Reinforcement Learning

Lecture 7:
Model-free Learning (2)

Yishay Mansour, Tel-Aviv University

On-policy versus Off-policy

On-policy

- Algorithm selects actions
- One algorithm for
  - Actions
  - Updates

Off-policy

- Separation between selecting
  - Actions
    - Behavioral policy
    - Logs
  - Updates
    - Current belief

Lecture 7: outline

- Temporal Differences:
  - TD(0)
  - TD(λ)
  - $SARSA(\lambda)$
- Importance Sampling
  - Behavioral policy
  - Evaluated policy
- Actor-Critic

Q-learning: Off-Policy Algorithm

- Initialization: arbitrary
  - $Q_0(s, a) = 0$
- Observe (s_t, a_t, r_t, s_{t+1})
- Update:
  - $Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_t(s_t, a_t)\Gamma_t$
  - $\Gamma_t = r_t + \gamma \max_a \{Q_t(s_{t+1}, a)\} - Q_t(s_t, a_t)$

SARSA: On-Policy Algorithm

- Initialization: arbitrary
  - $Q_0(s, a) = 0$
- Observe $(s_t, a_t, r_t, s_{t+1}, a_{t+1})$
  - $a_{t+1} = \pi(s_{t+1}; Q_t)$ output of the on-policy
- Update:
  - $Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_t(s_t, a_t)\Gamma_t$
  - $\Gamma_t = r_t + \gamma Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t)$

Monte Carlo: Algorithm

- Mainly for episodic MDP
- Given observations
  - $G_1(s), \ldots, G_m(s)$
  - Estimate $\hat{V}^{\pi}(s) = \frac{1}{m}\sum_{i=1}^{m} G_i\,(s)$
- First Visit
- Every Visit

Temporal Differences

TD: methods

- Learns value function
  - directly from experience
- Model-free
- Uses incomplete episodes
  - Updates before the end
- Updates the estimate
  - Given the change in estimates

TD(0): motivation

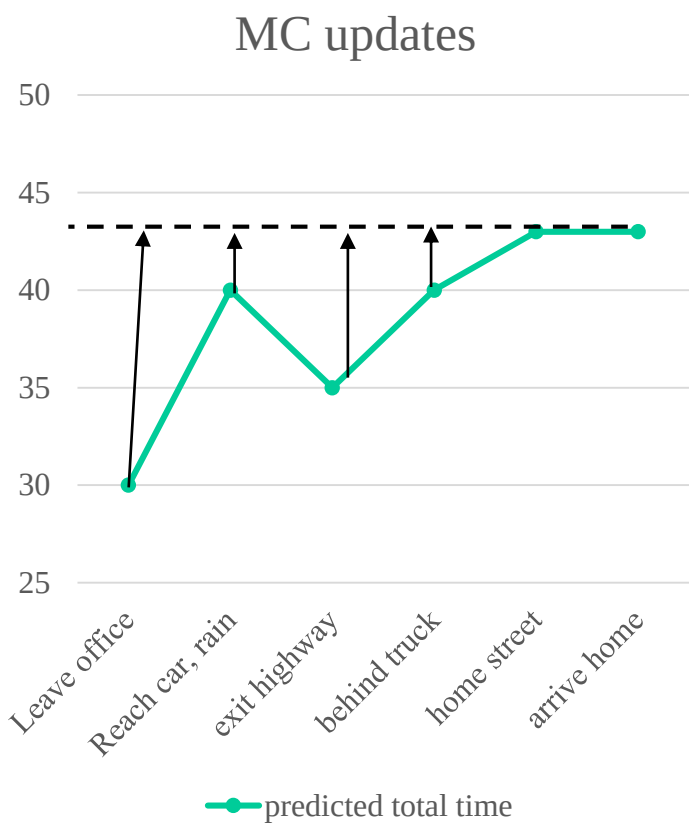
- Fix a policy π
- Value Iteration
  - $V_{t+1}(s) = E^{\pi}[r(s,a) + \gamma V_t(s')]$
    - Convergence $V_t \to V^{\pi}$
- Assume we sample (s_t, a_t, r_t, s_{t+1})
  - $E^{\pi}\left[r_t + \gamma \hat{V}_t(s_{t+1}) \middle| s_t\right] = E^{\pi}[r(s,a) + \gamma \hat{V}_t(s')]$
    - Where $s = s_t$, and $a = a_t = \pi(s_t)$
  - TD(0): Do an update in that direction

