Recitation 5

A comparative overview of DP vs. MC vs. TD

3.5.2021

Notations

	Today's slides	Other equivalence
Current state, action, reward	s, a, r	s_t, a_t, r_t S_t, A_t, R_t
Next / successor state, action, reward	s', a', r'	$s_{t+1}, a_{t+1}, r_{t+1}$ $S_{t+1}, A_{t+1}, R_{t+1}$
True value function of a policy	$v_\pi(\cdot)$	
Estimate of value function	$V(\cdot)$	

Fundamental Concepts

Model-free / Model-based

- Model-free method requires no knowledge of an MDP’s rewards / dynamics.
- Model-based method does.

Fundamental Concepts

On-policy / Off-policy

- <u>On-policy</u> means behavior policy is the same as target policy.
- <u>Off-policy</u> means behavior policy is **not** the same as target policy.

> **What is behavior policy?**

Behavior policy is the one used to *select actions*.

> **What is target policy?**

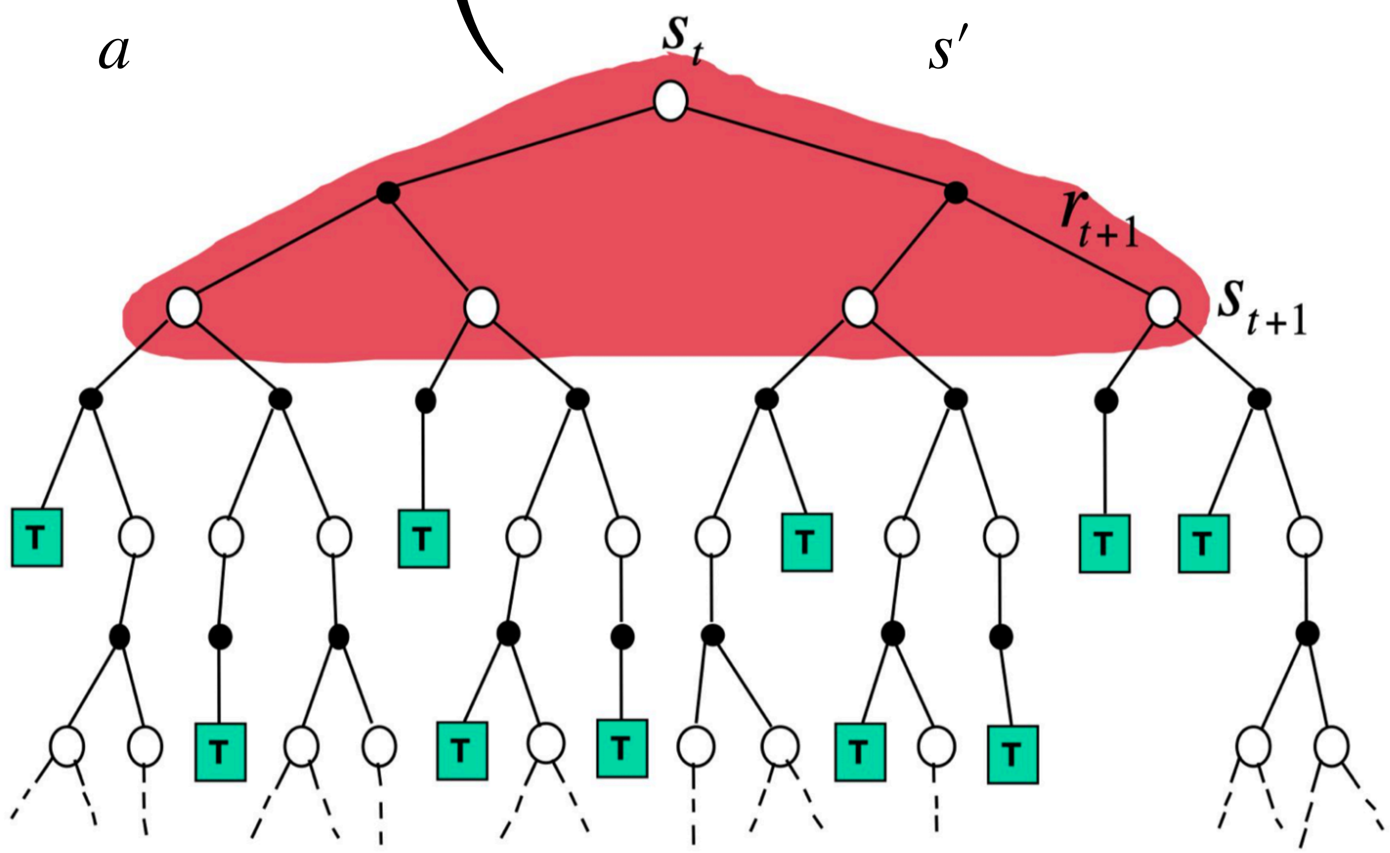
Target policy is the policy that an agent is *trying to learn*, i.e agent is learning value function for this policy.

