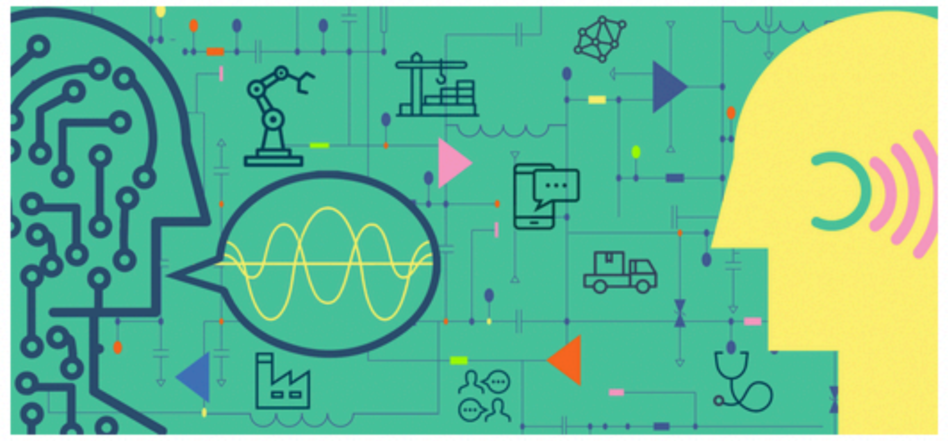
Universal and data-adaptive algorithms for model selection in linear contextual bandits

Vidya Muthukumar, Georgia Tech ECE and ISyE

(Joint work with Akshay Krishnamurthy, Microsoft Research; work conducted at Simons Institute)

Programs | Fall 2020



Theory of Reinforcement Learning

Aug. 19 – Dec. 18, 2020

The model selection problem

Setting: K-armed linear contextual bandit problem…with potential simple MAB structure

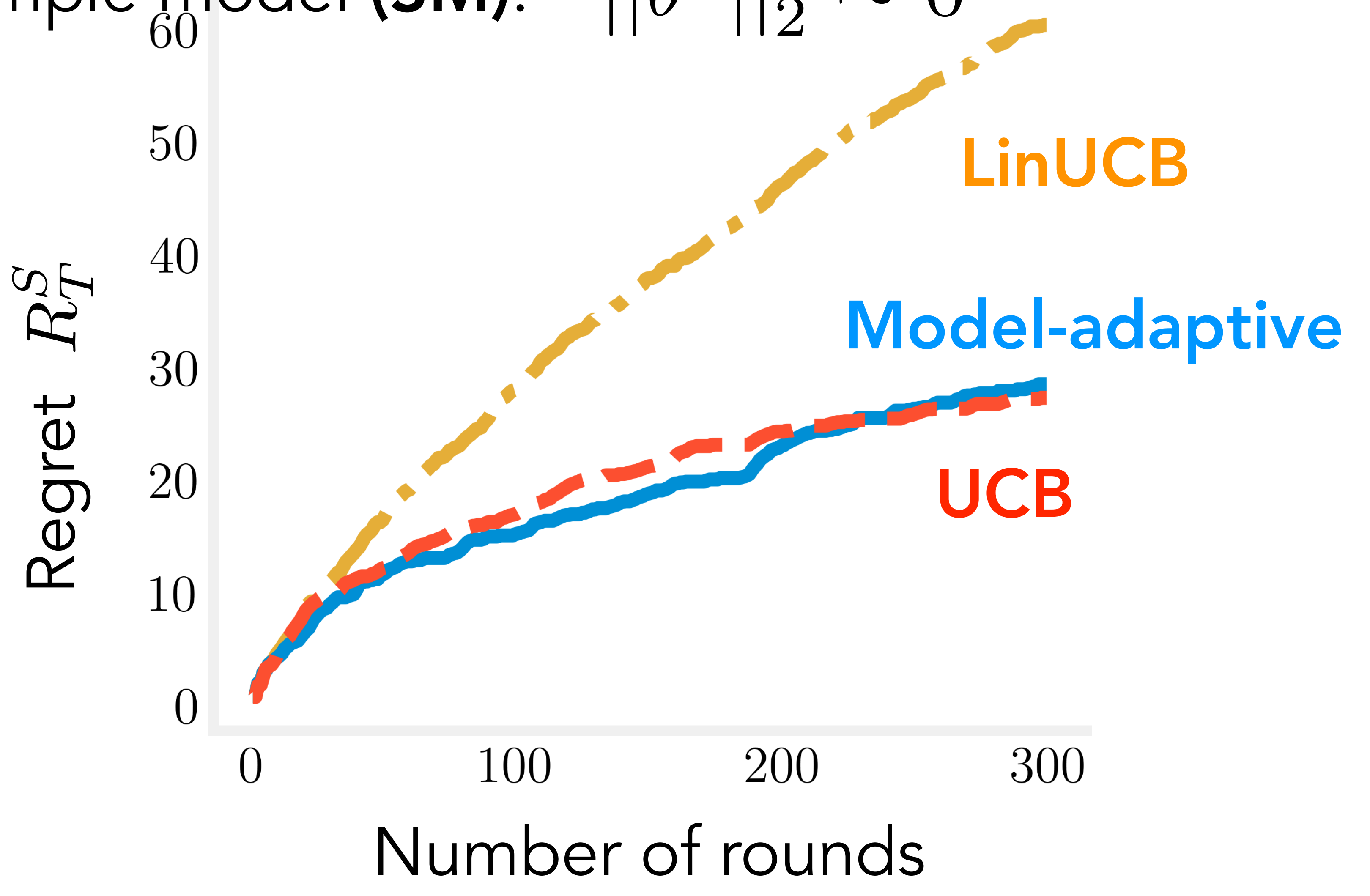
$$G_{i,t} = \mu_i + \langle \boldsymbol{x}_{i,t}, \boldsymbol{\theta}^* \rangle + W_{i,t}$$