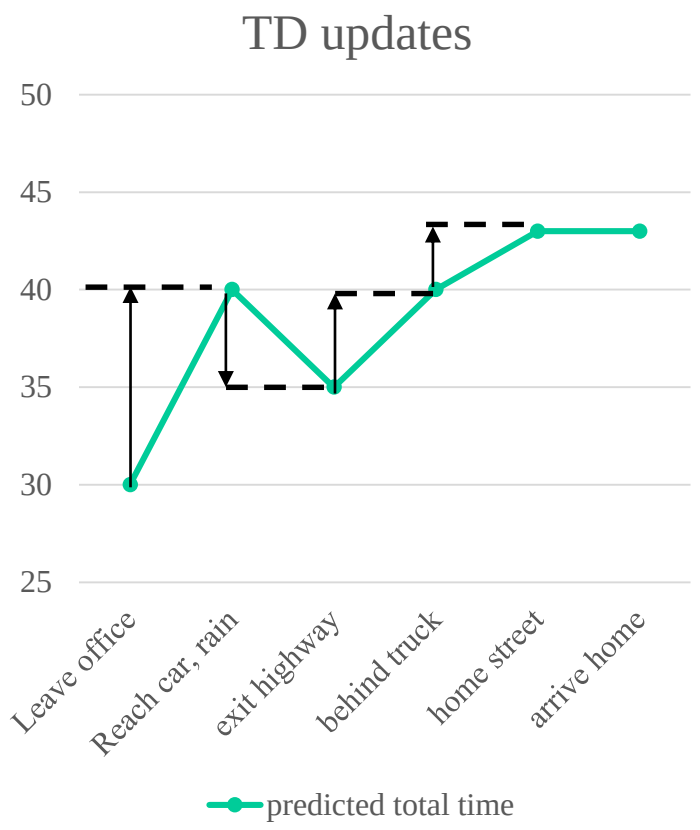
TD(0): Algorithm

- Maintains $\hat{V}$
- Update using (s_t, a_t, r_t, s_{t+1})
  - $\hat{V}(s_t) = (1-\alpha_t)\hat{V}(s_t) + \alpha_t[r_t + \gamma\hat{V}(s_{t+1})]$
  - $\hat{V}(s_t) = \hat{V}(s_t) + \alpha_t[r_t + \gamma\hat{V}(s_{t+1}) - \hat{V}(s_t)]$
    - Step size $\alpha_t(s, a)$
  - $\Delta_t = r_t + \gamma\hat{V}(s_{t+1}) - \hat{V}(s_t)$
- TD(0): $\hat{V}(s_t) = \hat{V}(s_t) + \alpha_t\Delta_t$

TD: Motivating example

state	Elapsed time (minutes)	Predicted Time to Go	Predicted Total Time
Leaving office	0	30	30
Reach car, rain	5	35	40
Exit highway	20	15	35
Behind truck	30	10	40
Home street	40	3	43
Arrive home	43	0	43


MC updates
50
45
40
35
30
25
Leave office
Reach car, rain
exit highway
behind truck
home street
arrive home
predicted total time


TD updates
50
45
40
35
30
25
Leave office
Reach car, rain
exit highway
behind truck
home street
arrive home
predicted total time

TD(0) vs MC

- Example:
  - Assume we observe a Markov Chain episodes:
    - (A,0,B,0,C) once
    - (B,1,D) six times
    - (B,0,C) once
  - Monte-Carlo estimates:
    - $V(B) = \frac{1 \cdot 6 + 0 \cdot 2}{8} = \frac{3}{4}$
    - $V(A) = \frac{0}{1} = 0$

TD(0) vs MC

- TD estimates
  - Init $\hat{V}(A) = \hat{V}(B) = 0$
    - $\gamma = 1; \alpha = 0.1$
  - After $(A, 0, B, 0, C)$

$\hat{V}(B) = 0 + \alpha\,(0 + \gamma \cdot 0 - 0) = 0$

  - After $(B, 0, C)$

$\hat{V}(B) = 0 + \alpha\,(0 + \gamma \cdot 0 - 0) = 0$

  - After $(B, 1, D)$

$\hat{V}(B) = 0 + \alpha\,(1 + \gamma \cdot 0 - 0) = \alpha$

- Running to stabilization
  - Resample episodes

$\frac{6}{8}\alpha\left(1 + \gamma \cdot 0 - \hat{V}(B)\right) +$

$\frac{2}{8}\alpha\left(0 + \gamma \cdot 0 - \hat{V}(B)\right) = 0$

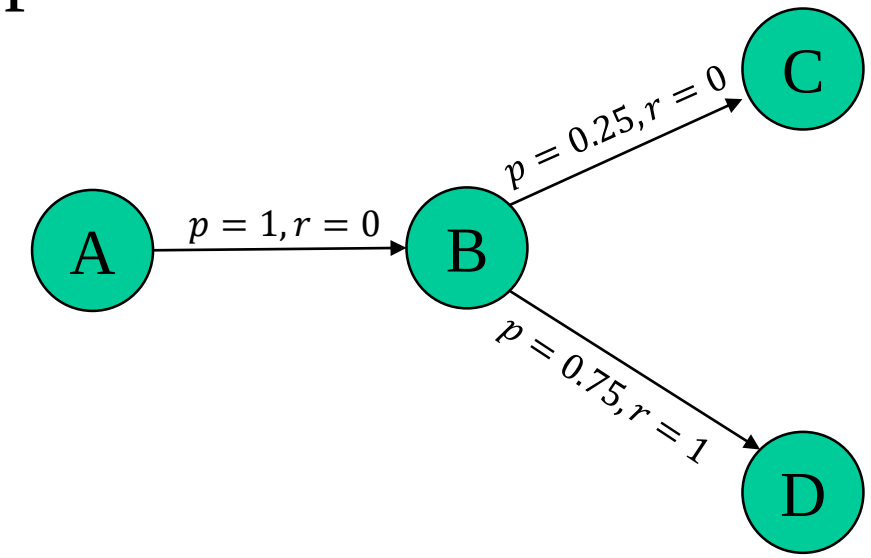
  - $\hat{V}(B) = 0.75$

  - $\hat{V}(A) = \gamma\hat{V}(B) = 0.75\gamma$

$\alpha\left(0 + \gamma \cdot \hat{V}(B) - \hat{V}(A)\right) = 0$

TD and empirical model

- For the example:



Empirical model

- Empirical model
  - Given (s_t, a_t, r_t, s_{t+1})
    - $\hat{r}(s,a) = \frac{1}{n(s,a)} \sum_{s_t=s, a_t=a} r_t$
      - $n(s,a) = |\{t: s_t = s, a_t = a\}|$
    - $\hat{p}(s'|s,a) = \frac{n(s,a,s')}{n(s,a)}$
      - $n(s,a,s') = |\{t: s_t = s, a_t = a, s_{t+1} = s'\}|$

TD(0) and empirical model

- Theorem:
  - run TD(0) on the sample until convergence
    - Re-sampling from sample,
  - Then TD(0) estimate $\hat{V}^{\pi}$ equal to V^{π} in the empirical model

- Recall that MC used the reduced empirical model !

