Improved Sample Complexity for Incremental Autonomous Exploration in MDPs

Jean Tarbouriech[1,2], Matteo Pirotta[1], Michal Valko[3], Alessandro Lazaric[1]   [1] FACEBOOK AI   [2] Inria   [3] DeepMind

NEURAL INFORMATION PROCESSING SYSTEMS

Take-Away

What should an RL agent do in a *reward-free* and *open-ended* unknown environment?

▶ Our `DisCo` algorithm provably **1)** *discovers* all states within its *"reach"* in an *incremental* fashion, and **2)** learns a *near-optimal goal-conditioned policy* to reach *each* of them

▶ We provide theoretical analysis for concepts in deep RL such as exploration on the *frontier of the so far visited states*

Incremental Autonomous Exploration

▶ *Environment* $\mathcal{E}$: reward-free, possibly very large, resettable to s_0

▶ *Desired objective*: explore $\mathcal{E}$ and stop when:

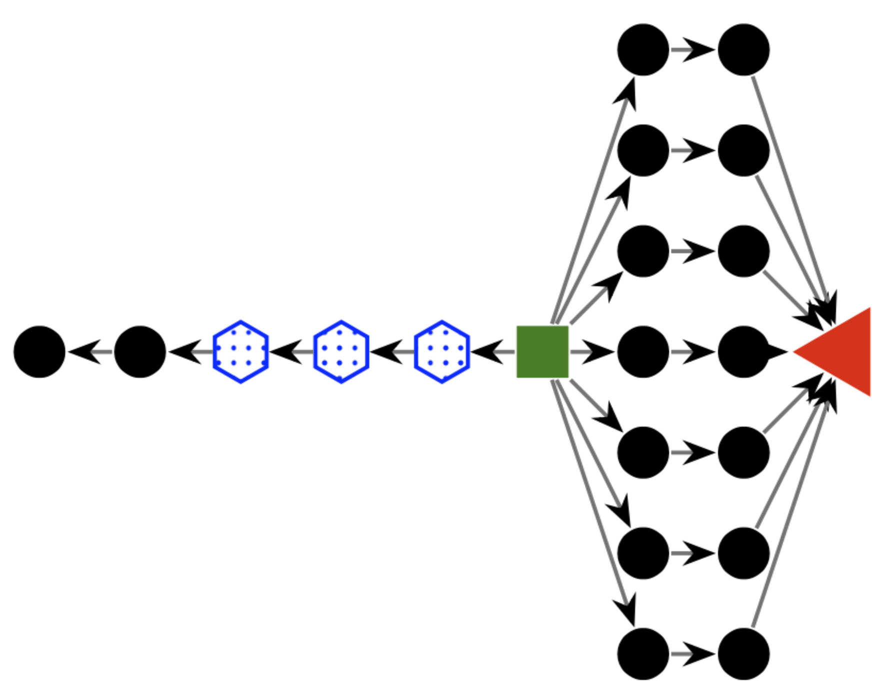
- it identifies all the *L-controllable* states
- it learns an ϵ-optimal goal-reaching policy for *each* of them

state s is L-controllable if: $\underbrace{\min_\pi V_\pi(s_0 \to s)}_{\text{shortest-path distance}} \leq L$

⚠ May require an *exponential* number of steps

☞ Find the *incrementally* controllable states

[Lim & Auer, COLT 2012]



$\mathcal{S}_L^{\rightarrow}$: set of incrementally L-controllable states
⚠ **unknown**

Objective: For *every* goal state $g \in \mathcal{S}_L^{\rightarrow}$, find a policy $\widehat{\pi}_g$ such that

$$V_{\widehat{\pi}_g}(s_0 \to g) \leq \min_{\pi \in \Pi(\mathcal{S}_L^{\rightarrow})} V_\pi(s_0 \to g) + \epsilon$$

$\Pi(\mathcal{S}_L^{\rightarrow})$: class of policies that take the RESET action in states outside of $\mathcal{S}_L^{\rightarrow}$