- $G_{i,t}$: Reward of arm k in round t
- μ_i: (unknown) mean of arm k
- $\boldsymbol{x}_{i,t}$: Context for arm k ($\in \mathbb{R}^d$)
- $\boldsymbol{\theta}^*$: (unknown) parameter ($\in \mathbb{R}^d$)
- $W_{i,t}$: noise

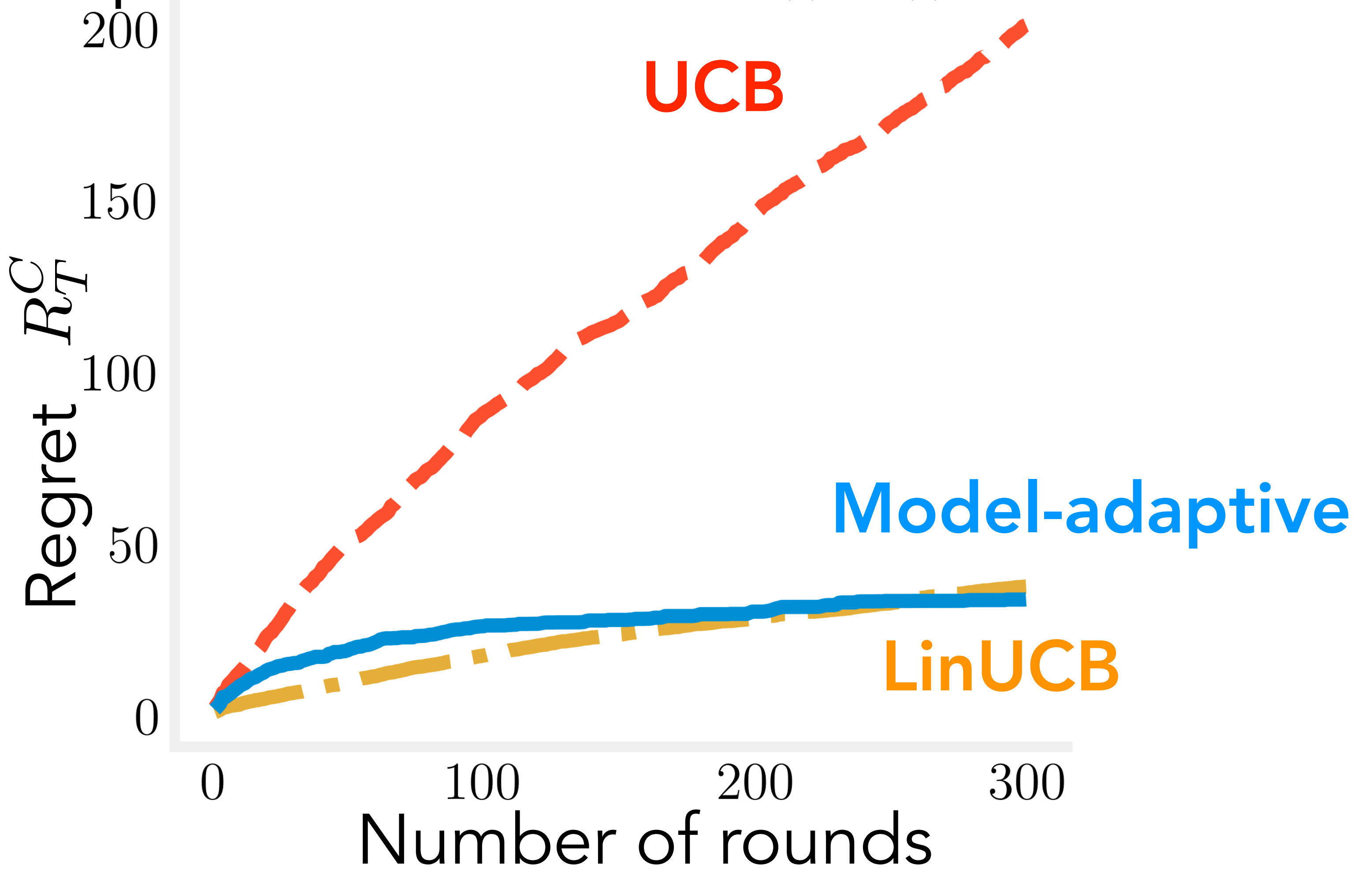
Model selection, Objective 1 (optimal)**:** Design a single algorithm that achieves regret rates

