NeurIPS 2024
The 38th Annual Conference on Neural Information Processing Systems

A Unified Confidence Sequence for Generalized Linear Models, with Applications to Bandits

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

KAIST AI
Kim Jaechul Graduate School

Optimization and Statistical Inference LAB

OptiML
Optimization & Machine Learning

Computer Science

Online Learning and Bandits

An Introduction

- The learner *sequentially* interacts with the environment, with *limited feedback*
- The goal is to **adapt to the environment in a very fast manner!**

contextual vs. non-contextual

- for $t = 1,2,\cdots,T$
  - an action set $\mathscr{A}_t$, possibly with other contextual information $\mathscr{X}_t$ are revealed to the learner
  - learner chooses some action $a_t \in \mathscr{A}_t$ **possibly dependent on the previous history!**
  - environment reveals a reward $r_t = r_t(a_t)$

bandit feedback

  - environment (partially) reveals $r_t(\cdot)$

semi-bandit/full feedback

(~online learning)

Online Learning and Bandits

Real-world applications

- **Clinical trials**
- Recommender systems (news, advertisement, etc)
- Resource allocation (e.g., wireless networks, routing)
- Social network influence maximization
- Navigation system, Shortest path routing
- ....etc

Two Types of Bandits

Stochastic and Adversarial

- **Stochastic bandits.**
  - r_t follows a fixed distribution, i.e., for each $a \in \mathcal{A}, \;\; r_t(a) \,|\, \sigma(\mathcal{H}_{t-1}) \sim \mathcal{D}_a$
  - Here, $\mathcal{H}_{t-1} := (a_1, r_1, \cdots, a_{t-1}, r_{t-1})$ is the history up to previous time
  - Usually, this can be rewritten as $r_t(a) = \mu_a + \eta_{t,a}$, where $\eta_{t,a}$ is a *martingale difference noise*
  - There are two main goals in stochastic bandits: **regret minimization** and **pure exploration**

- **Adversarial bandits.** — not considered in this talk
  - The environment ("adversary") *arbitrarily* chooses $(r_1(\cdot), r_2(\cdot), \cdots, r_T(\cdot))$ in advance!
  - The learner then plays against the adversary (~ two-player zero-sum game) ==> randomisation!!

Multi-armed Bandits

Most Basic Bandit Setting!

- $\mathscr{A} = \{a_1, a_2, \cdots, a_K\},\ K < \infty,$ suppose $\mu_{a_1} \leq \cdots \leq \mu_{a_{K-1}} < \mu_{a_K} =: \mu_\star$.
- **Suboptimality gap:** $\Delta_a := \mu_\star - \mu_a$ ~ *difficulty of the bandit instance!*

- For K-armed bandits, we have the following **Regret decomposition lemma:**

$$\mathrm{Reg}^\pi(T) = \sum_{a \in \mathscr{A}} \Delta_a \mathbb{E}[N_a(T)], \quad N_a(T) := \sum_{t=1}^{T} 1[a_t = a]$$

- In other words, we need to look out for *number of pulls of suboptimal arms!!*

Multi-armed Bandits

Regret lower bounds

- A policy π is **consistent** if $\text{Reg}^\pi(T) = o(T^\alpha),\ \forall \alpha > 0$.
- **Instance-wise Lower Bound (Lai & Robbins, 1985).** For any consistent π,
$$\liminf_{T\to\infty} \frac{\text{Reg}^\pi(T)}{\log T} \gtrsim \sum_{a\in\mathcal{A},\Delta_a>0} \frac{1}{\Delta_a}$$
- **Minimax Lower Bound (Vogel, 1960).** For unit variance Gaussian K-armed bandits,
$$\min_\pi \max_B \text{Reg}^\pi(T;B) \geq \frac{1}{27}\sqrt{(K-1)T}\,.$$
    - **pf.** *change-of-measure, Le-Cam's method, Bregtanolle-Huber inequality!! (~ info theory, nonparametric statistics)*

Multi-armed Bandits

Optimism Principle for Stochastic Bandits and UCB (Auer et al., Mach. Learn. 2002)

- *Exploration* ~ try to ***estimate*** the environment as efficiently as possible
  => ***constructing some "confidence sequence"***
- *Exploitation* ~ *"act as if our estimates are as nice as **plausibly possible**"*
  => ***Optimism in the Face of Uncertainty (OFU)***

exploration bonus for arms not pulled sufficiently enough

Upper Confidence Bound (UCB) Algorithm:

$$a_t = \text{argmax}_{a \in \mathcal{A}, \left\{\mu_{a'} \in \mathcal{C}_{a',t},\ \forall a' \in \mathcal{A}\right\}} \mu_a = \text{argmax}_{a \in \mathcal{A}} \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta_t)}{N_a(t-1)}}$$

$$\mathcal{C}_{a',t} := \left\{ \mu_{a'} : \mu_{a'} \leq \hat{\mu}_{a'}(t-1) + \sqrt{\frac{2\log(1/\delta_t)}{N_a(t-1)}} \right\}$$

