Reinforcement Learning

Computer Engineering Department
Sharif University of Technology

Mohammad Hossein Rohban, Ph.D.

Spring 2024

Courtesy: Some slides are adopted from CS 285 Berkeley, and CS 234 Stanford, and Pieter Abbeel's compact series on RL.

Outline

- Recap of Monte Carlo
- Temporal Difference Learning (Prediction)
- TD vs. MC
- n-Step TD
- TD(λ) (Forward and Backward view)
- Temporal Difference Learning (Control)
- SARSA
- On and Off-Policy Learning
- Q-Learning

Disadvantages of MC Learning

- We have seen MC algorithms can be used to learn value predictions
- But when episodes are long, learning can be slow
  - ...we have to **wait until an episode ends** before we can learn
  - ...return can have **high variance**
- Are there alternatives? (Yes)

Temporal Difference Learning

Prediction

TD Overview

TD methods learn directly from episodes of experience

TD is *model-free*: no knowledge of MDP transitions / rewards

TD learns from *incomplete* episodes, by *bootstrapping*

TD updates a guess towards a guess

Temporal Difference Learning by Sampling Bellman Equations

- Bellman update equations:

$$v_{k+1}(s) = \mathbb{E}\left[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s, A_t \sim \pi(S_t)\right]$$

- We can sample this!

$$v_{t+1}(S_t) = R_{t+1} + \gamma v_t(S_{t+1})$$

- Samples could be averaged, in a similar way to MC:

$$v_{t+1}(S_t) = v_t(S_t) + \alpha_t \underbrace{\left(\underbrace{R_{t+1} + \gamma v_t(S_{t+1})}_{\text{target}} - v_t(S_t) \right)}_{\text{temporal difference error } \delta_t}$$

Temporal Difference Learning

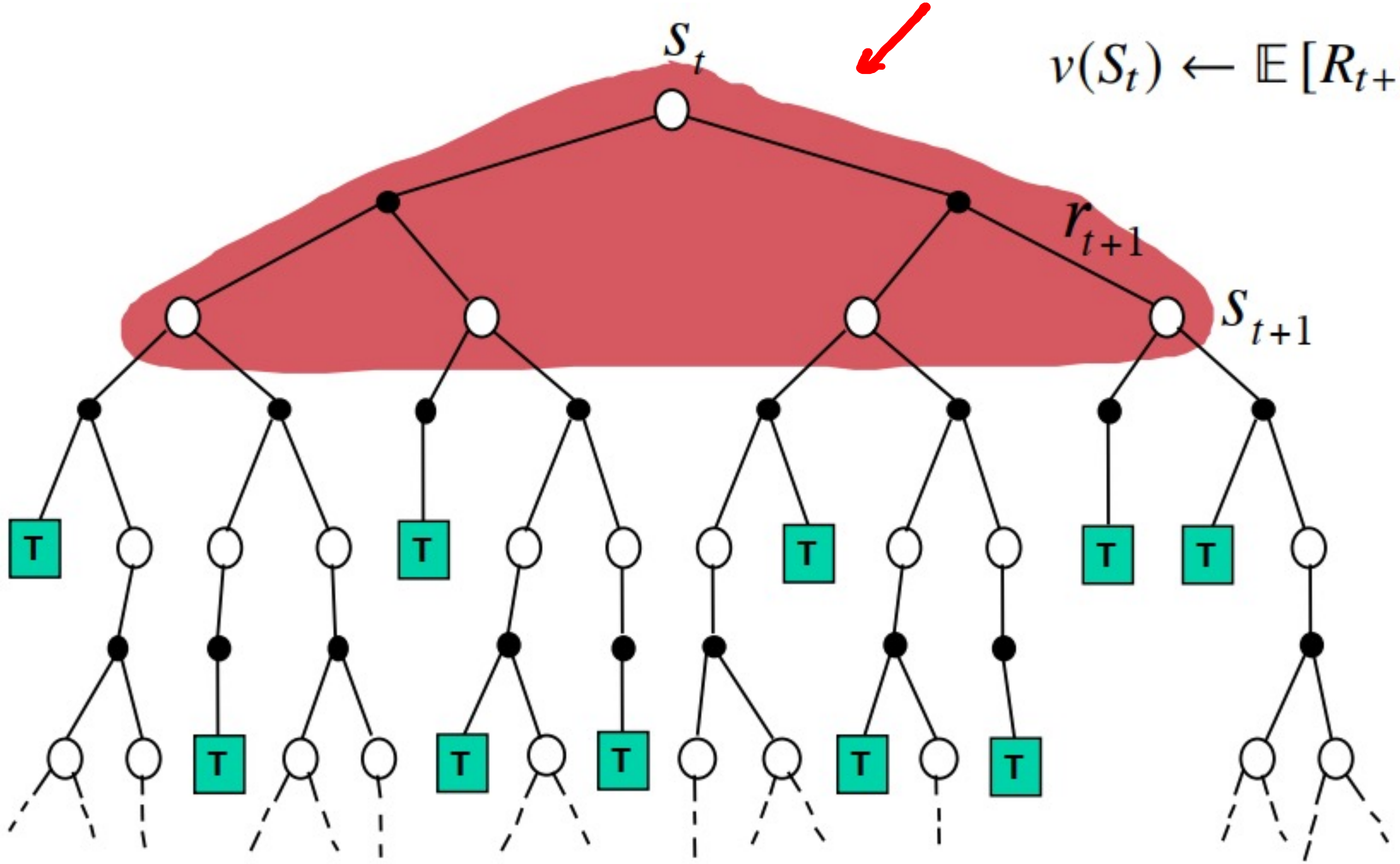
- Prediction setting: learn v_π online from experience under policy π
- Monte Carlo
  - Update value $v_n(S_t)$ towards sampled return G_t

$$v_{n+1}(S_t) = v_n(S_t) + \alpha\left(G_t - v_n(S_t)\right)$$

- TD Learning
  - Update value $v_t(S_t)$ towards estimated return $R_{t+1} + \gamma v(S_{t+1})$

