AIL 722: Reinforcement Learning

Lecture 11: Fitted Value Iteration

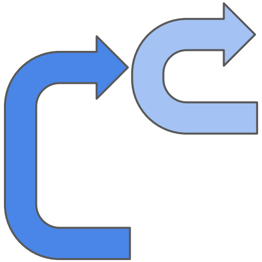
Raunak Bhattacharyya

YARDI SCHOOL OF ARTIFICIAL INTELLIGENCE
INDIAN INSTITUTE OF TECHNOLOGY DELHI

Outline

- Approximating the value function
- Fitted value iteration
- Fitted Q iteration

Policy Iteration: Using Q-values

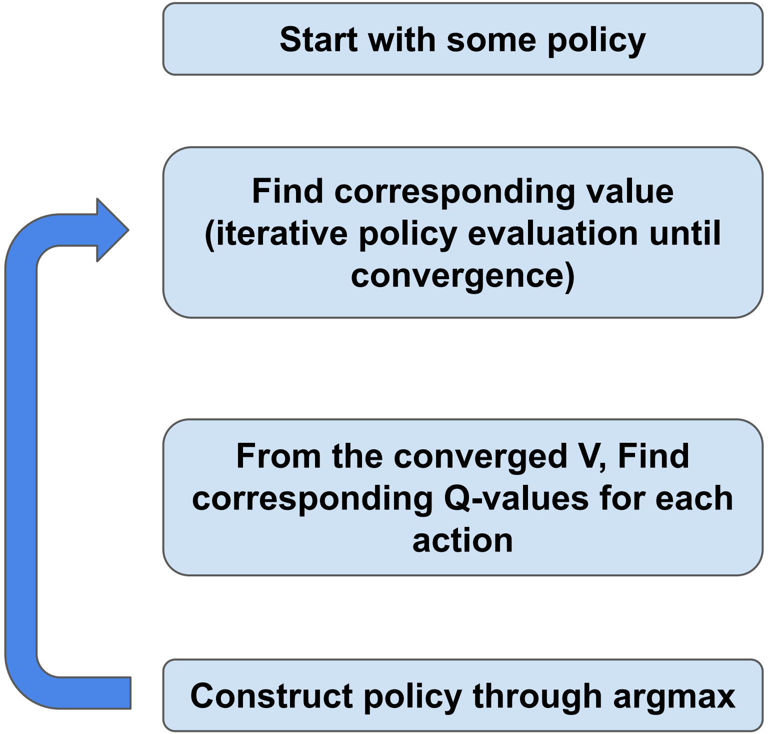


1. $V^\pi(s) = r\Big(s, \pi(s)\Big) + \gamma \cdot \mathbb{E}_{p(s'|s,\pi(s))}\Big[V^\pi(s')\Big]$

2. Set $\pi \longleftarrow \pi_{\text{new}}$

$$\pi_{\text{new}} = \begin{cases} 1 & \text{if } a = \arg\max_a Q^\pi(s,a) \\ 0 & \text{otherwise} \end{cases}$$

Workflow

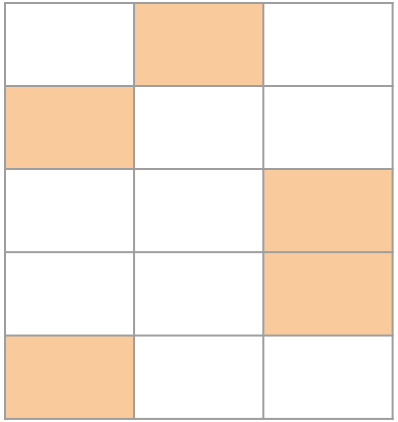

Start with some policy
Find corresponding value (iterative policy evaluation until convergence)
From the converged V, Find corresponding Q-values for each action
Construct policy through argmax

