MODULAR MULTITASK RL w/ POLICY SKETCHES

If we give RL agents a description of high-level actions to take in order to compute a complex task, w/o specifying the low-level actions that comprise the high level actions, can the agent learn those low-level associations?

GIVEN $\mathcal{H}$, the set of high level tasks.

Provide agent $\aleph$ (aleph) with list of tasks $h \in \mathcal{H}$, can $\aleph$ learn low level actions & policy associated with every h?

Structured Deep models ← Compositional reasoning & Interaction

↓

Used for QA systems
Relational reasoning

Page 3 ¶ 3 ⟵ not sure about the task-specific training signals.

NCP
instruction-following: string ⟶ symbolic action sequences ⟶ Hard Coded actions
(string: Natural lang.)

THIS PAPER
Symbolic Action Sequences ⟶ Learned Action Policies (discovering on own)

3. LEARNING MODULAR POLICIES FROM SKETCHES.

$(\mathcal{S}, \mathcal{A}, P, \gamma)$

- $\mathcal{S}$: set of states
- $\mathcal{A}$: set of LOWLEVEL actions.
- $P: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ transition prob. dist.
- γ : discount factor

} standard inf. horizon discounted MDP

task $\tau \in T$ specified by (R_τ, ρ_τ):

$R_\tau : \mathcal{S} \to \mathbb{R}$ (task specific reward function)

$\rho_\tau : \mathcal{S} \to \mathbb{R}$ initial distribution over states.

$\{(s_i, a_i)\}$ sequence of states & actions

$$q_i = \sum_{j=i+1}^{\infty} \gamma^{j-i-1} R(s_j)$$ empirical return.

ANNOTATE TASKS $\tau \in T$ with sketches K_τ: (task specific.)

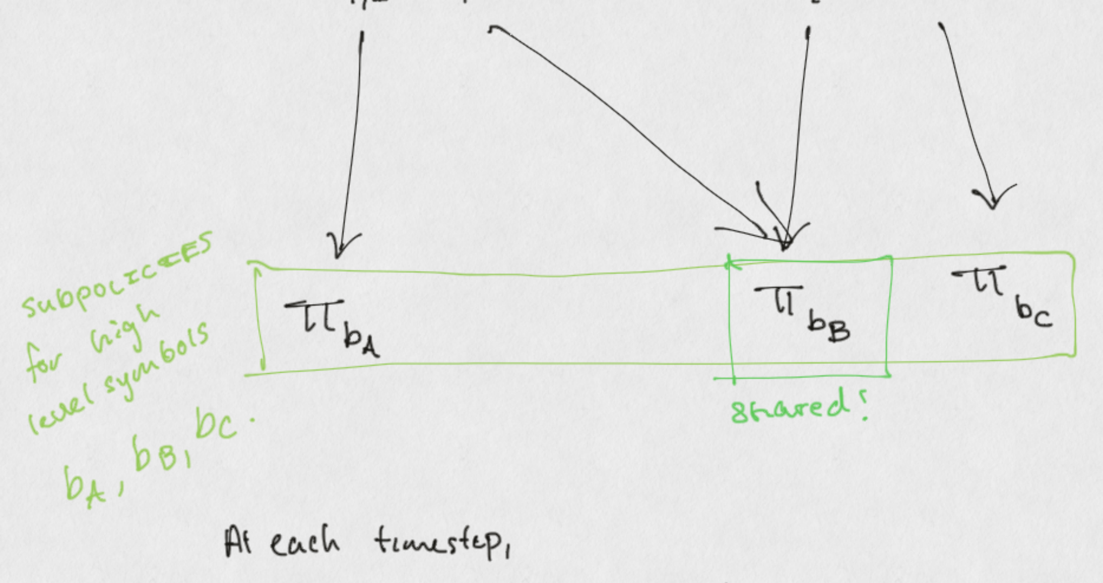
$$K_\tau = (b_{\tau 1}, b_{\tau 2}, \dots)$$

$b_\tau \in \mathcal{B}$ fixed vocab of symbolic labels

3.1 MODEL

construct for each symbol b, subpolicy π_b.

$\tau_1 = (b_{\tau_1 1}, b_{\tau_1 2} \cdots)$ $\quad \tau_2 = (b_{\tau_2 1}, b_{\tau_2 2} \cdots)$



At each timestep,

π_b can choose $a \in \mathcal{A}$ or STOP

$\Rightarrow \mathcal{A}^+ \doteq \mathcal{A} \cup \{STOP\}$.

