A Meta-MDP Approach to Improve Exploration in Reinforcement Learning

Francisco M. Garcia, Philip S. Thomas

University of Massachusetts - Amherst

Problem

- Explorations techniques are crucial for an agent to be able to solve novel complex problems.
- Many algorithms rely on exploration methods based on the task the agent is **currently** trying to solve, ignoring the possibility of **previous experience** with **related tasks.**
- Previous experience with a set of similar tasks can be **leveraged** to guide an agent on how to best explore when solving new related tasks

Our Approach

- We propose a *meta-learning* approach where one agent, called the **advisor**, learns a policy to guide other agents on how to explore.
- We separate the agent's behavior into two policies: **exploration** and **exploitation.** Assume ε-greedy exploration schedule.
- Advisor maintains a general **exploration** policy, μ, for all related tasks.
- Agent maintains a task-specific **exploitation** policy, π, for each new task.

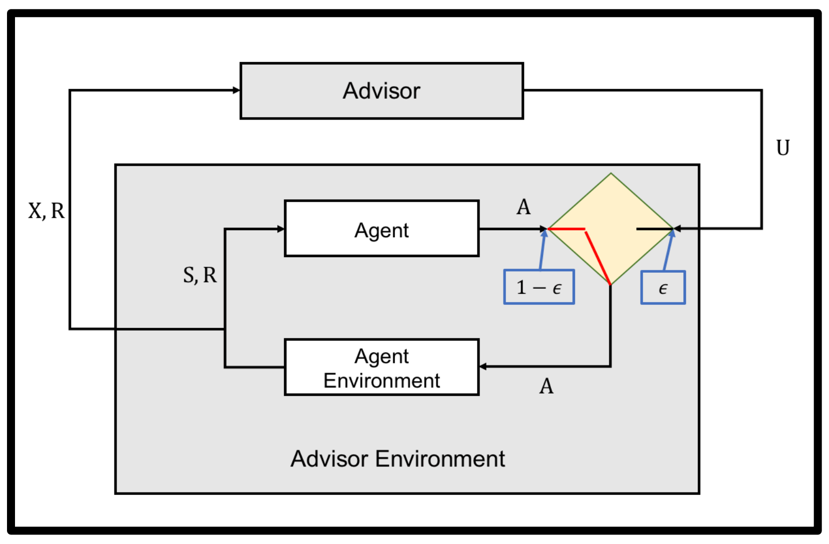


Diagram depicting interaction between advisor and agent. Advisor's policy suggests action U and agent's policy suggest action A.

If agent explores it executes action U otherwise it executes A

- **The Performance** of exploration policy μ in task c is given by sum of returns:

$$\rho(\mu, c) = \mathbf{E}\left[\sum_{i=0}^{I}\sum_{t=0}^{T} R_t^i \middle| \mu, c\right]$$

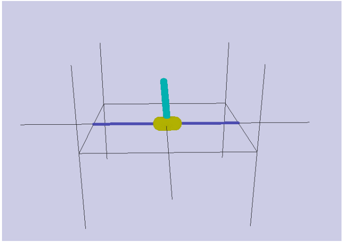
- The **objective** learn an optimal exploration policy that maximizes expected performance over tasks, define as:

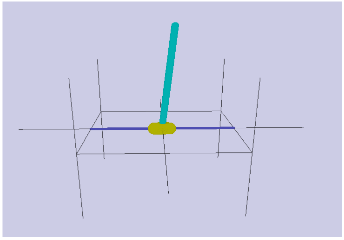
$$\mu^* \in \arg\max_{\mu} \mathbf{E}\left[\rho(\mu, C)\right]$$

- Learn exploration policy with standard RL techniques.

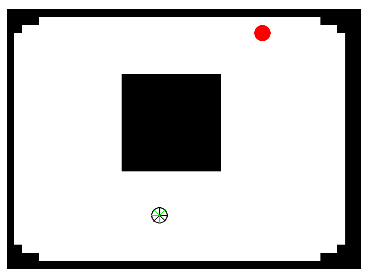
Experiments

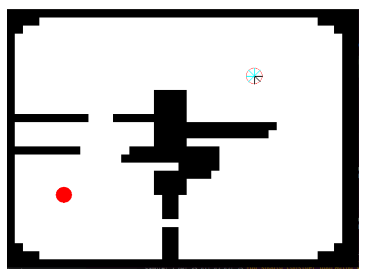
- **Discrete Action Space:**
  - **Cartpole:** task variations correspond to poles of different length and mass.





  - **Animat [1]:** task variations correspond to different mazes and goal locations.