TD(0) and empirical model

- Proof (sketch):
- Recall that TD(0): $\hat{V}(s_t) = \hat{V}(s_t) + \alpha_t \Delta_t$
  - $\Delta_t = r_t + \gamma \hat{V}(s_{t+1}) - \hat{V}(s_t)$
  - At convergence we have $E[\Delta_t] = 0$
  - $\hat{V}(s_t) = E\left[r_t + \gamma \hat{V}(s_{t+1})\right] = \hat{r}(s,a) + \gamma E_{s' \sim \hat{p}(s'|s,a)}[\hat{V}(s')]$
    - Where $s = s_t; a = a_t$ and the expectation is w.r.t. the empirical model

TD(0): Convergence

- **Theorem:**
  - Step size, for every *(s,a):*

$$\sum_t \alpha_t(s,a) = \infty \quad \text{and} \quad \sum_t \alpha_t^2(s,a) = O(1)$$

  - **$V_t(s)$ Converges with probability 1 to $V^{\pi}(s)$**

- Very similar to Q-learning.

TD(0): contraction

- Operator (for a given policy π)
  - $(Hv)(s) = r(s,a) + \gamma \sum_{s'} p(s'|s,a)v(s')$
    - Where $a = \pi(s)$
- Contraction:
  - $\|Hv_1 - Hv_2\|_\infty = \gamma \max_{s,a} |\sum_{s'} p(s'|s,a)[v_1(s') - v_2(s')]|$
  - $\leq \gamma \max_{s,a} \max_{s'} |v_1(s') - v_2(s')|$
  - $\leq \gamma \|v_1 - v_2\|_\infty$

TD(0): convergence

- Rewrite TD(0):
  - $V_{t+1}(s_t) = (1-\alpha_t)V_t(s_t, \pi(s_t)) + \alpha_t \Phi_t$
  - $\Phi_t = r_t + \gamma V_t(s_{t+1})$
  - $E[\Phi_t] = (HV_t)(s_t)$
  - $w_t(s_t) = \Phi_t - (HV_t)(s_t)$
  - $E[w_t] = 0$, and $|w_t| \leq \frac{R_{max}}{1-\gamma}$
  - $\Phi_t = (HV_t)(s_t) + w_t$

TD(0): Convergence

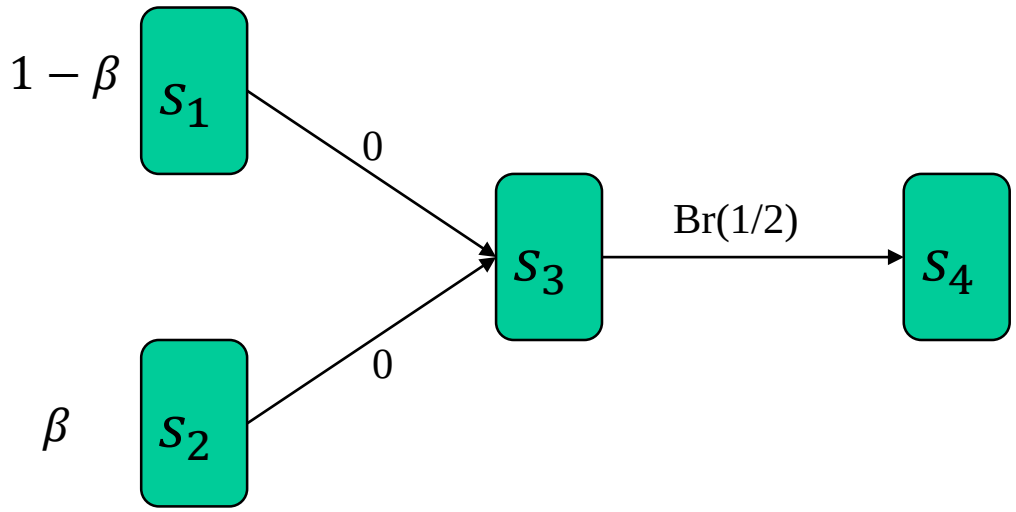
- Use *stochastic approximation* convergence
  - Step size: by assumption
  - Noise: $E[w_t] = 0$ & $|w_t| \leq V_{max} = \frac{R_{max}}{1-\gamma}$
  - Contraction:
    - operator H is γ-contracting with V^{π} fixed-point
- Result: V_t converges to V^{π} with prob. 1
  - Assumes "exploration"

TD(0) vs MC vs DP

- Three ways to look at the value function $V^{\pi}(s)$
  - (MC): $E^{\pi}[R_t|s_t = s]$
  - TD(0): $E^{\pi}[r(s,a) + \gamma V^{\pi}(s')|s_t = s, a_t = a, s' = s_{t+1}]$
  - (DP): $\sum_{a\in A}\pi(a|s)[r(s,a) + \gamma \sum_{s'} p(s'|s,a)V^{\pi}(s')]$

TD(0) vs MC: converges rate

- Time to reach $|\hat{V}(s_2) - \frac{\gamma}{2}| \approx \epsilon$
  - MC: $\frac{1}{\beta\epsilon^2}$
  - TD(0):$\frac{1}{\epsilon^2} + \frac{1}{\beta}$



Lecture 7: outline

- Temporal Differences:
  - TD(0)
  - TD(λ)
  - $SARSA(\lambda)$
- Importance Sampling
  - Behavioral policy
  - Evaluated policy
- Actor-Critic