Value Iteration

$a_?$
$a_?$
$a_?$
$a_?$
$a_?$

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

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

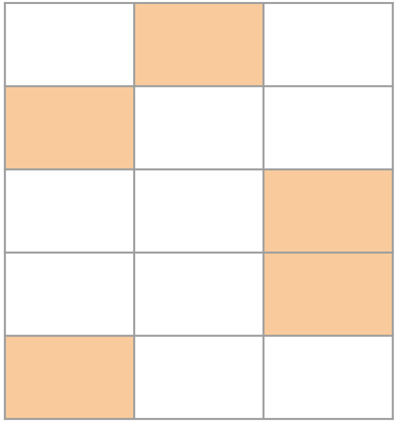


a_2
a_1
a_3
a_3
a_1

$a_?$
$a_?$
$a_?$
$a_?$
$a_?$

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

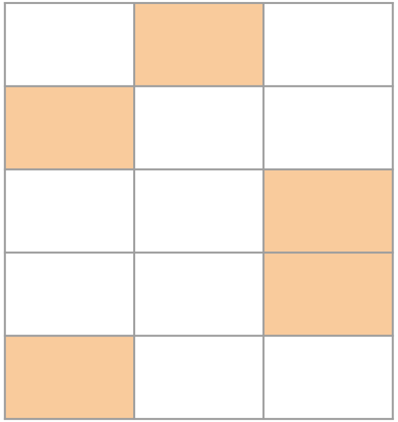
$Q(s_1,a_1)$	$Q(s_1,a_2)$	$Q(s_1,a_3)$
$Q(s_2,a_1)$	$Q(s_2,a_2)$	$Q(s_2,a_3)$
$Q(s_3,a_1)$	$Q(s_3,a_2)$	$Q(s_3,a_3)$
$Q(s_4,a_1)$	$Q(s_4,a_2)$	$Q(s_4,a_3)$
$Q(s_5,a_1)$	$Q(s_5,a_2)$	$Q(s_5,a_3)$



