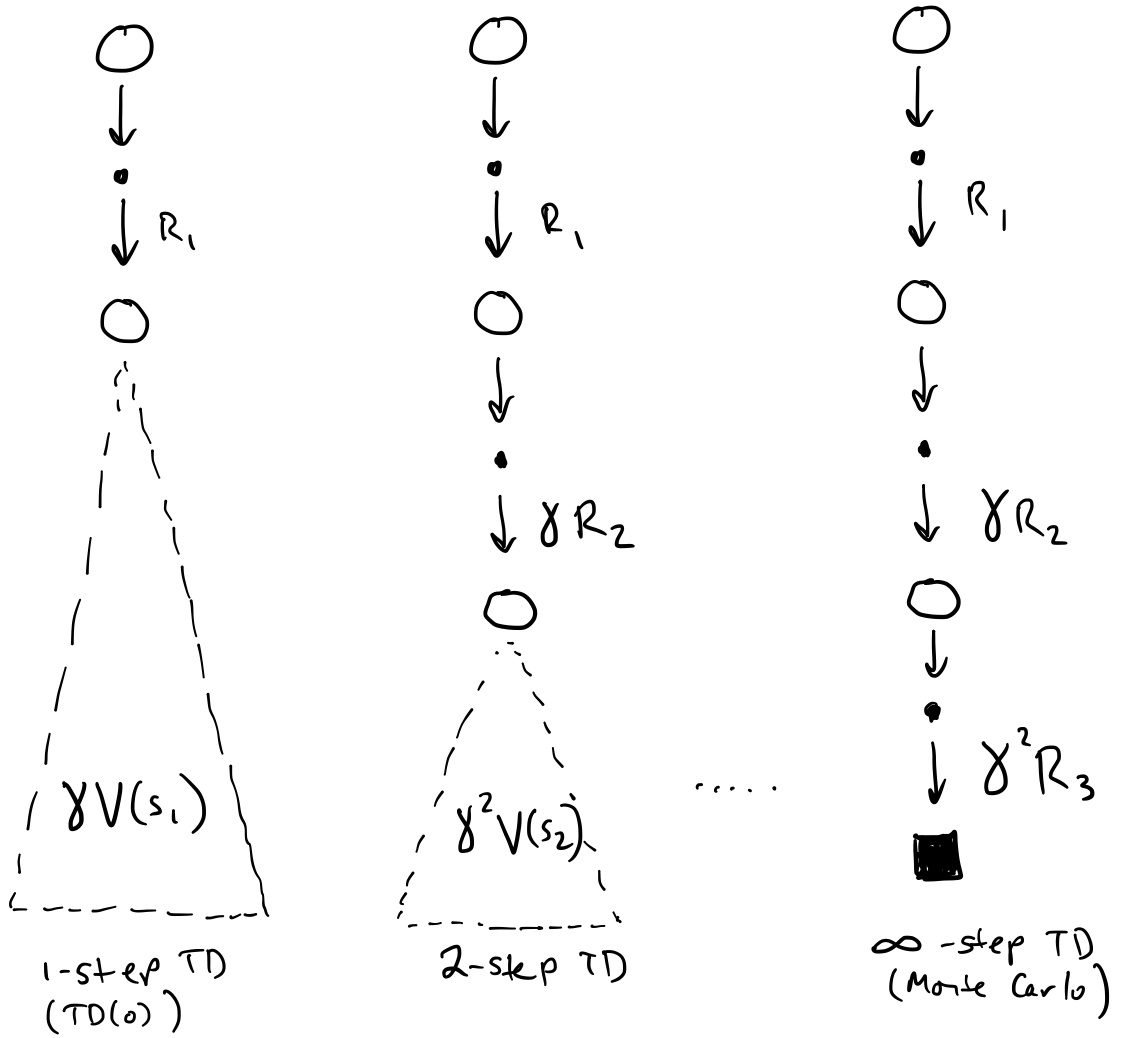


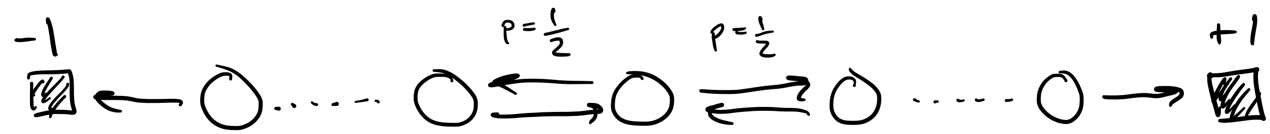

## Targets

TD(0) : $R_t + \gamma V(S_{t+1})$

n-step TD : $R_t + \gamma R_{t+1} + \cdots + \gamma^{n-1} R_{t+n-1} + \gamma^n V(S_{t+n})$