$\pi_b : \mathcal{S} \to \mathcal{A}^+$ any representation works. ⟶ π is NN in this PAPER

$K_\tau = (b_1, b_2, \dots) \longrightarrow \Pi_\tau = \text{CONCAT}(\pi_{b_1}, \pi_{b_2} \cdots)$

maintain subpolicy index i
when π_{b_i} emits STOP, inc. i.

Π_τ is a MARKOV CHAIN over $\mathcal{S} \times \mathcal{B}$.

$(s, b_i) \to (s', b_i)$ with pr. $\sum_{a \in \mathcal{A}} \pi_{b_i}(a|s) \cdot P\{s'|s,a\}$ (pr. of action · pr. of s' given action; within subpolicy)

$\to (s, b_{i+1})$ with pr $\pi_{b_i}(STOP|s)$ (transition between subpolicies.)

$\Pi \doteq \bigcup_\tau \{\Pi_\tau\}$, π_b is arbitrary func. $\mathcal{S} \to \mathcal{A}$ param. by θ_b.

optimize over all θ_b

$$J(\Pi) \doteq \sum_\tau J(\Pi_\tau) \doteq \sum_\tau \mathbb{E}_{s_i \sim \Pi_\tau}\left[\sum_i \gamma^i R_\tau(s_i)\right]$$

maximize expected discounted reward

3.2 POLICY OPTIMIZATION

$$\nabla_\theta J(\pi) = \sum_i \left(\nabla_\theta \log \pi(a_i|s_i)\right)\left(q_i - c(s_i)\right)$$

($c(s_i)$: critic's score)

one critic per TASK τ

subpolicies partake in multiple tasks w/ different reward functions
⟹ no well defined subpolicy STATE-VALUE function.

3.3 CURRICULUM LEARNING

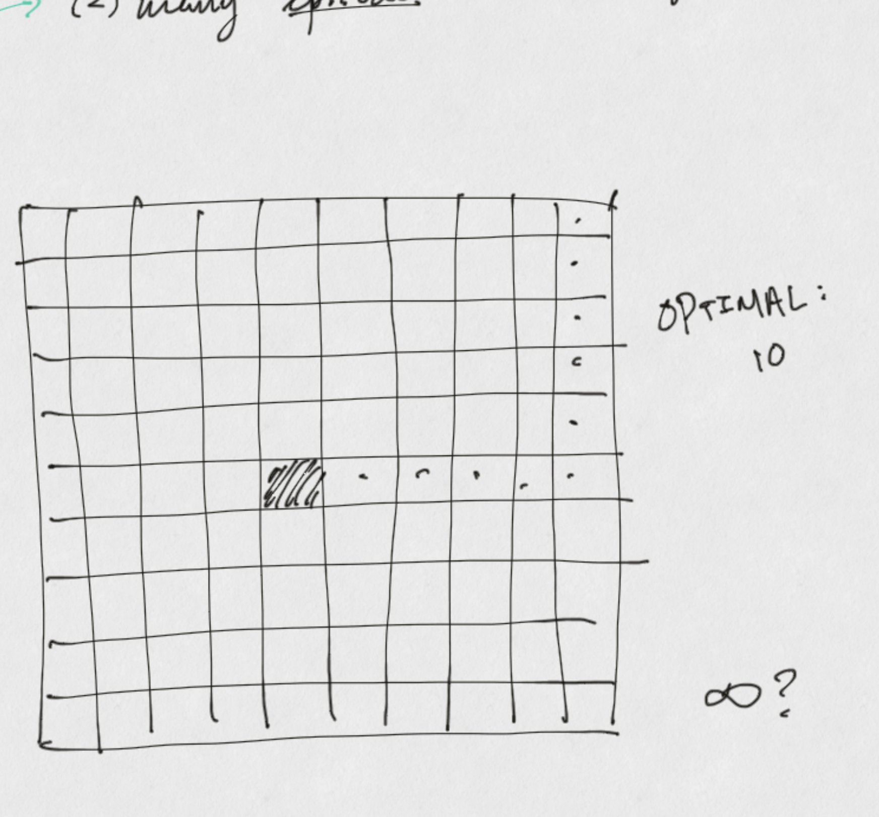
Complex tasks — impossible to earn ANY positive reward w/o already establishing some subpolicies.

Curriculum learning balances "EASY" tasks w/ hard ones w/o allowing subpolicies to overfit to the easy tasks.

