When Do Skills Help Reinforcement Learning?

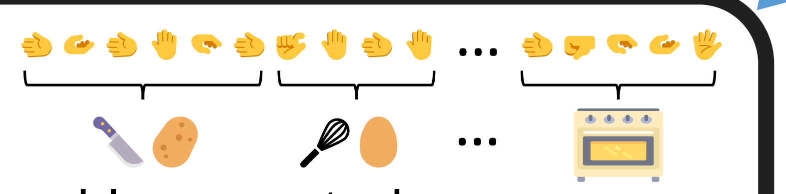
SuperUROP
Advanced Undergraduate Research Opportunities Program

Zhening Li[1] Gabriel Poesia[2] Armando Solar-Lezama[1] [1] MIT CSAIL [2] Stanford University
[1] {zli11010,asolar}@csail.mit.edu [2] poesia@stanford.edu

Paper

Introduction

Base actions:
RL skills:



Background: RL skills — components reusable across tasks

- Improve exploration and planning in environments where the agent learns to accomplish a goal ("sparse reward")

But RL skills are not widely used as they often don't work...

Goal: Theoretically understand when and how RL skills work.

Main result: See center panel A.

Preliminary definitions

- We focus on ***deterministic sparse-reward MDPs (DSMDPs)***, which are deterministic Markov decision processes (MDPs) with a single ***goal state***. Getting to the goal state is the only way to receive a reward, which is +1 by default.
- A ***solution*** to a state is a successful trajectory (sequence of actions leading to the goal state), cf. symbolic reasoning domains.
- A ***solution-separable*** DSMDP is one where every action sequence solves at most one state.
- A ***deterministic skill*** (from now on, ***skill***) in a DSMDP is a function from states to finite action sequences, i.e., we specify the sequence of actions for each possible initial state of the skill.
- A ***macroaction*** is a skill that produces the same sequence of actions regardless of initial state.

Notation: Subscript "**+**" denotes DSMDP augmented with skills; "**0**" denotes base DSMDP.

Modeling RL sample complexity *(episodic setting)*

Two stages: **Note:** *We assume size of state space ($|S|$) is constant. (Skills do not affect S.)*

1) Explore to gather experience

- $q(s)$: Pr[uniformly random policy solves s within H steps]
- Samples needed to solve every state once with a uniformly random policy: $\propto \frac{1}{|S|}\sum_s \frac{1}{q(s)}$
- Generalize to weighted mean: $\mathbb{E}_{s\sim p}[1/q(s)]$
- Switch to geometric mean to compensate for overestimation: $\exp \mathbb{E}_{s\sim p}[\log(1/q(s))]$

p-exploration difficulty:

$$J_{\text{explore}}(\mathcal{M};p,\delta) = \mathbb{E}_{s\sim p}[-\log q_{\mathcal{M},\delta}(s)]$$

(MDP initial state distribution)

Pr[uniformly random policy that at every step terminates w.p. $\delta \sim 1/H$ solves s]

2) Learn from gathered experience

Lemma. In value iteration with discount rate 1 and learning rate α, convergence of the value of state s takes time

$$\Theta(\alpha^{-1}|S||A|(d_{\mathcal{M}}(s) + \log(1/\varepsilon))).$$

- Constant $|S|, \alpha, \varepsilon$: $\Theta(|A| d_{\mathcal{M}}(s))$.
- Weigh states according to MDP initial state distribution p.

p-learning difficulty:

$$J_{\text{learn}}(\mathcal{M};p) = |A|\mathbb{E}_{s\sim p}[d_{\mathcal{M}}(s)]$$

size of action space

length of shortest solution to s

Experiments: A weighted average of J_{learn} and $\exp J_{\text{explore}}$ vs. sample complexity N: correlation at least $\sim$0.7 most of the time, over 32 action space variants of each of 4 base environments and 4 RL algorithms. See paper Section 3.3.

A

Compressible	skills	Abstracted solutions
DLLLDRURDL	L+:1	D1DR2
URDLRURDLLLL	URDL:2	2R21
LLDULLLLLLL		1DU1

Incompressible	skills	
RDUDLRUDRLULD		*[No compression achieved]*
LRDLLDURD		
UURRDLRLDDR		

Q: When does RL in a deterministic sparse-reward MDP benefit from skills?

A: When successful trajectories ("solutions") are ***compressible*** in the information-theoretic sense.

B

entropy of abstracted solutions

(a parameter)

$$\mathbf{IC}(\mathcal{M};p,P,\varepsilon) = \frac{\mathrm{H}[P] - \log\left(\frac{1-\varepsilon}{\varepsilon}\right)}{\mathbb{E}_{s\sim p}[d_{\mathcal{M}}(s)]\log\left(\frac{|A|}{1-\varepsilon}\right)}$$

p-incompressibility

size of action space

average length of a shortest solution

C

Theorem 4.2. *Skills benefit **learning from experience** less when **p-incompressibility** is high.*

$$\frac{J_{\text{learn}}(\mathcal{M}_+;p)}{J_{\text{learn}}(\mathcal{M}_0;p)} \geq \frac{|A_+|\log|A_0|}{|A_0|\log|A_+|}\mathbf{IC}(\mathcal{M}_0;p,P_{A_+},\varepsilon)$$

