Understanding (Unsupervised) Exploration for Goal-Based RL

Alessandro LAZARIC (FAIR)

September 21, 2022 - EWRL - Milan, Italy

Meta

Understanding Unsupervised Exploration for Goal-Based RL

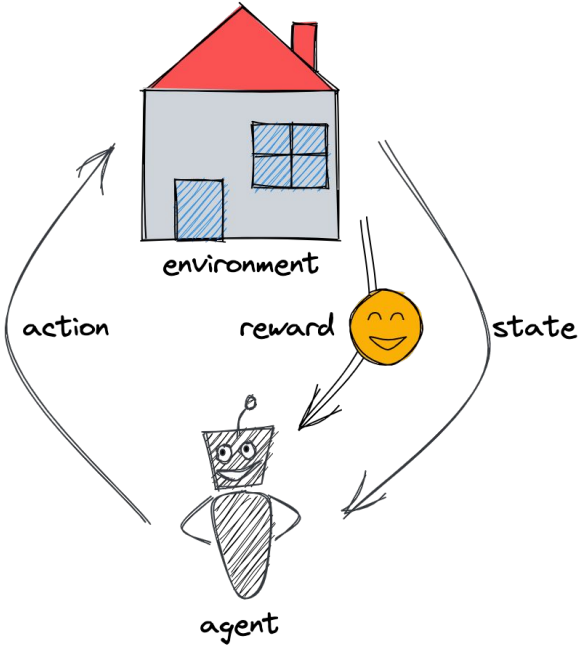
Alessandro LAZARIC (FAIR)

September 21, 2022 - EWRL - Milan, Italy

Meta

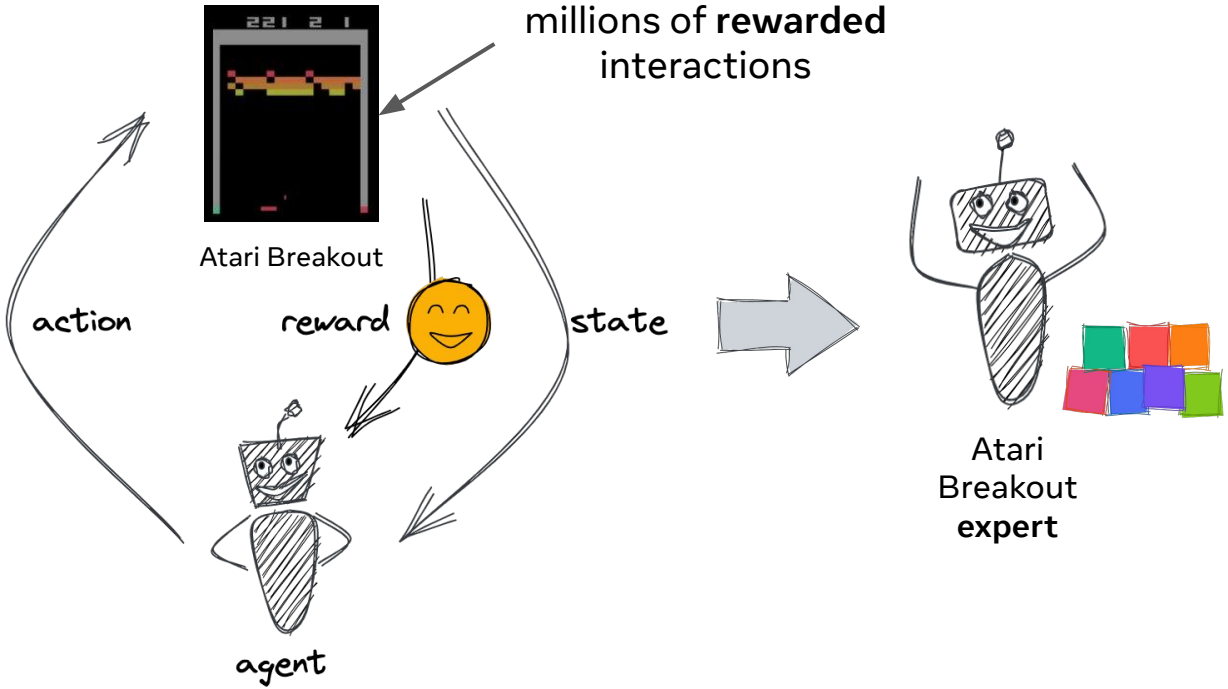
From Specialized to Universally Controllable Agents

Robotics, recommender systems, portfolio management, (computer) games, autonomous cars, ...

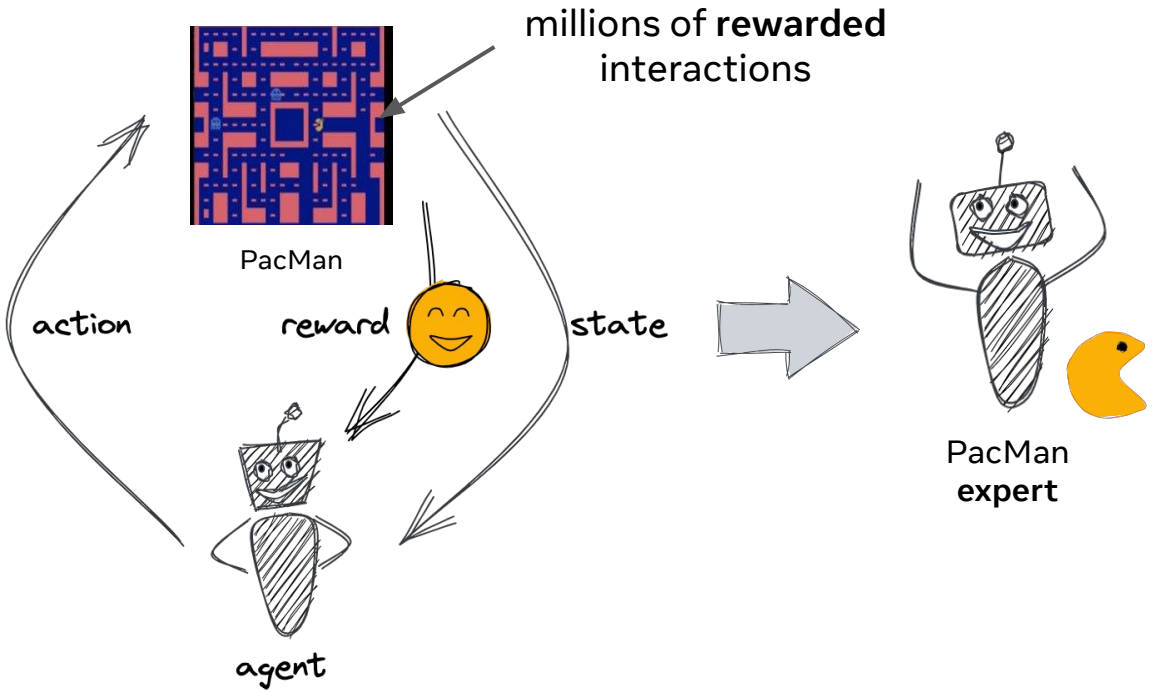


Data-driven sequential decision making under uncertainty = RL

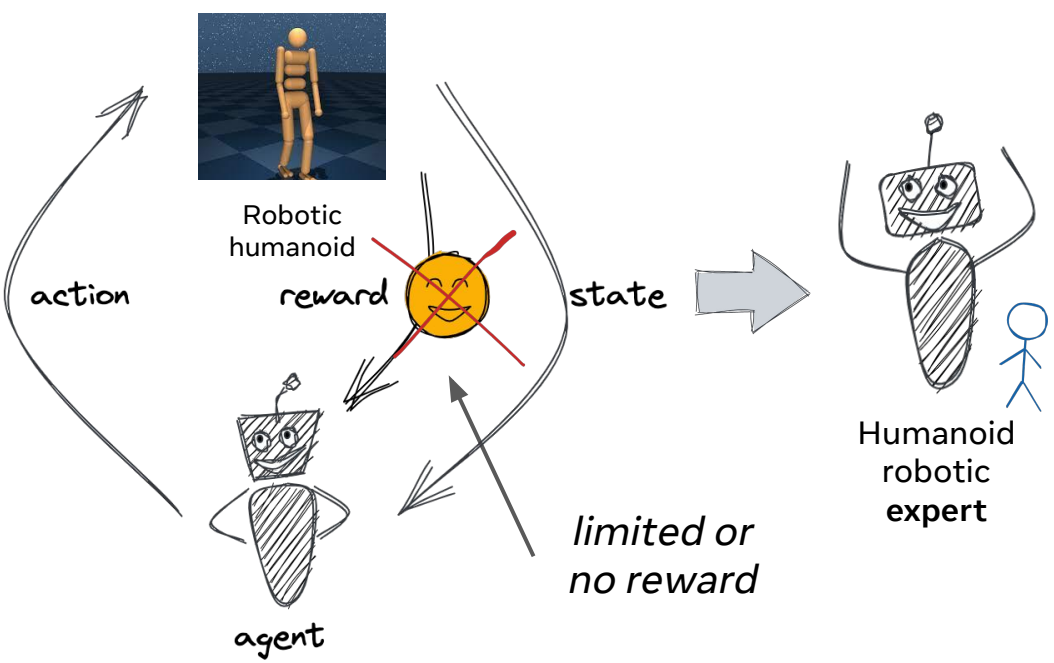
From **Specialized** to Universally Controllable Agents



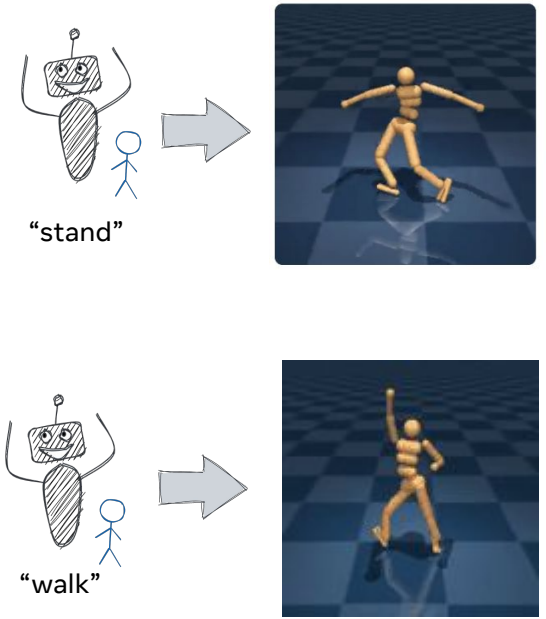
From **Specialized** to Universally Controllable Agents



From Specialized to **Universally Controllable** Agents

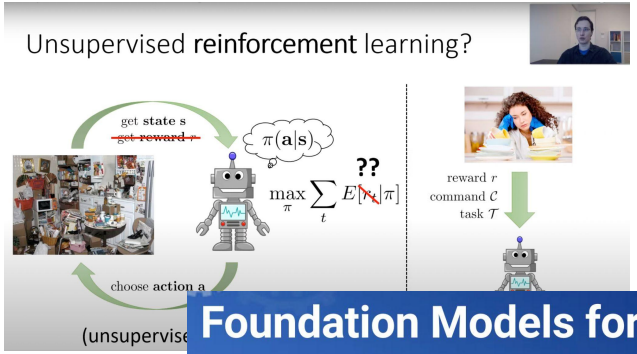


Unsupervised RL



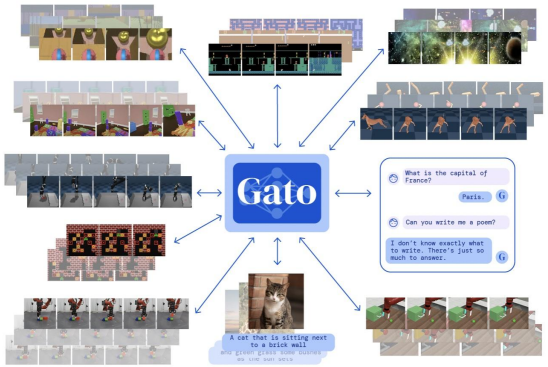
Zero/few-shot learning

From Specialized to **Universally Controllable** Agents



URLB: Unsupervised Reinforcement Learning Benchmark

Michael Laskin* UC Berkeley mlaskin@berkeley.edu
Denis Yarats* NYU, FAIR denisyarats@cs.nyu.edu
U
Albert Zhan UC Berkeley
Kevin Lu UC Berkeley
Catherine Cang UC Berkeley
Lerrel Pint NYU



Foundation Models for Decision Making

NeurIPS 2022 Workshop, New Orleans, USA (in Person)

December 3, 2022 (Saturday) 08:00 - 18:00 (ET)

Self-supervision for Reinforcement Learning (SSL-RL)

May 7, 2021 // ICLR Workshop

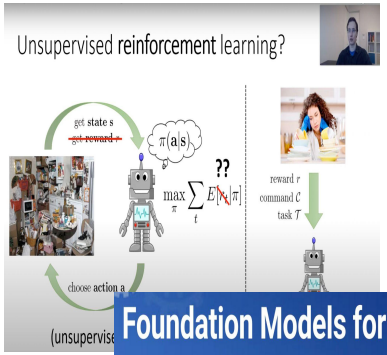
Workshop on

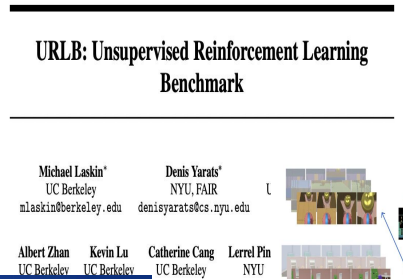
Pre-training Robot Learning

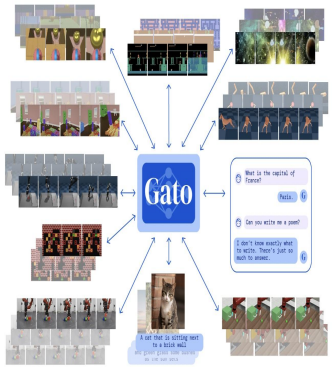
at the Conference on Robot Learning, 2022

Thursday, December 15th, 2022

From Specialized to **Universally Controllable** Agents







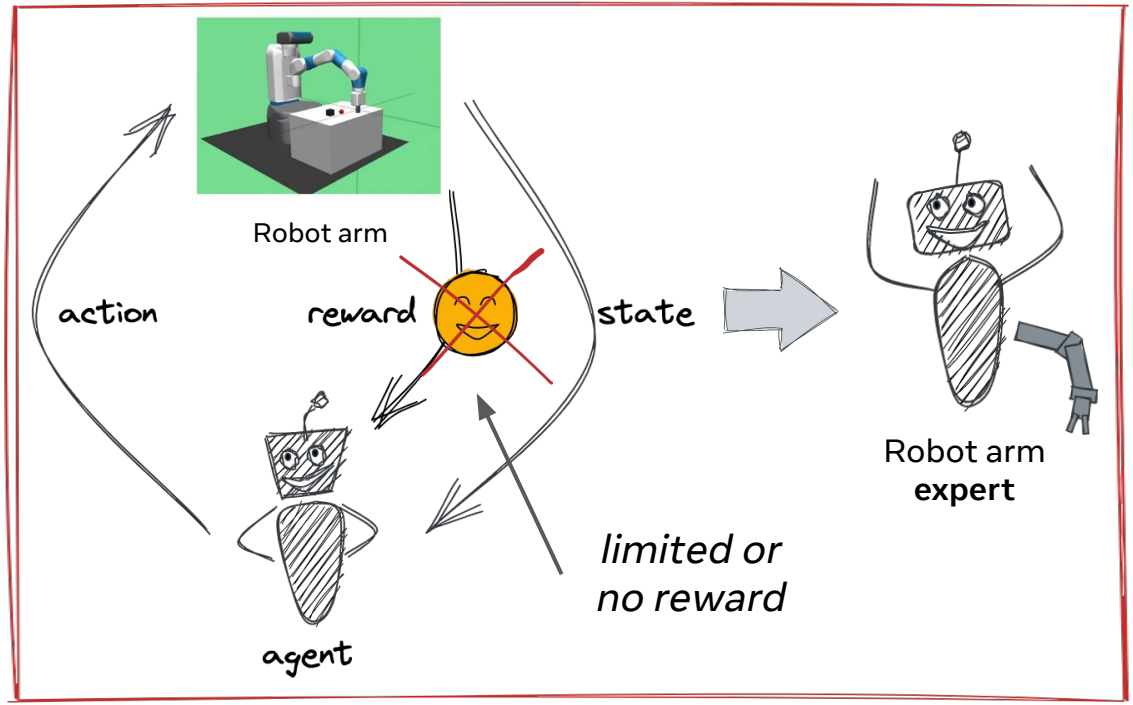
Connections to

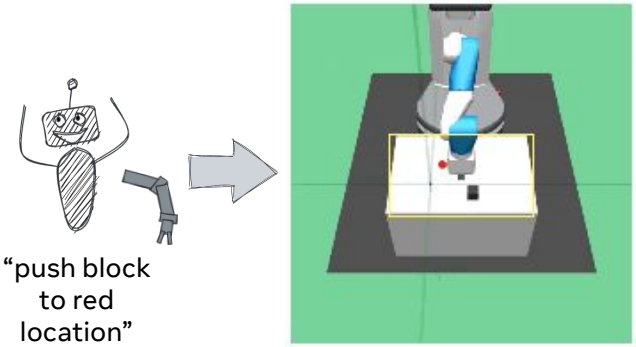
- Meta / Continual-RL
- Multi-task / Transfer-RL
- …

Self-supervision for Reinforcement Learning (SSL-RL)

May 7, 2021 // ICLR Workshop

This Talk: Unsupervised Exploration for Goal-Based RL

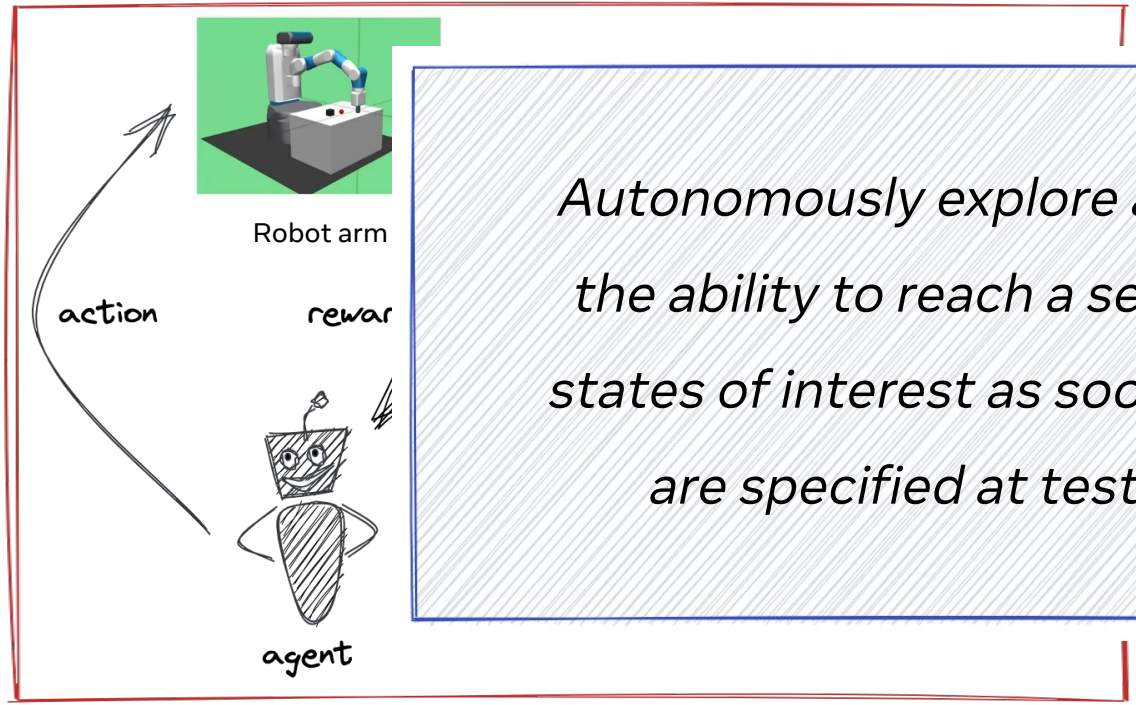




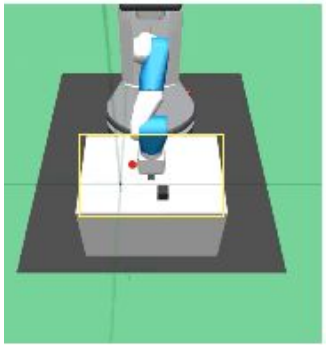
Unsupervised Exploration

Goal-Based RL

This Talk: Unsupervised Exploration for Goal-Based RL



Autonomously explore and learn the ability to reach a set of goal states of interest as soon as they are specified at test time



