AIL 722: Reinforcement Learning

Lecture 13: Fitted Q-Iteration

Raunak Bhattacharyya

ScAI YARDI SCHOOL OF ARTIFICIAL INTELLIGENCE
INDIAN INSTITUTE OF TECHNOLOGY DELHI

Outline

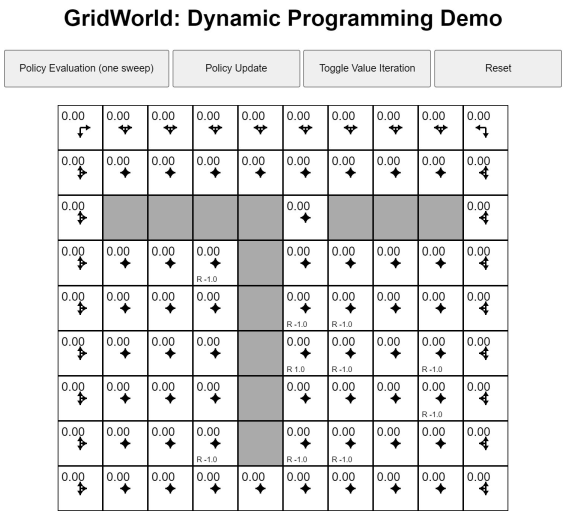
- Fitted value iteration

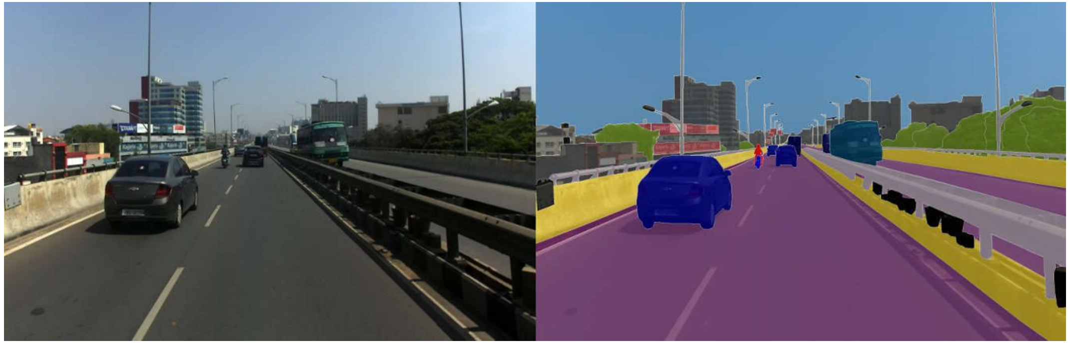
- Towards model-free algorithms

- Fitted Q iteration

Approximating the Value Function

Toy Domains to Reality



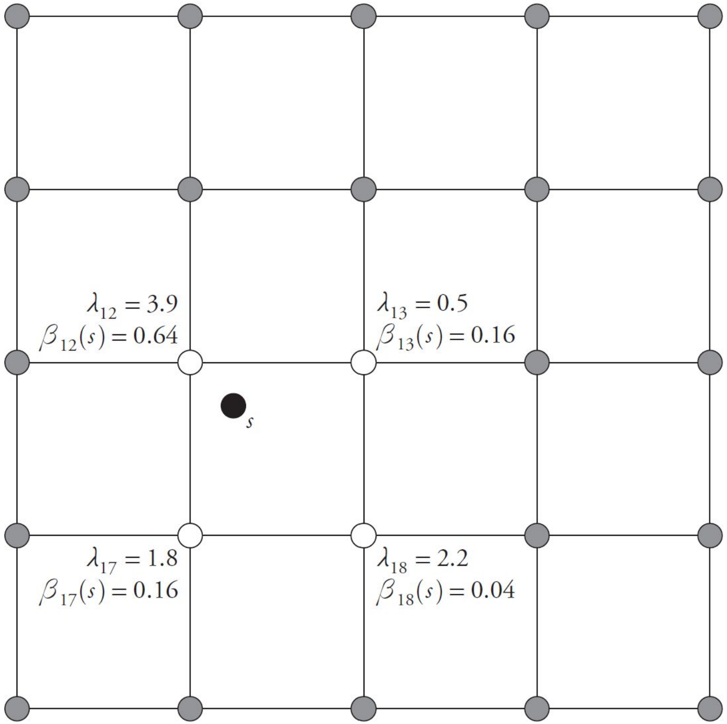


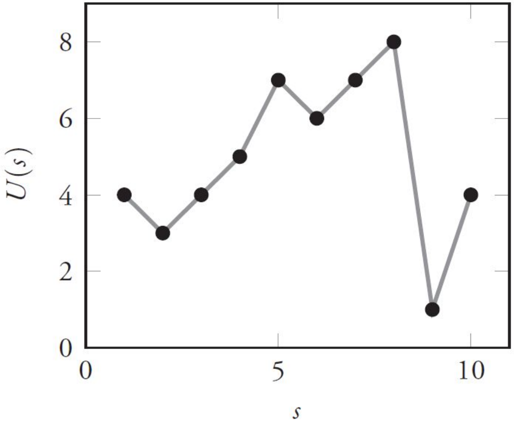
