Last Time

- What tools do we have to solve MDPs with continuous S and A?

Course Map

- Outcome Uncertainty, Immediate vs Future Rewards (MDP)
- Model Uncertainty (Reinforcement Learning)
- State Uncertainty (POMDP)
- Interaction Uncertainty (Game)

Course Map

- Outcome Uncertainty, Immediate vs Future Rewards (MDP)
- Model Uncertainty (Reinforcement Learning)
- State Uncertainty (POMDP)
- Interaction Uncertainty (Game)

Course Map

- Outcome Uncertainty, Immediate vs Future Rewards (MDP)
- Model Uncertainty (Reinforcement Learning)
- State Uncertainty (POMDP)
- Interaction Uncertainty (Game)

Guiding Questions

Guiding Questions

- What is Reinforcement Learning?
- What are the main challenges in Reinforcement Learning?

Guiding Questions

- What is Reinforcement Learning?
- What are the main challenges in Reinforcement Learning?
- How do we categorize RL approaches?

Problem from HW2

Question 2. (25 pts) Consider a game with 3 squares in a horizontal line drawn on paper, a token, and a die. Each turn, the player can either reset or roll the die. If the player rolls and the die shows an odd number, the token is moved one square to the right, and if an even number is rolled, the token is moved two squares to the right (in both cases stopping at the rightmost square[1]). If the player resets, the token is always moved to the leftmost square. If the reset occurs when the token is in the middle square, two points are added; if the player resets when the token is on the right square, a point is subtracted.

c) Suppose you are not sure that the die is fair (i.e. whether it will yield odd and even with equal probability). Give finite upper and lower bounds for the accumulated discounted score that you can expect to receive with discount $\gamma = 0.95$.

$$\frac{\bar{r}}{1-\gamma}$$

Reinforcement Learning

Reinforcement Learning

Previously: (S, A, T, R, γ)

Reinforcement Learning

Previously: $(S, A, \cancel{T}, \cancel{R}, \gamma)$

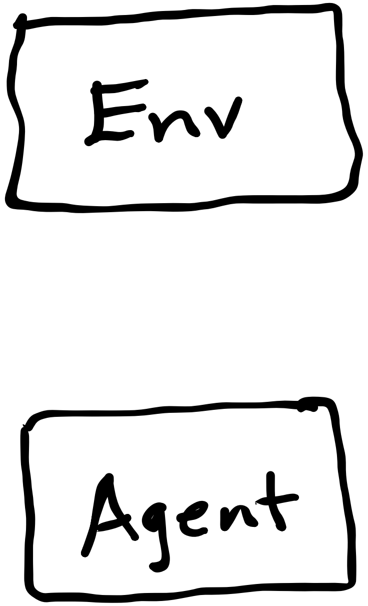
Unknown!

Reinforcement Learning

Previously: $(S, A, \cancel{T}, \cancel{R}, \gamma)$

Unknown!