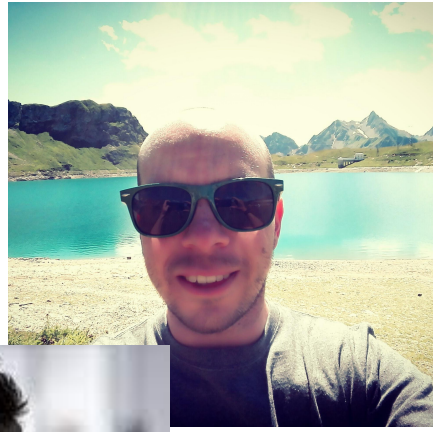
Unsupervised Exploration

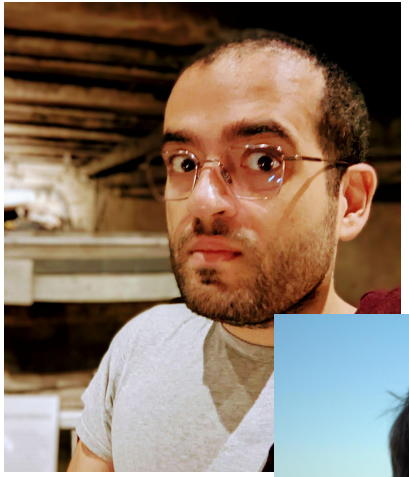
Goal-Based RL

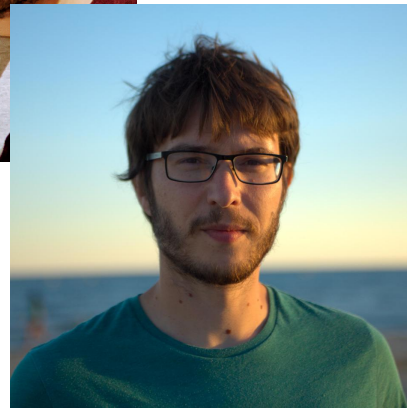
Outline

- Unsupervised Exploration for **Controllable** States
- Unsupervised Exploration for **Incrementally** Controllable States
- Discussion

Collaborators







Credit: Francis Vachon / Cedric Colas

Unsupervised Exploration for **Controllable States**

Meta

Unsup. Exploration: What is the **Question**?

From a **theory** point of view **[not comprehensive!]**

- Active exploration for MDP estimation [Tarbouriech, **Lazaric**; 2019 / Tarbouriech, Ghavamzadeh, **Lazaric**; 2020]
- “Simulated” generative model [Tarbouriech, Pirotta, Valko, **Lazaric**; 2021]
- Maximum entropy [Hazan et al.; 2019 / Mutti et al., 2022]
- Reward-free exploration [Jin et al.; 2020 / …]

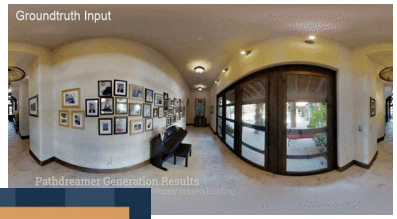
Unsup. Exploration: What is the **Question**?

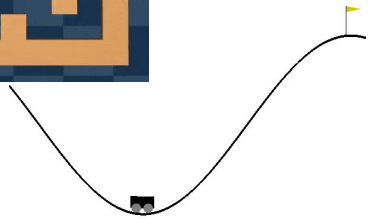
From an **algorithmic** point of view **[not comprehensive!]**

- Intrinsically motivated RL [Schmidhuber, 1991 / Bellmare et al., 2016 / Deepak et al., 2017 / …]
- Goal generation [Colas et al., 2017 / Held et al., 2017 / Péré et al., 2018 / Laversanne-Finot et al., 2018 / Pong et al., 2020 / Zhang et al., 2020 / Ecoffet et al., 2021 / Mezghani et al., 2022 / …]
- Maximum entropy [Silviu et al., 2020 / Mutti et al.,2021 / …]

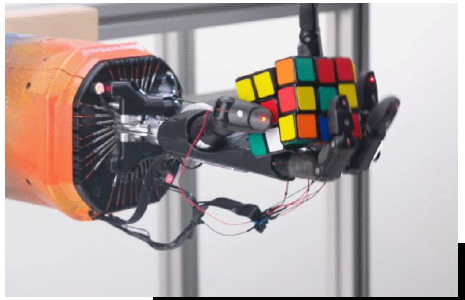
Goal-Based Reinforcement Learning

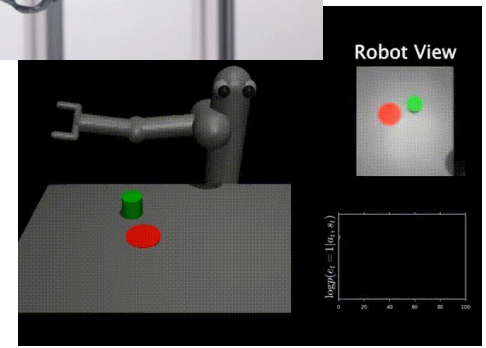
Navigation



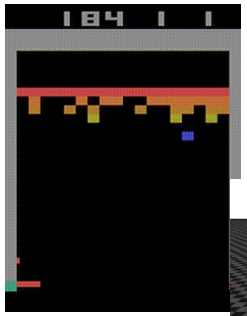


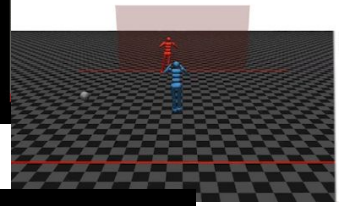
Robotics

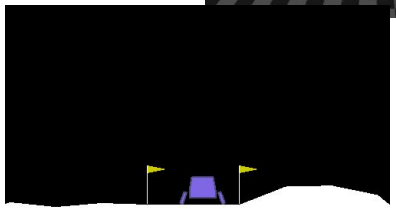




Games



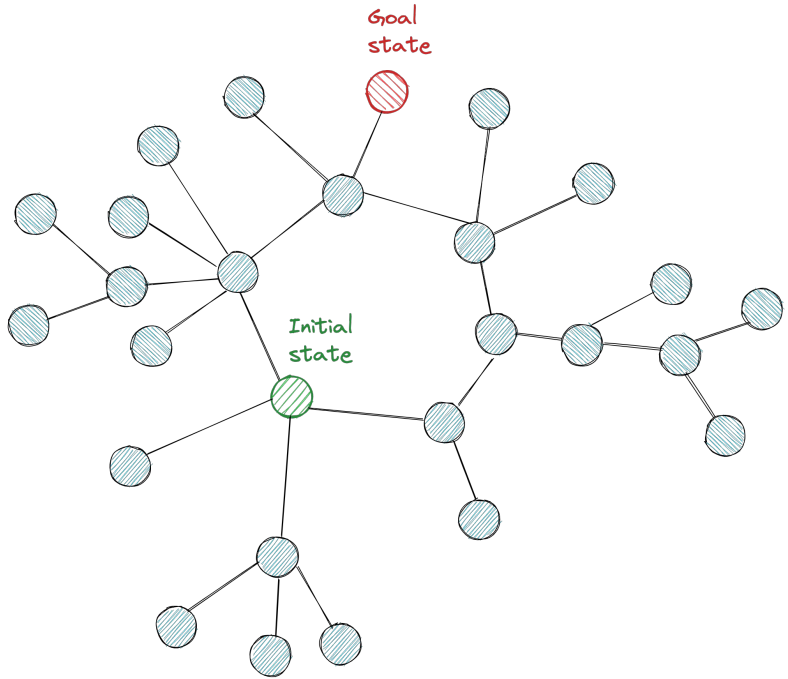




Formalizing Goal-Based RL [see Matteo's tutorial]

Goal-Based MDP (specific instance of SSP)

- State space $\mathcal{S}$
- Initial state s_0
- Goal state g
- Action space $\mathcal{A}$
- Transition model $p(s'|s,a)$
- Cost function $c(s,a) = 1 \quad c(g) = 0$

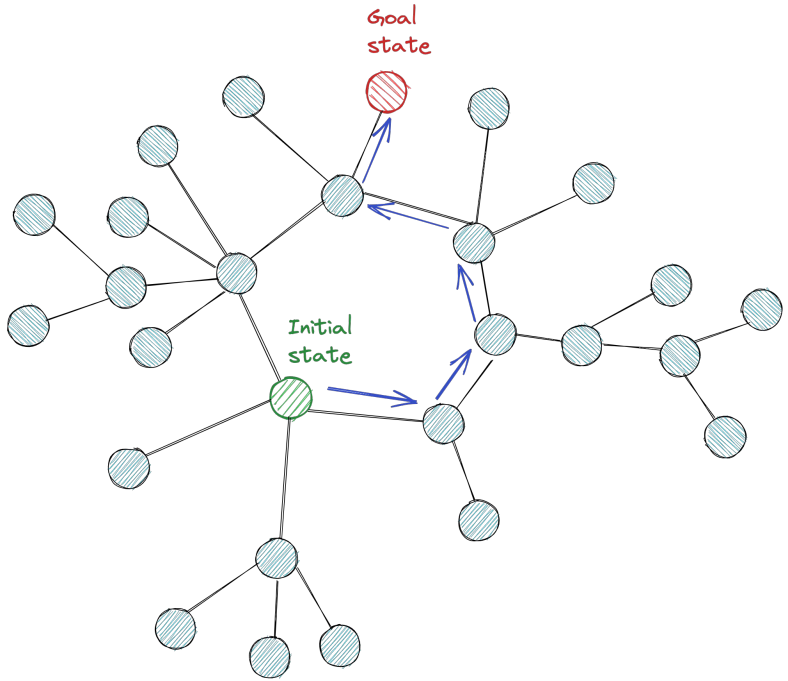


Formalizing Goal-Based RL

Goal-Based MDP (specific instance of SSP)

- Policy $\pi : \mathcal{S} \to \mathcal{A}$
- Hitting time $\tau_\pi(s \to s')$
- Value function = expected hitting time

$$V^\pi(s \to s') = \mathbb{E}\left[\tau_\pi(s \to s')\right]$$



Exploration for Goal-Based RL (see Matteo's tutorial)

Thm: Sample Complexity [Chen et al., 2022 (similar results in Tarbouriech et al., 2020)]

There exists an algorithm that returns an ϵ-optimal policy with a sample complexity

$$\widetilde{O}\left(\frac{T_\star^3 S A}{\epsilon^2}\right)$$

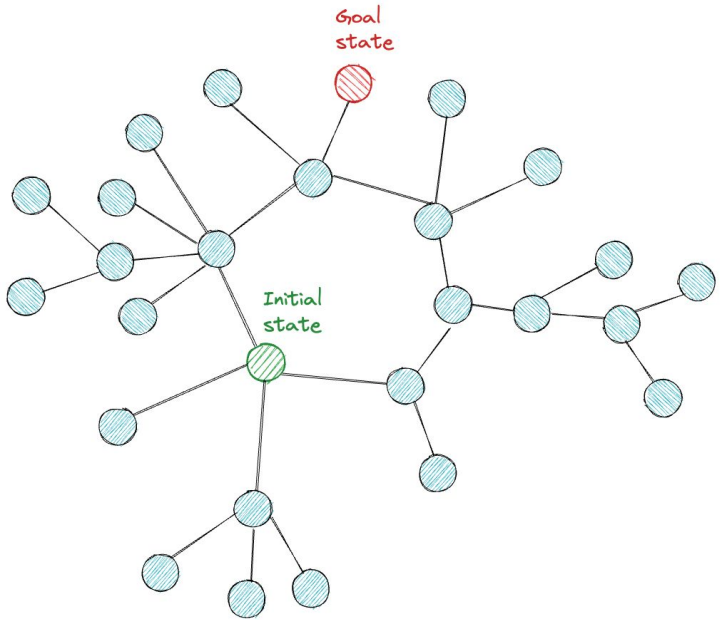
Remarks

- Similar to finite-horizon and discounted bounds
- "Binary" cost function
- $c_{\min} = 1$
- $B_\star = T_\star$

From Single-Goal to **Multi-Goal**

Multi-Goal MDP

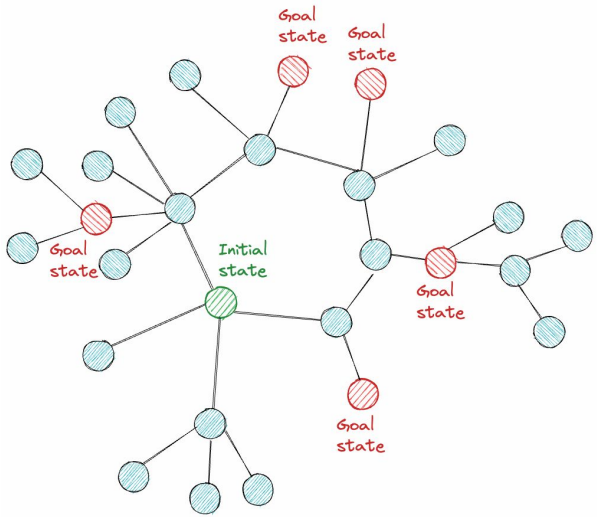
- Set of Goals $\mathcal{G} \subseteq \mathcal{S}$
- Goal-Based Policy $\pi : \mathcal{S} \times \mathcal{G} \rightarrow \mathcal{A}$



A General Principle for Multi-Goal Exploration

SYOG: **S**et **Y**our **O**wn **G**oals

1. Select a **relevant goal** g_k
2. Execute an **exploratory** version of $\pi(\cdot|s, g_k)$
3. **Improve** $\pi(\cdot|s, g_k)$ with the collected experience
4. *If* $\pi(\cdot|s, g_k)$ is **good** *then* stop
   otherwise jump to 1.

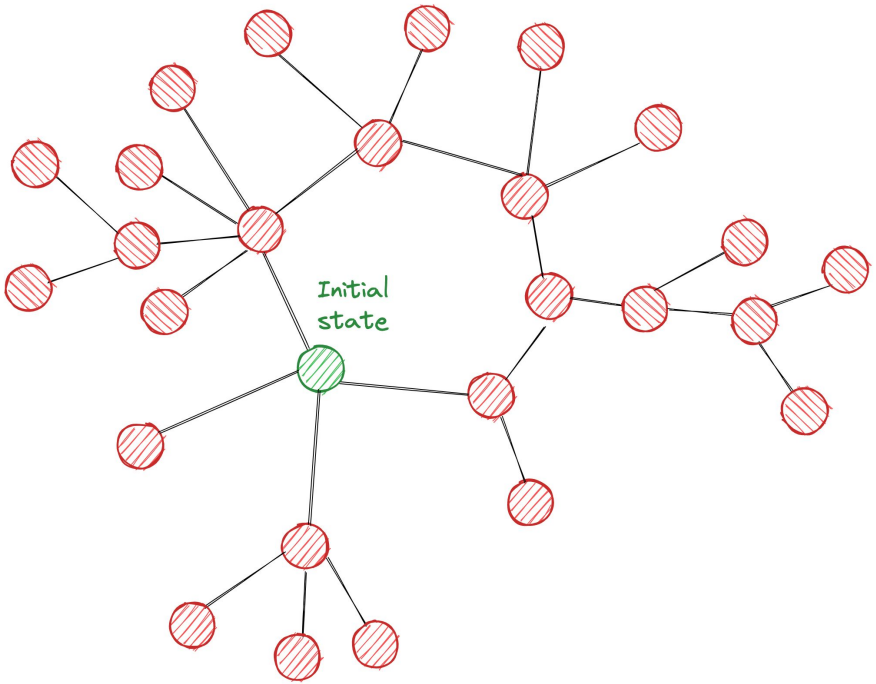


Similar to many schemes defined in literature but rarely provide a well-formalized objective and guarantees

What are “Relevant” Unsupervised Goals?

All possible states $\mathcal{G}_{\text{test}} \equiv \mathcal{G} \equiv \mathcal{S}$

- Prior knowledge of the “valid” states
- Possibly very difficult goals



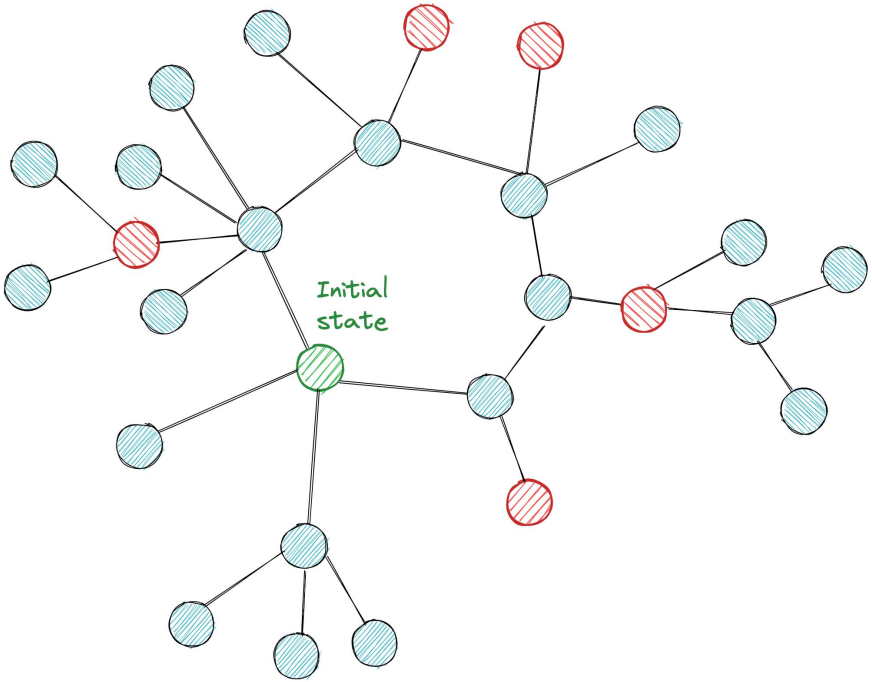
What are “Relevant” Unsupervised Goals?