$$|S| = (255^3)^{600 \times 600}$$

Curse of dimensionality

From INSAAN, India Driving Dataset

Approximate DP



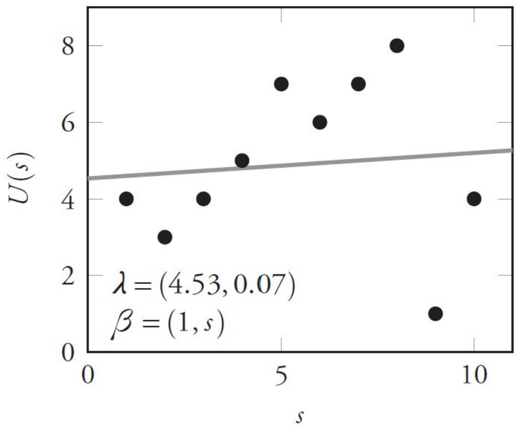


(a) Linear interpolation.

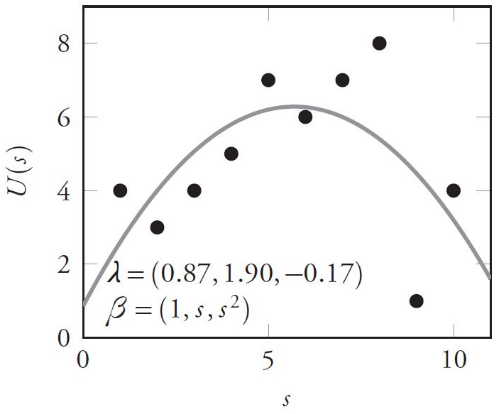
From [Decision Making under Uncertainty](), Mykel Kochenderfer

Global Approximation: Different Basis Functions

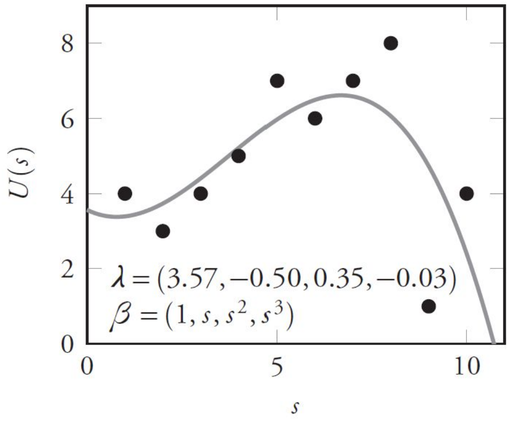
One dim state space



(b) Linear regression (linear basis).



(c) Linear regression (quadratic basis).

