A Model Selection Approach for Corruption Robust Reinforcement Learning

Chen-Yu Wei, (USC)

Chris Dann, (Google)

Julian Zimmert (Google)

Online Decision Making

Given: policy set Π

For $t = 1, \ldots, T$:

The learner chooses $\pi_t \in \Pi$

The learner receives $r_t \in [0,1]$ with $\mathbb{E}[r_t] = f(\pi_t)$

Examples: multi-armed bandits, contextual bandits, episodic MDP, etc.

$$\text{Reg} = \sum_{t=1}^{T} \left(\max_{\pi \in \Pi} f(\pi) - f(\pi_t) \right)$$

Online Decision Making with Corrupted Samples

Given: policy set Π

For $t = 1, \ldots, T$:

The adversary decides f_t: $\Pi \to [0, 1]$

The learner chooses $\pi_t \in \Pi$

The learner receives $r_t \in [0, 1]$ with $\mathbb{E}[r_t] = f_t(\pi_t)$

Examples: multi-armed bandits, contextual bandits, episodic MDP, etc.

$$\text{Reg} = \sum_{t=1}^{T} \left(\max_{\pi \in \Pi} f(\pi) - f(\pi_t) \right)$$

$C := \sum_{t=1}^{T} \max_{\pi} |f_t(\pi) - f(\pi)|$ (For MDPs, $C :=$ transition corruption $+$ reward corruption)

C is unknown to the learner

Corrupted Multi-Armed Bandits and MDPs

(omitting log terms, #actions, #states)

Lykouris et al. (2018)

$$C \min\left(\frac{1}{\Delta_{\mathcal{A}}}, \sqrt{T}\right)$$

↓

Lykouris et al. (2021)

$$C \min\left(\frac{1}{\Delta_{\mathcal{A}}}, \sqrt{T}\right) + C^2$$

Gupta et al. (2019)

$$\min\left(\frac{1}{\Delta_{\mathcal{A}}}, \sqrt{T}\right) + C$$

↓

Chen et al. (2021)

$$\min\left(\frac{1}{\Delta_{\Pi}}, \sqrt{T}\right) + C^2$$

(only allow $C \le \sqrt{T}$)

Zimmert and Seldin (2019)

$$\min\left(\frac{1}{\Delta_{\mathcal{A}}}, \sqrt{T}\right) + C$$

↓

Jin et al. (2021)

$$\min\left(\frac{1}{\Delta_{\mathcal{A}}}, \sqrt{T}\right) + C$$

(only allow corruption in reward)

$\Delta_{\mathcal{A}} =$ action value gap

$\Delta_{\Pi} =$ policy value gap $= f(\pi^\star) - \max_{\pi \neq \pi^\star} f(\pi)$

$\Delta_{\Pi} \le \Delta_{\mathcal{A}}$

Observations

There are obstacles in extending from MAB to MDP: if transition is corrupted, previous works only tolerate $C \leq \sqrt{T}$

But if C is known, the tight bound $\min\left(\frac{1}{\Delta_{\mathcal{A}}}, \sqrt{T}\right) + C$ can be easily achieved.

(Lykouris et al. (2021): UCB + widened confidence interval)

→ Difficulties come from "C is unknown"

Contributions

Reduction from "unknown C" to "known C"

$\sqrt{T} + C$ for known C $\Rightarrow$ **Meta Algorithm 1** $\Rightarrow$ $\sqrt{T} + C$ for unknown C

