A Simple Reward-Free Approach to Constrained Reinforcement Learning

International Conference on Machine Learning (ICML) 2022

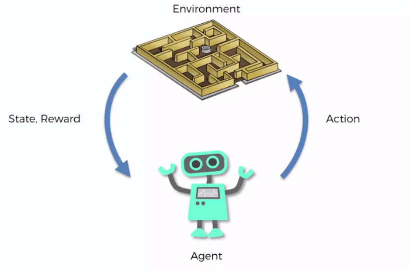
Sobhan Miryoosefi, Chi Jin

July 2022

PRINCETON UNIVERSITY

Standard RL

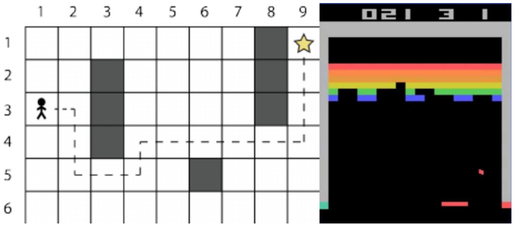
Agent interactively takes some action in the Environment and receives some scalar reward for the action taken.



Goal: find policy that maximizes the cumulative scalar reward

Reward Models Desired Behavior

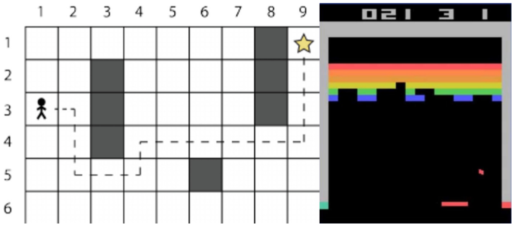
Reward Models Desired Behavior



Desired behavior is simple

- Agent needs to reach the "goal" as "quickly" as possibly (e.g., gridworld)

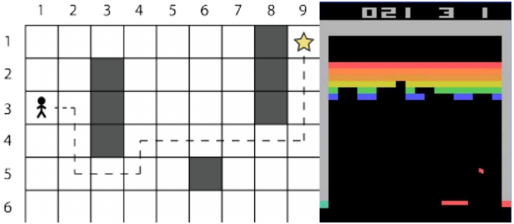
Reward Models Desired Behavior



Desired behavior is simple

- Agent needs to reach the "goal" as "quickly" as possibly (e.g., gridworld)
- Agent playing a game where score is explicitly given (e.g., Atari game)

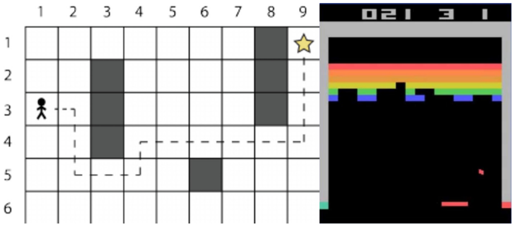
Reward Models Desired Behavior



Desired behavior is simple

- Agent needs to reach the "goal" as "quickly" as possibly (e.g., gridworld)
- Agent playing a game where score is explicitly given (e.g., Atari game)
- . . .

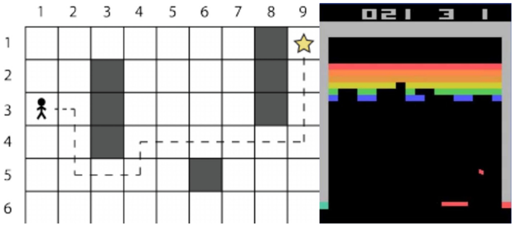
Reward Models Desired Behavior



Desired behavior is simple

- Agent needs to reach the "goal" as "quickly" as possibly (e.g., gridworld)
- Agent playing a game where score is explicitly given (e.g., Atari game)
- ...

Reward Models Desired Behavior



Desired behavior is simple

- Agent needs to reach the "goal" as "quickly" as possibly (e.g., gridworld)
- Agent playing a game where score is explicitly given (e.g., Atari game)
- . . .