Simple model **(SM):**	$R_T^S = \mathcal{O}(\sqrt{KT})$
Complex model **(CM):**	$R_T^C = \mathcal{O}((\sqrt{d} + \sqrt{K})\sqrt{T})$

Simple model **(SM)**: $||\theta^*||_2^2 \sim 0$



Complex model **(CM)**: $||\theta^*||_2^2 \gg 0$



The model selection problem

Setting: K-armed linear contextual bandit problem…with potential simple MAB structure

(unknown) mean of arm k

(unknown) parameter $(\in \mathbb{R}^d)$

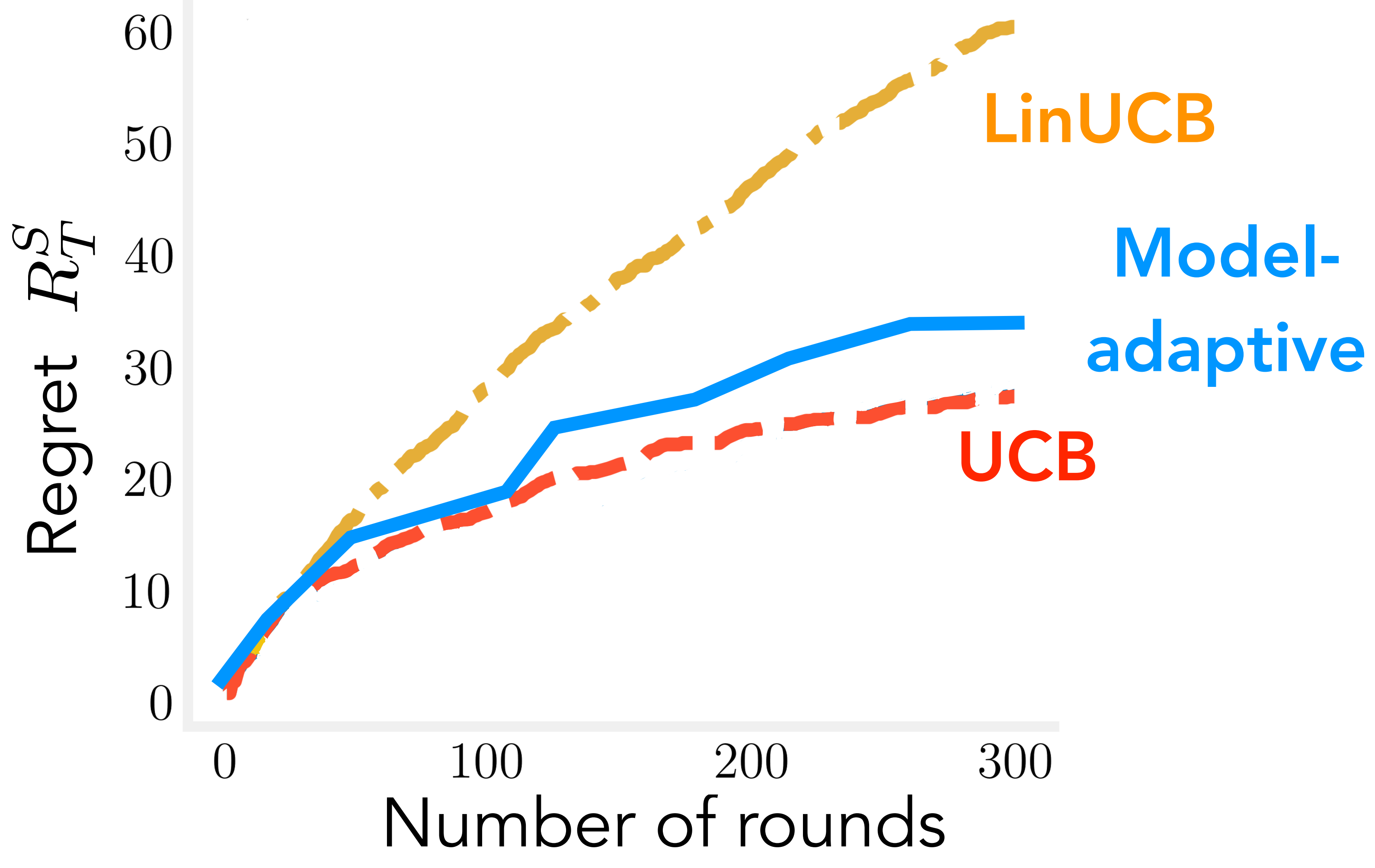