Now: Episodic Simulator

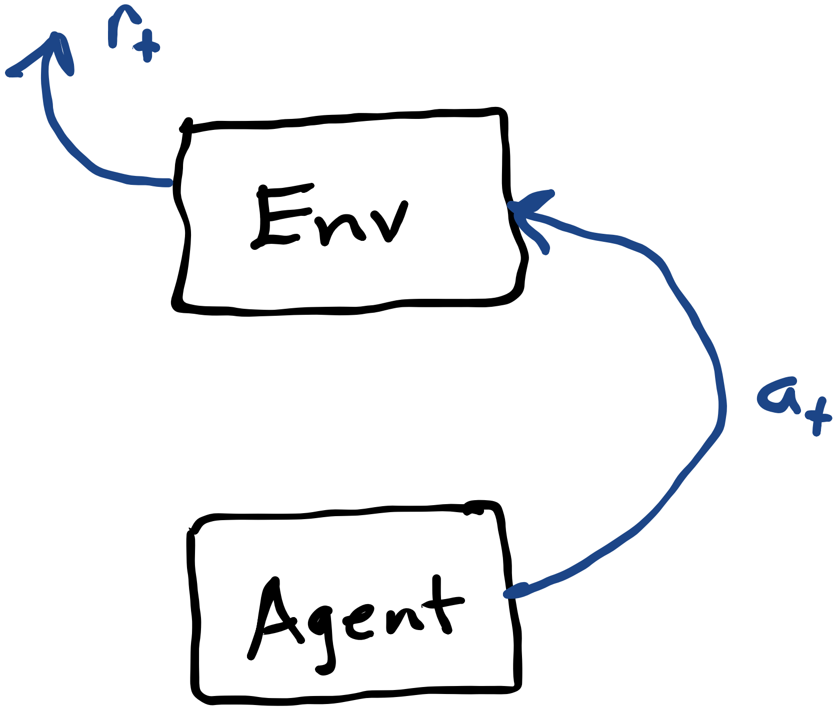


Reinforcement Learning

Previously: $(S, A, \cancel{T}, \cancel{R}, \gamma)$

Unknown!

Now: Episodic Simulator



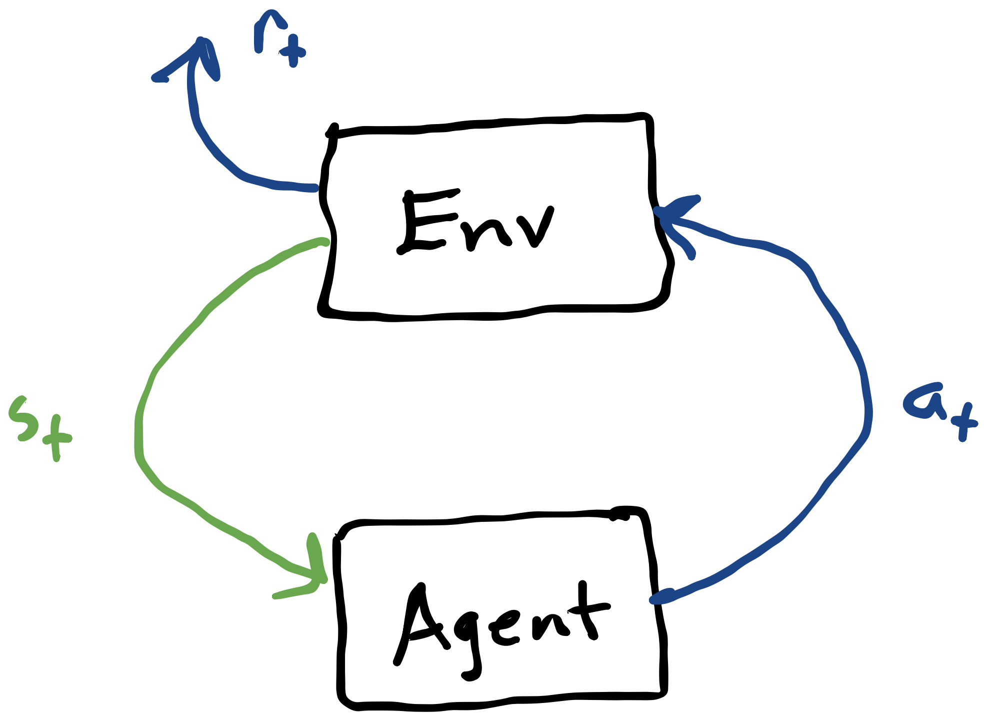

```
r = act!(env, a)
```

Reinforcement Learning

Previously: $(S, A, \cancel{T}, \cancel{R}, \gamma)$

Unknown!

