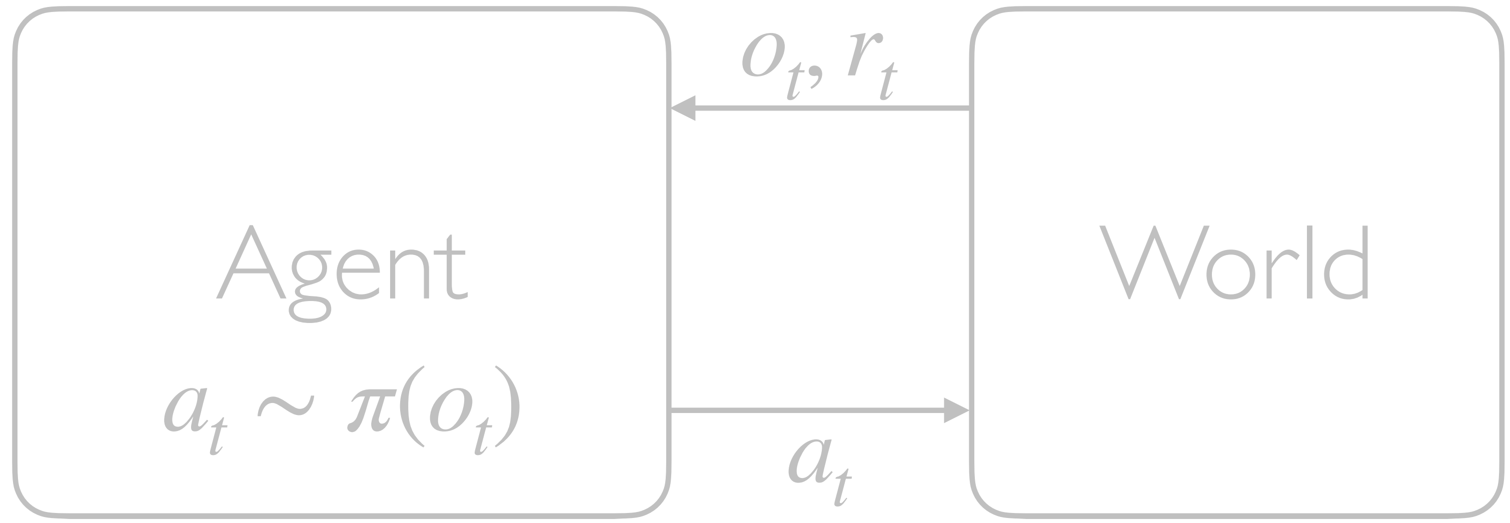
Reinforcement Learning

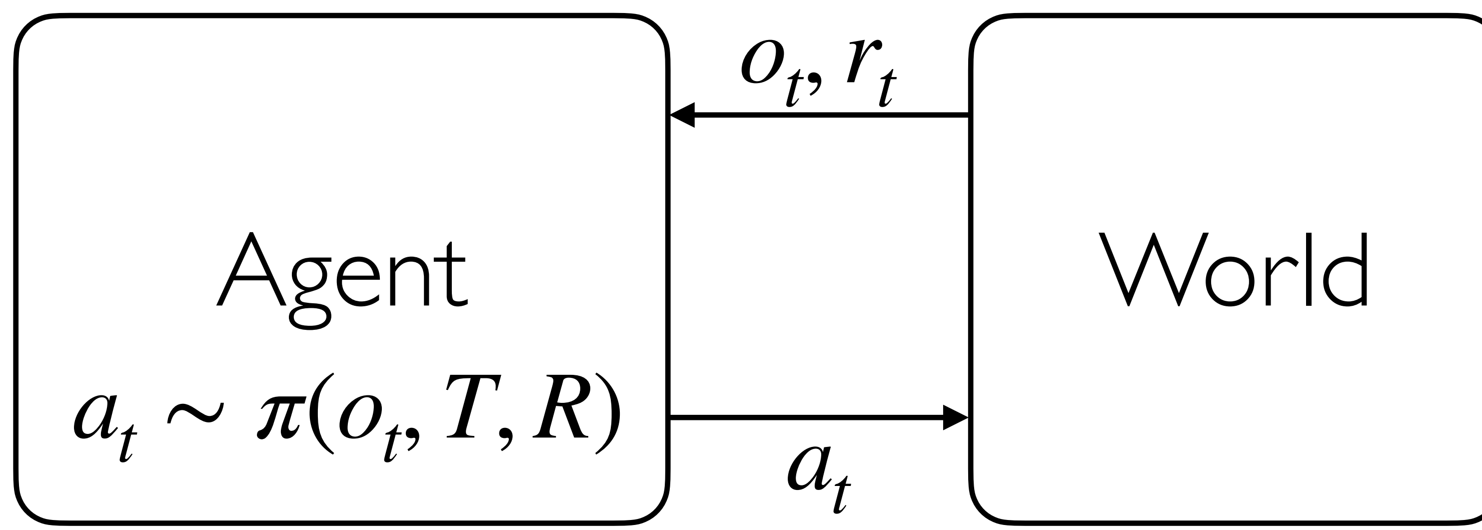
Solving MDPs

Policy: $a_t \sim \pi(o_t)$

Most General Case



More Specific Case



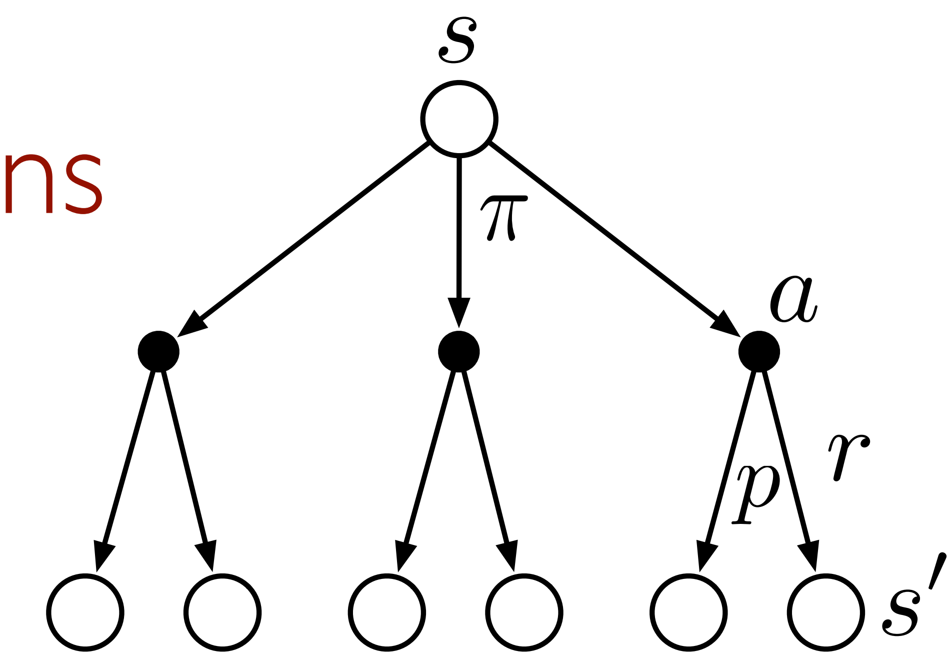