Monte Carlo : $R_t + \gamma R_{t+1} + \cdots + \gamma^{T-t-1} R_{T-t}$

# Chain Markov Reward Process (Length = 19)




## Mini 5-chain $\alpha = 0.5$ $\gamma = 1$

True $V(s)$: $-\frac{2}{3}$ $-\frac{1}{3}$ $0$ $\frac{1}{3}$ $\frac{2}{3}$

1-step TD

| | | A | B | C | D | E | |
|---|---|---|---|---|---|---|---|
| 1-step TD | -1 ← | 0 | 0 | 0 | 0 | 0 | → +1 |
| 3-step TD | -1 ← | 0 | 0 | 0 | 0 | 0 | → +1 |
| ∞-step TD | -1 ← | 0 | 0 | 0 | 0 | 0 | → +1 |

### Experiences

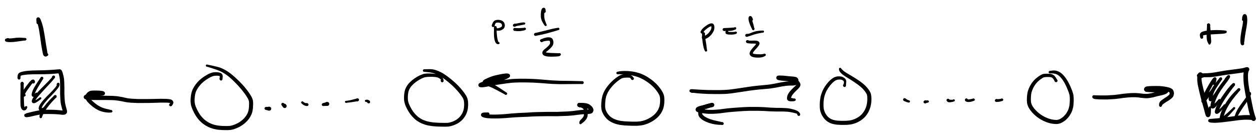
B, C, D, E, +1

C, B, A, -1

A, B, C, D, E, +1

$$V_{t+n}(S_t) \leftarrow V_{t+n-1}(S_t) + \alpha\left[G_{t:t+n} - V_{t+n-1}(S_t)\right]$$

# Chain Markov Reward Process (Length = 19)




## Mini 5-chain $\alpha = 0.5$ $\gamma = 1$

| | | A | B | C | D | E | |
|---|---|---|---|---|---|---|---|
| True $V(s)$: | | $-\frac{2}{3}$ | $-\frac{1}{3}$ | 0 | $\frac{1}{3}$ | $\frac{2}{3}$ | |
| 1-step TD | $-1 \leftarrow$ | 0 | 0 | 0 | 0 | ~~0~~ 0.5 | $\rightarrow +1$ |
| 3-step TD | $-1 \leftarrow$ | 0 | 0 | ~~0~~ 0.5 | ~~0~~ 0.5 | ~~0~~ 0.5 | $\rightarrow +1$ |
| $\infty$-step TD | $-1 \leftarrow$ | 0 | ~~0~~ 0.5 | ~~0~~ 0.5 | ~~0~~ 0.5 | ~~0~~ 0.5 | $\rightarrow +1$ |

### Experiences

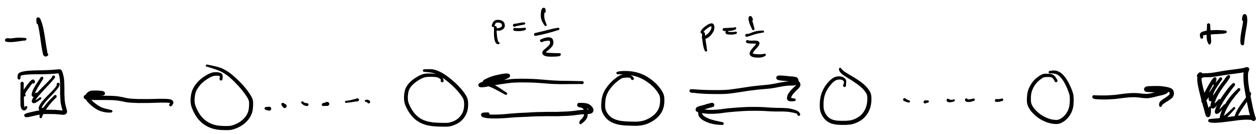
✓ B, C, D, E +1

C, B, A, -1

A, B, C, D, E, +1

$$V_{t+n}(S_t) \leftarrow V_{t+n-1}(S_t) + \alpha\left[G_{t:t+n} - V_{t+n-1}(S_t)\right]$$

