b. Describe the TD(0) algorithm for policy prediction. Why, in general, does this compute a different solution from the Monte Carlo estimate?[10 marks]

TD(0) only looks one step ahead and can converge to some optimal solution given a fixed policy and a constrained step size.

Monte Carlo methods minimize the MSE on training data, whereas the TD(0) optimizes for the maximum-likelhood model of the Markov process.

While both are optimal, they optimize for different conditions. TD(0) estimates the structureof the FMDP (or essentially, estimates a model) and subsequently performs policy evaluation on said model.

New Addition:

TD(0) algorithm for policy prediction looks one step ahead at the next state and its respective reward to update an estimate of at the current state and will converge to some optimal solution given a fixed policy and a constrained step size.

The TD(0) algorithm pseudocode is shown in figure 1.0. Firstly, the constant step-size parameter is defined along with an arbitrary initialisation of for all states except for the terminal state. A nested loop is then run for each ‘episode’ where for each state S, an action defined by a policy is run. The reward and next state is recorded and the value estimate is updated by substituting in all observations into the previously determined equation (4). The algorithm then continues by iterating to the next state .

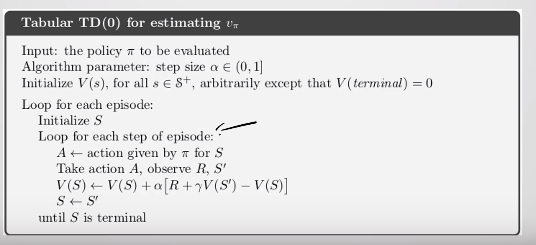


Figure 1.0 – TD(0) algorithm for policy prediction

The TD(0) algorithm can also be further optimized for experiences of finite episodes or if number of episodes needs to be reduced. To do this, batch updating is used where an episode is iterated through multiple times, training data through TD(0) each time and the equation 2 is continually run for the different states. This is repeated for a fixed number of iterations or until convergence and finally the a new value can be estimated from this.

The TD method computes a different solution to the Monte Carlo estimate because the algorithms optimize the converged solution based off different criterion:

* Monte Carlo methods perform well (low/zero error) on existing data as it minimizes the mean squared error on training data.
* TD(0) methods are able to perform better (lower error) than MC methods on future data because it optimizes the maximum-likelihood model of the Markov process. It does this by learning the certainty-equivalent estimate. What this means is that it learns the structure of the FMDP (or essentially, estimates a model) and subsequently performs policy evaluation on said model while assuming the model it has developed is exactly the correct model rather than an approximate.

In practice, TD methods converge faster than MC methods on stochastic tasks