Strategy: applying standard RL using positive and negative **scalar rewards**.

Complex Desired Behavior

Desired behavior is complex and it consists of many subgoals and restrictions

Complex Desired Behavior

Desired behavior is complex and it consists of many subgoals and restrictions

Complex Desired Behavior

Desired behavior is complex and it consists of many subgoals and restrictions

- Autonomous vehicle: not only get to the destination, but should also respect safety, fuel efficiency, and human comfort.

Complex Desired Behavior

Desired behavior is complex and it consists of many subgoals and restrictions

- Autonomous vehicle: not only get to the destination, but should also respect safety, fuel efficiency, and human comfort.
- Robot should not only fulfill its task, but should also control its wear and tear (e.g., limiting torque on its motors)

Complex Desired Behavior

Desired behavior is complex and it consists of many subgoals and restrictions

- Autonomous vehicle: not only get to the destination, but should also respect safety, fuel efficiency, and human comfort.
- Robot should not only fulfill its task, but should also control its wear and tear (e.g., limiting torque on its motors)
- . . .

Approach

One Approach: applying standard RL

- Boiling down learning goal into a single scalar is challenging
- Agent might maximize the reward without satisfying our desired behavior
- Gets harder as desired behavior gets more complex

Better Approach: Constraint-Based RL

In many settings, it's more natural and easier to express some behaviors by constraints.

Our Contribution

Framework that bridges Reward-free RL and Constrained RL

- Direct translation of any progress in Reward-free RL to Constrained RL
- While being modular, provides sharp sample complexity for the tabular setting
- Providing first sample complexity results for linear setting

Model: Episodic VMDP

Episodic vector-valued Markov Decision Process (VMDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, \boldsymbol{r})$ Same as MDP model except $\boldsymbol{r}$ is d-dimensional

- vector-valued value function $\boldsymbol{V}_h^\pi$ and $\boldsymbol{Q}_h^\pi$

Scalarized MDP: For any $\boldsymbol{\theta} \in \mathbb{R}^d$, define $\mathcal{M}_{\boldsymbol{\theta}} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r_{\boldsymbol{\theta}})$ where $r_{\boldsymbol{\theta}} = \langle \boldsymbol{\theta}, \boldsymbol{r} \rangle$

- scalarized value functions $V_h^\pi(\cdot; \boldsymbol{\theta})$ and $Q^\pi(\cdot, \cdot; \boldsymbol{\theta})$
- optimal value function $V_h^\star(\cdot; \boldsymbol{\theta})$ and $Q_h^\star(\cdot, \cdot; \boldsymbol{\theta})$

Learning Objective(s)

Assume $\mathcal{C}$ is a convex and compact set in $\mathbb{R}^d$.

Learning Objective(s)

Assume $\mathcal{C}$ is a convex and compact set in $\mathbb{R}^d$.

Task 3: Constrained RL

Assume utility function

$$u = \{u_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}_{h=1}^{H}$$

$$\max_{\pi} \quad \mathbb{E}_{\pi}[\sum_{h} u_h(s_h, a_h)]$$

$$\text{s.t.} \quad \boldsymbol{V}_1^{\pi}(s_1) \in \mathcal{C}$$

Learning Objective(s)

Assume $\mathcal{C}$ is a convex and compact set in $\mathbb{R}^d$.

Task 2: Approachability

$$\min_{\pi} \quad \mathrm{dist}\Big(\boldsymbol{V}_1^{\pi}(s_1), \mathcal{C}\Big)$$

Task 3: Constrained RL

Assume utility function $u = \{u_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}_{h=1}^{H}$

$$\max_{\pi} \quad \mathbb{E}_{\pi}[\sum_h u_h(s_h, a_h)]$$

$$\text{s.t.} \quad \boldsymbol{V}_1^{\pi}(s_1) \in \mathcal{C}$$

Learning Objective(s)

Assume $\mathcal{C}$ is a convex and compact set in $\mathbb{R}^d$.