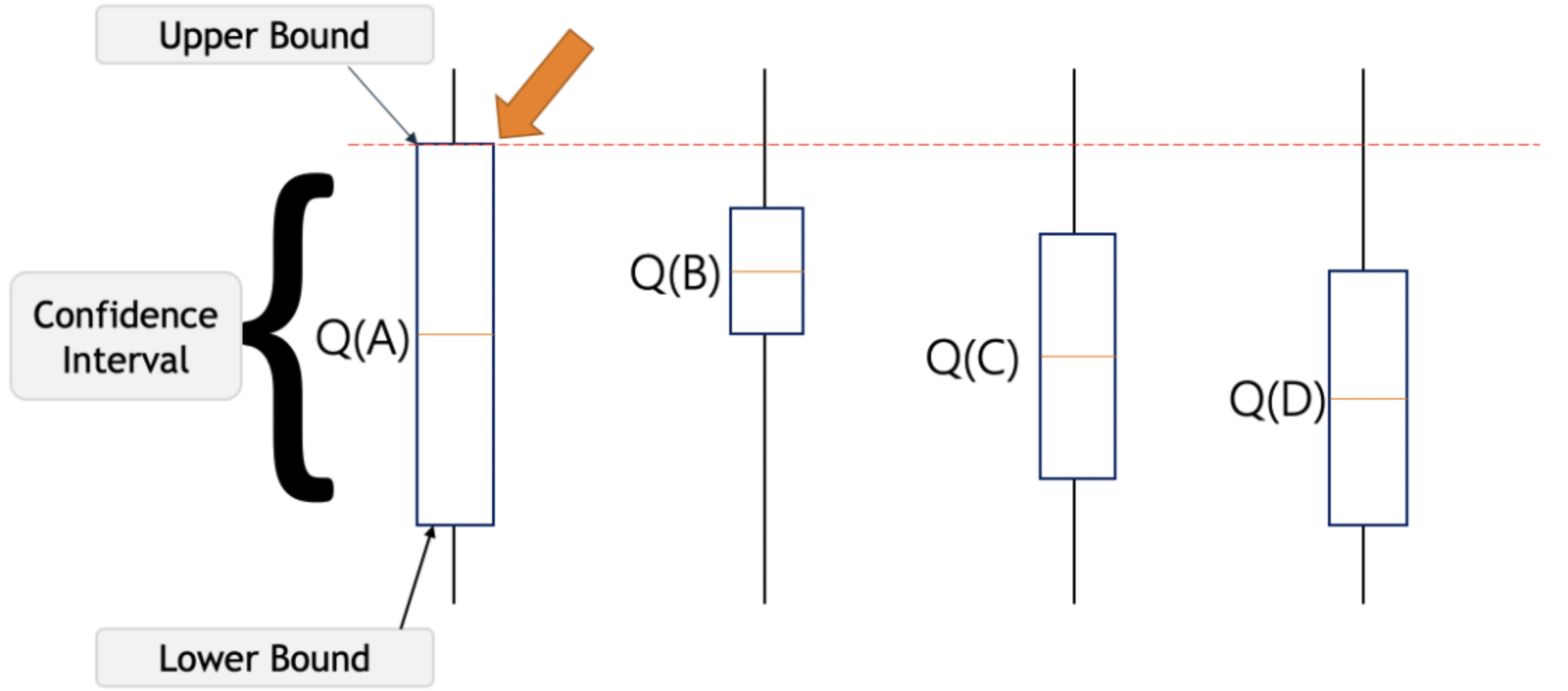
Multi-armed Bandits

Optimism Principle for Stochastic Bandits and UCB (Auer et al., Mach. Learn. 2002)

Upper Confidence Bound (UCB) Algorithm:

$$a_t = \text{argmax}_{a \in \mathscr{A}} \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta_t)}{N_a(t-1)}}$$

$$\mathscr{C}_{a',t} := \left\{ \mu_{a'} : \mu_{a'} \le \hat{\mu}_{a'}(t-1) + \sqrt{\frac{2\log(1/\delta_t)}{N_a(t-1)}} \right\}$$



Regret of UCB (Auer, 2002). With $\delta_t^{-1} = 1 + t(\log t)^2$,

$$\text{Reg}^{UCB}(T) \lesssim \sum_{a \in \mathscr{A}, \Delta_a > 0} \frac{\log T}{\Delta_a}$$

Instance-wise asymptotically optimal!

(recall our lower bound)

Linear Bandits

Auer (Mach. Learn. 2002); Dani, Hayes, and Kakade (COLT'08)

- $\mathscr{A} \subset \mathbb{R}^d$ that is compact and possibly infinite!
- **Linear realizability.** There exists a fixed $\theta_\star \in \mathscr{B}^d(S)$ such that $r_t(a) = \langle \theta_\star, a \rangle + \eta_{t,a}$

- This can be interpreted as *contextual linear bandit! (Chu et al., AISTATS'11)*
  - The learner observes a **context vector** $x_{a,t} \in \mathbb{R}^d$ for each action $a \in [K]$
  - **Linear realizability.** $r_t(a) = \langle \theta_\star, x_{t,a} \rangle + \eta_{t,a}$, with $\mathbb{E}[\eta_{t,a} \mid x_{t,a}] = 0$

- **Minimax regret lower bounds.** $\Omega(d\sqrt{T})$ $(|\mathscr{A}| \le \infty)$ $\Omega(\sqrt{dT})$ $(|\mathscr{A}| = K < \infty)$

LinUCB/OFUL: OFU for Linear Bandits

Chu, Li, Reyzin, and Schapire (AISTATS'11); Abbasi-Yadkori, Pal, and Szepesvari (NIPS'11)

- Estimate mean of each arm ==> **Estimate** $\theta_\star$ ~ ***confidence sequence (CS)***

  => *A random sequence of sets* $\left\{\mathscr{C}_t(\delta)\right\}_{t\geq 1}$ *s.t.* $\mathbb{P}\left(\exists t \geq 1 : \theta_\star \notin \mathscr{C}_t(\delta)\right) \leq \delta$

- **Theorem (*Elliptical* CS for linear bandits).**

  $\mathscr{C}_t(\delta) := \left\{\theta : \|\theta - \hat{\theta}_t\|_{V_t} \lesssim \beta_t(\delta) \triangleq \sqrt{\log\frac{1}{\delta} + d\log\left(1 + \frac{ST}{d}\right)}\right\}$, where

  $V_t := \frac{1}{S^2}I_d + \sum_{s=1}^{t-1} x_s x_s^\top$ is the **design matrix** and $\hat{\theta}_t := V_t^{-1}\sum_{s=1}^{t-1} r_s x_s$ is the **(regularized) MLE.**

- *Pf.* ***self-normalized vector martingale*** *(Method of mixtures, supermartingale construction)*

LinUCB/OFUL: OFU for Linear Bandits

Chu, Li, Reyzin, and Schapire (AISTATS'11); Abbasi-Yadkori, Pal, and Szepesvari (NIPS'11)

- Recall the UCB for K-armed bandits:

$$a_t = \mathrm{argmax}_{a \in \mathcal{A}, \{\mu_{a'} \in \mathcal{C}_{a',t}, \forall a' \in \mathcal{A}\}} \mu_a = \mathrm{argmax}_{a \in \mathcal{A}} \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta_t)}{N_a(t-1)}}$$

- Take the first formulation and convert it to our linear bandit setting:

$$x_t = \mathrm{argmax}_{a \in \mathcal{A}, \theta \in \mathcal{C}_t(\delta)} \langle a, \theta \rangle$$ <= **LinUCB/OFUL**

- Thanks to the ellipsoidal form, above can be *equivalently* rewritten as follows:

$$x_t = \mathrm{argmax}_{a \in \mathcal{A}} \langle x_a, \widehat{\theta}_t \rangle + \beta_t(\delta) \|x_a\|_{V_t^{-1}}$$

exploration bonus for arms not pulled sufficiently enough

LinUCB/OFUL: OFU for Linear Bandits

Chu, Li, Reyzin, and Schapire (AISTATS'11); Abbasi-Yadkori, Pal, and Szepesvari (NIPS'11)

- **Regret of OFUL.** $\mathcal{O}(d\sqrt{T}\log T)$ **for** $|\mathcal{A}| \leq \infty$**,**
  - **pf.** Relies on the *confidence sequence* + *Cauchy-Schwartz* + *elliptical potential lemma*

