AIL 722: Reinforcement Learning

Lecture 38: Continuous Actions, Exploration

Raunak Bhattacharyya

Today’s Outline

- DQN with continuous actions
- Exploration vs. Exploitation
- Formulating exploration

Deep Q-Learning with Continuous Actions

$$\pi'(a_t|s_t) = \begin{cases} 1, & \text{if } a_t = \arg\max_{a_t} A^{\pi}(s_t, a_t) \\ 0, & \text{otherwise} \end{cases}$$

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

How do we find the max when we have continuous actions?

Sampling-based Approaches

$$\max_a Q(s,a) \simeq \max\{Q(s,a_1),...,Q(s,a_N)\}$$

- Derivative free stochastic optimisation

High dimensions. Samples won't be representative enough

Learn Approximate Maximiser

Take some action a_i and observe (s_i, a_i, s'_i, r_i) and add it to $\mathcal{B}$

Sample mini-batch $\{s_j, a_j, s'_j, r_j\}$

Compute $y_j = r_j + \gamma Q_{\phi'}(s'_j, a'_j)$

$\phi \leftarrow \phi - \alpha \Sigma_j \frac{dQ_\phi}{d\phi}(s_j, a_j)(Q_\phi(s_j, a_j) - y_J)$

$\theta \leftarrow \theta + \beta \sum_j \frac{d\mu}{d\theta}(s_j) \frac{dQ_\phi}{da}(s_j, a)$

update ϕ' and θ'

Can also be viewed as deterministic AC algo

Cont control with deep RL, Lillicrap, ICML 2015

Trivia

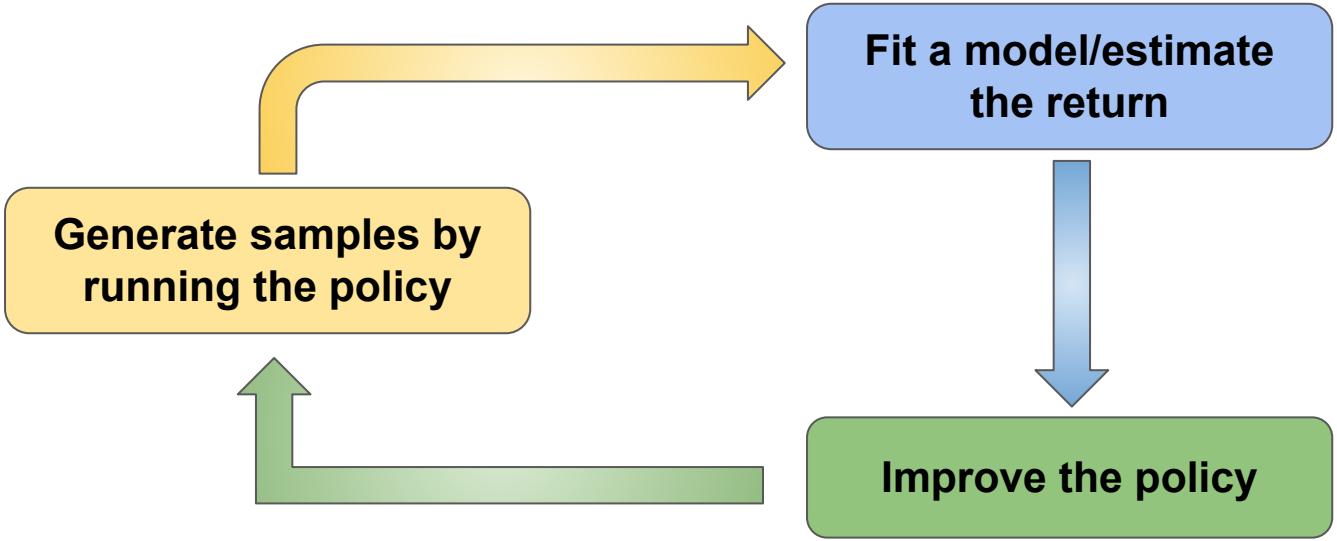
$$\Delta w_{ij} = \alpha_{ij}(r - b_{ij})e_{ij},$$

where α_{ij} is a *learning rate factor*, b_{ij} is a *reinforcement baseline*, and $e_{ij} = \partial \ln g_i / \partial w_{ij}$ is called the *characteristic eligibility* of w_{ij}. Suppose further that the reinforcement baseline b_{ij} is conditionally independent of y_i, given $\mathbf{W}$ and $\mathbf{x}^i$, and the rate factor α_{ij} is nonnegative and depends at most on $\mathbf{w}^i$ and t. (Typically, α_{ij} will be taken to be a constant.) Any learning algorithm having this particular form will be called a *REINFORCE* algorithm. The name is an acronym for "*RE*ward *I*ncrement = *N*onnegative *F*actor × *O*ffset *R*einforcement × *C*haracteristic *E*ligibility," which describes the form of the algorithm.