TD: multiple look-ahead

TD: multiple look-ahead

- TD(0) update:
  - Given (s_t, a_t, r_t, s_{t+1})
  - $\Delta_t = R_t^{(1)}(s_t) - \hat{V}(s_t)$
    - $R_t^{(1)}(s_t) = r_t + \gamma \hat{V}(s_{t+1})$
- Two step look-ahead
  - Given $(s_t, a_t, r_t, s_{t+1}, a_{t+1}, r_{t+1}, s_{t+2})$
  - $R_t^{(2)}(s_t) = r_t + \gamma r_{t+1} + \gamma^2 \hat{V}(s_{t+2})$

TD: multiple look-ahead

- Generalize it to *n* steps:
  - $R_t^{(n)}(s_t) = \sum_{i=0}^{n-1} \gamma^i r_{t+i} + \gamma^n \hat{V}(s_{t+n})$
- New updates:
  - $\Delta_t^{(n)} = R_t^{(n)}(s_t) - \hat{V}(s_t)$
  - $\Delta_t^{(n)} = \sum_{i=0}^{n-1} \gamma^i \Delta_{t+i}$
    - $= \sum_{i=0}^{n-1} \gamma^i r_{t+i} + \sum_{i=0}^{n-1} \gamma^{i+1} \hat{V}_{t+i}(s_{t+i+1}) - \sum_{i=0}^{n-1} \gamma^i \hat{V}_{t+i}(s_{t+i})$
    - $= \sum_{i=0}^{n-1} \gamma^i r_{t+i} + \gamma^n \hat{V}(s_{t+n}) - \hat{V}(s_t) = \Delta_t^{(n)}$

TD: multiple look-ahead

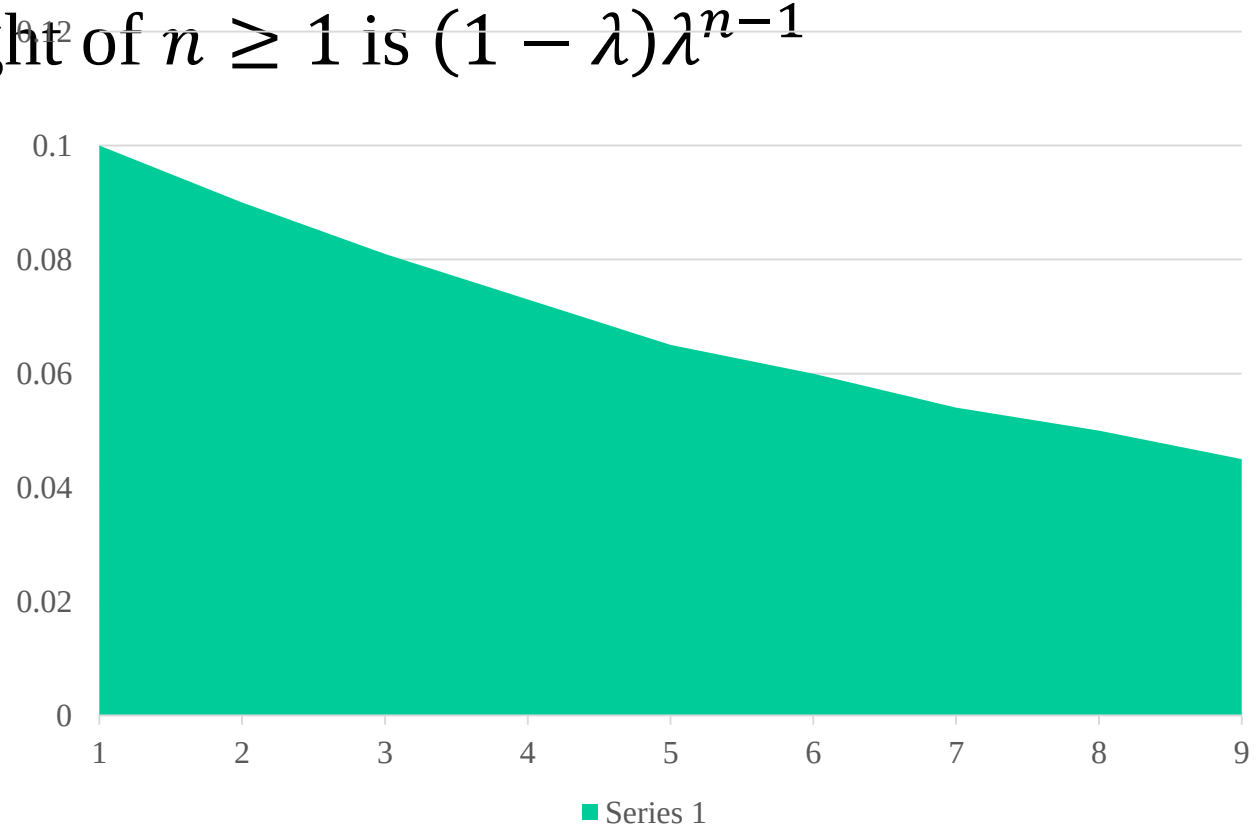
- Using n step look-ahead
  - $\hat{V}(s_t) = \hat{V}(s_t) + \alpha_t \Delta_t^{(n)}$
  - $\Delta_t^{(n)} = R_t^{(n)}(s_t) - \hat{V}(s_t)$
- The operator $R_t^{(n)}$ is γ^n-contracting
  - $\left\| R_t^{(n)}(V_1) - R_t^{(n)}(V_2) \right\| \leq \gamma^n \|V_1 - V_2\|_\infty$

TD: how to select the look-ahead

- We can use any parameter n
  - If episodes ends before, pad with zeros
  - Actually, $n = \infty$ is simply MC
- Basic idea: we can average over n
  - What is the best way to average?
  - The simplest is exponential averaging!