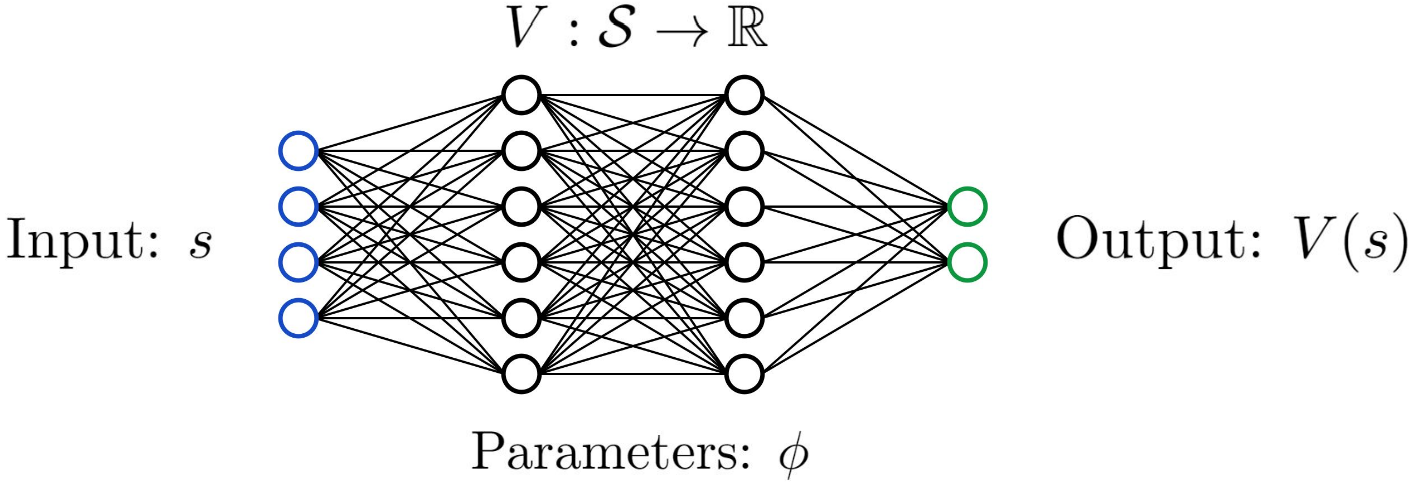


(d) Linear regression (cubic basis).

From Decision Making under Uncertainty, Mykel Kochenderfer

Fitted Value Iteration

Approximating V



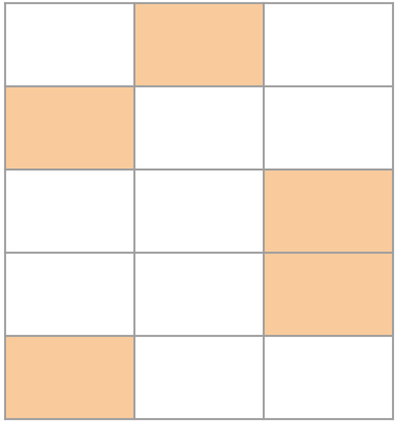
Foundation: Value Iteration

$Q(s_1,a_1)$	$Q(s_1,a_2)$	$Q(s_1,a_3)$
$Q(s_2,a_1)$	$Q(s_2,a_2)$	$Q(s_2,a_3)$
$Q(s_3,a_1)$	$Q(s_3,a_2)$	$Q(s_3,a_3)$
$Q(s_4,a_1)$	$Q(s_4,a_2)$	$Q(s_4,a_3)$
$Q(s_5,a_1)$	$Q(s_5,a_2)$	$Q(s_5,a_3)$

Start with a random value function $V(s)$

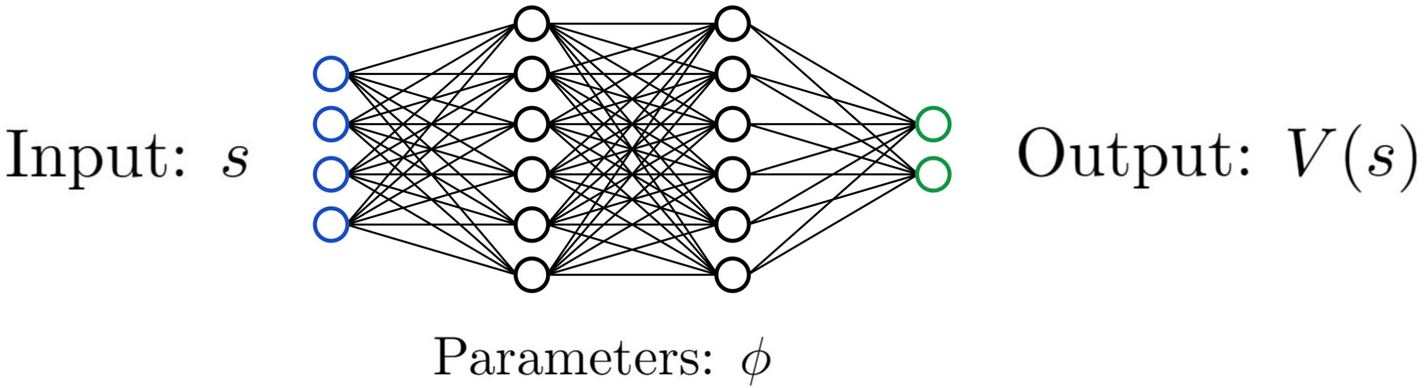
1. Set $Q(s,a) \longleftarrow r(s,a) + \gamma . \mathbb{E}_{p(s'|s,a)}\left[V^{\pi}(s')\right]$