Simple stat grad following algos for connectionist RL, Williams, ML 1992

Unified View

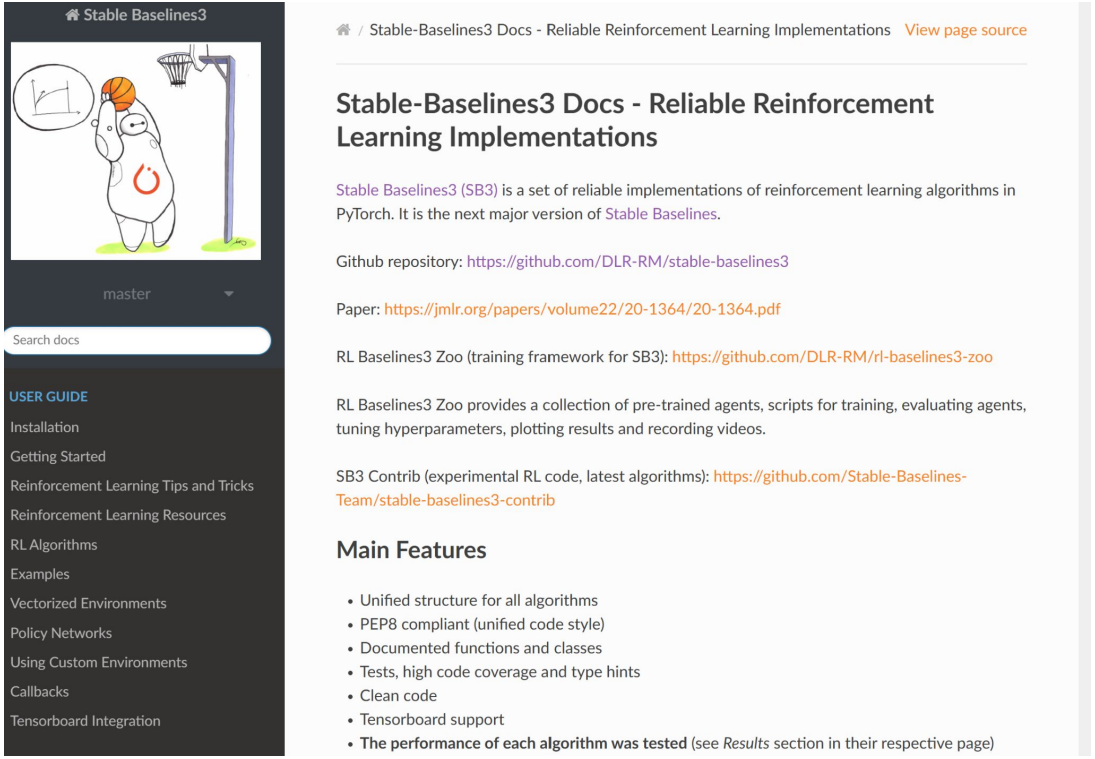
$$Q(s,a) \longleftarrow r(s,a) + \gamma.\mathbb{E}_{p(s'|s,a)}\left[V^{\pi}(s')\right]$$



