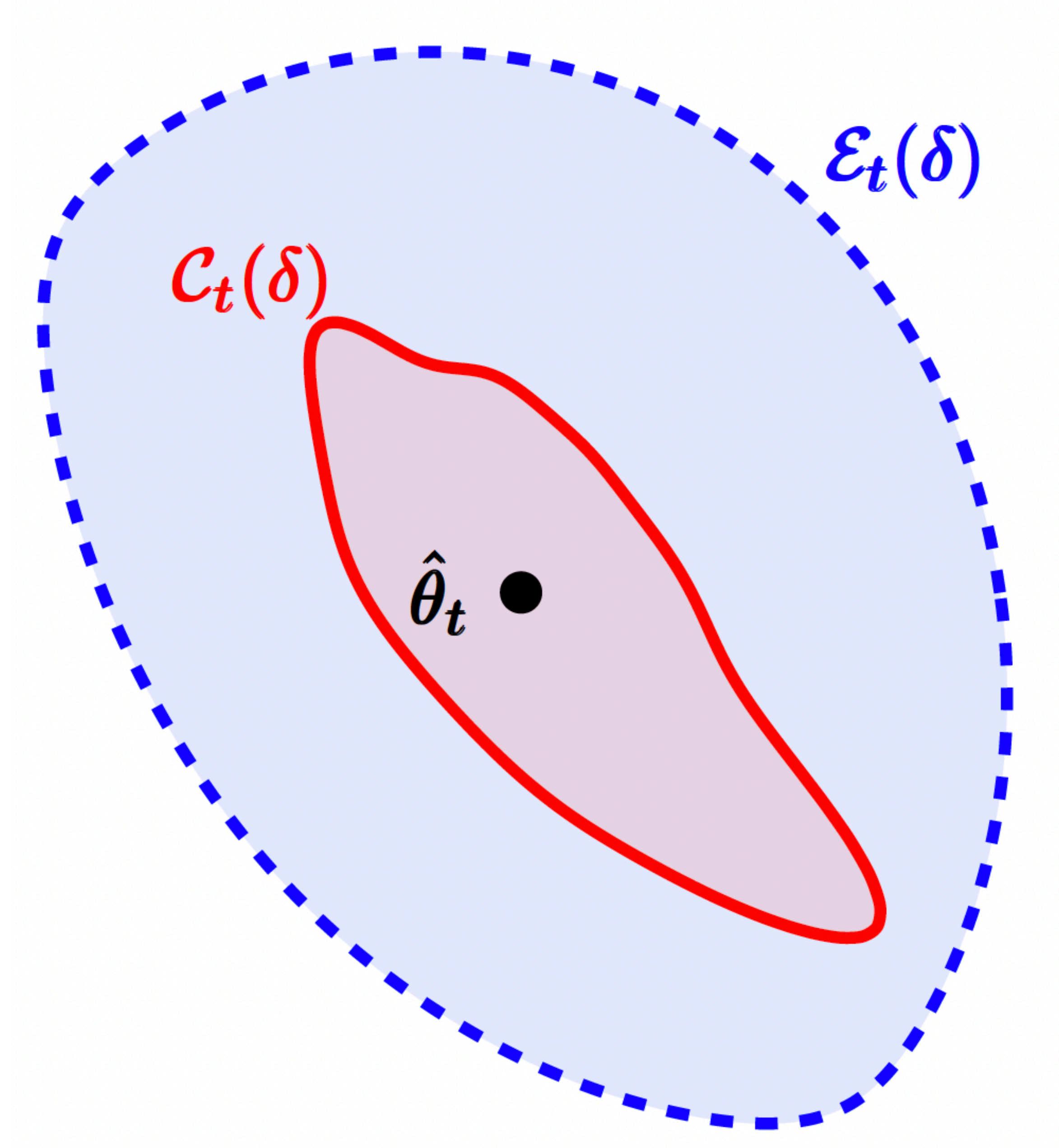
- **OFULog-r** **[Abeille et al., AISTATS'21]:**

$$\mathcal{E}_t(\delta) := \left\{ \theta \in \mathbb{B}^d(S) : \mathcal{L}_t(\theta) - \mathcal{L}_t(\widehat{\theta}_t) \leq \mathcal{O}\left(\sqrt{dS^3 \log t}\right) \right\}$$

