Breaking the $\sqrt{T}$ Barrier: Instance-Independent Logarithmic Regret in Stochastic Contextual Linear Bandits

Avishek Ghosh and Abishek Sankararaman

ICML 2022, Baltimore, USA

Contextual Linear Bandits

- Standard framework in online learning—Abe et. al ’03, Auer et. al ’02, Li et. al ’10, Chu et. al ’11, Yadkori et. al ’16 …
- Learning Model: At time $t \in [T]$, Agent observes K contexts $[\beta_{1,t}, \ldots, \beta_{K,t}]$, each dim d
- Plays Algorithm $\mathscr{A}$, chooses arm $i \in [K]$, observes reward $r_t = \langle \beta_{i,t}, \theta^* \rangle + \xi_t$

Contextual Linear Bandits

- Standard framework in online learning—Abe et. al ’03, Auer et. al ’02, Li et. al ’10, Chu et. al ’11, Yadkori et. al ’16 …
- Learning Model: At time $t \in [T]$, Agent observes K contexts $[\beta_{1,t}, \ldots, \beta_{K,t}]$, each dim d
- Plays Algorithm $\mathscr{A}$, chooses arm $i \in [K]$, observes reward $r_t = \langle \beta_{i,t}, \theta^* \rangle + \xi_t$
- Here, θ^* is unknown, $\|\theta^*\| \leq 1$, and $\{\xi_t\}_{t=1}^{T}$ zero mean sub-Gaussian noise

Objective: Minimize regret: $$R_{\mathscr{A}}(T) = \sum_{t=1}^{T} \max_{j \in [K]} \langle \beta_{j,t}, \theta^* \rangle - \langle \beta_{i_t,t}, \theta^* \rangle$$

The $\sqrt{T}$ Barrier

- If we have no control over the context generation — Adversarial Contexts
- Chu et. al '11 obtain $R(T) = \Omega(\sqrt{T})$
- No hope to break the $\sqrt{T}$ barrier for adversarial contexts!!

The $\sqrt{T}$ Barrier

- If we have no control over the context generation— Adversarial Contexts
- Chu et. al '11 obtain $R(T) = \Omega(\sqrt{T})$
- No hope to break the $\sqrt{T}$ barrier for adversarial contexts!!
- Question: What if contexts are stochastic with some additional structures?

The $\sqrt{T}$ Barrier

- If we have no control over the context generation — Adversarial Contexts
- Chu et. al '11 obtain $R(T) = \Omega(\sqrt{T})$
- No hope to break the $\sqrt{T}$ barrier for adversarial contexts!!
- Question: What if contexts are stochastic with some additional structures?
- In this work: We break $\Omega(\sqrt{T})$ barrier, and propose an algorithm $\mathscr{A}$ such that

$$R_{\mathscr{A}}(T) = \mathcal{O}(\text{polylog}(T))$$

Structured Stochastic Contexts

- $\beta_{i,t}$ drawn independent of the past and $\{\beta_{j,t}\}_{j \neq i}$
- $\mathbb{E}_{t-1}[\beta_{i,t}] = 0$, $\mathbb{E}_{t-1}[\beta_{i,t}\beta_{i,t}^\top] \succeq \rho_{\min} I$ with $\rho_{min} > 0$ ($\mathbb{E}_{t-1}$: cond. exp. upto $t-1$)
- Some additional (mild) technical condition on variance of projected contexts…

Structured Stochastic Contexts

- $\beta_{i,t}$ drawn independent of the past and $\{\beta_{j,t}\}_{j\neq i}$
- $\mathbb{E}_{t-1}[\beta_{i,t}] = 0$, $\mathbb{E}_{t-1}[\beta_{i,t}\,\beta_{i,t}^{\top}] \succeq \rho_{\min} I$ with $\rho_{min} > 0$ ($\mathbb{E}_{t-1}$: cond. exp. upto $t-1$)
- Some additional (mild) technical condition on variance of projected contexts…
- No new assumptions !! Same assumptions in following works:
- Gentile et.al '17—Clustering, Chatterjee et.al '20 —Model Selection, Ghosh et.al '21 - Personalization, Ghosh et.al '21 — Adaptation