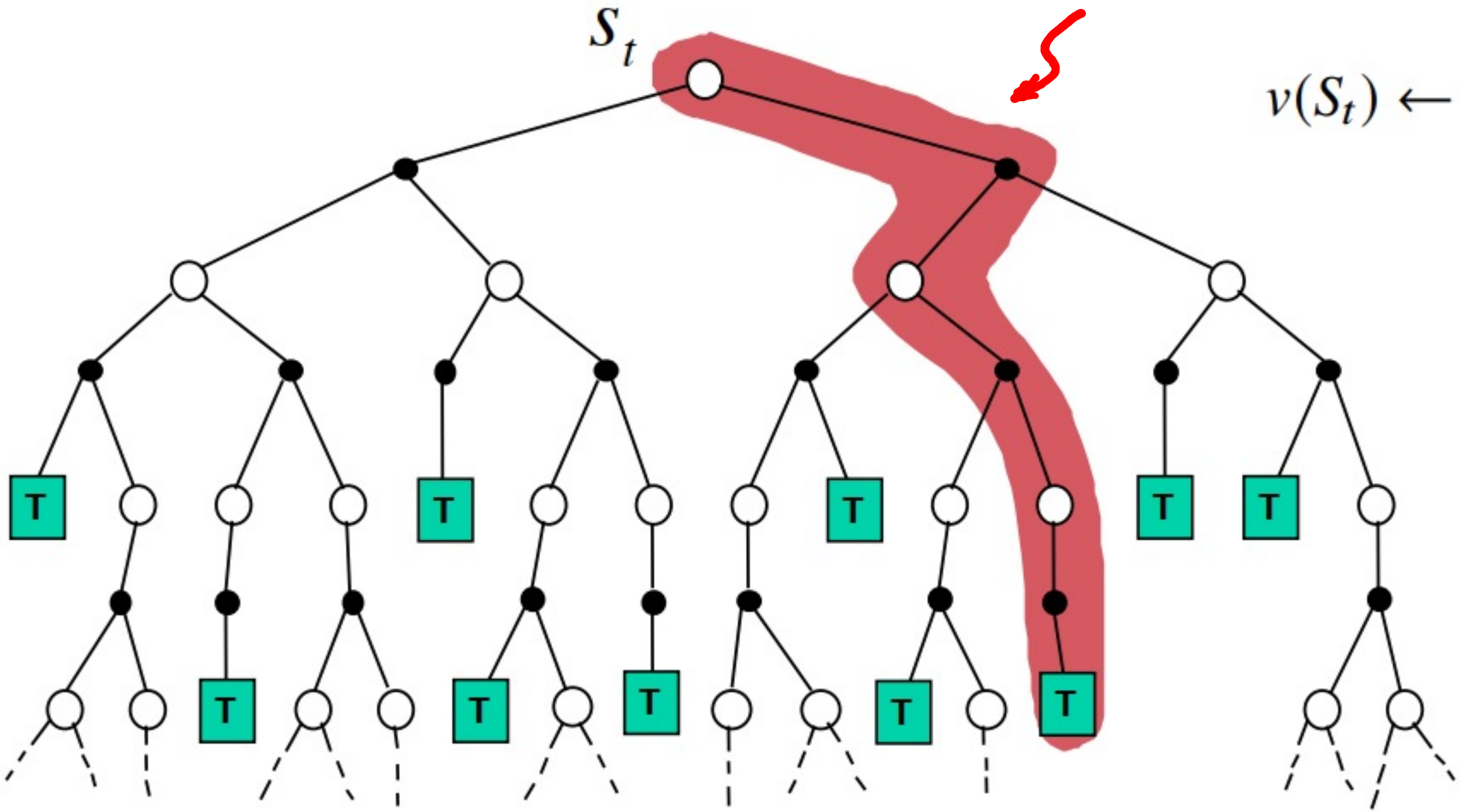
$$v_{t+1}(S_t) \leftarrow v_t(S_t) + \alpha\left(\overbrace{\underbrace{R_{t+1} + \gamma v_t(S_{t+1})}_{\text{target}} - v_t(S_t)}^{\text{TD error}}\right)$$

Backup (Dynamic Programming)



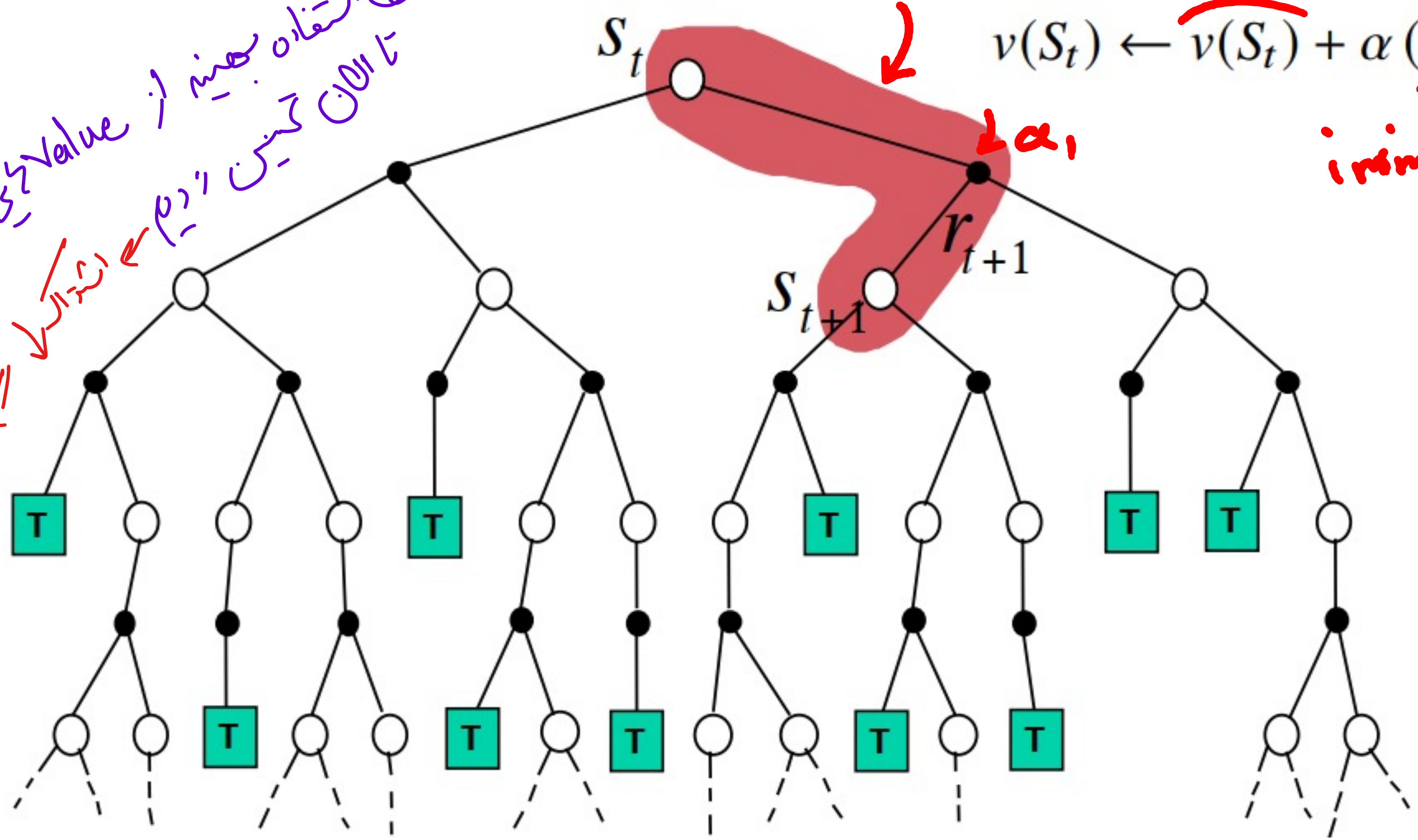
$$v(S_t) \leftarrow \mathbb{E}\left[R_{t+1} + \gamma v(S_{t+1}) \mid A_t \sim \pi(S_t)\right]$$

Backup (Monte Carlo)



$$v(S_t) \leftarrow v(S_t) + \alpha\,(G_t - v(S_t))$$

Backup (Temporal Difference)



① استفاده بهینه از Valueهایی که تا الان تخمین زدیم ⇐ اشتراک ⇐ واریانس

estimated

$$v(S_t) \leftarrow v(S_t) + \alpha\,(\underbrace{R_{t+1}}_{\text{immediate reward}} + \underbrace{\gamma v(S_{t+1})}_{\text{estimated disc. reward to go}} - v(S_t))$$