noise

$$G_{i,t} = \mu_i + \langle \boldsymbol{x}_{i,t}, \boldsymbol{\theta^*} \rangle + W_{i,t}$$

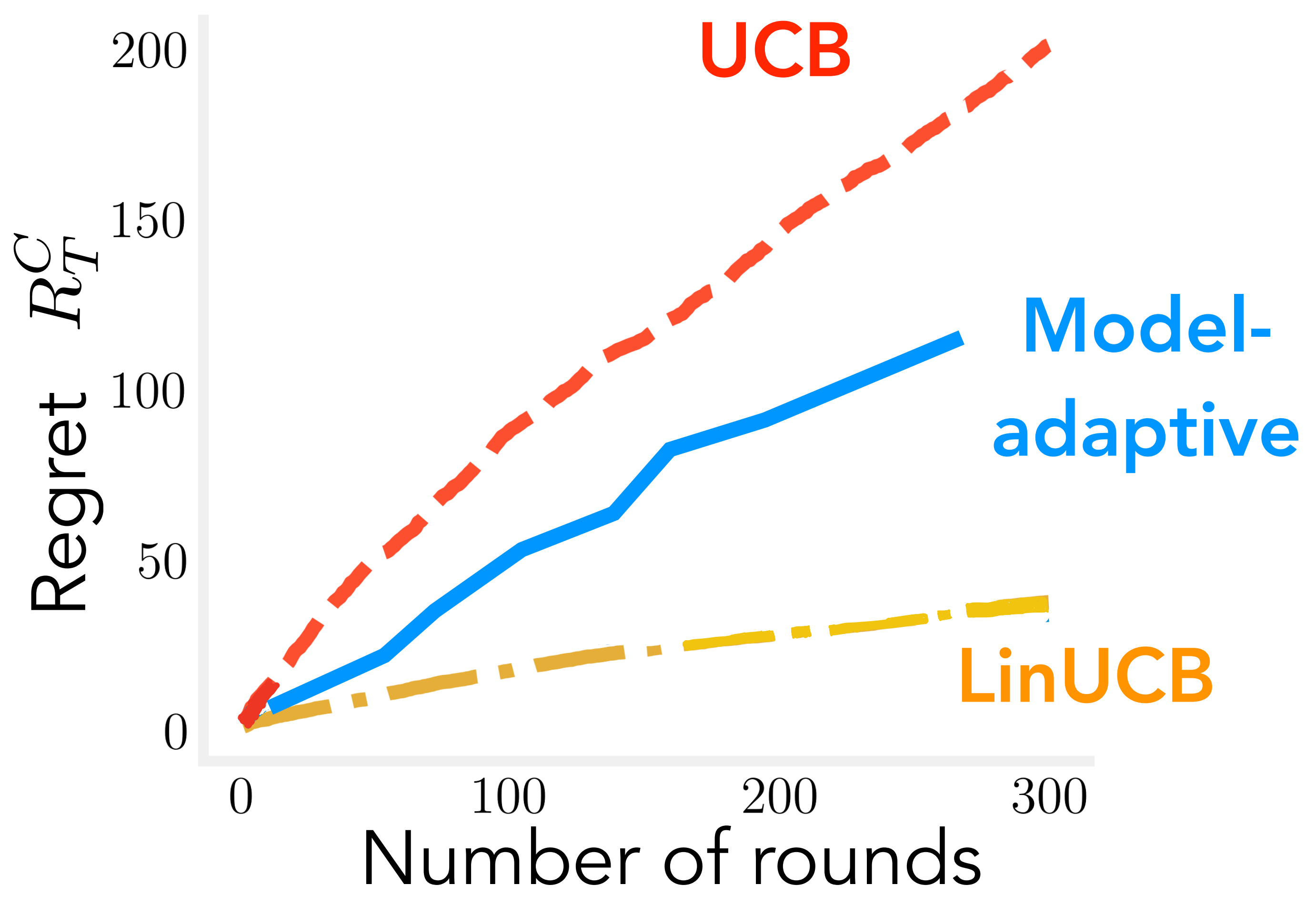
Reward of arm k in round t

Context for arm k $(\in \mathbb{R}^d)$

Simple model **(SM)**: $||\theta^*||_2^2 \sim 0$



Complex model **(CM)**: $||\theta^*||_2^2 \gg 0$



Model selection, Objective 2 (suboptimal but non-trivial)**:** Design a single algorithm that achieves regret rates (for $\alpha < 1/2$)