Structured Stochastic Contexts

- $\beta_{i,t}$ drawn independent of the past and $\{\beta_{j,t}\}_{j \neq i}$
- $\mathbb{E}_{t-1}[\beta_{i,t}] = 0$, $\mathbb{E}_{t-1}[\beta_{i,t} \beta_{i,t}^\top] \succeq \rho_{\min} I$ with $\rho_{min} > 0$ ($\mathbb{E}_{t-1}$: cond. exp. upto $t-1$)
- Some additional (mild) technical condition on variance of projected contexts…
- No new assumptions !! Same assumptions in following works:
- Gentile et.al '17—Clustering, Chatterjee et.al '20 —Model Selection, Ghosh et.al '21 - Personalization, Ghosh et.al '21 — Adaptation
- Key Insight: Assumptions imply inference (estimation) on θ^* and reg. min. simultaneously !!
- We use this inference to obtain the $\mathcal{O}(\text{polylog}(T))$ regret

Key Components

- Norm adaptive linear bandit algorithm (ALB)—Ghosh et.al '21
- For contextual linear bandits with parameter θ^*, a (modified) ALB of Ghosh et.al '21

$$R(T) \leq \mathcal{O}(\|\theta^*\|\sqrt{dT})$$

- If we make $\|\theta^*\|$ small, regret $R(T)$ is small

Key Components

- Norm adaptive linear bandit algorithm — Ghosh et.al '21

- For contextual linear bandits with parameter θ^*, a (modified) adaptive algo of Ghosh et.al '21

$$R(T) \leq \mathcal{O}(\|\theta^*\|\sqrt{dT})$$

- If we make $\|\theta^*\|$ small, regret $R(T)$ is small

- Parameter Inference (estimation) — Structured Contexts allow this !!

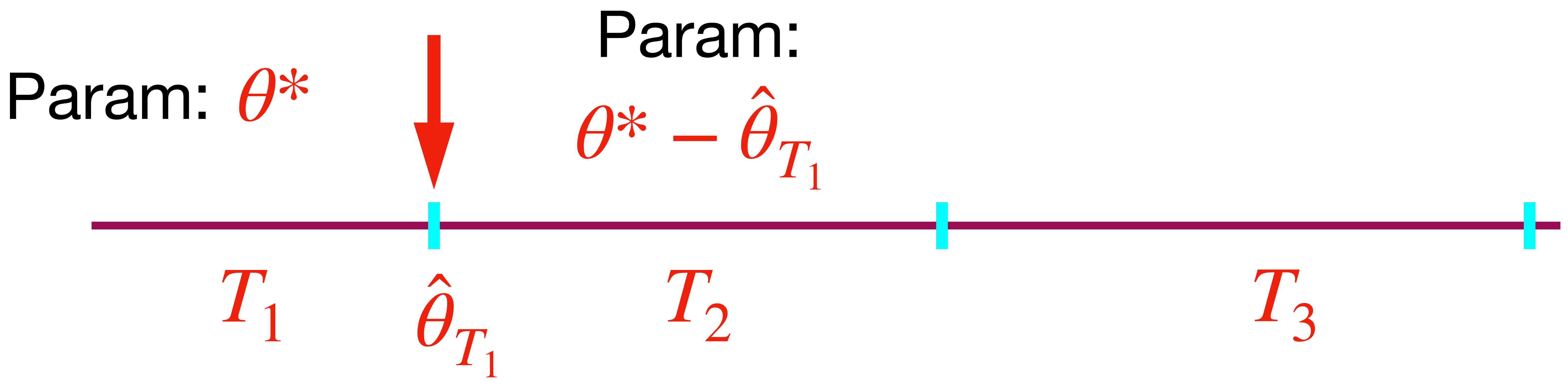
Key Components

- Norm adaptive linear bandit algorithm — Ghosh et.al '21
- For contextual linear bandits with parameter θ^*, a (modified) adaptive algo of Ghosh et.al '21