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

Logistic Bandits 101

Gradient and Confidence set



- **Gradient**
  - $$\nabla \mathscr{L}_t(\theta_\star) = \underbrace{\sum_{s=1}^{t-1} \left(\mu(\langle x_s, \theta_\star \rangle) - r_s\right) x_s}_{\text{Martingale Sum}} + 2\lambda_t \theta$$
  - The gradient at $\theta_\star$ should be near zero!
  - The confidence check can be done with the inverse of Hessian (covariance)
    $$H(\theta_\star) = \sum_{s=1}^{t-1} \dot{\mu}(\langle x_s, \theta_\star \rangle) x_s x_s^\top + \lambda I$$
  - However, we should compute gradient and Hessian for all θ
  - Gradient -> Loss conversion can formulate a convex confidence set, albeit not a tightly bound one

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 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 $\tilde{\theta}_s$:

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

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

- $\mathrm{Reg}^O(t)$ is the online regret up to time t, and $\zeta(t)$ 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 while $\tilde{\theta}_s$ is online

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

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*.

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

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

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!

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 prediction problem 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!

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

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

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

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

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

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

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

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

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

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

New Confidence Set!

Likelihood Ratio-Based Confidence Set

- Let L_t be the Lipschitz constant of $\mathscr{L}_t$ which is bounded above by $(1 + \frac{S}{2})(t-1)$

Theorem 3. **[Lee et al., pre-print]** We have $\mathbb{P}\left[\forall t \geq 1,\ \theta_\star \in C_t(\delta)\right] \geq 1 - \delta$, where

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

$$\beta_t(\delta) := 1 + \log\frac{1}{\delta} + d\log\frac{2SL_t}{d}$$

Remove S dependency!

Martingale Log-Likelihood

Proof Sketch of Theorem 3

Let $M_t(\theta) = \exp\big(\mathscr{L}_t(\theta_\star) - \mathscr{L}_t(\theta)\big)$. Then, it is easy to check $M_t(\theta)$ is a non-negative Martingale.

Lemma. For any data-independent prior $\mathbb{Q}$, the following holds:

$$\mathbb{P}\left(\exists t : \mathbb{E}_{\theta\sim\mathbb{Q}}[M_t(\theta)] \geq \log\frac{1}{\delta}\right) \leq \delta$$

Time-Uniform PAC-Bayesian Bound

Proof Sketch of Theorem 3

We follow the usual recipes for deriving time-uniform PAC-Bayesian bound (Alquier, 2024; Chugg et al., 2023):

Lemma. For any data-independent prior $\mathbb{Q}$ and any sequence of adapted posterior distributions (possibly learned from the data) $\{\mathbb{P}_t\}$, the following holds:

$$\mathbb{P}\left(\exists t : \mathscr{L}_t(\theta_\star) - \mathbb{E}_{\theta\sim\mathbb{P}_t}[\mathscr{L}_t(\theta)] \geq \log\frac{1}{\delta} + D_{KL}(\mathbb{P}_t \| \mathbb{Q})\right) \leq \delta$$

Our novelty is the choice of $\mathbb{Q}$ and $\{\mathbb{P}_t\}$

$$\mathbb{Q} = Unif(\Theta), \quad \mathbb{P}_t = Unif(\widetilde{\Theta}_t \triangleq (1-c)\widehat{\theta}_t + c\Theta)$$