Fully Observed System $o_t = s_t$

Known Transition Function $s_{t+1} \sim T(s_t, a_t)$

Known Reward Function $R(s_{t+1}, s_t, a_t)$

Recap

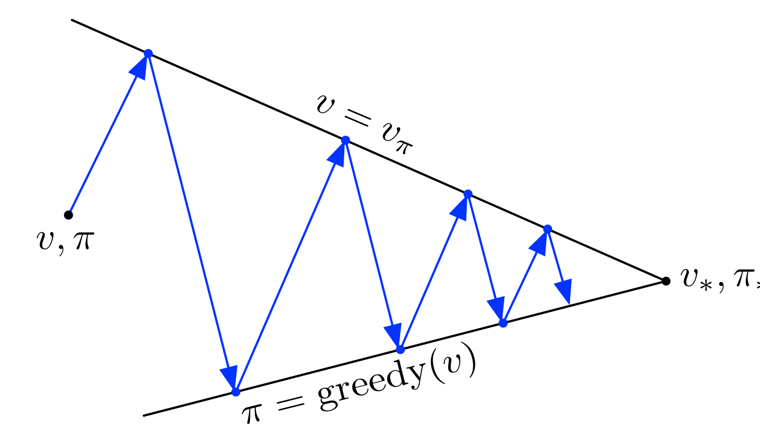
Computing $V_*(s)$ and $Q_*(s, a)$ for known MDPs.