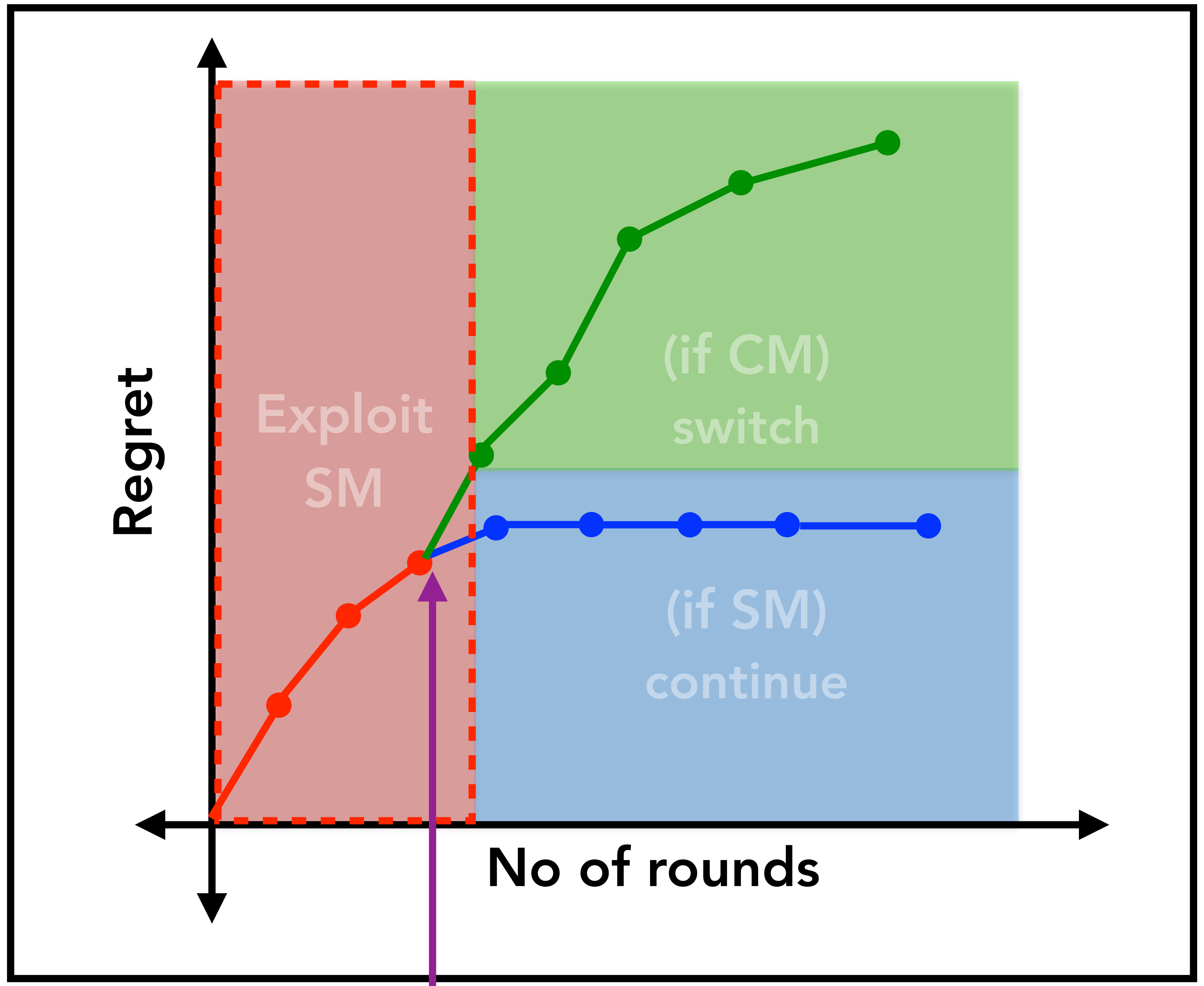
Simple model **(SM):** $R_T^S = \mathcal{O}(K^\beta T^{1-\alpha})$

Complex model **(CM):** $R_T^C = \mathcal{O}(K^\beta d^\alpha T^{1-\alpha})$

Existing algorithms for model selection (and their limitations)