$\min\left(\frac{1}{\Delta_\Pi}, \sqrt{T}\right) + C$ for known C $\Rightarrow$ **Meta Algorithm 2** $\Rightarrow$ $\min\left(\frac{1}{\Delta_\Pi}, \sqrt{T}\right) + C$ for unknown C

$$\Delta_\Pi = f(\pi^\star) - \max_{\pi \neq \pi^\star} f(\pi)$$

Implications of the Reduction

Tabular MDP

Algorithm	Regret	Limitations
Lykouris et al., 2021	$C\min\{\frac{1}{\Delta_{\mathcal{A}}}, \sqrt{T}\} + C^2$	
Chen et al., 2021	$\min\{\frac{1}{\Delta_{\Pi}}, \sqrt{T}\} + C^2$	computationally inefficient
Jin et al., 2021	$\min\{\frac{1}{\Delta_{\mathcal{A}}}, \sqrt{T}\} + C$	only for corrupted reward
Ours	$\min\{\frac{1}{\Delta_{\Pi}}, \sqrt{T}\} + C$	

First result that tolerate all $C \leq T$ with corrupted transition without knowing C

Linear bandit

Algorithm	Regret	Limitations
Li et al., 2019	$\frac{1}{\Delta^2} + \frac{C}{\Delta}$	
Bogunovic et al., 2020/2021	$\sqrt{T} + C^2$ and $C\sqrt{T}$	
Lee, et al., 2021	$\min\{\frac{1}{\Delta}, \sqrt{T}\} + C$	only for linearized corruption
Ours	$\min\{\frac{1}{\Delta}, \sqrt{T}\} + C$	

Linear contextual bandit

Foster et al. 2020	$\sqrt{CT}$	
Ours	$\sqrt{CT}$	
	$\sqrt{T} + C$	Computationally inefficient

Linear MDP

Lykouris et al., 2021	$C^2\sqrt{T}$	
Ours	$\sqrt{CT}$	
	$\sqrt{T} + C$	Computationally inefficient

MDPs with low Bellman-Eluder dimension

Algorithm	Regret	Limitations
Ours	$\sqrt{CT}$	computationally inefficient

Meta Algorithm 1 $(\sqrt{T} + C)$

Simplified Formulation

Assume that either $C = c_1$ or $C = c_2$ $(\sqrt{T} \leq c_1 \ll c_2)$.

How to achieve Reg $\lesssim \sqrt{T} + C$?

BaseAlg(c)

If the total corruption $C \leq c$, then

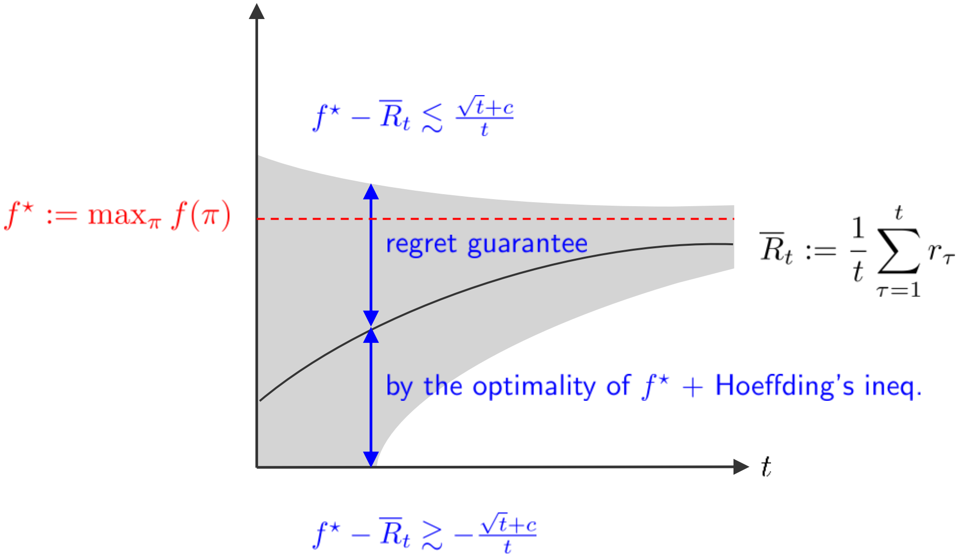
best policy's total reward in $[1...t]$ - learner's total reward in $[1...t] \lesssim \sqrt{t} + c$.

