Deep Reinforcement Learning

[Human-Level Control through deep reinforcement learning, Nature 2015]

CS 486/686
University of Waterloo
Lecture 20: July 10, 2017

Outline

- Value Function Approximation
  - Linear approximation
  - Neural network approximation
    - Deep Q-network

Quick recap

- Markov Decision Processes: value iteration

$$V(s) \leftarrow \max_a R(s) + \gamma \sum_{s'} \Pr(s'|s,a)\, V(s')$$

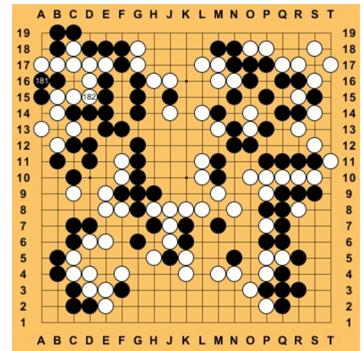
- Reinforcement Learning: Q-Learning

$$Q(s,a) \leftarrow Q(s,a) + \alpha[R(s) + \gamma \max_{a'} Q(s',a') - Q(s,a)]$$

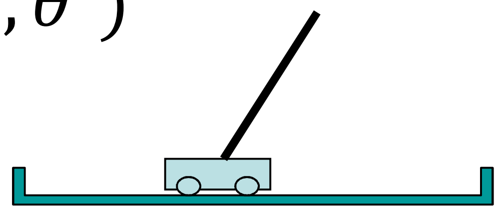
- Complexity depends on number of states and actions

Large State Spaces

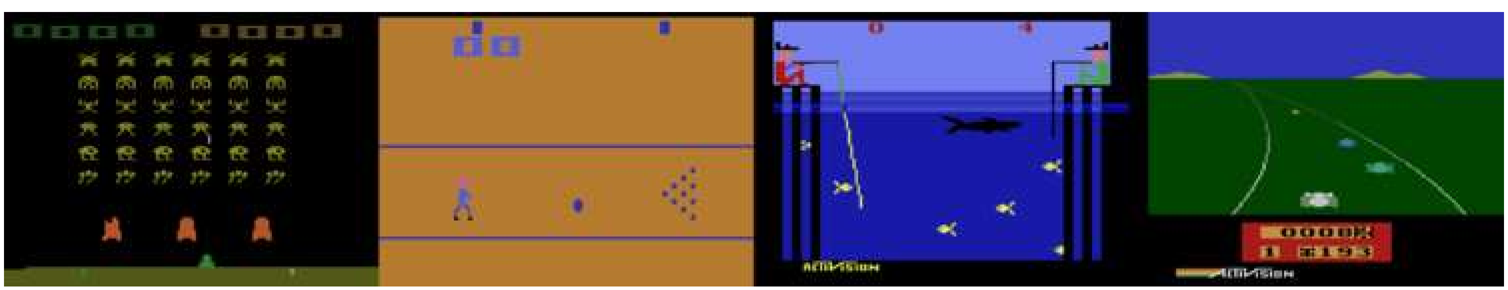
- Computer Go: 3^{361} states



- Inverted pendulum: (x, x', θ, θ')
  - 4-dimensional continuous state space



- Atari: 210x160x3 dimensions (pixel values)



Functions to be Approximated

- Policy: $\delta(s) \rightarrow a$
- Q-function: $Q(s, a) \in \Re$
- Value function: $V(s) \in \Re$

Q-function Approximation

- Let $s = (x_1, x_2, \dots, x_n)^T$

- Linear

  $Q(s,a) \approx \sum_i w_{ai} x_i$

- Non-linear (e.g., neural network)

  $Q(s,a) \approx g(\boldsymbol{x}; \boldsymbol{w})$

Gradient Q-learning

- Minimize squared error between Q-value estimate and target
  - Q-value estimate: $Q_{\boldsymbol{w}}(s,a)$
  - Target: $R(s) + \gamma \max_{a'} Q_{\overline{\boldsymbol{w}}}(s',a')$

$\overline{\boldsymbol{w}}$ fixed

- Squared error:

$$Err(\boldsymbol{w}) = \frac{1}{2}\left[Q_{\boldsymbol{w}}(s,a) - R(s) - \gamma \max_{a'} Q_{\overline{\boldsymbol{w}}}(s',a')\right]^2$$

- Gradient

$$\frac{\partial Err}{\partial \boldsymbol{w}} = \left[Q_{\boldsymbol{w}}(s,a) - R(s) - \gamma \max_{a'} Q_{\boldsymbol{w}}(s',a')\right] \frac{\partial Q_{\boldsymbol{w}}(s,a)}{\partial \boldsymbol{w}}$$

