Learns models of reward and environment. Builds from bootstrapping and Monte Carlo methods. Uses Monte Carlo scheme for non-stationary environments using a bootstrapped estimate of the return which doesn't require a model of a complete episode.

Bootstrapping uses estimates of successor estimates of state and action values --> predicts future returns at a certain state to estimate the return.

Advantages over both

Monte Carlo and PI/VI have to have episodic tasks

- episodes can be very long, or just not episodic

...over stationary environments

- environments continuously change

New addition:

Temporal difference methods builds from bootstrapping and Monte Carlo methods. It uses the property of not needing to wait until a full episode is truncated through from bootstrapping, and it uses the property of not requiring a pre-specified model from Monte Carlo.

From the Monte Carlo methods, to account for non-stationary environments, it updates the value function estimation from calculating the average of returns (Equation (1)), which is very unsensitive if changes to the property of the environment occurs when the number of states is large:

(1)

To a new value function estimation procedure (Equation (2)):

(2)

Where is the value function, is the return and is some scalar between (0,1) to ensure the algorithm can converge. The can be treated as an error signal, which we want to minimize and this error is what makes the estimation of the value function react to changes in the system property as alpha, which is constant, keeps the value function fluctuating but also keeping to the trend of the optimal value function.

However from equation (2) is still dependent on , where the definition of and so it can be seen that the value function estimation is still reliant on knowing every return of future states, limiting Monte Carlo to be episodic. Therefore, bootstrapping is used to remove the need for future terms Bootstrapping uses the estimate of future returns to estimate the return at a specific state.

Therefore the expected return can be replaced with:

(3)

Where is the estimated value of return, is the reward, is the bootstrap estimate of the value function at the next step.

As such, substituting in equation (3) into (2), we obtain equation (4) which is the temporal difference learning estimate, specifically, TD(0).

(4)

And so the key points are that the allows the temporal difference method to work with non-stationary systems and allows the system to be non-episodic.

Therefore, it is clear an advantage over Monte Carlo methods is that it does not need the environment to be episodic or stationary.

Its advantage over policy and value iteration is that as well as the advantages it achieves over Monte Carlo, it doesn’t need a pre-specified model of the environment, or rewards.