Lemma 11. *Let* $\{X_t\}_{t=1}^{\infty}$ *be a sequence in* $\mathbb{R}^d$, V *a* $d \times d$ *positive definite matrix and define* $\overline{V}_t = V + \sum_{s=1}^{t} X_s X_s^\top$. *Then, we have that*

$$\log\left(\frac{\det(\overline{V}_n)}{\det(V)}\right) \leq \sum_{t=1}^{n} \|X_t\|^2_{\overline{V}_{t-1}^{-1}} .$$

Further, if $\|X_t\|_2 \leq L$ *for all* t*, then*

$$\sum_{t=1}^{n} \min\left\{1, \|X_t\|^2_{\overline{V}_{t-1}^{-1}}\right\} \leq 2(\log\det(\overline{V}_n) - \log\det V) \leq 2(d\log((\mathrm{trace}(V) + nL^2)/d) - \log\det V) ,$$

and finally, if $\lambda_{\min}(V) \geq \max(1, L^2)$ *then*

$$\sum_{t=1}^{n} \|X_t\|^2_{\overline{V}_{t-1}^{-1}} \leq 2\log\frac{\det(\overline{V}_n)}{\det(V)} .$$

- cf. **Regret of SupLinUCB.** $\mathcal{O}(\sqrt{dT\log(KT)})$ **for** $|\mathcal{A}| = K < \infty$

This is a **elimination-based algorithm**

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]

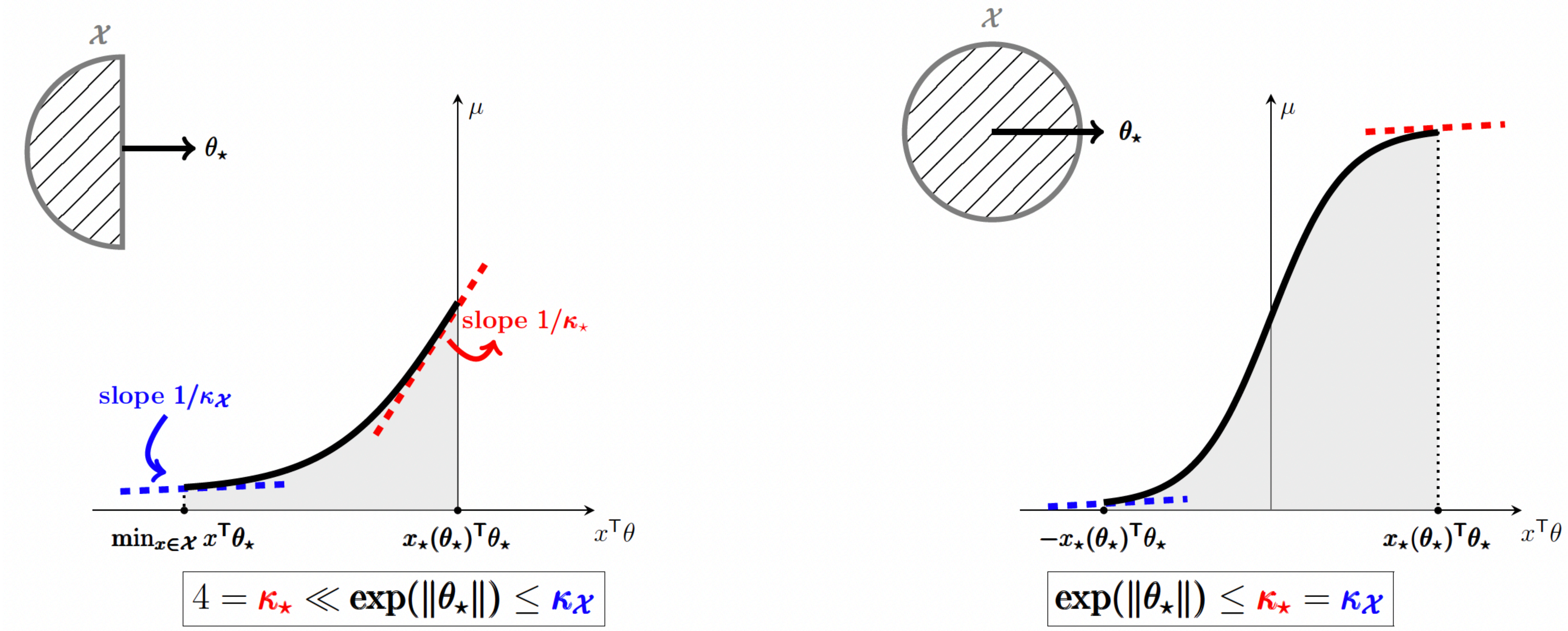
Logistic Bandits 101

$d\sqrt{T/\kappa_\star(T)}$ is minimax optimal (taken from L. Faury's slides)

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:

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

- More linear (smaller $\dot{\mu}$), 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

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

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

Generalized Linear Models

Problem Setting

Generalized Linear Models

Problem Setting

Consider the **Generalized Linear Model (GLM)**:

$$dp(r\,|\,x;\theta_\star) = \exp\left(\frac{r\langle x,\theta_\star\rangle - m(\langle x,\theta_\star\rangle)}{g(\tau)} + h(r,\tau)\right) d\nu,$$

with dispersion parameter $\tau > 0$, base measure ν, **context** $x \in X$, and **unknown parameter** $\theta_\star \in \Theta$.

Generalized Linear Models

Problem Setting

Consider the **Generalized Linear Model (GLM)**:

$$dp(r \mid x; \theta_\star) = \exp\left(\frac{r\langle x, \theta_\star \rangle - m(\langle x, \theta_\star \rangle)}{g(\tau)} + h(r, \tau) \right) d\nu,$$

with dispersion parameter $\tau > 0$, base measure ν, **context** $x \in X$, and **unknown parameter** $\theta_\star \in \Theta$.

Assumptions. $X \subseteq \mathbb{B}^d(1)$, $\varnothing \neq \Theta \subseteq \mathbb{B}^d(S)$, Θ compact & convex, $m(\cdot)$ is convex and three-times differentiable.

Properties. $\mathbb{E}[r \mid x, \theta_\star] = m'(\langle x, \theta_\star \rangle) =: \mu(\langle x, \theta_\star \rangle)$, $\mathrm{Var}[r \mid x, \theta_\star] = g(\tau)\dot{\mu}(\langle x, \theta_\star \rangle)$

Examples. $\mu(z) = z$: Gaussian, $\mu(z) = (1 + e^{-z})^{-1}$: **Bernoulli**, $\mu(z) = e^z$: Poisson

Generalized Linear Models

Problem Setting

Consider the **Generalized Linear Model (GLM)**:

$$dp(r \mid x; \theta_\star) = \exp\left(\frac{r\langle x, \theta_\star\rangle - m(\langle x, \theta_\star\rangle)}{g(\tau)} + h(r, \tau)\right) d\nu,$$

with dispersion parameter $\tau > 0$, base measure ν, **context** $x \in X$, and **unknown parameter** $\theta_\star \in \Theta$.

