Wednesday, December 2, 2020

Lecture 28

CS 59000-RL

Agenda

- Competetive games
- Bellman Residual minimization
- Deep Q Network (DQN)
  - DDQN
  - BDQN
- Wrap up

So far single agent MDP. $\pi \to \eta(\pi) \to \max_{\pi} \eta(\pi)$

what about two player game?, e.g. competing against each other?

Two players play chess

- Play ping pong
- competing for the same resources
- competing for a same radio channel.
- racing
- ⋮

→ Competetive Markov Decision processes.

- zero sum games.

Two players at each time step make their decision after observing X_t

$P1 \rightarrow A^1 \sim \pi^1(X_t)$

$P2 \rightarrow A^2 \sim \pi^2(X_2)$ $\Rightarrow X_{t+1}$, r_t^1 r_t^2

$\Rightarrow$ zero sum game $r_t^1 + r_t^2 = 1$

$\Rightarrow r_t = r_t^1 = -r_t^2$

$\eta^1(\pi^1, \pi^2) \Rightarrow$ objective of P1

$\eta^2(\pi^1, \pi^2) \Rightarrow$ objective of P2

$$\eta(\pi^1, \pi^2) = \eta^1(\pi^1, \pi^2) = -\eta^2(\pi^1, \pi^2) = E_{\pi^1, \pi^2}\left[\sum r_t\right]$$

P1; first player aims to maximize η^1 and respectively η

P2; second player aims to maximize η^2, ergo minimize η

$$\max_{\pi_1} \eta(\pi^1, \pi^2) , \quad \min_{\pi_2} \eta(\pi^1, \pi^2)$$

How we develope policy gradient method for such setting?

Wednesday, December 2, 2020

How we did it for single player MDP?

$$\eta(\pi_\theta) \to \theta' = \theta + \underset{\Delta\theta}{\arg\max}\ \underline{\Delta\theta^T \nabla \eta(\theta)} - \underline{\frac{1}{2\alpha}\|\Delta\theta\|^2}$$

linear approximation of the objective

Tylor examsion

$$\Rightarrow \theta' = \theta + \alpha \nabla \eta(\theta)$$

$\theta^1 \to \pi^1_{\theta^1}$; $\theta^2 \to \pi^2_{\theta^2}$

Naive: $\theta^1 \leftarrow \theta^1 + \underset{\Delta\theta^1}{\arg\max}\ \Delta\theta^{1^T} \nabla_{\theta^1}\eta(\theta^1,\theta^2) - \frac{1}{2\alpha}\|\Delta\theta^1\|^2$

$$\theta^2 \leftarrow \theta^2 + \arg\min \Delta\theta^{2^T} \nabla_{\theta^2}\eta(\theta^1,\theta^2) + \frac{1}{2\alpha}\|\Delta\theta^2\|^2$$

Approximat $\eta(\theta^1,\theta^2)$ such that the local game is linear in each player' parameter.

$$\eta(\theta^1+\Delta\theta^1, \theta^2+\Delta\theta^2) \approx \eta(\theta^1,\theta^2)$$

$$+ \underline{\Delta\theta^1} \nabla_{\theta_1}\eta + \Delta\theta^2 \nabla_{\theta^2}\eta + \Delta\theta_1^T \nabla_{\theta_1\theta_2}\eta\, \Delta\theta_2$$

$$+ \Delta\theta^{2^T} \nabla_{\theta_2\theta_1}\eta\, \Delta\theta^1 + \underline{\text{other}}$$

$\Rightarrow$ This is a bilinear approximation of the game

Wednesday, December 2, 2020

$$\rightarrow \theta^1 \leftarrow \theta^1 + \arg\max_{\Delta\theta^1} \Delta\theta^{1T} \nabla_{\theta^1} \eta + \Delta\theta^1 \nabla_{\theta^1\theta^2} \eta \, \Delta\theta_2 - \frac{1}{2\alpha} \|\Delta\theta_1\|^2$$

$$\rightarrow$$

$$\rightarrow \theta^2 \leftarrow \theta^2 + \arg\min_{\Delta\theta^2} \Delta\theta^{2T} \nabla_{\theta^2} \eta + \Delta\theta^2 \nabla_{\theta^2\theta^1} \eta \, \Delta\theta_1 + \frac{1}{2\alpha} \|\Delta\theta_2\|^2$$

bilinear local game $\rightarrow$
has close form solution, which is the
Nash equilibrium of this bilinear game

$$\Rightarrow \theta^1 \leftarrow \theta^1 + \alpha \left(I + \alpha^2 \nabla_{\theta^1\theta^2} \eta \, \nabla_{\theta^2\theta^1} \eta \right)^{-1} \left(\nabla_{\theta^1} \eta - \alpha \nabla_{\theta^1\theta^2} \nabla_{\theta^2} \eta \right)$$

$$\theta^2 \leftarrow \theta^2 - \alpha \left(I + \alpha^2 \nabla_{\theta^2\theta^1} \eta \, \nabla_{\theta^1\theta^2} \eta \right)^{-1} \left(\nabla_{\theta^2} \eta + \alpha \nabla_{\theta^2\theta^1} \nabla_{\theta^1} \eta \right)$$