$$R(T) \leq \mathcal{O}(\|\theta^*\|\sqrt{dT})$$

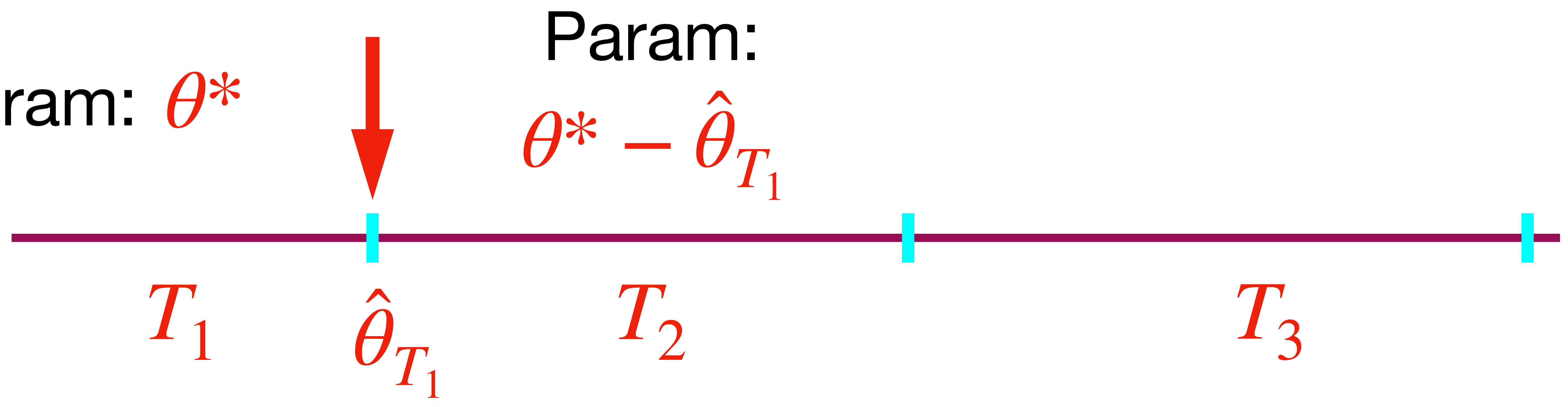
- If we make $\|\theta^*\|$ small, regret $R(T)$ is small
- Parameter Inference (estimation) — Structured Contexts allow this !!
- Question: Can we combine the above?
- Yes!!! we estimate θ^* and use it to reduce $R(T)$ using norm adaptive algo
- We break the learning horizon and do this over multiple epochs

Algorithm- Low Regret Stochastic Contextual Bandits (LR-SCB)



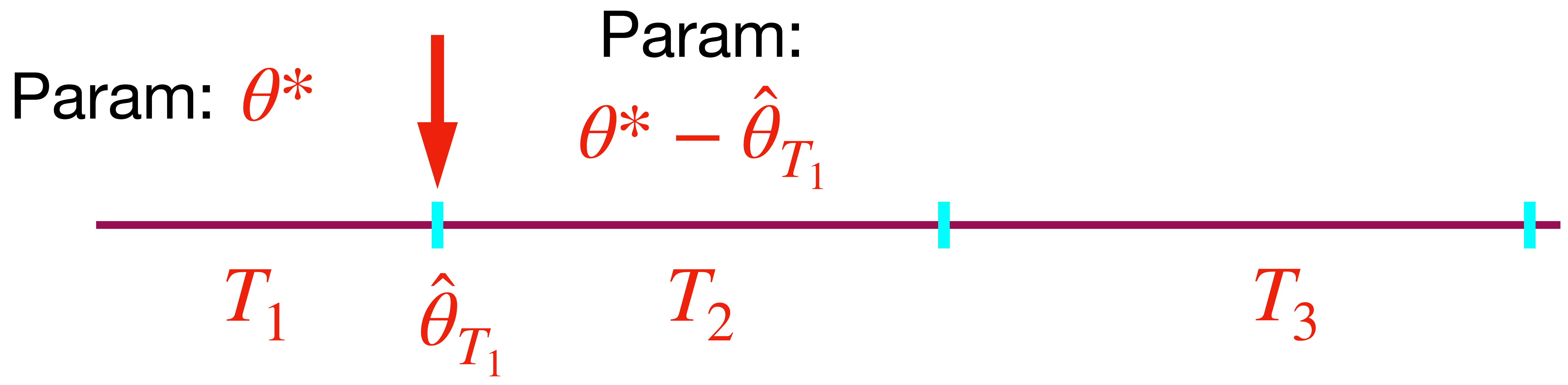
- Epoch 1: Play Adaptive ALB of Ghosh et. al '21 with $\|\theta^*\| \leq 1$
- Regret in first epoch: $R_1 \leq \mathcal{O}(\|\theta^*\|\sqrt{T_1}) \leq \mathcal{O}(\sqrt{T_1})$

