

RL²:

Fast Reinforcement Learning via Slow Reinforcement Learning

Presenter: Leo

10/12/2023



Goal

1. Faster learning: **sample efficient + better performing**
2. Faster learning on a **new** game / task by utilizing learned skills: **generalization**

So, what are current limitations?

Motivation: missing prior

Bayesian reinforcement learning:

- ✓ enforce prior
- ✗ **intractable** computation of the Bayesian update

Bayesian + domain-specific ideas — e.g. PILCO (Deisenroth & Rasmussen, 2011)

- ✓ enforce prior
- ✓ tractable to (slightly) more complex problem
- ✗ Make assumptions about the environment
- ✗ Not scalable to high-dimensional setting

Related Work

- Automatic tuning of hyper-parameters
 - learning rate and temperature
 - Shin Ishii, Wako Yoshida, and Junichiro Yoshimoto. Control of exploitation-exploration meta-parameter in reinforcement learning. *Neural networks*, 2002.
 - Nicolas Schweighofer and Kenji Doya. Meta-learning in reinforcement learning. *Neural Networks*, 16(1):5–9, 2003
- Hierarchical Bayesian methods
 - Aaron Wilson, Alan Fern, Soumya Ray, and Prasad Tadepalli. Multi-task reinforcement learning: a hierarchical bayesian approach. In *ICML*, 2007.
- Hierarchical Reinforcement Learning
 - Satinder Pal Singh. Transfer of learning by composing solutions of elemental sequential tasks. *Machine Learning*, 1992.
 - Theodore J Perkins, Doina Precup, et al. Using options for knowledge transfer in reinforcement learning. *Tech. Report*, 1999.

Related Work

- Broader context of Machine Learning
 - one-shot learning for object classification
 - Ricardo Vilalta and Youssef Drissi. A perspective view and survey of meta-learning. *Artificial Intelligence Review*, 2002.
 - Li Fei-Fei, [...], Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 2006.
 - **meta-learning as an optimization problem**
 - Adam Santoro, [...], and Timothy Lillicrap. Oneshot learning with memory-augmented neural networks. *arXiv:1605.06065*, 2016.
 - Oriol Vinyals, [...] Daan Wierstra. Matching networks for one shot learning. *arXiv:1606.04080*, 2016.
 - meta-learning over the optimization process
 - Sepp Hochreiter, [...], and Peter R Conwell. Learning to learn using gradient descent. In *ICANN*, Springer, 2001
 - Marcin Andrychowicz, [...], Nando de Freitas. Learning to learn by gradient descent by gradient descent. *arXiv:1606.04474*, 2016.

Related Work

- Multi-task / Transfer Learning
 - Andrei A Rusu, [...], Raia Hadsell. Policy distillation. ICLR 2016
 - Andrei A Rusu, [...], Raia Hadsell. Progressive neural networks. arXiv:1606.04671
 - Andrei A Rusu, [...], Raia Hadsell. Sim-to-real robot learning from pixels with progressive nets. ICML 2017

Previous Problem Setting:

