AIL 722: Reinforcement Learning

Lecture 24: Towards Deep Q-Learning

Raunak Bhattacharyya

ScAI YARDI SCHOOL OF ARTIFICIAL INTELLIGENCE
INDIAN INSTITUTE OF TECHNOLOGY DELHI

Outline

- From fitted Q-iteration to Online Q-Learning
- Exploration rules: Intro
- Assignment 2 overview

Arriving Here

- Bootstrap and sample the expected return
- Off-policy algorithm
- Q-Learning

- Approximate the value function
- Fitted VI was not model-free
- Fitted QI

Fitted QI

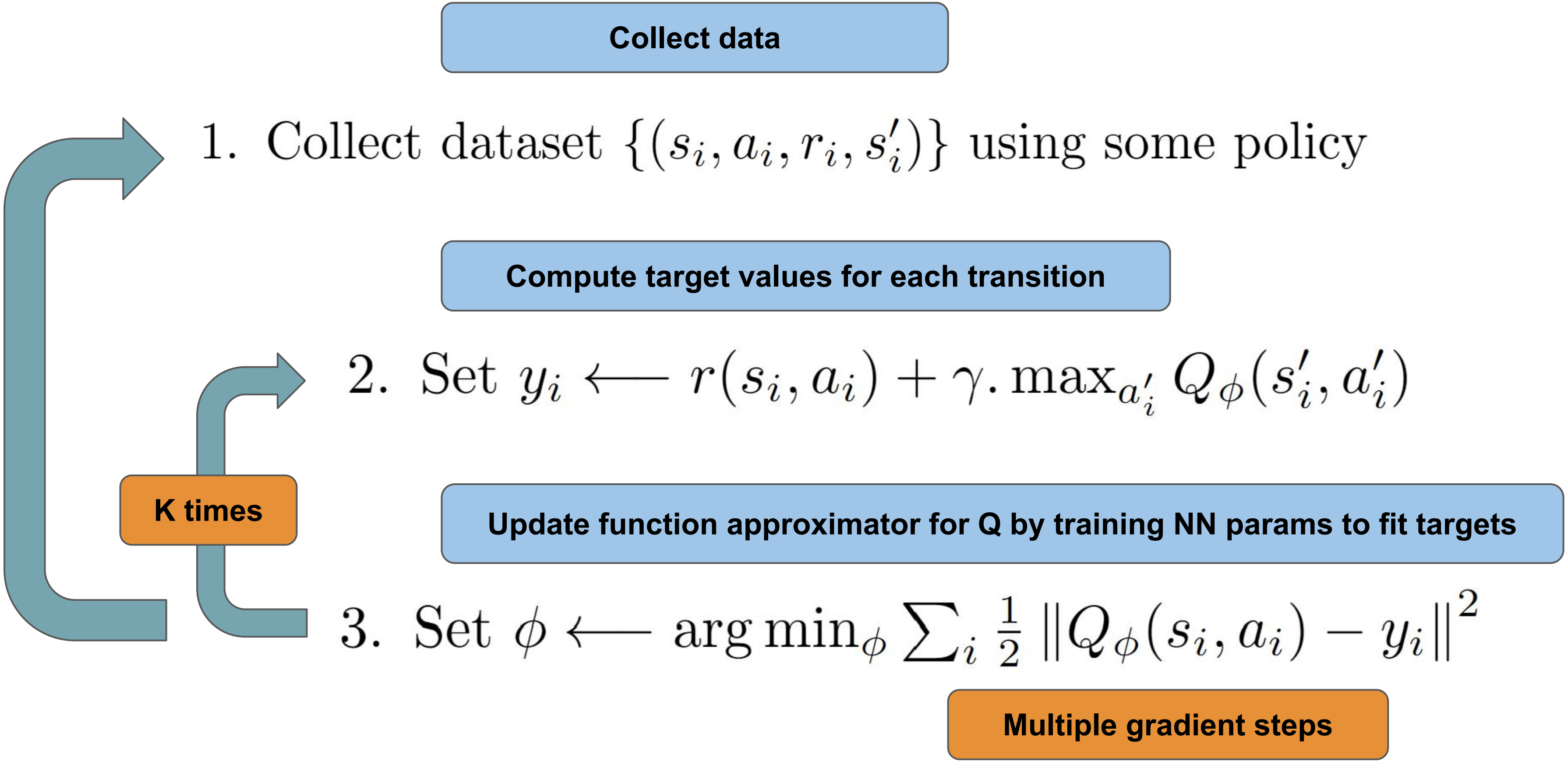
Properties

- Did not require knowledge of transition dynamics

- Did not require explicit representation of policy

- Off policy

Fitted QI



Collect data

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

Compute target values for each transition

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

K times

Update function approximator for Q by training NN params to fit targets

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

Multiple gradient steps

Fitted QI

Properties

- Did not require knowledge of transition dynamics
- Did not require explicit representation of policy
- Off policy

What's the spectrum?

Choices

- How many transitions to collect?
- How many gradient steps to update parameters
- How many times to alternate between new target creation and new function fitting before collecting more data

Completely Online Algorithm

1 step of data collection

1. Take some action a_i and obtain (s_i, a_i, s'_i, r_i)

1 step before collecting more data

2. $y_i = r(s_i, a_i) + \gamma \cdot \max_{a'_i} Q(s'_i, a'_i)$

3. $\phi \longleftarrow \phi - \alpha \cdot \frac{dQ_\phi}{d\phi}(s_i, a_i) \cdot \left(Q_\phi(s_i, a_i) - y_i \right)$

1 gradient step

Look familiar?

Exploration

1. Take some action a_i and obtain (s_i, a_i, s'_i, r_i)

2. $y_i = r(s_i, a_i) + \gamma \cdot \text{max}_{a'_i} Q(s'_i, a'_i)$

3. $\phi \longleftarrow \phi - \alpha \cdot \frac{dQ_\phi}{d\phi}(s_i, a_i) \cdot \left(Q_\phi(s_i, a_i) - y_i \right)$

How do we pick an action in step 1?

Exploration

- Epsilon-greedy
- Boltzmann exploration

Summary & Announcements

- Summary
  - Online Q learning
  - Exploration ideas

- Announcements
  - Midterm marks clarification session
    - Office hours today