Now: Episodic Simulator



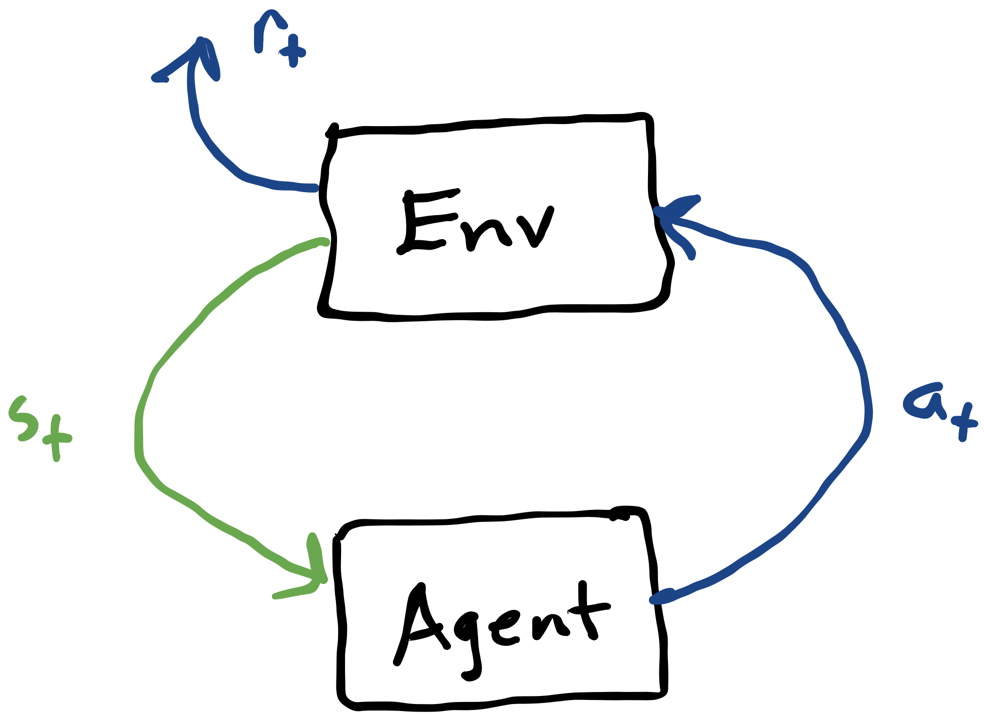

```
r = act!(env, a)

s = observe(env)
```

Reinforcement Learning

Previously: $(S, A, \cancel{T}, \cancel{R}, \gamma)$

Unknown!

Now: Episodic Simulator




```
r = act!(env, a)

s = observe(env)
```

In python, typically

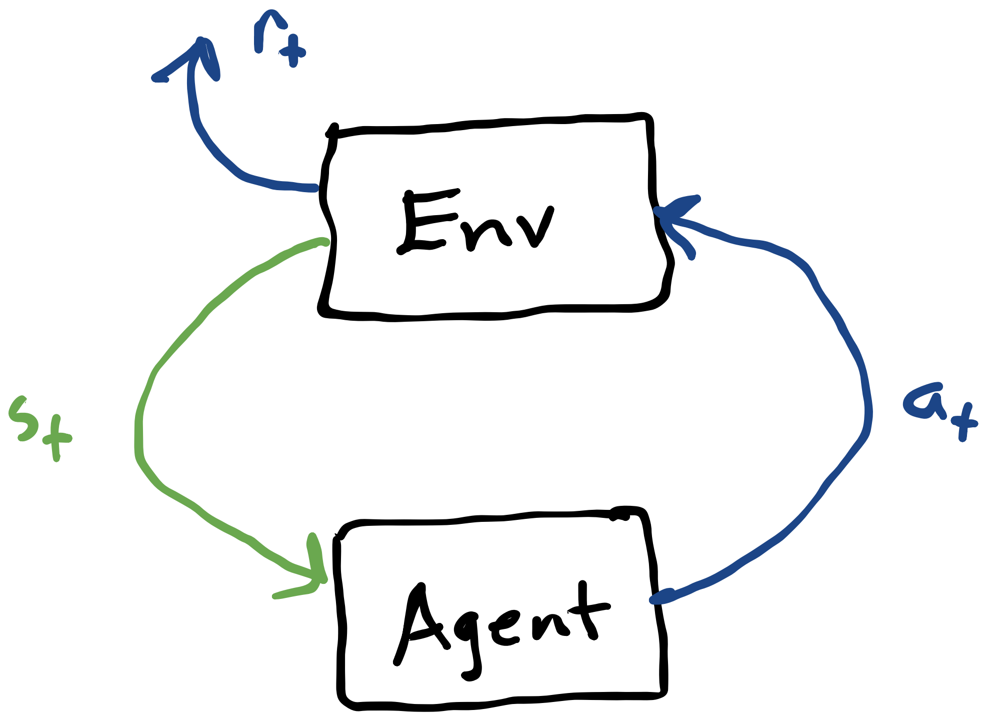
```
s, r = env.step(a)
```

Reinforcement Learning

Previously: $(S, A, \cancel{T}, \cancel{R}, \gamma)$

Unknown!