Backup diagrams, Bellman equations

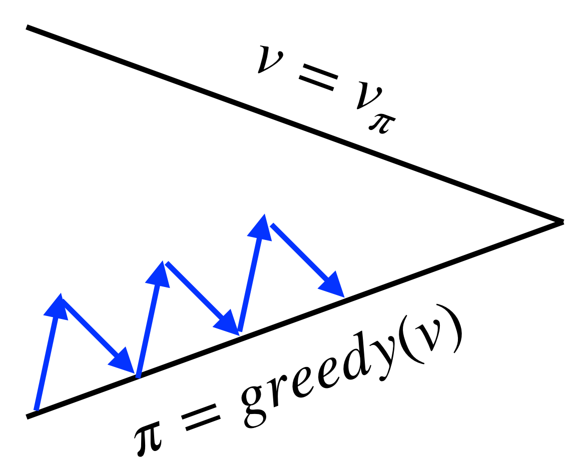
$$V_\pi(s) = \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)\big(r + \gamma V_\pi(s')\big)$$

Policy Evaluation, Improvement



Value Iteration

$$V_{k+1}(s) = max_a \sum_{s',r} p(s', r \mid s, a)\big(r + \gamma V_k(s')\big)$$



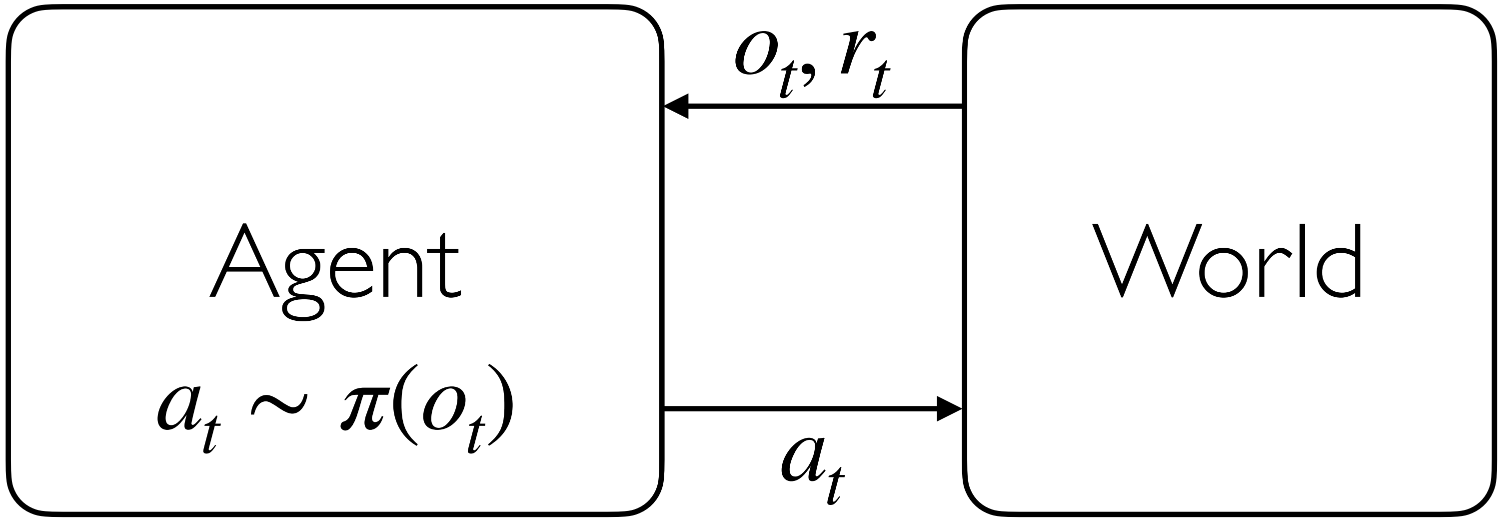
Solving MDPs

Policy: $a_t \sim \pi(o_t)$

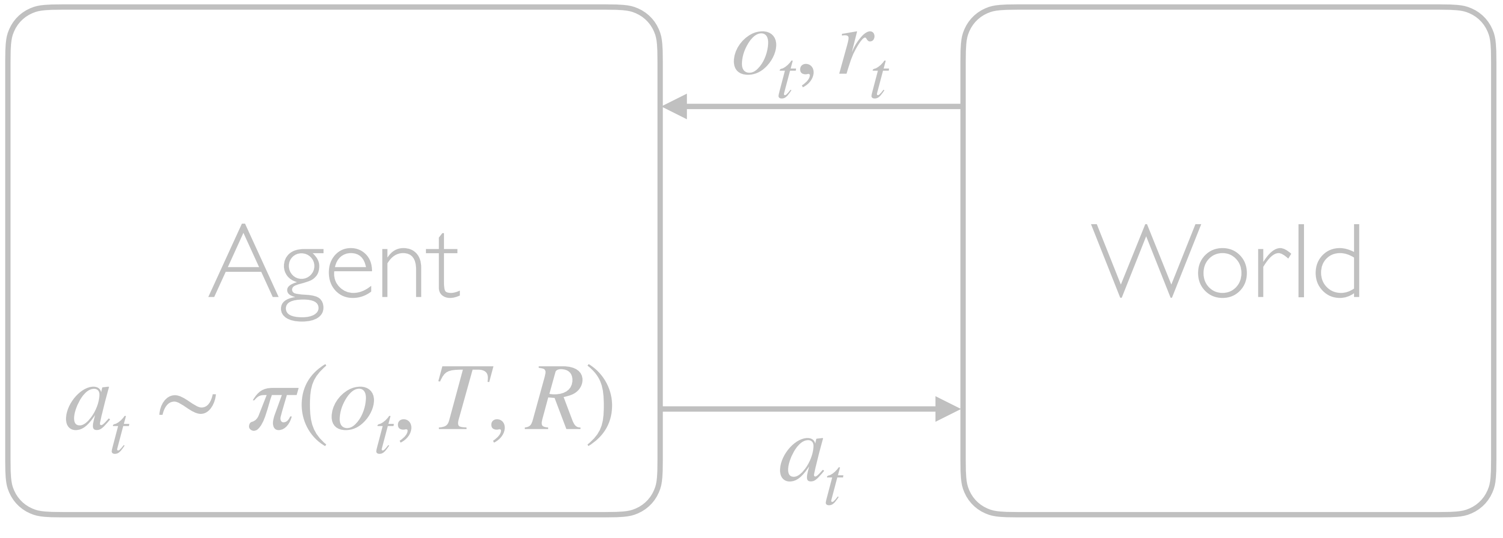
Most General Case

Fully Observed System:

$o_t = s_t$



More Specific Case

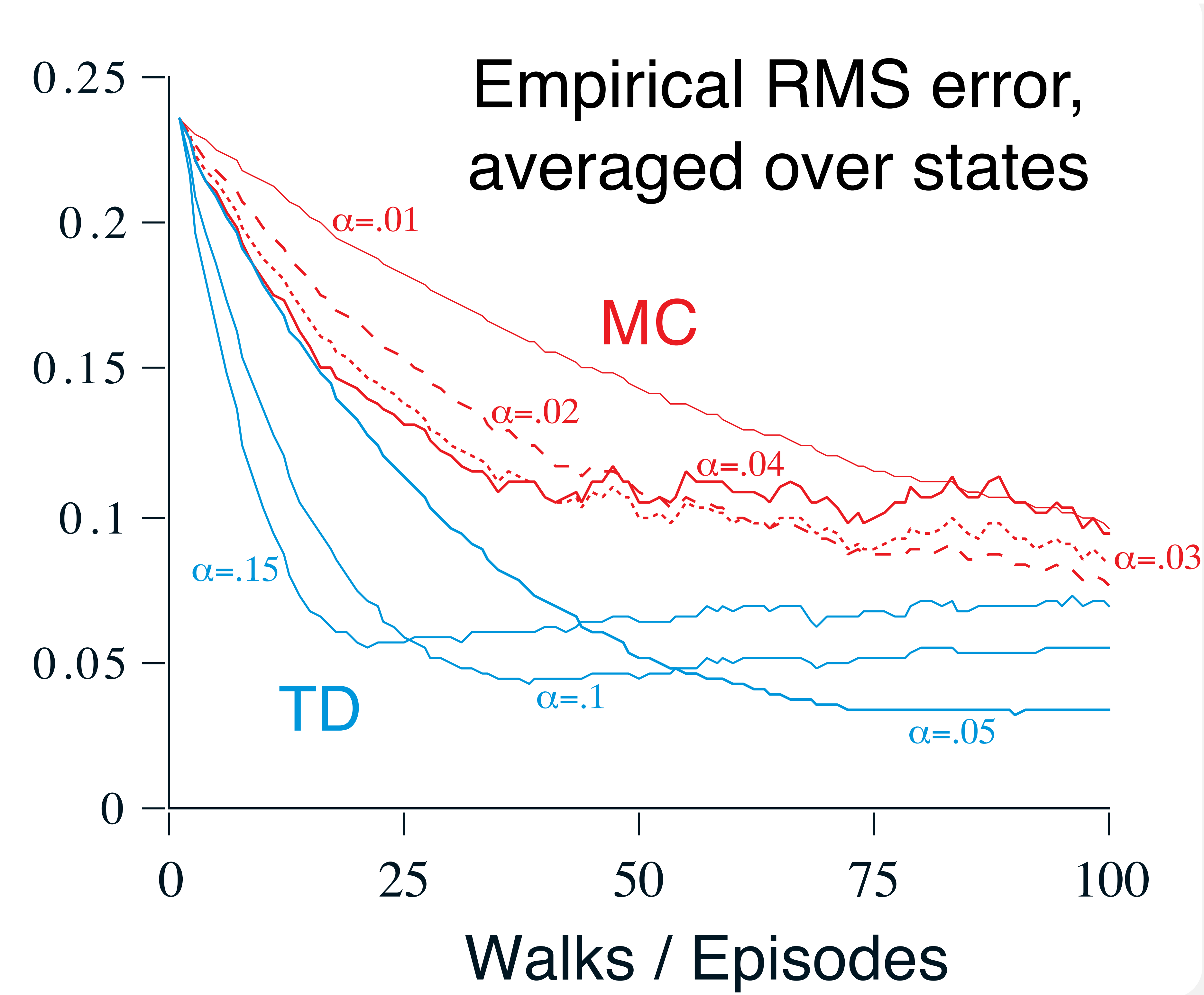


Fully Observed System $o_t = s_t$