TD: exponential averaging

- Have a parameter $\lambda \in (0,1)$
  - Weight of $n \geq 1$ is $(1-\lambda)\lambda^{n-1}$



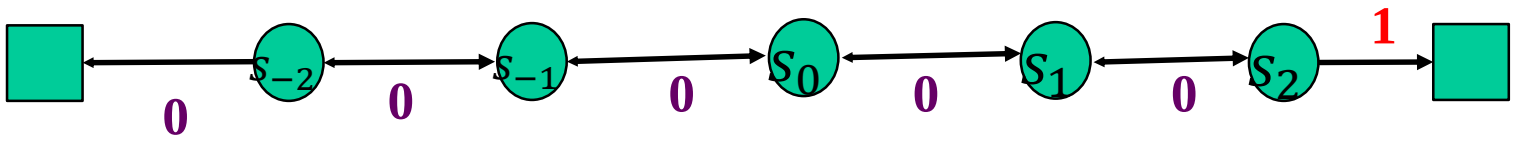
$TD(\lambda)$

- Recall:
  - $\Delta_t^{(n)} = R_t^{(n)}(s_t) - \hat{V}(s_t)$
- $TD(\lambda)$ update:
  - $\hat{V}(s_t) = \hat{V}(s_t) + \alpha_t(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\ \Delta_t^{(n)}$

- Do not confuse γ and λ

Random walk example

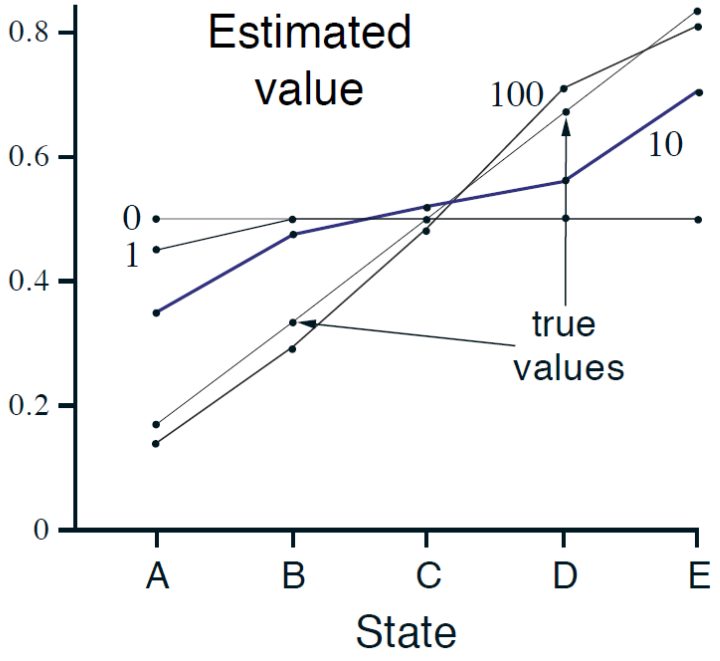
- Random walk on a line

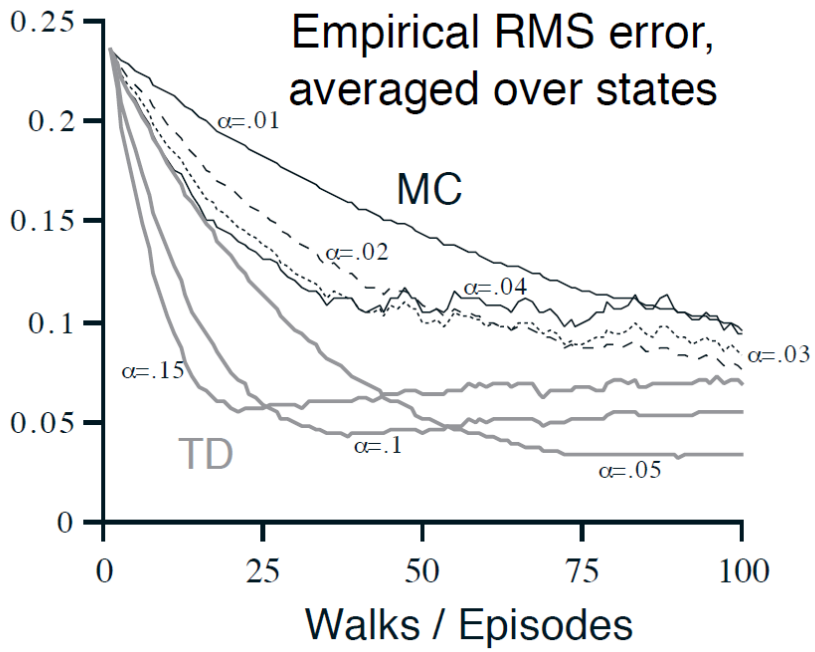


- Policy: random
  - Left/right with prob 0.5
- Values:

$$V(s_{-2}) = \frac{1}{6}; V(s_{-1}) = \frac{2}{6}; V(s_0) = \frac{3}{6}; V(s_1) = \frac{4}{6}; V(s_2) = \frac{5}{6}$$