distribution of shortest abstracted solutions to states $\sim p$

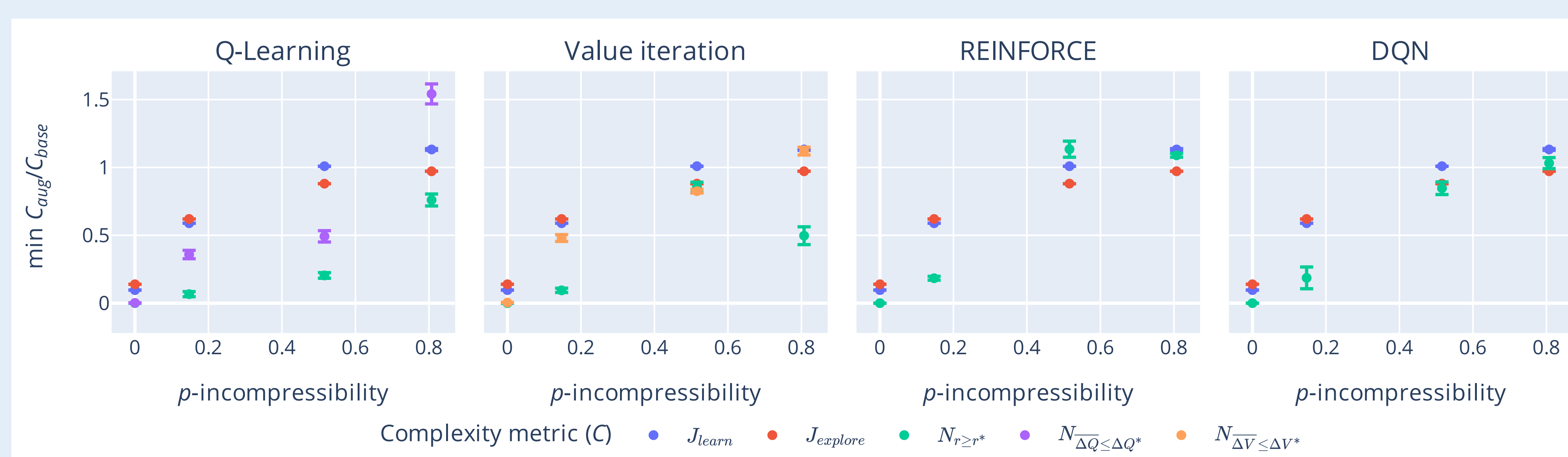
Corollary 5.3. *Macroactions benefit **gathering experience** less when **p-incompressibility** is high.*

$$\frac{J_{\text{explore}}(\mathcal{M}_+;p,\delta)}{J_{\text{explore}}(\mathcal{M}_0;p,\delta)} \geq \mathbf{IC}(\mathcal{M}_0;p,p,\delta)$$

if $\mathcal{M}_0$ is solution-separable.

D

Experimental results: Environments with higher ***p-incompressibility*** (x-axis) see higher **sample complexity** of hierarchical RL with macroactions relative to vanilla RL.



Modeling incompressibility of solutions

Most general form – see center panel B. *But what exactly is P? It depends – see paper.*

Relationship to MDL skill learning objectives

- Find skills s.t. the distribution P of abstracted solutions minimizes p-incompressibility $\approx$ minimize $\mathrm{H}[P]$ = LOVE objective (Jiang et al. 2022)
- Find skills that maximize the p-incompressibility of the *abstracted* MDP $\approx$ minimize $\mathbb{E}_{s\sim p}[d_{\mathcal{M}_+}(s)]\log|A_+|$ = LEMMA objective (Li et al. 2021)

Theoretical results *(see paper for precise statements)*

- ***Unhelpfulness of skills** is lower-bounded by **p-incompressibility**.* — Theorem 4.2 (for J_{learn}) and Corollary 5.3 (for J_{explore}). See center panel C.
- ***There are environments where macroactions <u>always harm</u>***, *e.g.*, solution-separable DSMDPs where p-incompressibility is "high enough." — Corollary 4.5 (for J_{learn}); Theorems 5.6 & 5.7 (for J_{explore}).
- ***More expressive skills have more potential to be helpful.*** Suggested by Theorem 4.2, Corollary 4.4, formalized by appendix Theorem F.5 (for J_{learn}); suggested by Theorem 5.2 (for J_{explore}).
- ***Skills are better at improving exploration ($J_{explore}$) than learning from gathered experience (J_{learn}).*** — Theorem 5.4 and Corollary 5.5.

Experiments

Setup 1:

- 4 base environments
- Choose best hRL sample complexity of 31 macroaction augmentations* and compare with vanilla RL sample complexity.

* 1 learnt (LEMMA, Li et al. 2021), 5 derived from learnt, 25 random.

Results 1:

See center panel D.

Here, p-incompressibility is $\sup_{0<\varepsilon<1} \mathbf{IC}(\mathcal{M}_0;p,p,\varepsilon)$.

Setup 2:

- 4 base environments
- Learn LOVE (Jiang et al. 2022) neural options and compare hRL sample complexity with vanilla RL sample complexity.

Results 2: See table below.

Environment	N_+/N_0	$\mathrm{IC}(\mathcal{M}_0;p)$
CliffWalking	0.000007 ± 0.000007	0.0000
CompILE2	0.00023 ± 0.00011	0.1475
8Puzzle	0.64 ± 0.19	0.5157
RubiksCube222	0.73 ± 0.17	0.8072

Conclusion and implications

- First theoretical characterization of when and how RL skills benefit sample complexity.
  - Developed RL difficulty metrics and related their improvement by skills to incompressibility of successful trajectories.
  - There are environments where unexpressive skills like macroactions provably increase RL difficulty.
- We hope our insights will guide research in automatic skill discovery and help RL practitioners better decide when and how to use skills.