a_2
a_1
a_3
a_3
a_1

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

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



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

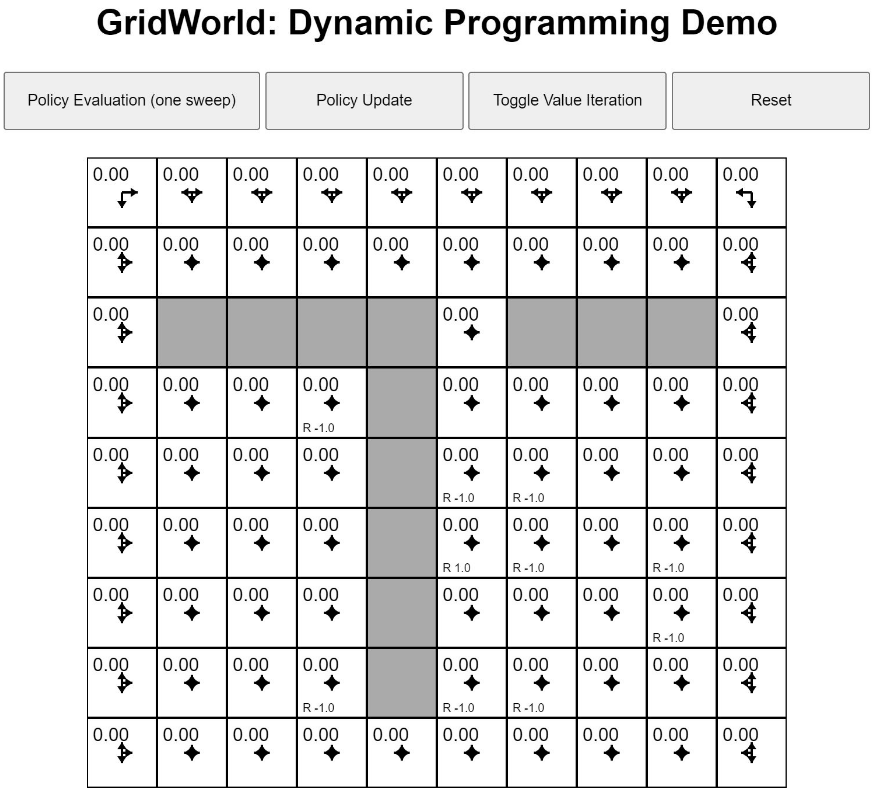
Value Iteration

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

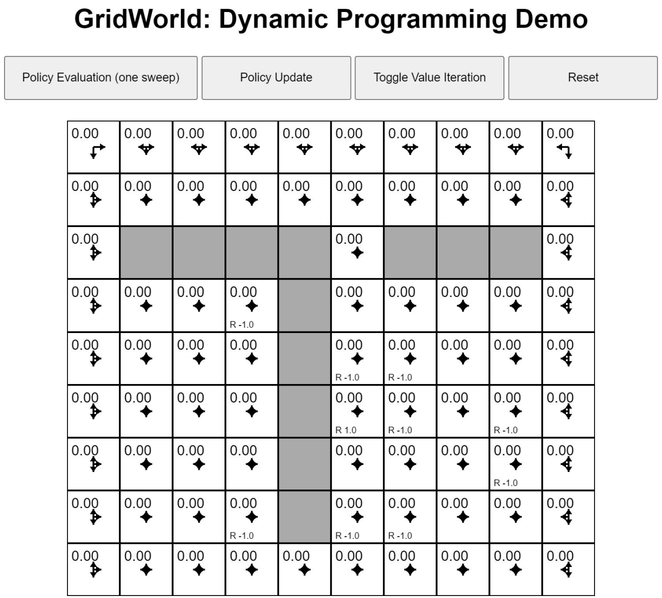
Value Iteration Demo



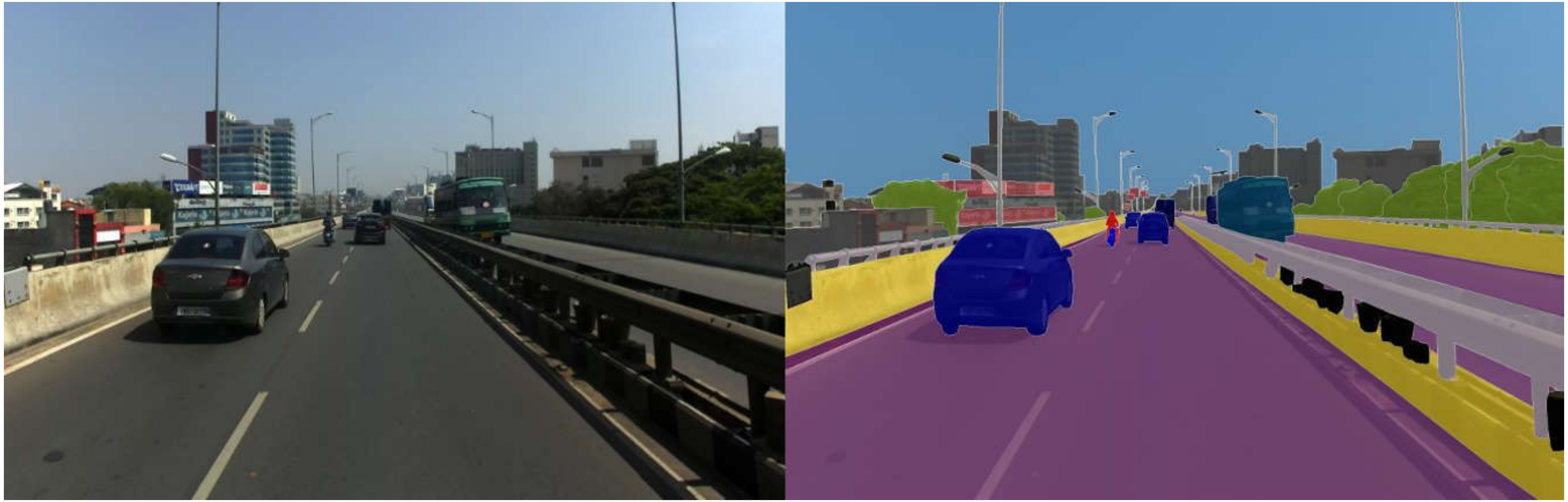