Algorithm- Low Regret Stochastic Contextual Bandits (LR-SCB)



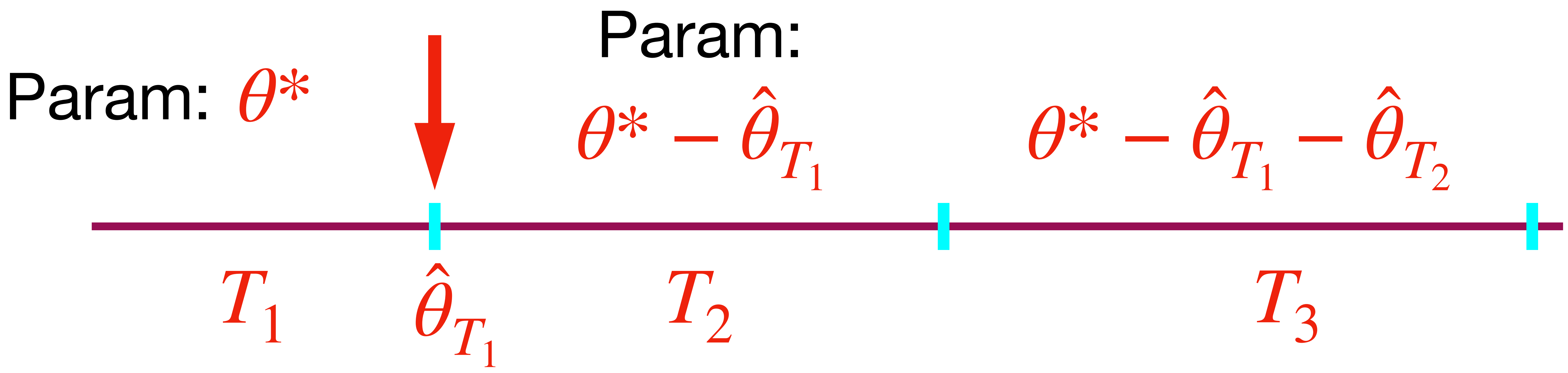
- Epoch 1: Play Adaptive ALB of Ghosh et. al '21 with $\|\theta^*\| \leq 1$
- Regret in first epoch: $R_1 \leq \mathcal{O}(\|\theta^*\|\sqrt{T_1}) \leq \mathcal{O}(\sqrt{T_1})$
- After first epoch: $\hat{\theta}_{T_1}$ is the estimate of θ^* with $\|\hat{\theta}_{T_1} - \theta^*\| \leq \mathcal{O}(1/\sqrt{T_1})$

Algorithm- Low Regret Stochastic Contextual Bandits (LR-SCB)