\mathcal{S} : state set **Given**

\mathcal{A} : action set

\mathcal{P} : transition probability distribution

r : reward function

ρ_0 : initial state distribution

γ : discount factor

T : time horizon

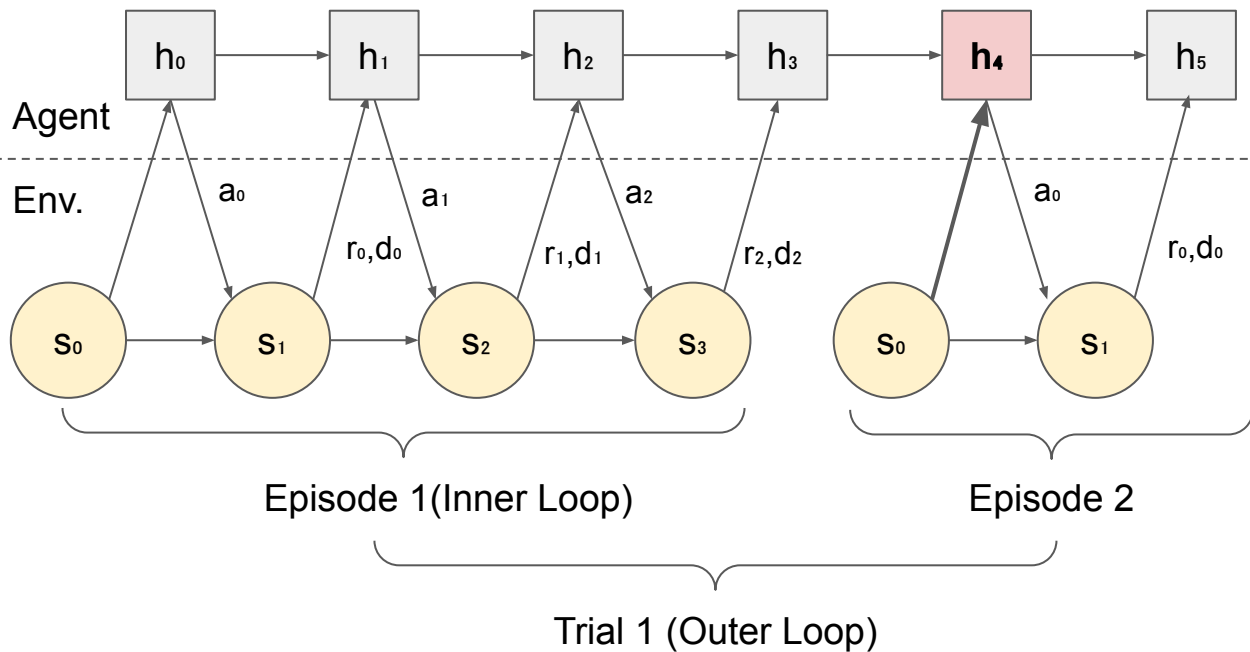
$$a_t \sim \pi_\theta(a_t | s_t)$$

$$\eta(\pi_\theta) = \mathbb{E}_\tau \left[\sum_{t=0}^T \gamma^t r(s_t, a_t) \right]$$