Bootstrapping and Sampling

- Bootstrapping: update involves an estimate
  - MC does not bootstrap
  - DP bootstraps
  - TD bootstraps
- Sampling: update samples an expectation
  - MC samples
  - DP does not sample
  - TD samples

TD Learning for action values

TD(1)

TD(0)

$S_0, A_0, R_1, S_1, A_1, \cdots \sim \pi_t$

SARSA

- We can apply the same idea to action values
- Temporal-difference learning for action values:

  - Update value $q_t(S_t, A_t)$ towards estimated return $R_{t+1} + \gamma q(S_{t+1}, A_{t+1})$

$\sum \alpha_t = \infty$

$\sum \alpha_t^2 < \infty$

$\alpha \to 0$

$$q_{t+1}(S_t, A_t) \leftarrow q_t(S_t, A_t) + \alpha \left(\overbrace{\underbrace{R_{t+1} + \gamma q_t(S_{t+1}, A_{t+1})}_{\text{target}} - q_t(S_t, A_t)}^{\text{TD error}} \right)$$

این همگرا شه

سایز

current est.

sample ← ولی این نه لزوماً چون همواره یک sample

TD vs. MC

- **TD can learn before knowing the final outcome**
  - TD can learn online after every step
  - MC must wait until end of episode before return is known
- **TD can learn without the final outcome**
  - MC must wait until end of episode before return is known
  - MC can only learn from complete sequences
  - TD works in continuing (non-terminating) environments
  - MC only works for episodic (terminating) environments
- **TD is independent of the temporal span of the prediction**
  - TD can learn from single transitions
  - MC must store all predictions (or states) to update at the end of an episode
- **TD needs reasonable value estimates** → بدیشه → اولش اگه خیلی بد باشه مشکله....

فرق RL در این هست که اگر در یه جای اشتباه گیر کنی دیگه مگه با خدا بیاری بیرون
لا
مومنتوم اون زیاده

Bias/Variance Tradeoff

- MC return $G_t = R_{t+1} + \gamma R_{t+2} + \cdots$ is an unbiased estimate of $v_\pi(S_t)$

- TD target $R_{t+1} + \gamma v_t(S_{t+1})$ is a biased estimate of $v_\pi(S_t)$
  - unless $\mathbb{E}[v_t(S_{t+1}) | S_{t+1}] = v_\pi(S_{t+1})$

- But the TD target has lower variance:

MC

  - Return depends on many random actions, transitions, rewards
  - TD target depends on one random action, transition, reward

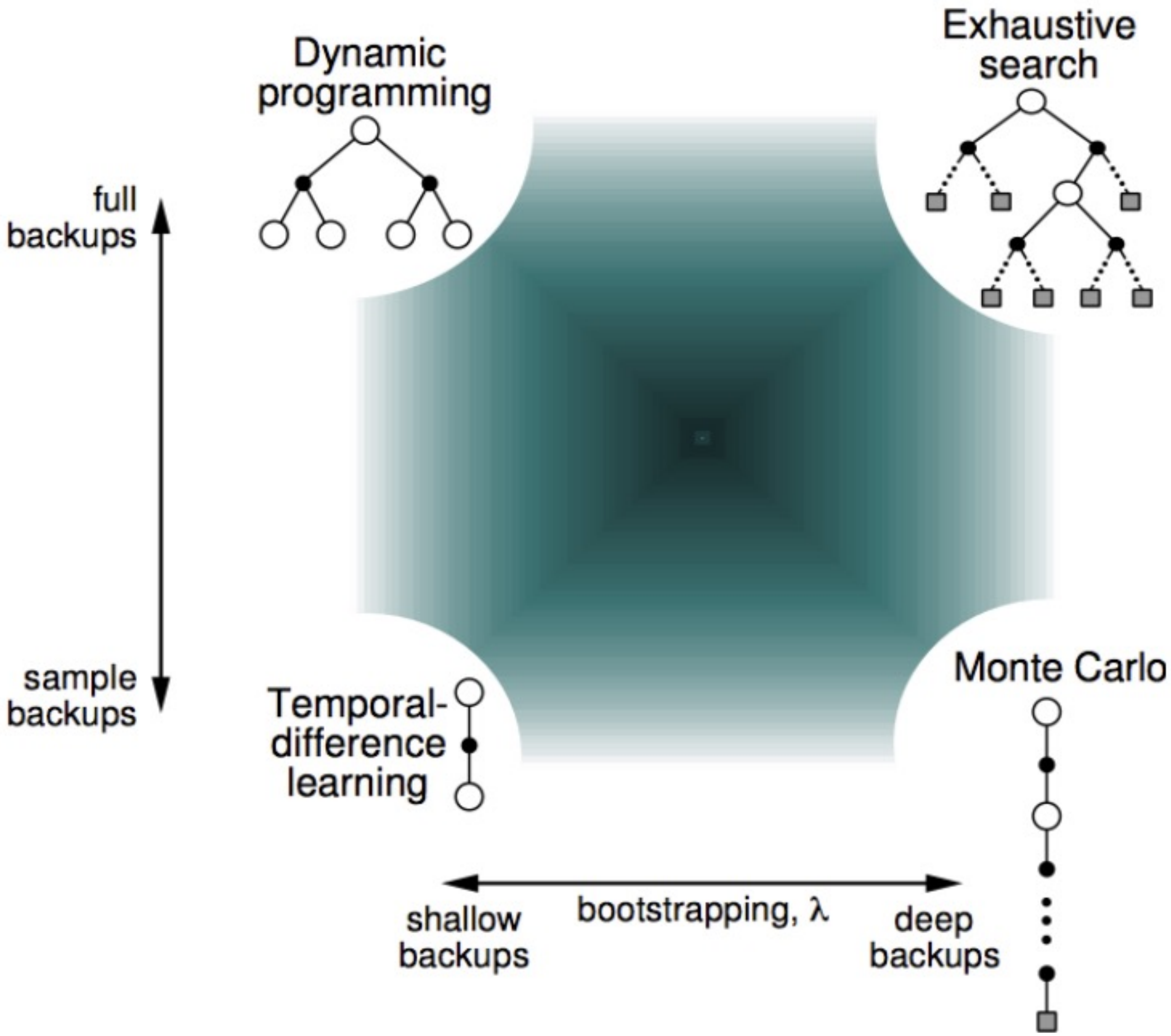
TD

$\Rightarrow$ منبع واریانس کمتر شده

TD(1) TD(0)