Task 1: Reward-free

After exploration phase, $\{\boldsymbol{r}_h(s_h^k, a_h^k)\}_{(k,h)\in[K]\times[H]}$ are revealed

for any $\boldsymbol{\theta} \in \mathbb{R}^d$, algorithm should outputs the (near-)optimal policy $\pi_{\boldsymbol{\theta}}$ that maximizes $V_1^{\pi_{\boldsymbol{\theta}}}(s_1; \boldsymbol{\theta})$

Task 2: Approachability

$$\min_{\pi} \quad \mathrm{dist}\left(\boldsymbol{V}_1^{\pi}(s_1), \mathcal{C}\right)$$

Task 3: Constrained RL

Assume utility function $u = \{u_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}_{h=1}^H$

$$\max_{\pi} \quad \mathbb{E}_{\pi}[\sum_h u_h(s_h, a_h)]$$
$$\text{s.t.} \quad \boldsymbol{V}_1^{\pi}(s_1) \in \mathcal{C}$$

Reductions

Task 1: Reward-free

Task 2: Approachability

Task 3: Constrained RL

Reductions

Task 1: Reward-free

Task 2: Approachability $\xrightarrow{\checkmark}$ **Task 3: Constrained RL**

It's easy to show that

$$\textbf{Sample Complexity of Task 3} \leq \tilde{\mathcal{O}}\Big(\textbf{Sample Complexity of Task 2} + \underbrace{H^2 \log[d]/\epsilon^2}_{\text{usually lower order term}} \Big)$$

Reductions

Task 1: Reward-free $\xrightarrow{\checkmark}$ **Task 2: Approachability** $\xrightarrow{\checkmark}$ **Task 3: Constrained RL**

We design a Meta Algorithm that satisfies

Main Result

$$\textbf{Sample Complexity of Task 2} \leq \tilde{O}\Big(\textbf{Sample Complexity of Task 1} + \underbrace{H^2 \log[d]/\epsilon^2}_{\text{usually lower order term}} \Big)$$

Sample Complexity Guarantees

	Algorithm	T1: Reward-free	T2: Approachability	T3: CMDP
Tabular	[WBY20]	$\tilde{\mathcal{O}}(\min\{d,S\}H^4SA/\epsilon^2)$	-	-
	[BDL+20]	-	-	$\tilde{\mathcal{O}}(d^2H^3S^2A/\epsilon^2)$
	[YTZS21]	-	$\tilde{\mathcal{O}}(\min\{d,S\}H^3SA/\epsilon^2)$	$\tilde{\mathcal{O}}(\min\{d,S\}H^3SA/\epsilon^2)$
	This work	$\tilde{\mathcal{O}}(\min\{d,S\}H^4SA/\epsilon^2)$	$\tilde{\mathcal{O}}(\min\{d,S\}H^4SA/\epsilon^2)$	$\tilde{\mathcal{O}}(\min\{d,S\}H^4SA/\epsilon^2)$
Linear	This work	$\tilde{\mathcal{O}}(d_{\text{lin}}^3H^6/\epsilon^2)$	$\tilde{\mathcal{O}}(d_{\text{lin}}^3H^6/\epsilon^2)$	$\tilde{\mathcal{O}}(d_{\text{lin}}^3H^6/\epsilon^2)$

References

[BDL+20] Kianté Brantley, Miro Dudik, Thodoris Lykouris, Sobhan Miryoosefi, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings. In *Advances in Neural Information Processing Systems*, volume 33, pages 16315–16326. Curran Associates, Inc., 2020.

[WBY20] Jingfeng Wu, Vladimir Braverman, and Lin F Yang. Accommodating picky customers: Regret bound and exploration complexity for multi-objective reinforcement learning. *arXiv preprint arXiv:2011.13034*, 2020.

[YTZS21] Tiancheng Yu, Yi Tian, Jingzhao Zhang, and Suvrit Sra. Provably efficient algorithms for multi-objective competitive rl. *arXiv preprint arXiv:2102.03192*, 2021.