Now: Episodic Simulator




```
r = act!(env, a)

s = observe(env)
```

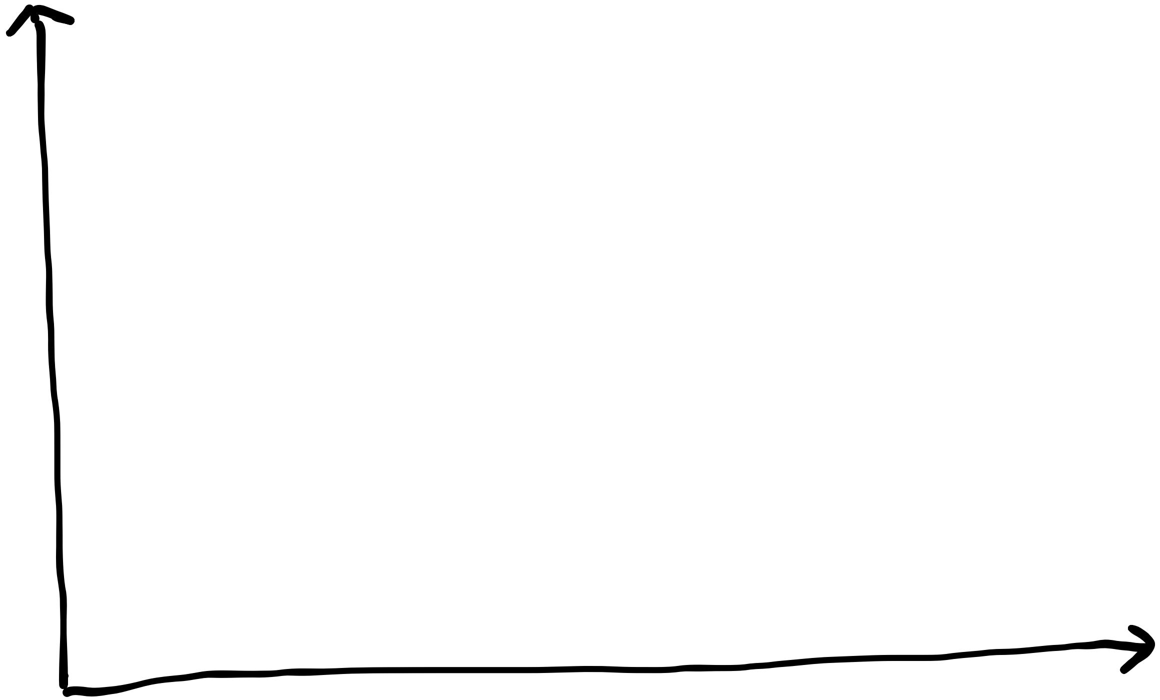
In python, typically

```
s, r = env.step(a)
```

