CSCI 1470/2470
Spring 2024
Ritambhara Singh
April 19, 2024
Friday

Deep Learning

Reinforcement Learning : Q-Learning
Review: Q-value and V-value Tables (made up)

<table>
<thead>
<tr>
<th>State #1</th>
<th>Action #1</th>
<th>Action #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>State #2</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>State #3</td>
<td>-1</td>
<td>-10</td>
</tr>
<tr>
<td>State #4</td>
<td>0</td>
<td>1.9</td>
</tr>
<tr>
<td>State #5</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>State #6</td>
<td>-10</td>
<td>-10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>State</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>State #1</td>
<td>0</td>
</tr>
<tr>
<td>State #2</td>
<td>1</td>
</tr>
<tr>
<td>State #3</td>
<td>-1</td>
</tr>
<tr>
<td>State #4</td>
<td>1.9</td>
</tr>
<tr>
<td>State #5</td>
<td>10</td>
</tr>
<tr>
<td>State #6</td>
<td>-10</td>
</tr>
</tbody>
</table>
Review: Value iteration pseudocode

1. For all $s$, set $V(s) := 0$.
2. Repeat until convergence:
   1. For all $s$:
      1. For all $a$, set $Q(s, a) := \sum_{s' \in S} T(s, a, s')[R(s, a, s') + \gamma V(s')]$
      2. $V(s) := \max_a Q(s, a)$
3. Return $Q$
Review: Frozen Lake – final value table & optimal policy

Environment

Example Final Value Table

Final Policy
## Organizing RL problems/algorithms

<table>
<thead>
<tr>
<th>Simple/discrete</th>
<th>Complex/continuous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Know $T$ and $R$</td>
<td>Don’t know $T$ and $R$</td>
</tr>
<tr>
<td>Value iteration</td>
<td>Q-Learning</td>
</tr>
<tr>
<td>Deep Q-Networks</td>
<td>REINFORCE</td>
</tr>
<tr>
<td>Actor-Critic</td>
<td></td>
</tr>
</tbody>
</table>

For a more complete taxonomy of RL algorithms, see [https://spinningup.openai.com/en/latest/spinningup/rl_intro2.html#citations-below](https://spinningup.openai.com/en/latest/spinningup/rl_intro2.html#citations-below)
Tabular Q-learning
Motivation: Why Not Value Iteration?

In value iteration, we assume that we know the transition and reward functions, but what if this isn’t the case?

\[ \{ S, A, R, T, \gamma \} \]

How can we learn in this scenario?
**Examples of Unknown T and/or R:**

**Self-Driving Cars**

Can’t design T for a self-driving car. Too complex to model directly

**Frozen Lake**

What if we don’t know how slippery the lake is? (i.e. T is unknown)

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>H</td>
<td>F</td>
<td>H</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>H</td>
</tr>
<tr>
<td>H</td>
<td>F</td>
<td>F</td>
<td>G</td>
</tr>
</tbody>
</table>
First Attempt

Start in Frozen Lake with no knowledge of T or R

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>F</th>
<th>F</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>H</td>
<td>F</td>
<td>F</td>
<td>H</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>H</td>
</tr>
<tr>
<td>H</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>G</td>
</tr>
</tbody>
</table>

Take random actions to estimate T and R

Run value iteration using estimates of T and R

\[ Q(s, a) = \sum_{s'} T(s, a, s') R(s, a, s') + \gamma V(s') \]

\[ V(s) = \arg\max_a (Q(s, a)) \]
OpenAI Gym ([https://gym.openai.com/](https://gym.openai.com/))

Gym

Gym is a toolkit for developing and comparing reinforcement learning algorithms. It supports teaching agents everything from walking to playing games like Pong or Pinball.

View documentation ›
View on GitHub ›
Farama Foundation
(https://gymnasium.farama.org)
Frozen Lake in OpenAI Gym

This environment is part of the Toy Text environments. Please read that page first for general information.