Between MC and TD: Multi-step TD

General View of Reinforcement Learning



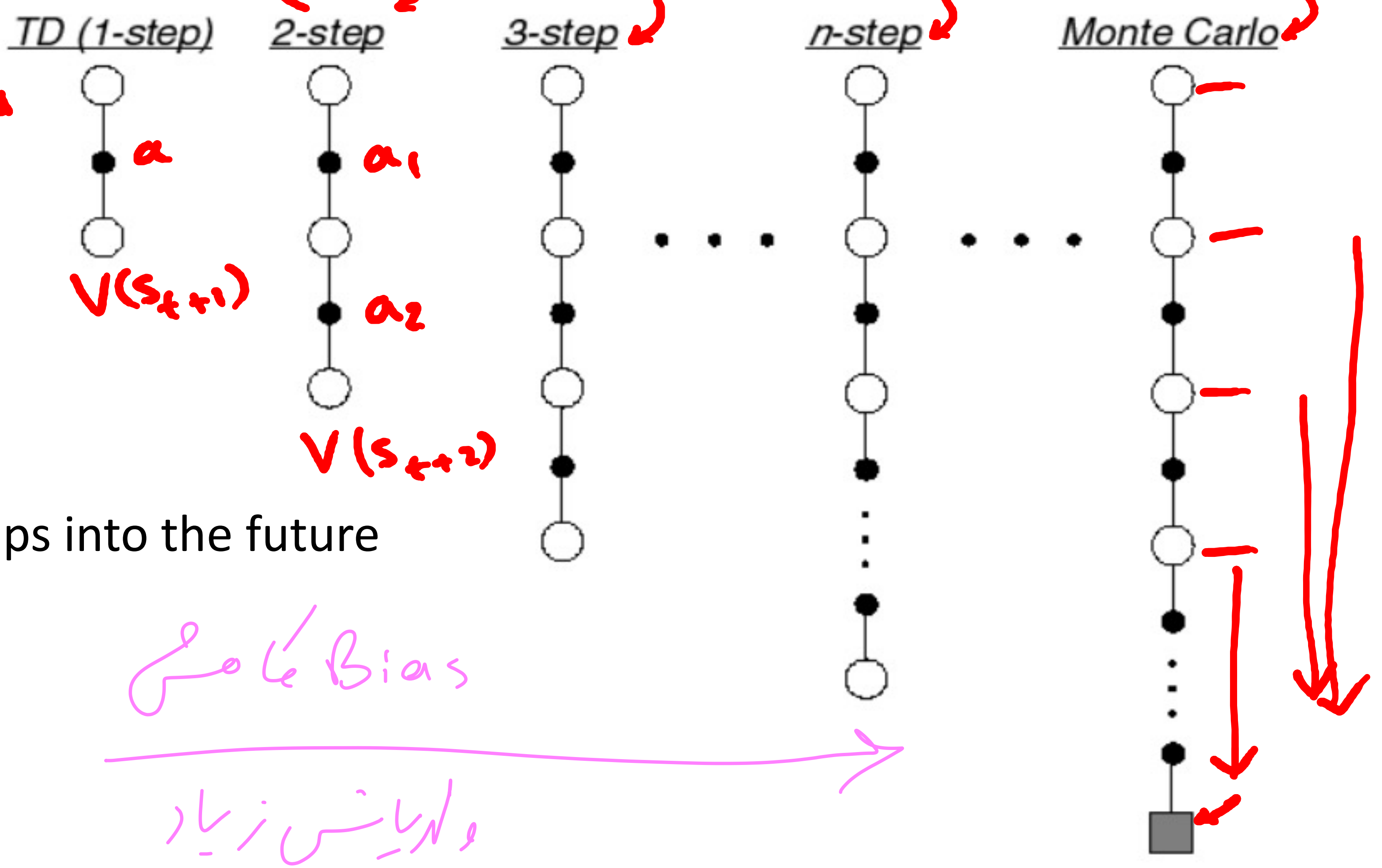
Motivation

- TD uses value estimates which might be inaccurate
- In addition, information can propagate back quite slowly
- In MC information propagates faster, but the updates are noisier
- We can go in between TD and MC

n-Step Prediction

$\overbrace{R_{t+1} + \gamma R_{t+2}}^{\text{Sample}} + \gamma^2 V(S_{t+2})$

$\underbrace{R_{t+1} + \gamma}_{\text{Sample}} V(S_{t+1})$



- Let TD target look n steps into the future

صفر Bias

واریانس زیاد

n-Step Returns → میانگین وزن دار از شاخه ها

- Consider the following n-step returns for n = 1; 2; ...:

from time t onwards & expanding only one action

$$
\begin{aligned}
n &= 1 \quad (TD) \quad & G_t^{(1)} &= R_{t+1} + \gamma v(S_{t+1}) \\
n &= 2 & G_t^{(2)} &= R_{t+1} + \gamma R_{t+2} + \gamma^2 v(S_{t+2}) \\
&\vdots & &\vdots \\
n &= \infty \quad (MC) & G_t^{(\infty)} &= R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{T-t-1} R_T
\end{aligned}
$$

- In general, the n-step return is defined by

$$G_t^{(n)} = R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{n-1} R_{t+n} + \gamma^n v(S_{t+n})$$

- Multi-step temporal-difference learning

$$v(S_t) \leftarrow v(S_t) + \alpha \left(G_t^{(n)} - v(S_t) \right)$$

λ-Returns

$G_t^{(n)} + \sum_{i=1}^{\infty} \lambda^i G_t^{(i+1)}$

- The λ-return G_t^{λ} combines all n-step returns $G_t^{(n)}$
- Using weight $(1-\lambda)\lambda^{n-1}$

$0 \le \lambda \le 1$

وجی‌ها به سمت MC برابر وزن اون

تکه‌شو مشکلی نکنه

TD(0)

$$\underbrace{G_t^{\lambda}}_{TD(1)} = (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_t^{(n)} \rightarrow \sum_{i=0}^{\infty}\lambda^i = \frac{1}{1-\lambda}$$

جمع وزن‌ها 1 میشه

- Forward view TD(λ)