All possible states $\mathcal{G}_{\text{test}} \equiv \mathcal{G} \equiv \mathcal{S}$

- Prior knowledge of the “valid” states
- Possibly very difficult goals

Predefined set of states $\mathcal{G}_{\text{test}} \equiv \mathcal{G} \subset \mathcal{S}$

- Prior knowledge
- No generalization to unknown states at downstream time



What are “Relevant” Unsupervised Goals?

All possible states $\mathcal{G}_{\text{test}} \equiv \mathcal{G} \equiv \mathcal{S}$

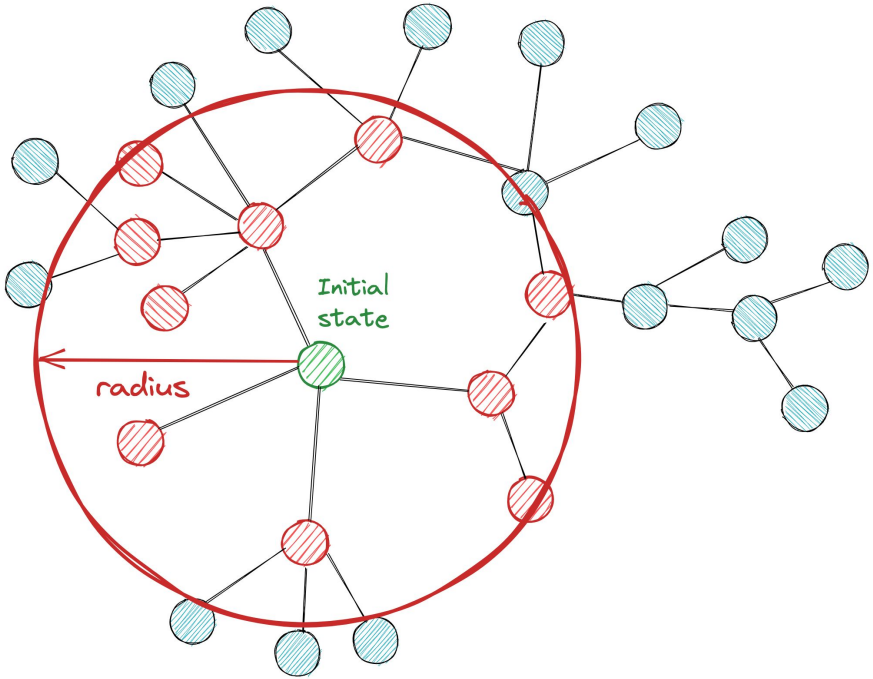
- Prior knowledge of the “valid” states
- Possibly very difficult goals

Predefined set of states $\mathcal{G}_{\text{test}} \equiv \mathcal{G} \subset \mathcal{S}$

- Prior knowledge
- No generalization to unknown states at downstream time

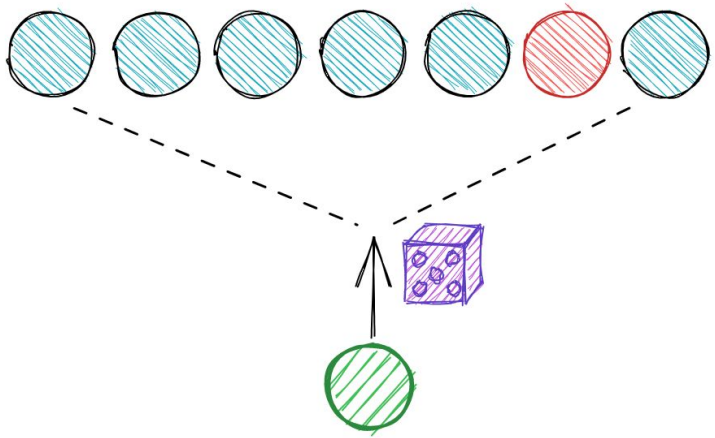
Radius of “competence” $\mathcal{G}_{\text{test}} \neq \mathcal{G} \subseteq \mathcal{S}$

- No prior knowledge
- More natural to “express”
- Enable curriculum learning
- *Unknown to the agent*



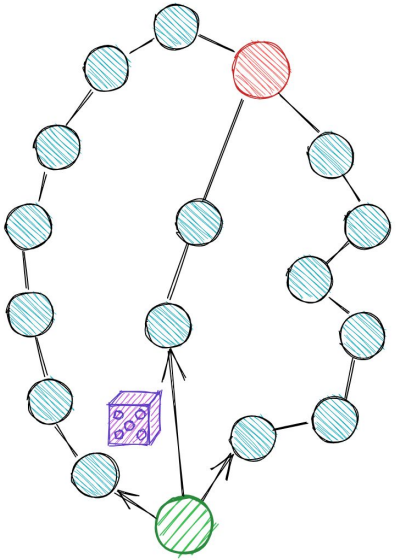
Controllable States

Reachable State



$$\mathbb{P}\left[\tau_\pi(s_0 \to s) < \infty\right] > 0$$

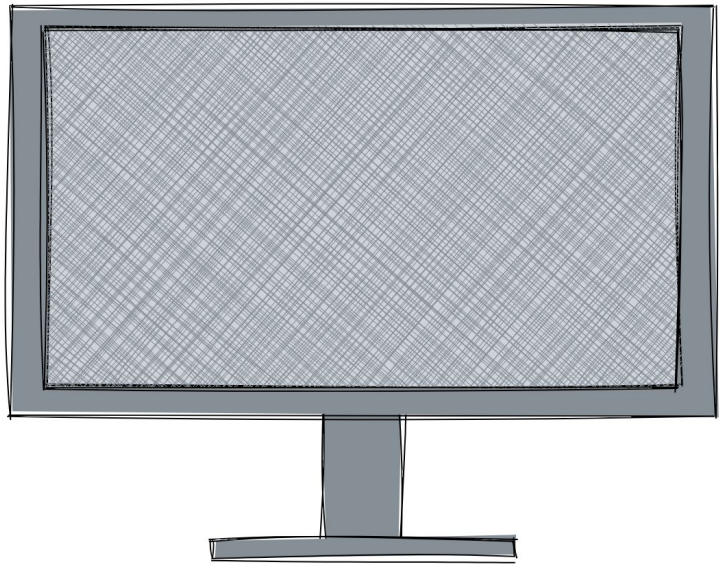
Controllable State



$$\mathbb{E}\left[\tau_\pi(s_0 \to s)\right] < \infty$$

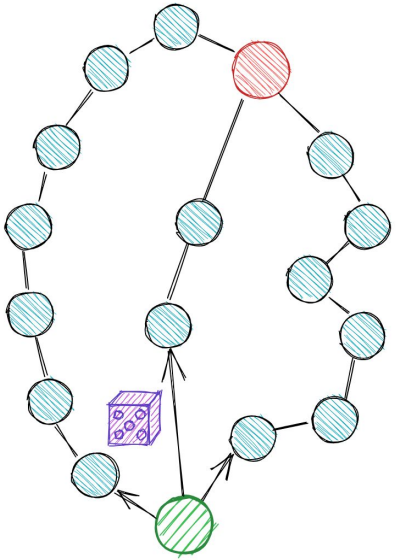
Controllable States

Noisy TV



$$\mathbb{P}\left[\tau_\pi(s_0 \to s) < \infty\right] > 0$$

Controllable State



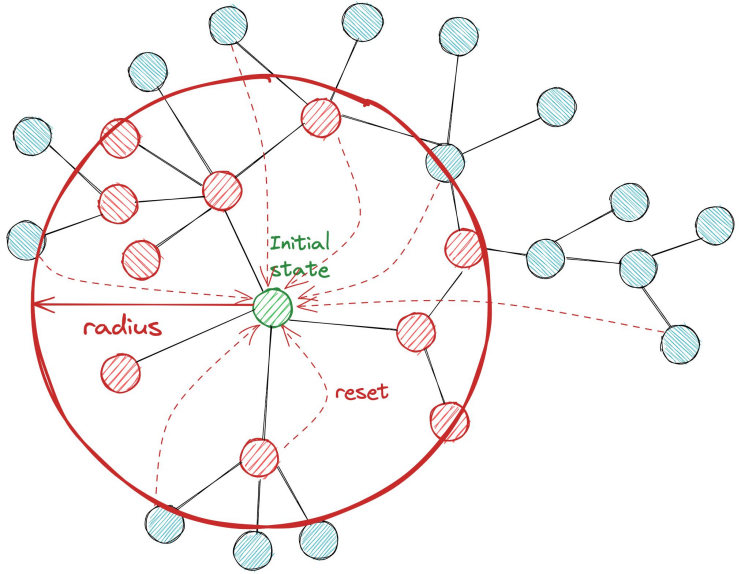
$$\mathbb{E}\left[\tau_\pi(s_0 \to s)\right] < \infty$$

Unsupervised **M**ulti-**G**oal **E**xploration

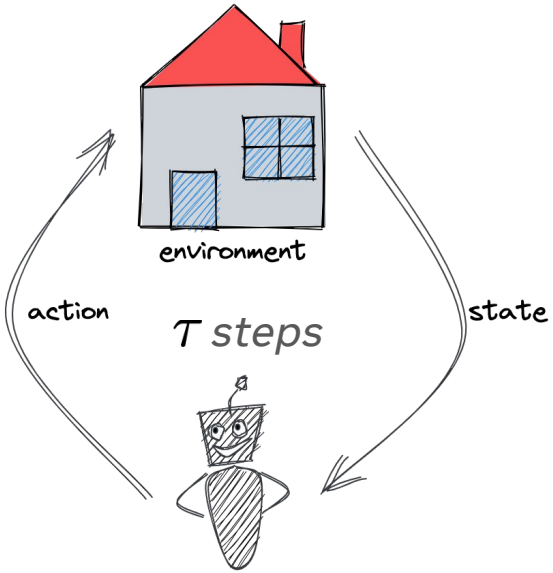
*Definition of **MGE***

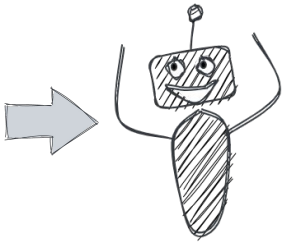
- Reset action a_{reset} s.t. $p(s_0|s, a_{\text{reset}}) = 1$
- Goal radius L
- Accuracy level ϵ
- Goal set

$$\mathcal{G}_L = \left\{ g \in \mathcal{S} : \min_{\pi} \mathbb{E}_{\pi} \left[\tau_{\pi}(s_0 \to g) \right] = V^{\star}(s_0 \to g) \leq L \right\}$$

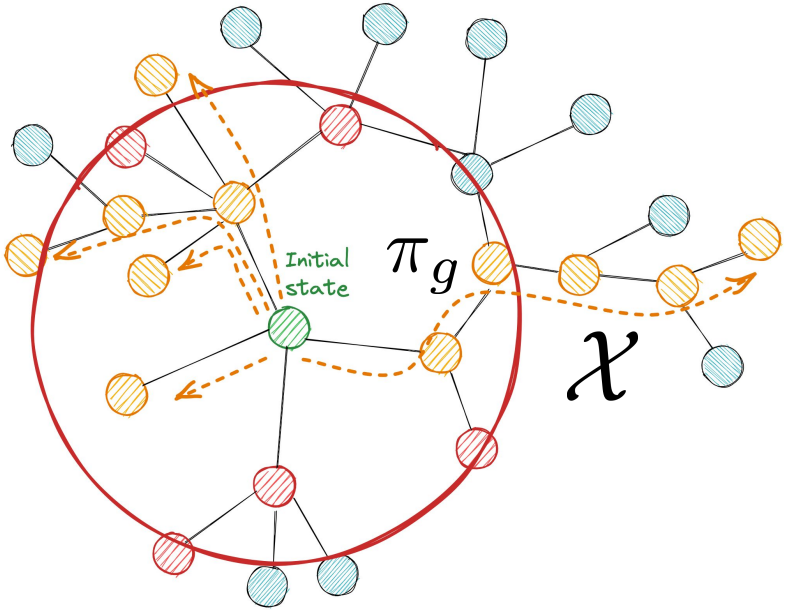


Unsupervised **M**ulti-**G**oal **E**xploration





Goal-based policy *that the agent is confident to accurately* ***control a set of goal states***



Unsupervised **M**ulti-**G**oal **E**xploration

(ϵ, δ, L)-*PAC Learner*

Agent stops after $\tau = \text{poly}(S, A, L, \log(1/\delta), 1/\epsilon)$ steps and

$$\mathbb{P}\left[\begin{array}{c} \textit{Accurate goal set identification} \\ \mathcal{G}_L \subseteq \mathcal{X} \subseteq \mathcal{G}_{L+\epsilon} \\ \\ \textit{Near-optimal goal-based policy} \\ V^{\pi_g}(s_0 \to g) \leq V^{\star}(s_0 \to g) + \epsilon \quad \forall g \in \mathcal{X} \end{array}\right] \geq 1 - \delta$$

Unsupervised **M**ulti-**G**oal **E**xploration

(ϵ, δ, L)-*PAC Learner*

Agent stops after $\tau = \text{poly}(S, A, L, \log(1/\delta), 1/\epsilon)$ steps and

$$\mathbb{P}\left[\begin{array}{c} \textit{Accurate goal set identification} \\ \mathcal{G}_L \subseteq \mathcal{X} \subseteq \mathcal{G}_{L+\epsilon} \\ \\ \textit{Near-optimal goal-based policy} \\ V^{\pi_g}(s_0 \to g) \leq V^{\star}(s_0 \to g) + \epsilon \quad \forall g \in \mathcal{X} \end{array}\right] \geq 1 - \delta$$

Unsupervised **M**ulti-**G**oal **E**xploration

(ϵ, δ, L)-*PAC Learner*

Agent stops after $\tau = \text{poly}(S, A, L, \log(1/\delta), 1/\epsilon)$ steps and

$$\mathbb{P}\left[\begin{array}{c} \textit{Accurate goal set identification} \\ \mathcal{G}_L \subseteq \mathcal{X} \subseteq \mathcal{G}_{L+\epsilon} \\ \\ \textit{Near-optimal goal-based policy} \\ V^{\pi_g}(s_0 \to g) \leq V^{\star}(s_0 \to g) + \epsilon \quad \forall g \in \mathcal{X} \end{array}\right] \geq 1 - \delta$$

Unsupervised **M**ulti-**G**oal **E**xploration

(ϵ, δ, L)-*PAC Learner*

Agent stops after $\tau = \text{poly}(S, A, L, \log(1/\delta), 1/\epsilon)$ steps and

$$\mathbb{P}\left[\begin{array}{c} \textit{Accurate goal set identification} \\ \mathcal{G}_L \subseteq \mathcal{X} \subseteq \mathcal{G}_{L+\epsilon} \\ \\ \textit{Near-optimal goal-based policy} \\ V^{\pi_g}(s_0 \to g) \leq V^{\star}(s_0 \to g) + \epsilon \quad \forall g \in \mathcal{X} \end{array}\right] \geq 1 - \delta$$

Unsupervised Multi-Goal Exploration

Thm: Lower Bound [Tarbouriech et al., 2022]

For any (ϵ, δ, L)-PAC learner, there exists an MDP such that

$$\mathbb{E}[\tau] = \Omega\left(\frac{L^3 S A}{\epsilon^2}\right)$$

Remarks

- Horizon is “known”
- Goal states are unknown
- Dependencies match finite-horizon/discounted

Adaptive Goal Selection Scheme - AdaGoal

SYOG: **S**et **Y**our **O**wn **G**oals → AdaGoal

1. Select a **relevant goal** g_k
2. Execute an **exploratory** version of $\pi(\cdot|s, g_k)$
3. **Improve** $\pi(\cdot|s, g_k)$ with the collected experience
4. *If* $\pi(\cdot|s, g_k)$ is **good** *then* STOP and return
   otherwise jump to 1.

J. Tarbouriech, O. Darwiche Domingues, P. Ménard, M. Pirotta, M. Valko, **A. Lazaric**
"Adaptive Multi-Goal Exploration", AI&Stats-2022.

AdaGoal - Two Main Ingredients

Optimistic controllability

$$\mathcal{D}_k(g) \leq V^\star(s_0 \to g)$$

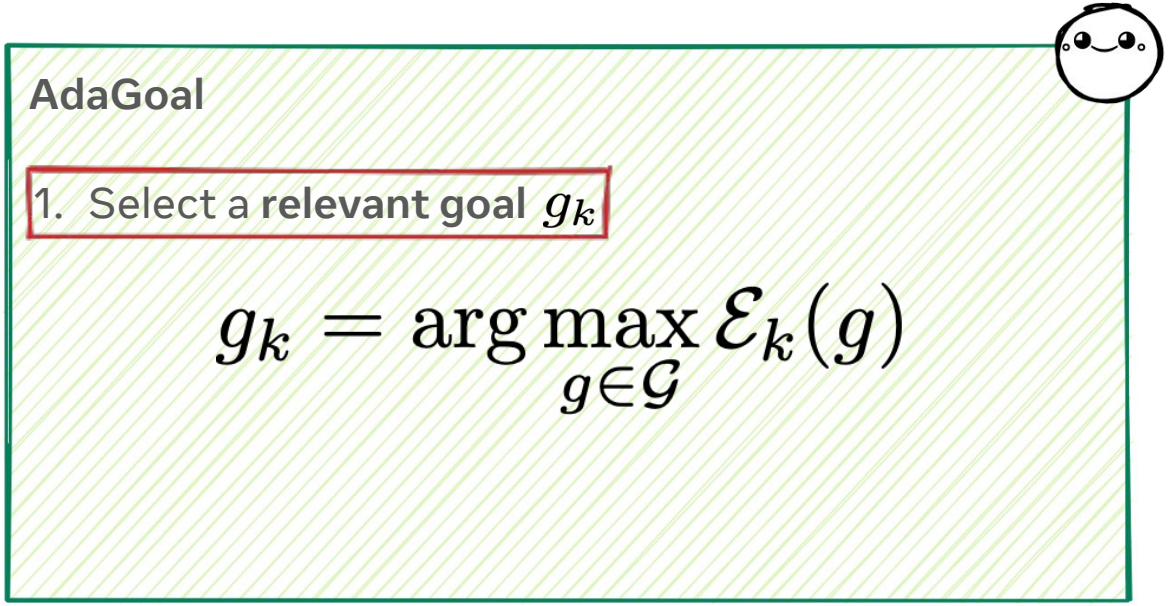
true optimal value

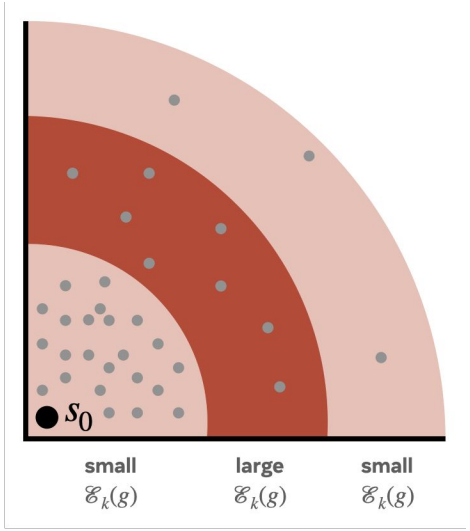
Uncertainty (or *regret* or *performance loss*)

$$V^{\pi_k}(s_0 \to g) - \mathcal{D}_k(g) \leq \mathcal{E}_k(g)$$

true value of current policy

Adaptive **Goal** Selection Scheme



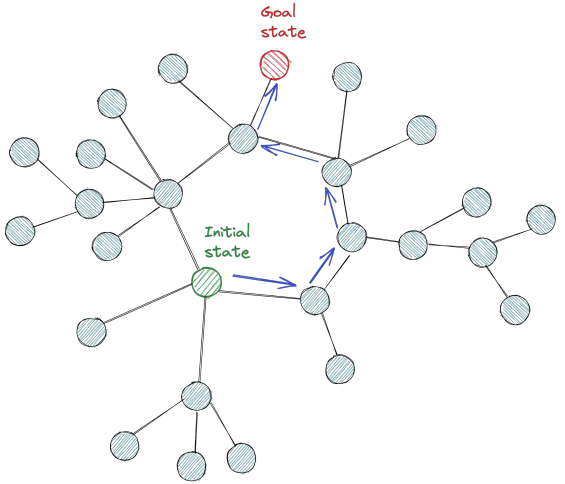


Difficult to control ≠ uncertain

Adaptive Goal Selection Scheme

AdaGoal

1. Select a **relevant goal** g_k
2. Execute an **exploratory** version of $\pi(\cdot|s, g_k)$
3. **Improve** $\pi(\cdot|s, g_k)$ with the collected experience