<table>
<thead>
<tr>
<th>Action Space</th>
<th>Discrete(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observation Space</td>
<td>Discrete(16)</td>
</tr>
<tr>
<td>Import</td>
<td>gym.make(&quot;FrozenLake-v1&quot;)</td>
</tr>
</tbody>
</table>
Frozen Lake ‘Wandering’ Demo

https://colab.research.google.com/drive/1yBCDzAXlu9j0A2aT8fH1zm7beBgTB9yY#scrollTo=sI_x2TsCp15L
Problems with This Approach

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>F</th>
<th>F</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>32709</td>
<td>13151</td>
<td>5951</td>
<td>2868</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>H</th>
<th>F</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12335</td>
<td>7095</td>
<td>1760</td>
<td>1141</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>F</th>
<th>F</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4646</td>
<td>1579</td>
<td>1015</td>
<td>252</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>H</th>
<th>F</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1380</td>
<td>643</td>
<td>501</td>
<td>132</td>
</tr>
</tbody>
</table>

Values are the number of times each state was visited over 10000 episodes run in OpenAI Gym.

We’re extremely unlikely to reach the goal state through random wandering, so our estimates of T and R will probably be bad.

So, what is the issue here?
How to Improve?

We can interleave policy improvement with wandering (Get better at exploring *by exploring*)

But how do we improve our policy?
Q-Learning

- Every time we take an action, use the observed reward to update our estimate of $Q$
- $V(s)$ is still $\max_a Q(s, a)$, so it only matters how we update $Q$
- Importantly, we use our estimates of $Q$ at each step to pick what we think is the best action, instead of just moving randomly
  - action = $\arg\max_a Q(s, a)$

$$Q^\pi(s, a) = E[R(s, a, s') + \gamma V^\pi(s')]$$
$$Q^\pi(s, a) = \sum_{s' \in S} P(s'|s, a)[R(s, a, s') + \gamma V^\pi(s')]$$
Q-Learning

- Every time we take an action, use the observed reward to update our estimate of $Q$
- $V(s)$ is still $\max_a Q(s, a)$, so it only matters how we update $Q$
- Importantly, we use our estimates of $Q$ at each step to pick what we think is the best action, instead of just moving randomly
  - action = $\arg\max_a Q(s, a)$

\[
Q^\pi(s, a) = E[R(s, a, s') + \gamma V^\pi(s')] \\
Q^\pi(s, a) = \sum_{s' \in S} P(s'|s, a)[R(s, a, s') + \gamma V^\pi(s')] 
\]
Q-Learning

- Every time we take an action, use the observed reward to update our estimate of \( Q \)
- \( V(s) \) is still \( \max_a Q(s, a) \), so it only matters how we update \( Q \)
- Importantly, we use our estimates of \( Q \) at each step to pick what we think is the best action, instead of just moving randomly
  - action = \( \arg\max_a Q(s, a) \)

\[
Q^\pi(s, a) = E[R(s, a, s') + \gamma V^\pi(s')] \\
Q^\pi(s, a) = \sum_{s' \in S} P(s'|s, a)[R(s, a, s') + \gamma V^\pi(s')] 
\]
**Q-Learning**

- Every time we take an action, use the observed reward to update our estimate of $Q$
  
- $V(s)$ is still $\max_a Q(s, a)$, so it only matters how we update $Q$

- Importantly, we use our estimates of $Q$ at each step to pick what we think is the best action, instead of just moving randomly
  
  - action $= \arg\max_a Q(s, a)$

$$Q^\pi(s, a) = E[R(s, a, s') + \gamma V^\pi(s')]$$

$$Q^\pi(s, a) = \sum_{s' \in S} P(s'|s, a)[R(s, a, s') + \gamma V^\pi(s')]$$

Use $Q$ to pick action $a_0$. Observe transition and get reward $r_0$. Use $r_0$ to update $Q(s_0, a_0)$ immediately.
Q-Learning

