UC SANTA BARBARA
Computer Science

$\tilde{O}(\sqrt{T})$ regret

Sample-Efficient Reinforcement Learning with loglog(T) Switching Cost

Optimal !

Dan Qiao (1), Ming Yin (1,2), Ming Min (2), Yu-Xiang Wang (1)

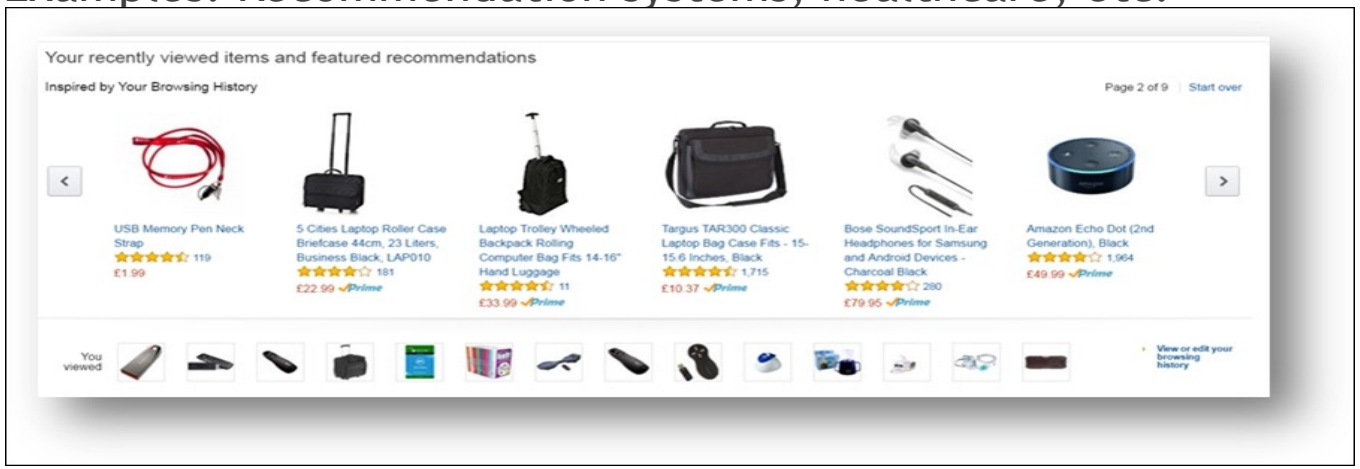
1. Department of Computer Science, UC Santa Barbara

2. Department of Statistics and Applied Probability, UC Santa Barbara

UC SANTA BARBARA
Department of Statistics and Applied Probability

Motivation and goal

- Motivation: In many real-world reinforcement learning tasks, it is costly to run fully adaptive algorithms that update the exploration policy frequently.
- Examples: Recommendation systems, healthcare, etc.



- Our goal: Minimize the number of policy switching while maintaining

(nearly) the same regret bounds as its fully-adaptive counterparts.

Problem setup: Finite Horizon Episodic MDP with Low Switching Cost

- Discrete/ finite state and actions, i.e., Tabular setting.
- Nonstationary (aka. time-inhomogeneous).
- Minimize regret and global switching cost, which is defined as:

$$N_{switch} := \sum_{k=1}^{K-1} \mathbb{1}\{\pi_k \neq \pi_{k+1}\}.$$

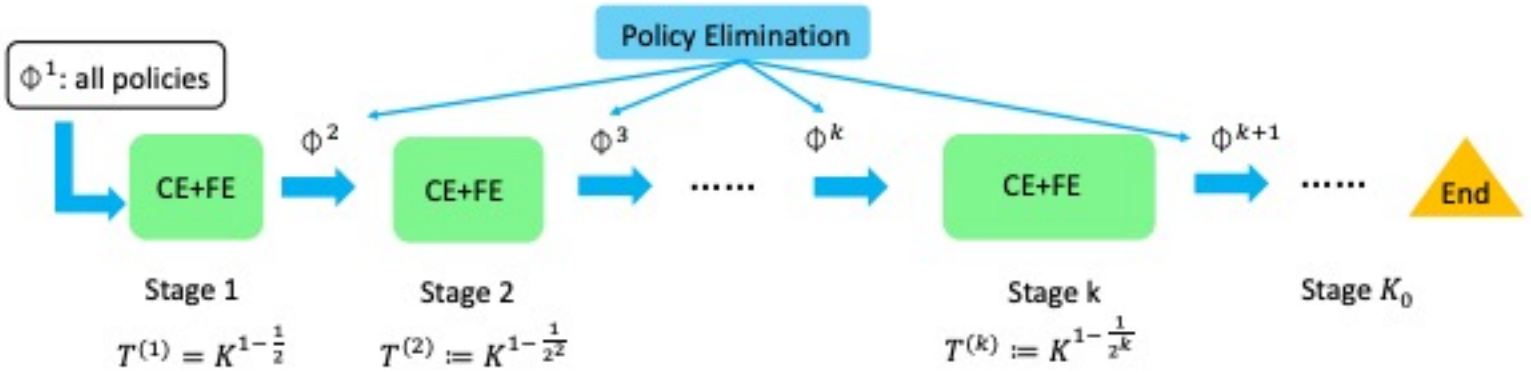
Summary of results: Upper and lower bounds

- We propose a new algorithm based on stage-wise exploration and adaptive policy elimination that achieves a regret of $\tilde{O}(\sqrt{H^4S^2AT})$ while requiring a switching cost of only O(HSA loglog T).
- A reward-free exploration algorithm with a switching cost of O(HSA).
- Lower bound: Any no-regret algorithm must have a switching cost of Ω(HSA).
- Lower bound: Any $\tilde{O}(\sqrt{T})$ regret algorithm must incur a switching cost of Ω(HSA loglog T).

Comparison to previous works

Algorithms for regret minimization	*Regret*	*Switching cost*
UCB2-Bernstein [Bai et al., 2019]	$\tilde{O}(\sqrt{H^3SAT})$	Local: $O(H^3SA\log T)$
UCB-Advantage [Zhang et al., 2020c]	$\tilde{O}(\sqrt{H^2SAT})$	Local: $O(H^2SA\log T)$
Algorithm 1 in [Gao et al., 2021] *	$\tilde{O}(\sqrt{d^3H^3T})$	Global: $O(dH\log T)$
APEVE (Our Algorithm 1)	$\tilde{O}(\sqrt{H^4S^2AT})$	Global: $O(HSA\log\log T)$
Explore-First w. LARFE (Our Algorithm 4)	$\tilde{O}(T^{2/3}H^{4/3}S^{2/3}A^{1/3})$	Global: $O(HSA)$
Lower bound (Our Theorem 4.2)	if $\tilde{O}(\sqrt{T})$ ("Optimal regret")	Global: $\Omega(HSA\log\log T)$
Lower bound (Our Theorem 4.3)	if $o(T)$ ("No regret")	Global: $\Omega(HSA)$
Algorithms for reward-free exploration	*Sample (episode) complexity*	*Switching cost*
Algorithm 2&3 in [Jin et al., 2020a]	$\tilde{O}(\frac{H^5S^2A}{\epsilon^2})$	Global: $\tilde{O}(\frac{H^7S^4A}{\epsilon})^{\ddagger}$
RF-UCRL [Kaufmann et al., 2021]	$\tilde{O}(\frac{H^4S^2A}{\epsilon^2})$	Global: $\tilde{O}(\frac{H^4S^2A}{\epsilon^2})$
RF-Express [Ménard et al., 2021]	$\tilde{O}(\frac{H^3S^2A}{\epsilon^2})$	Global: $\tilde{O}(\frac{H^3S^2A}{\epsilon^2})$
SSTP [Zhang et al., 2020b]	$\tilde{O}(\frac{S^2A}{\epsilon^2})^{\star}$	Global: $\tilde{O}(SA\log(\frac{S^2A}{\epsilon^2}))^{\dagger}$
Algorithm 3&4 in [Huang et al., 2022]	$\tilde{O}(dH(\frac{d^{3c_K}H^{6c_K+1}}{\epsilon^{2c_K}})^{\frac{1}{c_K-1}})$	Global: c_KdH+1
LARFE (Our Algorithm 4)	$\tilde{O}(\frac{H^5S^2A}{\epsilon^2})$	Global: $O(HSA)$

Our approach: Policy elimination with two-stage exploration



- Crude layer-wise exploration
  - Get a crude approximation of the model.
- Fine stagewise exploration
  - Use the crude model to identify “representative” policies.
- Policy elimination
  - Disqualify policies that are certifiably suboptimal.