Note: Different from $s', r = G(s, a)$

Learning Curve

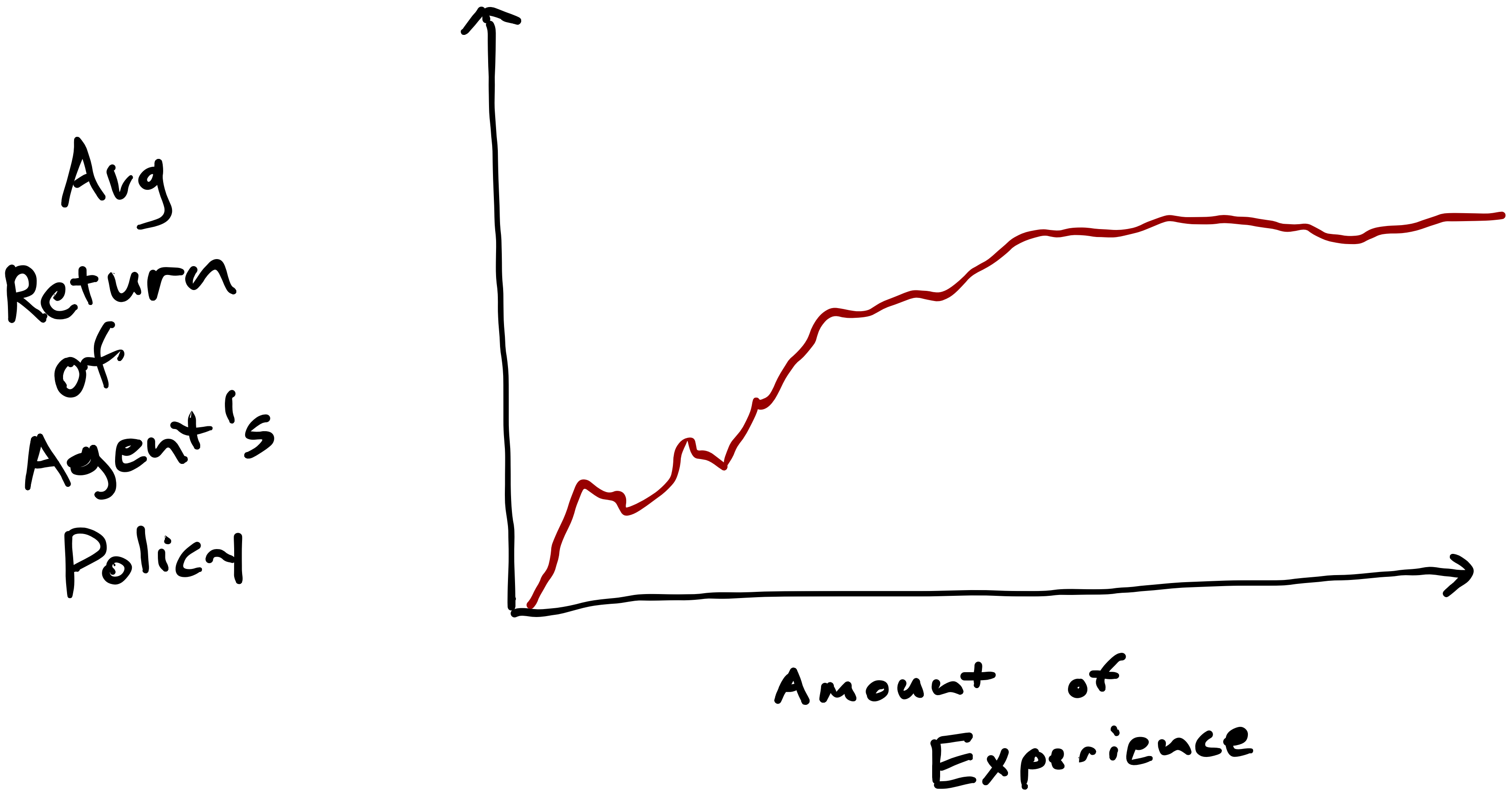
Learning Curve



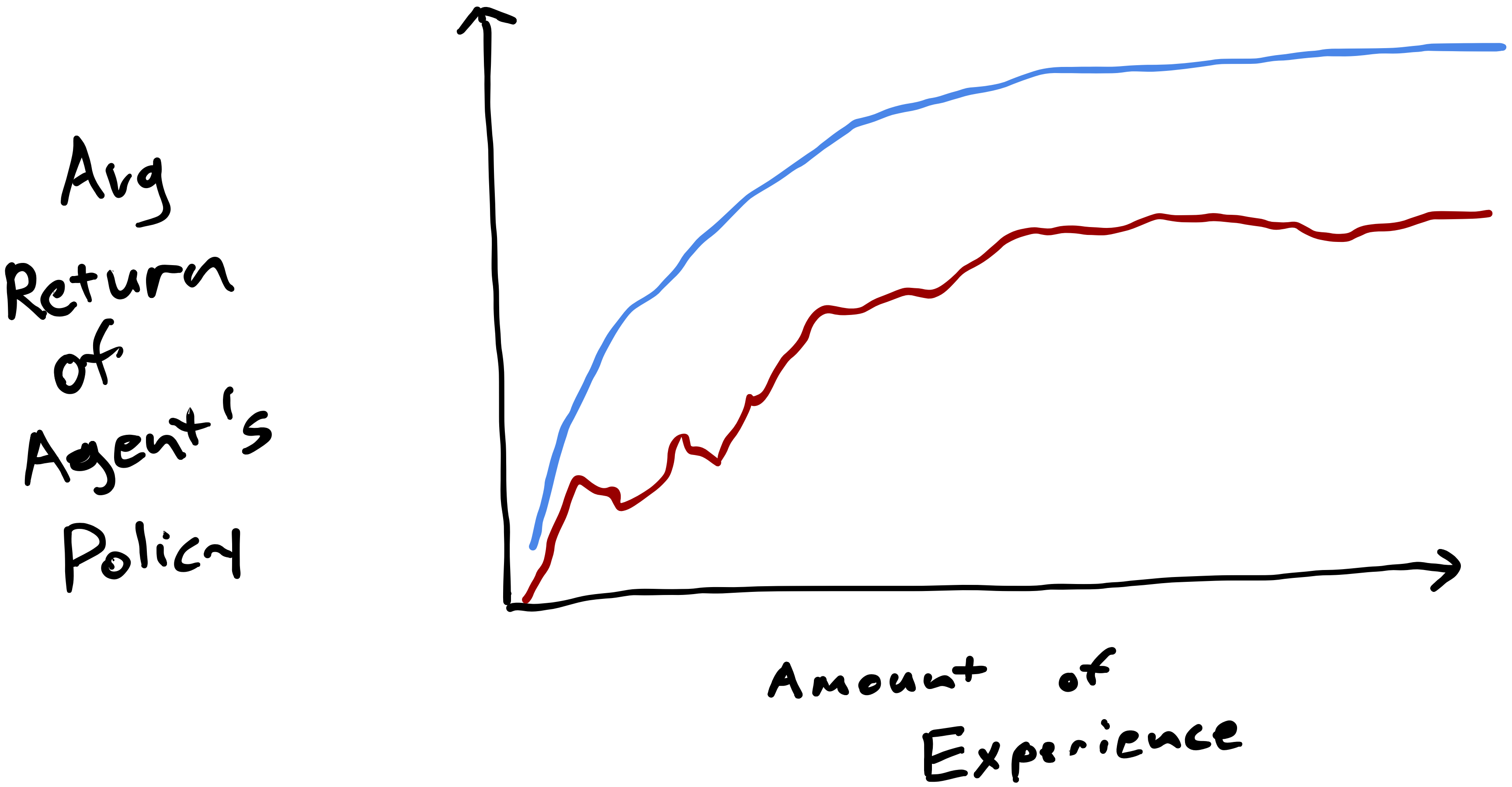
Learning Curve

Learning Curve

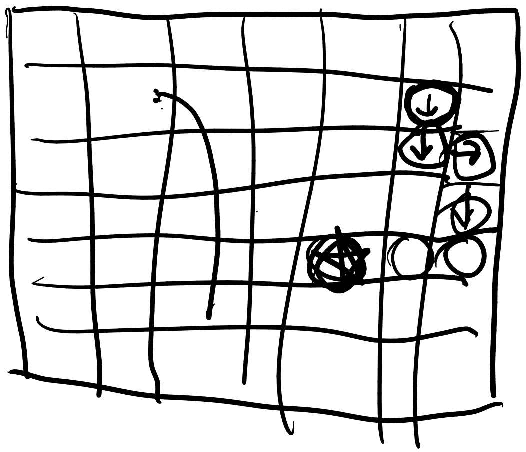
Learning Curve



Learning Curve



Break



know S, A

How to interact to maximize reward given no knowledge of T and R.

(1) $\hat{R}(s,a) = \Big\{$ heuristic that steers toward reward states.

$\hat{U}$ $\quad \pi(s) = \text{argmax} \left(\hat{R}(s,a) + \gamma E\left[\hat{U}(s')\right] \right)$

$\hat{T}$?

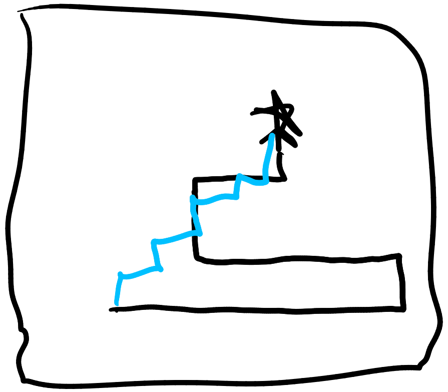
$\hat{Q}(s,a)$ $\quad \pi(s) = \text{argmax}\, \hat{Q}(s,a)$

$\hat{\pi}(s)$ = action that ~~has~~ led to the most future reward in previous episodes