Gradient Q-learning

Initialize weights $\boldsymbol{w}$ at random in $[-1,1]$

Observe current state s

Loop

Select action a and execute it

Receive immediate reward r

Observe new state s'

Gradient: $\frac{\partial Err}{\partial \boldsymbol{w}} = \left[Q_{\boldsymbol{w}}(s,a) - r - \gamma \max_{a'} Q_{\boldsymbol{w}}(s',a')\right] \frac{\partial Q_{\boldsymbol{w}}(s,a)}{\partial \boldsymbol{w}}$

Update weights: $\boldsymbol{w} \leftarrow \boldsymbol{w} - \alpha \frac{\partial Err}{\partial \boldsymbol{w}}$

Update state: $s \leftarrow s'$

Recap: Convergence of Tabular Q-learning

- Tabular Q-Learning converges to optimal Q-function under the following conditions:

  $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$

- Let $\alpha(s,a) = 1/N(s,a)$
  - Where $N(s,a)$ is # of times that (s,a) is visited

- Q-learning

$$Q(s,a) \leftarrow Q(s,a) + \alpha(s,a)[R(s) + \gamma \max_{a'} Q(s',a') - Q(s,a)]$$

Convergence of Linear Gradient Q-Learning

- Linear Q-Learning converges under the same conditions:

$$\sum_{t=0}^{\infty} \alpha_t = \infty \text{ and } \sum_{t=0}^{\infty} \alpha_t^2 < \infty$$

- Let $\alpha_t = 1/t$
- Let $Q_{\boldsymbol{w}}(s,a) = \sum_i w_i x_i$

- Q-learning

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + \alpha_t \left[Q_{\boldsymbol{w}}(s,a) - r - \gamma \max_{a'} Q_{\boldsymbol{w}}(s',a') \right] \frac{\partial Q_{\boldsymbol{w}}(s,a)}{\partial \boldsymbol{w}}$$

Divergence of non-linear Q-learning

- Even when the following conditions hold

  $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$

  non-linear Q-learning may diverge

- Intuition:
  - Adjusting $\boldsymbol{w}$ to increase Q at (s, a) might introduce errors at nearby state-action pairs.

Mitigating divergence

- Two tricks are often used in practice:

1. Experience replay
2. Use two networks:
   - Q-network
   - Target network

Experience Replay

- Idea: store previous experiences (s, a, s', r) into a buffer and sample a mini-batch of previous experiences at each step to learn by Q-learning

- Advantages
  - Break correlations between successive updates (more stable learning)
  - Fewer interactions with environment needed to converge (greater data efficiency)

Target Network

- Idea: Use a separate target network that is updated only periodically

repeat for each (s, a, s', r) in mini-batch:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + \alpha_t \Big[\underbrace{Q_{\boldsymbol{w}}(s,a)}_{\text{update}} - r - \gamma \max_{a'} \underbrace{Q_{\overline{\boldsymbol{w}}}(s',a')}_{\text{target}}\Big] \frac{\partial Q_{\boldsymbol{w}}(s,a)}{\partial \boldsymbol{w}}$$

$$\overline{\boldsymbol{w}} \leftarrow \boldsymbol{w}$$

- Advantage: mitigate divergence

Target Network

- Similar to value iteration:

repeat for all s

$$\underbrace{V(s)}_{\text{update}} \leftarrow \max_a R(s) + \gamma \sum_{s'} \Pr(s'|s,a) \underbrace{\bar{V}(s')}_{\text{target}} \quad \forall s$$

$\bar{V} \leftarrow V$

repeat for each (s,a,s',r) in mini-batch:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + \alpha_t [\underbrace{Q_{\boldsymbol{w}}(s,a)}_{\text{update}} - r - \gamma \max_{a'} \underbrace{Q_{\bar{\boldsymbol{w}}}(s',a')}_{\text{target}}] \frac{\partial Q_{\boldsymbol{w}}(s,a)}{\partial \boldsymbol{w}}$$

$\bar{\boldsymbol{w}} \leftarrow \boldsymbol{w}$

Deep Q-network

- Google Deep Mind:

- Deep Q-network: Gradient Q-learning with
  - Deep neural networks
  - Experience replay
  - Target network

- Breakthrough: human-level play in many Atari video games

Deep Q-network

Initialize weights $\boldsymbol{w}$ and $\overline{\boldsymbol{w}}$ at random in $[-1,1]$

Observe current state s

Loop

Select action a and execute it

Receive immediate reward r

Observe new state s'