Idea 1. Confidence region via regret bounds

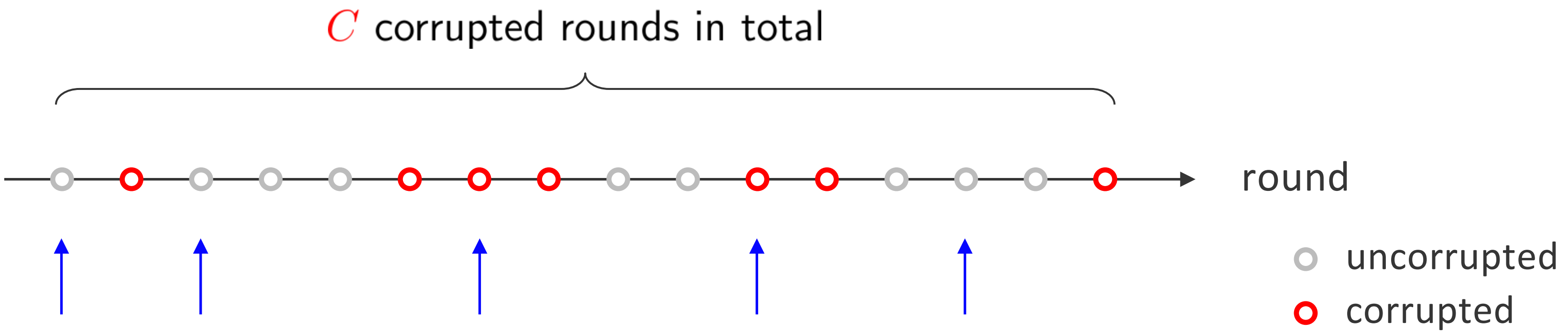
Idea 2. Robustness through sub-sampling

Idea 1. Confidence region via regret bounds

Suppose that we run **BaseAlg**(c) under $C \le c$



Idea 2. Robustness through sub-sampling (Lykouris et al., 2018)

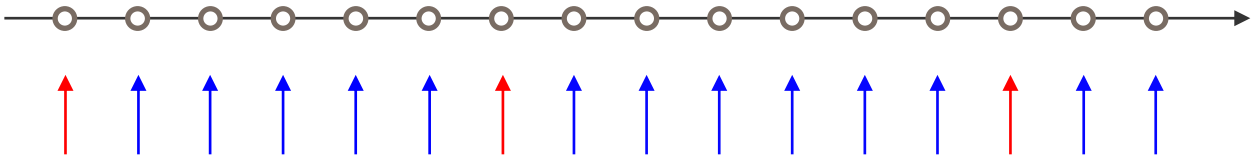


If a learner executes every round only w.p. α
(and skip w.p. $1-\alpha$)

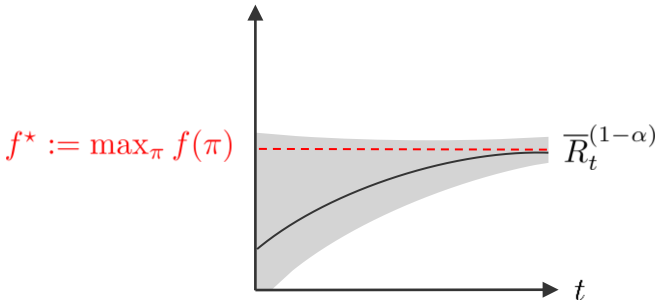
Then in expectation, the learner only *experiences* $\alpha \cdot C$ corrupted rounds

$\alpha := \frac{c_1}{c_2} \ll 1$

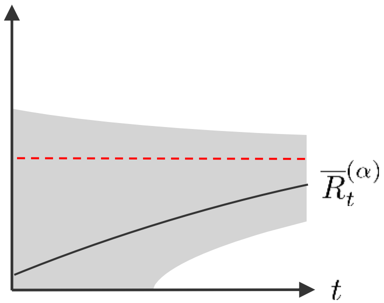
Execute two independent BaseAlg(c_1), w.p. $1-\alpha$ and α respectively.