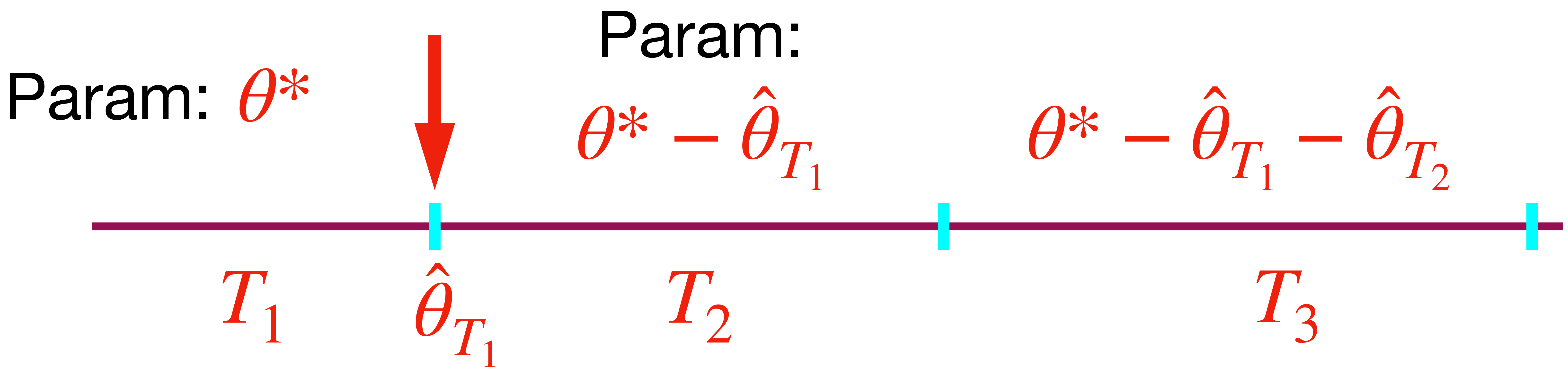
- Epoch 1: Play Adaptive ALB of Ghosh et. al '21 with $\|\theta^*\| \leq 1$
- Regret in first epoch: $R_1 \leq \mathcal{O}(\|\theta^*\|\sqrt{T_1}) \leq \mathcal{O}(\sqrt{T_1})$
- After first epoch: $\hat{\theta}_{T_1}$ is the estimate of θ^* with $\|\hat{\theta}_{T_1} - \theta^*\| \leq \mathcal{O}(1/\sqrt{T_1})$
- Epoch 2: Shift the Learning to $\hat{\theta}_{T_1}$: Play shifted ALB with parameter $\theta^* - \hat{\theta}_{T_1}$
- Crucial: $\|\hat{\theta}_{T_1} - \theta^*\| \leq \mathcal{O}(1/\sqrt{T_1})$ and Regret: $R_2 \leq \mathcal{O}(\|\theta^* - \hat{\theta}_{T_1}\|\sqrt{T_1}) \leq \mathcal{O}(\sqrt{\frac{T_2}{T_1}})$

Algorithm- Low Regret Stochastic Contextual Bandits (LR-SCB)



- Continue shift and estimation over successive epochs

Algorithm- Low Regret Stochastic Contextual Bandits (LR-SCB)



- Continue shift and estimation over successive epochs
- Total Regret over N epochs:

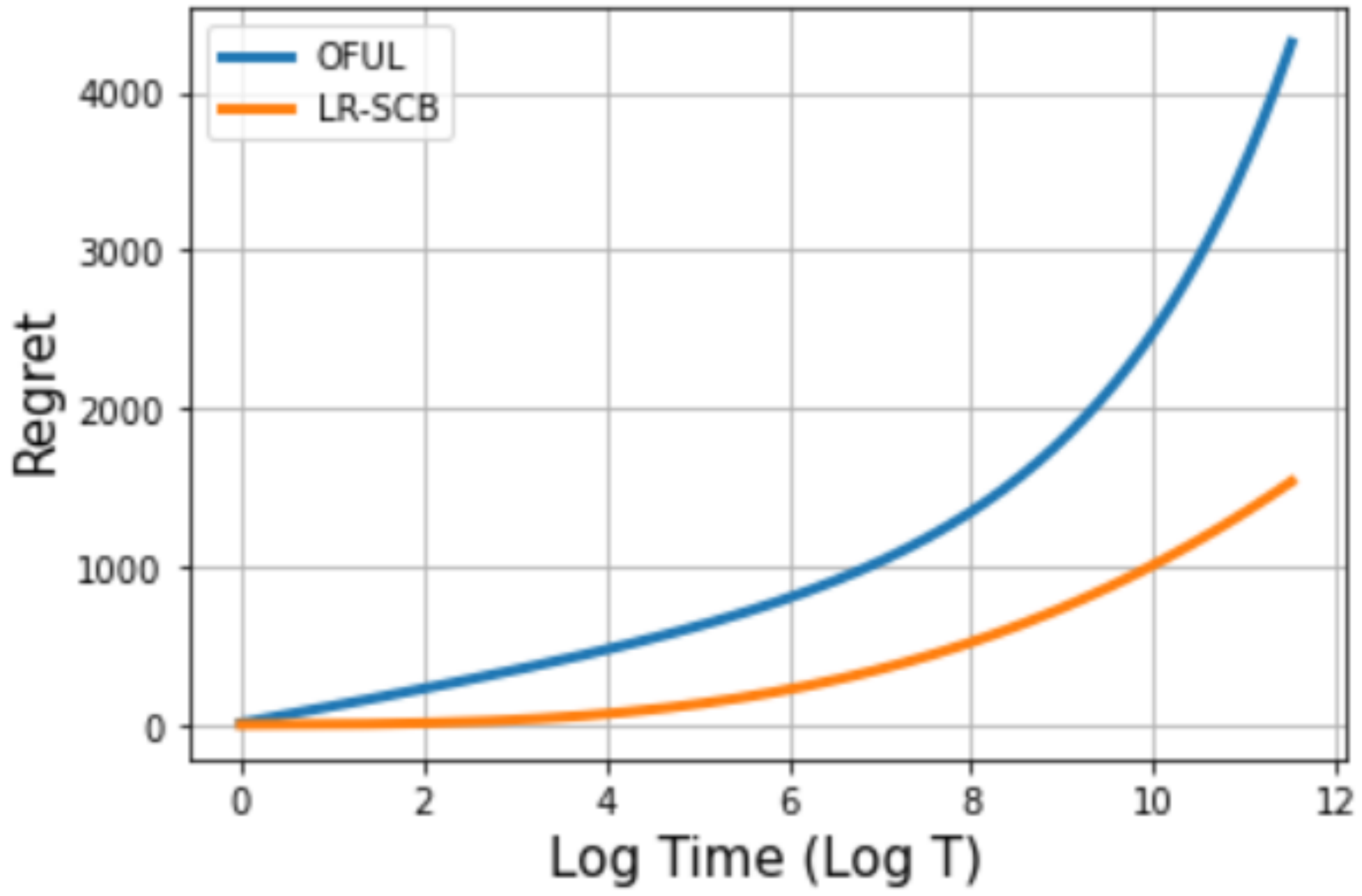
$$R(T) \leq \mathcal{O}\left(\sqrt{T_1} + \sqrt{\frac{T_2}{T_1}} + \sqrt{\frac{T_3}{T_2}} + ..\right)$$