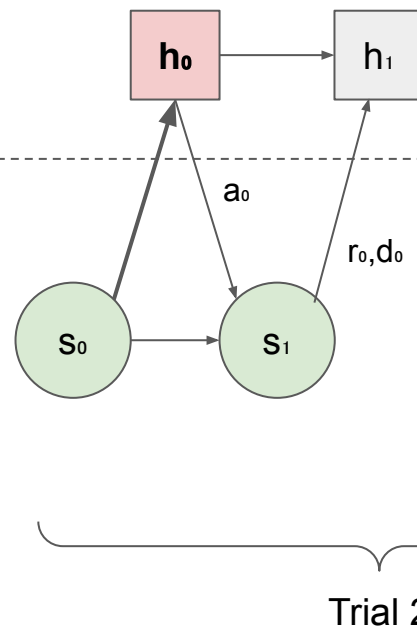
Expected Discounted Return

Proposed Formulation

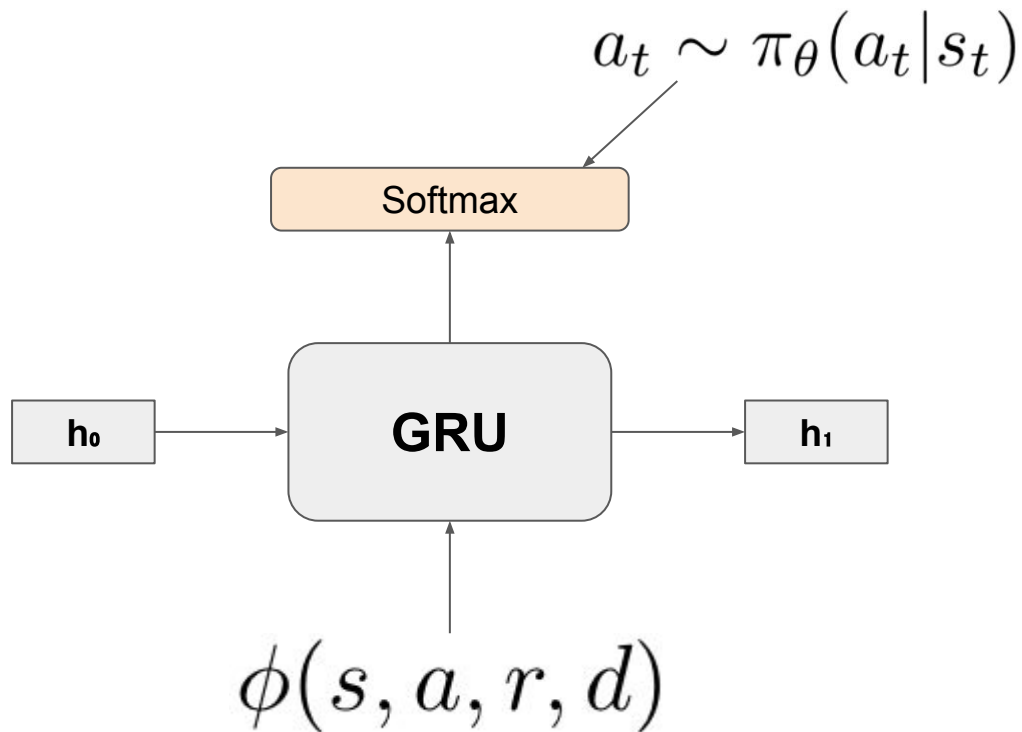
MDP 1



MDP 2



Policy Representation



Policy Optimization

Objective: $\mathbb{E}_{\tau} [\sum_{t=0}^T \gamma^t r(s_t, a_t)] + \mathbb{E}_{\tau} [\sum_{t=0}^T \gamma^t r(s_t, a_t)] + \dots$

Episode 1 Episode 2

Optimization: TRPO + GAE

Policy Optimization: TRPO + GAE

$$\delta_t^V = r_t + \gamma V(s_{t+1}) - V(s_t)$$

Initialize policy parameter θ_0 and value function parameter ϕ_0 .

for $i = 0, 1, 2, \dots$ **do**

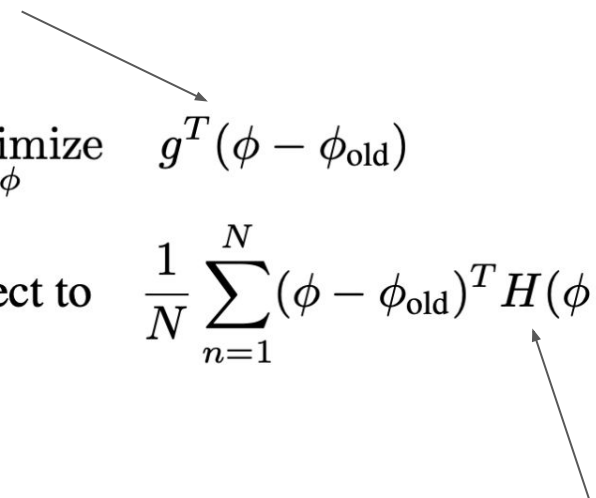
end for

Policy Optimization: TRPO + GAE (Eq. 30)

Gradient of $\underset{\phi}{\text{minimize}} \sum_{n=1}^N \|V_{\phi}(s_n) - \hat{V}_n\|^2$