$\boxed{\text{If } C = c_1}$



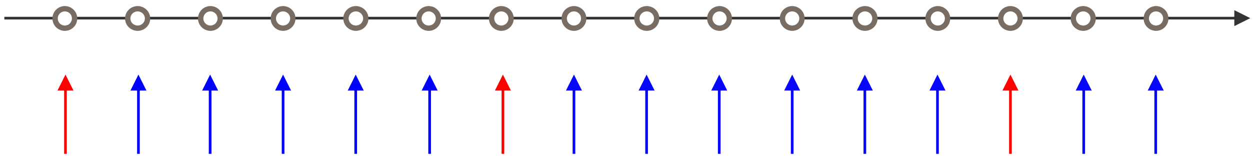
Fast learner



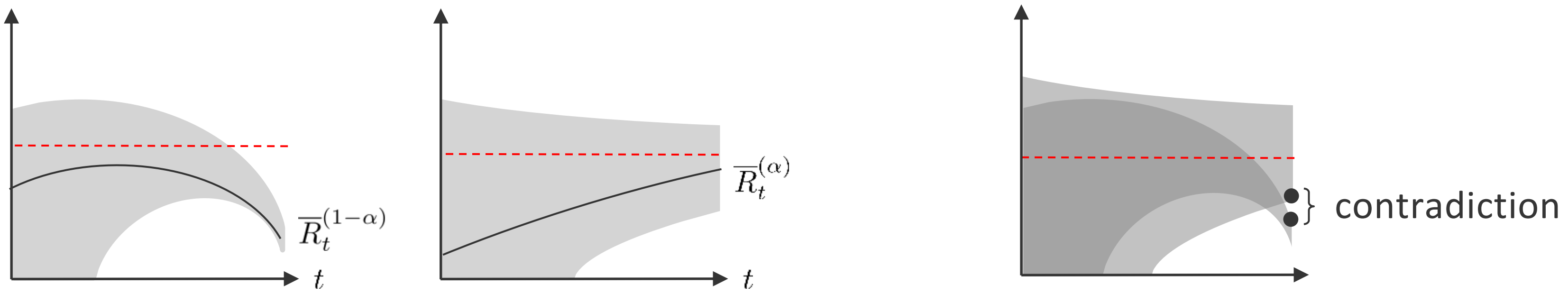
Slow learner

$\alpha := \frac{c_1}{c_2} \ll 1$

Execute two independent BaseAlg(c_1), w.p. $1-\alpha$ and α respectively.



$\boxed{\text{If } C = c_2}$ Slow learner only experiences corruption $\leq \alpha c_2 = c_1$



Fast learner

Slow learner

Algorithm

Initiate two independent BaseAlg(c_1).
Execute them w.p. $\alpha = \frac{c_1}{c_2}$ (slow learner) and $1 - \alpha$ (fast learner).

At every t, if $\overline{R}_t^{\text{fast}} \lesssim \overline{R}_t^{\text{slow}} - \Omega\left(\frac{1}{\sqrt{\alpha t}} + \frac{c_1}{t}\right)$, (must be $C = c_2$)
terminate the algorithms; start running BaseAlg(c_2).