# Chain Markov Reward Process (Length = 19)




## Mini 5-chain $\alpha = 0.5$ $\gamma = 1$

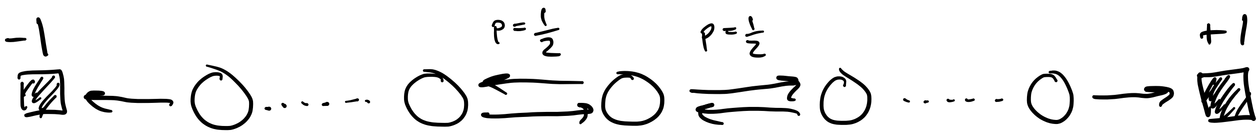
| | | A | B | C | D | E | |
|---|---|---|---|---|---|---|---|
| True v(s): | | $-\frac{2}{3}$ | $-\frac{1}{3}$ | $0$ | $\frac{1}{3}$ | $\frac{2}{3}$ | |
| 1-step TD | -1 ← | ~~0~~ -0.5 | 0 | 0 | 0 | ~~0~~ 0.5 | → +1 |
| 3-step TD | -1 ← | ~~0~~ -0.5 | ~~0~~ -0.5 | ~~0~~ ~~0.5~~ -0.25 | ~~0~~ 0.5 | ~~0~~ 0.5 | → +1 |
| ∞-step TD | -1 ← | ~~0~~ -0.5 | ~~0~~ ~~0.5~~ -0.25 | ~~0~~ ~~0.5~~ -0.25 | ~~0~~ 0.5 | ~~0~~ 0.5 | → +1 |

### Experiences

- ✓ B, C, D, E, +1
- ✓ C, B, A, -1
- A, B, C, D, E, +1

$$V_{t+n}(S_t) \leftarrow V_{t+n-1}(S_t) + \alpha\left[G_{t:t+n} - V_{t+n-1}(S_t)\right]$$

# Chain Markov Reward Process (Length = 19)




## Mini 5-chain $\alpha = 0.5$ $\gamma = 1$

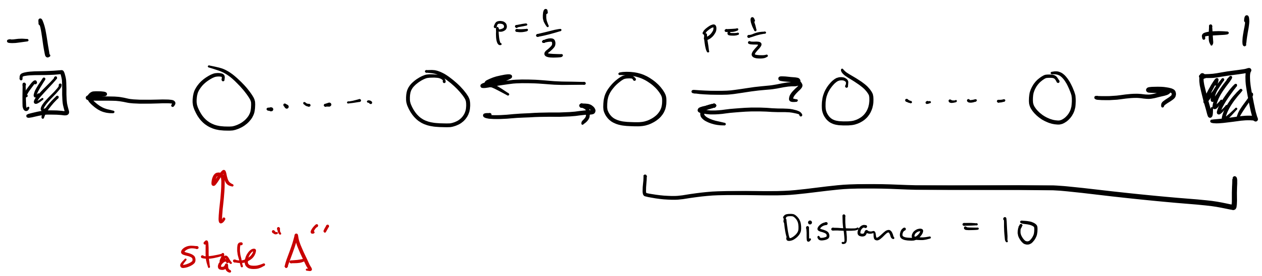
| | | A | B | C | D | E | |
|---|---|---|---|---|---|---|---|
| True V(S): | | $-\frac{2}{3}$ | $-\frac{1}{3}$ | 0 | $\frac{1}{3}$ | $\frac{2}{3}$ | |
| 1-step TD | -1 ← | ~~0~~ ~~-0.5~~ -0.25 | ~~0~~ 0 | ~~0~~ 0 | ~~0~~ 0.25 | ~~0~~ ~~0.5~~ 0.75 | → +1 |
| 3-step TD | -1 ← | ~~0~~ ~~-0.5~~ 0 | ~~0~~ ~~-0.5~~ 0 | ~~0~~ ~~0.5~~ ~~-0.25~~ 0.375 | ~~0~~ ~~0.5~~ 0.75 | ~~0~~ ~~0.5~~ 0.75 | → +1 |
| ∞-step TD | -1 ← | ~~0~~ ~~-0.5~~ 0.25 | ~~0~~ ~~0.5~~ ~~-0.25~~ 0.375 | ~~0~~ ~~0.5~~ ~~-0.25~~ 0.375 | ~~0~~ ~~0.5~~ 0.75 | ~~0~~ ~~0.5~~ 0.75 | → +1 |