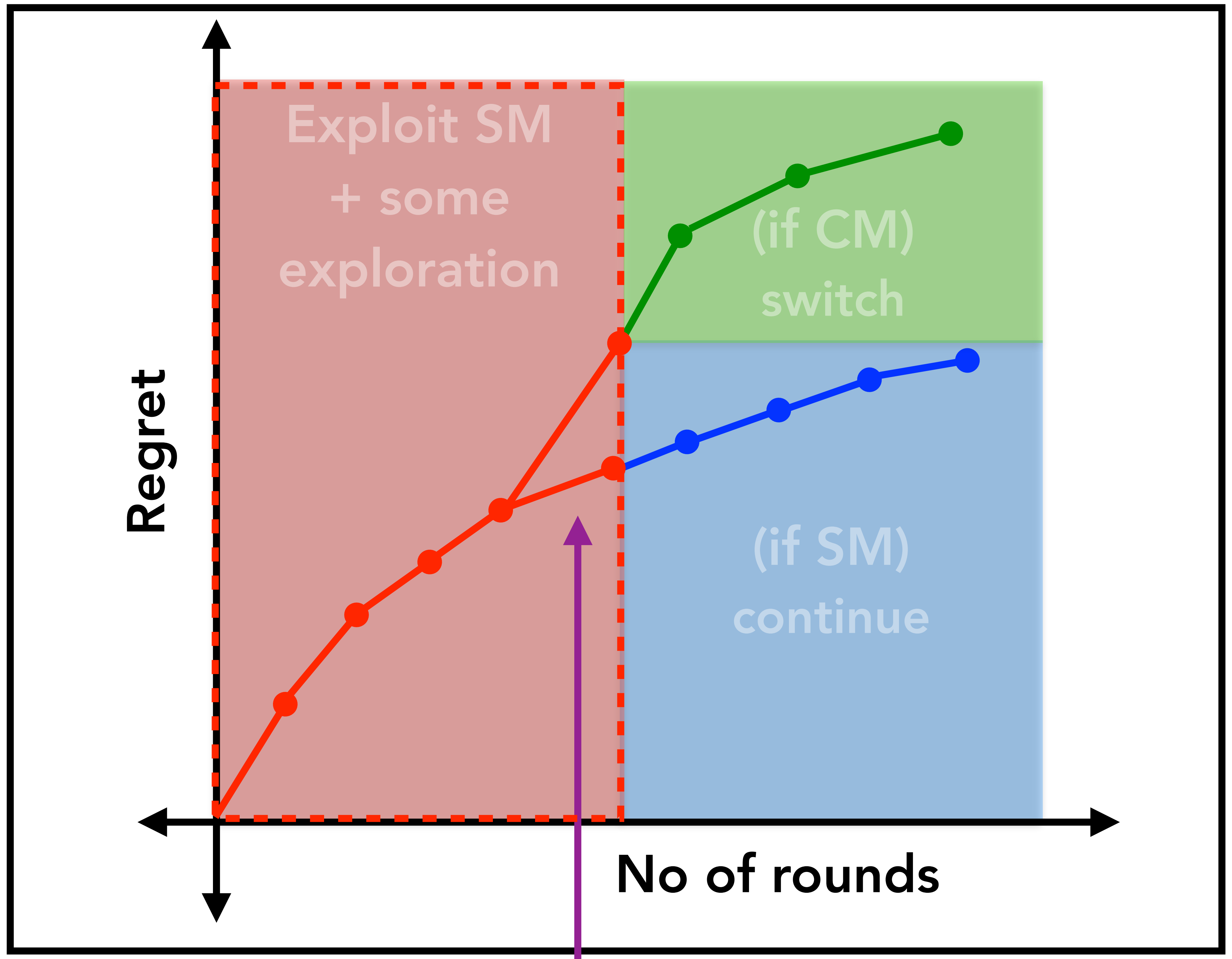
Meta exploration-vs-exploitation tradeoff: ensure success of test v.s. exploit simple model

OSOM (Chatterji, Muthukumar and Bartlett, AISTATS 2020)



Statistical test on:
CM regret under SM algorithm

ModCB (Foster, Krishnamurthy and Luo, NeurIPS 2019)