It is called competetive policy optimization principle
(CoPO) Prajapt et al 2020

Bellman residual minimization

we turn the problem of learning Q function
to some what clasical regression problem.

$\rightarrow$ follow a policy $\rightarrow$ x_t, a_t, r_t, x_{t+1}

$$\rightarrow \left\| f(x_t, a_t) - \underline{\left(r_t + \gamma \max_{a'} f(x_{t+1}, a') \right)} \right\|^2$$

$f(x) \rightarrow y \quad \|f(x) - y\|^2 \rightarrow r_t + \gamma f(x_{t+1})$

Wednesday, December 2, 2020

Notes:> in general the learned f is a biased estimate of Q.

↳ well behaved in Linear Q

This very objective function is the same used in Deep Q Network (DQN)

→ To tackle Atari games

→ To tackle Go

↳ Boosts of the field Deep RL

DQN

Initialize Q

1) run epsilon greedy policy

$$\begin{cases} \arg\max_a Q(x_t, a) & 1-\varepsilon \\ \text{uniform random} & \varepsilon \end{cases}$$

2) x_t, a_t, r_t, x_{t+1}

3) update Q → $\left\| Q(x_t, a_t) - r_t - \gamma \max_{a'} Q(x_{t+1}, a') \right\|^2$

4) using gradient descent

Wednesday, December 2, 2020

Improve. Q^{target}

For line 3 we use $\|Q(x_t, a_t) - r_t - \gamma \max_{a'} Q^{target}(x_{t+1}, a')\|^2$

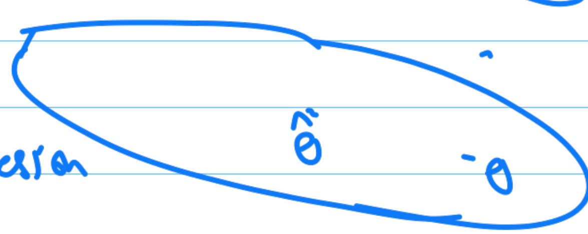
line 5) update $\underline{Q^{target}} = \underline{Q}$ once in a while.

Mnih 2015

In linear bandit → Thompson sampling

$\theta^T \phi(x, a)$ $\qquad \theta \sim \text{Normal}(\underline{\text{mean}}, \underline{\text{cov}})$

Bayesian linear regression



$x \to W\phi(x) \to Q(x, a)$

learn ϕ using gradient descent

learn W using Bayesian linear regression.