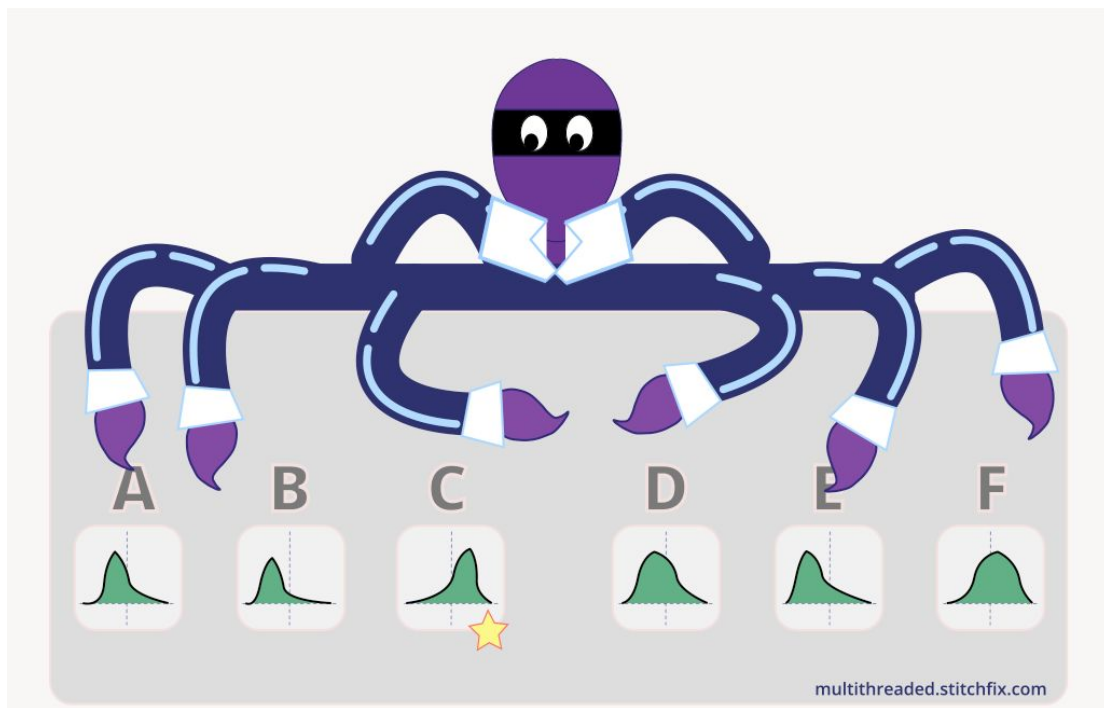
$\underset{\phi}{\text{minimize}} \quad g^T(\phi - \phi_{\text{old}})$

subject to $\frac{1}{N} \sum_{n=1}^N (\phi - \phi_{\text{old}})^T H(\phi - \phi_{\text{old}}) \leq \epsilon.$ (30)



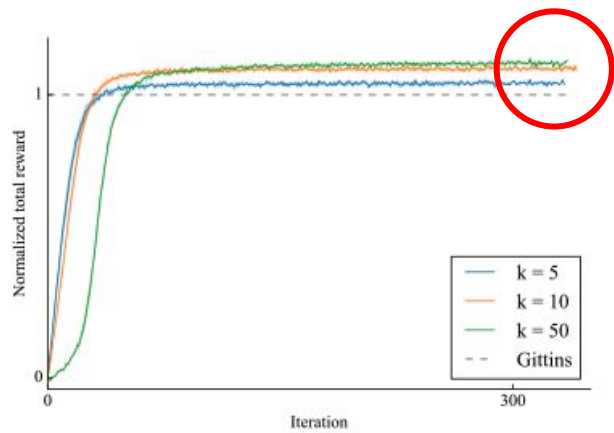
$$H = \frac{1}{N} \sum_n j_n j_n^T, \text{ where } j_n = \nabla_{\phi} V_{\phi}(s_n)$$

Multi-Arm Bandit

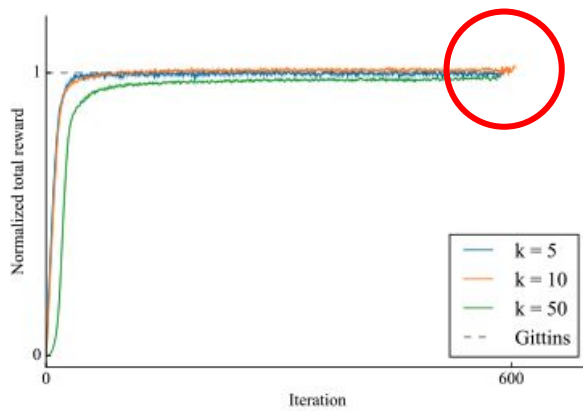


$\text{reward}_i \sim \text{Bernoulli}(p_i)$
 $p_i \sim \text{Uniform}[0, 1]$

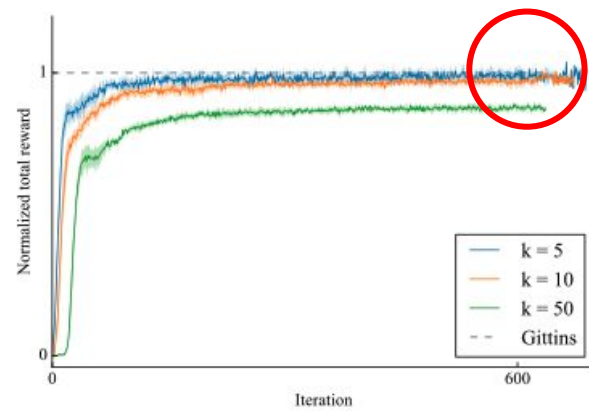
Result



(a) $n = 10$



(b) $n = 100$



(c) $n = 500$

$k \in \{5, 10, 50\}$ bandits and $n \in \{10, 100, 500\}$ episodes

✓ policy learns to trade off between exploration and exploitation.