From ReinforceJS, Andrej Karpathy

Fitted Value Iteration

Toy Domains to Reality



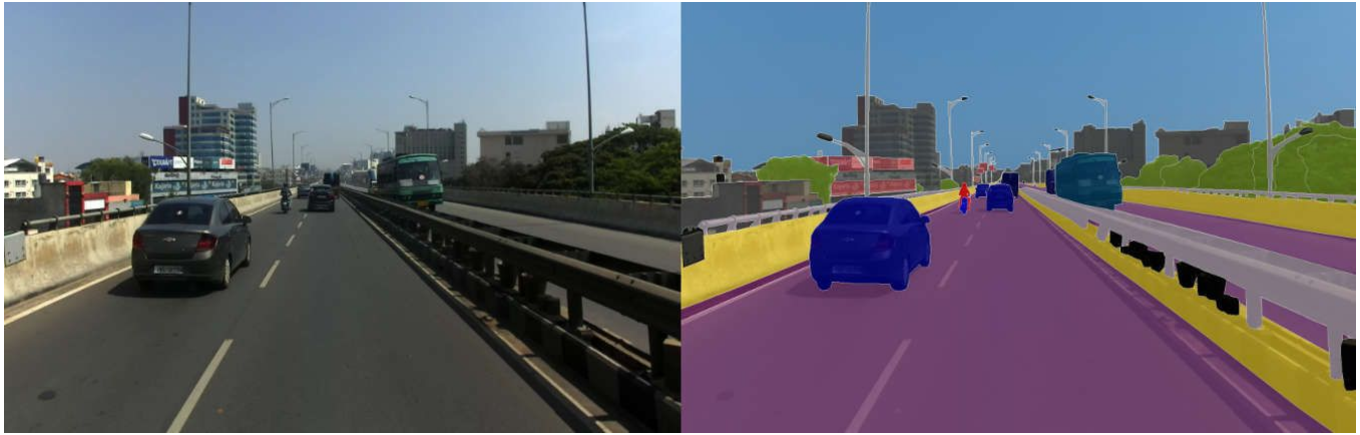
Toy Domains to Reality



How do we represent V?

From INSAAN, India Driving Dataset

Toy Domains to Reality

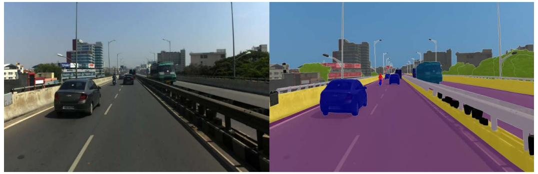


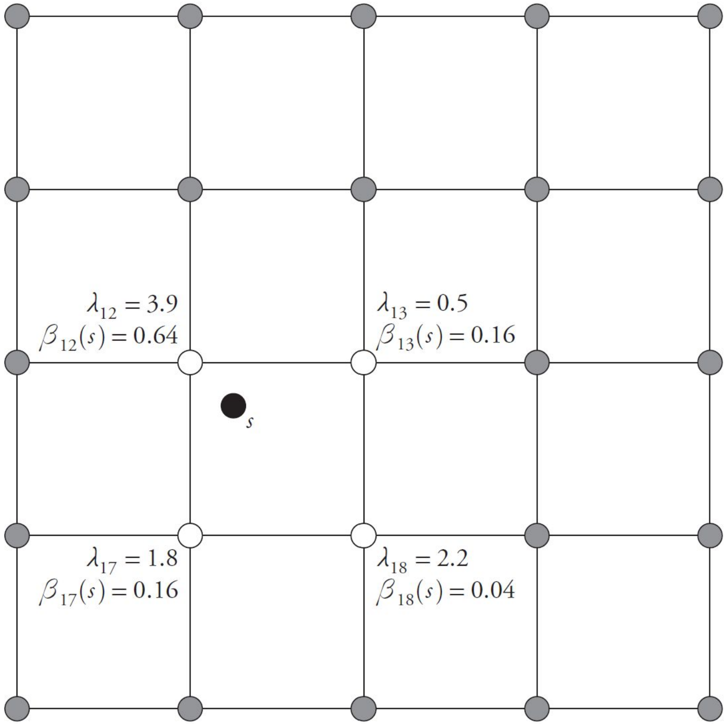
Curse of dimensionality

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

From [INSAAN, India Driving Dataset]