Theorem

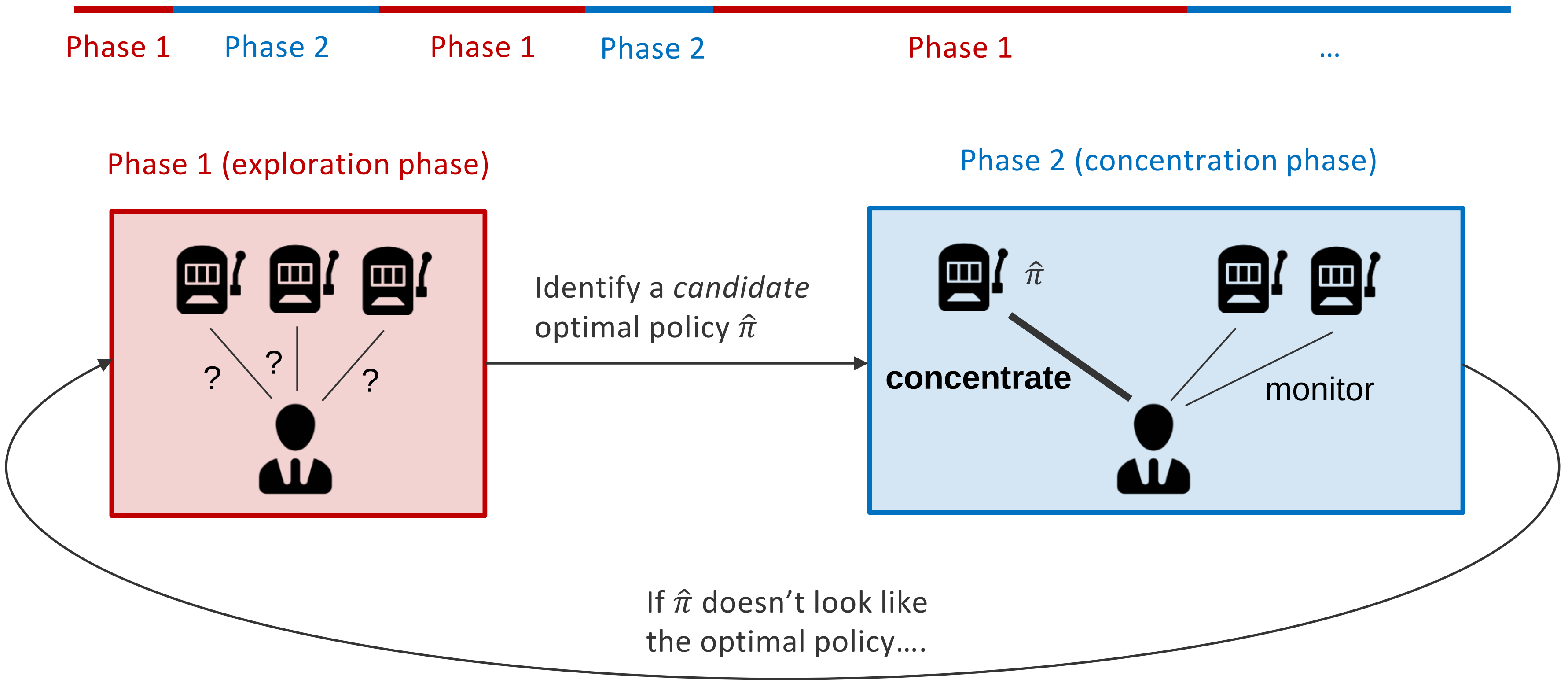
Let $C \in \{c_1, c_2\}$. The algorithm above ensures Regret $= \tilde{\mathcal{O}}(\sqrt{T} + C)$.