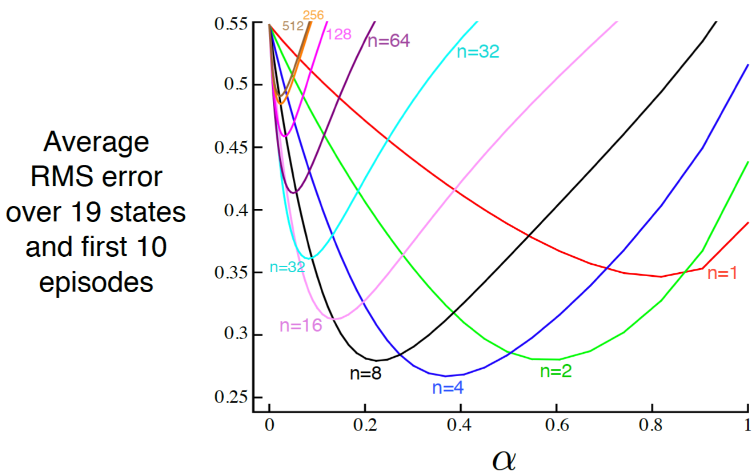
## Experiences

- ✓ B, C, D, E, +1
- ✓ C, B, A, -1
- ✓ A, B, C, D, E, +1

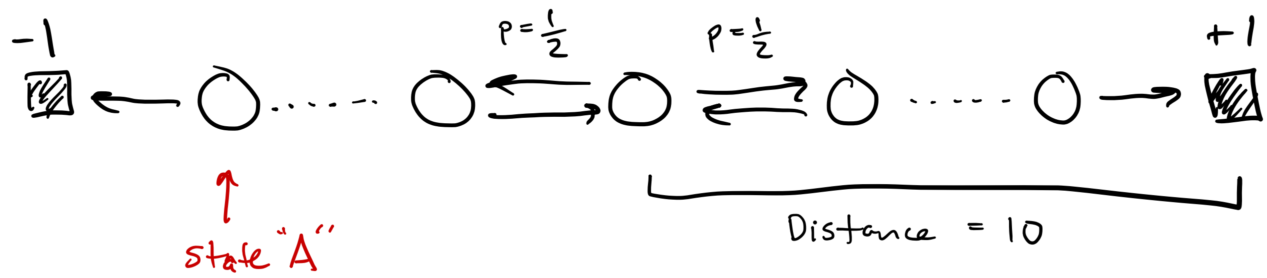
$$V_{t+n}(S_t) \leftarrow V_{t+n-1}(S_t) + \alpha\left[G_{t:t+n} - V_{t+n-1}(S_t)\right]$$

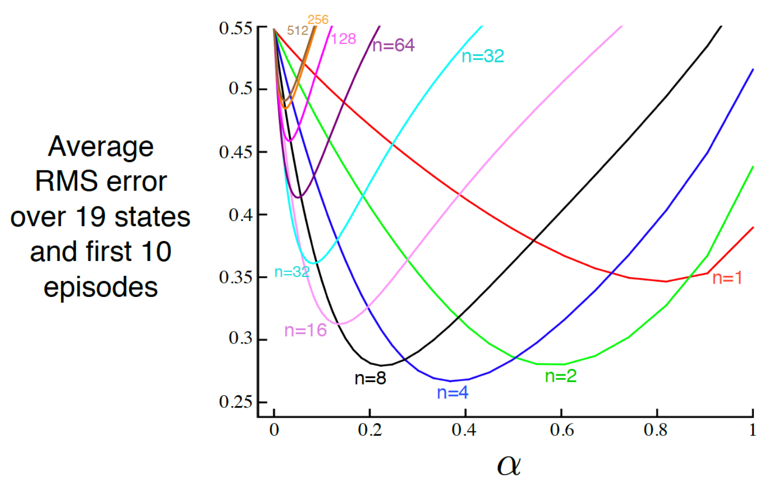
# Chain Markov Reward Process (Length=19)