Comparison

DP vs. MC vs. TD

Depth / Width of Backup

$$V(s) \leftarrow \sum_a \pi(a \mid s)\left(r(s,a) + \gamma \sum_{s'} p(s' \mid s,a) V(s') \right)$$



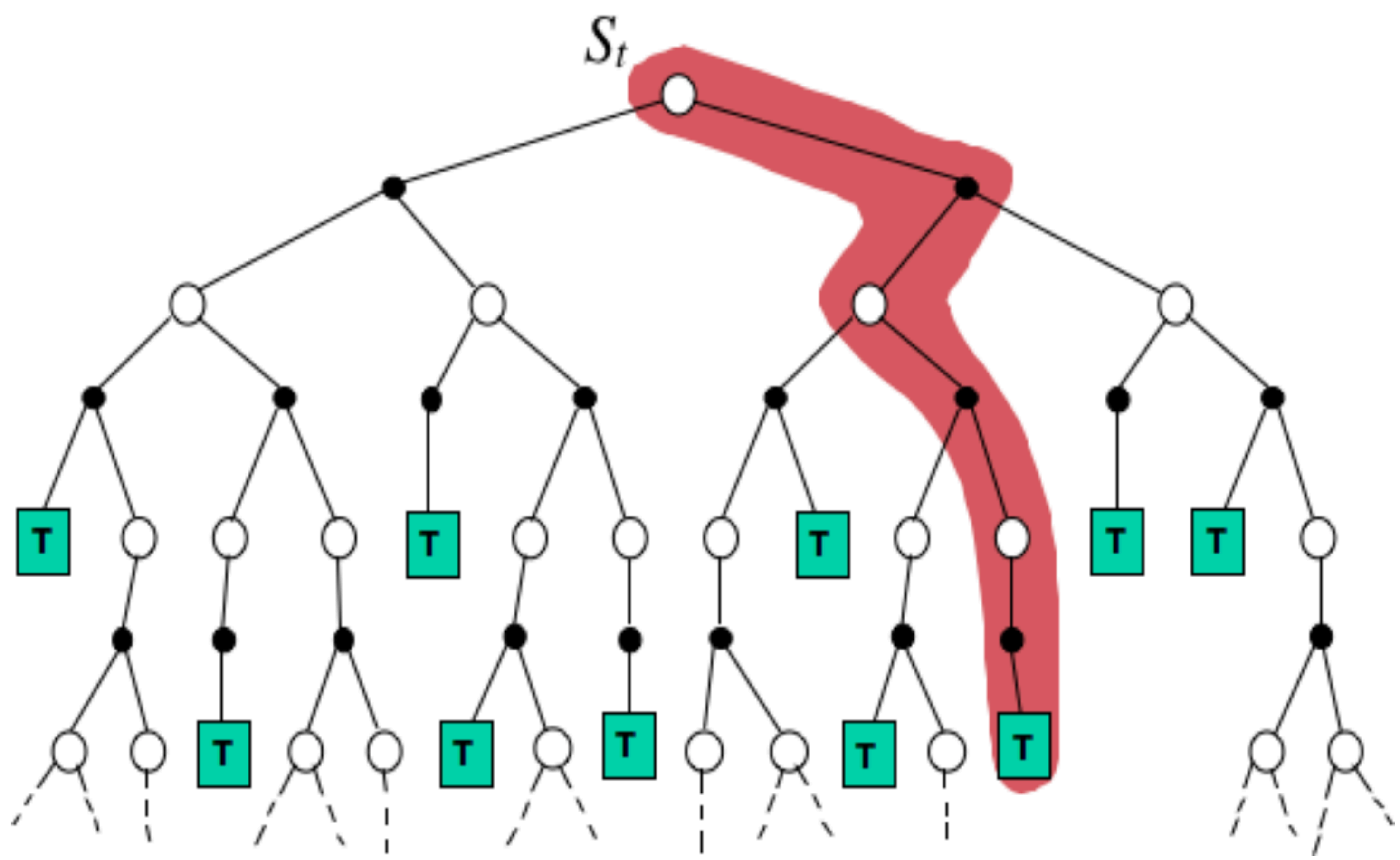
Back-up diagram of DP
Reference: David Silver RL Slides