Any “good”
SSP exploration algorithm

Adaptive Goal Selection Scheme

AdaGoal

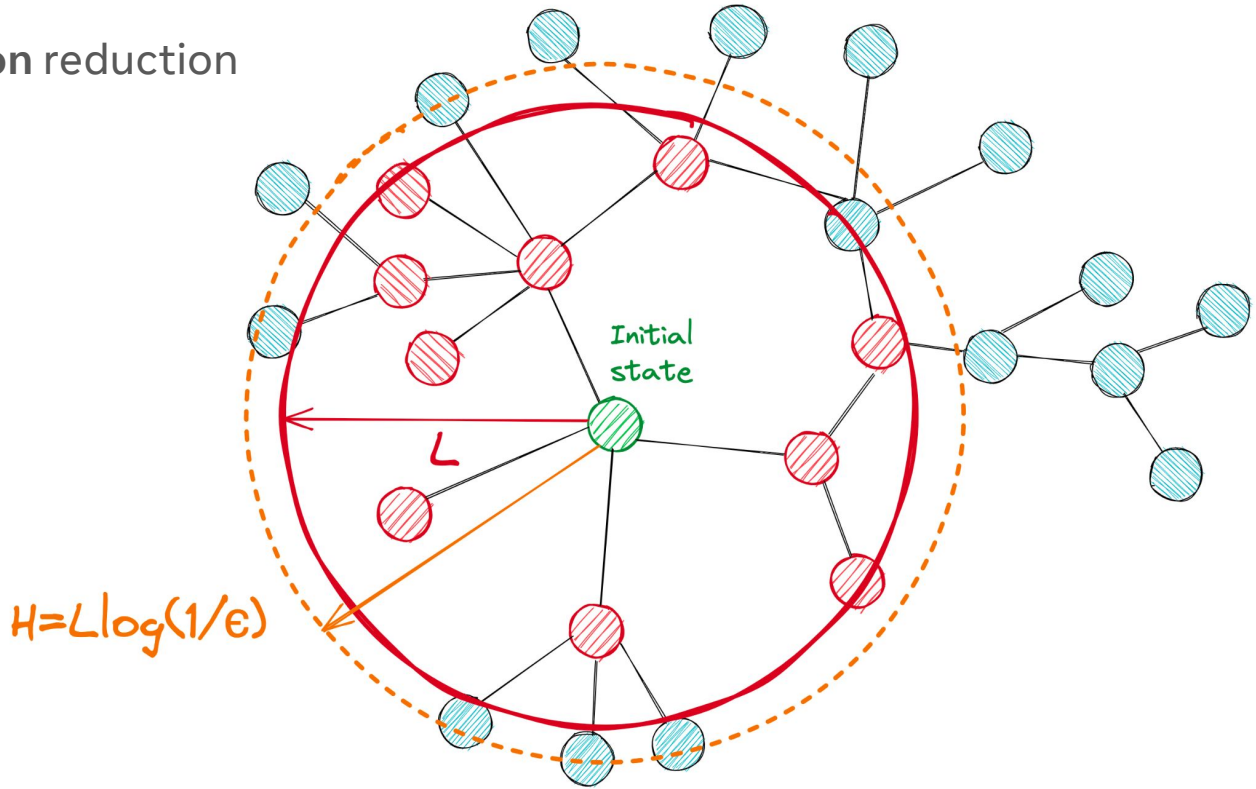
1. Select a **relevant goal** g_k
2. Execute an **exploratory** version of $\pi(\cdot|s, g_k)$
3. **Improve** $\pi(\cdot|s, g_k)$ with the collected experience
4. *If* $\pi(\cdot|s, g_k)$ is **good** *then* stop
   otherwise jump to 1.

$$\max_{g:\mathcal{D}_k(g)\leq L} \mathcal{E}_k(g) \leq \epsilon$$

$$\mathcal{X} = \{g \in \mathcal{G} : \mathcal{D}_k(g) \leq L\}$$

Tabular-AdaGoal

Finite-horizon reduction



Tabular-AdaGoal

Model-based **upper-confidence** estimate

$$\overline{Q}_h(s,a;g) = \mathrm{clip}\Big(\mathbb{1}(s \neq g) + \widehat{p}(\cdot|s,a)\overline{V}_{h+1}(\cdot;g) - \; ; 0, H\Big)$$

$\overline{V}_{h+1} - \underline{V}_{h+1}$

$\mathrm{Var}_{\widehat{p}(\cdot|s,a)}\big(\overline{V}_{h+1}\big)$

refined empirical Bernstein bound

optimism

clipping

$$\mathcal{D}_k(g) = \min_a \overline{Q}_1(s_0, a)$$

Adapted from *P. Ménard et al. "Fast active learning for pure exploration in reinforcement learning", ICML-2021.* (see also [Azar et al., 2017], [Zanette and Brunskill, 2019]).

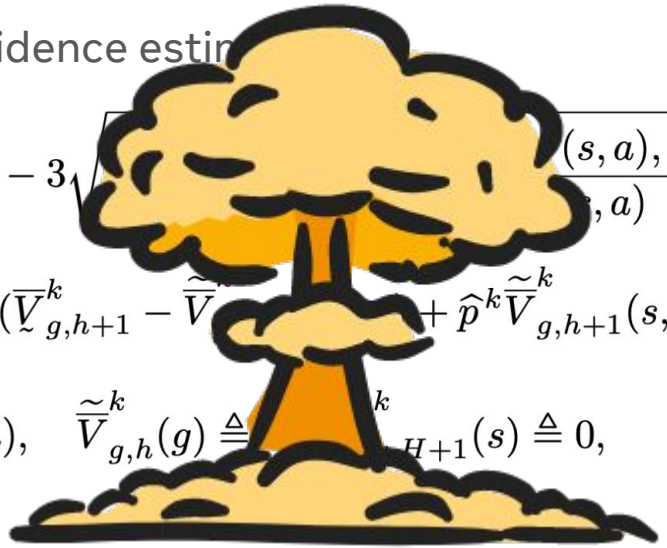
Tabular-AdaGoal

Model-based upper-confidence estim

$$\widetilde{\overline{Q}}^k_{g,h}(s,a) \triangleq \mathrm{clip}\Bigg(\mathbb{1}[s \neq g] - 3\sqrt{\qquad\qquad (s,a),\delta)} - 14H^2\frac{\beta(n^k(s,a),\delta)}{n^k(s,a)}$$

$$- \frac{1}{H}\widehat{p}^k(\overline{V}^k_{g,h+1} - \widetilde{\overline{V}}^k \qquad + \widehat{p}^k\widetilde{\overline{V}}^k_{g,h+1}(s,a),\ 0,\ H\Bigg),$$

$$\widetilde{\overline{V}}^k_{g,h}(s) \triangleq \min_{a\in\mathcal{A}} \widetilde{\overline{Q}}^k_{g,h}(s,a), \quad \widetilde{\overline{V}}^k_{g,h}(g) \triangleq \qquad k \qquad H+1(s) \triangleq 0,$$



Adaptation of

Tabular-AdaGoal

Cumulative error estimates

$$\overline{U}_h(s,a;g) = \mathrm{clip}\left(\left(1+\frac{3}{H}\right)\sum_{s'}\widehat{p}(s'|s,a)\sum_{a'}\pi_{h+1}(a'|s';g)\overline{U}_{h+1}(s',a';g) + \text{empirical Bernstein bound}\ ; H\right)$$

$$\mathcal{E}_k(g) = \sum_a \pi_k(a|s_0;g)\overline{U}_1(s_0,a;g)$$

$\mathrm{Var}_{\widehat{p}(\cdot|s,a)}\left(\overline{V}_{h+1}\right)$

empirical Bernstein bound

Propagation of error estimates

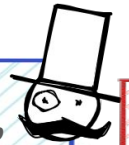
Adapted from *P. Ménard et al. "Fast active learning for pure exploration in reinforcement learning", ICML-2021.* (see also [Azar et al., 2017], [Zanette and Brunskill, 2019]).

Tabular-AdaGoal: **Sample Complexity** Bounds

Thm: Sample Complexity [Tarbouriech et al., 2022]

AdaGoal is (ϵ, δ, L)-PAC and

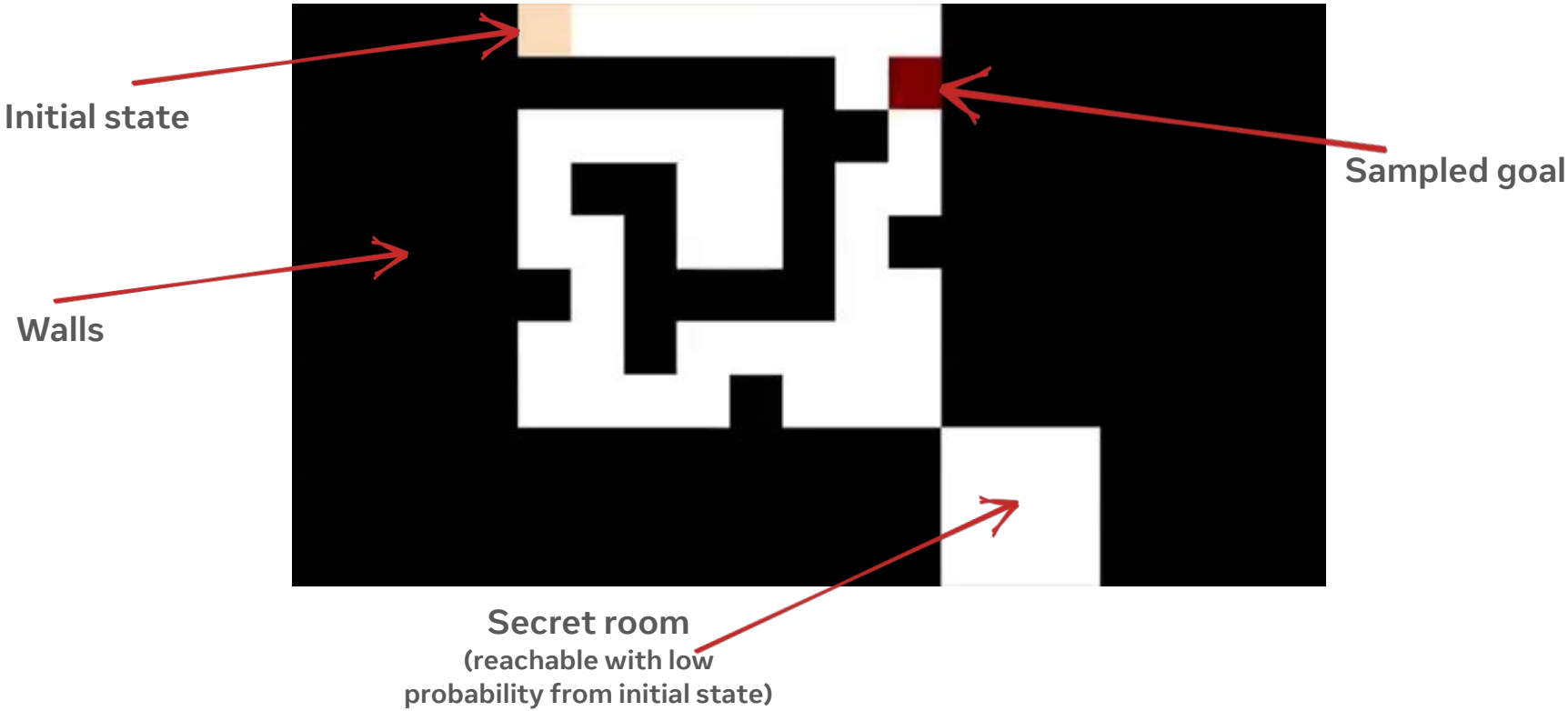
$$\mathbb{E}[\tau] = \widetilde{O}\left(\frac{L^3 S A}{\epsilon^2}\right)$$



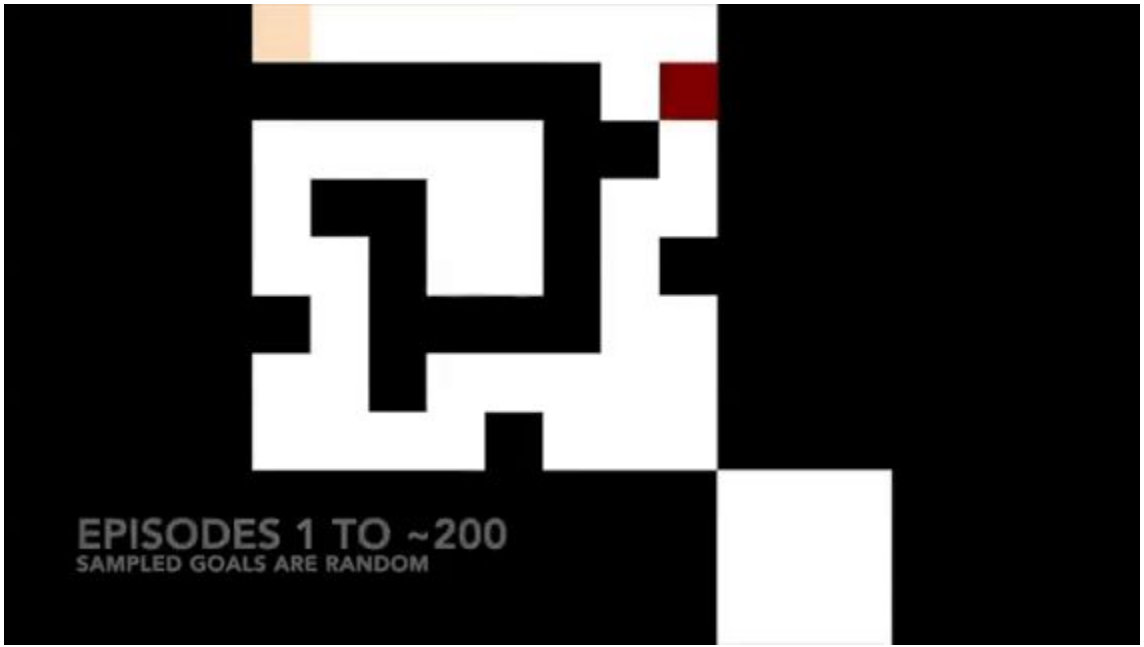
Remarks

- Stopping when confident to return accurate goal set and goal-based policy
- Minimax optimal
- Generalizable to linear MDPs