# Chain Markov Reward Process (Length = 19)



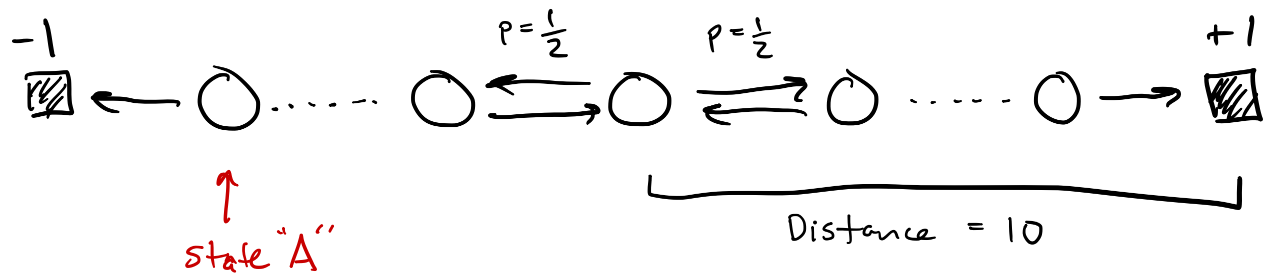




<u>Influence state A value via a +1 reward:</u>

MC: must get to +1 goal at least once after visiting A

TD(0): must visit +1 goal at least once <u>anytime</u> and then wait for value to propagate.

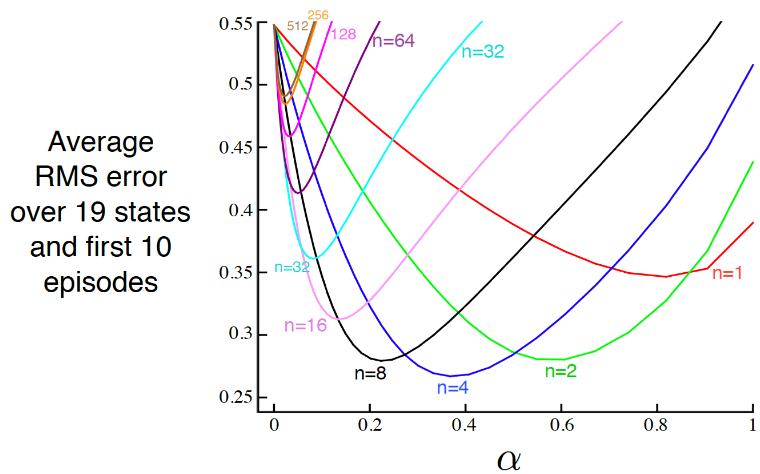
# Chain Markov Reward Process (Length = 19)




## Influence state A value via a +1 reward:

MC: must get to +1 goal at least once after visiting A

TD(0): must visit +1 goal at least once anytime and then wait for value to propagate.




## Statistical Properties:

MC: Unbiased, high variance
Fast propagation of reward (all visited states)

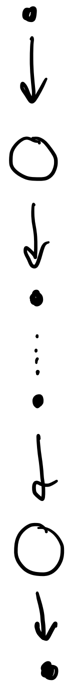
TD(0): Biased updates, low variance
Slow propagation of reward (most recent state)

n-step TD: Medium variance
Medium propagation of reward

Correct balance problem specific!