Comparison

DP vs. MC vs. TD

Depth / Width of Backup

$$V(s) \leftarrow V(s) + \alpha(G_t - V(s))$$



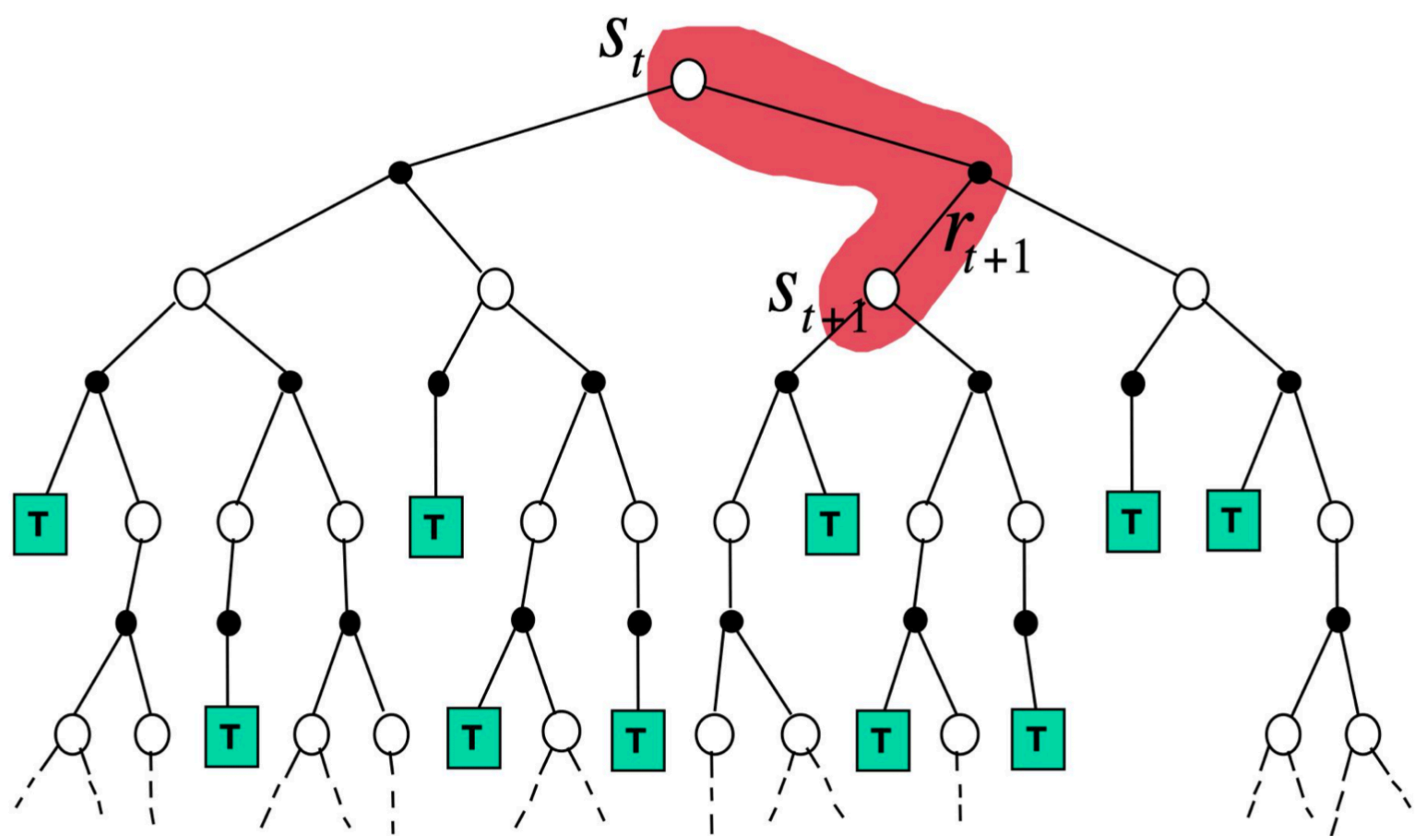
Back-up diagram of MC
Reference: David Silver RL Slides

Comparison

DP vs. MC vs. TD

Depth / Width of Backup

$$V(s) \leftarrow V(s) + \alpha(r' + \gamma V(s') - V(s))$$



Back-up diagram of TD(0)
Reference: David Silver RL Slides

Comparison

DP vs. MC vs. TD

(MC & TD) vs. DP

- (MC & TD) are model-free. DP is model-based.
- (MC & TD) learn directly by interacting with environment. DP doesn’t need to interact with environment.

Comparison

DP vs. MC vs. TD

Advantage of MC over TD

- Unbiased.
- Less sensitive to initial value.
- Good convergence
- Easy to understand

Comparison

DP vs. MC vs. TD.

Advantage of TD over MC

- Lower variance.
- Can learn from incomplete episode.
- Can apply to non-terminating environment.
- Usually more efficient than MC.