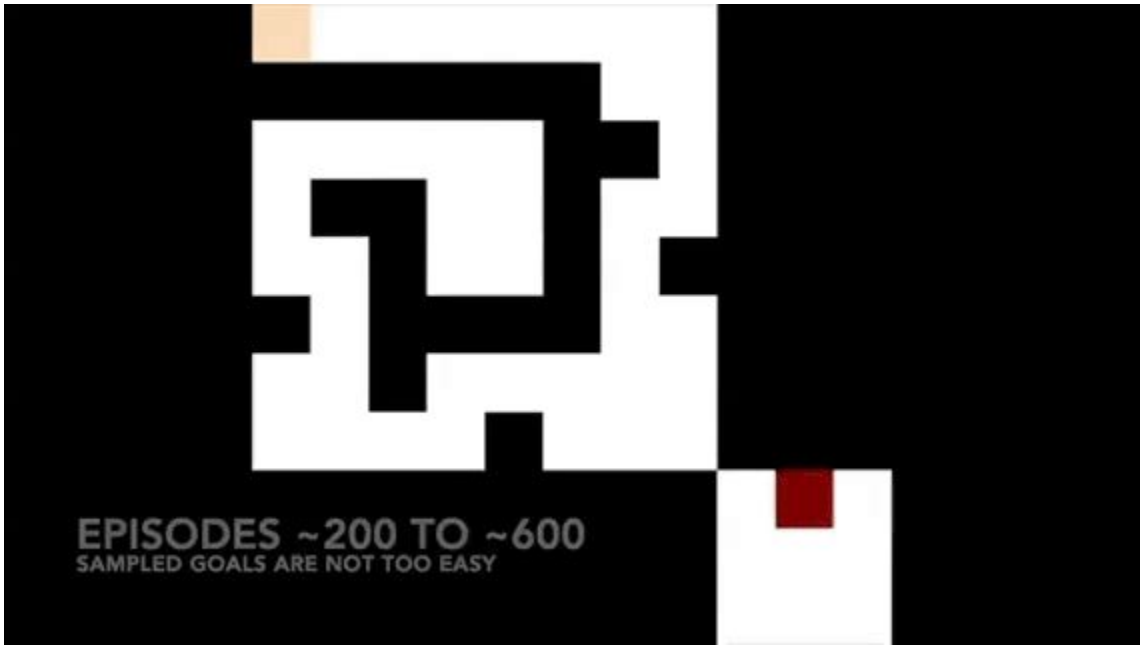
Tabular-AdaGoal: A Simple Example



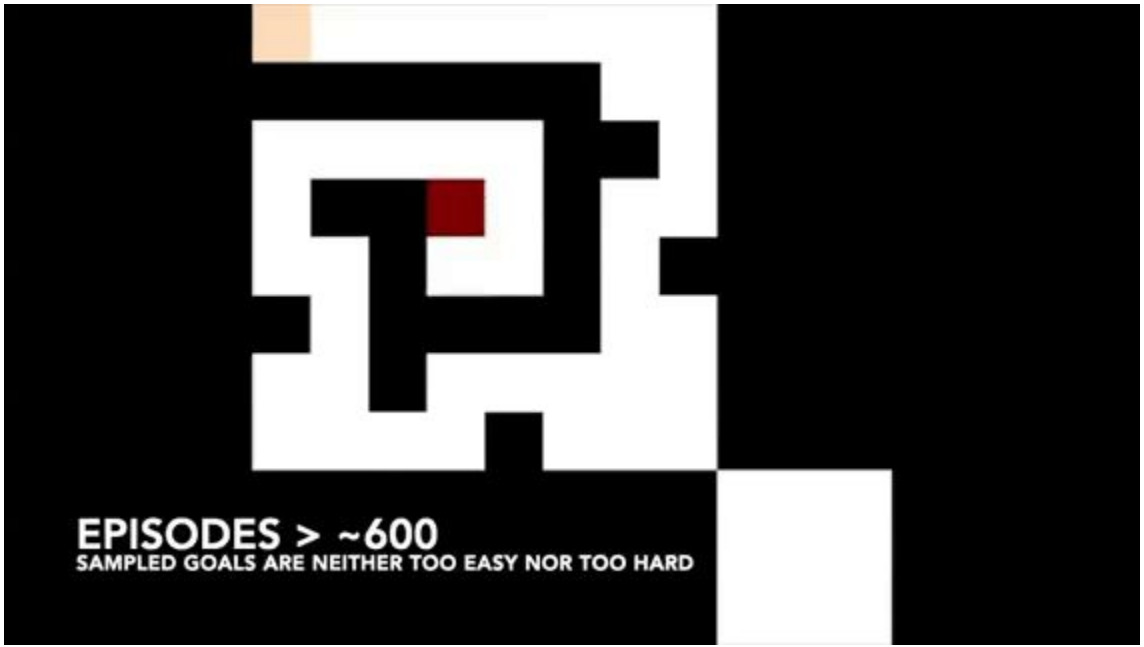
Tabular-AdaGoal: A Simple Example



Tabular-AdaGoal: A Simple **Example**

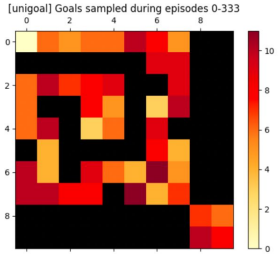


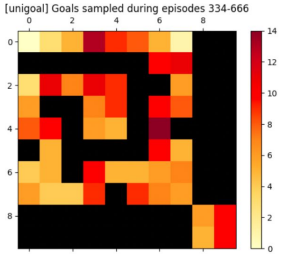
Tabular-AdaGoal: A Simple **Example**

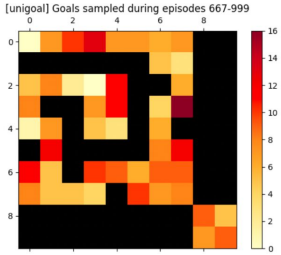


Tabular-AdaGoal: A Simple Example

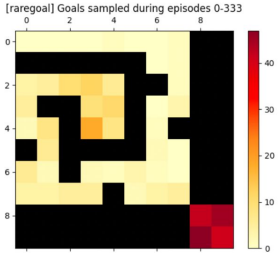
Uniform

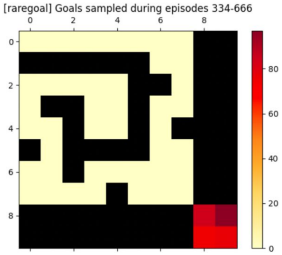


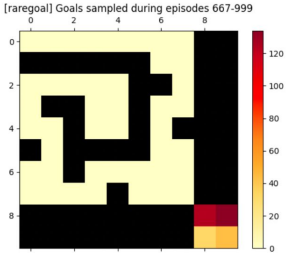




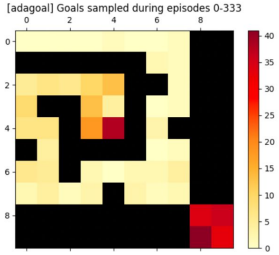
Rare goals

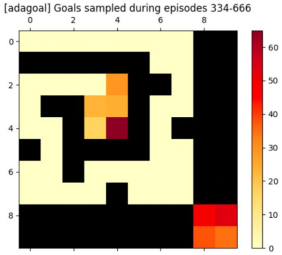


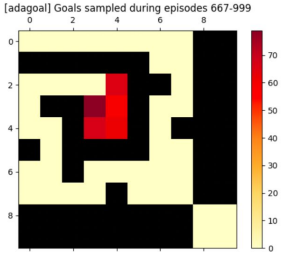




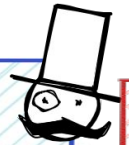
Ada goals







Tabular-AdaGoal: Sample Complexity Bounds



Thm: Sample Complexity of a Curriculum

[Tarbouriech et al., 2022]

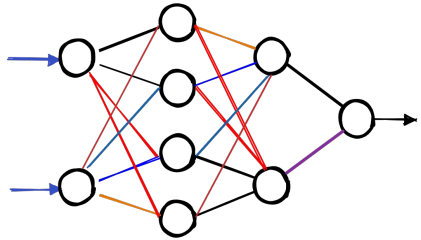
Let $L_{\max} = 2^{\ell}$, then AdaGoal run over multiple phases with $L = 2, 2^2, \ldots, 2^{\ell}$ is $(\epsilon, \delta, L_{\max})$-PAC with sample complexity

$$\mathbb{E}[\tau] = \widetilde{O}\left(\frac{L_{\max}^3 SA}{\epsilon^2}\right)$$

Remarks

- Performing a curriculum over the radius only adds a logarithmic factor

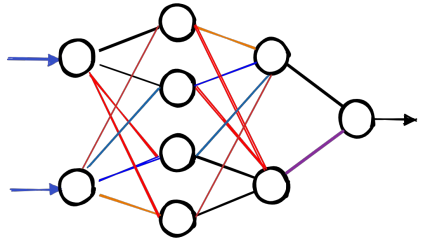
Deep-AdaGoal



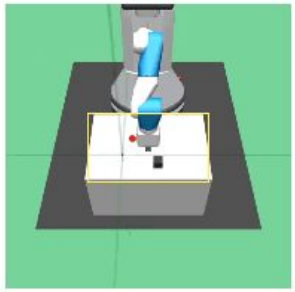
Similar to **value disagreement** [Zhang et al., 2020]

$$\mathcal{E}_k(g) = \mathrm{std}\left\{ V_1^{\pi}(s_0; g), \ldots, V_J^{\pi}(s_0; g) \right\}$$

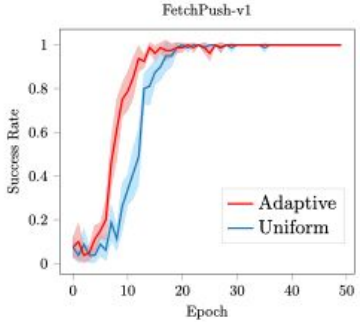
Deep-AdaGoal

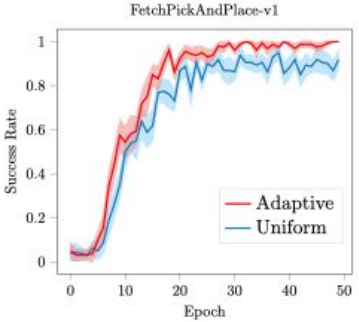


Goal
prior knowledge

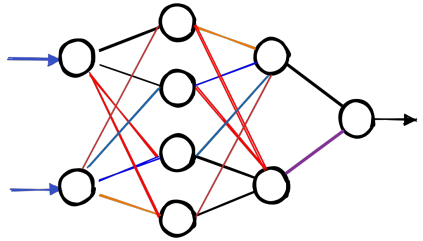


$\mathcal{G}_{train} = \mathcal{G}_{test}$

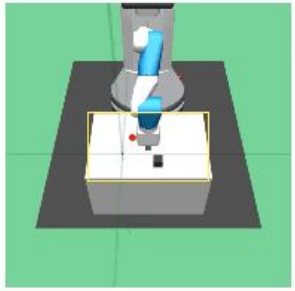




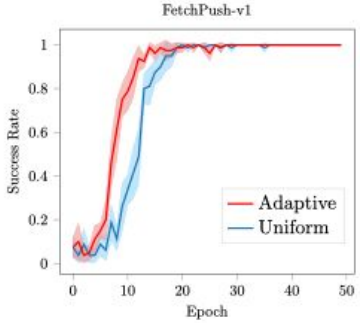
Deep-AdaGoal

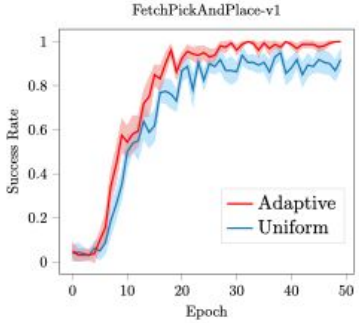


Goal prior knowledge

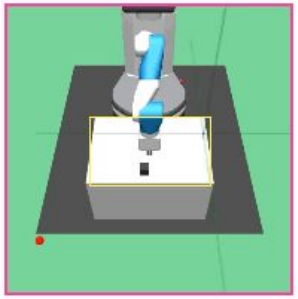


$\mathcal{G}_{train} = \mathcal{G}_{test}$

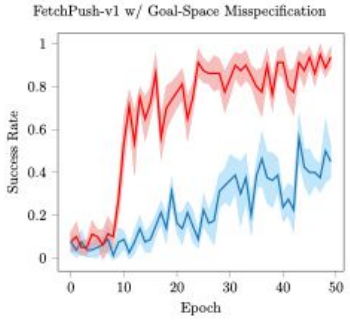


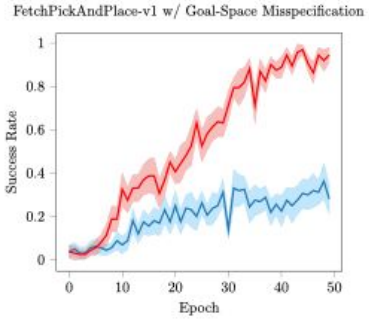


Goal misspecification



$\mathcal{G}_{train} \supset \mathcal{G}_{test}$





Summary

- MGE formalizes **unsupervised goal-based exploration**
- AdaGoal formalizes the popular **SYOG** principle
- AdaGoal is **minimax optimal in tabular MDPs** and **sample efficient in linear MDPs**
- AdaGoal can be implemented as a **deepRL** algorithm with **encouraging empirical results**

Unsupervised Exploration for **Incrementally** Controllable States

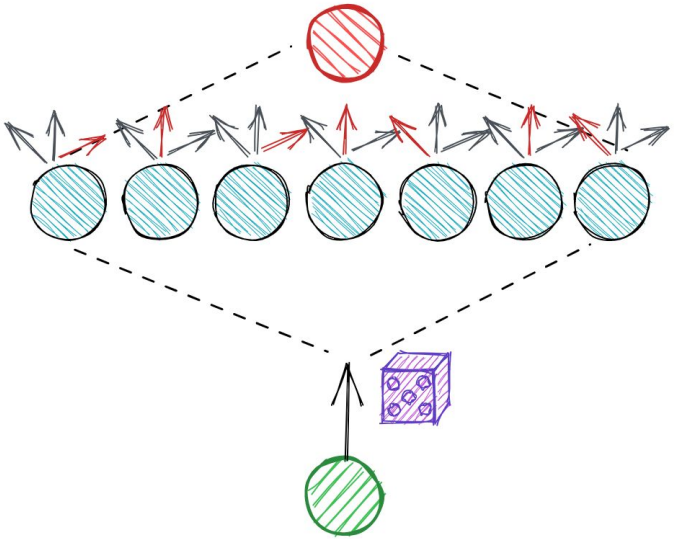
Meta

Limitations of UnsupExp of Controllable States

Thm: Sample Complexity [Tarbouriech et al., 2022]

AdaGoal is (ϵ, δ, L)-PAC and $S \gg S_L$

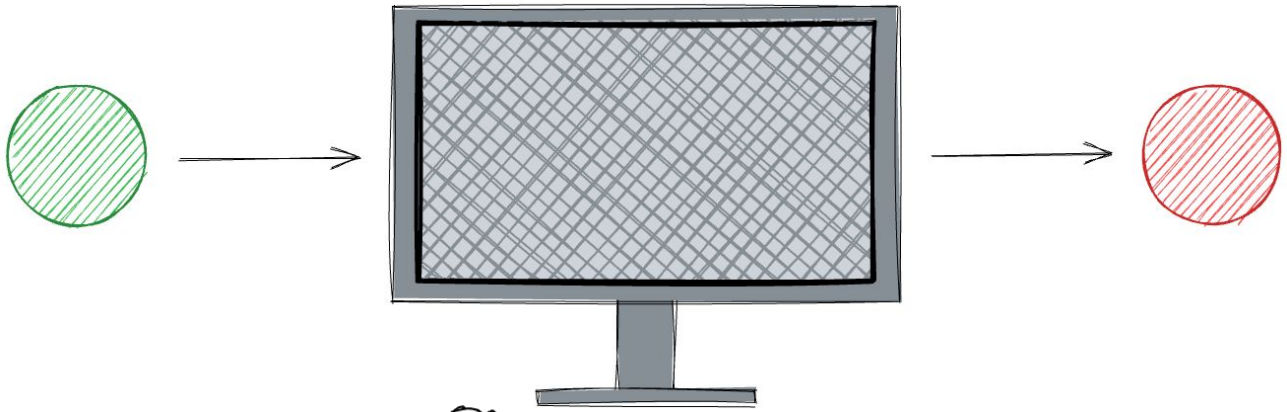
$$\mathbb{E}[\tau] = \widetilde{O}\left(\frac{L^3 S A}{\epsilon^2}\right)$$



2-step controllable state

(H>>1)-step controllable states

Limitations of UnsupExp of Controllable States



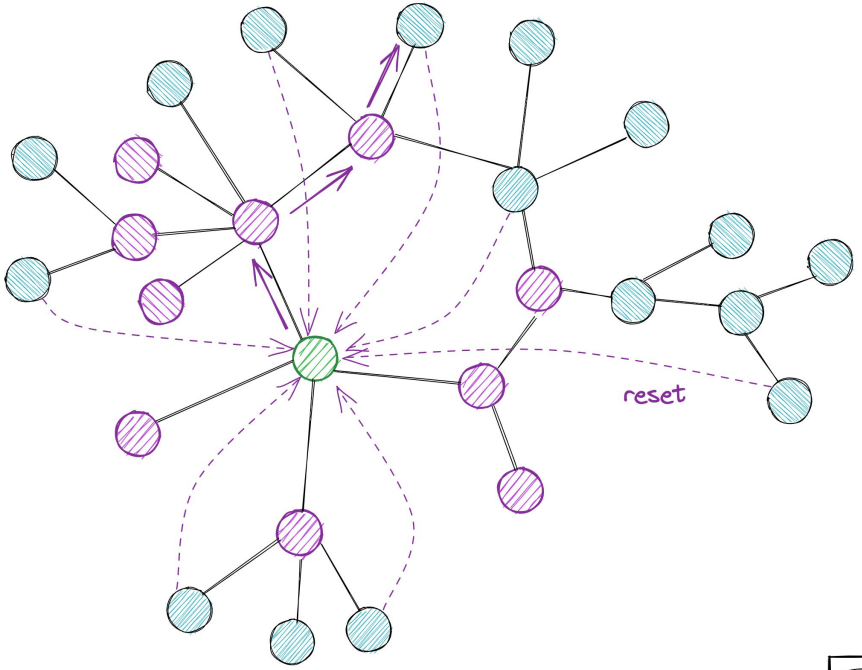
Rare goal sampling samples the noisy TV and ignore the red goal
⇒ “goal” inefficient

AdaGoal prioritizes the red goal but still needs to learn the optimal action at noisy TV states
⇒ “sample” inefficient

Incrementally Controllable States

Policy π *restricted on* $\mathcal{S}'$

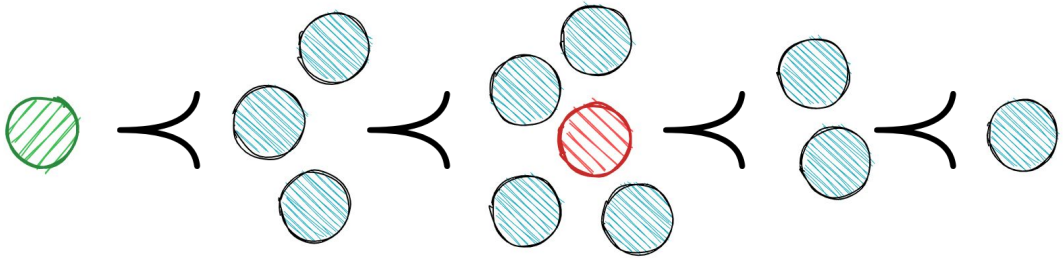
$$\pi(s) = a_{\text{reset}}$$
$$\text{for all } s \notin \mathcal{S}'$$



S. Lim & P. Auer, *Autonomous Exploration For Navigating In MDPs*, COLT-2012.

Incrementally Controllable States

Given a partial order $\prec$ *on* $\mathcal{S}$

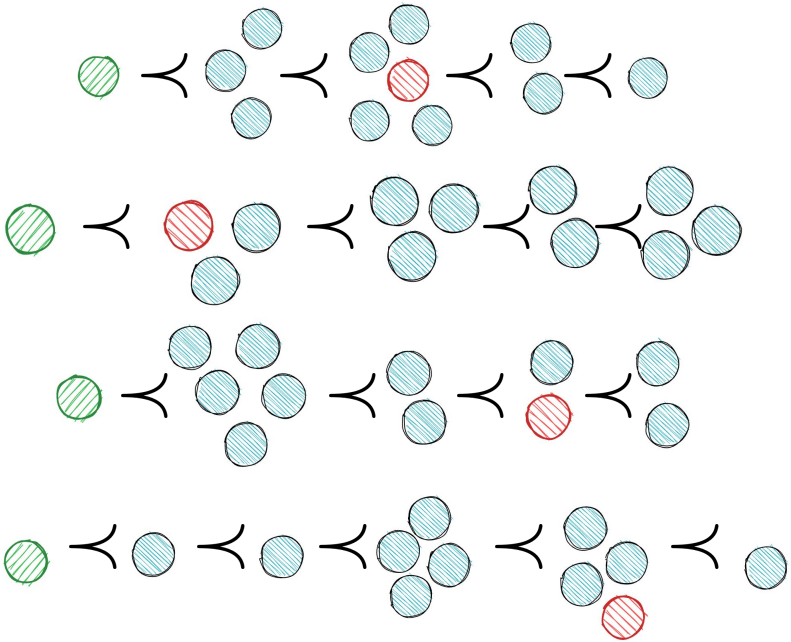


$s \in \mathcal{S}_L^{\prec}$ L-incrementally controllable state on partial order $\prec$

$\exists \pi$ *restricted on* $\left\{ s' \in \mathcal{S}_L^{\prec} : s' \prec s \right\}$ $\quad V^{\pi}(s_0 \to s) \leq L$

S. Lim & P. Auer, *Autonomous Exploration For Navigating In MDPs*, COLT-2012.

Incrementally Controllable States



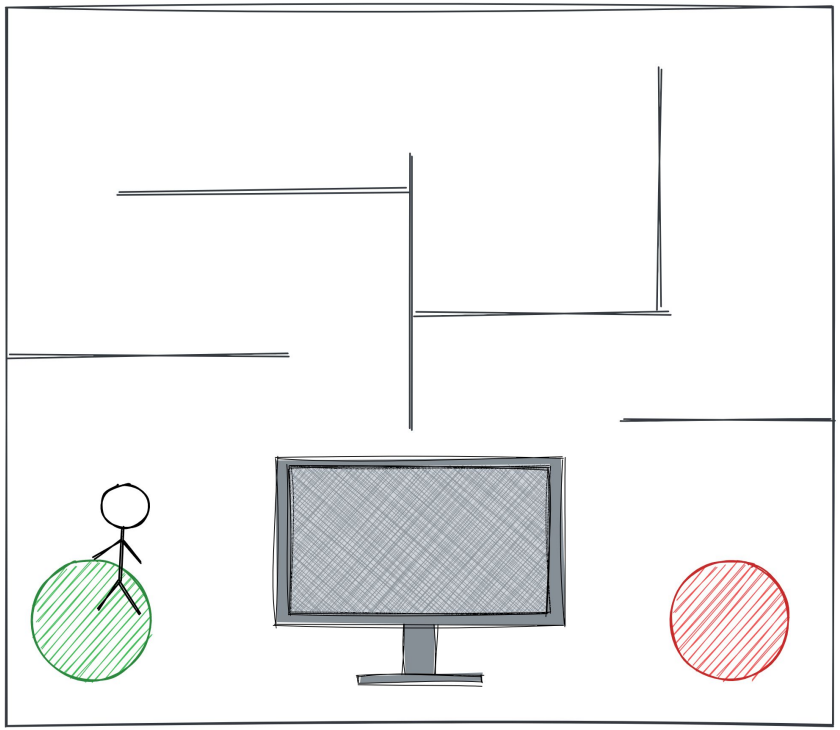
$$s \in \mathcal{S}_L^{\rightarrow} = \bigcup_{\prec} \mathcal{S}_L^{\prec}$$