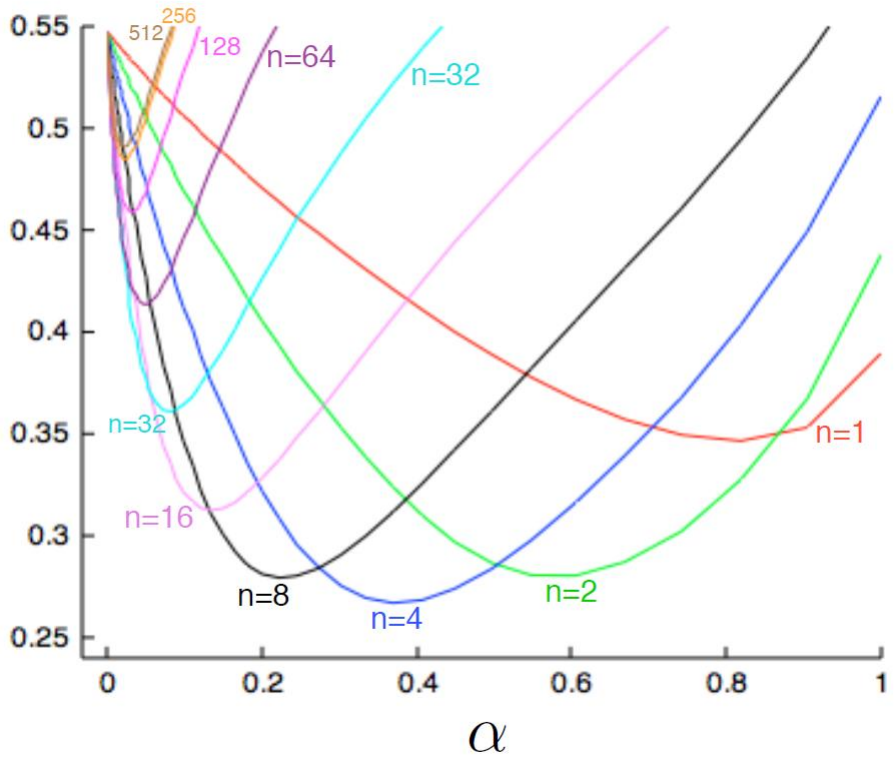
Random walk: TD(0) vs MC



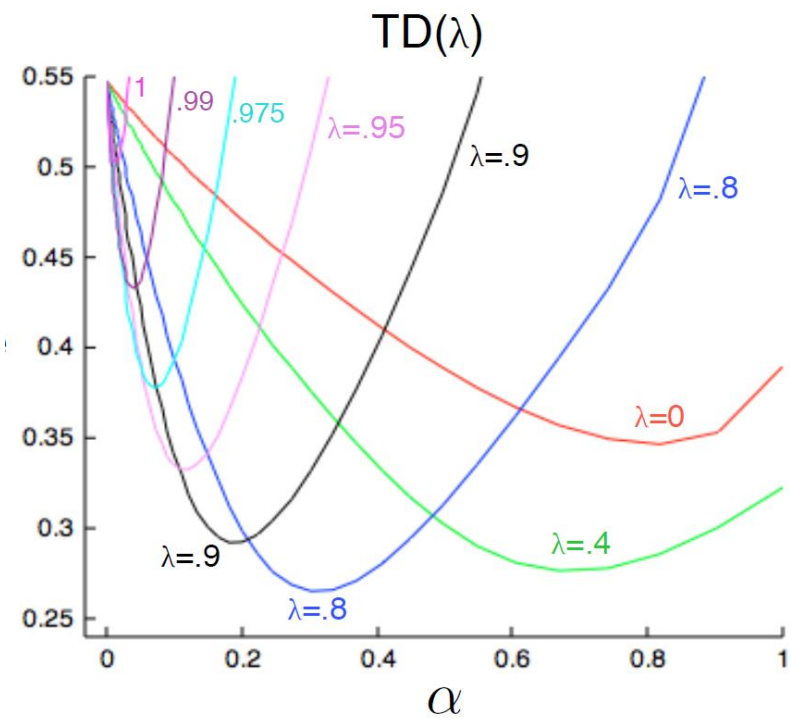


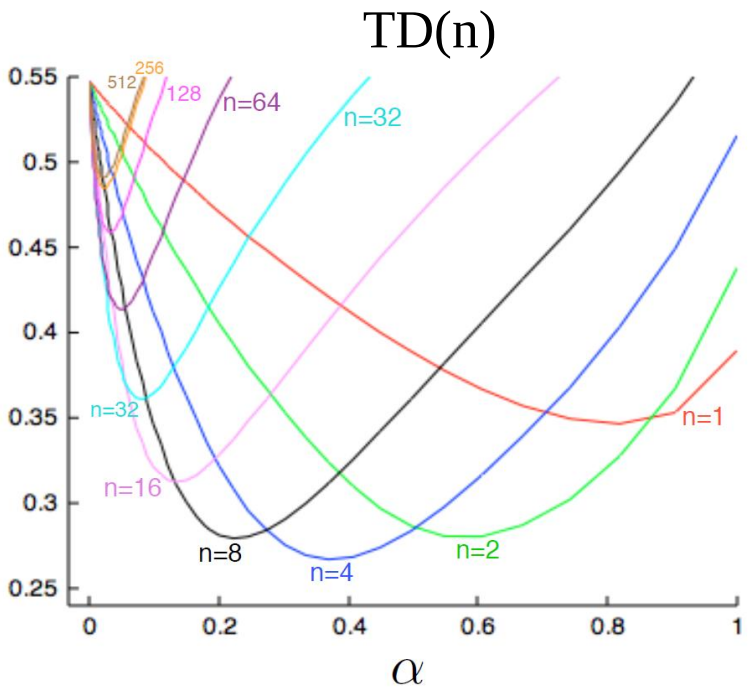
Random walk: $TD(n)$

- Random walk on 19 states
- Average RMS error
- 10 episodes



Random walk: $TD(\lambda)$ and $TD(n)$





Lecture 7: outline

- Temporal Differences:
  - TD(0)
  - TD(λ)
  - $SARSA(\lambda)$
- Importance Sampling
  - Behavioral policy
  - Evaluated policy
- Actor-Critic

$TD(\lambda)$: Forward view

- $TD(\lambda)$ update:
  - $\hat{V}(s_t) = \hat{V}(s_t) + \alpha(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\ \Delta_t^{(n)}$
  - $\Delta_t^{(n)} = R_t^{(n)}(s_t) - \hat{V}(s_t)$
- Look-ahead:
  - How can we know future rewards
  - Do we need to wait?
  - For $TD(\lambda)$ we depend on all future rewards!