$$V(S_t) \leftarrow V(S_t) + \alpha\left(G_t^{\lambda} - V(S_t)\right)$$

چون MSE بیشتر موثر از

واریانس هست

برابر MC میل کنه به سمت وجی‌ها

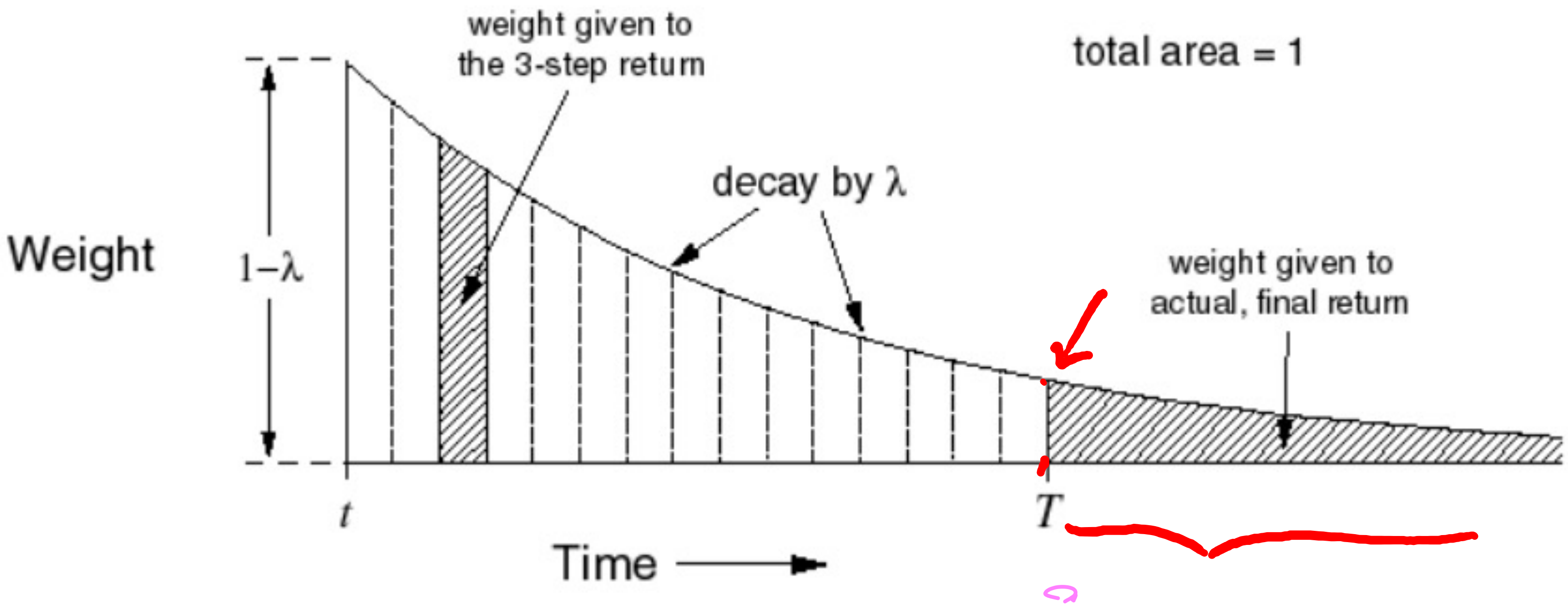
ولی واریانس بیشتر میشه

$\Rightarrow$ وزن اون کمتر میشه

$\lambda \to 1 \to MC$

$\lambda \to 0 \to TD$

TD(λ) Weighting Function



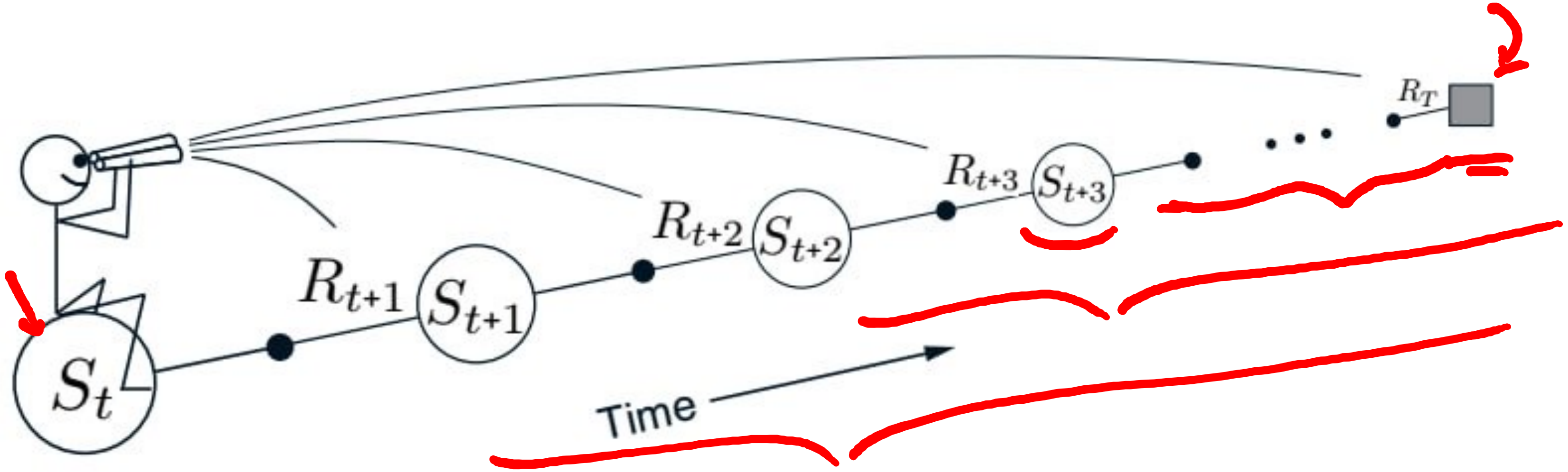
$$G_t^{\lambda} = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} G_t^{(n)}$$

بزرگترش کردیم به انتها از یه جا می‌بریمش
و بقیه‌اش رو حساب می‌کنیم

Forward-view TD(λ)

متد MC باید تا ته بریم

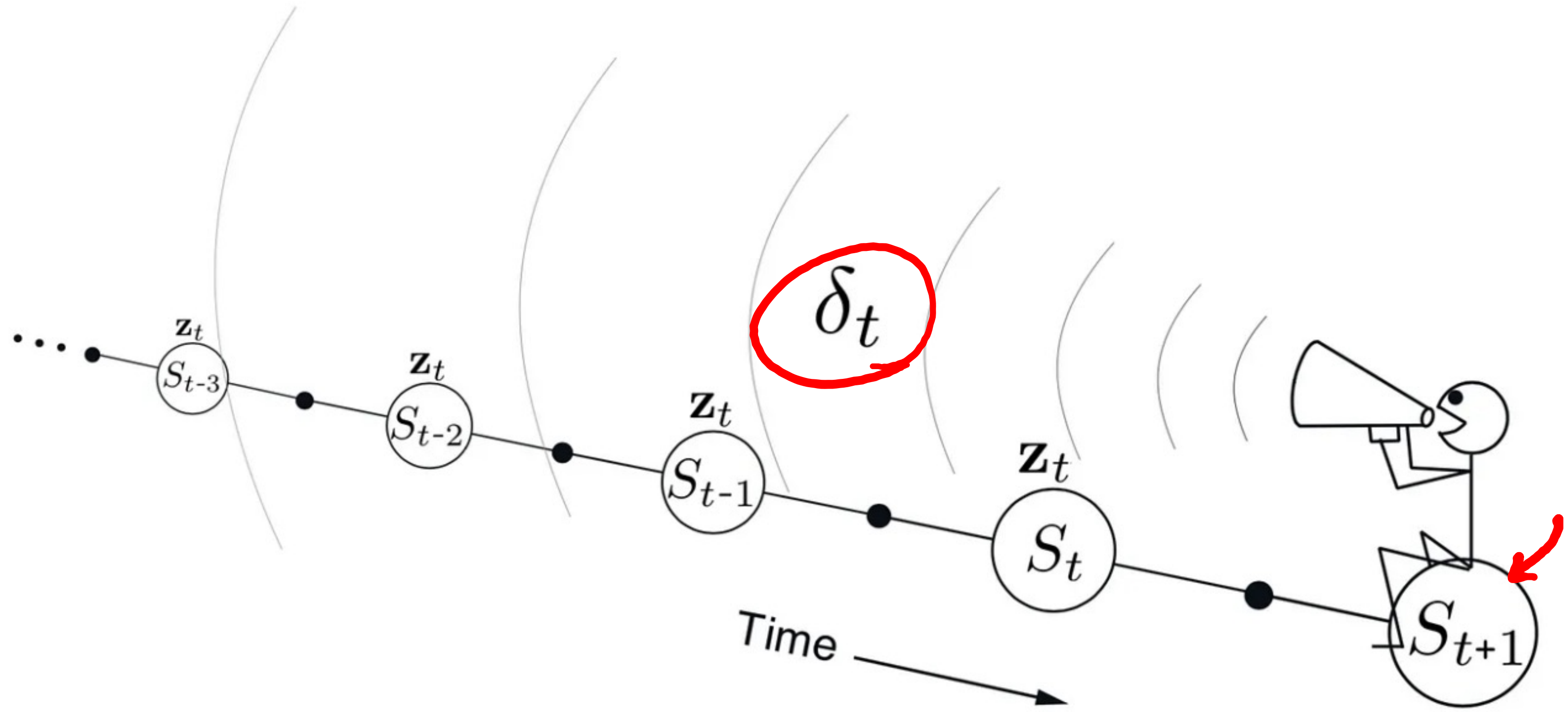
- Update value function towards the λ-return
- Forward-view looks into the future to compute G_t^{λ}
- Like MC, can only be computed from complete episodes