- Every time we take an action, use the observed reward to update our estimate of $Q$
- $V(s)$ is still $\max_a Q(s, a)$, so it only matters how we update $Q$
- Importantly, we use our estimates of $Q$ at each step to pick what we think is the best action, instead of just moving randomly
  - action $= \arg\max_a Q(s, a)$

$Q^\pi(s, a) = E[R(s, a, s') + \gamma V^\pi(s')]$

$Q^\pi(s, a) = \sum_{s' \in S} P(s'|s, a)[R(s, a, s') + \gamma V^\pi(s')]$
How do we Update Q?

**Basic Strategy:**
Take a weighted average of our old Q estimate with our new Q estimate

\[ Q(s, a) = (1 - \alpha)Q_{old}(s, a) + \alpha Q_{new}(s, a) \]

Where the hyperparameter \( \alpha \) controls how quickly we learn

But what should \( Q_{new}(s, a) \) be?
Determining $Q_{new}(s, a)$

In value iteration, we took an expectation over all possible next states and actions to get a new estimate for $Q(s, a)$

$$Q(s, a) = \sum_{s'} T(s, a, s')(R(s, a, s') + \gamma V(s'))$$

In Q-Learning, we only take one action and see one new state, and use that one transition to update our estimate for $Q(s, a)$, so our update rule becomes

$$Q(s, a) = R(s, a, s') + \gamma V(s')$$

i.e. the reward we observed for moving into state $s'$, plus whatever our current value function estimate thinks is the value of being in state $s'$

How do we change this?
The Q-Learning Update Rule

Combining our new estimate for $Q(s, a)$ with the weighted average equation, the final update rule for Q-Learning becomes

$$Q(s, a) = (1 - \alpha)Q_{old}(s, a) + \alpha Q_{new}(s, a)$$

$$Q(s, a) = (1 - \alpha)Q(s, a) + \alpha(R(s, a, s') + \gamma V(s'))$$
Q-Learning in Frozen Lake

- Initially, the agent is acting randomly, because all states have a value of zero.
  - It would likely fall into the holes many times before reaching the goal state for the first time.

\[ \alpha = .5, \gamma = .99 \]

- Values in the grid are $Q(s,a)$.
- Blank quadrants have a Q-value of 0.
Q-Learning in Frozen Lake

- Initially, the agent is acting randomly, because all states have a value of zero
  - It would likely fall into the holes many times before reaching the goal state for the first time, but for this example assume it got lucky
- At every transition in this path(each <s,a,r,s’> tuple), the agent sees a reward of zero up until the final one, so none of the Q-estimates change except the state before the goal
  - $Q(s, a) = (1 - \alpha)0 + \alpha(0) = 0$

- $\alpha = .5, \gamma = .99$
- Values in the grid are $Q(s,a)$
- Blank quadrants have a Q-value of 0
Q-Learning in Frozen Lake

Initially, the agent is acting randomly, because all states have a value of zero
- It would likely fall into the holes many times before reaching the goal state for the first time, but for this example assume it got lucky
- At every transition in this path (each <s,a,r,s’> tuple), the agent sees a reward of zero up until the final one, so none of the Q-estimates change except the state before the goal
  - $Q(s, a) = (1 - \alpha)0 + \alpha(0) = 0$
- The last transition has a positive reward though, so the previous state action pair is changed
  - $Q(s, a) = (1 - \alpha)0 + \alpha(1) = .5$

• $\alpha = .5, \gamma = .99$
• Values in the grid are Q(s,a)
• Blank quadrants have a Q-value of 0
Q-Learning in Frozen Lake

- In another episode, the agent will still act randomly (since most values are still zero), unless it reaches the state next to the goal, which now has a positive value.
  - Assume it got lucky once again

- $\alpha = .5, \gamma = .99$
- Values in the grid are $Q(s,a)$
- Blank quadrants have a $Q$-value of 0
Q-Learning in Frozen Lake