Tabular MDP

$$|\mathcal{S}| = 10$$

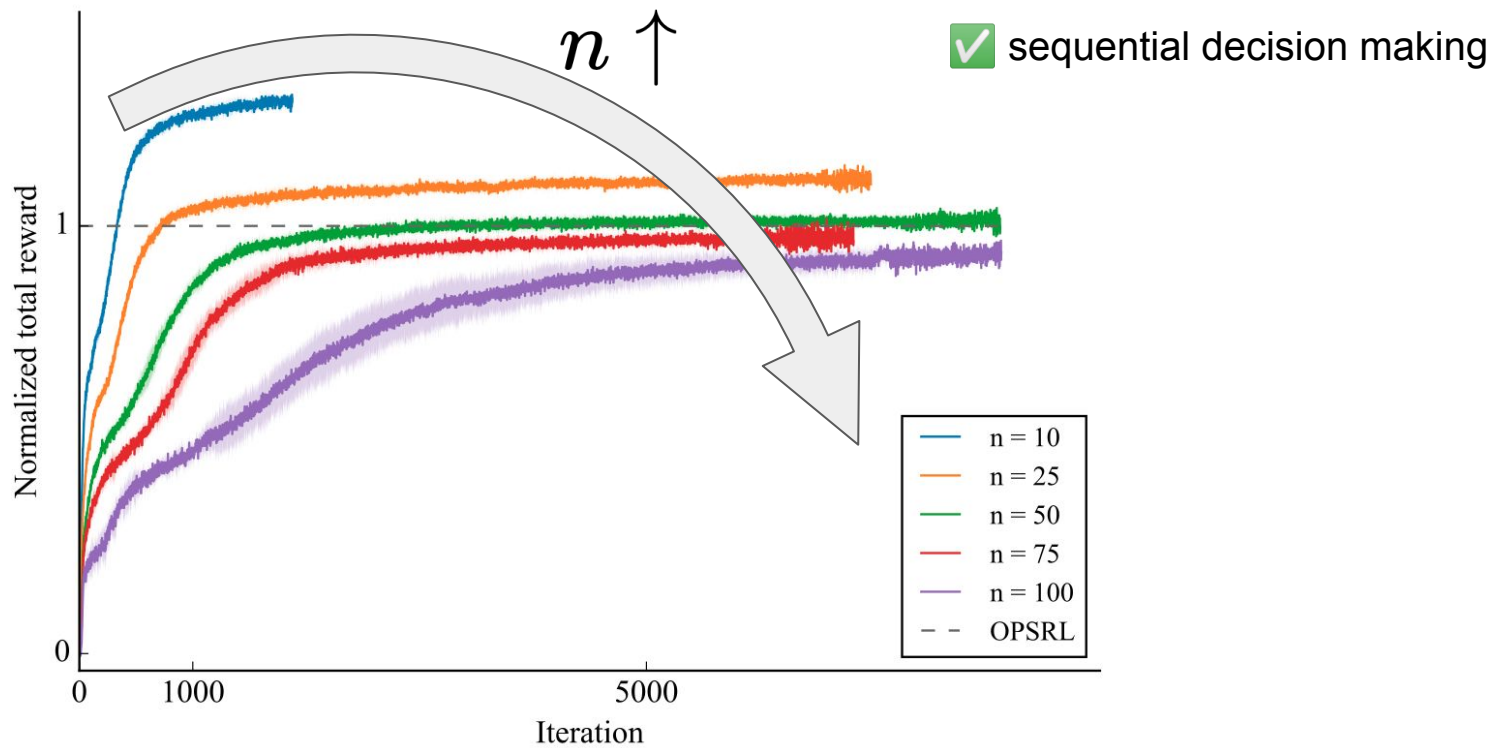
$$|\mathcal{A}| = 5$$

$$r(s, a) \sim \mathcal{N}(\mu, 1), \mu \sim \mathcal{N}(1, 1)$$

$$P(s' | s, a) \sim \text{Dir}(\boldsymbol{\alpha})$$

$$T = 10$$

Results



Discussion of Results

1-2 slides

What conclusions are drawn from the results by the authors?