Challenges

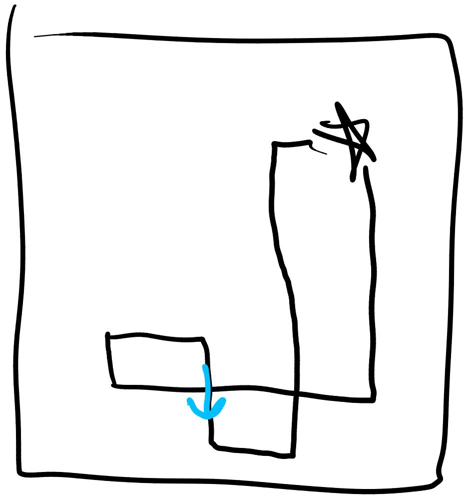
Challenges

1. Exploration vs Exploitation



Challenges

1. Exploration vs Exploitation
2. Credit Assignment



Challenges

1. Exploration vs Exploitation
2. Credit Assignment
3. Generalization

Classifications

Classifications

- **Model Based**: Attempt to learn T and R, then find π^* by solving MDP

Classifications

- **Model Based**: Attempt to learn T and R, then find π^* by solving MDP
- **Model Free**: Attempt to find Q^* or π^* directly without estimating T or R

Classifications

- **Model Based**: Attempt to learn T and R, then find π^* by solving MDP
- **Model Free**: Attempt to find Q^* or π^* directly without estimating T or R

- **On-Policy**: Learn only using experience generated with the current policy.

Classifications

- **Model Based**: Attempt to learn T and R, then find π^* by solving MDP
- **Model Free**: Attempt to find Q^* or π^* directly without estimating T or R

- **On-Policy**: Learn only using experience generated with the current policy.
- **Off-Policy**: Learn using experience generated from the current policy *and* previous policies.

Classifications

- **Model Based**: Attempt to learn T and R, then find π^* by solving MDP
- **Model Free**: Attempt to find Q^* or π^* directly without estimating T or R

- **On-Policy**: Learn only using experience generated with the current policy.
- **Off-Policy**: Learn using experience generated from the current policy *and* previous policies.
- **Batch**: Learn only from previously-generated experience.

Classifications

- **Model Based**: Attempt to learn T and R, then find π^* by solving MDP
- **Model Free**: Attempt to find Q^* or π^* directly without estimating T or R

- **On-Policy**: Learn only using experience generated with the current policy.
- **Off-Policy**: Learn using experience generated from the current policy *and* previous policies.
- **Batch**: Learn only from previously-generated experience.

- **Tabular**: Keep track of learned values for each state in a table

Classifications

- **Model Based**: Attempt to learn T and R, then find π^* by solving MDP
- **Model Free**: Attempt to find Q^* or π^* directly without estimating T or R

- **On-Policy**: Learn only using experience generated with the current policy.
- **Off-Policy**: Learn using experience generated from the current policy *and* previous policies.
- **Batch**: Learn only from previously-generated experience.

- **Tabular**: Keep track of learned values for each state in a table
- **Deep**: Use a neural network to approximate learned values

Tabular Maximum Likelihood Model-Based RL

TMLMBRL
MLMBTRL

Given env, S, A

$N \leftarrow 0$
$\rho \leftarrow 0$
$s \leftarrow$ observe(env)
$\pi \leftarrow$ random policy
loop
$\quad a \leftarrow \begin{cases} \text{rand}(A) & \text{w.p. } \varepsilon \\ \pi(s) & \text{w.p. } 1-\varepsilon \end{cases}$
$\quad r \leftarrow$ act!(env, a)
$\quad s' \leftarrow$ observe(env)
$\quad N[s,a,s'] \mathrel{+}= 1$
$\quad \rho[s,a] \mathrel{+}= r$
$\quad T[s,a,s'] \leftarrow \dfrac{N[s,a,s']}{\sum_{s'} N[s,a,s']}$ $\quad \forall\ s,a,s'$
$\quad R[s,a] \leftarrow \dfrac{\rho[s,a]}{\sum_{s'} N[s,a,s']}$ $\quad \forall\ s,a$
$\quad \pi \leftarrow$ solve(T, R) ← expensive
$\quad s \leftarrow s'$

$N[s,a,s']$ ← # a taken in s resulting in s'
$\rho[s,a]$ ← cumulative reward

(s,a,r,s')

Guiding Questions

- What is Reinforcement Learning?
- What are the main challenges in Reinforcement Learning?
- How do we categorize RL approaches?

Exploitation vs Exploration
Credit Assignment
Gen.