Backward-view TD(λ)

- Forward view provides theory
- Backward view provides mechanism
- Update online, every step, from incomplete sequences

Backward-view TD(λ)



Backward-view TD(λ)

این لحظه به بعد یعنی اطلاع داده شه که از این استفاده کنه

چون این s باید از ⟹

شایستگی

شایستگی این یکی واحد بیشتر میشه

هر وقت یک s دیده میشه

- Keep an eligibility trace for every state s

تأثیر عمل جدید

$$E_0(s) = 0$$
$$E_t(s) = \gamma\lambda E_{t-1}(s) + \mathbf{1}(S_t = s)$$

هر بار یه فاکتور روش می‌خوره

$E_t = [\ ]$

به اندازه تعداد استیت‌ها

- Update value V (s) for every state s in proportion to TD-error δ_t and eligibility trace $E_t(s)$

Original TD

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$$
$$V(s) \leftarrow V(s) + \alpha\delta_t E_t(s)$$

چون تأثیری که یک پاداش به ... گام قبل برسه باید خیلی کمتر باشه

چقدر این s شایسته تأثیره

مقدار تأثیر یک گام برابر همه

TD(λ) and TD(0)

- When $\lambda = 0$, only current state is updated

$$E_t(s) = \mathbf{1}(S_t = s)$$
$$V(s) \leftarrow V(s) + \alpha \delta_t E_t(s)$$

- This is exactly equivalent to TD(0) update

$$V(S_t) \leftarrow V(S_t) + \alpha \delta_t$$

Why Eligibility Traces?

G_i^{λ}

$G_i^{(t)}$

$$\mathcal{T}^\pi_\lambda V(x_0) = (1-\lambda)\mathbb{E}\Big[\sum_{t\geq 0}\lambda^t\Big(\sum_{i=0}^{t}\gamma^i r^\pi(x_i) + \gamma^{t+1}V(x_{t+1})\Big)\Big]$$

$$= \mathbb{E}\Big[\cancel{(1-\lambda)}\sum_{i\geq 0}\gamma^i r^\pi(x_i)\sum_{t\geq i}\lambda^t + \sum_{t\geq 0}\gamma^{t+1}V(x_{t+1})(\lambda^t - \lambda^{t+1})\Big]$$

$$= \mathbb{E}\Big[\sum_{i\geq 0}\lambda^i\big(\gamma^i r^\pi(x_i) + \gamma^{i+1}V(x_{i+1}) - \gamma^i V(x_i)\big)\Big] + V_n(x_0)$$

$$= \mathbb{E}\Big[\sum_{i\geq 0}(\gamma\lambda)^i d_i\Big] + V(x_0),$$

$$\sum_{t\geq i}\lambda^t = \lambda^i \underbrace{\sum_{t\geq 0}\lambda^t}_{\frac{1}{1-\lambda}}$$

Online and Offline Updates

$V(S_t) \leftarrow V(S_t) + \alpha\left(G_t^\lambda - V(S_t)\right)$

- Offline updates
  - Updates are accumulated within episode but applied in batch at the end of episode
- Online updates
  - updates are applied online at each step within episode

Theorem

The sum of offline updates is identical for forward-view and backward-view TD(λ)

$$\sum_{t=1}^{T} \alpha \delta_t E_t(s) = \sum_{t=1}^{T} \alpha \left(G_t^\lambda - V(S_t)\right) \mathbf{1}(S_t = s)$$

Equivalence of Forward and Backward TD in Online and Offline

تنها در نهایت ؟ آنلاین میرن جلو

Offline updates	$\lambda = 0$	$\lambda \in (0,1)$	$\lambda = 1$
Backward view Forward view	TD(0) $\parallel$ TD(0)	TD(λ) $\parallel$ Forward TD(λ)	TD(1) $\parallel$ MC
Online updates	$\lambda = 0$	$\lambda \in (0,1)$	$\lambda = 1$
Backward view Forward view Exact Online	TD(0) $\parallel$ TD(0) $\parallel$ TD(0)	TD(λ) $\nparallel$ Forward TD(λ) $\parallel$ Exact Online TD(λ)	TD(1) $\nparallel$ MC $\parallel$ Exact Online TD(1)

$=$ here indicates equivalence in total update at end of episode.

Temporal Difference Learning

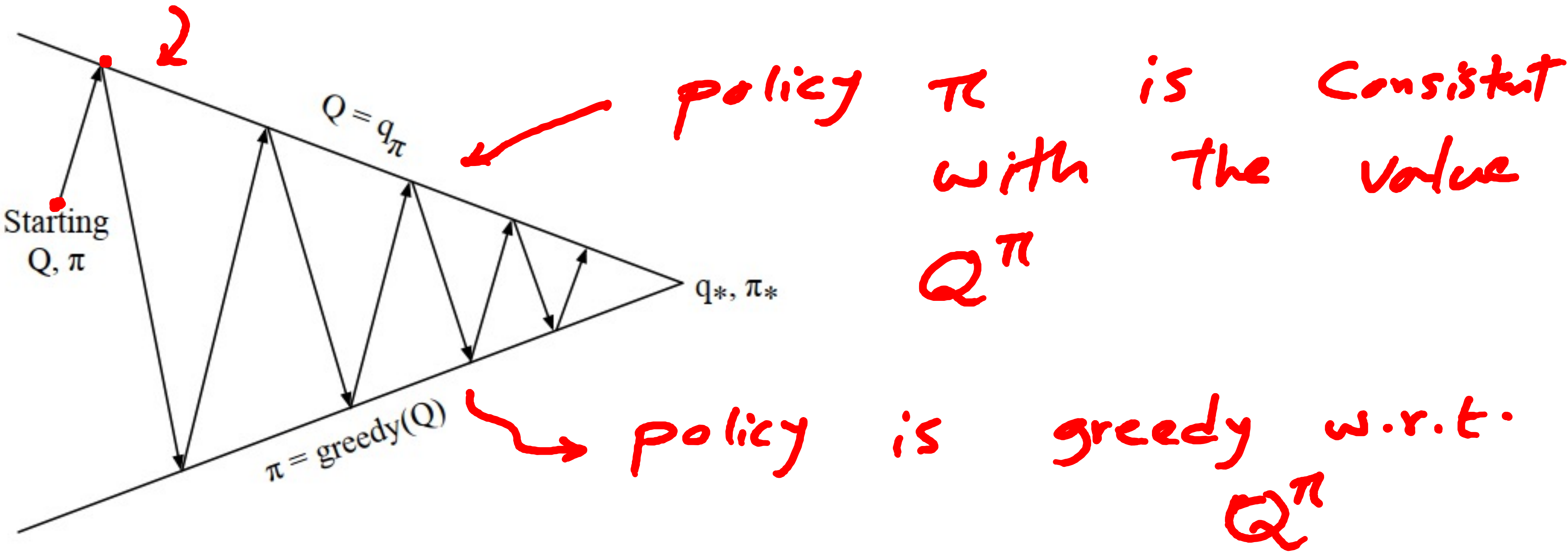
Control

On and Off-Policy Learning

$\pi_b = \pi_e$

- On-policy learning
  - "Learn on the job“
  - Learn about policy π from experience sampled from π
- Off-policy learning $\pi_b \neq \pi_e$
  - "Look over someone’s shoulder“
  - Learn about policy π from experience sampled from μ

Generalized Policy Iteration with MC Evaluation (Review)



Policy evaluation Monte-Carlo policy evaluation, $Q = q_\pi$

Policy improvement Greedy policy improvement?