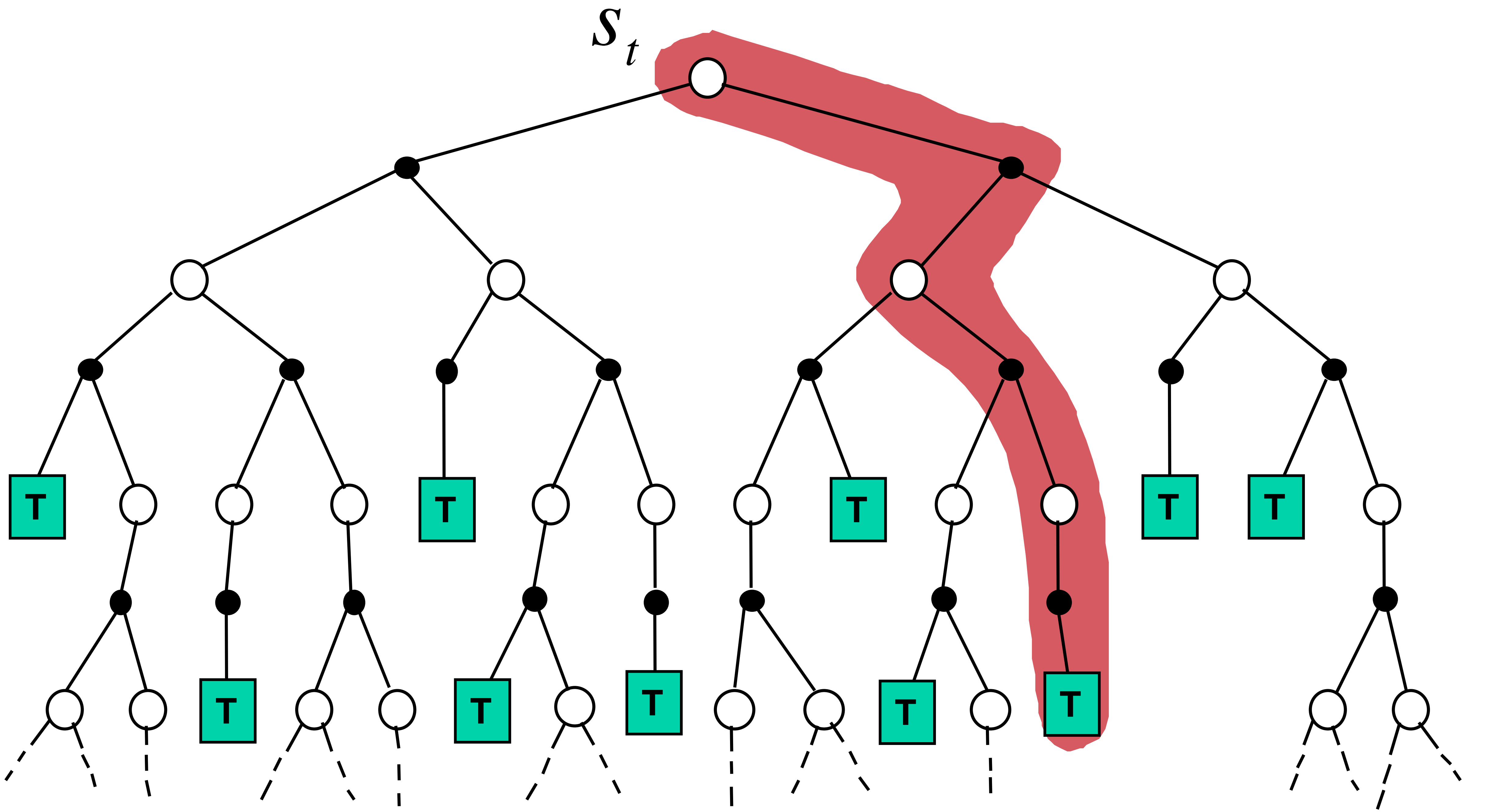
Known Transition Function $s_{t+1} \sim T(s_t, a_t)$