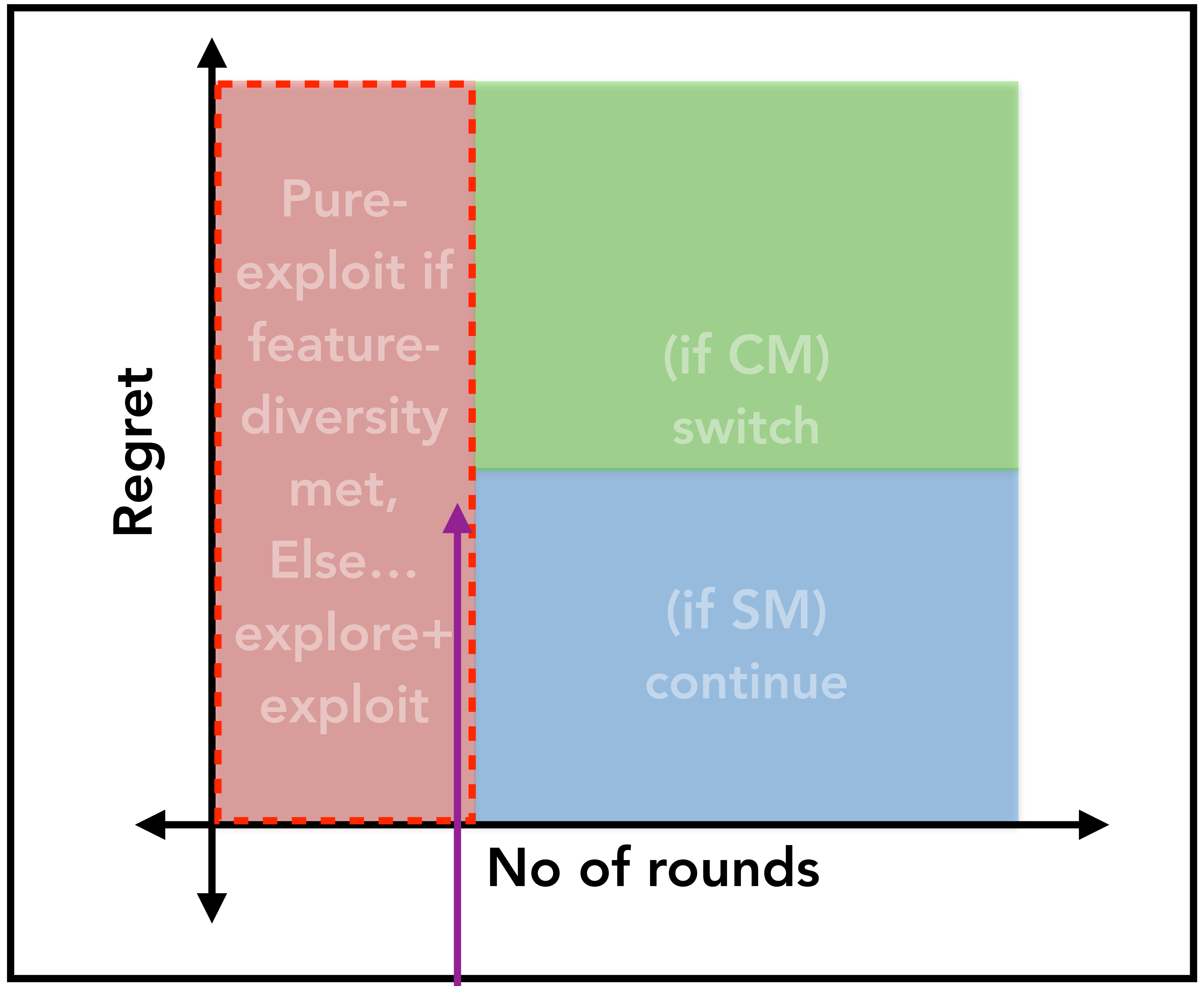
Statistical test on:
(upper bound on) gap between
best performance of CM/SM

Achieves Objective 1 only under feature diversity condition **for all arms** (very strong)

Achieves Objective 2 (for $\alpha = 1/3$), but only under feature diversity condition **averaged over arms**