2. Set $V(s) \longleftarrow \max_a Q(s,a)$



$V(s_1)$
$V(s_2)$
$V(s_3)$
$V(s_4)$
$V(s_5)$

Loss Function



$$L(\phi) = \frac{1}{2} \left\| V_\phi(s) - \max_a Q^\pi(s,a) \right\|^2$$

Fitted Value Iteration

1. Set $y_i \longleftarrow \max_{a_i} \left(r(s_i, a_i) + \gamma.\mathbb{E}\left[V_\phi(s_i') \right] \right)$

2. Set $\phi \longleftarrow \arg\min_\phi \sum_i \frac{1}{2} \left\| V_\phi(s_i) - y_i \right\|^2$

Fitted Value Iteration

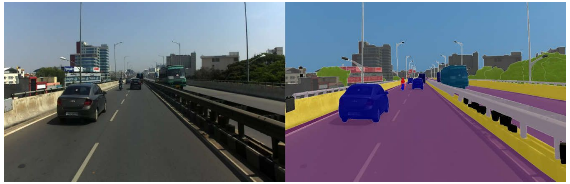
$$1.\ \text{Set } y_i \longleftarrow \max_a \left(r(s_i, a_i) + \gamma . \mathbb{E}\left[V_\phi(s_i') \right] \right)$$

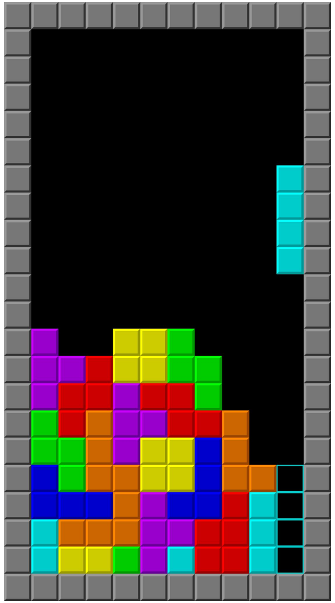
$$2.\ \text{Set } \phi \longleftarrow \arg\min_\phi \textstyle\sum_i \frac{1}{2} \left\| V_\phi(s_i) - y_i \right\|^2$$

- We will work with samples
- We have a (finite) sampled set of states
- At each state, we compute the Q values corresp to each action, then take the max over those to create our target y_i
- Compute NN parameters through linear regression to make V close to maxQ

What data do we need?

Towards Real World Problems





- State:
  - Board configuration
  - Shape of block (tetromino)
- Board is 10x20. And every square could be filled/not filled
- Action: Placement
- Reward: Number of rows eliminated
- Dynamics:
  - Wall change
  - Random next tetromino

How do we use fitted VI?

Unknown Transitions

Fitted VI: Restrictive Assumption

1. Set $y_i \longleftarrow \max_a \left(r(s_i, a_i) + \gamma . \mathbb{E}\left[V_\phi(s_i') \right] \right)$

2. Set $\phi \longleftarrow \arg\min_\phi \sum_i \frac{1}{2} \left\| V_\phi(s_i) - y_i \right\|^2$

There are two places where we require knowledge of the transition dynamics

Compute expected value at next state

Take max over actions (needs us to be able to try out all possible actions from the same state)

Need an MDP simulator: to try out every action, get next state and reward

Does not match up to experience-based learning in general. Cannot go back to exact same state to try out new actions.