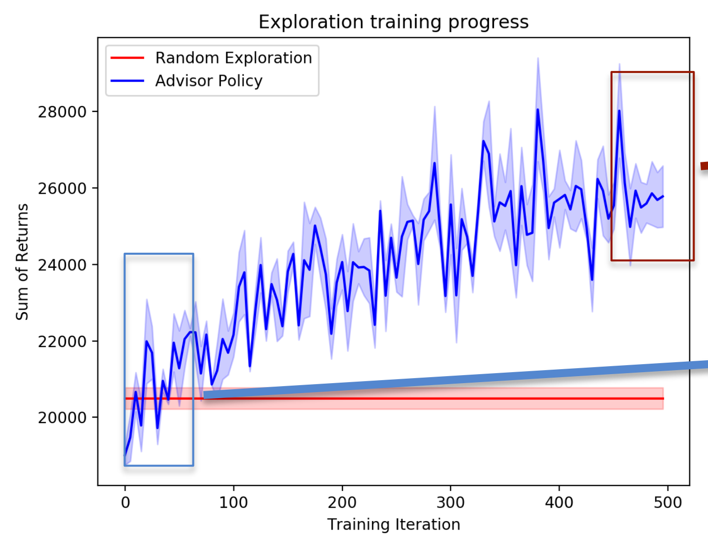




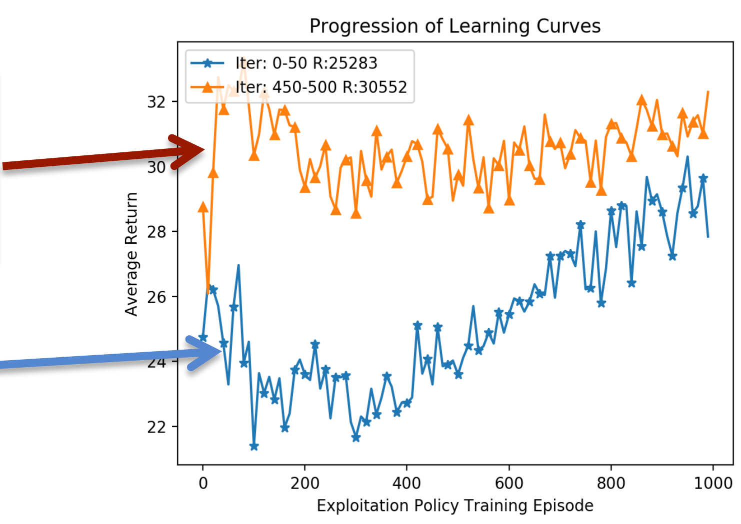
- **Continuous Action Space:**
  - **Cartpole:** continuous version of cartpole
  - **Hopper:** task variations correspond to agents of different sizes
  - **Ant:** task variations correspond to agents of different sizes

Results

Cumulative return (cartpole)
Each iteration corresponds to an agent lifetime



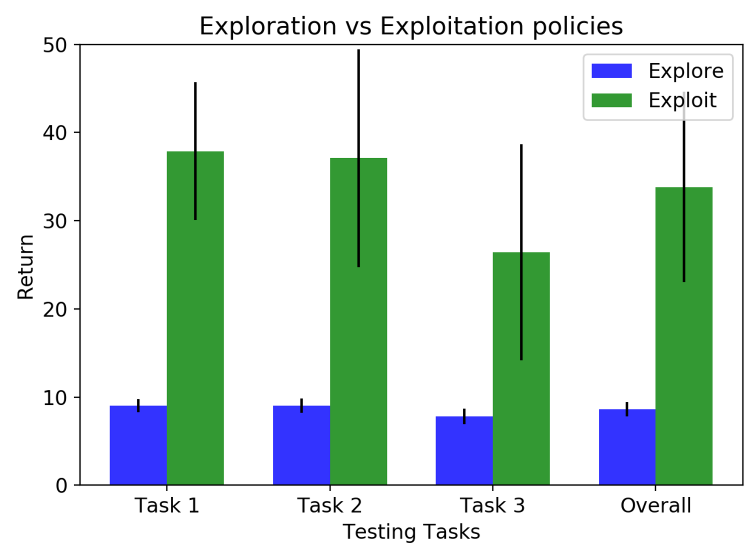
Average learning curve (cartpole)
Average over first 50 iterations (blue)
Average over last 50 iterations(orange)



- **Advisor** improves agent's overall performance through exploration (left).
- **Agent** is able to learn faster in novel tasks (right).

Is an exploration policy simply a general exploitation policy?

- We compared the performance of the learned **exploration policy** (blue) with the **task-specific policy** (green) on novel problems.
- **Exploration policy** fails to find a good solution to any problem, indicating it is **not** simply an **exploitation policy**.



Comparison of proposed approach to MAML [2] in benchmark problems

Problem Class	R	R+Advisor	PPO	PPO+Advisor	MAML
Pole-balance (d)	20.32 ± 3.15	28.52 ± 7.6	27.87 ± 6.17	$\mathbf{46.29 \pm 6.30}$	39.29 ± 5.74
Animat	-779.62 ± 110.28	$\mathbf{-387.27 \pm 162.33}$	-751.40 ± 68.73	-631.97 ± 155.5	-669.93 ± 92.32
Pole-balance (c)	—	—	29.95 ± 7.90	$\mathbf{438.13 \pm 35.54}$	267.76 ± 163.05
Hopper	—	—	13.82 ± 10.53	$\mathbf{164.43 \pm 48.54}$	39.41 ± 7.95
Ant	—	—	-42.75 ± 24.35	83.76 ± 20.41	$\mathbf{113.33 \pm 64.48}$

Table 1. Average performance (and standard deviations) on discrete and continuous control unseen tasks over the last 50 episodes.

Conclusion

- We show that **experience** with similar tasks can be use to adapt a policy specifically for exploration.
- The problem of learning an exploration policy can be modeled as a reinforcement learning problem itself.
- A key feature needed for this approach to work is that related problems **must provide** some structure which can be exploited.
- There is a clear direction for future work. At present, we are able to learn **how** an agent should behave when exploring, but we are ignoring **when** an agent should explore. This is also a crucial component for intelligent behavior.

References

- [1] Thomas, P., and Barto, A. Conjugate Markov Decision Processes. *Proceedings of the 28th International Conference on Machine Learning.*
- [2] Finn, C., Abeel, P., and Levine S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *Proceedings of the 34th International Conference on Machine Learning.*