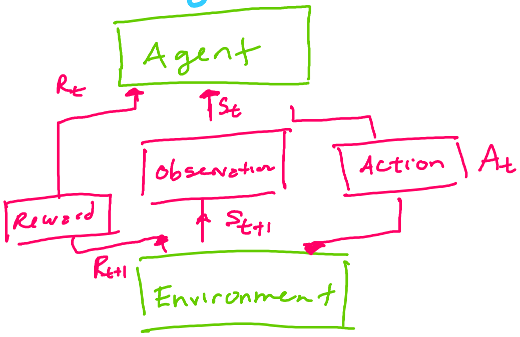
The RL Framework: The Problem

Friday, November 27, 2020 6:20 PM

3.1 The Agent-Environment Interface



At t = 0:

- Observation: Situation presented to Agent (S_t)
- Action: Response to Observation (A_t)
- One time step later: Reward is presented (R_{t+1}) along with new state, (S_{t+1})

Assumption: Agent is able to fully observe state of environment

Sequence:

t	Sequence
t = 0	S_0 A_0
t = 1	R_1 S_1 A_1
t = 2	R_2 S_2 A_2 . . .

→ Reward is most important.

Goal: ' Maximize expected culmalitive Reward

Episodic vs. Continuing Tasks

Episodic: "Well-defined Ending Point"

↳ e.g → game: win / lose
→ car : car crashes

- When end point reached:
  - Consider reward
- over many lives, agent gets better!
- Target aims to ↑ culminative reward.

Continuing: Interaction continues without limit

- $S_0 A_0, R_1, S_1, A_1, \ldots\ldots$
- More complex (e.g Stock Market)

Chess Example:

→ E.g of Action: Moving a Piece

→ E.g of State: Config of Board

→ On first game, you're winning by 5 pieces whats the reward? Ans: 0

3.2 Goals & Rewards.

- "Reward Hypothesis": Maximize expected Culminative Reward
- Rewarding is subjective to the task
  - → e.g reward in context of robot learning to walk?
    - ↳ what makes walking good?
- We want rewards to be a scientific concept!
- Scenario: Robot Walking
  - Actions: { Forces applied to joints }
  - States: { Position & Velocity of joints,
    Measurements of the ground,
    Contact Sensor Data }
  - Reward { Feedback Mechanism }

$$r = \overbrace{\min(v_x, v_{max})}^{\text{prop. to foward velocity}} - \overbrace{0.005(v_y^2 + v_z^2)}^{\text{penalize deviation from forward direction}}$$

$$- \underbrace{0.05y^2}_{\text{penalize: deviation from center of track}} - \underbrace{0.02\|u\|^2}_{\text{penalize torque}} + \underbrace{0.02}_{\text{Constant: Reward for not falling}}$$

- What are we rewarding for?

1. foward velocity: walk fast
2. forward direction: walk foward
3. torque: walk smoothly
4. Constant: walk as long as possible

- General: In general, rewarding can be as simple as +1 for win or ↑ a scoreboard

Questions:

Q1: How would you reward escaping quickly in a maze escape game

A: -1 for every step taken (Part of reward)

Q2: What reward encourages board gamers to win?