Fitted Q-Iteration

Fitted Value Iteration

1. Set $y_i \longleftarrow \max_{a_i} \left(r(s_i, a_i) + \gamma . \mathbb{E}\left[V_\phi(s_i') \right] \right)$

2. Set $\phi \longleftarrow \arg\min_\phi \sum_i \frac{1}{2} \left\| V_\phi(s_i) - y_i \right\|^2$

We don't have a simulator. We only have a start state. We can sample trajectories

Underlying Idea: VI

1. Set $y_i \longleftarrow \max_a \left(r(s_i, a_i) + \gamma . \mathbb{E}\left[V_\phi(s_i') \right] \right)$

2. Set $\phi \longleftarrow \arg\min_\phi \sum_i \frac{1}{2} \left\| V_\phi(s_i) - y_i \right\|^2$

1. Set $Q(s, a) \longleftarrow r(s, a) + \gamma . \mathbb{E}_{p(s'|s,a)} \left[V^\pi(s') \right]$

2. Set $V(s) \longleftarrow \max_a Q(s, a)$

How do we get to a model-free algorithm?

Fitted Q-Iteration

1. Set $Q(s,a) \longleftarrow r(s,a) + \gamma.\mathbb{E}_{p(s'|s,a)}\left[V^{\pi}(s')\right]$
2. Set $V(s) \longleftarrow \max_a Q(s,a)$

$$\text{Set } V(s) \longleftarrow \max_a Q(s,a)$$

The crux element

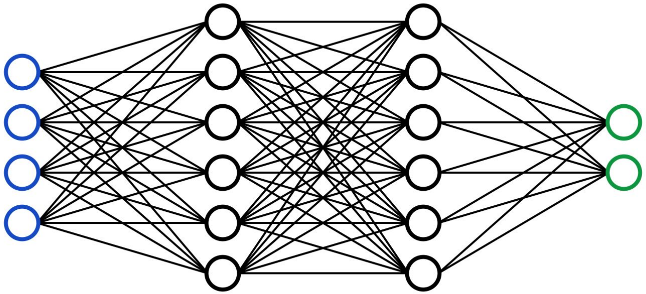
$$\text{Set } Q(s,a) \longleftarrow r(s,a) + \gamma.\mathbb{E}_{p(s'|s,a)}\left[V^{\pi}(s')\right]$$

$$\text{Set } Q(s,a) \longleftarrow r(s,a) + \gamma.\max_{a'} Q(s',a')$$

No longer exact. What's the approximation?

Fitted QI: Key Elements

Work with Q, instead of V



Input: s

Output: $Q(s, a)$

Parameters: ϕ

$$\arg\min_\phi \sum_i \frac{1}{2} \|Q_\phi(s_i, a_i) - y_i\|^2$$

Fitted QI: What's the Algorithm Then?

1. Collect dataset

2. Set $y_i \longleftarrow r(s_i, a_i) + \gamma . \max_{a'_i} Q_\phi(s'_i, a'_i)$

3. Set $\phi \longleftarrow \arg\min_\phi \sum_i \frac{1}{2} \left\| Q_\phi(s_i, a_i) - y_i \right\|^2$

What data do we need?

Fitted QI: What's the Algorithm Then?

1. Collect dataset $\{(s_i, a_i, r_i, s_i')\}$ using some policy

2. Set $y_i \longleftarrow r(s_i, a_i) + \gamma . \max_{a_i'} Q_\phi(s_i', a_i')$

3. Set $\phi \longleftarrow \arg\min_\phi \sum_i \frac{1}{2} \left\| Q_\phi(s_i, a_i) - y_i \right\|^2$

How Model Free? Fitted VI vs QI

1. Collect dataset $\{(s_i, a_i, r_i, s'_i)\}$ using some policy

Set $y_i \longleftarrow \max_a \left(r(s_i, a_i) + \gamma . \mathbb{E}\left[V_\phi(s'_i) \right] \right)$