- ❖ What insights are gained from the experiments?
- ❖ What strengths and weaknesses of the proposed method are illustrated by the results?

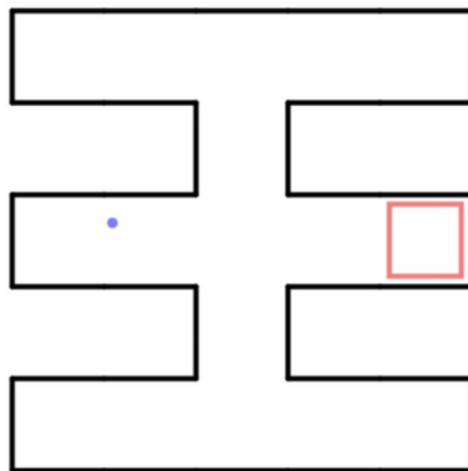
Are the stated conclusions fully backed by the results and references?

- ❖ If so, why? (Recap the relevant supporting evidences from the given results + refs)
- ❖ If not, what are the additional experiments / comparisons that can further support/repudiate the conclusions of the paper?

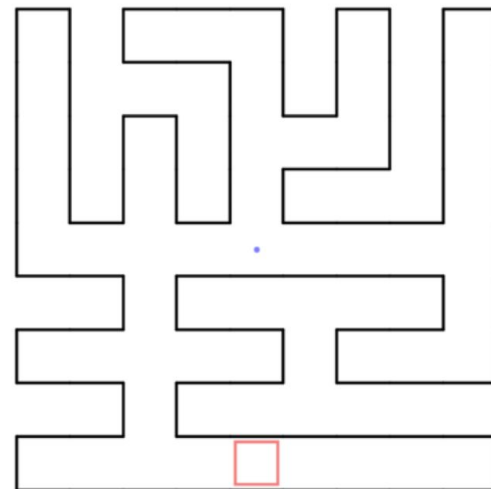
Visual Navigation



(a) Sample observation



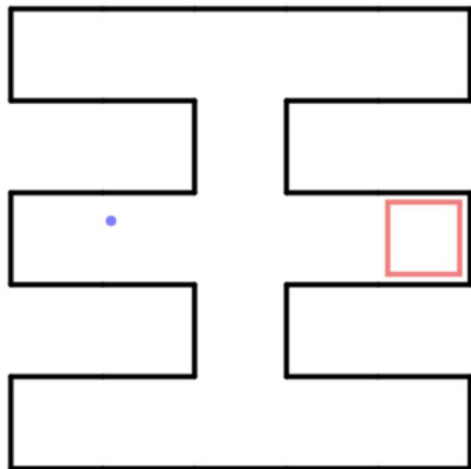
(b) Layout of the 5×5 maze in (a)



(c) Layout of a 9×9 maze

Figure 4: Visual navigation. The target block is shown in red, and occupies an entire grid in the maze layout.