Known Reward Function $R(s_{t+1}, s_t, a_t)$

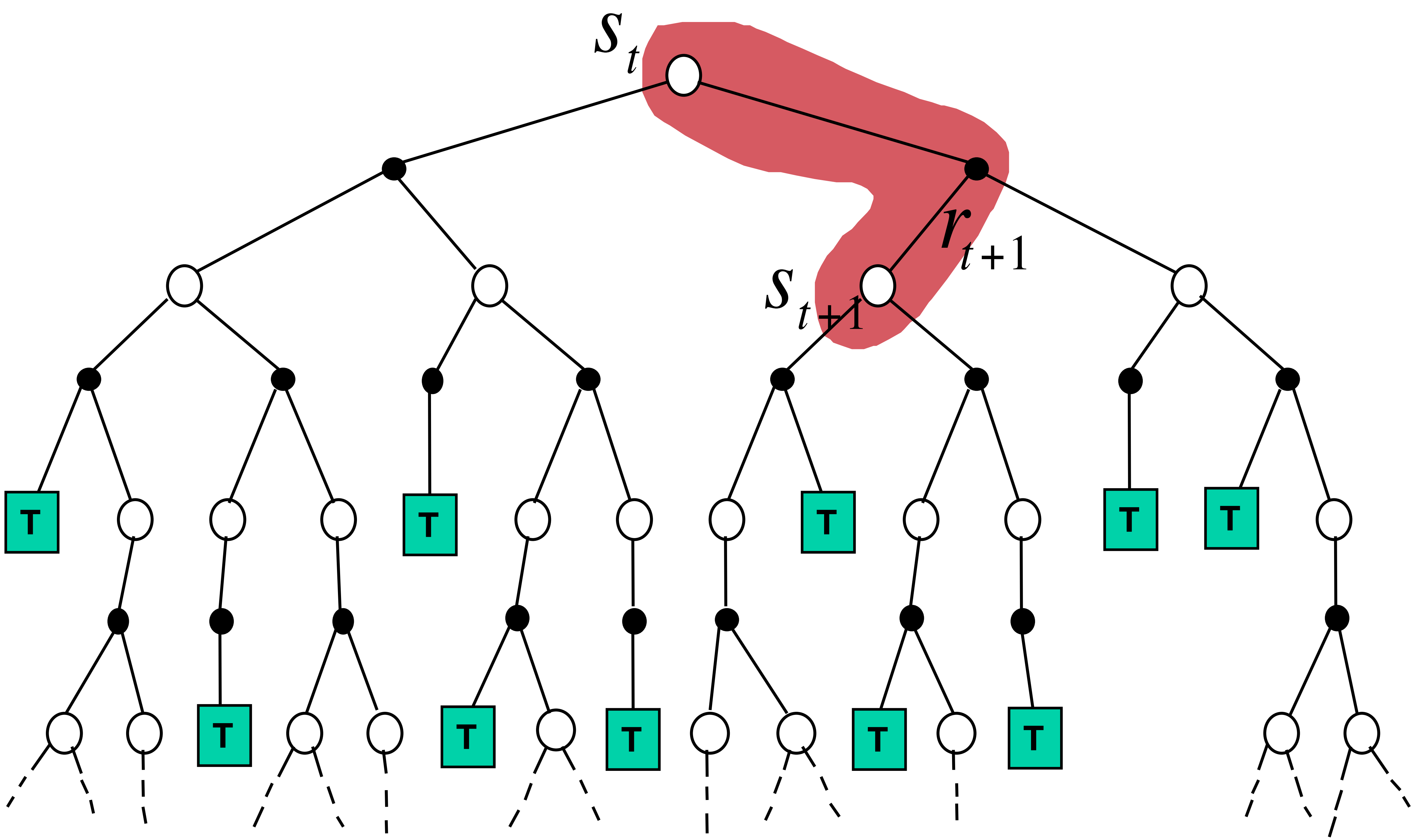
TD vs MC


Empirical RMS error, averaged over states
0.25
0.2
0.15
0.1
0.05
0
α=.01
α=.02
α=.04
α=.03
MC
α=.15
α=.1
α=.05
TD
0
25
50
75
100
Walks / Episodes