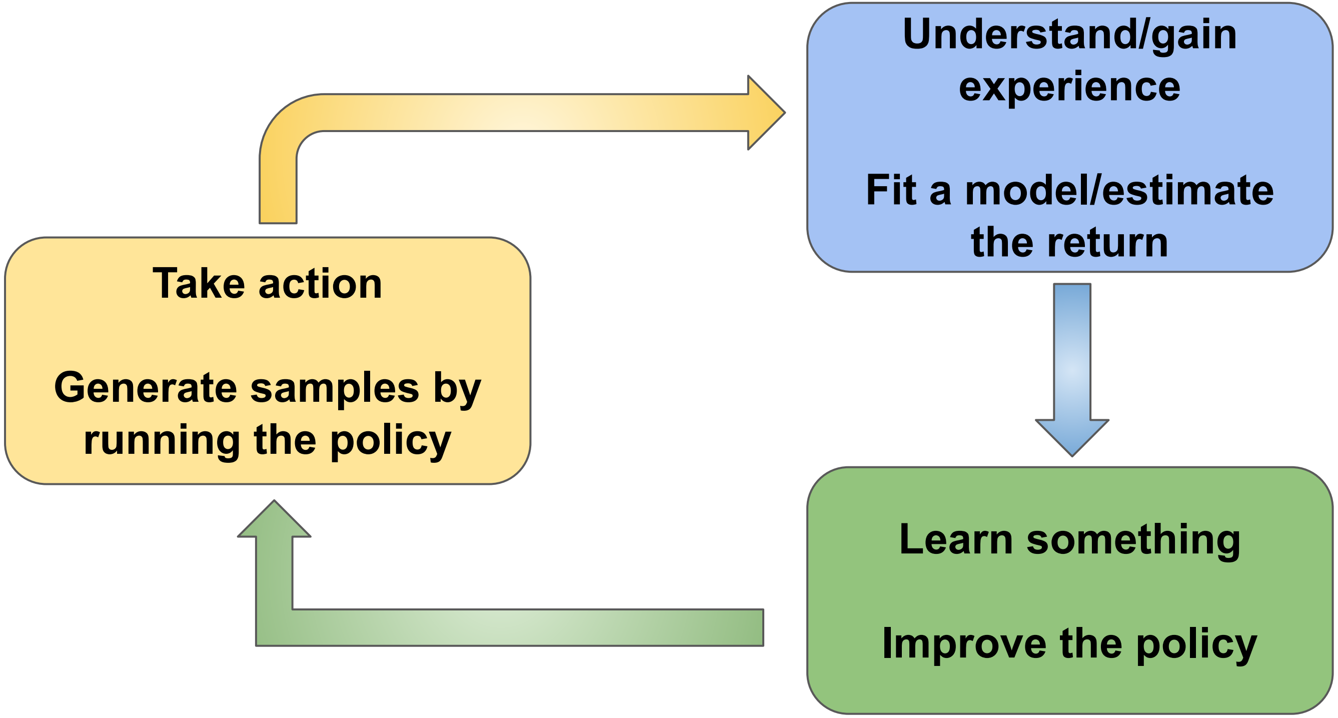
Set $y_i \longleftarrow r(s_i, a_i) + \gamma . \max_{a'_i} Q_\phi(s'_i, a'_i)$

Set $\phi \longleftarrow \arg\min_\phi \sum_i \frac{1}{2} \left\| V_\phi(s_i) - y_i \right\|^2$

Set $\phi \longleftarrow \arg\min_\phi \sum_i \frac{1}{2} \left\| Q_\phi(s_i, a_i) - y_i \right\|^2$