Assumptions. $X \subseteq \mathbb{B}^d(1)$, $\varnothing \neq \Theta \subseteq \mathbb{B}^d(S)$, Θ compact & convex, $m(\cdot)$ is convex and three-times differentiable.

Properties. $\mathbb{E}[r \mid x, \theta_\star] = m'(\langle x, \theta_\star\rangle) =: \mu(\langle x, \theta_\star\rangle)$, $\mathrm{Var}[r \mid x, \theta_\star] = g(\tau)\dot{\mu}(\langle x, \theta_\star\rangle)$

Examples. $\mu(z) = z$: Gaussian, $\mu(z) = (1 + e^{-z})^{-1}$: **Bernoulli**, $\mu(z) = e^z$: Poisson

Generalized Linear Bandits

Confidence Sequence (CS) for the Unknown Parameter

Generalized Linear Bandits

Confidence Sequence (CS) for the Unknown Parameter

Goal: For $\delta \in (0,1)$, **obtain** $\{\mathscr{C}_t(\delta)\}_{t \geq 1}$ **s.t.** $\mathbb{P}\left(\exists t \geq 1 : \theta_\star \notin \mathscr{C}_t(\delta)\right) \leq \delta$

Generalized Linear Bandits

Confidence Sequence (CS) for the Unknown Parameter

Goal: For $\delta \in (0,1)$**, obtain** $\{\mathscr{C}_t(\delta)\}_{t \geq 1}$ **s.t.** $\mathbb{P}\left(\exists t \geq 1 : \theta_\star \notin \mathscr{C}_t(\delta)\right) \leq \delta$

Setting. $\{(x_s, r_s)\}_{s \geq 1}$: adaptively collected observations satisfying $\mathbb{E}[r_s \mid \Sigma_s] = \mu(\langle x_s, \theta_\star \rangle)$, where $\Sigma_s := \sigma(\{x_1, r_1, \cdots, x_{s-1}, r_{s-1}, x_s\})$.

Generalized Linear Bandits

Confidence Sequence (CS) for the Unknown Parameter

Goal: For $\delta \in (0,1)$, obtain $\{\mathscr{C}_t(\delta)\}_{t\geq 1}$ s.t. $\mathbb{P}\left(\exists t \geq 1 : \theta_\star \notin \mathscr{C}_t(\delta)\right) \leq \delta$

Setting. $\{(x_s, r_s)\}_{s\geq 1}$: adaptively collected observations satisfying $\mathbb{E}[r_s \mid \Sigma_s] = \mu(\langle x_s, \theta_\star \rangle)$, where $\Sigma_s := \sigma(\{x_1, r_1, \cdots, x_{s-1}, r_{s-1}, x_s\})$.

We consider **CS** of the form $\mathscr{C}_t(\delta) := \left\{\theta \in \Theta : \mathscr{L}_t(\theta) - \mathscr{L}_t(\widehat{\theta}_t) \leq \beta_t(\delta)^2\right\}$, where

$$\mathscr{L}_t(\theta) := \sum_{s=1}^{t-1} \left\{ \ell_s(\theta) \triangleq \frac{-r_s\langle x_s, \theta\rangle + m(\langle x_s, \theta\rangle)}{g(\tau)} \right\}, \quad \widehat{\theta}_t := \mathrm{argmin}_{\theta\in\Theta} \mathscr{L}_t(\theta)\,.$$

where $\mathscr{L}_t(\theta)$ is the cumulative log-likelihood loss til time $t-1$, with **Lipschitz constant L_t.**

New, State-of-the-Art CS for GLMs!

Contribution #1

Proof via PAC-Bayes

Theorem 3.1. We have $\mathbb{P}\left(\exists t \geq 1 : \theta_\star \notin \mathscr{C}_t(\delta)\right) \leq \delta$, where

$$\mathscr{C}_t(\delta) := \left\{\theta \in \Theta : \mathscr{L}_t(\theta) - \mathscr{L}_t(\widehat{\theta}_t) \leq \beta_t(\delta)^2\right\}$$

$$\beta_t(\delta)^2 := \log\frac{1}{\delta} + d\log\left(e \vee \frac{2eSL_t}{d}\right)$$

Bernoulli: $\beta_t(\delta)^2 \lesssim_\delta d\log\frac{St}{d}$ => poly(S)-free for **Bernoulli**!!!

<=> prior work [Lee et al., AISTATS'24]: $\mathcal{O}_\delta\left(S + d\log\frac{St}{d}\right)$

Rmk. For self-concordant GLMs, one can have an *ellipsoidal form* of the CS.

Proof of Theorem 3.1

Step 1. Time-Uniform PAC-Bayes Bound

Proof of Theorem 3.1

Step 1. Time-Uniform PAC-Bayes Bound

Lemma 3.3. 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:

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

Proof of Theorem 3.1

Step 1. Time-Uniform PAC-Bayes Bound

Lemma 3.3. 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:

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

pf. Consider the likelihood ratio $M_t(\theta) = \exp(\mathscr{L}_t(\theta_\star) - \mathscr{L}_t(\theta))$.

Proof of Theorem 3.1

Step 1. Time-Uniform PAC-Bayes Bound

Lemma 3.3. 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:

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

pf. Consider the likelihood ratio $M_t(\theta) = \exp(\mathscr{L}_t(\theta_\star) - \mathscr{L}_t(\theta))$.

1. $M_t(\theta)$ is a nonnegative martingale, and so is $\mathbb{E}_{\theta \sim \mathbb{Q}}[M_t(\theta)]$ by Tonelli's theorem

Proof of Theorem 3.1

Step 1. Time-Uniform PAC-Bayes Bound

Lemma 3.3. 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:

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

pf. Consider the likelihood ratio $M_t(\theta) = \exp(\mathscr{L}_t(\theta_\star) - \mathscr{L}_t(\theta))$.

1. $M_t(\theta)$ is a nonnegative martingale, and so is $\mathbb{E}_{\theta \sim \mathbb{Q}}[M_t(\theta)]$ by Tonelli's theorem

2. By Ville's inequality [Ville, 1939], we have $\mathbb{P}\left(\exists t \geq 1 : \mathbb{E}_{\theta \sim \mathbb{Q}}[M_t(\theta)] \geq \frac{1}{\delta}\right) \leq \delta$

Anytime-valid *Markov's inequality* for supermartingales

Proof of Theorem 3.1

Step 1. Time-Uniform PAC-Bayes Bound

Lemma 3.3. 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:

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

pf. Consider the likelihood ratio $M_t(\theta) = \exp(\mathscr{L}_t(\theta_\star) - \mathscr{L}_t(\theta))$.