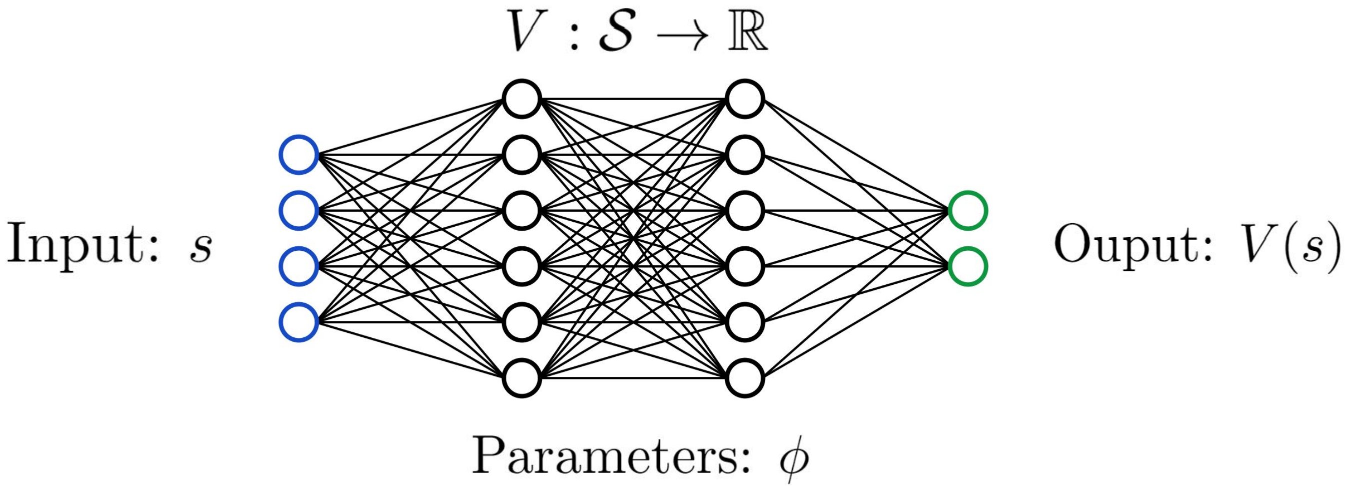
Approximating the Value Function





From Decision Making under Uncertainty, Mykel Kochenderfer

Approximating V



What should we train it on?

Value Iteration

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

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

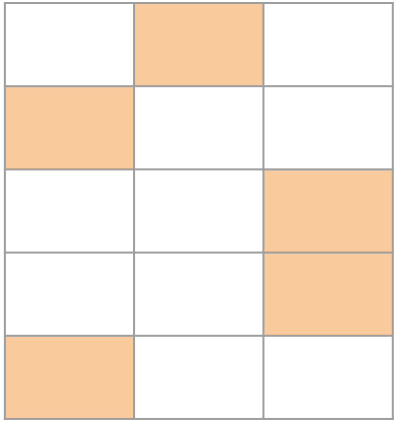
2. Set $V(s) \longleftarrow \max_a Q(s,a)$

Value Iteration

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

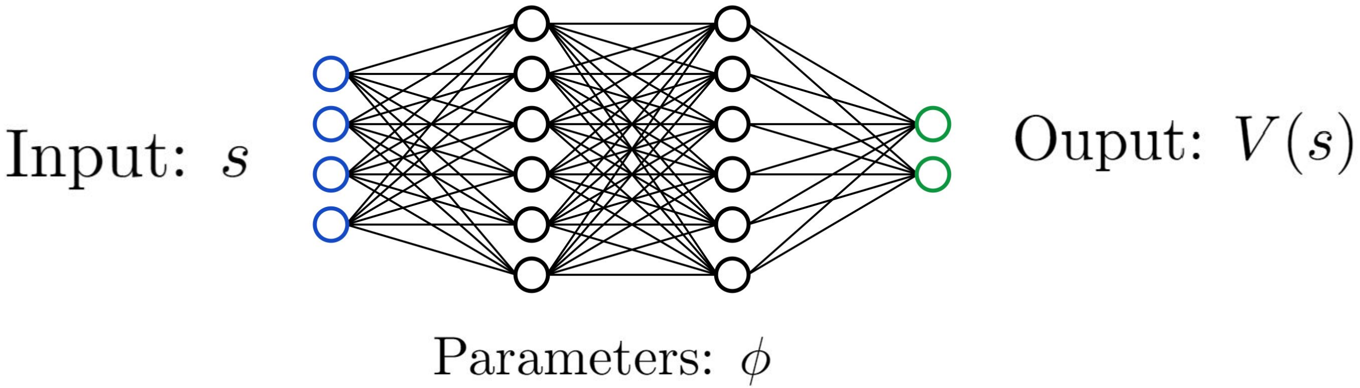
$$2.\ \text{Set } V(s) \longleftarrow \max_a Q(s,a)$$

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



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

Loss Function



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

How do we instantiate this?

Fitted Value Iteration

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

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

Fitted Value Iteration

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

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

Why did we not need samples in PI and VI?

- 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

Fitted Value Iteration

$$\text{Dataset: } \Big\{ (s_i, a_i, s'_i, r_i) \Big\}$$

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

Announcements

- Pytorch tutorial tomorrow

[Course webpage]