▶ *Tighter* variant of the objective originally considered by Lim & Auer

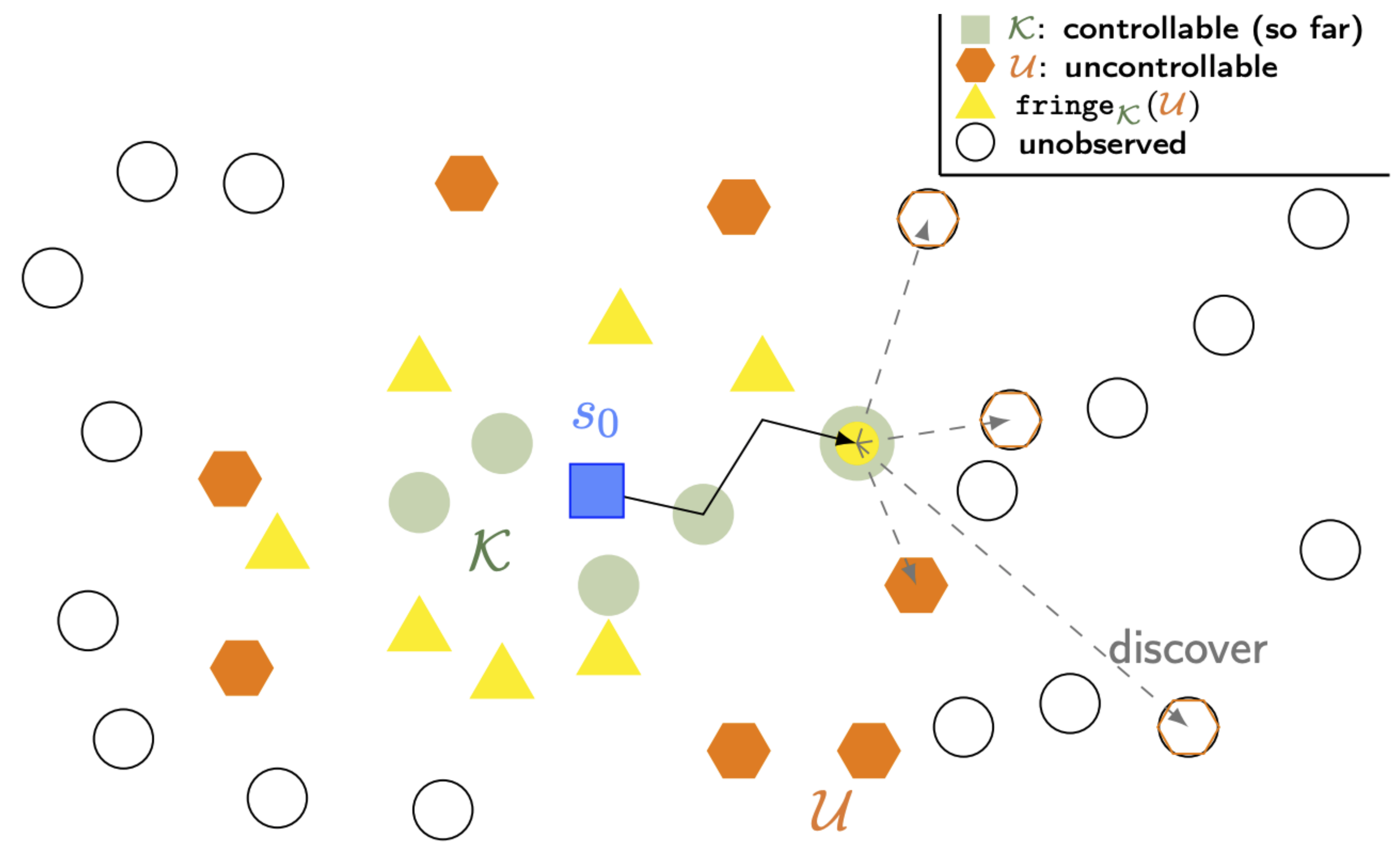
DisCo Algorithm — discover and control

[0] Initialize $\mathcal{K} \leftarrow \{s_0\}$, $\mathcal{U} \leftarrow \{\}$

[1] Execute goal-reaching π_g for each $g \in \mathcal{K}$ to *improve model estimate* and *discover new states* to add to $\mathcal{U}$