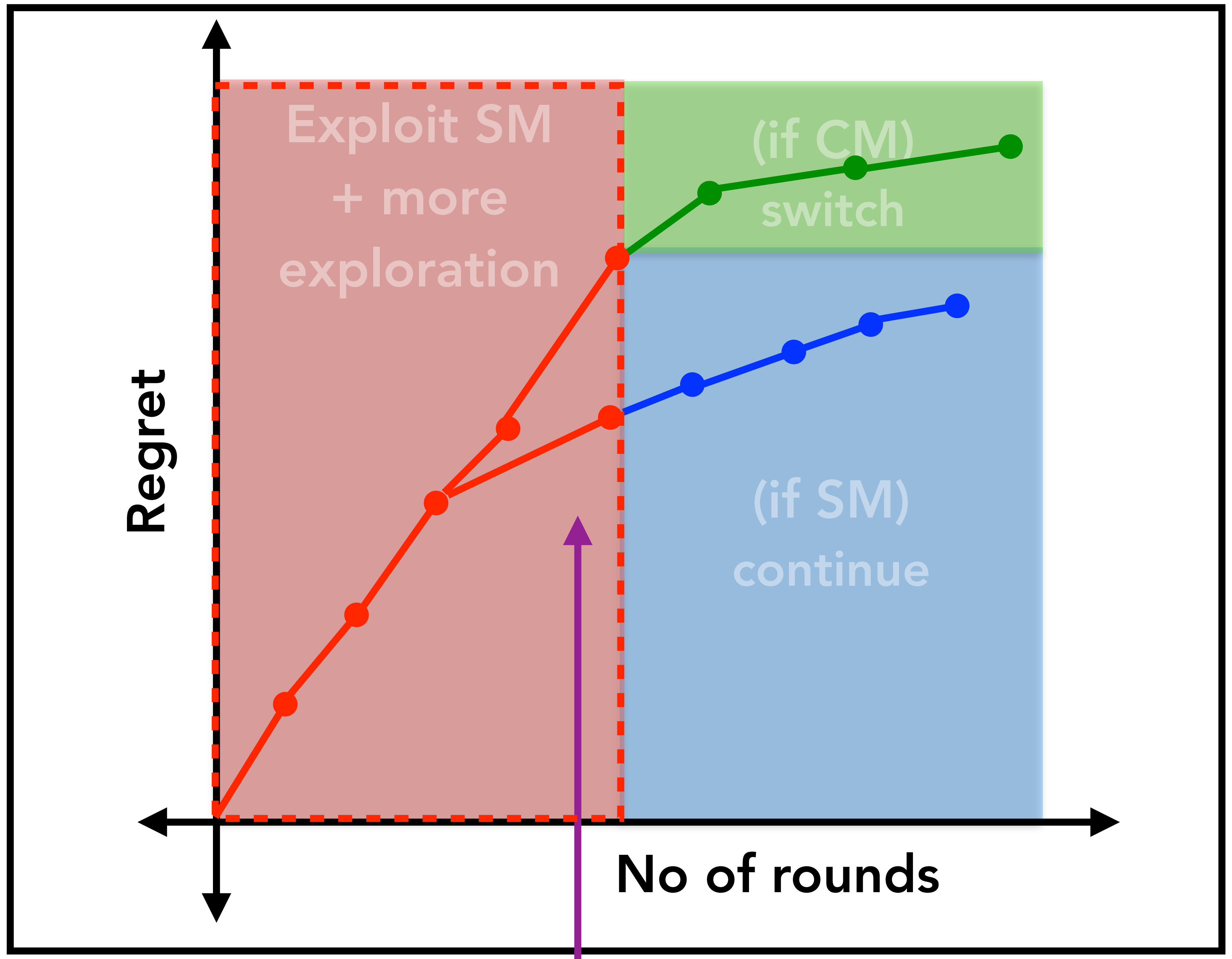
Our [universal, data-adaptive] algorithms

ModCB.A (adaptive)



Statistical test on:
(upper bound on) gap between best performance of CM/SM

ModCB.U (universal)



New universal statistical test on:
(upper bound on) gap between best performance of CM/SM

Achieves Objective 1 under arm-specific diversity **and Objective 2** (for $\alpha = 1/3$) under arm-averaged diversity

Achieves Objective 2 (for $\alpha = 1/6$) under **no feature diversity conditions whatsoever**

Summary of main results

Algorithm	Obj. 1 (optimal rates)	Obj. 2 ($d^\alpha T^{1-\alpha}$ rates)	context assumption
OSOM (Chatterji et al., 2020)	Yes	Yes ($\alpha = 1/2$)	$\forall i \in [K] : \boldsymbol{\Sigma}_i \succeq \gamma \boldsymbol{I}_d$
ModCB (Foster et al., 2019)	No	Yes ($\alpha = 1/3$)	$\boldsymbol{\Sigma} \succeq \gamma \boldsymbol{I}_d$
ModCB.U	No	Yes ($\alpha = 1/6$)	iid contexts only
Corral-style	No	No	iid contexts only

Table of results, **universality**

Algorithm	Arm-specific diversity	Arm-averaged diversity
OSOM	$\log(T)/\text{gap}$ and $\sqrt{dT}$	None
ModCB	$T^{2/3}$ and $d^{1/3}T^{2/3}$	$T^{2/3}$ and $d^{1/3}T^{2/3}$
ModCB.A	$\log(T)/\text{gap}$ and $\sqrt{dT}$	$T^{2/3}$ and $d^{1/3}T^{2/3}$

Table of results, **data-adaptivity**

Future work

- Universality *and* data-adaptivity in one algorithm
- Nested linear contextual bandits (beyond restrictive block-diagonal assumption)
- Beyond linear models
- Model selection under misspecification

Thank you!

Please see our poster and our full paper on arXiv: 2111.04688