$TD(\lambda)$: Backward view

- Forward view
  - Theoretical justification
- Backward view:
  - Online updates
    - Every time
    - Using incomplete information
- Performance: They will be equivalent
  - At the end of the episode

$TD(\lambda)$ backward updates

- Basic idea
  - The updates are linear in Temporal Differences
  - Fix a time t and state $s = s_t$
    - TD: $\Delta_t = r_t + \gamma V_t(s_{t+1}) - V_t(s_t)$
    - How is it weighted by previous time steps?
  - Influences only $s_\tau = s = s_t$
    - influences by $(\gamma\lambda)^{t-\tau}\Delta_t$
  - We can sum all the influences
    - $e_t(s) = \sum_{\tau=0}^{t}(\gamma\lambda)^{t-\tau} I(s_\tau = s)$

$TD(\lambda)$: Eligibility traces

- Define eligibility trace
  - $e_t(s) = \sum_{\tau=0}^{t} (\gamma\lambda)^{t-\tau} I(s_\tau = s)$
- Compute it online:
  - $e_t(s) = \lambda\gamma e_{t-1}(s) + I(s_t = s)$
- Update
  - $\hat{V}_{t+1}(s) = \hat{V}_t(s) + \alpha_t e_t(s)\Delta_t$
- TD(0): $\hat{V}_{t+1}(s_t) = \hat{V}_t(s_t) + \alpha_t I(s_t = s)\Delta_t$

$TD(\lambda)$: algorithm

- Initialization
  - Arbitrary $\hat{V}(s)$, $e_0(s) = 0$
- Update: observe (s_t, a_t, r_t, s_{t+1})
  - $\Delta_t = r_t + \gamma \hat{V}_t(s_{t+1}) - \hat{V}_t(s_t)$
  - $e_t(s) = \gamma \lambda e_{t-1}(s) + I(s_t = s)$
  - $\hat{V}_{t+1}(s) = \hat{V}_t(s) + \alpha_t e_t(s) \Delta_t$

$TD(\lambda)$: Equivalence of Forward and backward view

- Forward view
  - $\Delta V_t^F(s) = \alpha[R_t^\lambda - V_t(s)]$
    - $R_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} R_t^{(n)}$
- Backward view
  - $\Delta V_t^B(s) = \alpha \Delta_t e_t(s)$
    - $e_t(s) = \sum_{k=0}^{t} (\lambda\gamma)^{t-k} I(s = s_k)$
- Theorem: $\sum_{t=0}^{T-1} \Delta V_t^B(s) = \sum_{t=0}^{T-1} \Delta V_t^F(s) I(s_t = s)$

- Theorem:
  - $\sum_{t=0}^{T-1} \Delta V_t^B(s) = \sum_{t=0}^{T-1} \Delta V_t^F(s) I(s_t = s)$
- Proof: Forward view
  - $\sum_{t=0}^{T-1} \Delta V_t^F(s)$
  - $= \sum_{t=0}^{T-1} \alpha (1-\lambda) \sum_{n=t}^{T-1} \lambda^{n-t} \Delta_t^{(n)} I(s = s_t)$
  - $= \sum_{n \geq t} \alpha (\gamma\lambda)^{n-t} \Delta_n(s) I(s = s_t)$
- Backward view
  - $\sum_{t=0}^{T-1} \Delta V_t^B(s)$
  - $= \sum_{t=0}^{T-1} \alpha \Delta_t(s) \sum_{n=0}^{t} (\gamma\lambda)^{t-n} I(s = s_n)$
  - $= \sum_{t \geq n} \alpha (\gamma\lambda)^{t-n} \Delta_t(s) I(s = s_n)$
- Interchange: $n \leftrightarrow t$

$TD(\lambda)$: Summary

- Updates more frequently frequent states
  - Higher eligibility traces e_t
- Single parameter λ
  - Has the potential to improve over both:
    - TD(0)
    - MC
- Convergence:
  - can be shown

$SARSA$: n-step look-ahead

- Another application of eligibility traces
- SARSA update
  - $r(s_t, a_t) + \gamma Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t)$
- SARSA with n-step look-ahead
  - $q_t^{(n)} = \sum_{i=0}^{n-1} \gamma^i r(s_{t+i}, a_{t+i}) + \gamma^n Q_t(s_{t+n}, a_{t+n})$
  - $Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_t(q_t^{(n)} - Q_t(s_t, a_t))$

$SARSA(\lambda)$: Forward view

- Let q_t^{λ} be a weighted average of $q_t^{(n)}$
  - Using exponential weights
  - $q_t^{\lambda} = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1}\, q_t^{(n)}$
- Forward view of $SARSA(\lambda)$
  - $Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_t(q_t^{\lambda} - Q_t(s_t, a_t))$

$SARSA(\lambda)$: Backward view

- Use eligibility traces
  - $e_0(s,a) = 0$
  - $e_t(s,a) = \gamma\lambda e_{t-1}(s,a) + I(s_t = s, a_t = a)$
- Update of $Q(s,a)$
  - $\Delta_t = r_t + \gamma Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t)$
  - $Q_{t+1}(s,a) = Q_t(s_t, a_t) + \alpha_t \Delta_t e_t(s,a)$

Importance sampling

Importance sampling

- Using one distribution to evaluate another
  - Sampling using Q
  - Evaluation expectation w.r.t. P
- Derivation
  - $E_{X\sim P}[f(X)] = \sum_x P(x)f(x)$
  - $= \sum_x Q(x)\frac{P(x)}{Q(x)}f(x)$
  - $E_{X\sim Q}[\frac{P(X)}{Q(X)}f(X)]$