ε-Greedy Exploration

- Simplest idea for ensuring continual exploration
- All m actions are tried with non-zero probability
- With probability 1 − ε choose the greedy action
- With probability ε choose an action at random

$$\pi(a|s) = \begin{cases} \epsilon/m + 1 - \epsilon & \text{if } a^* = \underset{a \in \mathcal{A}}{\text{argmax}}\, Q(s,a) \\ \epsilon/m & \text{otherwise} \end{cases}$$

ε-Greedy Policy Improvement

Theorem

For any ϵ-greedy policy π, the ϵ-greedy policy π' with respect to q_π is an improvement, $v_{\pi'}(s) \geq v_\pi(s)$

$$
\begin{aligned}
q_\pi(s, \pi'(s)) &= \sum_{a \in \mathcal{A}} \pi'(a|s) q_\pi(s, a) \\
&= \epsilon/m \sum_{a \in \mathcal{A}} q_\pi(s, a) + (1 - \epsilon) \max_{a \in \mathcal{A}} q_\pi(s, a) \\
&\geq \epsilon/m \sum_{a \in \mathcal{A}} q_\pi(s, a) + (1 - \epsilon) \sum_{a \in \mathcal{A}} \frac{\pi(a|s) - \epsilon/m}{1 - \epsilon} q_\pi(s, a) \\
&= \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s, a) = v_\pi(s)
\end{aligned}
$$

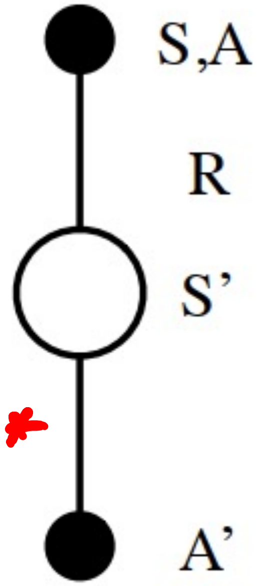
Therefore from policy improvement theorem, $v_{\pi'}(s) \geq v_\pi(s)$

MC vs. TD Control

- Temporal-difference (TD) learning has several advantages over Monte-Carlo (MC)
  - Natural idea: use TD instead of MC in our control loop
  - Online
  - Incomplete sequences
- Natural idea: use TD instead of MC in our control loop
  - Apply TD to Q(S, A)
  - Use ε-greedy policy improvement
  - Update every time-step

Updating Action-Value with SARSA

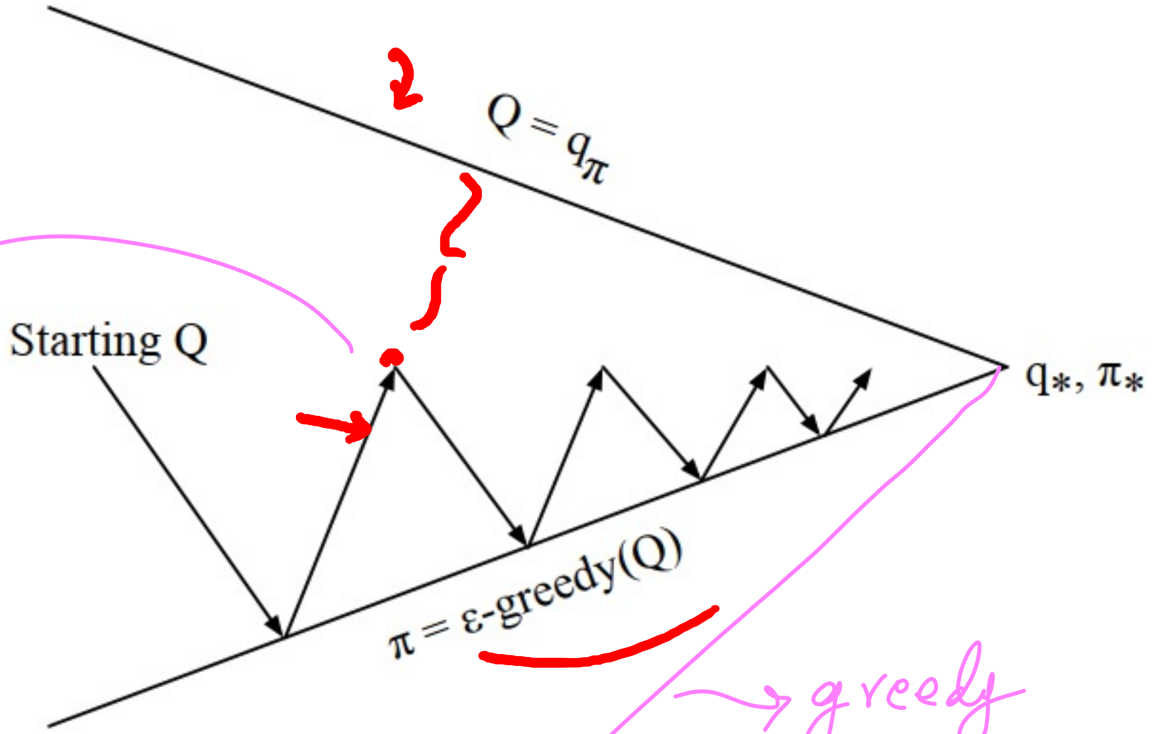
I want to estimate $Q^*(s,a)$ by using some random policy $\pi \neq \pi^*$ for sampling episodes.



$Q^*(\underline{s},\underline{a}) = \mathbb{E}_{s}\left(R(s) + \gamma \max_{a'} Q^*(s',a') \right.$

$$Q(S,A) \leftarrow Q(S,A) + \alpha \left(R + \gamma Q(S',A') - Q(S,A) \right)$$

On-Policy Control with SARSA



یه تخمین از واقعیت هستیم

بین این دو خط ما نوسان میکنیم

Every time-step:

Policy evaluation Sarsa, $Q \approx q_\pi$

Policy improvement ϵ-greedy policy improvement

SARSA Algorithm for On-Policy Control

SARSARSA

Initialize $Q(s, a), \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$, arbitrarily, and $Q(\textit{terminal-state}, \cdot) = 0$
Repeat (for each episode):
  Initialize S
  Choose A from S using policy derived from Q (e.g., ε-greedy)
  Repeat (for each step of episode):
    Take action A, observe R, S'
    Choose A' from S' using policy derived from Q (e.g., ε-greedy)
    $Q(S, A) \leftarrow Q(S, A) + \alpha\big[\underbrace{R + \gamma Q(S', A')}_{\text{sample}} - Q(S, A)\big]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until S is terminal