For line 1) we do Thompson sampling

3) $\|(W_{mean}\phi(x_t))_{a_t} - r_t - \max_{a'} Q^{target}(x_{t+1}, a')\|^2$

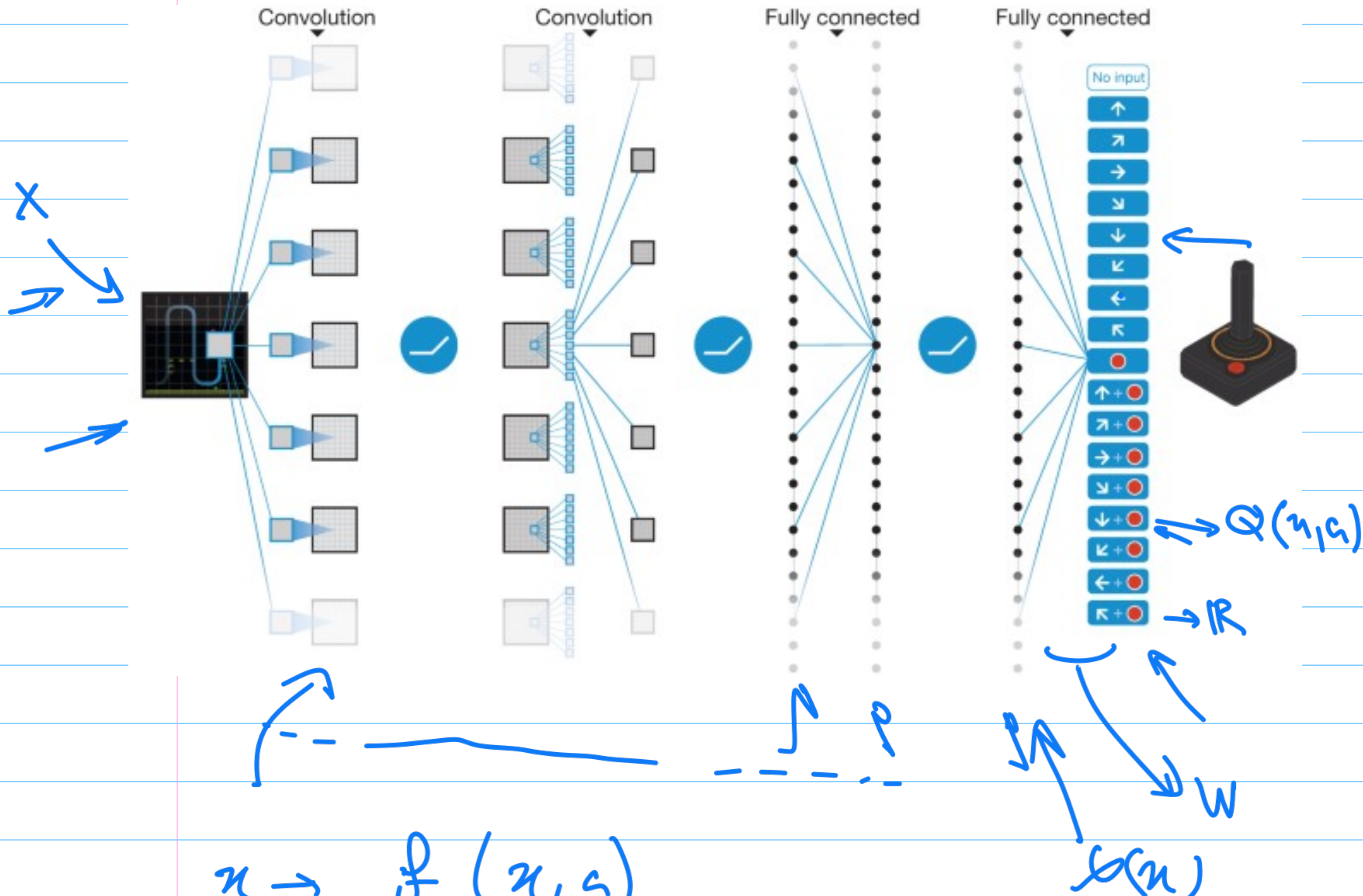
learn ϕ

Wednesday, December 2, 2020

This is called Bayessian DQN (BDQN)

- Measure theory
- Bandits.
- MDP
- POMDP
- Control setting
- Model based
- Model free value based
- Model free policy based
- Competetion optimization
- Deep RL
- off policy learning