Meta Algorithm 2 $(\min\{\frac{1}{\Delta_\Pi}, \sqrt{T}\} + C)$

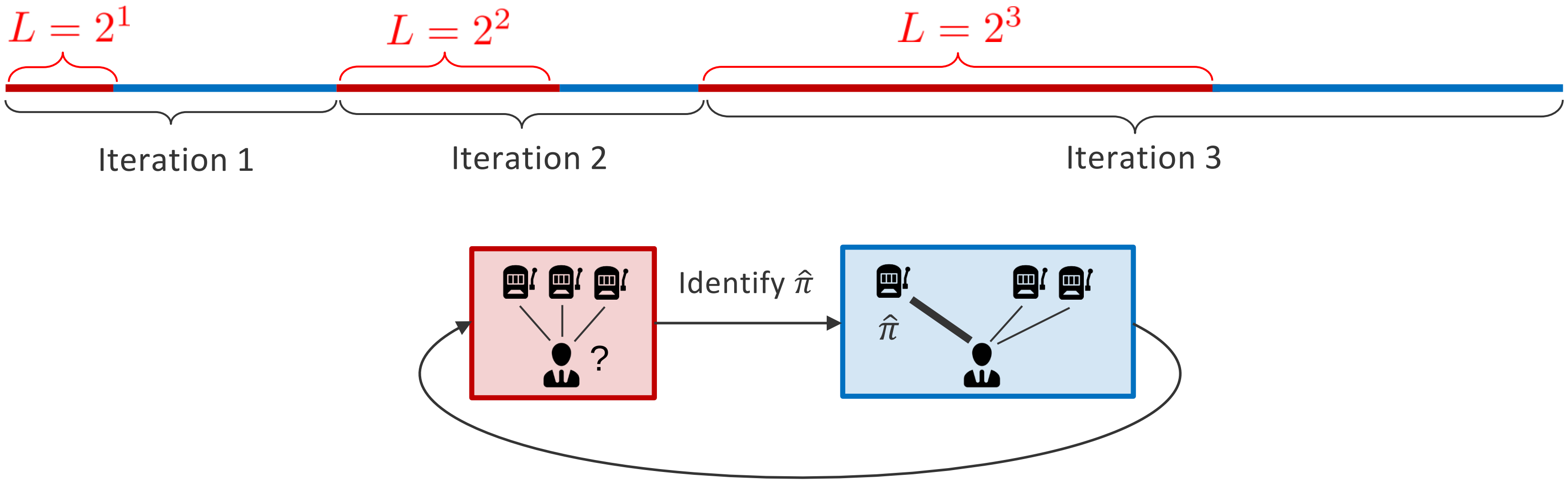
Limitation of Meta Algorithm 1

Always have $\mathrm{Reg} = \Omega\left(\sqrt{T}\right)$

(using Hoeffding's ineq. to construct confidence interval)