Add (s, a, s', r) to experience buffer

Sample mini-batch of experiences from buffer

For each experience $(\hat{s}, \hat{a}, \hat{s}', \hat{r})$ in mini-batch

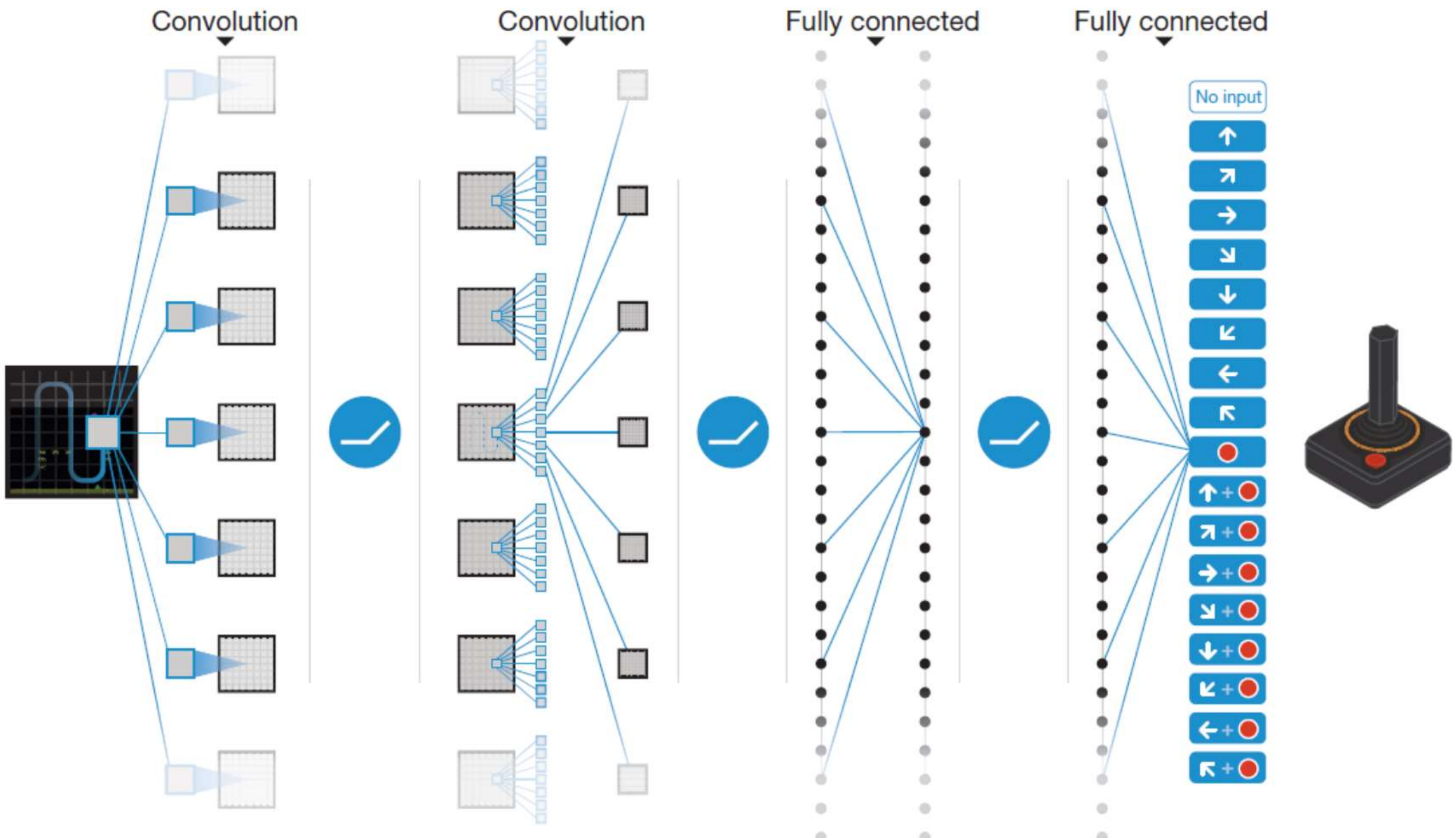
Gradient: $\frac{\partial Err}{\partial \boldsymbol{w}} = \left[Q_{\boldsymbol{w}}(\hat{s}, \hat{a}) - r - \gamma \max_{\hat{a}'} Q_{\overline{\mathbf{w}}}(\hat{s}', \hat{a}')\right] \frac{\partial Q_{\boldsymbol{w}}(\hat{s},\hat{a})}{\partial \boldsymbol{w}}$

Update weights: $\boldsymbol{w} \leftarrow \boldsymbol{w} - \alpha \frac{\partial Err}{\partial \boldsymbol{w}}$

Update state: $s \leftarrow s'$

Every c steps, update target: $\overline{\boldsymbol{w}} \leftarrow \boldsymbol{w}$

Deep Q-Network for Atari



DQN versus Linear approx.