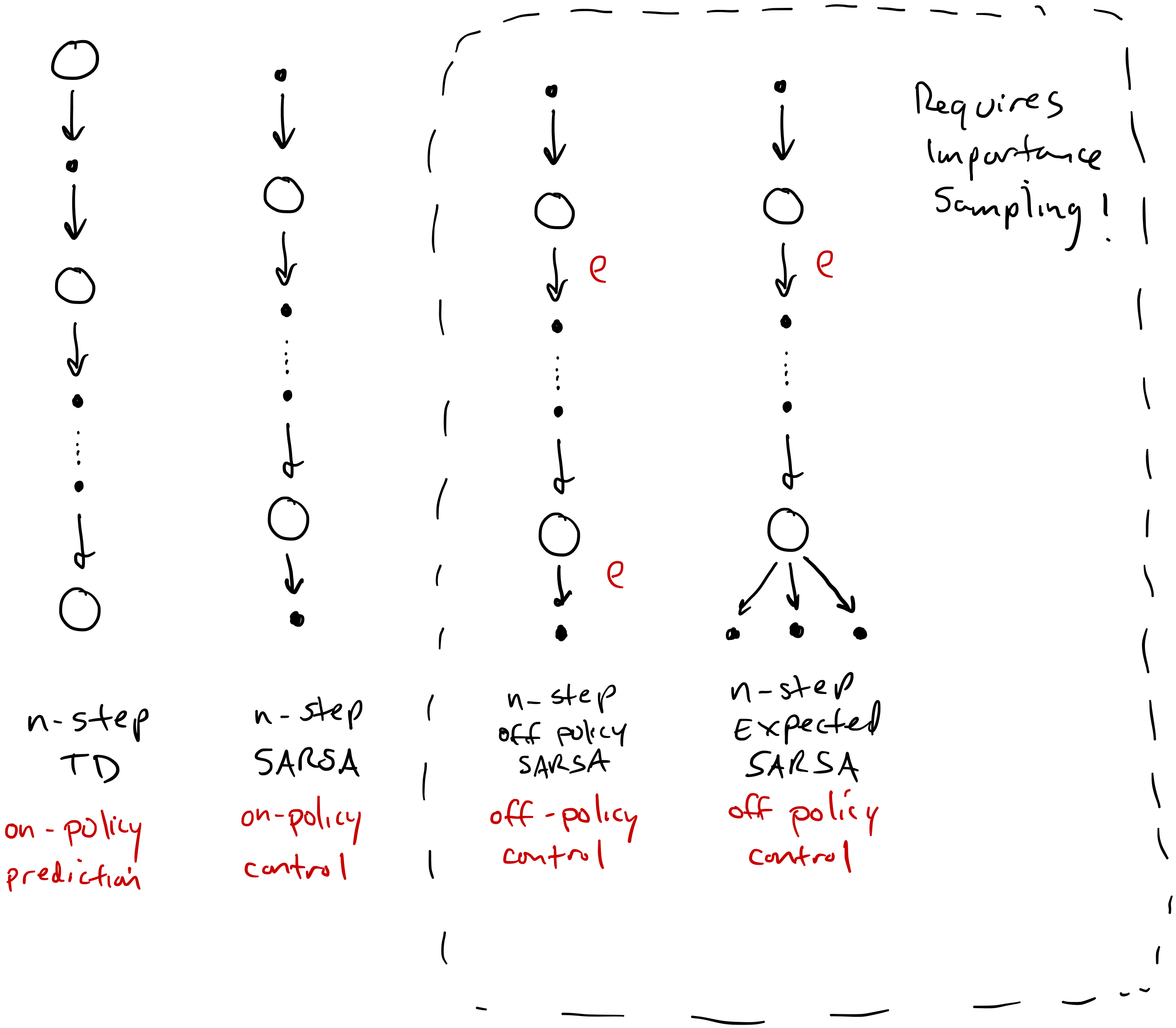
n-step
TD

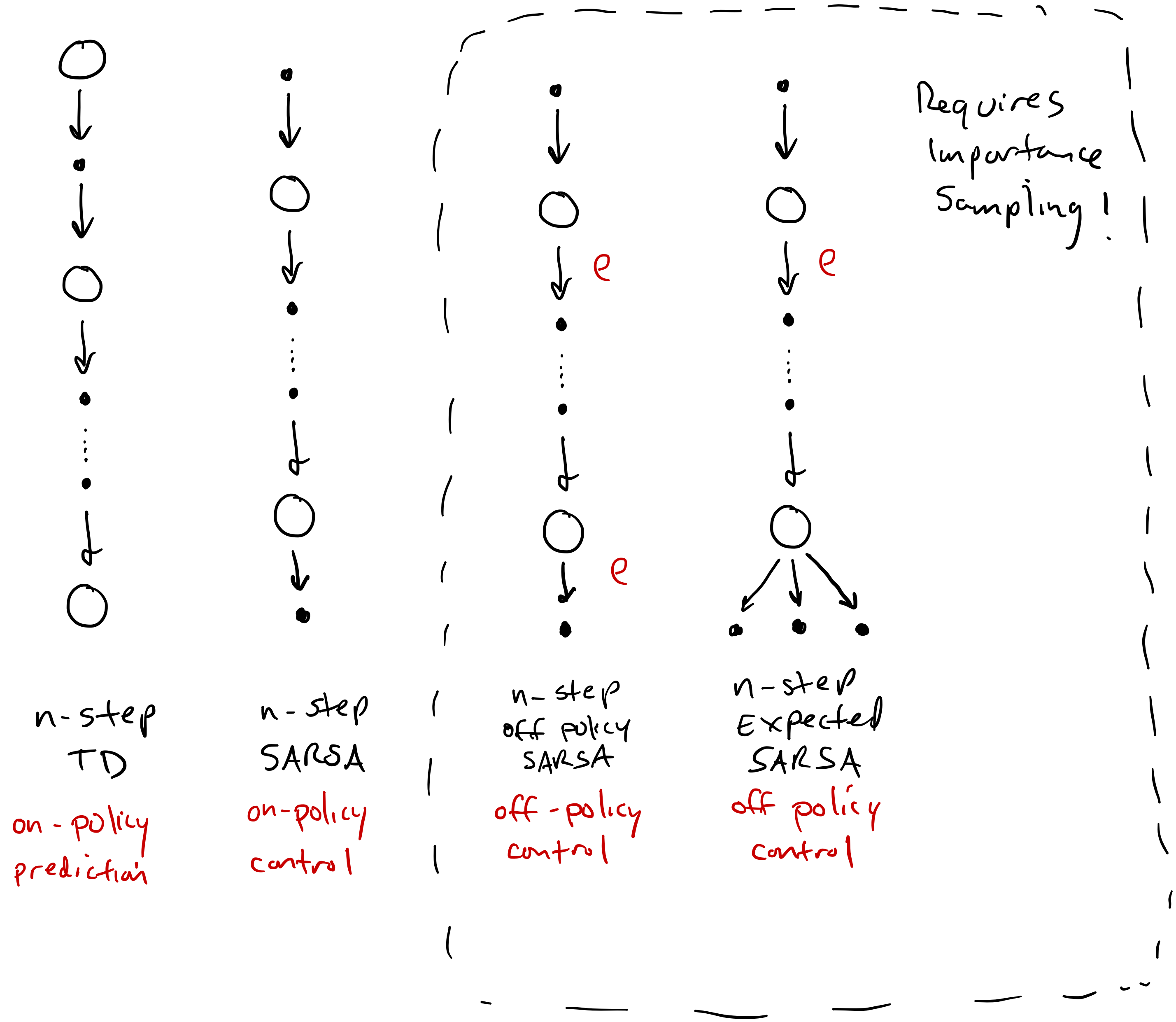
on-policy
prediction



n-step
SARSA

on-policy
control

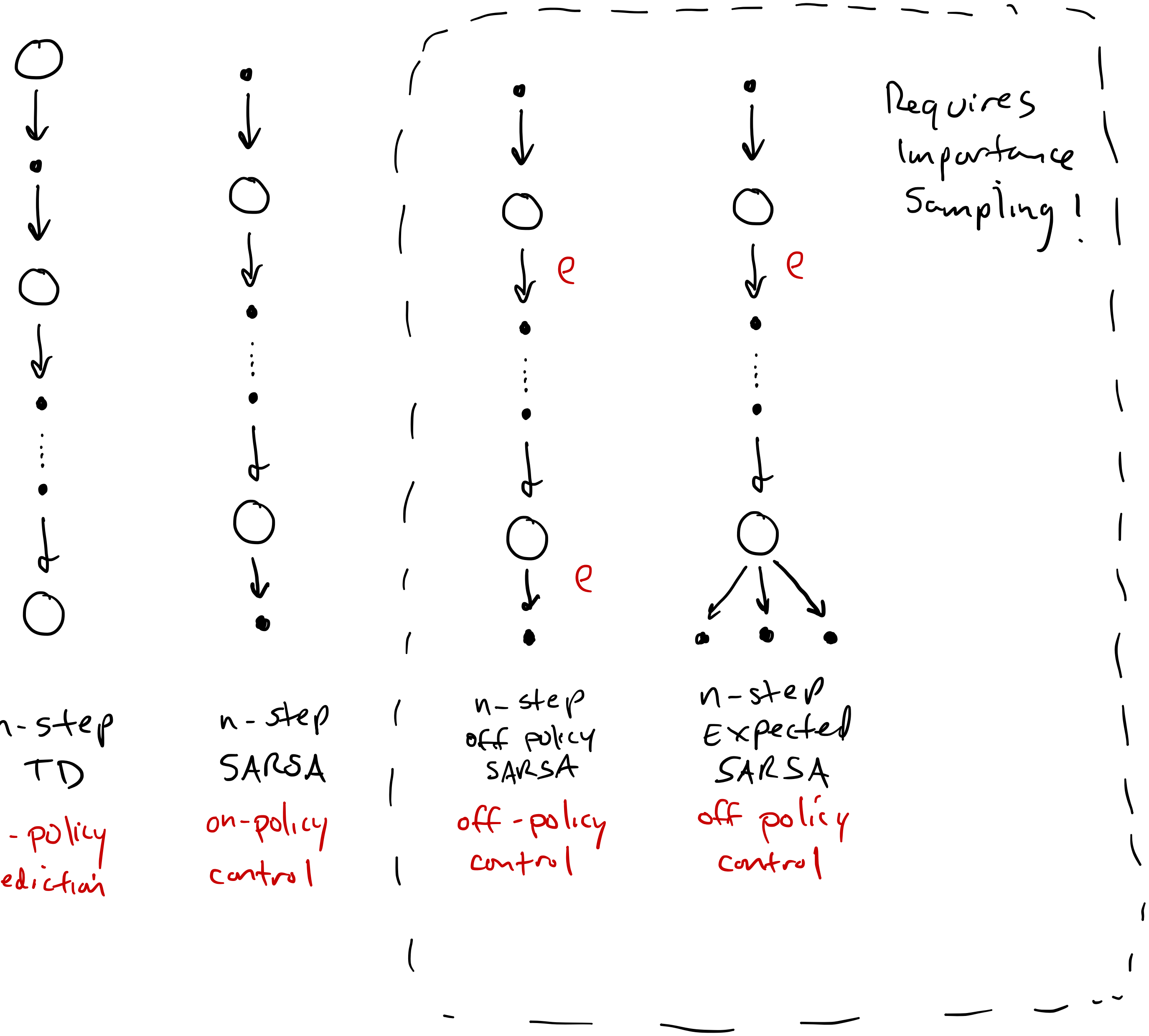

Requires Importance Sampling!
e
e
e
n-step TD
on-policy prediction
n-step SARSA
on-policy control
n-step off policy SARSA