Anytime-valid *Markov's inequality* for supermartingales

1. $M_t(\theta)$ is a nonnegative martingale, and so is $\mathbb{E}_{\theta\sim\mathbb{Q}}[M_t(\theta)]$ by Tonelli's theorem
2. By Ville's inequality [Ville, 1939], we have $\mathbb{P}\left(\exists t \geq 1 : \mathbb{E}_{\theta\sim\mathbb{Q}}[M_t(\theta)] \geq \frac{1}{\delta}\right) \leq \delta$
3. "Change" $\mathbb{Q}$ to $\mathbb{P}_t$ via **Donsker-Varadhan variational representation of KL** [Donsker & Varadhan, 1983].

$$\mathrm{KL}(\mathbb{P}_t || \mathbb{Q}) = \sup_{g:\Theta\to\mathbb{R}} \mathbb{E}_{\theta\sim\mathbb{P}_t}[g(\theta)] - \log \mathbb{E}_{\theta\sim\mathbb{Q}}[e^{g(\theta)}]$$

Proof of Theorem 3.1

Step 1. Time-Uniform PAC-Bayes Bound

Journal of Machine Learning Research 24 (2023) 1-61 Submitted 3/23; Revised 10/23; Published 12/23

A Unified Recipe for Deriving (Time-Uniform) PAC-Bayes Bounds

Ben Chugg BENCHUGG@CMU.EDU
Hongjian Wang HJNWANG@CMU.EDU
Aaditya Ramdas ARAMDAS@STAT.CMU.EDU
Departments of Statistics and Machine Learning
Carnegie Mellon University

Proof of Theorem 3.1

Step 1. Time-Uniform PAC-Bayes Bound

Journal of Machine Learning Research 24 (2023) 1-61 Submitted 3/23; Revised 10/23; Published 12/23

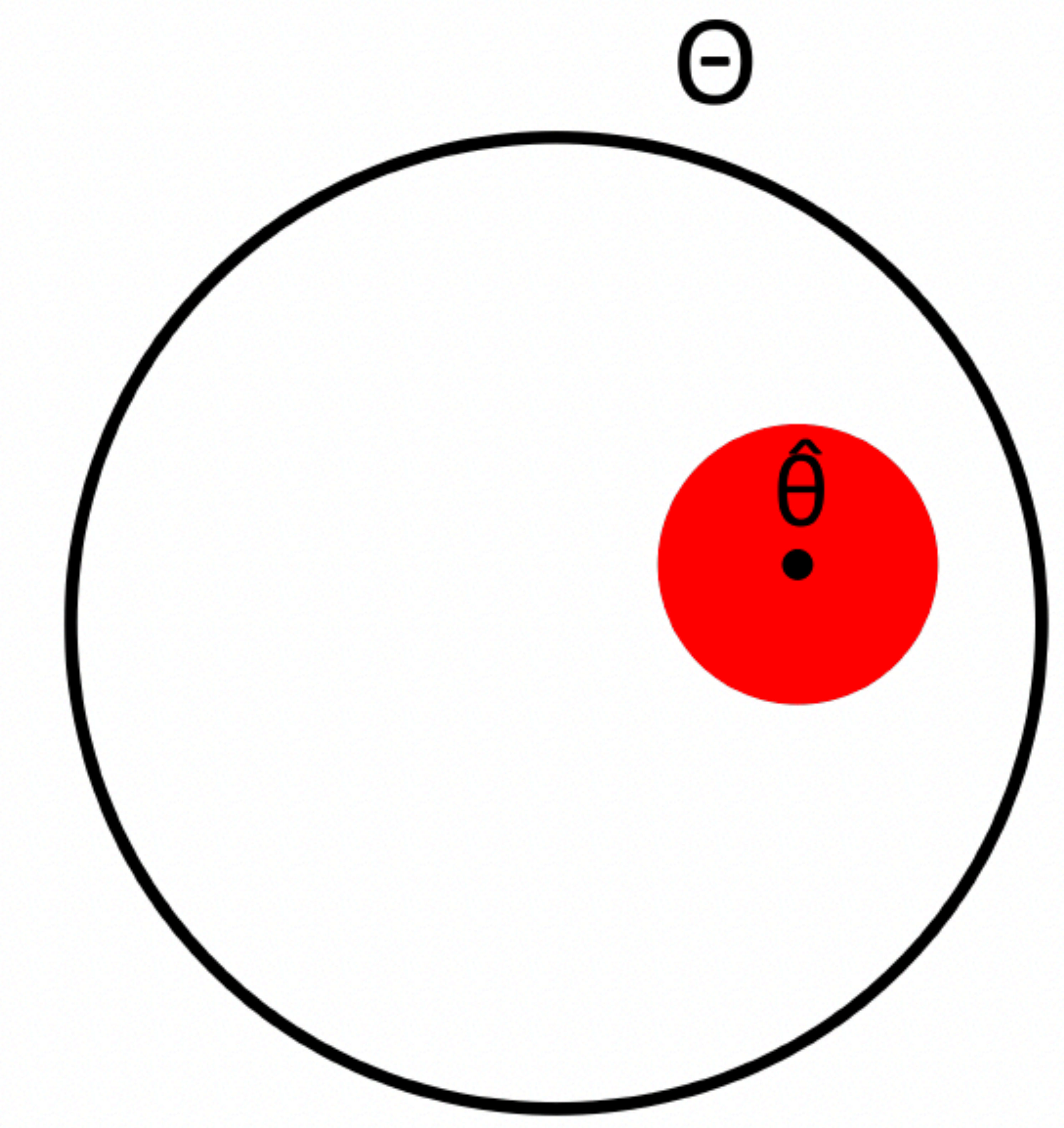
SURVEY

A Unified Recipe for Deriving (Time-Uniform) PAC-Bayes Bounds

Ben Chugg BENCHUGG@CMU.EDU
Hongjian Wang HJNWANG@CMU.EDU
Aaditya Ramdas ARAMDAS@STAT.CMU.EDU
Departments of Statistics and Machine Learning
Carnegie Mellon University

