A brief history of Semi Gradient

Semi gradient (Watkins 1989) was developed alongside Q learning

Experiment Design

10 random Seeds

Grid Search

Epsilon greedy, 0.0.01,.05,.3

Discount Factor 0.5,.8,1

DQN 128 hidden nodes

Learning rate .001

Experience Replay 10000

1000 episodes

The DQN parameters were kept fairly constant, as long as we were able to find a solution given the the architecture in one setting, we didn’t feel it would contribute to the comparison of SG vs Full G

Cartpole vs Mountain Car

Why does one method outperform the other. A potential clue is based on the experiment/environment.

What does full gradient mean in terms of how we decide to weight certain actions. It backpropagates the full return for the whole history of actions as opposed to the semi gradient which only backpropagates the future return to our current action.

In an environment setting such as cart pole, where every timestep we receive a positive reward, all the actions up to that point receive that reward. This produces repetitive undiversified behaviour as there is no reason to explore other actions given this action has already given you a positive reward i.e. weight towards that action increases.

In contrast the mountain car problem where per timestep we get a negative reward full gradient fails. Why? Any action we take has a negative reward backpropagated to it, creating a loop/rut where we get stuck and can’t make it out of the valley. In contrast semi gradient doesn’t suffer from this allows us to explore the policy space.

Areas of further exploration

Environments which we have not looked at , include the stochastic versus deterministic setting.

On policy versus off policy

How do they work with weighted sampling/SARSA

However <http://proceedings.mlr.press/v108/tosatto20a/tosatto20a.pdf>

A Nonparametric Off-Policy Policy Gradient

Full grad seems to perform well

or semi-gradient, still converges to the optimal policy in a discrete MDP setting (Degris et al., 2012; Imani et al., 2018).

<http://www.cs.utsa.edu/~bylander/cs6243/baird95residual.pdf>

Full gradient of TD error is not the same as full gradient of Bellman error which is the error we actually want to minimize. But the latter does not have admit an easy unbiased Monte Carlo estimator.

Full gradient of TD will lead to wrong optima in stochastic environments. And isn't any better with stability either in stochastic environments.