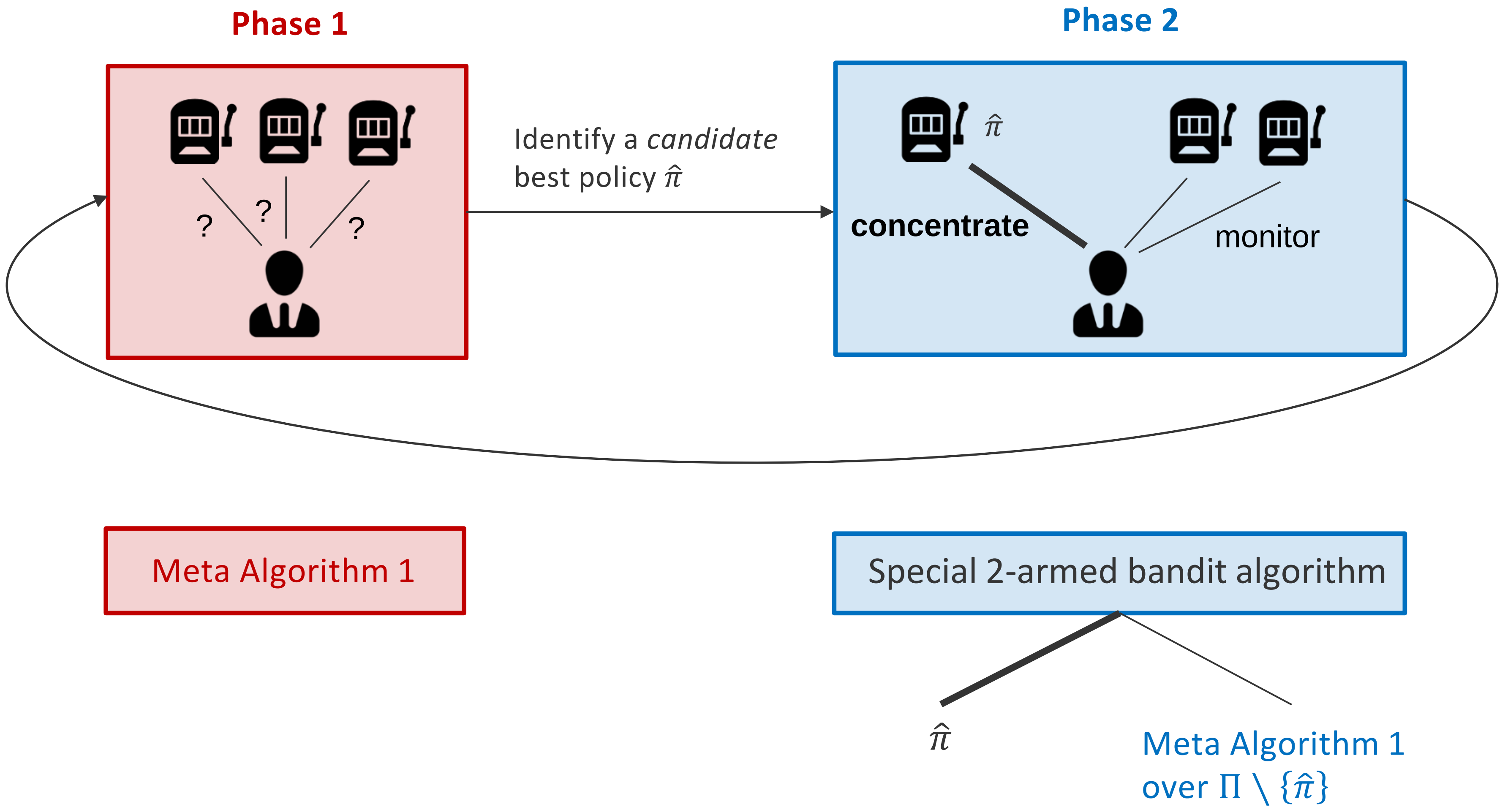
Inspired by [Bubeck and Slivkins 2012, Auer and Chiang, 2016, Lee et al., 2021]



Properties we show:

	Phase 1	Phase 2
In all iterations	Regret $\lesssim \sqrt{L} + C$	Regret $\lesssim \sqrt{L} + C$ (before termination)
If $\sqrt{L} \gtrsim \frac{1}{\Delta_\Pi} + C$	$\hat{\pi} = \pi^\star$ w.h.p.	Phase 2 **will not terminate**

Overall regret $\lesssim$ #iterations $\times \left(\sqrt{L_{\text{final-iteration}}} + C\right) \lesssim \log T \times \min\left(\frac{1}{\Delta_\Pi} + C, \sqrt{T} + C\right)$



Phase 1
Phase 2
?
?
?
Identify a *candidate* best policy $\hat{\pi}$
$\hat{\pi}$
concentrate
monitor
Meta Algorithm 1
Special 2-armed bandit algorithm
$\hat{\pi}$
Meta Algorithm 1 over $\Pi \setminus \{\hat{\pi}\}$

Summary and Open Problem

$\sqrt{T} + C$ for known C ⇨ **Meta Algorithm 1** ⇨ $\sqrt{T} + C$ for unknown C

$\min\left(\frac{1}{\Delta_\Pi}, \sqrt{T}\right) + C$ for known C ⇨ **Meta Algorithm 2** ⇨ $\min\left(\frac{1}{\Delta_\Pi}, \sqrt{T}\right) + C$ for unknown C

$\min\left(\frac{1}{\Delta_\mathcal{A}}, \sqrt{T}\right) + C$ or $\min\left(\text{inst}, \sqrt{T}\right) + C$

for unknown C?