Illustration of the policy elimination step

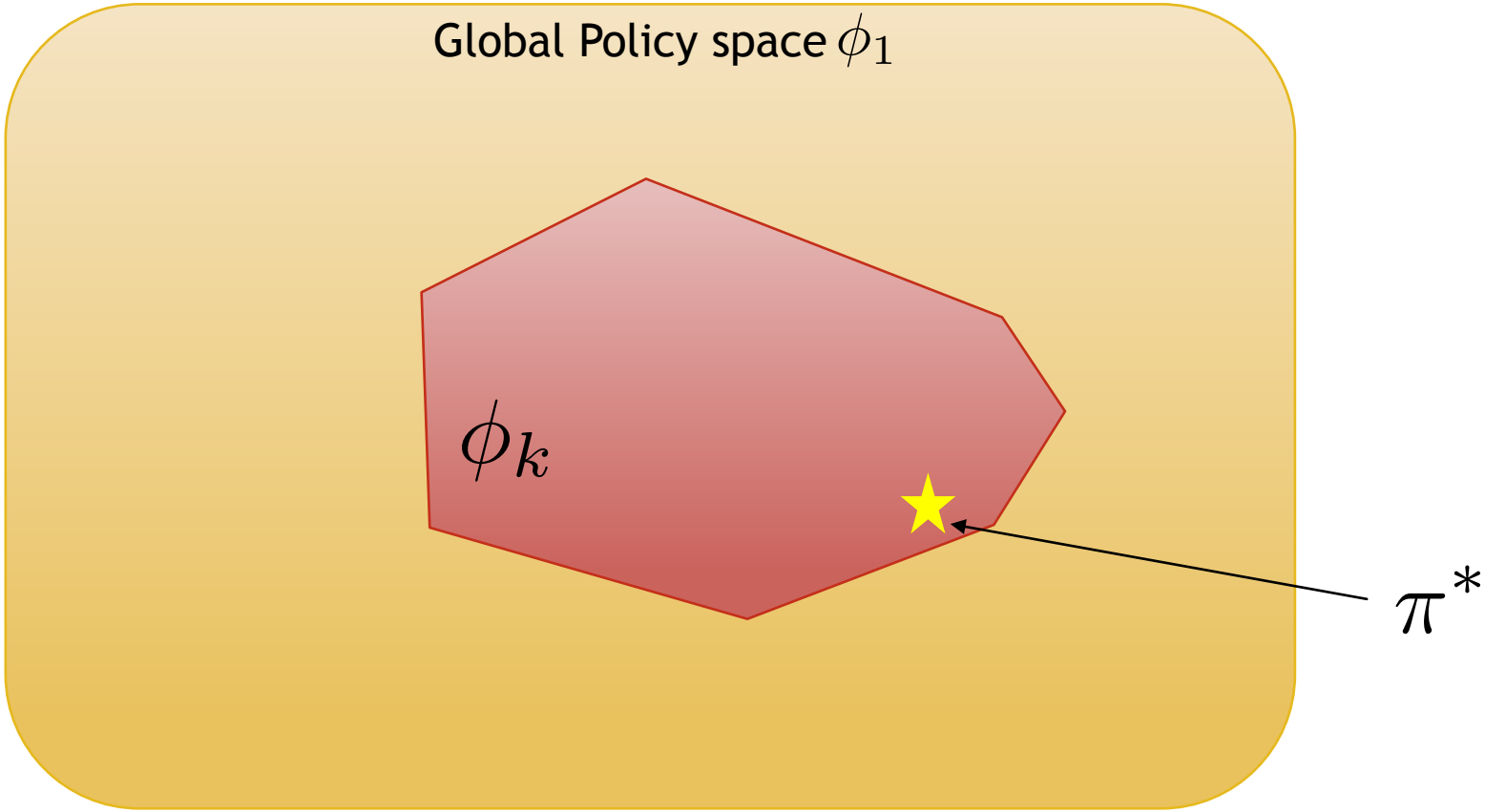


Illustration of the policy elimination step