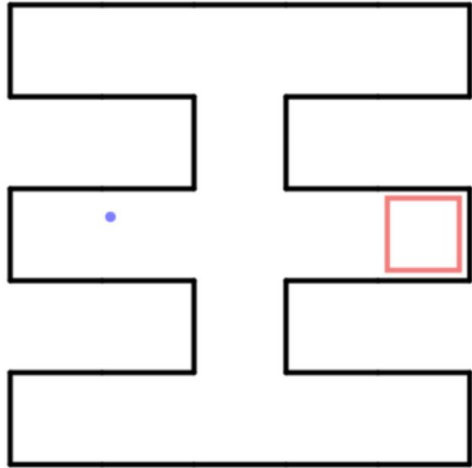
Visual Navigation



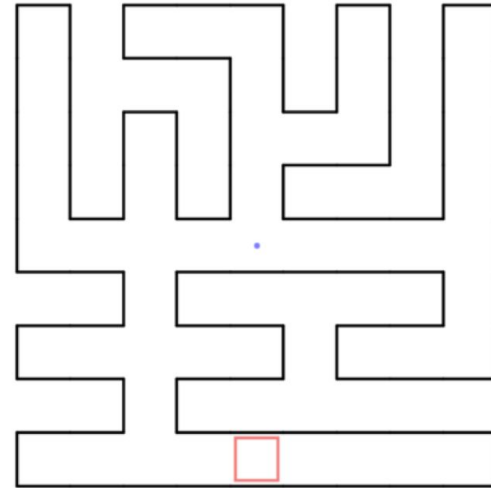
(b) Layout of the 5×5 maze in (a)

Train: 2 episode, up to 250 step;
MDP = 1 maze out of 1000

Test



(b) Layout of the 5×5 maze in (a)



(c) Layout of a 9×9 maze

Result

(a) Average length of successful trajectories

Episode	Small	Large
1	52.4 ± 1.3	180.1 ± 6.0
2	39.1 ± 0.9	151.8 ± 5.9
3	42.6 ± 1.0	169.3 ± 6.3
4	43.5 ± 1.1	162.3 ± 6.4
5	43.9 ± 1.1	169.3 ± 6.5

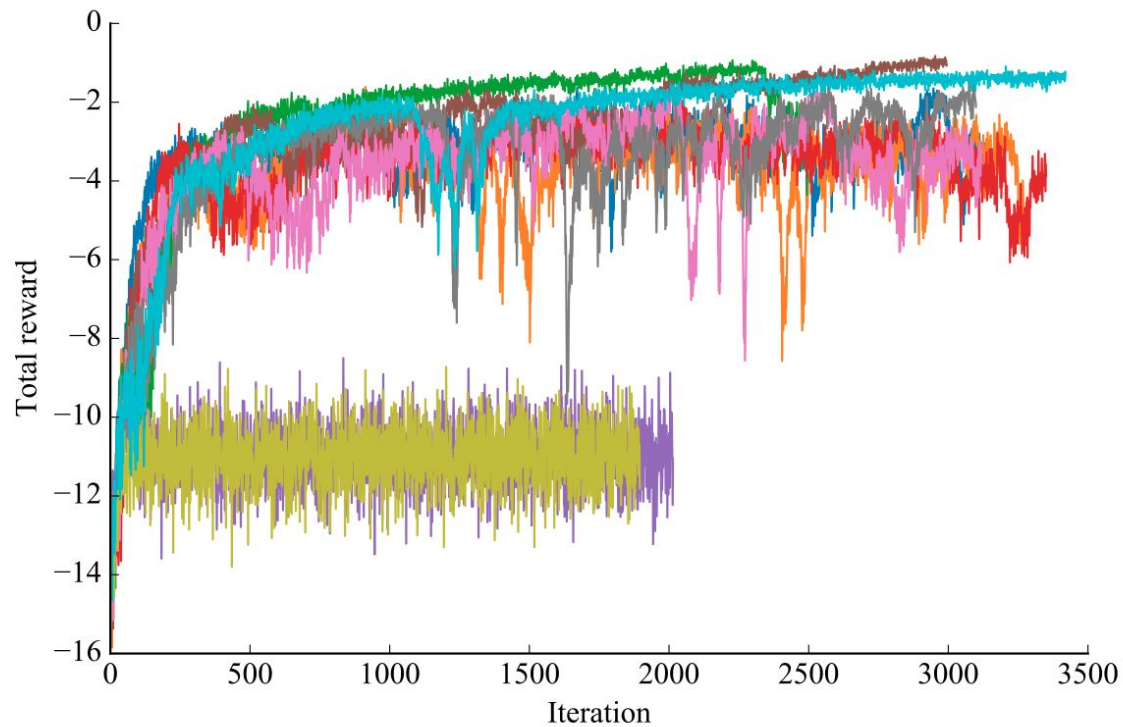
(b) %Success

Episode	Small	Large
1	99.3%	97.1%
2	99.6%	96.7%
3	99.7%	95.8%
4	99.4%	95.6%
5	99.6%	96.1%