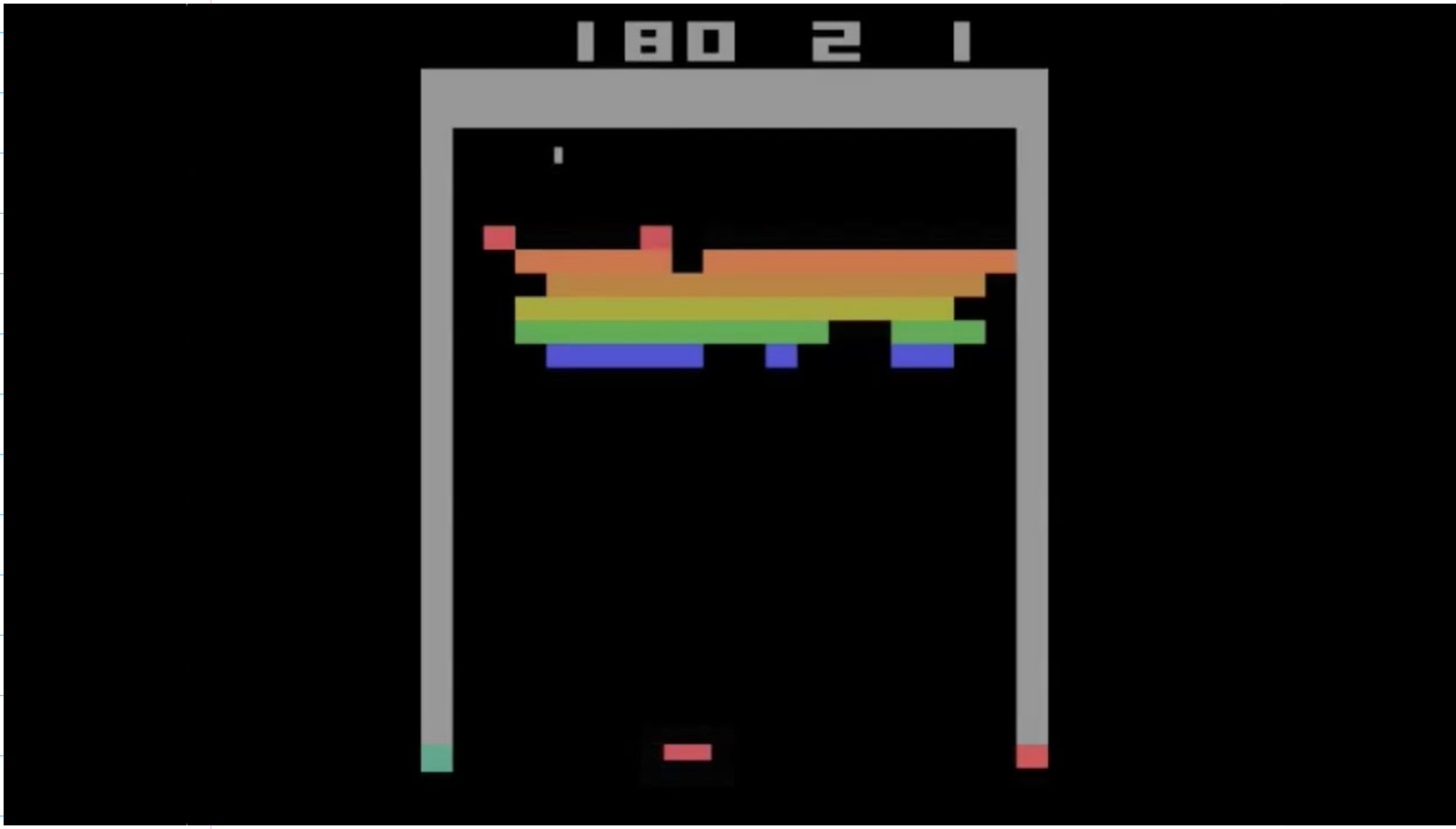
Wednesday, December 2, 2020



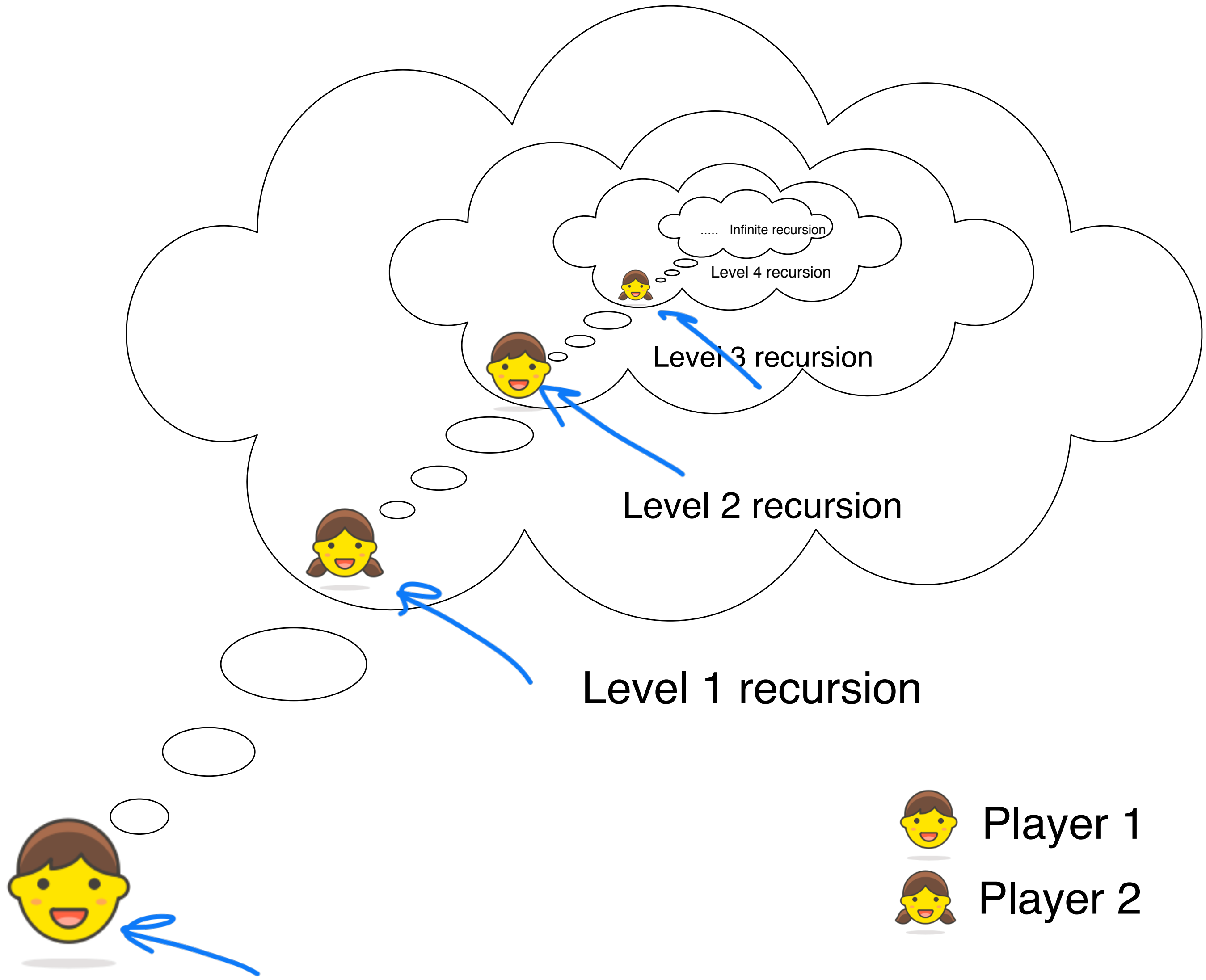
$$x \to f_\theta(x,a)$$

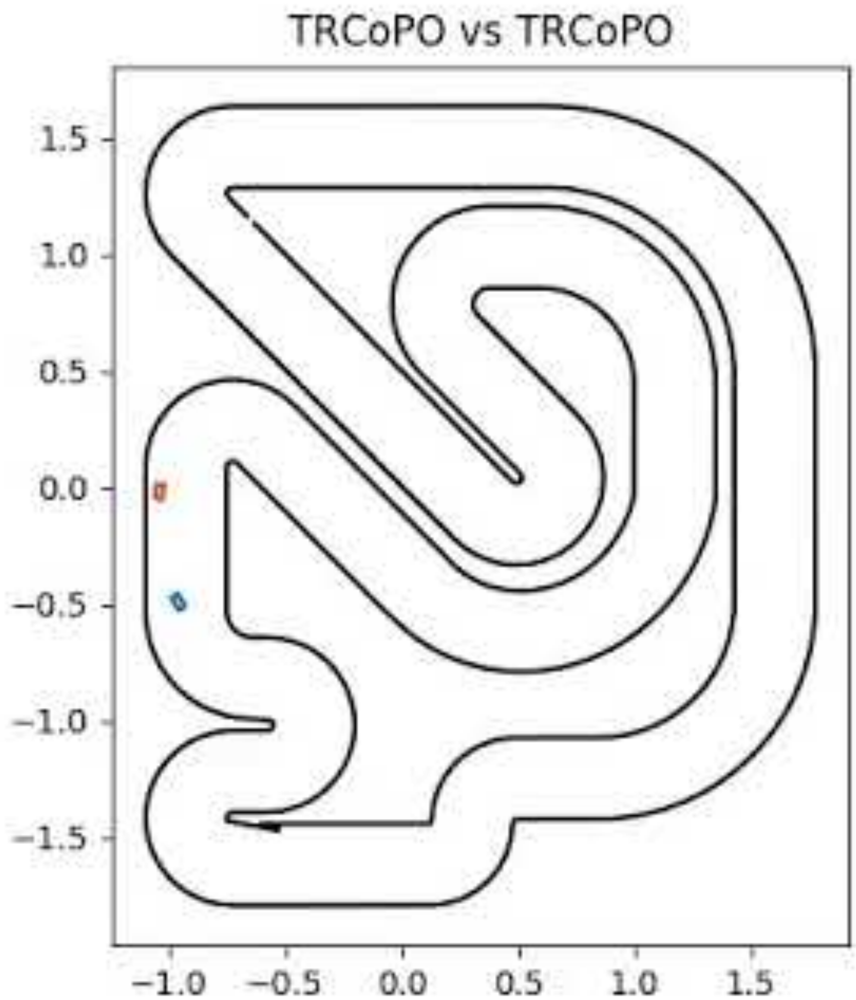
$$f_\theta(x,a) = \left(W^T \phi(x)\right)_a$$

Wednesday, December 2, 2020



Wednesday, December 2, 2020




TRCoPO vs TRCoPO
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5