Set of L-incrementally controllable states

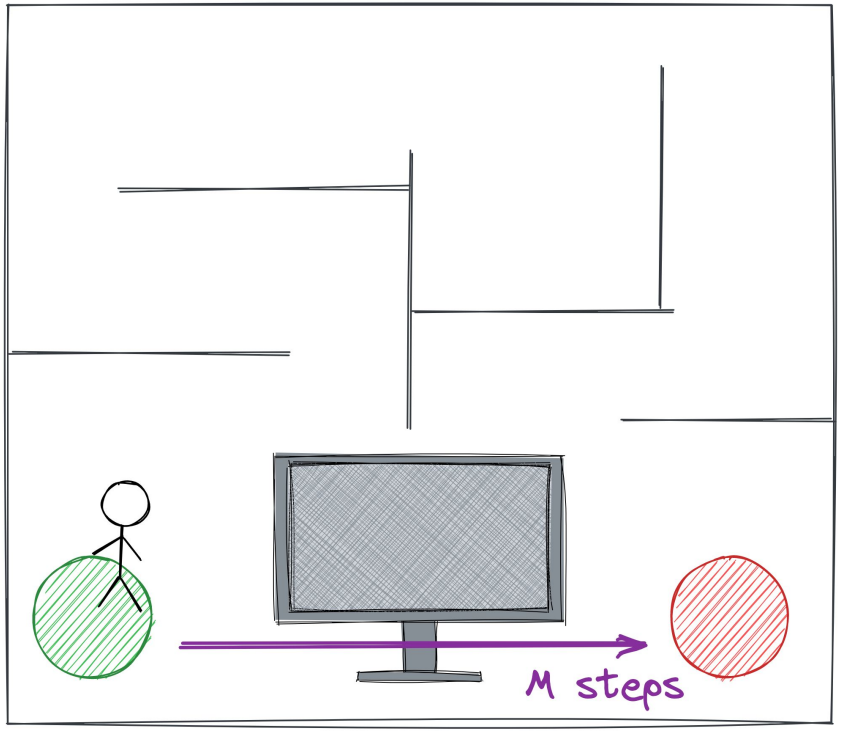
A state is L-incrementally controllable if it can be reached in L steps on average by only traversing states that are incrementally controllable

S. Lim & P. Auer, *Autonomous Exploration For Navigating In MDPs*, COLT-2012.

Incrementally Controllable States

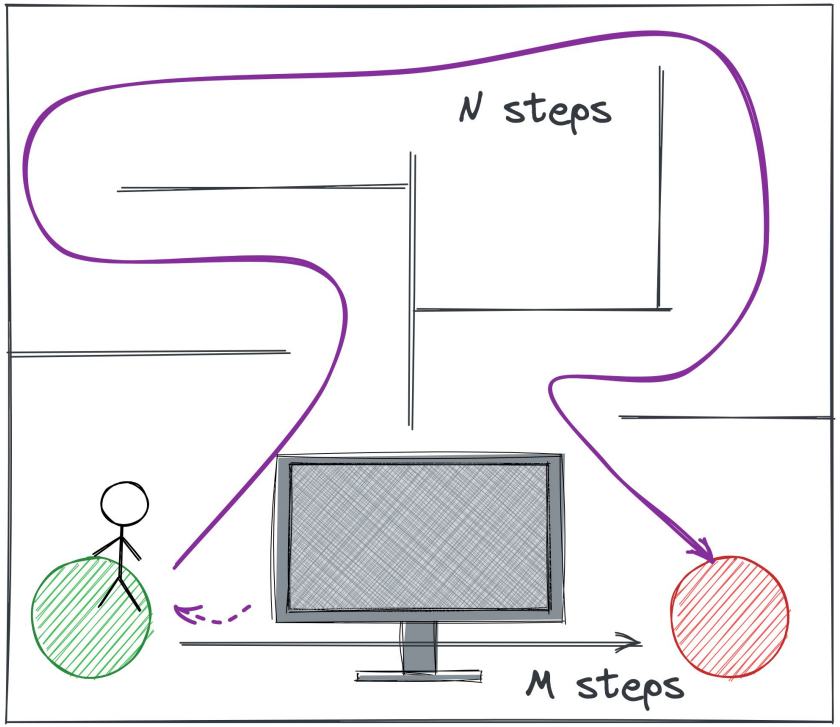


Incrementally Controllable States



M-step
controllable

Incrementally Controllable States

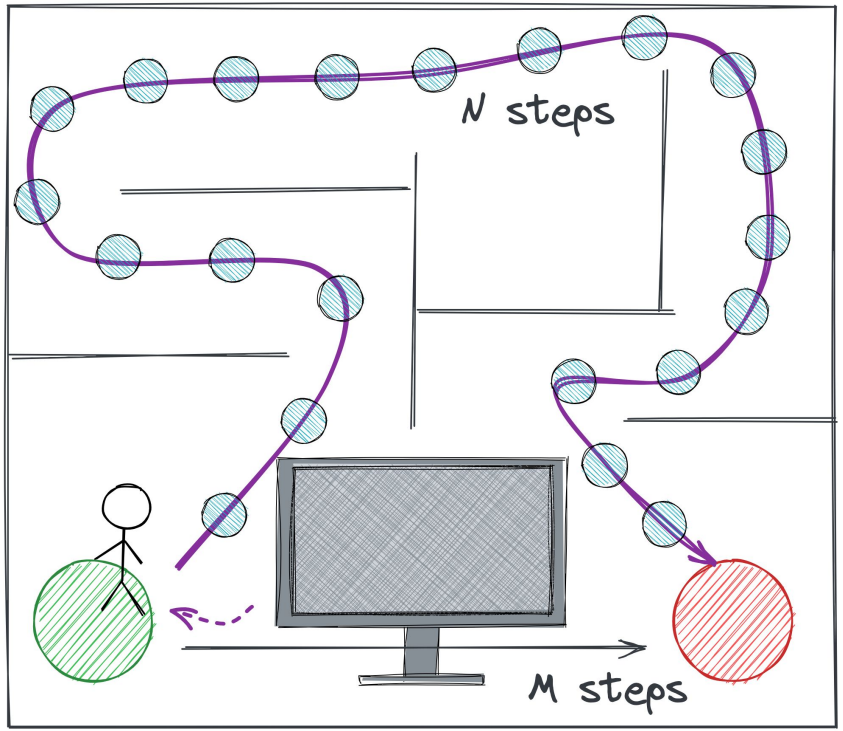


M-step controllable

M << N

N-step incrementally controllable

Incrementally Controllable States



M-step controllable

M << N

N-step incrementally controllable

Depending on the value of L, the state may be in $\mathcal{S}_L$ but not in $\mathcal{S}_L^{\rightarrow}$

Incremental Unsup. Exploration (aka Autonomous Exploration)

*Definition of **AX***

- Reset action a_{reset} s.t. $p(s_0|s, a_{\text{reset}}) = 1$
- Goal radius L
- Accuracy level ϵ
- Goal set $\mathcal{G}_L^{\rightarrow}$

Incremental Unsup. Exploration (aka Autonomous Exploration)

(ϵ, δ, L)-***AX* Learner***

Agent stops after $\tau = \text{poly}\left(S_L^{\rightarrow}, A, \log(1/\delta), 1/\epsilon, L, \log(S)\right)$ steps and

$$\mathbb{P}\left[\begin{array}{c} \textit{Accurate goal set identification} \\ \mathcal{G}_L^{\rightarrow} \subseteq \mathcal{X} \subseteq \mathcal{G}_{L+\epsilon}^{\rightarrow} \\ \textit{Near-optimal goal-based policy} \\ V^{\pi_g}(s_0 \rightarrow g) \leq V^{\star}_{\mathcal{G}_L^{\rightarrow}}(s_0 \rightarrow g) + \epsilon \quad \forall g \in \mathcal{X} \end{array}\right] \geq 1 - \delta$$

Incremental Unsup. Exploration (aka Autonomous Exploration)

(ϵ, δ, L)-***AX* Learner***

Agent stops after $\tau = \text{poly}(S_L^{\rightarrow}, A, \log(1/\delta), 1/\epsilon, L, \log(S))$ steps and

$$\mathbb{P}\left[\begin{array}{c} \textit{Accurate goal set identification} \\ \mathcal{G}_L^{\rightarrow} \subseteq \mathcal{X} \subseteq \mathcal{G}_{L+\epsilon}^{\rightarrow} \\ \\ \textit{Near-optimal goal-based policy} \\ V^{\pi_g}(s_0 \rightarrow g) \leq V^{\star}_{\mathcal{G}_L^{\rightarrow}}(s_0 \rightarrow g) + \epsilon \quad \forall g \in \mathcal{X} \end{array}\right] \geq 1 - \delta$$

Incremental Unsup. Exploration (aka Autonomous Exploration)

(ϵ, δ, L)-***AX* Learner***

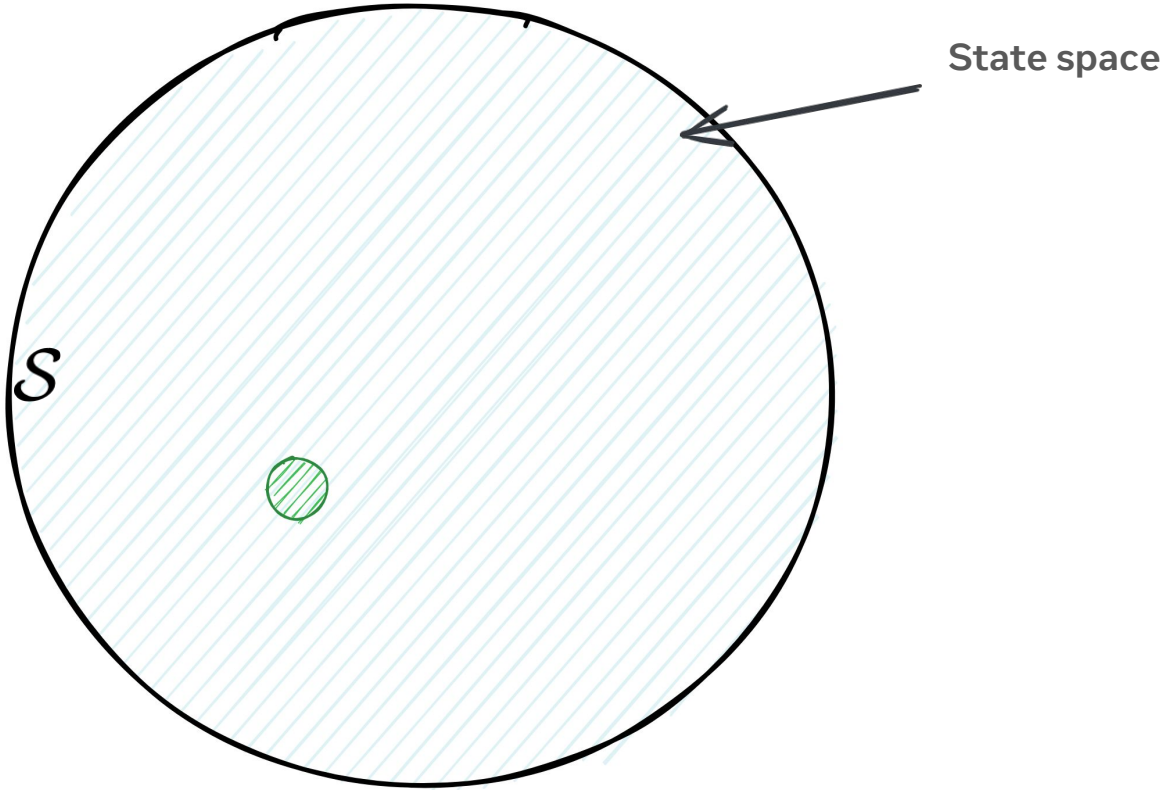
Agent stops after $\tau = \text{poly}(S_L^{\rightarrow}, A, \log(1/\delta), 1/\epsilon, L, \log(S))$ steps and

$$\mathbb{P}\Big[\ \textit{Accurate goal set identification}\quad \mathcal{G}_L^{\rightarrow} \subseteq \mathcal{X} \subseteq \mathcal{G}_{L+\epsilon}^{\rightarrow}$$

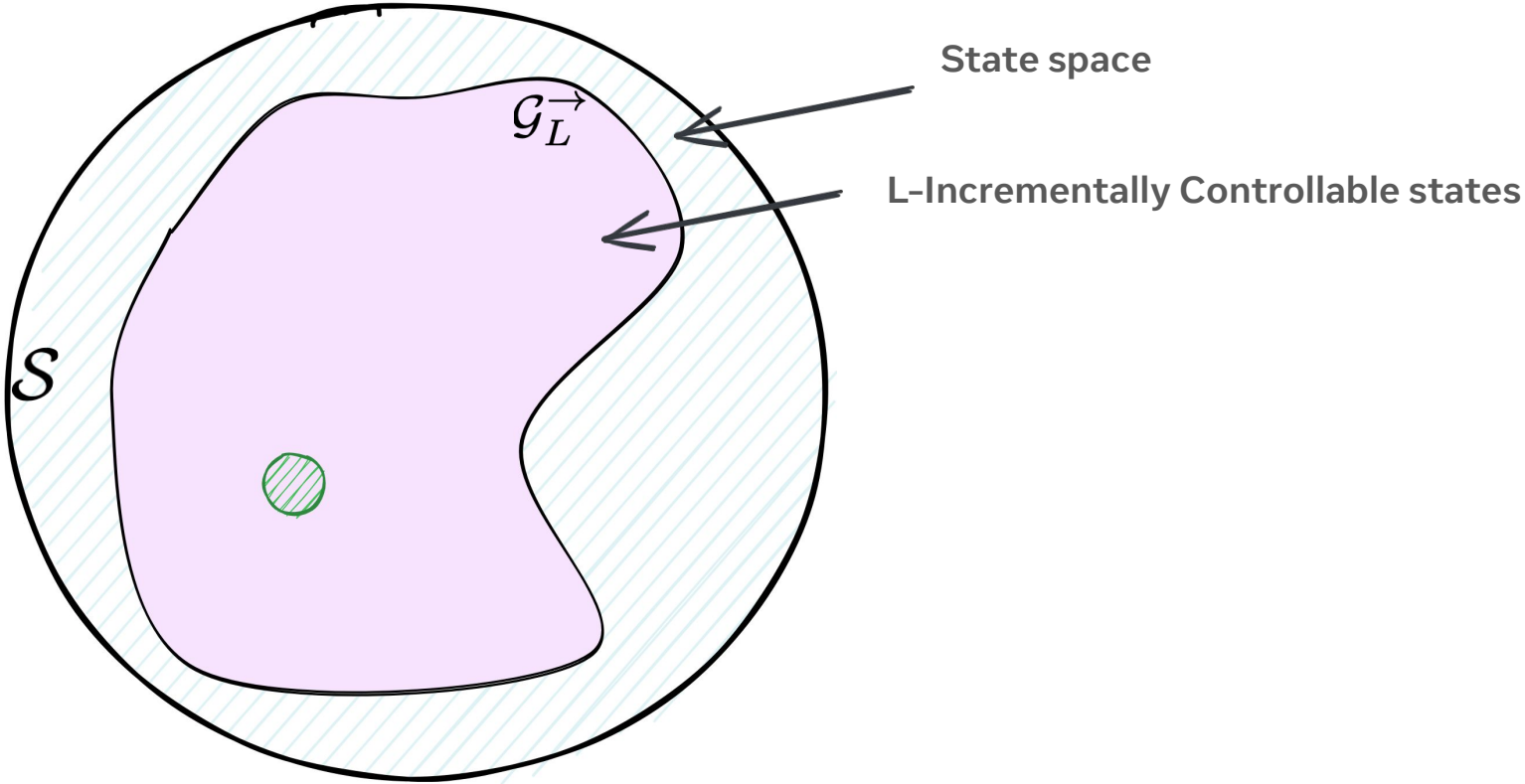
$$\textit{Near-optimal goal-based policy}\quad V^{\pi_g}(s_0 \to g) \leq V^{\star}_{\mathcal{G}_L^{\rightarrow}}(s_0 \to g) + \epsilon \quad \forall g \in \mathcal{X}\ \Big] \geq 1-\delta$$