[2] *Compute optim. goal-reaching* $\pi_{\widetilde{g}}$ for each $\widetilde{g} \in \mathtt{fringe}_{\mathcal{K}}(\mathcal{U})$

[3] **If** $\widetilde{V}_{\pi_{\widetilde{g}}}(s_0 \to \widetilde{g}) \leq L$, then add $\widetilde{g}$ to $\mathcal{K}$ and go back to step [1]; **else** *terminate*



Sample Complexity Guarantee

`DisCo` requires

$$\widetilde{O}\left(\frac{L^5 \Gamma_{L+\epsilon} S_{L+\epsilon} A}{\epsilon^2}\right)$$

time steps to find policies $\{\widehat{\pi}_g\}_{g \in \mathcal{S}_L^{\rightarrow}}$

▷ $S_{L+\epsilon} = |\mathcal{S}_{L+\epsilon}^{\rightarrow}|$: number of incrementally $(L+\epsilon)$-controllable states

▷ $\Gamma_{L+\epsilon}$: branching factor on $\mathcal{S}_{L+\epsilon}^{\rightarrow}$ (it is always $\leq S_{L+\epsilon}$, often times $= O(1)$)

▶ **`DisCo` is robust w.r.t. the total number of states** S

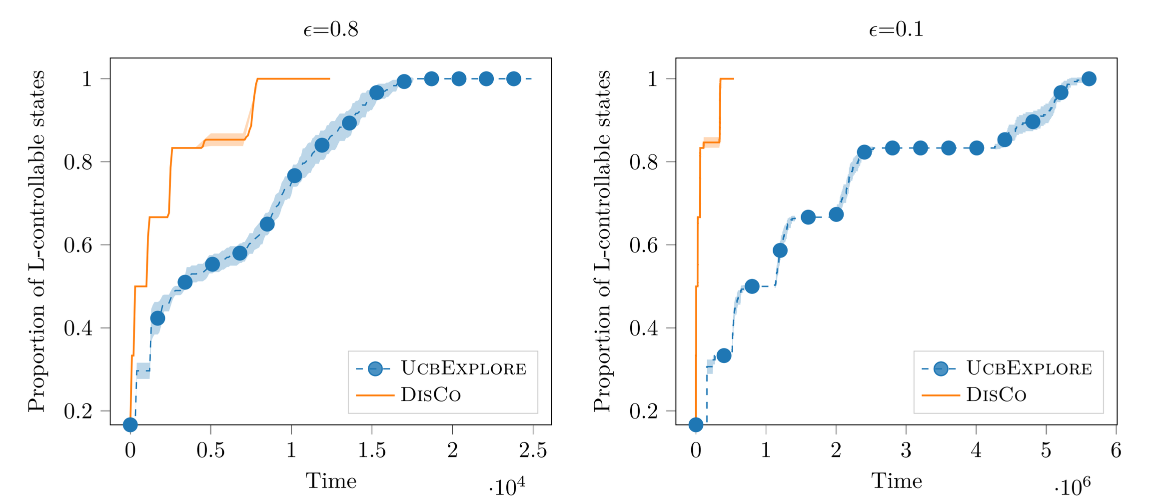
▷ Sample complexity: only in $\log(S)$

▷ Comput. complexity: indep. of S

Comparison with Prior Approach

	`DisCo` (this work)	`UcbExplore` (Lim & Auer, 2012)
Policies	ϵ-optimal	"accurate enough"
Rate	$\widetilde{O}(\epsilon^{-2})$	$\widetilde{O}(\epsilon^{-3})$

Numerical simulation: `DisCo` outperforms `UcbExplore`, especially as $\epsilon \downarrow$



Goal-Free Cost-Free Exploration on $\mathcal{S}_L^{\rightarrow}$ with `DisCo`

▶ Post-exploration, `DisCo` can compute an $(\epsilon/c_{\min})$-optimal policy for **any** goal-oriented problem restricted on $\mathcal{S}_L^{\rightarrow}$ with **any** cost function in $[c_{\min}, 1]$

▶ Goal-conditioned counterpart to the "reward-free" framework in *finite-horizon* [Jin et al., ICML 2020]