Unifying Lens on Algorithms

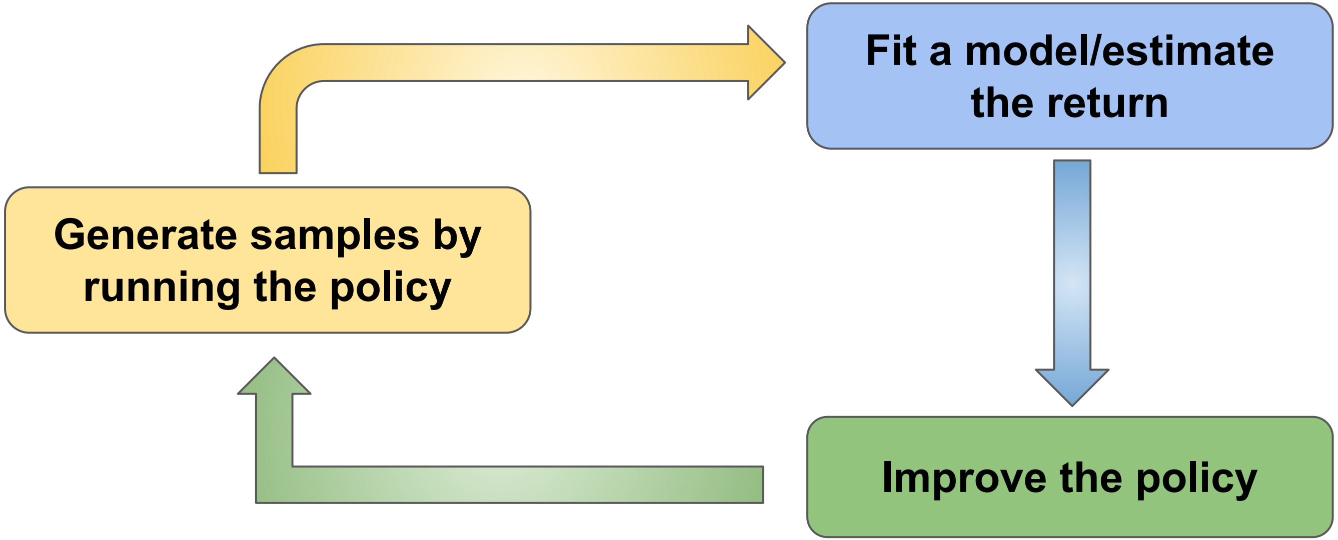
Unifying Anatomy of RL Algorithms



Adapted from [Berkeley CS 285, Instructor: Sergey Levine]

Anatomy of Fitted Value Iteration

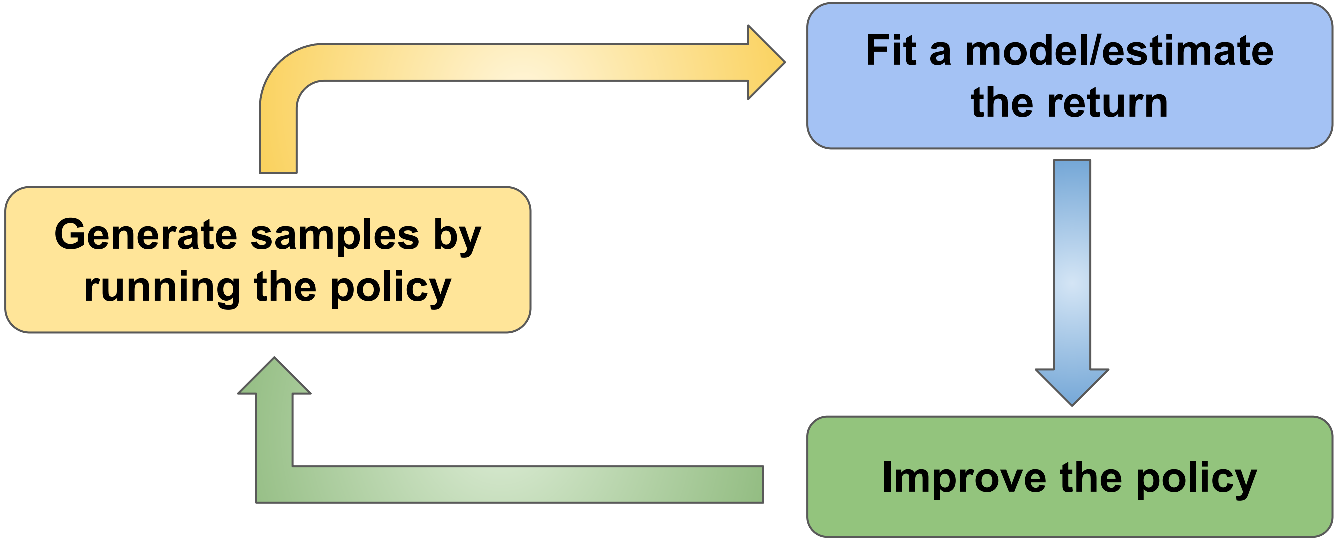
$$Q(s,a) \longleftarrow r(s,a) + \gamma.\mathbb{E}_{p(s'|s,a)}\Big[V^{\pi}(s')\Big]$$



$$V(s) \longleftarrow \max_a Q(s,a)$$

Anatomy of Fitted Q-Iteration

$$Q(s,a) \longleftarrow r(s,a) + \gamma . \max_{a'} Q(s',a')$$



Announcements

- Assignment 1 deadline
  - Sunday, 25 Aug, 11.55 pm

Course webpage