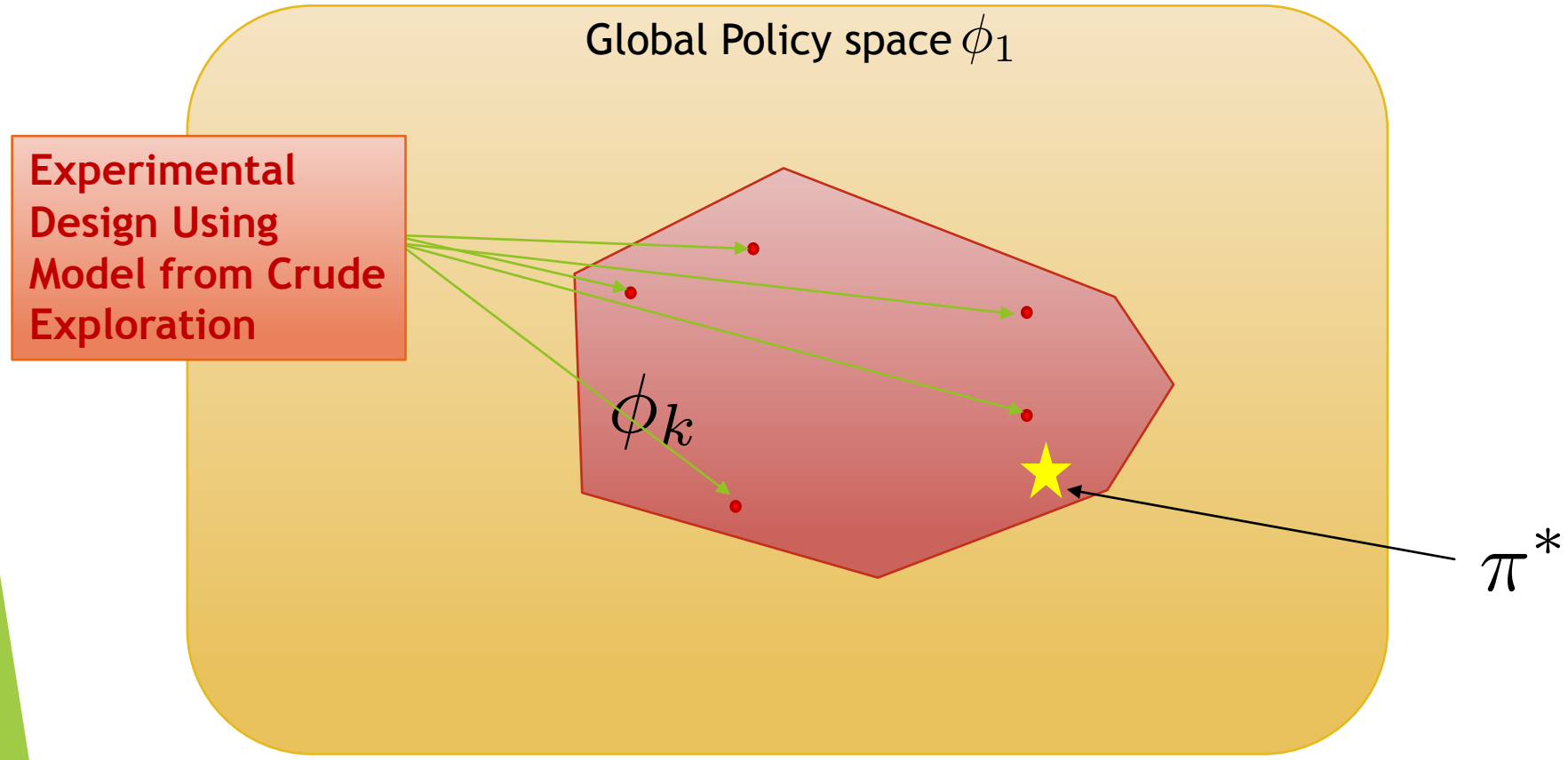


Illustration of the policy elimination step

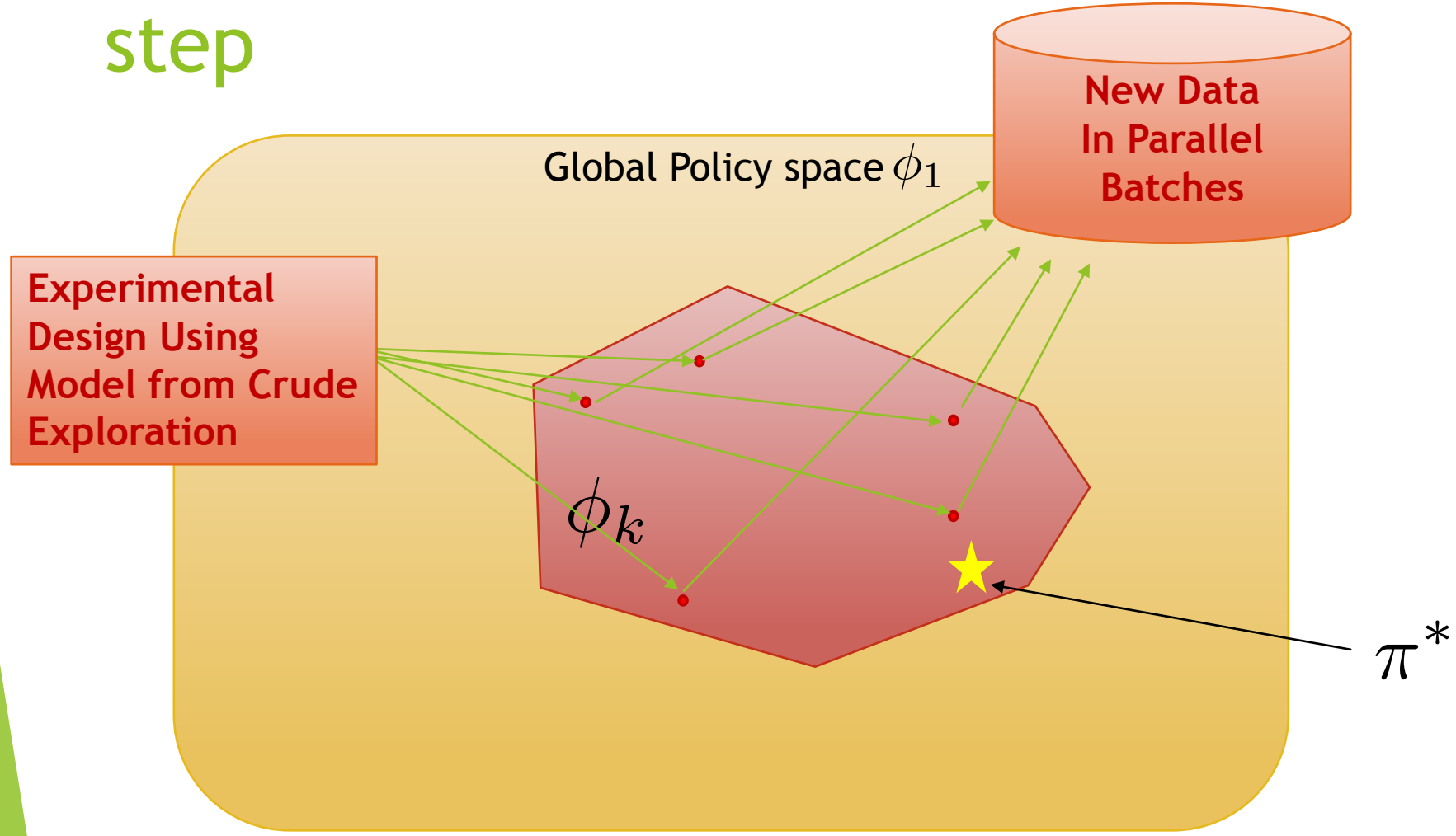
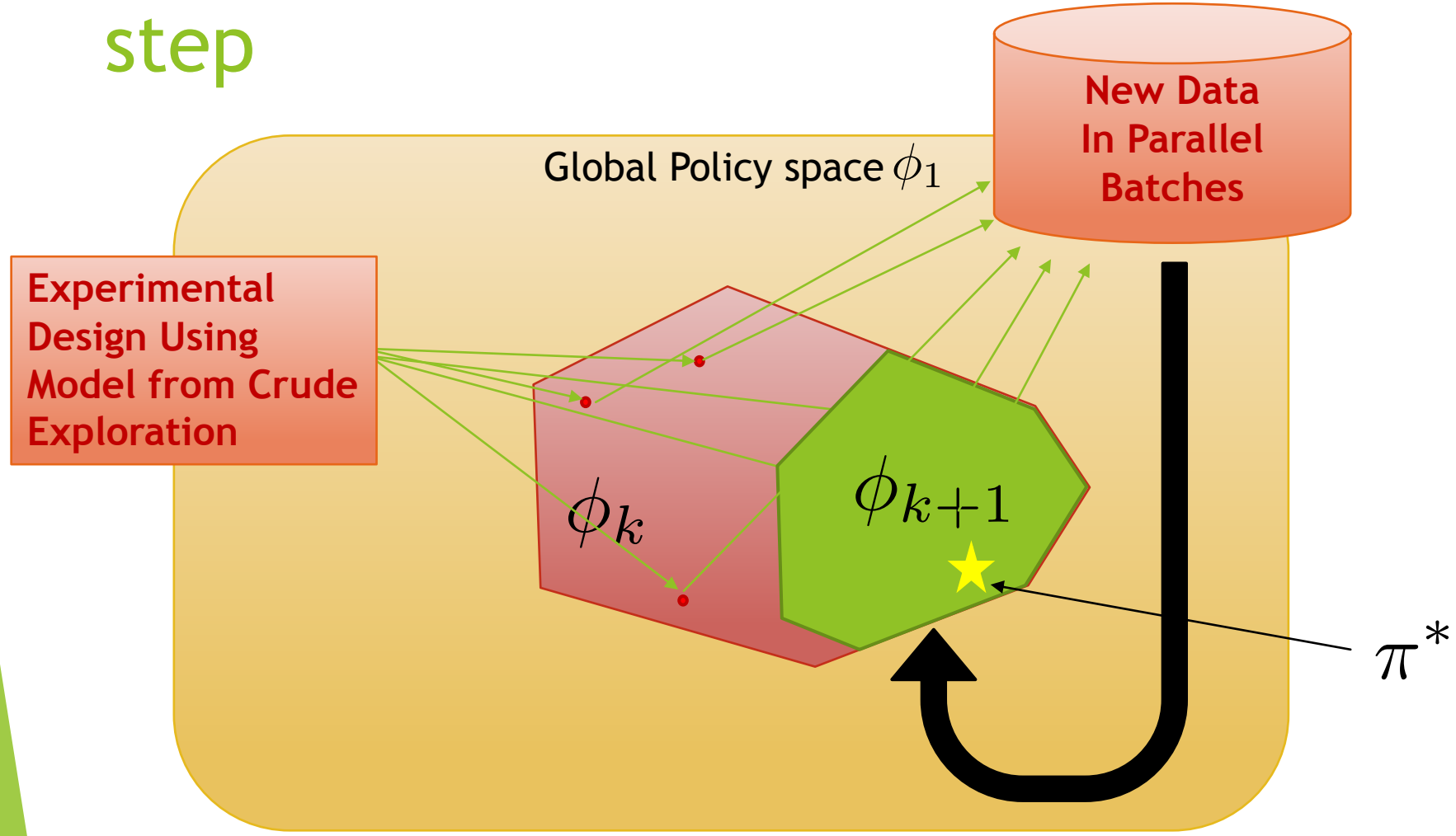


Illustration of the policy elimination step



Take home message

- A setting between online and offline RL.
- We characterized the optimal switching cost among algorithms with optimal regret.
- Promising new directions with many practical / theoretical opportunities.
- Welcome to my poster!