- Choose $T_i = T_1(\log T)^{i-1}$ and $T_1 = \mathcal{O}(1)$

Regret: $R(T) \leq \mathcal{O}(\text{polylog}(T))$

Simulations

Setup: $d = K = 20$

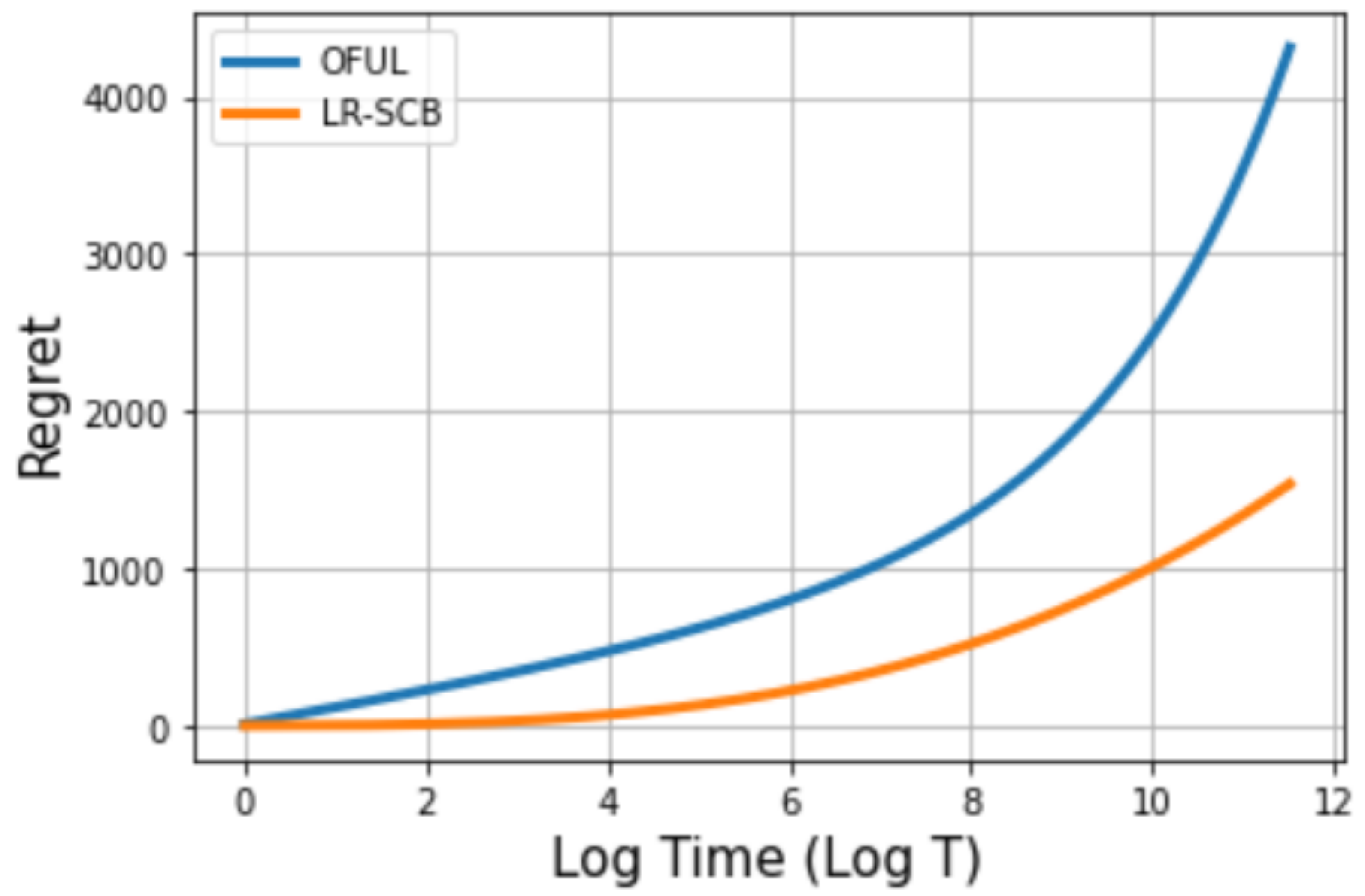


Plot: Regret vs $\log T$

- LR-SCB grows slowly (polynomially); unlike standard Linear bandit algo, OFUL (exponentially)

