20”x30” (landscape)

Minimum font 24pt

Image Titles

1. Avg Online Reward
2. Max Online Reward
3. Eval Batch Reward

Layout: 3 columns

Images in center column

Sections in order listed here

Exploration in Online and Batch Q-Learning

Shayne Miel smiel@stanford.edu

## Questions

* How does varying the exploration strategy affect the online learning process?
* How does the experience gathered by different exploration strategies affect the batch learning process?

## Why

Understanding how different exploration strategies affect batch learning can help us:

* Design better online learning algorithms, because experience replay is similar to batch learning in the limit.
* Design better experience collection strategies for batch learning tasks that cannot be simulated.

## Setup

Atari Breakout

DQN (for both online and batch)

2.5M training steps

## Process

For each exploration strategy:

1. Train and evaluate online DQN.
2. Save experience replay buffer.
3. Train and evaluate batch DQN using saved buffer.

## Exploration Types

### When to explore

Linear (eqn 1)-greedy

Linearly decrease (eqn 1) over the course of 1M training steps.

At each step, (eqn 2)

Shaped (eqn 1)-greedy

Linearly decrease (eqn 1) over the course of 1M training steps.

Take into account how far into the episode you are by tracking average episode length (eqn 3) and current episode step (eqn 4).

For even episodes, (eqn 5)

For odd episodes, (eqn 6)

### How to explore

Randomly

(eqn 7)

Inverse Count

Use locality sensitive hashing to cluster the state observations.

Estimate (eqn 8), the number of times the agent has taken action a in state s.

(eqn 9)

### How much to explore

Regular

Linearly decrease (eqn 1) from 1.0 to 0.1 over the course of 1M steps.

High

Linearly decrease (eqn 1) from 1.0 to 0.5 over the course of 1M steps.

## Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | Avg Reward | Max Reward | Eval Reward | Batch Eval Reward |
| Linear Random | 12.12 ± 0.99 | 46 | 13.34 ± 1.26 | **20.21 ± 1.26** |
| Shaped Random | 12.91 ± 2.21 | 238 | 12.52 ± 1.51 | 14.28 ± 2.29 |
| Linear Count | 11.73 ± 0.98 | 49 | 13.36 ± 0.99 | 19.86 ± 1.37 |
| Shaped Count | **14.88 ± 1.36** | 341 | 13.91 ± 0.78 | 15.90 ± 2.78 |
| Linear Random High | 05.39 ± 0.45 | 29 | 12.61 ± 0.82 | 19.01 ± 1.02 |
| Shaped Random High | 14.53 ± 1.28 | **391** | 13.47 ± 1.26 | 15.18 ± 2.56 |
| Linear Count High | 05.27 ± 0.53 | 26 | 12.52 ± 1.26 | 19.03 ± 1.19 |
| Shaped Count High | 14.82 ± 1.43 | 377 | **14.05 ± 1.17** | 15.55 ± 2.14 |

## Answers

* In the online setting, Shaped (eqn 1)-greedy with Inverse Count action selection is a promising exploration strategy that appears to outperform simple Linear (eqn 1)-greedy with Random action selection.
* In the batch setting, Linear (eqn 1)-greedy with Random action selection is as good as any other method, and better than any of the Shaped (eqn 1)-greedy methods.
* The amount of exploration has a much larger effect on rewards in the online setting than it does on the performance of a batch mode learner using the same experience.

## Next Steps

* Deeper understanding of why Shaped (eqn 1)-greedy exploration leads to suboptimal batch mode learning can help optimize the performance in both online and batch settings.
* Prioritized replay or other intelligent sampling methods to try and take advantage of the occasional high-performing episodes available in Shaped (eqn 1)-greedy rollouts.