STRONGEST LEARNING SIGNAL:

(1) on avg, agent doesn't achieve near upper-bound. ← don't waste time on almost solved tasks.

(2) many episodes result in high reward ← don't work on tasks until there is chance of success



OPTIMAL: 10

∞?

$$\mathbb{E}[H_{12}] = \frac{1}{2}\mathbb{E}[H_{11}] + \frac{1}{2}\mathbb{E}[H_{13}]$$

$H_{19} = H_{18}$

$H_{18} = \frac{1}{2}H_{19} + \frac{1}{2}H_{17}$

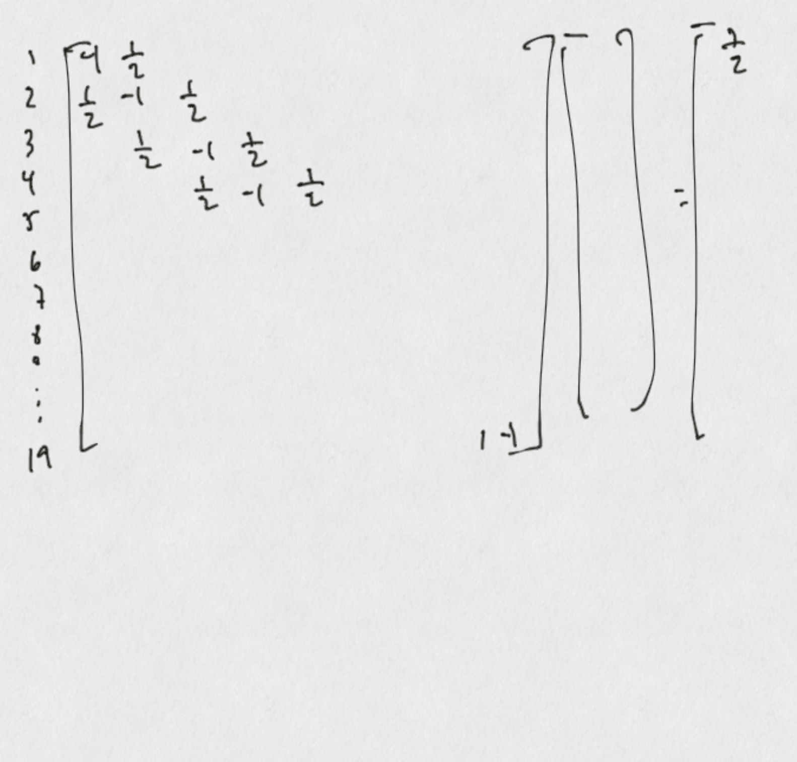
$H_{17} = \frac{1}{2}H_{18} + \frac{1}{2}H_{16}$

$\vdots$

$H_3 = \frac{1}{2}H_4 + \frac{1}{2}H_2$

$H_2 = \frac{1}{2}H_3 + \frac{1}{2}H_1$

$H_1 = \frac{1}{2} + \frac{1}{2}H_2$



$H_2 = \frac{1}{2}H_1 + \frac{1}{2}$ ($H_1 = H_2 + 1$)

$H_1 = H_2$

$H_2 = \frac{1}{2}(H_2 + 1) + \frac{1}{2}$

$\frac{1}{2}H_2 = 1$

$H_2 = 2$

~ 400?

$H_4 = 1 + \frac{1}{2}H_3$

$H_3 = \frac{1}{2}H_2 + \frac{1}{2}H_4 + 1$

$= \frac{1}{2}H_2 + \frac{1}{2} + \frac{1}{4}H_3 + 1 \Rightarrow H_3 = \frac{4}{3}\left(\frac{1}{2}H_2 + 2\right)$

$H_2 = \frac{1}{2}H_1 + \frac{1}{2}H_3 + 1 = \frac{1}{2}H_1 + \frac{2}{3}\left(\frac{1}{2}H_2 + 2\right) + 1$

$H_1 = 1 + H_2 \quad \Longrightarrow H_2 = \frac{3}{2}\left(\frac{1}{2}H_1 + \frac{7}{3}\right)$

$H_1 = 1 + \frac{3}{2}\left(\frac{1}{2}H_1 + \frac{7}{3}\right)$

$\frac{1}{4}H_1 = 1 + \frac{7}{2} = \frac{9}{2}$

$H_1 = 18$

↑ looks like it grows $\Theta(n^2)$?