Optimal policy restricted on $\mathcal{G}_L^{\rightarrow}$

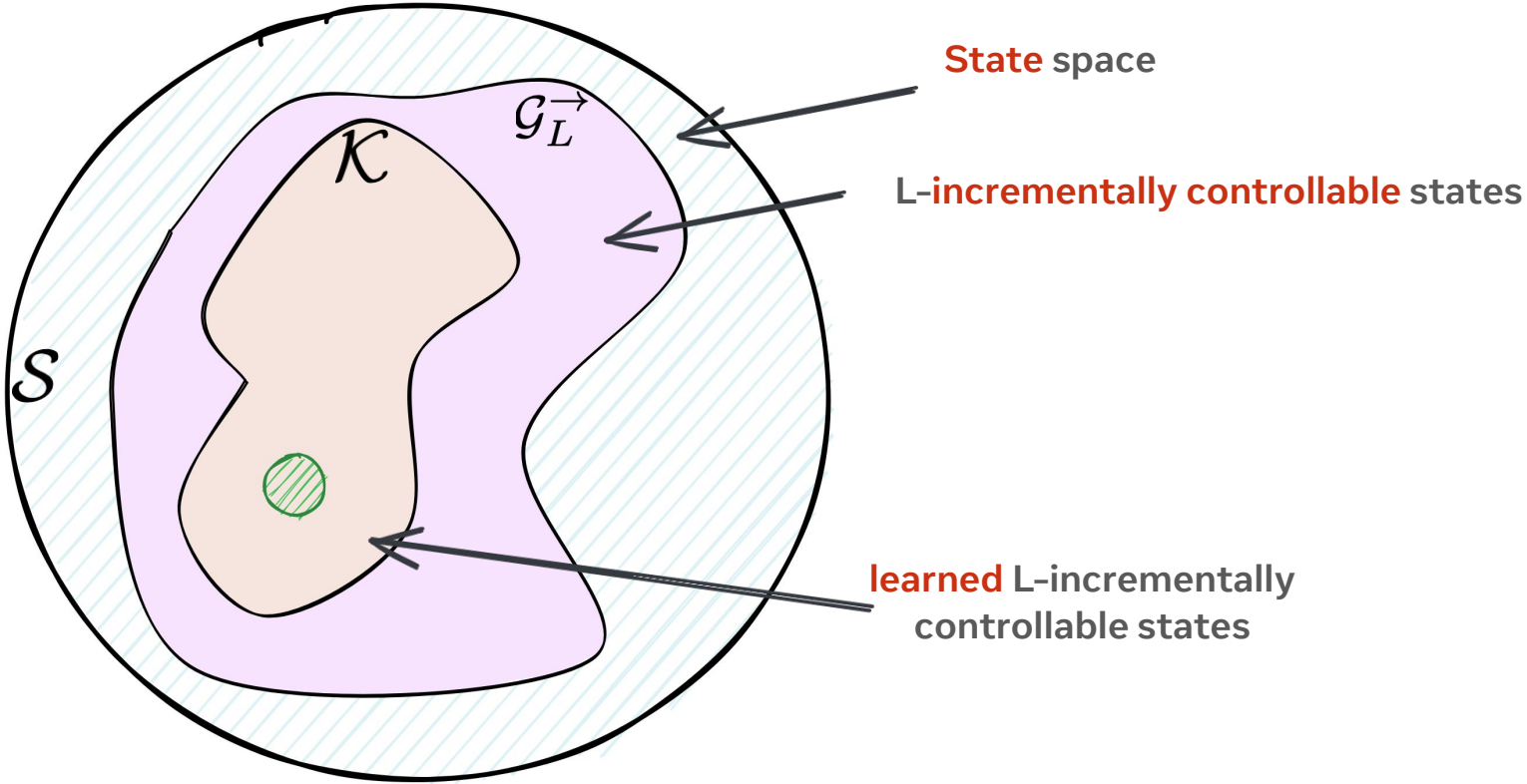
Discover and Control – DISCO



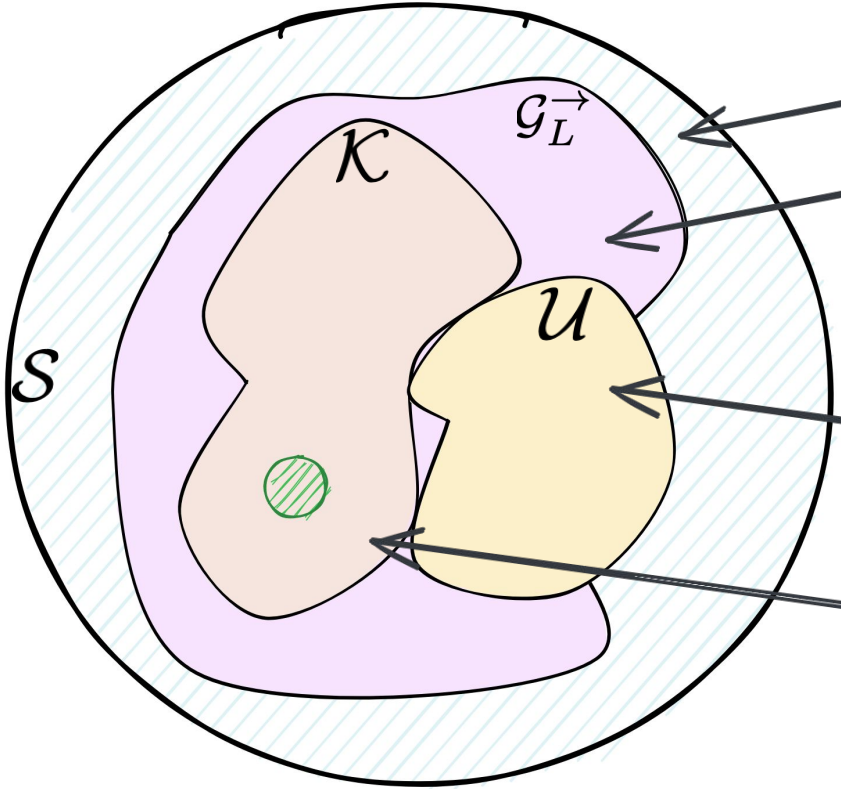
Discover and Control – DISCO



Discover and Control – DISCO



Discover and Control – DISCO



State space

L-incrementally controllable states

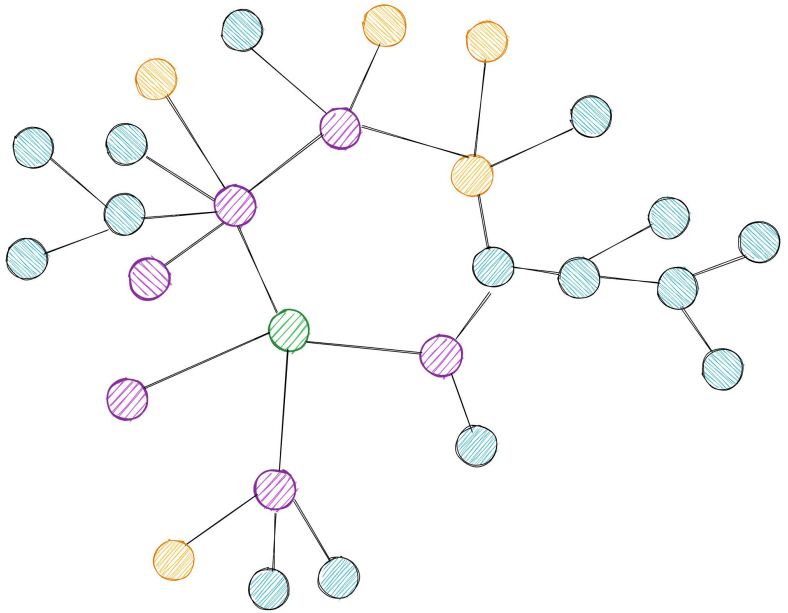
candidate L-incrementally controllable states

learned L-incrementally controllable states

Discover and **Co**ntrol – DISCO

Discover & Control

1. **Refine** model and **discover** states
2. **Update** policy and **learned** states
3. *If* policy is **good** *then* STOP and return
   otherwise jump to 1.

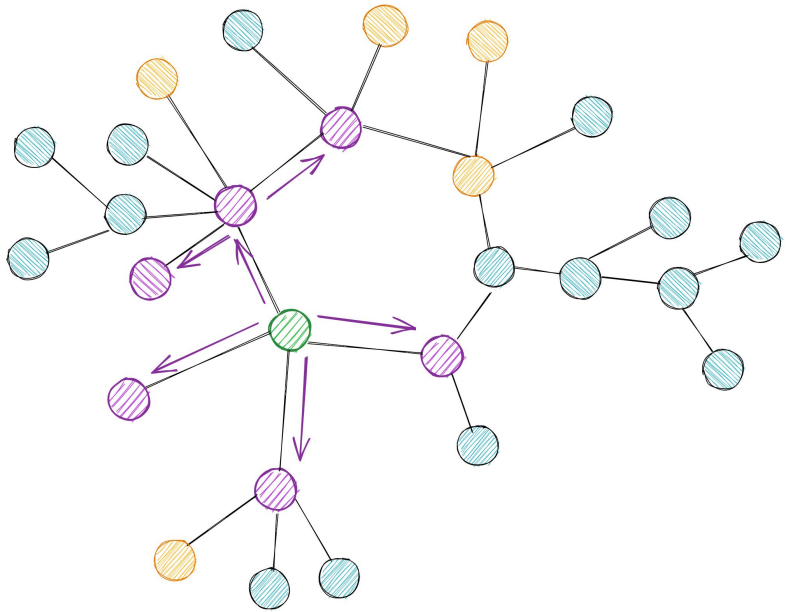


Discover and Control – DISCO

Discover & Control

1. **Refine** model and **discover** states
2. **Update** policy and **learned** states
3. *If* policy is **good** *then* STOP and return
   otherwise jump to 1.

$$\forall s \in \mathcal{K}_k \quad V^{\pi_k}_{\mathcal{K}_k}(s_0 \to s) \leq L + \epsilon$$



generative model for states in $\mathcal{K}_k$

with **cost (L+eps)** per sample

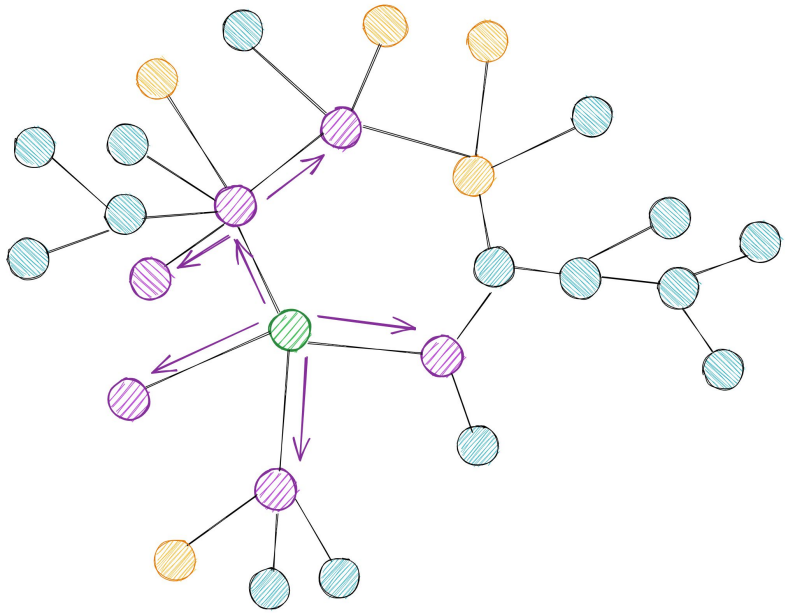
Discover and Control – DISCO

Discover & Control

1. **Refine** model and **discover** states
2. **Update** policy and **learned** states
3. *If* policy is **good** *then* STOP and return *otherwise* jump to 1.

$$\forall s \in \mathcal{K}_k, a \in \mathcal{A}$$

Collect samples until $N_k(s,a) \geq \phi(\mathcal{K}_k)$



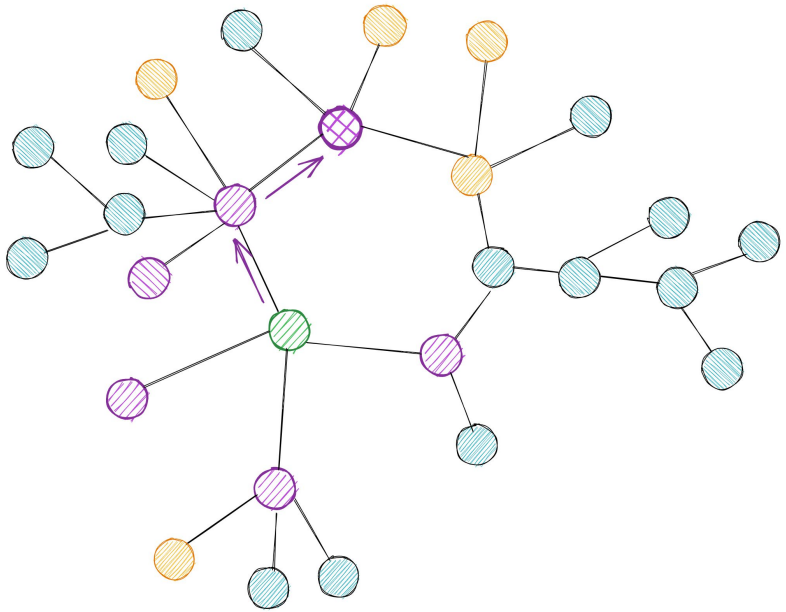
Discover and Control – DISCO

Discover & Control

1. **Refine** model and **discover** states
2. **Update** policy and **learned** states
3. *If* policy is **good** *then* STOP and return *otherwise* jump to 1.

$$\forall s \in \mathcal{K}_k, a \in \mathcal{A}$$

Collect samples until $N_k(s,a) \geq \phi(\mathcal{K}_k)$



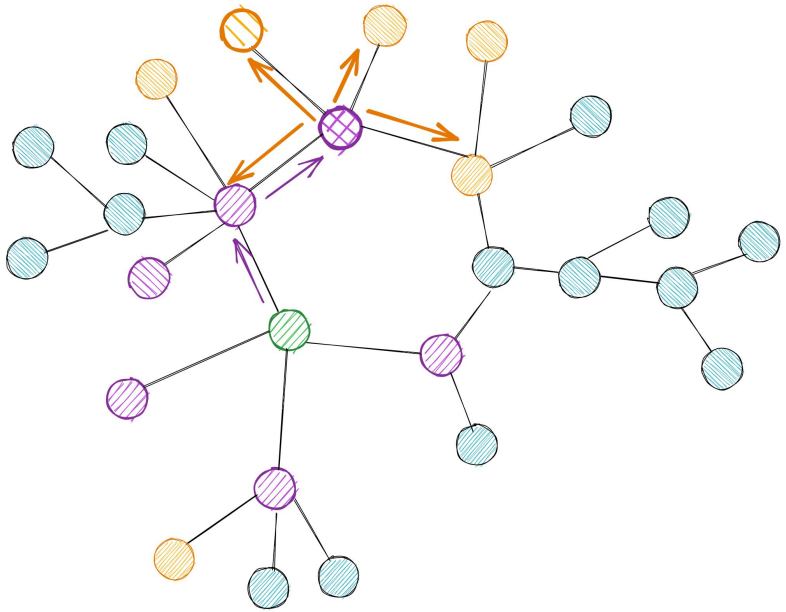
Discover and Control – DISCO

Discover & Control

1. **Refine** model and **discover** states
2. **Update** policy and **learned** states
3. *If* policy is **good** *then* STOP and return *otherwise* jump to 1.

$$\forall s \in \mathcal{K}_k, a \in \mathcal{A}$$

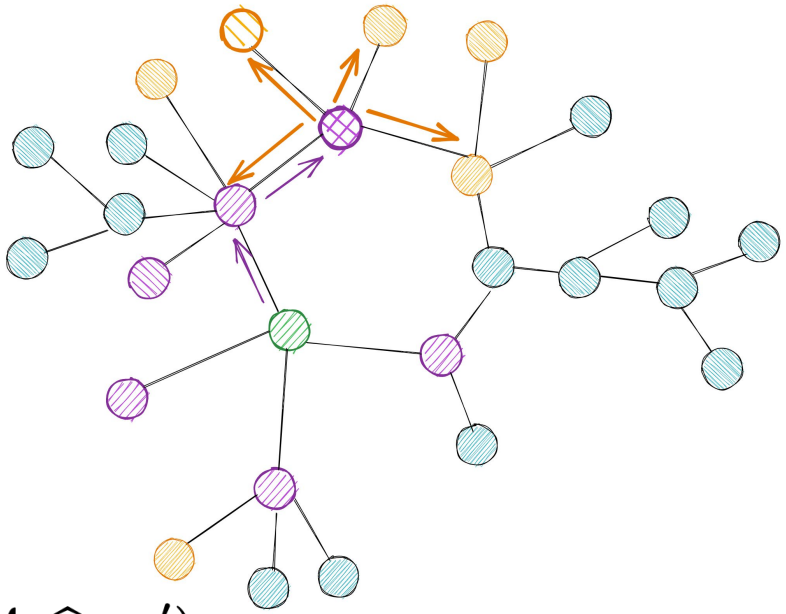
Collect samples until $N_k(s,a) \geq \phi(\mathcal{K}_k)$



Discover and Control – DISCO

Discover & Control

1. **Refine** model and **discover** states
2. **Update** policy and **learned** states
3. *If* policy is **good** *then* STOP and return *otherwise* jump to 1.



$$\forall s' \in \mathcal{U}_k$$

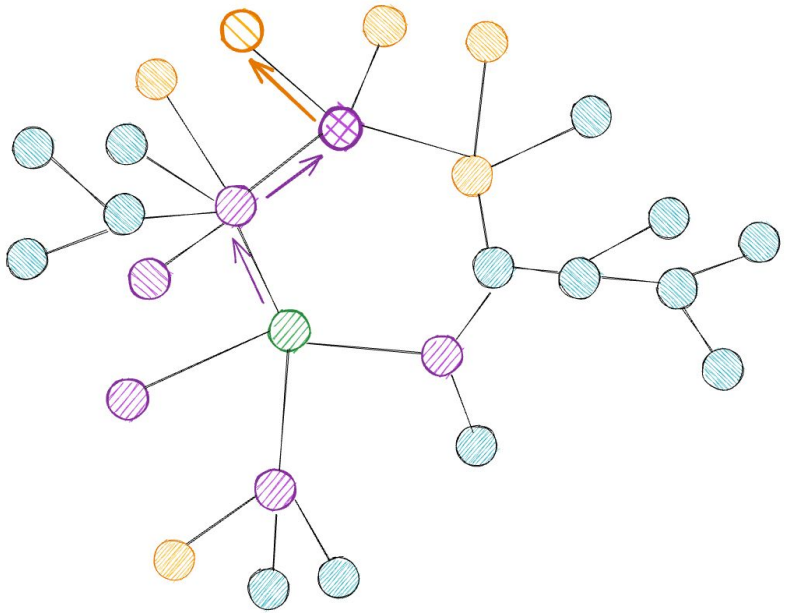
$$\left(\pi_{k+1}(s'); \overline{V}_{\mathcal{K}_k}^{\pi_{k+1}}(s_0 \to s')\right) = \mathrm{OVI}(\mathcal{K}_k, \mathcal{A}, \widehat{p}_k; s')$$

Optimistic policy and value function

Discover and Control – DISCO

Discover & Control

1. **Refine** model and **discover** states
2. **Update** policy and **learned** states
3. *If* policy is **good** *then* STOP and return
   otherwise jump to 1.



$$s^\dagger = \arg\min_{s' \in \mathcal{U}_k} \overline{V}^{\pi_{k+1}}_{\mathcal{K}_k}(s_0 \to s')$$

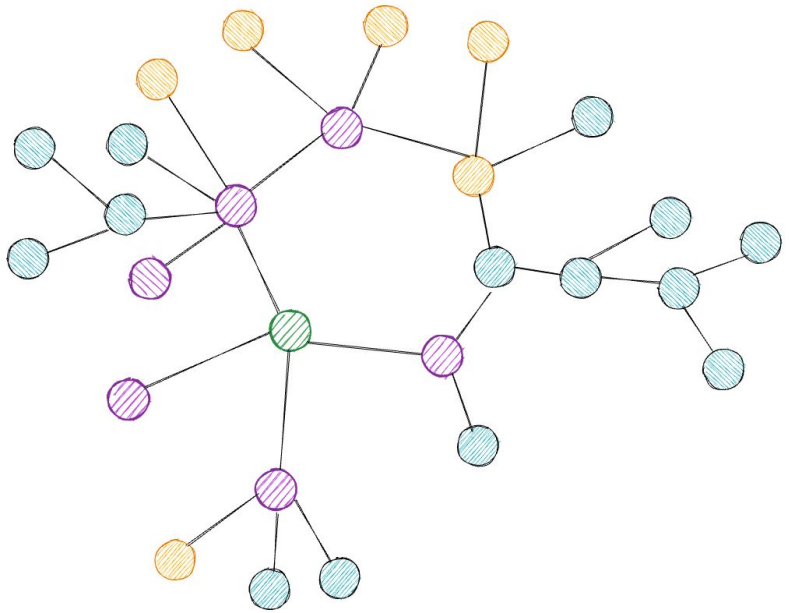
If $\overline{V}^{\pi_{k+1}}_{\mathcal{K}_k}(s_0 \to s^\dagger) \leq L$ then $\mathcal{K}_{k+1} = \mathcal{K}_k \cup \{s^\dagger\}$

Consolidate new state

Discover and Control – DISCO

Discover & Control

1. **Refine** model and **discover** states
2. **Update** policy and **learned** states
3. *If* policy is **good** *then* STOP and return
   otherwise jump to 1.