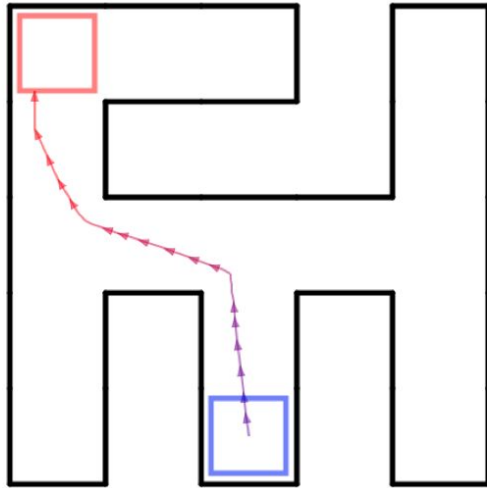
(c) %Improved

Small	Large
91.7%	71.4%

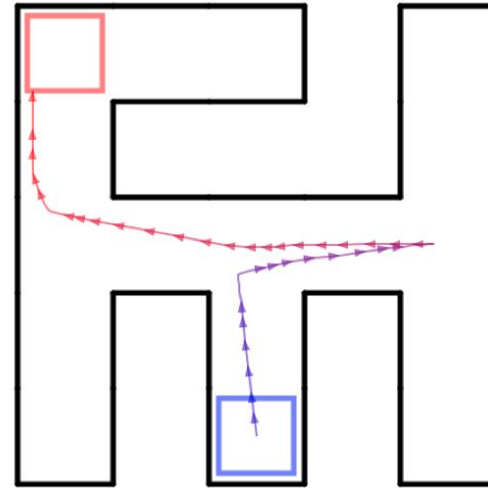
Sensitive to random initialization



Error Analysis



(c) Bad behavior, 1st episode



(d) Bad behavior, 2nd episode

Discussion of Results

1-2 slides

What conclusions are drawn from the results by the authors?

- ❖ What insights are gained from the experiments?
- ❖ What strengths and weaknesses of the proposed method are illustrated by the results?

Are the stated conclusions fully backed by the results and references?

- ❖ If so, why? (Recap the relevant supporting evidences from the given results + refs)
- ❖ If not, what are the additional experiments / comparisons that can further support/repudiate the conclusions of the paper?

Critique / Limitations / Open Issues

1-2 slides

What are the key limitations of the proposed approach / ideas? (e.g. does it require strong assumptions that are unlikely to be practical? Computationally expensive? Require a lot of data?)

Are there any practical challenges in deploying the approach on physical robots in the real world? Are there any safety or ethical concerns of using such approach?

If follow-up work has addressed some of these limitations, include pointers to that. But don't limit your discussion only to the problems / limitations that have already been addressed.

Limitations

- Likely to be **impractical** in robotics
 - Sign: Util now, no one uses this method in robotics
- Motivation looks good, but **massive simplification** is made:
 - For robotics tasks, representation – detection, segmentation, is still a problem
 - In robotics setting, MDP is more complex or multiple MDP comes into play randomly

Question for Future Work

What interesting questions does it raise for future work?

- ❖ Does this algorithm scale to more realistic / complex scenario (e.g. Robotics)? Why not?
- ❖

Extended Readings

Most similar:

Learning to reinforcement learn. Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. *arXiv:1611.05763*, 2016.

Most popular:

Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks, *Chelsea Finn, Pieter Abbeel, Sergey Levine*
Proceedings of the 34th International Conference on Machine Learning, PMLR 70:1126-1135, 2017.

Most recent:

RL3: Boosting Meta Reinforcement Learning via RL inside RL2, Abhinav Bhatia, Samer B. Nashed, Shlomo Zilberstein.
arXiv:2306.15909v1

Summary

- ❖ Problem:
 - New framing to adapt existing RL algorithm for fast learning new task
- ❖ Why hard?
 - Agent does not have the primitives for efficient exploration at the beginning of training
- ❖ What is the key limitation of prior work?
 - Not scalable, prior assumption
- ❖ What is the key insight(s) of the proposed work
 - This inner-outer-loop formulation and modeling choice works at least in the simplest case
 - RL techniques seems to be a bottleneck for “meta-learning” new tasks



Thanks!