CME241 Assignment11

Quinn Hollister

February 2022

1 MC Prediction

```
def tabular_mc_prediction(
    traces,
    approx_0,
    gamma,
    episode_length_tolerance = 1e-6
):
    ''' Evaluate an MRP using the monte carlo method for tabular MRP's. '''

    episodes: Iterator[Iterator[mp.ReturnStep[S]]] = \
        (returns(trace, gamma, episode_length_tolerance) for trace in traces)
    f = approx_0
    yield f

    count = defaultdict(lambda: 0)

    for episode in episodes:
        for RS in episode:
            s = RS.state
            r = RS.return_
            count[s] += 1
            f[s] += (1/count[s])*(r-f[s])
        yield f
```

2 TD Prediction

```
def tabular_td_prediction(
        transitions,
        approx_0,
        gamma: float
) -> Iterator[ValueFunctionApprox[S]]:
    ''' Evaluate a tabular MRP using TD(0) using the given sequence of transitions '''
    counts = defaultdict(lambda: 0)
    alpha = 0.03
    H = 1000
    beta = 0.5
    def step(v, transition: mp.TransitionStep[S]):
        counts[transition.state] += 1
        alpha_n = alpha / (1 + ((counts[transition.state] - 1)/H)**beta)
        v[transition.state] += (alpha_n)*(transition.reward + \
```

```
        gamma*v[transition.next_state] - v[transition.state])
        return v

    return iterate.accumulate(transitions, step, initial = approx_0)
```

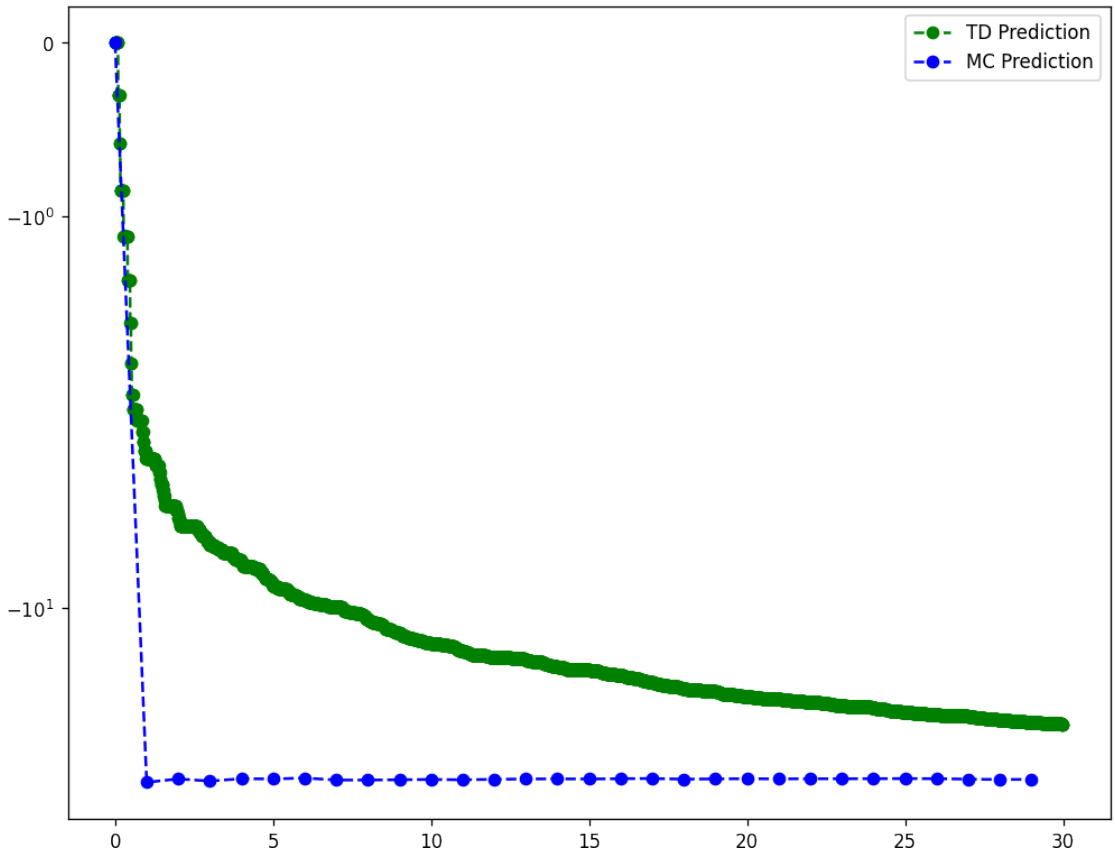
3 Value Function Comparison

True Values:
NonTerminal(state=InventoryState($on_h and = 0, on_o rder = 0$)) : -35.511,
$NonTerminal(state = InventoryState(on_h and = 0, on_o rder = 1))$: -27.932,
$NonTerminal(state = InventoryState(on_h and = 0, on_o rder = 2))$: -28.345,
$NonTerminal(state = InventoryState(on_h and = 1, on_o rder = 0))$: -28.932,
$NonTerminal(state = InventoryState(on_h and = 1, on_o rder = 1))$: -29.345,
$NonTerminal(state = InventoryState(on_h and = 2, on_o rder = 0))$: -30.345

MC Prediction Values:
NonTerminal(state=InventoryState($on_h and = 0, on_o rder = 0$)) : -35.503,
$NonTerminal(state = InventoryState(on_h and = 0, on_o rder = 1))$: -27.923,
$NonTerminal(state = InventoryState(on_h and = 0, on_o rder = 2))$: -28.336,
$NonTerminal(state = InventoryState(on_h and = 1, on_o rder = 0))$: -28.925,
$NonTerminal(state = InventoryState(on_h and = 1, on_o rder = 1))$: -29.339,
$NonTerminal(state = InventoryState(on_h and = 2, on_o rder = 0))$: -30.342

TD Prediction Values:
NonTerminal(state=InventoryState($on_h and = 0, on_o rder = 0$)) : -35.399,
$NonTerminal(state = InventoryState(on_h and = 0, on_o rder = 1))$: -27.942,
$NonTerminal(state = InventoryState(on_h and = 0, on_o rder = 2))$: -28.245,
$NonTerminal(state = InventoryState(on_h and = 1, on_o rder = 0))$: -28.888,
$NonTerminal(state = InventoryState(on_h and = 1, on_o rder = 1))$: -29.207,
$NonTerminal(state = InventoryState(on_h and = 2, on_o rder = 0))$: -30.205

4 Testing on Simple Inventory



We can see how the MC Prediction algorithm pretty much immediately goes to the optimal value function which might appear peculiar at first. However, our state space consists of only 6 non-terminal states, so when we run an episode of length 100, we get plenty of return, state pairs that can be averaged out for just one run. That's why there is only very small perturbations to this number after more runs of MC. TD on the other hand exhibits more of what we expect from the algorithm: good incremental improvements from a steady stream of atomic experiences.