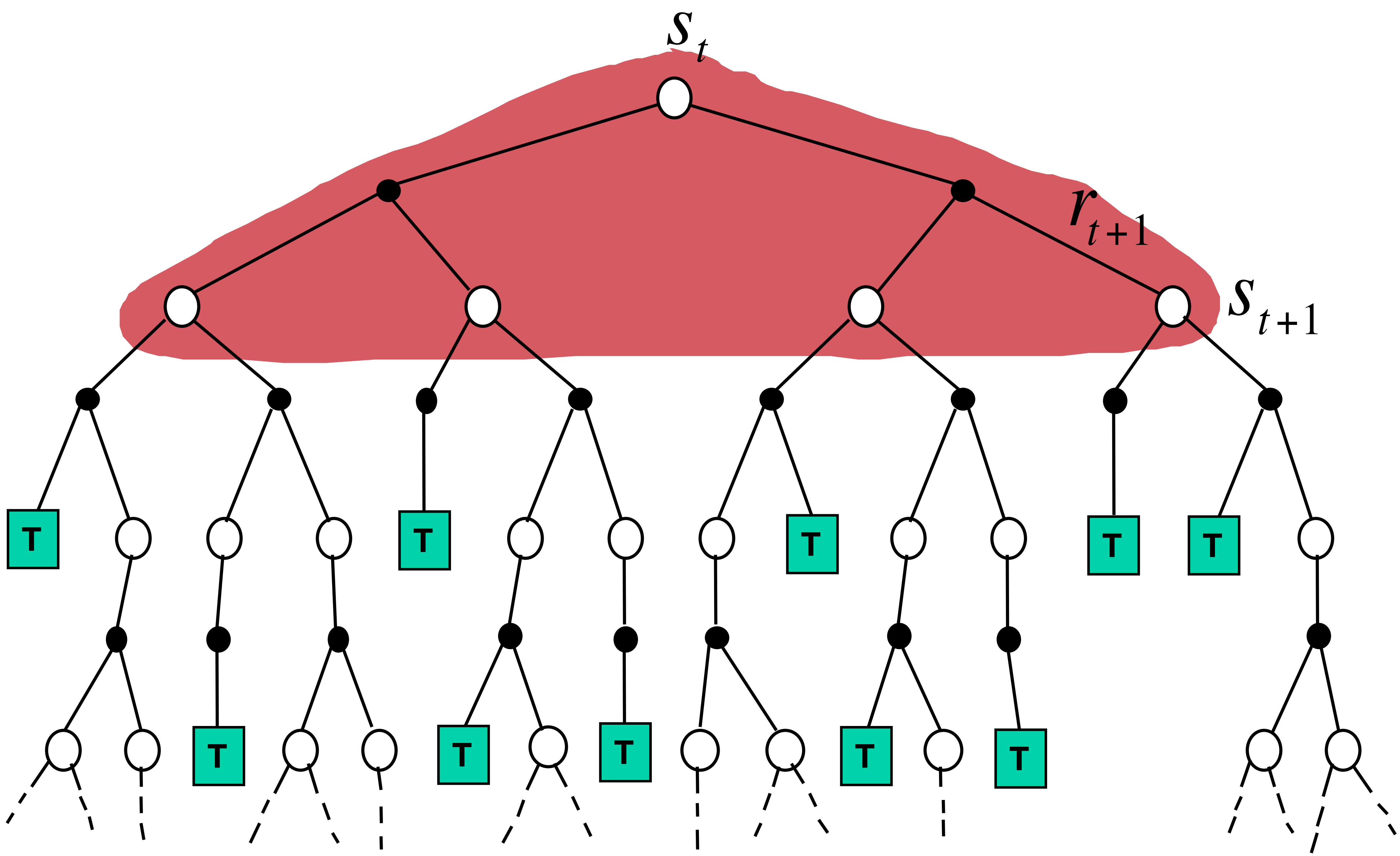
MC Backup

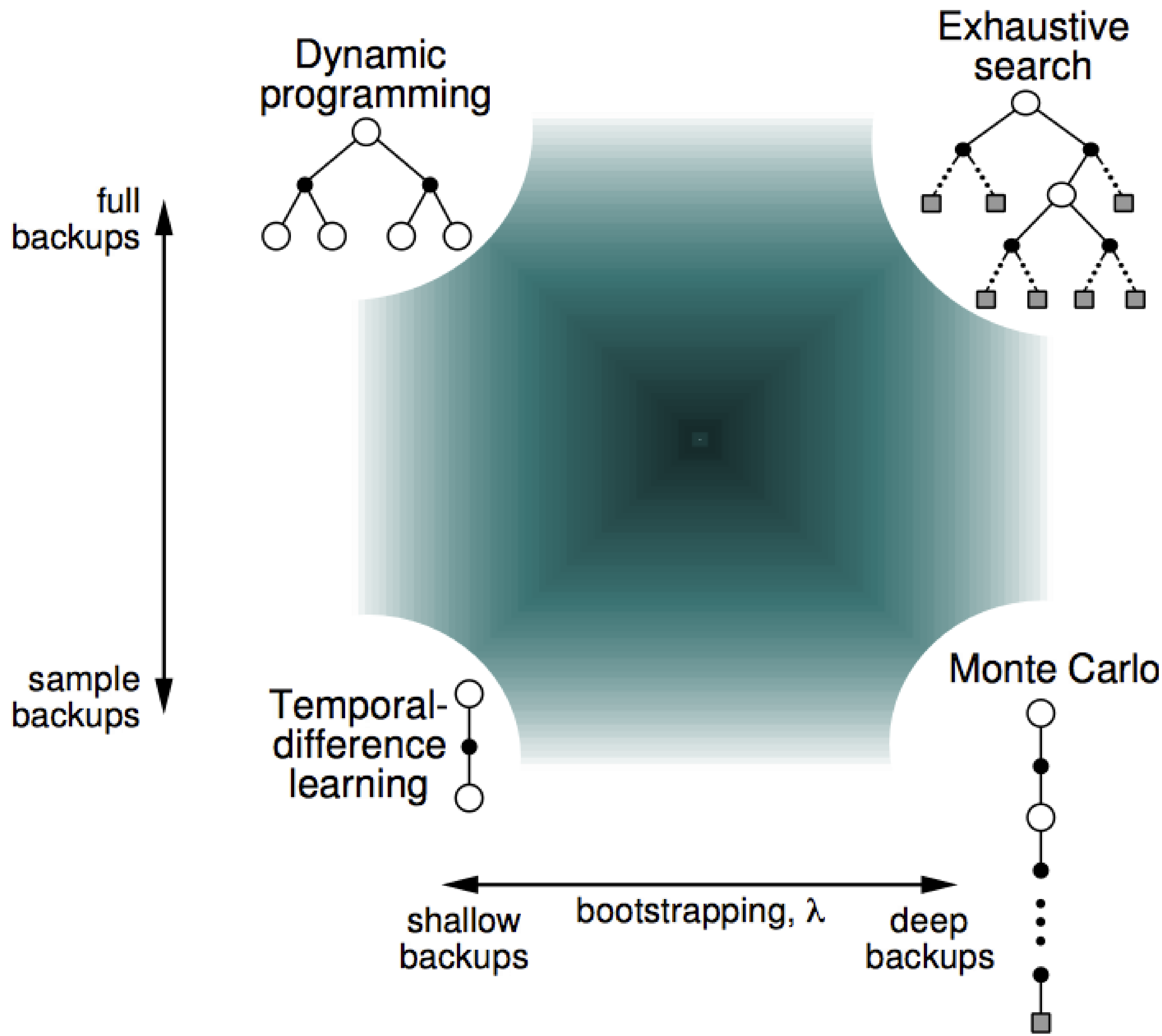


TD(0) Backup



Dynamic Programming Backup




Dynamic programming
Exhaustive search
full backups
sample backups
Temporal-difference learning
Monte Carlo
shallow backups
bootstrapping, λ
deep backups

Recap

- Model Free Policy Evaluation
  - Monte Carlo, TD(0), TD(λ)
- Model Free Control
  - On-policy: ϵ-greedy, SARSA, SARSA(λ)
  - Off-policy: Q-Learning

Model Free RL

Model Free Policy Evaluation

Model Free Control

Playing Atari with Deep Reinforcement Learning

Volodymyr Mnih Koray Kavukcuoglu David Silver Alex Graves Ioannis Antonoglou
Daan Wierstra Martin Riedmiller