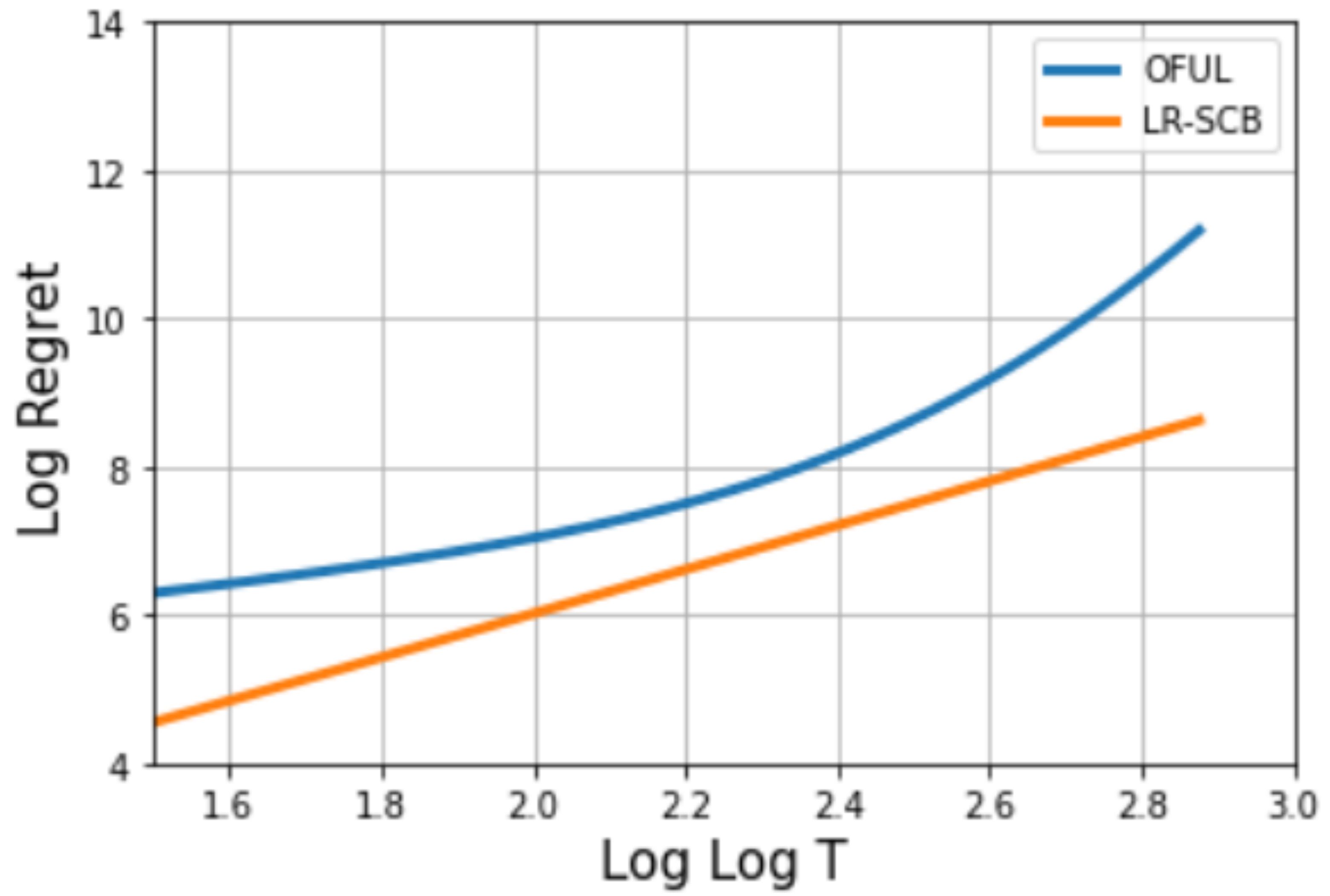
Simulations

Setup: $d = K = 20$



Plot: Regret vs $\log T$

- LR-SCB grows slowly (polynomially); unlike standard Linear bandit algo, OFUL (exponentially)



Plot: Log Regret vs $\log \log T$

- Slope of LR-SCB constant;
- Slope of OFUL: growing
- Regret of LR-SCB— Polylogarithmic

Thank you