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., pre-print 24]. Tight loss-based confidence set

$$d\sqrt{\frac{T}{\kappa_\star(T)}} + d^2\kappa_{\mathcal{X}}(T)$$

with Many New Tricks

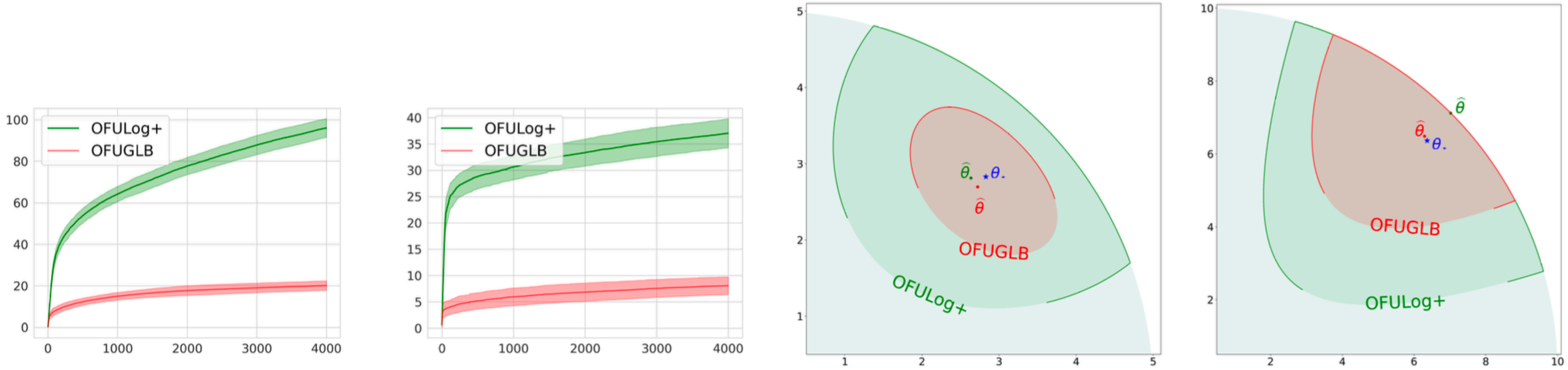
Proof Sketch of Regret

- **regret :** $\sum_{t}^{T} \mu(x_{t,\star}^{\top}\theta_\star) - \mu(x_t^{\top}\theta_\star)$
- **Upper Bound :** $\sum_{t}^{T} \mu(x_t^{\top}v_t) - \mu(x_t^{\top}\hat{\theta}_t)$ **where** v_t **is the point maximizing the gap in confidence set**
- **Taylor:** $\sum_{t}^{T} \dot{\mu}(x_t^{\top}\hat{\theta}_t)x_t^{\top}(v_t - \hat{\theta}_t)$
- **With** $H_t = \sum_{i=1}^{T} \dot{\mu}(x_i^{\top}\hat{\theta}_i)x_t x_t^{\top}$, **Cauchy-Schwartz** $\sum_{t}^{T} \dot{\mu}(x_t^{\top}\hat{\theta}_t)\|x_t\|_{H_t^{-1}}\|v_t - \hat{\theta}_t\|_{H_t}$
- $\|v_t - \hat{\theta}_t\|_{H_t}$ **is bounded by the confidence radius**
- **Cauchy-Schwartz** $\sum_{t}^{T} \dot{\mu}(x_t^{\top}\hat{\theta}_t)\|x_t\|_{H_t^{-1}} \leq \sqrt{\sum_{t}^{T} \dot{\mu}(x_t^{\top}\hat{\theta}_t)}\sqrt{\sum_{t}^{T} \dot{\mu}(x_t^{\top}\hat{\theta}_t)\|x_t\|_{H_t^{-1}}^2}$
- **EPL can conclude this proof...**

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!!



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

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. PAC-Bayesian Bound can further enhance the confidence set!
4. We empirically show that our new confidence-set based UCB algorithm attains the best performance.

Thank You