Importance sampling

- Unbiased estimator !
  - Expectation is perfect
- Variance
  - Can be huge
    - Depends on $P(x)/Q(x)$

Importance sampling: MC

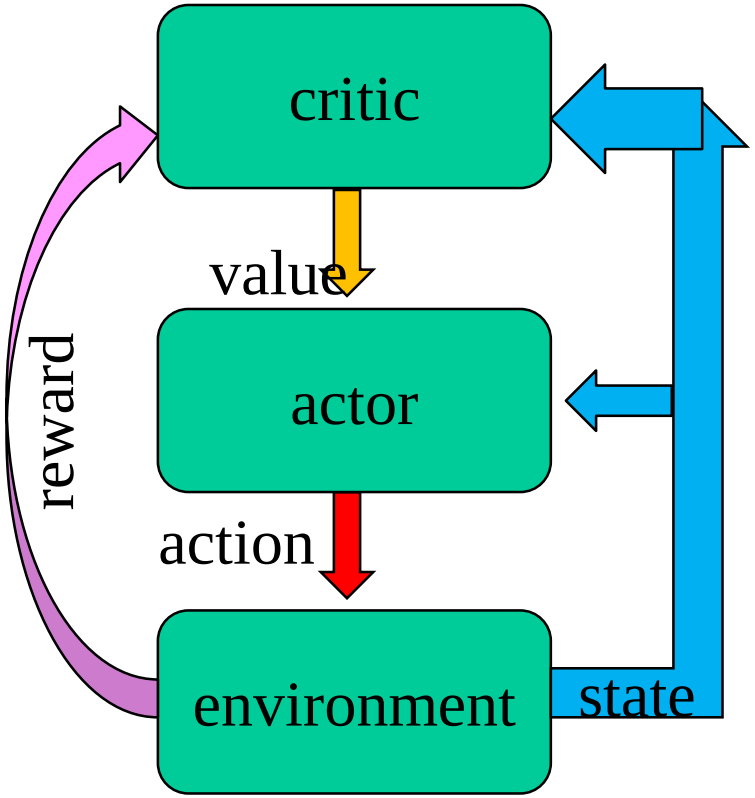
- Action selection policy π
- Evaluated policy ρ
- Estimated return
  - $\frac{\rho(s_1,a_1,r_1,\ldots,s_T,a_T,r_T)}{\pi(s_1,a_1,r_1,\ldots,s_T,a_T,r_T)} = \prod_{t=1}^{T} \frac{\rho(a_t|s_t)}{\pi(a_t|s_t)}$
  - $G^{\rho/\pi} = \prod_{t=1}^{T} \frac{\rho(a_t|s_t)}{\pi(a_t|s_t)} (\sum_{t=1}^{T} r_t)$
  - $\hat{V}^{\rho}(s_0) = \hat{V}^{\rho}(s_0) + \alpha(G^{\rho/\pi} - \hat{V}^{\rho}(s_0))$
- Might have huge updates

Importance sampling: TD

- Action selection policy π
- Evaluated policy ρ
- Re-weight the observation
  - $r_t + \gamma \hat{V}^{\rho}(s_{t+1})$
  - $\Delta_t = \frac{\rho(a_t|s_t)}{\pi(a_t|s_t)}\left[r_t + \gamma \hat{V}^{\rho}(s_{t+1})\right] - \hat{V}^{\rho}(\mathrm{s_t})$
- Much smaller variance!

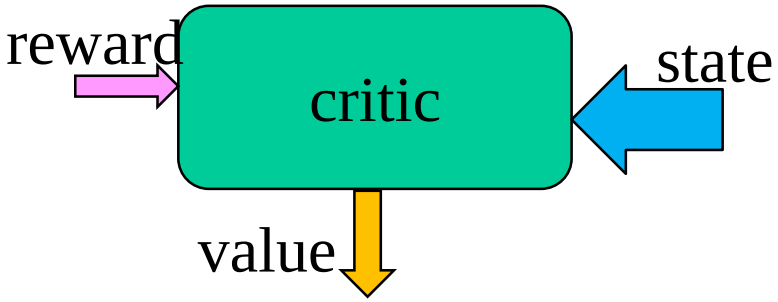
Actor-Critic Methodology

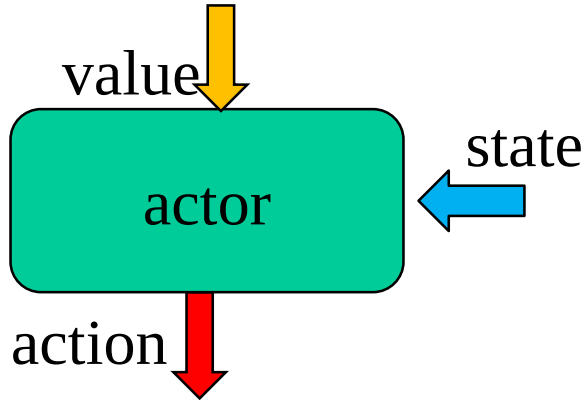
- A general methodology to design RL algorithms
- Critic:
  - Computes value function
- Actor
  - Selects action
  - Improves policy



Actor-Critic Methodology

- Critic:
  - Input: state and reward
  - Goal: evaluate current policy
  - Method: TD
- Actor
  - Input: value and state
  - Goal: improve policy
  - Method: Near-greedy





Lecture 7: outline

- Temporal Differences:
  - TD(0)
  - TD(λ)
  - $SARSA(\lambda)$
- Importance Sampling
  - Behavioral policy
  - Evaluated policy
- Actor-Critic