Reinforcement Learning: Model Free Prediction

Earl Wong

Model Free Prediction

- In practice, the MDP may not be known.
- Can the value function still be computed? Yes
- Suppose each rollout / episode was finite:

$$S_0,\ A_0,\ R_0,\ S_1,\ A_1,\ R_1, \ldots,\ S_T,\ A_T,\ R_T \sim \pi$$

- Let the return be written as:

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \ldots + \gamma^T R_{t+T}$$

- Next, recall the expression for the value function:

$$V(s) = E_\pi[G_t \mid S_t = s]$$

Monte Carlo Method

- The value function can be approximated by computing the empirical mean for every state:

$$V(S_t = s) = \frac{1}{N} \sum_{i=1}^{N} G_{i,t}$$

- The value function can also be updated after every rollout / episode.

$$V(S_t = s) \leftarrow V(S_t = s) + \alpha(G_{i,t} - V(S_t = s))$$

Temporal Difference Method

- Can we also perform updates after every step in a specific rollout / episode? Yes
- Recall: $S_0,\ A_0,\ R_0,\ S_1,\ A_1,\ R_1, \ldots,\ S_T,\ A_T,\ R_T \sim \pi$
- The previous equation becomes:

$$V(S_t = s) \leftarrow V(S_t = s) + \alpha(R_t + \gamma V(S_{t+1} = s') - V(S_t = s))$$

- This process is called temporal difference - specifically TD(0).
- The target is defined as: $R_t + \gamma V(S_{t+1} = s')$
- The error is defined as: $R_t + \gamma V(S_{t+1} = s') - V(S_t = s)$

Multistep TD and $TD(\lambda)$

- In general, for various n, the TD target is:
- N = 0 $G_t^0 = R_t + \gamma V(S_{t+1} = s')$
- N = 1 $G_t^1 = R_t + \gamma R_{t+1} + \gamma^2 V(S_{t+2} = s')$

- N = n $G_t^n = R_t + \gamma R_{t+1} + \ldots + \gamma^n R_{t+n} + \gamma^{n+1} V(S_{t+n+1} = s')$

- If N = n =T corresponds to the monte carlo rollout / episode, we can then combine the multistep TD results, producing $TD(\lambda)$

$$G_t^\lambda = (1-\lambda) \sum_{i=0}^{n} \lambda^n G_t^i$$

$$V(S_t = s) \leftarrow V(S_t = s) + \alpha(G_t^\lambda - V(S_t = s))$$

What Are the Advantages and Disadvantages?

- If we update the state value function using the complete rollout / episode (Monte Carlo), the updated estimate will have high variance (because of the numerous transitions in the episode).

- However, the estimate will be unbiased.

- If we update the state value function using only the first term in the rollout / episode (TD(0)), the updated estimate will be biased (because only a single transition is used).

- However, the estimate will have low variance.

- If we update the state estimate using a combination of the n terms, we will obtain an intermediate bias - variance tradeoff.

What Are the Advantages and Disadvantages?

- For the Monte Carlo approach, the entire rollout / episode needs to be observed, before an update can occur.

- For the TD(0) approach, only a single observation in the episode needs to be observed, before an update can occur.
  - Each “guess” is then updated by the next “guess” = bootstrapping.

- Because of bootstrapping, TD(0) is more sensitive to the starting condition than the Monte Carlo approach.

- However, unlike the Monte Carlo approach, TD(0) can be run online, can learn from incomplete sequences and does not require a terminating sequence to exist.

What Are the Advantages and Disadvantages?