Proof of Theorem 3.1

Step 2. Novel choice of of “prior” and “posterior” & Lipschitzness



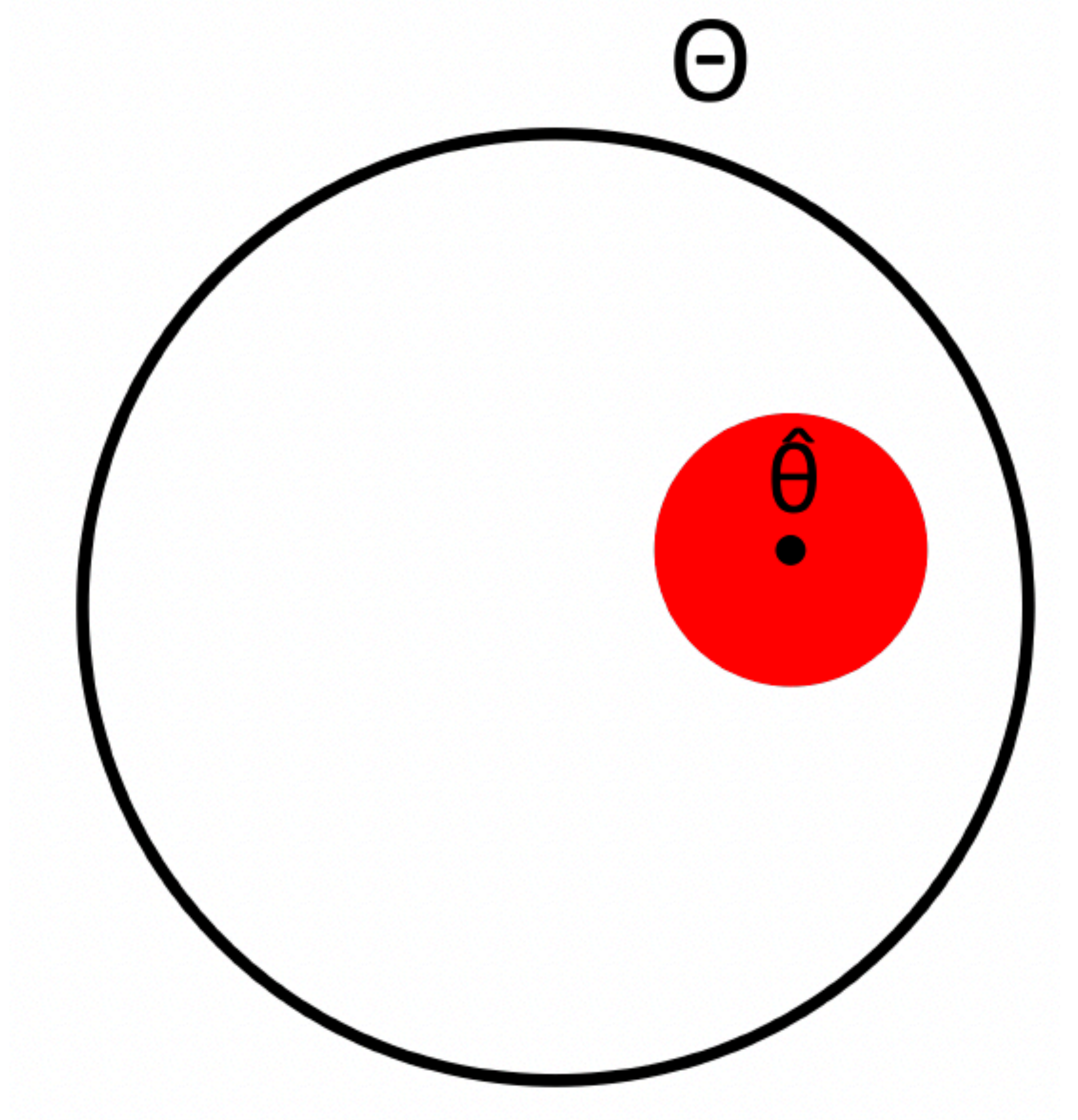
From P. Alquier’s MLSS lecture slides

Proof of Theorem 3.1

Step 2. Novel choice of of “prior” and “posterior” & Lipschitzness

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

Remark. Originally considered in portfolio optimization [Blum and Kalai, 1999] and fast rates in online learning [Hazan et al., 2007; Foster et al., COLT’18].



From P. Alquier’s MLSS lecture slides

Proof of Theorem 3.1

Step 2. Novel choice of of “prior” and “posterior” & Lipschitzness

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

Remark. Originally considered in portfolio optimization [Blum and Kalai, 1999] and fast rates in online learning [Hazan et al., 2007; Foster et al., COLT'18].

Proof of Theorem 3.1

Step 2. Novel choice of of "prior" and "posterior" & Lipschitzness

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

Remark. Originally considered in portfolio optimization [Blum and Kalai, 1999] and fast rates in online learning [Hazan et al., 2007; Foster et al., COLT'18].

=> $D_{KL}(\mathbb{P}_t || \mathbb{Q}) = \log \dfrac{\mathrm{vol}(\Theta)}{\mathrm{vol}(\widetilde{\Theta})} = \log \dfrac{\mathrm{vol}(\Theta)}{\mathrm{vol}(c\Theta)} = d \log \dfrac{1}{c}$

Also, $\mathbb{E}_{\theta\sim\mathbb{P}_t}[\mathscr{L}_t(\theta)] = \mathscr{L}_t(\widehat{\theta}_t) + \mathbb{E}_{\theta\sim\mathbb{P}_t}[\mathscr{L}_t(\theta) - \mathscr{L}_t(\widehat{\theta}_t)] \leq \mathscr{L}_t(\widehat{\theta}_t) + 2SL_t c$,

Proof of Theorem 3.1

Step 2. Novel choice of of "prior" and "posterior" & Lipschitzness

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

Remark. Originally considered in portfolio optimization [Blum and Kalai, 1999] and fast rates in online learning [Hazan et al., 2007; Foster et al., COLT'18].

=> $D_{KL}(\mathbb{P}_t || \mathbb{Q}) = \log \dfrac{\mathrm{vol}(\Theta)}{\mathrm{vol}(\widetilde{\Theta})} = \log \dfrac{\mathrm{vol}(\Theta)}{\mathrm{vol}(c\Theta)} = d \log \dfrac{1}{c}$

Also, $\mathbb{E}_{\theta\sim\mathbb{P}_t}[\mathscr{L}_t(\theta)] = \mathscr{L}_t(\widehat{\theta}_t) + \mathbb{E}_{\theta\sim\mathbb{P}_t}[\mathscr{L}_t(\theta) - \mathscr{L}_t(\widehat{\theta}_t)] \leq \mathscr{L}_t(\widehat{\theta}_t) + 2SL_tc$,

All in all, with probability at most δ, there exists a $t \geq 1$ such that

$$\mathscr{L}_t(\theta_\star) - \mathscr{L}_t(\widehat{\theta}_t) \geq \log\frac{1}{\delta} + d\log\frac{1}{c} + \mathbb{E}_{\theta\sim\mathbb{P}_t}[\mathscr{L}_t(\theta)] - \mathscr{L}_t(\widehat{\theta}_t) \geq \log\frac{1}{\delta} + d\log\frac{1}{c} + 2SL_tc$$

Choose $c = \min\left\{1, d/(2SL_t)\right\}$ and we are done.