$$V(s) \longleftarrow \max_a Q(s,a)$$

Adapted from Sergey Levine

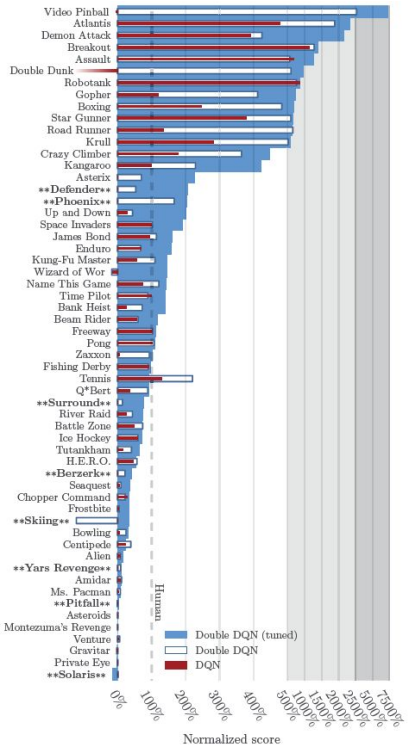
Stable Baselines

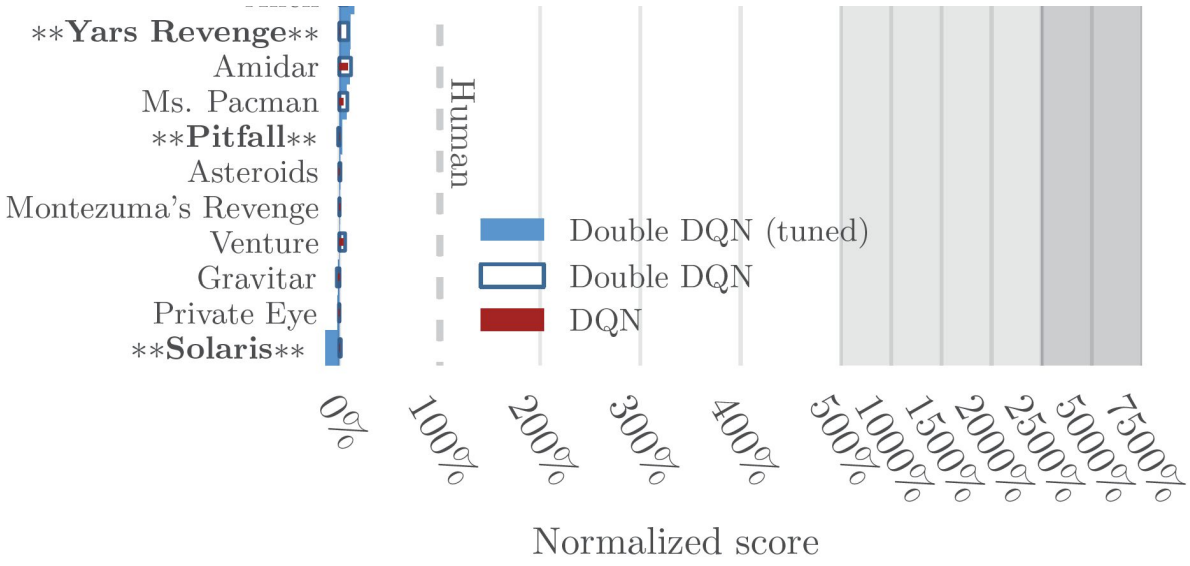


Stable Baselines 3

Exploration

Games where DDQN fails

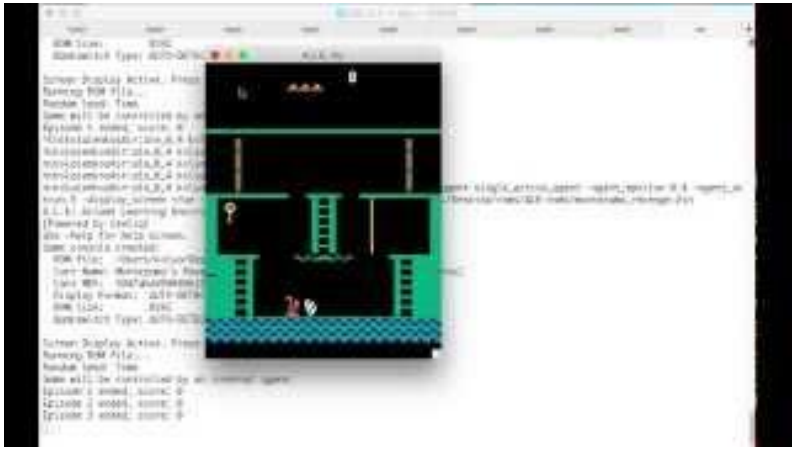




Montezuma’s Revenge, Source: Youtube

Pitfall, Source: Youtube

Venture, Source: Youtube



Breakout vs MR, Source: Youtube

Exploration vs Exploitation

How can an agent decide whether to attempt new behaviors (to discover ones with higher reward) or continue to do the best thing it knows so far?

Exploitation: Do what you think will yield the highest reward

Exploration: Do the things that you haven't done before hoping that it will yield even higher reward

Dynamic and persistent decision we have to keep making

What was the requirement for Q-learning to converge to the optimal state-action value function?

Multi-armed Bandits

One-arm bandit, Source: Youtube

Multi-armed Bandits, Source: Wikipedia