A: rewar

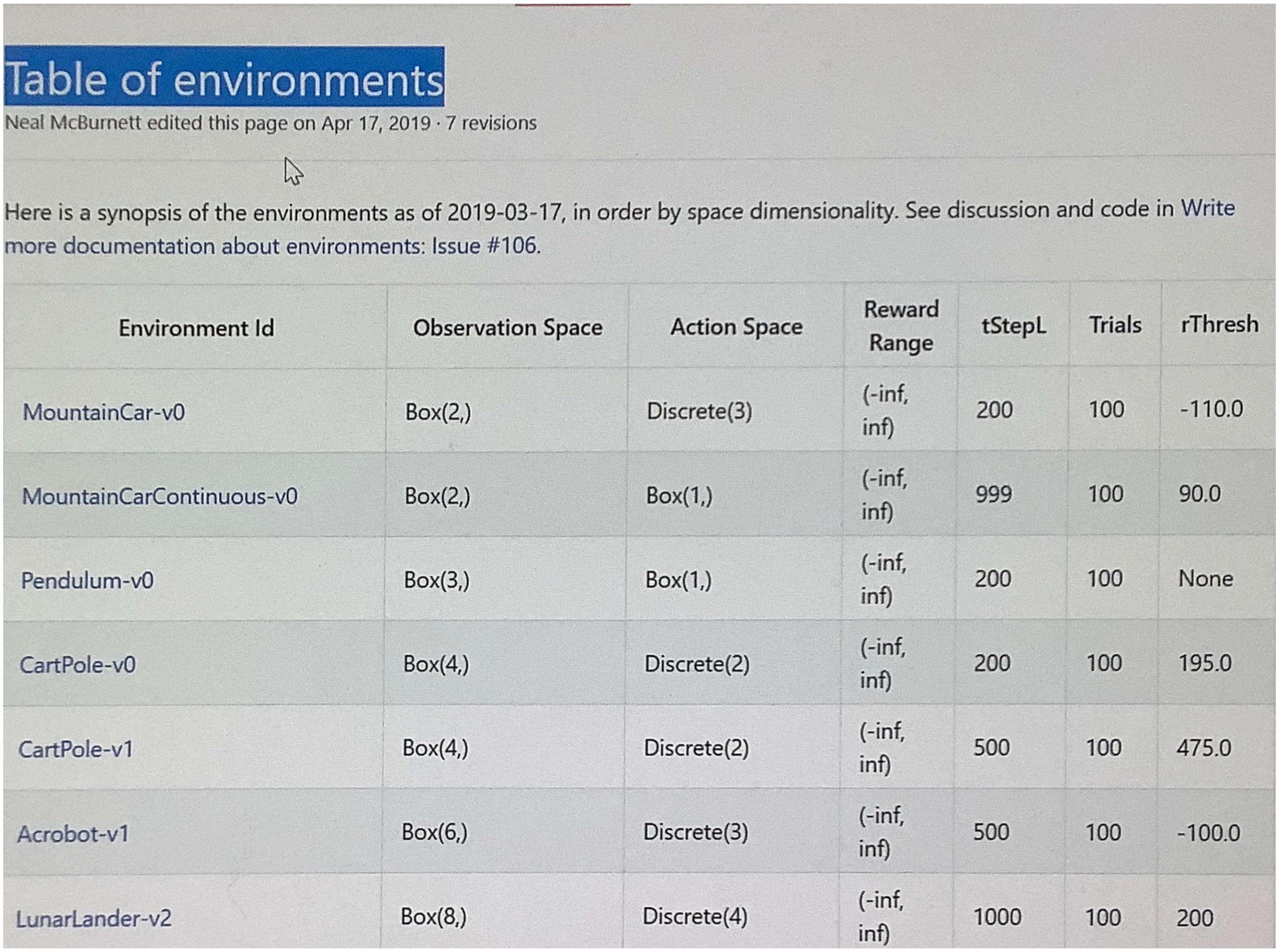


Table of environments

Neal McBurnett edited this page on Apr 17, 2019 · 7 revisions

Here is a synopsis of the environments as of 2019-03-17, in order by space dimensionality. See discussion and code in Write more documentation about environments: Issue #106.

Environment Id	Observation Space	Action Space	Reward Range	tStepL	Trials	rThresh
MountainCar-v0	Box(2,)	Discrete(3)	(-inf, inf)	200	100	-110.0
MountainCarContinuous-v0	Box(2,)	Box(1,)	(-inf, inf)	999	100	90.0
Pendulum-v0	Box(3,)	Box(1,)	(-inf, inf)	200	100	None
CartPole-v0	Box(4,)	Discrete(2)	(-inf, inf)	200	100	195.0
CartPole-v1	Box(4,)	Discrete(2)	(-inf, inf)	500	100	475.0
Acrobot-v1	Box(6,)	Discrete(3)	(-inf, inf)	500	100	-100.0
LunarLander-v2	Box(8,)	Discrete(4)	(-inf, inf)	1000	100	200