Generalized Linear Bandits

Problem Setting

For $t \in [T]$:

1. The learner observes a potentially infinite (contextual) arm-set $\mathcal{X}_t \subset X$
2. The learner chooses $x_t \in \mathcal{X}_t$ according to some policy
3. Receive a reward $r_t \sim GLM(x_t, \theta_\star; \mu(\cdot))$
   - $\theta_\star$ is unknown to the learner

Goal: Minimize the regret

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

Generalized Linear Bandits

Contribution #2

<u>OFUGLB: Optimism in the Face of Uncertainty for Generalized Linear Bandits</u>

1. Compute $\widehat{\theta}_t$ and $\mathcal{C}_t(\delta)$ - **tighter confidence sequence** (Theorem 3.1)!
2. $(x_t, \theta_t) = \text{argmax}_{x \in \mathcal{X}_t, \theta \in \mathcal{C}_t(\delta)} \mu(\langle x, \theta \rangle)$
3. Play x_t and observe/receive a reward $r_t \sim GLM(x_t, \theta_\star; \mu(\cdot))$

Theorem 4.1. OFUGLB attains the following regret bound for self-concordant generalized linear bandits w.p. at least $1 - \delta$:

Nontrivial proof!!

$$\text{Reg}(T) \lesssim \underbrace{d\sqrt{\frac{g(\tau)T}{\kappa_\star(T)} \log \frac{SL_T}{d} \log \frac{R_{\dot{\mu}}ST}{d}}}_{\text{permanent term}} + \underbrace{d^2 R_s R_{\dot{\mu}} \sqrt{g(\tau)} \kappa(T)}_{\text{transient term}}$$

Generalized Linear Bandits

OFUGLB: Optimism in the Face of Uncertainty for Generalized Linear Bandits

- **Linear Bandits:** $\tilde{\mathcal{O}}\left(\sigma d\sqrt{T}\right)$
  - => matches state-of-the-art [Flynn et al., NeurIPS'23]
- **Logistic Bandits:** $\tilde{\mathcal{O}}\left(d\sqrt{T/\kappa_\star(T)} + d^2\kappa(T)\right)$
  - => *first* poly(S)-free regret with **computationally tractable, purely optimistic approach**!!
  - => improves upon prior state-of-the-art [Lee et al., AISTATS'24]
  - => similar guarantee in a *concurrent* work [Sawarni et al., arXiv'24], but is intractable and involves explicit warmup + their guarantees only apply to *bounded* GLBs.
- **Poisson Bandits:** $\tilde{\mathcal{O}}\left(dS\sqrt{T/\kappa_\star(T)} + d^2 e^{2S}\kappa(T)\right)$
  - => *state-of-the-art* regret guarantee

Brief Proof Sketch of Theorem 4.1

OFUGLB: Optimism in the Face of Uncertainty for Generalized Linear Bandits

Brief Proof Sketch of Theorem 4.1

OFUGLB: Optimism in the Face of Uncertainty for Generalized Linear Bandits

Previously: use self-concordance control lemma to obtain

$\|\theta_\star - \hat{\theta}_t\|_{H_t(\hat{\theta}_t)} = \mathcal{O}(S\beta_T(\delta))$

Brief Proof Sketch of Theorem 4.1

OFUGLB: Optimism in the Face of Uncertainty for Generalized Linear Bandits

Previously: use self-concordance control lemma to obtain

$\|\theta_\star - \hat{\theta}_t\|_{H_t(\hat{\theta}_t)} = \mathcal{O}(S\beta_T(\delta))$

Here: maximally avoid self-concordance control => use "exact" Taylor expansion,

$\|\theta_\star - \hat{\theta}_t\|_{\tilde{G}_t(\hat{\theta}_t, \nu_t)} = \mathcal{O}(\beta_T(\delta))$, where $\tilde{G}_t(\hat{\theta}_t, \nu_t) = \lambda \mathbf{I} + \frac{1}{g(\tau)} \sum_{s=1}^{t-1} \tilde{\alpha}_s(\hat{\theta}_t, \nu_t) x_s x_s^\top$ and

$\tilde{\alpha}_s(\theta, \nu) = \int_0^1 (1 - v)\dot{\mu}_t(\theta + v(\nu - \theta))dv.$

Brief Proof Sketch of Theorem 4.1

OFUGLB: Optimism in the Face of Uncertainty for Generalized Linear Bandits

Brief Proof Sketch of Theorem 4.1

OFUGLB: Optimism in the Face of Uncertainty for Generalized Linear Bandits

BUT, the remaining term of Cauchy-Schwartz, $\sum_t \|x_t\|^2_{\tilde{G}_t(\hat{\theta}_t, \nu_t)^{-1}}$, how to apply ***elliptical potential lemma?***

$$\tilde{G}_t(\hat{\theta}_t, \nu_t) = \lambda \mathbf{I} + \frac{1}{g(\tau)} \sum_{s=1}^{t-1} \tilde{\alpha}_s(\hat{\theta}_t, \nu_t) x_s x_s^\top$$

Lemma B.2 (Elliptical Potential Lemma; EPL[5]). *Let* $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T \in \mathcal{B}^d(X)$ *be a sequence of vectors and* $\mathbf{V}_t := \lambda \boldsymbol{I} + \sum_{s=1}^{t-1} \boldsymbol{x}_s \boldsymbol{x}_s^\intercal$. *Then, we have that*

$$\sum_{t=1}^{T} \min \left\{1, \|\boldsymbol{x}_t\|_{\mathbf{V}_t^{-1}}^2 \right\} \leq 2d \log \left(1 + \frac{X^2 T}{d\lambda}\right). \tag{23}$$

BUT, the remaining term of Cauchy-Schwartz, $\sum_t \|x_t\|^2_{\tilde{G}_t(\hat{\theta}_t, \nu_t)^{-1}}$, how to apply ***elliptical potential lemma?***

$$\tilde{G}_t(\hat{\theta}_t, \nu_t) = \lambda \mathbf{I} + \frac{1}{g(\tau)} \sum_{s=1}^{t-1} \tilde{\alpha}_s(\hat{\theta}_t, \nu_t) x_s x_s^\top$$