↳ 1-step

n-Step SARSA

Consider the following n-step returns for $n = 1, 2, \infty$:

$$
\begin{array}{lll}
n = 1 & (Sarsa) & q_t^{(1)} = R_{t+1} + \gamma Q(S_{t+1}) \\
n = 2 & & q_t^{(2)} = R_{t+1} + \gamma R_{t+2} + \gamma^2 Q(S_{t+2}) \\
\vdots & & \vdots \\
n = \infty & (MC) & q_t^{(\infty)} = R_{t+1} + \gamma R_{t+2} + ... + \gamma^{T-1} R_T
\end{array}
$$

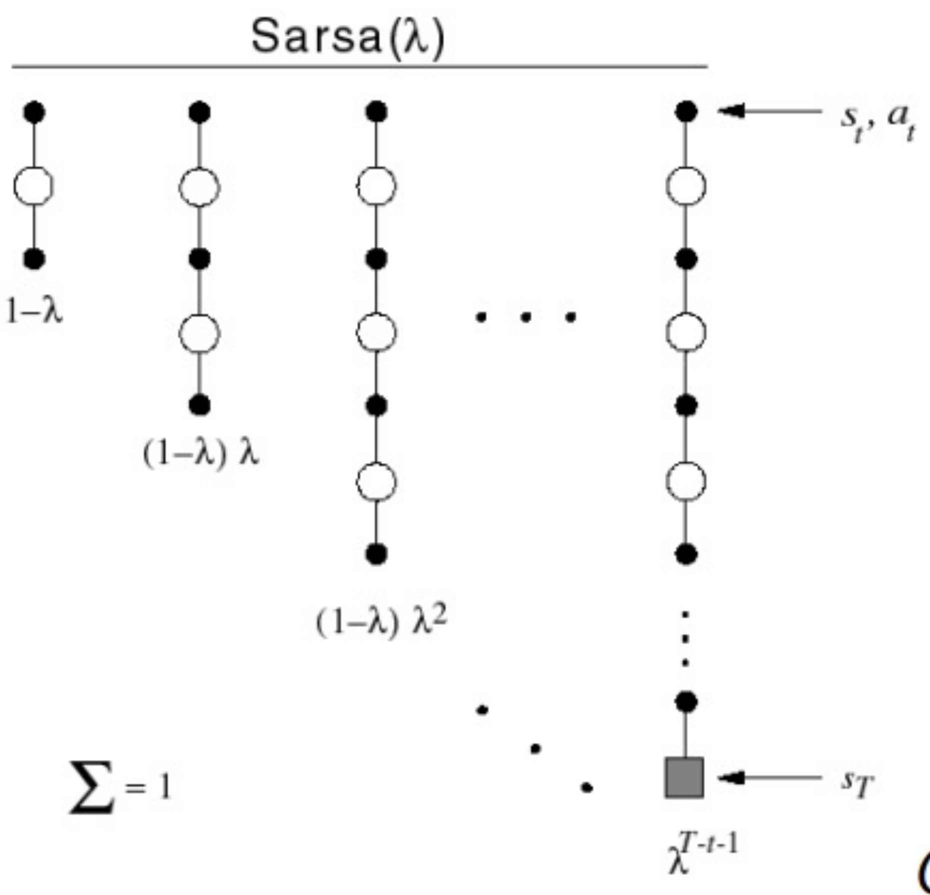
Define the n-step Q-return

$$q_t^{(n)} = R_{t+1} + \gamma R_{t+2} + ... + \gamma^{n-1} R_{t+n} + \gamma^n Q(S_{t+n})$$

n-step Sarsa updates $Q(s, a)$ towards the n-step Q-return

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left(q_t^{(n)} - Q(S_t, A_t) \right)$$

Forward View SARSA(λ)



- The q^λ *return* combines all *n*-step Q-returns $q_t^{(n)}$
- Using weight $(1-\lambda)\lambda^{n-1}$

$$q_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}q_t^{(n)}$$

- Forward-view Sarsa(λ)

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left(q_t^\lambda - Q(S_t, A_t)\right)$$

Backward View SARSA(λ)

- Just like TD(λ), we use eligibility traces in an online algorithm
- But SARSA(λ) has one eligibility trace for each state-action pair

$$E_0(s, a) = 0$$
$$E_t(s, a) = \gamma\lambda E_{t-1}(s, a) + \mathbf{1}(S_t = s, A_t = a)$$

- Q(s, a) is updated for every state s and action a in proportion to TD-error δ_t and eligibility trace $E_t(s, a)$

$$\delta_t = R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)$$
$$Q(s, a) \leftarrow Q(s, a) + \alpha\delta_t E_t(s, a)$$

SARSA(λ) Algorithm

Initialize $Q(s, a)$ arbitrarily, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$
Repeat (for each episode):
  $E(s, a) = 0$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$
  Initialize S, A
  Repeat (for each step of episode):
    Take action A, observe R, S'
    Choose A' from S' using policy derived from Q (e.g., ε-greedy)
    $\delta \leftarrow R + \gamma Q(S', A') - Q(S, A)$
    $E(S, A) \leftarrow E(S, A) + 1$
    For all $s \in \mathcal{S}, a \in \mathcal{A}(s)$:
      $Q(s, a) \leftarrow Q(s, a) + \alpha \delta E(s, a)$
      $E(s, a) \leftarrow \gamma \lambda E(s, a)$
    $S \leftarrow S'$; $A \leftarrow A'$
  until S is terminal

Off-Policy TD and Q-Learning

On and Off-Policy Learning

- On-policy learning
  - Learn about behavior policy π from experience sampled from π.
- Off-policy learning
  - Learn about target policy π from experience sampled from μ.
  - Learn ‘counterfactually’ about other things you could do: “what if...?”
  - e.g., “What if I would turn left?” =) new observations, rewards?

Off-Policy Learning

- Evaluate target policy $\pi(a|s)$ to compute $v_\pi(s)$ or $q_\pi(s,a)$
- While using behavior policy $\mu(a,s)$ to generate actions
- Why is this important?
  - Learn from observing humans or other agents (e.g., from logged data)
  - Re-use experience from old policies (e.g., from your own past experience)
  - Learn about multiple policies while following one policy
  - Learn about greedy policy while following exploratory policy

Q-Learning

- Q-learning estimates the value of the greedy policy

$$q_{t+1}(s,a) = q_t(S_t, A_t) + \alpha_t \left(\underbrace{R_{t+1} + \gamma \max_{a'} q_t(S_{t+1}, a')} - q_t(S_t, A_t) \right)$$

- Acting greedy all the time would not explore sufficiently

Theorem

Q-learning control converges to the optimal action-value function, $q \to q^*$, as long as we take each action in each state infinitely often.

- Note: no need for greedy behavior!
- Works for **any** policy that eventually selects all actions sufficiently often

Q-Learning for Off-Policy Control

Initialize $Q(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$, arbitrarily, and $Q(\textit{terminal-state}, \cdot) = 0$
Repeat (for each episode):
    Initialize S
    Repeat (for each step of episode):
        Choose A from S using policy derived from Q (e.g., ε-greedy)
        Take action A, observe R, S'
        $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$
        $S \leftarrow S'$;
    until S is terminal

از هر پالیسی
حتی رندوم شده
دیتا جمع کن بعد
با هاش value
بهینه رو تخمین بزن

اون وقت ⟸ پالیسی بهینه رو داری