If $\overline{V}^{\pi_{k+1}}_{\mathcal{K}_k}(s_0 \to s^\dagger) > L$ then 

Not even the most optimistic state is optimistically L-incrementally controllable

Tabular-DISCO

Thm: Sample Complexity [Tarbouriech et al., 2020]

DISCO is (ϵ, δ, L)-AX* with sample complexity

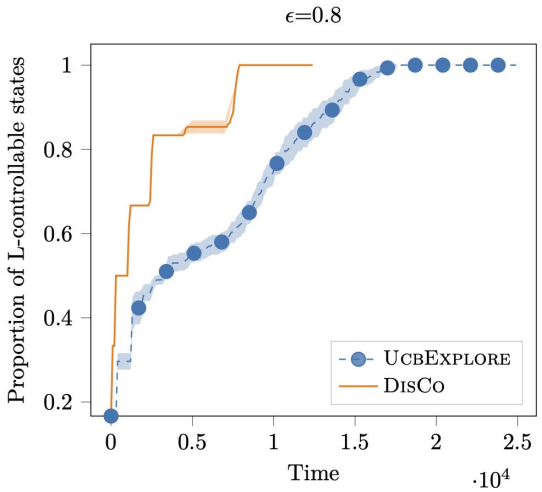
$$\mathbb{E}[\tau] = \widetilde{O}\left(\frac{L^5 \Gamma_{L+\epsilon} S_{L+\epsilon} A}{\epsilon^2}\right)$$

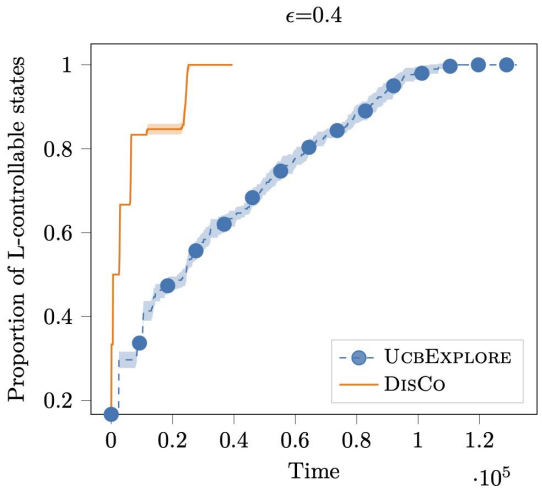
Remarks

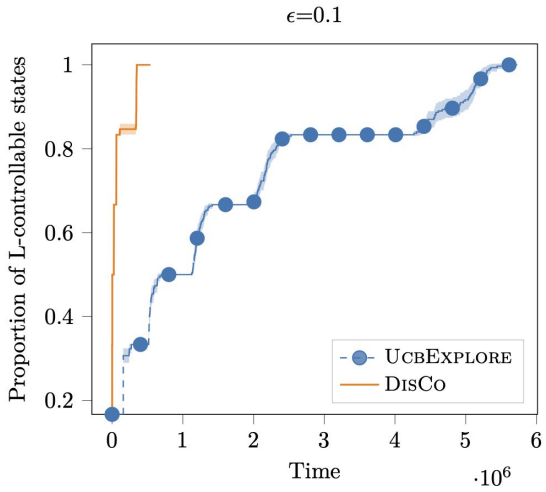
Compared to UCBExplore

- Stronger policy guarantees
- Better than $O(L^6 / \text{eps}^3)$
- Worse than $O(S_L)$

DISCO: A Simple **Example**







A **Recent** Result

Thm: Sample Complexity

DISCO+MGE= Re-MG-SSP is (ϵ, δ, L)-AX* with sample complexity

$$\mathbb{E}[\tau] = \widetilde{O}\left(\frac{L^3 S_{2L} A}{\epsilon^2}\right)$$

Remarks

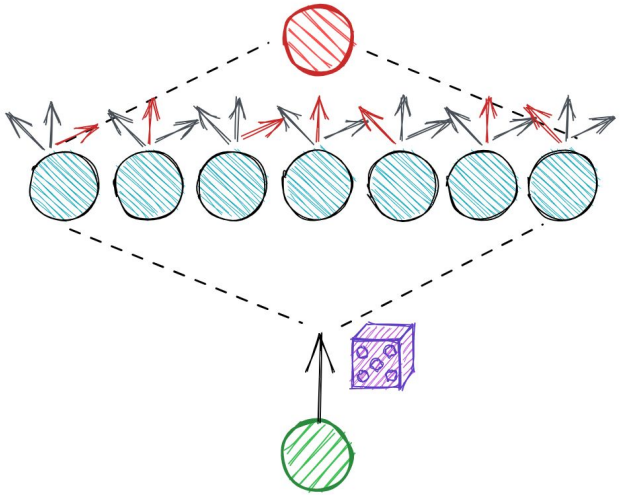
Compared to DISCO

- Better than $O(L^5 \Gamma)$
- Worse than $O(S_{L+eps})$

Cai et al., *Near-Optimal Algorithms for Autonomous Exploration and Multi-Goal Stochastic Shortest Path*, ICML-2022.

Limitations and Open Questions

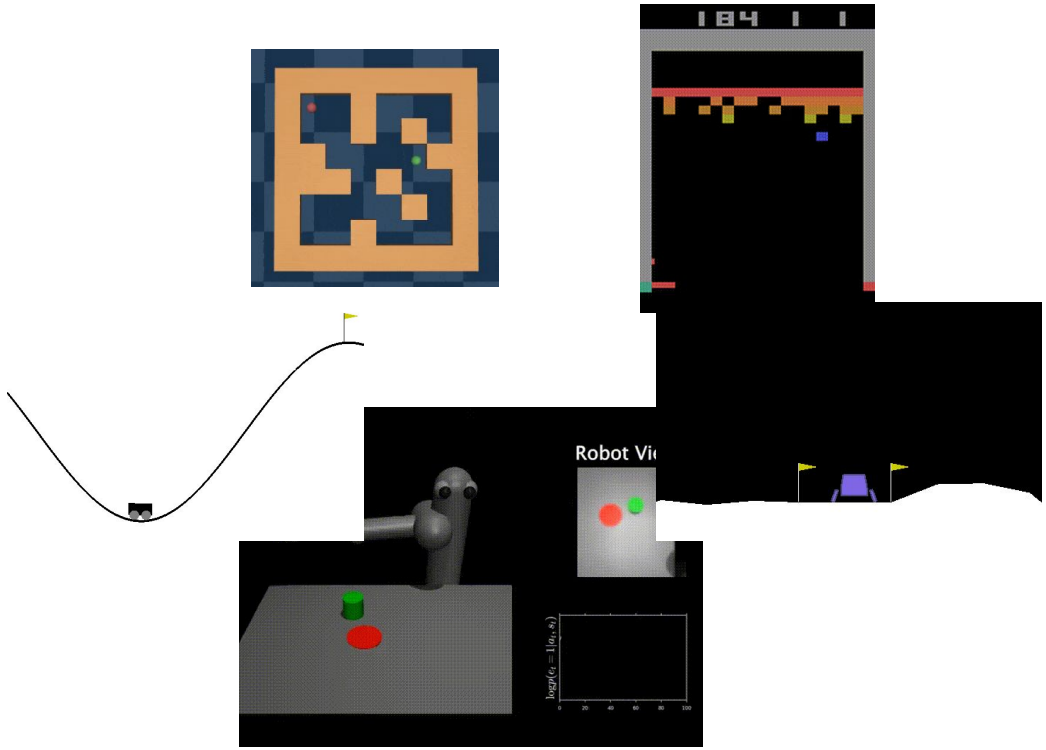
- **Deep-DISCO**: Unlike AdaGoal, DISCO is intrinsically tabular (e.g., listing consolidated and candidate states, prescribing number of samples)
- **Unified algorithm** for controllable and inc.controllable states
- Recent result improves (some aspects of) our bound but still **not minimax optimal**
- **Problem-dependent** analysis
- **SSP** with incrementally-controllable goal
- Incremental controllability at different levels of **temporal abstraction**



Limitations and **Open Questions** (cont'd)

Is really $\mathcal{S}_L \neq \mathcal{S}_L^{\rightarrow}$ in **"practice"**?

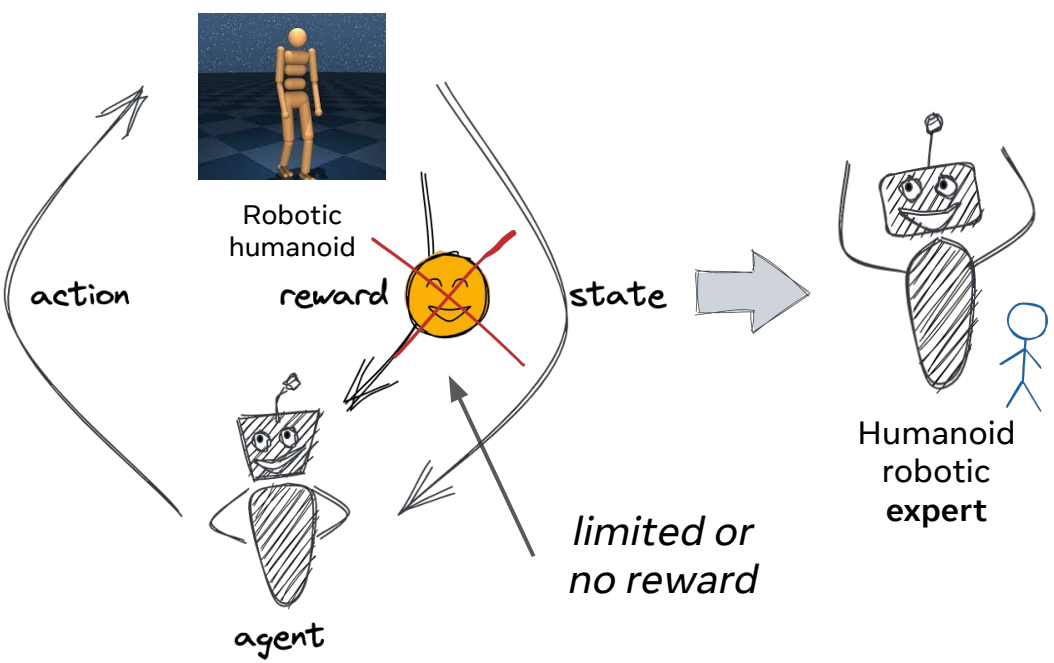
- **Deterministic** MDPs
- **Smooth** MDPs?



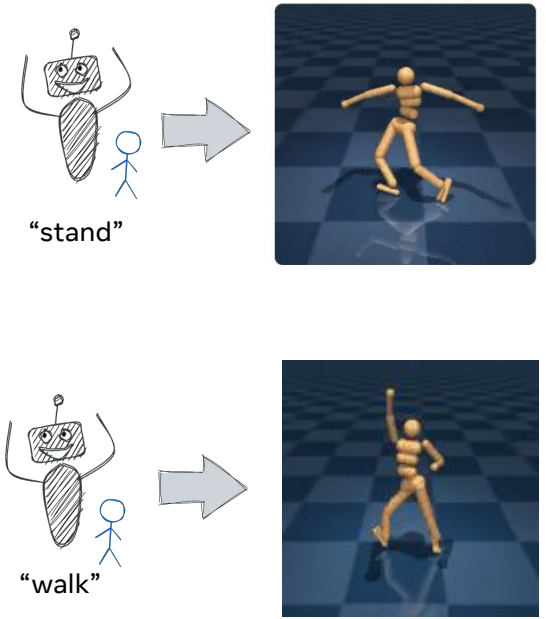
Discussion

Meta

From Specialized to **Universally Controllable** Agents



Unsupervised RL

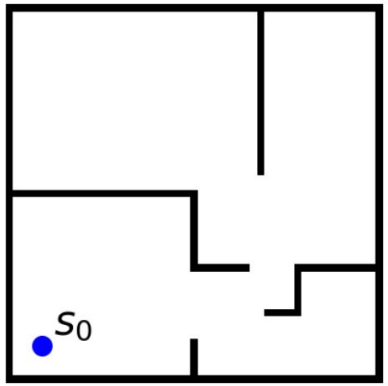


Zero/few-shot learning

From Learning to Control States to **Skill Discovery**

- Goal-based policy:
  - **Too “flat”**
  - **1 goal = 1 policy**
  - **No compositionality**
- Performance requirement **too strong (zero-shot)**

$$V^{\pi_g}(s_0 \to g) \leq V^{\star}(s_0 \to g) + \epsilon \quad \forall g \in \mathcal{X}$$

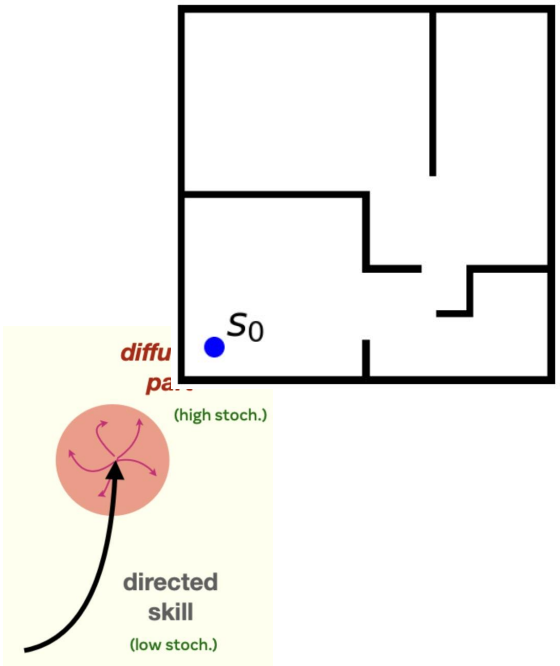


⇒ Generate a few policies (**options**) that **cover the goal space** and can be **efficiently fine-tuned**

Kamienny*, Tarbouriech*, **Lazaric**, Denoyer, “Direct then Diffuse: Incremental Unsupervised Skill Discovery for State Covering and Goal Reaching, ICLR-2022

From Learning to Control States to **Skill Discovery**

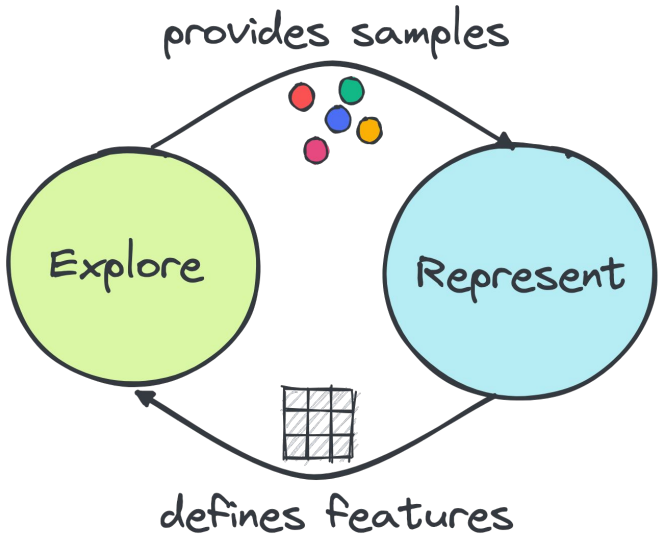
- **UPSIDE** – Three ingredients
  - Inductive bias in skills: **coverage and directedness**
  - **Tree** structure
  - Novel constrained version of **mutual information**
- Open questions
  - Formal definition of **skill discovery**
  - Connections between **MI** and **MGE**



Kamienny*, Tarbouriech*, **Lazaric**, Denoyer, "Direct then Diffuse: Incremental Unsupervised Skill Discovery for State Covering and Goal Reaching, ICLR-2022

The Role of **Representation** in Unsup. Exploration

- In **tabular** all states “equally” matter
- A **representation** defines what “matters”
- An **exploration** strategy provides “information”
- No “grounding” on reward

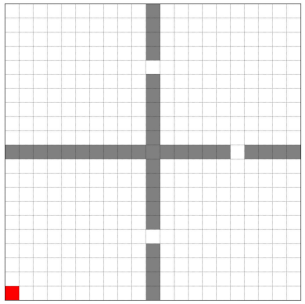


A. Erraqabi, M. Machado, M. Zhao, S. Sukhbaatar, **A. Lazaric**, D. Ludovic, Y. Bengio. *Temporal abstractions-augmented temporally contrastive learning: An alternative to the Laplacian in RL.* UAI-2022.

D. Yarats, R. Fergus, A. **Lazaric**, L. Pinto. *Reinforcement Learning with Prototypical Representations.* ICML-2021.

The Role of **Representation** in Unsup. Exploration

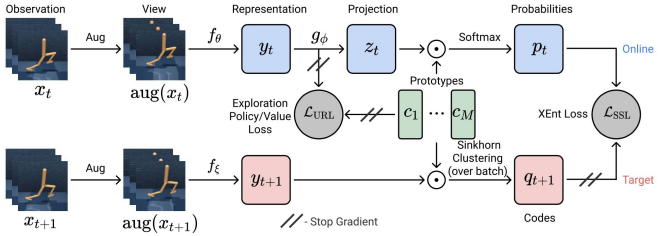
- TATC – Three steps
  - Follow exploration **driven by representation**
  - **Random exploration** at the end
  - **Update representation** with LapRep-like loss

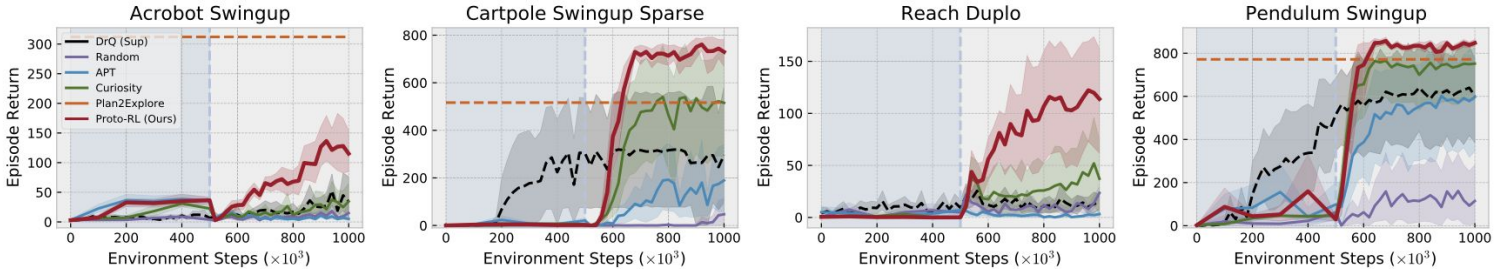


A. Erraqabi, M. Machado, M. Zhao, S. Sukhbaatar, **A. Lazaric**, D. Ludovic, Y. Bengio. *Temporal abstractions-augmented temporally contrastive learning: An alternative to the Laplacian in RL*. UAI-2022.

The Role of **Representation** in Unsup. Exploration

- **ProtoRL** – two ingredients
  - Representation learning (clustering-like)
    - Data augmentation + BYOL/SwaV-like rep.learning
    - Discrete prototypes
  - Exploration





D. Yarats, R. Fergus, A. **Lazaric**, L. Pinto. *Reinforcement Learning with Prototypical Representations.* ICML-2021.

From Goals to “Prompts”

- Beyond goals:
  - **Language-based** tasks (e.g., “set up living room environment for movie night”)
  - **Underspecified** tasks (e.g., “walk in a funny way”)
  - **Questions** (e.g., “what happens if I push the door?”)
- Change of protocol
  - Add **demonstrations** at train time
  - Add **corrections** at test time

Thank you!

“Walk in a funny way”

Meta