Lemma B.2 (Elliptical Potential Lemma; EPL[5]). *Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T \in \mathcal{B}^d(X)$ be a sequence of vectors and $\mathbf{V}_t := \lambda \boldsymbol{I} + \sum_{s=1}^{t-1} \boldsymbol{x}_s \boldsymbol{x}_s^\intercal$. Then, we have that*

$$\sum_{t=1}^{T} \min\left\{1, \|\boldsymbol{x}_t\|^2_{\mathbf{V}_t^{-1}}\right\} \leq 2d \log\left(1 + \frac{X^2 T}{d\lambda}\right). \tag{23}$$

BUT, the remaining term of Cauchy-Schwartz, $\sum_t \|x_t\|^2_{\tilde{G}_t(\hat{\theta}_t, \nu_t)^{-1}}$, how to apply ***elliptical potential lemma?***

$$\tilde{G}_t(\hat{\theta}_t, \nu_t) = \lambda \mathbf{I} + \frac{1}{g(\tau)} \sum_{s=1}^{t-1} \tilde{\alpha}_s(\hat{\theta}_t, \nu_t) x_s x_s^\top$$

Main proof novelty: designate the "worst-case" $\bar{\theta}_t$'s such that

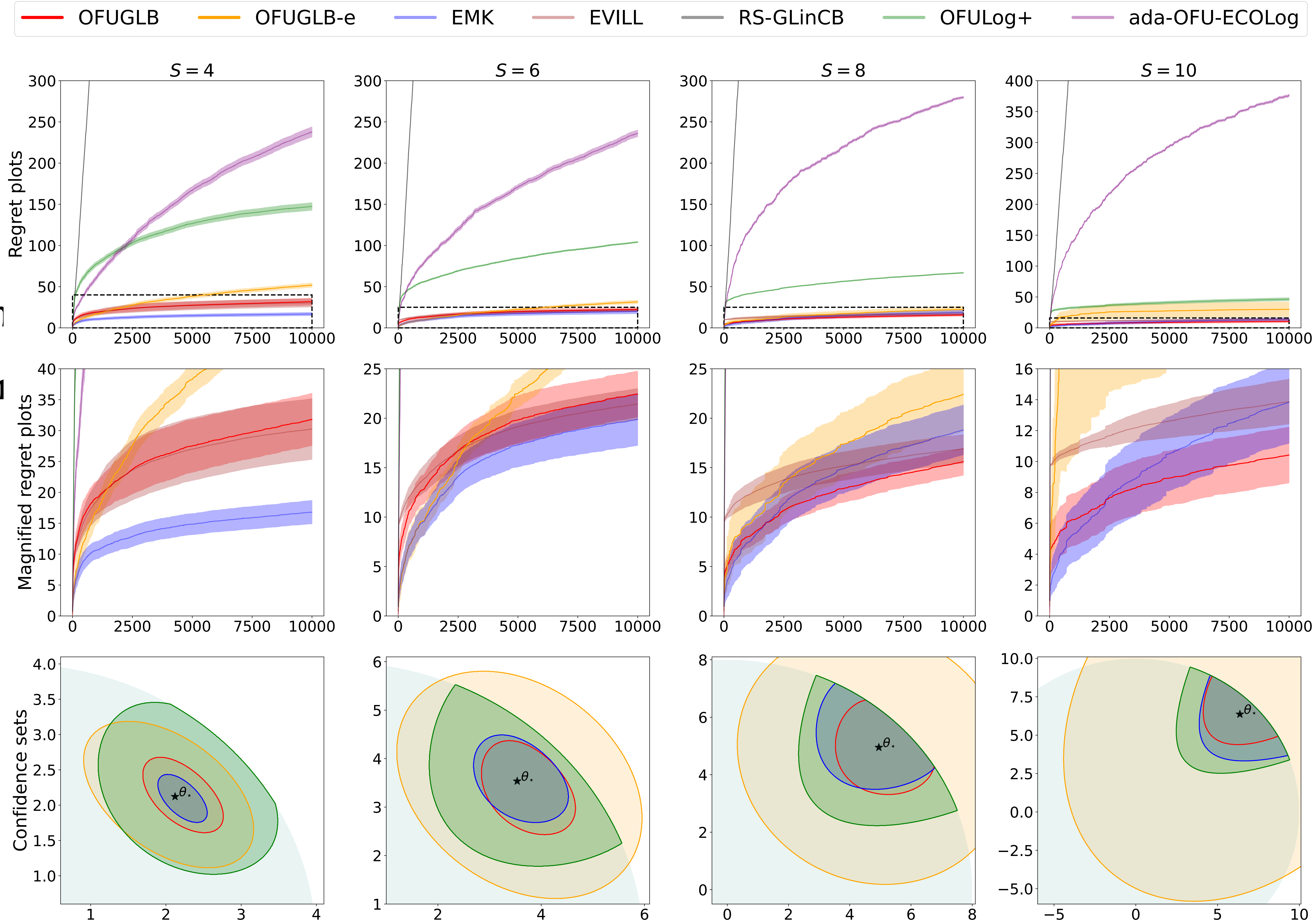
$$\sum_t \|x_t\|^2_{\tilde{G}_t(\hat{\theta}_t, \nu_t)^{-1}} \leq \sum_t \min\left\{1, \dot{\mu}(\bar{\theta}_s)\|x_t\|^2_{\bar{H}_t^{-1}}\right\}, \text{ where } \bar{H}_t = 2g(\tau)\lambda I + \sum_{s=1}^{t-1} \dot{\mu}_s(\bar{\theta}_s) x_s x_s^\top$$

Experiments for Logistic Bandits

Better than most of existing approaches

- 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!!

- On
- Sy



OFUGLB
OFUGLB-e
EMK
EVILL
RS-GLinCB
OFULog+
ada-OFU-ECOLog
$S = 4$
$S = 6$
$S = 8$
$S = 10$
Regret plots
Magnified regret plots
Confidence sets
$\theta_\star$

References

J. Lee, S.-Y. Yun, and K.-S. Jun. Improved Regret Analysis of (Multinomial) Logistic Bandits via Regret-to-Confidence-Set Conversion. In *AISTATS* 2024.

A. Sawarni, N. Das, S. Barman, and G. Sinha. Generalized Linear Bandits with Limited Adaptivity. In *NeurIPS* 2024.

H. Flynn, D. Reeb, M. Kandemir, and J. R. Peters. Improved Algorithms for Stochastic Linear Bandits Using Tail Bounds for Martingale Mixtures. In *NeurIPS* 2023.

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

M. D. Donsker and S. R. S. Varadhan. Asymptotic Evaluation of Certain Markov Process Expectations for Large Times. IV. *Communications on Pure and Applied Mathematics*, 36(2):183-212, 1983.

A. Blum and A. Kalai. Universal Portfolios With and Without Transaction Costs. *Machine Learning*, 35(3):193-205, 1999.

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

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