Comparison

DP vs. MC vs. TD

MC vs. TD
Bias / Variance Trade-Off

	Backup	Concern	Biased / Unbiased Estimate of $v_\pi(s)$?
MC	$V(s) \leftarrow V(s) + \alpha(G_t - V(s))$	$G_t = \sum_{k=t}^{T-1} \gamma^{k-t} r_{k+1}$	Unbiased
TD	$V(s) \leftarrow V(s) + \alpha(r' + \gamma V(s') - V(s))$	$r' + \gamma V(s')$ <TD target>	Biased
		$r' + \gamma v_\pi(s')$ <**True** TD target>	Unbiased

Comparison

DP vs. MC vs. TD

MC vs. TD
Bias / Variance Trade-Off

	Backup	Concern	Dependence	Variance
MC	$V(s) \leftarrow V(s) + \alpha(G_t - V(s))$	$G_t = \sum_{k=t}^{T-1} \gamma^{k-t} r_{k+1}$	Depends on **many** random actions, transitions, rewards	Higher
TD	$V(s) \leftarrow V(s) + \alpha(r' + \gamma V(s') - V(s))$	$r' + \gamma V(s')$	Depends on **one** random action, transition, reward	Lower

Comparison

DP vs. MC vs. TD

MC vs. TD
Off-Policy Importance Sampling

	Backup	Dependence	Variance
MC	$V(s) \leftarrow V(s) + \alpha(G_t^{\pi/\mu} - V(s)), \quad G_t^{\pi/\mu} = \prod_{k=t}^{T} \frac{\pi(a_k \mid s_k)}{\mu(a_k \mid s_k)} G_t$	Depends on **many** random actions, transitions, rewards	Higher
TD	$V(s) \leftarrow V(s) + \alpha \left(\frac{\pi(a \mid s)}{\mu(a \mid s)}(r' + \gamma V(s')) - V(s) \right)$	Depends on **one** random action, transition, reward	Lower

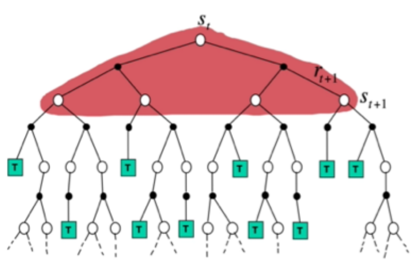
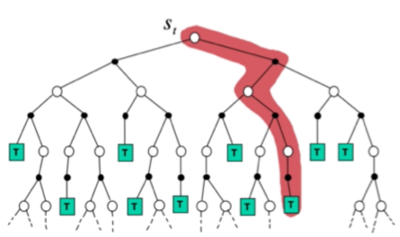
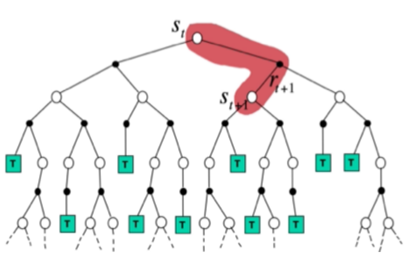
Summary

- Connection among DP, MC, TD
- Connection between MC, TD
- Connection between DP, TD

	Model-free / Model-based	# of action considered at each state	Involve next state? s'	Sub-topics	Backup
DP (Dynamic Programming)	Model-based	All, $\lvert\mathscr{A}\rvert$	Yes	• Asynchronous / Synchronous DP • Iterative Policy Evaluation / Policy Iteration / Value Iteration • Prediction & Control • Convergence	$V(s) \leftarrow \sum_{a} \pi(a \mid s)\left(r(s,a) + \gamma \sum_{s'} p(s' \mid s,a) V(s')\right)$ $V(s) \leftarrow \max_{a \in \mathscr{A}} \left(r(s,a) + \gamma \sum_{s'} p(s' \mid s,a) V(s')\right)$
MC (Monte Carlo)	Model-free	One	No	• Prediction & Control • Convergenc • First-visit / Every-visit • On-policy / Off-policy	$V(s) \leftarrow V(s) + \alpha(G_t - V(s))$
TD (Temporal Difference) **TD(0)**	Model-free	One	Yes	• Prediction & Control • Convergence • On-policy / Off-policy	$V(s) \leftarrow V(s) + \alpha(r' + \gamma V(s') - V(s))$

Summary

- Connection among DP, MC, TD
- Connection between MC, TD
- Connection between DP, TD

	Diagram	Bootstrapping? (update involves an estimate)	Sampling?	Bias / Variance Tradeoff	Computation
DP (Dynamic Programming)		Yes	No	X	• Costly when directly solving matrix solution • Costly when doing full sweep in iteration, especially when $\lvert \mathcal{S} \rvert$ is large.
MC (Monte Carlo)		No	Yes	• High variance, no bias	• Usually higher than TD
TD (Temporal Difference) TD(0)		Yes	Yes	• Low variance, some bias	• Usually better than MC • Less computation and less memory

Reference & Acknowledgement

Many of the slides and tables are summarized based on David Silver's RL slides, current and previous CMU 10403 & 10703 lecture and recitation slides

Important Disclaimer

Quiz 1 will **not** cover TD content!

Q & A

Thank you for listening.