• In another episode, the agent will still act randomly (since most values are still zero), unless it reaches the state next to the goal, which now has a positive value.
  • Assume it got lucky once again
  • When that happens, the previous state-action pair will be updated using the value of the state next to the goal
    • \( Q(s, a) = (1 - \alpha)0 + \alpha(0 + \gamma \times .5) = .5 \times .5 \times .99 \)

• \( \alpha = .5, \gamma = .99 \)
• Values in the grid are \( Q(s, a) \)
• Blank quadrants have a Q-value of 0
In another episode, the agent will still act randomly (since most values are still zero), unless it reaches the state next to the goal, which now has a positive value.

- Assume it got lucky once again
- When that happens, the previous state-action pair will be updated using the value of the state next to the goal
  - \( Q(s, a) = (1 - \alpha)0 + \alpha(0 + \gamma \times .5) = .5 \times .5 \times .99 \)
- When choosing its next action from the state next to the goal, the agent will choose \( \arg \max_a Q(s, a) \), picking the action corresponding to the value of .5 (since all the other action-values for that state are zero).

\( \alpha = .5, \gamma = .99 \)

Values in the grid are \( Q(s, a) \)

Blank quadrants have a Q-value of 0
In another episode, the agent will still act randomly (since most values are still zero), unless it reaches the state next to the goal, which now has a positive value.

- Assume it got lucky once again
- When that happens, the previous state-action pair will be updated using the value of the state next to the goal
  - \( Q(s, a) = (1 - \alpha)0 + \alpha(0 + \gamma \times 0.5) = 0.5 \times 0.5 \times 0.99 \)
- When choosing its next action from the state next to the goal, the agent will choose \( \text{argmax}_a Q(s, a) \), picking the action corresponding to the value of 0.5 (since all the other action-values for that state are zero).
- The value of the previous state is then updated again
  - \( Q(s, a) = (1 - \alpha)0.5 + \alpha(1) = 0.75 \)

\[ \alpha = 0.5, \gamma = 0.99 \]

- Values in the grid are \( Q(s, a) \)
- Blank quadrants have a \( Q \)-value of 0
In another episode, the agent will still act randomly (since most values are still zero), unless it reaches the state next to the goal, which now has a positive value.

- Assume it got lucky once again

- When that happens, the previous state-action pair will be updated using the value of the state next to the goal
  \[ Q(s, a) = (1 - \alpha)0 + \alpha(0 + \gamma \times .5) = .5 \times .5 \times .99 \]

- When choosing its next action from the state next to the goal, the agent will choose argmax_a Q(s, a), picking the action corresponding to the value of .5 (since all the other action-values for that state are zero).

- The value of the previous state is then updated again
  \[ Q(s, a) = (1 - \alpha) .5 + \alpha (1) = .75 \]

- This process repeats over and over again until the Q-values converge to \( Q^*(s, a) \), the optimal Q-values.

- \( \alpha = .5, \gamma = .99 \)
- Values in the grid are Q(s,a)
- Blank quadrants have a Q-value of 0
Problem: Exploration/Exploitation

• But what if Frozen Lake had two goal states?
Problem: Exploration/Exploitation

- But what if Frozen Lake had two goal states?
  - Q-Learning could learn a path to the .5 reward without ever getting to the 1 reward state.
  - Converges to a suboptimal solution
Problem: Exploration/Exploitation

• But what if Frozen Lake had two goal states?
  • Q-Learning could learn a path to the .5 reward without ever getting to the 1 reward state.
  • Converges to a suboptimal solution
• How can we balance exploiting knowledge we already have with exploring unseen parts of the state space?