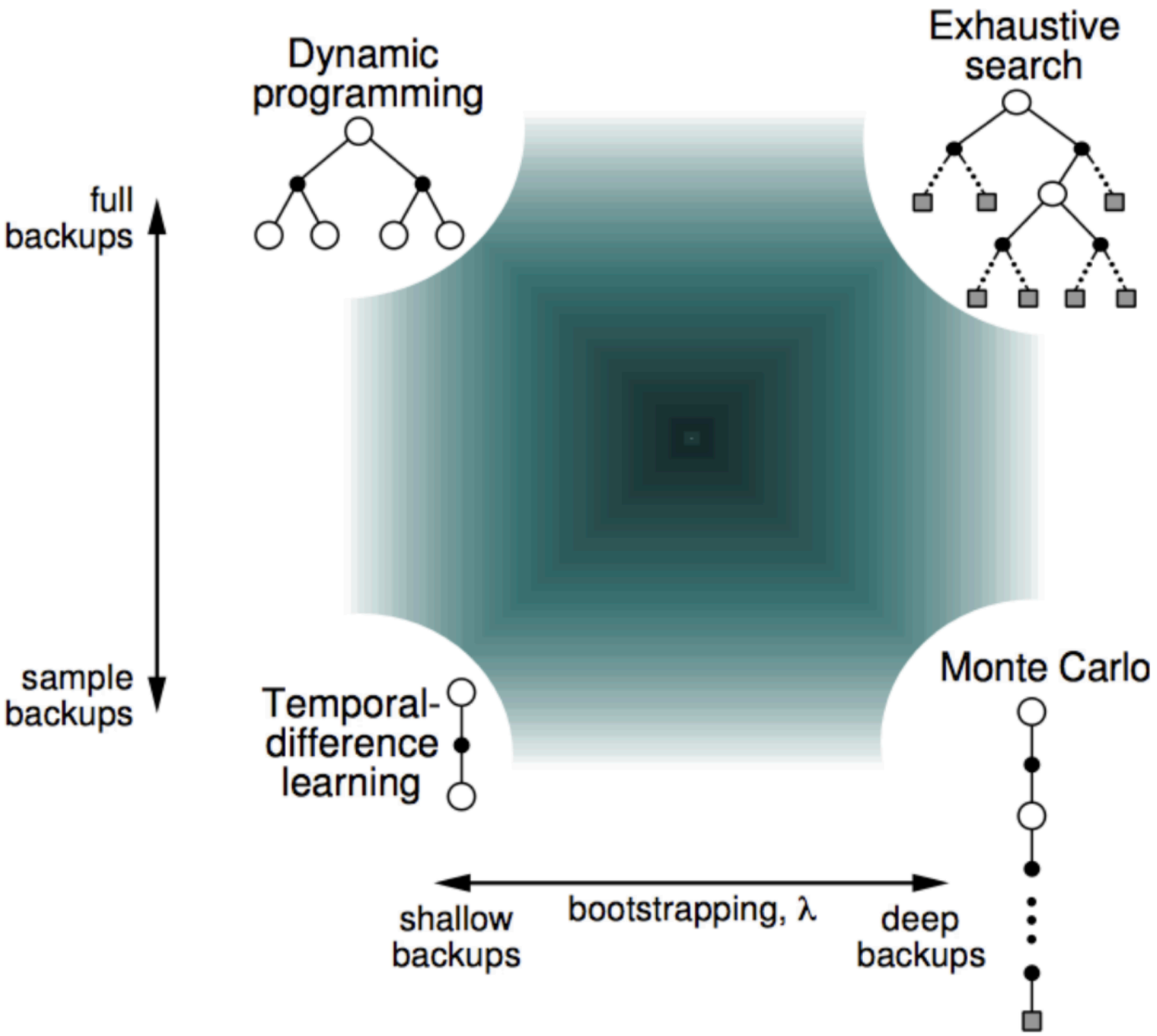
- By construction, the $TD(\lambda)$ approach shares some of the same disadvantages as the Monte Carlo approach, since it also needs to observe the entire episode.

- However, a backward view of $TD(\lambda)$ exists, where online updates can occur at every step, using incomplete sequences.

- This backward view involves computing eligibility traces.

- Qualitatively, eligibility traces track the frequency and the recency of a state visit.

- This information is then incorporated into the update.



This slide was taken from David Silver's RL lectures, and shows a unified view of Reinforcement Learning.

In the RL_PlanningAndDP slides, we discussed the DP backup diagram.
In this set of slides, we discussed the TD(0) and MC backup diagrams.