Any ideas?
Solution: Epsilon-Greedy Policies

- Instead of always following our estimate of Q, we instead take random actions $\epsilon$ percent of the time, where $\epsilon$ is a hyperparameter.
- We can also decrease $\epsilon$ over time, as our estimates of Q improve:
  - Ex: after each episode, set $\epsilon = \frac{E}{(E+i)}$, where $i$ is the number of episodes experienced so far.
  - This lets the agent act less randomly over time.
Q-Learning Update in Code

```python
if np.random.rand(1) < epsilon:
    act = env.action_space.sample()
else:
    act = np.argmax(Q[st])

nst, rwd, done, _ = env.step(act)
Q[st][act] = (1-alpha)Q[st][act] + alpha(rwd +
            gamma*V[nst])
```

Any questions?
Q-Learning Code Demo

https://colab.research.google.com/drive/1yBCDzAXlu9j0A2aT8fH1zm7beBgtB9yY#scrollTo=sI_x2TsCp15L
Organizing RL problems/algorithms

Know $T$ and $R$

Simple/discrete
- Value iteration

Complex/continuous
- Deep Q-Networks

Don’t know $T$ and $R$

- Q-Learning
- REINFORCE
- Actor-Critic

For a more complete taxonomy of RL algorithms, see https://spinningup.openai.com/en/latest/spinningup/rl_intro2.html#citations-below
Deep Q-learning
Limitations of Tabular Q-Learning

- Tabular methods can’t handle large or continuous state spaces
  - Can’t have a table for infinite Q or V values
- Lots of problems have these kinds of state spaces
  - Robotic navigation: The state of the robot could be (map of environment, position), where position is a 2-D vector. Infinite possible states in this setup.
  - Go: has $\sim10^{170}$ states

Too many possible values for a table
Beyond Tabular Learning

In Q-learning, we are learning a function $Q$ that maps from a state-action pair to a real number

$$Q: (s, a) \rightarrow \mathbb{R}$$

Or, equivalently, from a state to a vector of real numbers

$$Q: (s) \rightarrow \mathbb{R}^{|a|}$$

Instead of storing every $(s, a)$ in a table to learn this function, we can learn a function to approximate $Q$ using a (relatively) small set of parameters $\theta$

$$Q: (s, a, \theta) \rightarrow \mathbb{R}, \theta \ll |S \times A|$$
Neural networks are excellent function approximators, so we can use a deep network to learn this parameterized function.
Frozen Lake Example

How do we set up this using neural networks?
Frozen Lake Example

• Feed a one-hot representation of the state into a neural network to approximate the Q-value for each action
  • One-hot vector of length 16 for frozen lake
• Can also pass in a more complex state to the Q-network rather than a one hot representation.
  • Useful when the number of states is very large, making a one hot representation impractical
  • Or when states are continuous
TF Example Code for Frozen Lake

# Weights for Q-Network
Q = tf.Variable(tf.random.uniform([16,4],0,0.01))

# Q-value function
def qVals(inptSt):
    oneH = tf.one_hot(inptSt,16)
    qVals = tf.matmul([oneH],Q)
    return qVals

# argmax over q-values to get estimated best action
action = tf.argmax(qVals(st),1)
Atari Example

• 81x81 Images of the game
  • Even if each pixel is just on or off, that’s $2^{81\times81}$ possible states, way too many for a table
  • They’re actually colored, so real scenario is even worse
  • Need a different approach to learn Q and V values for large state spaces like these

https://www.youtube.com/watch?v=fnFfPlpjdgg
How To Train This Q Network?

The original Q-learning update is not a minimization problem

\[ Q(s, a) = (1 - \alpha)Q(s, a) + \alpha(R(s, a, s') + \gamma V(s')) \]

So how can we transform this into a loss function we can use?
Recap

Tabular Q-learning

- Wandering to estimate T and R
- Q-learning: Explore + Improve
- Epsilon-Greedy Policies
- Limitations of Tabular Q-learning

Deep Q-learning

- Neural nets for Q approximation
- Tensorflow code

For more reading: https://huggingface.co/learn/deep-rl-course/unit2/introduction?fw=pt