off-policy control
n-step Expected SARSA
off policy control

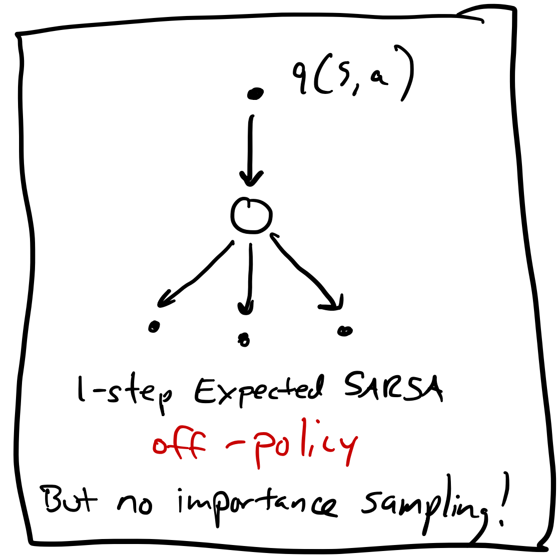



## Targets:

n-step SARSA : $R_t + \gamma R_{t+1} + \cdots + \gamma^{n-1} R_{t+n-1} + \gamma^n Q(S_{t+n}, a_{t+n})$

n-step off policy SARSA : $\rho_{t:t+n-1}\left[R_t + \gamma R_{t+1} + \cdots + \gamma^{n-1} R_{t+n-1} + \gamma^n Q(S_{t+n}, a_{t+n})\right]$




- No actions considered other than:
  1. action prescribed by $q(s,a)$
  2. actions in expectation w.r.t. $\pi_{eval}$

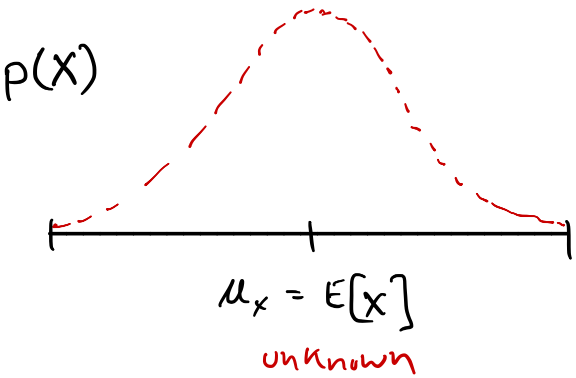
- N-step Expected SARSA takes N-1 actions from $\pi_b$ that require reweighting

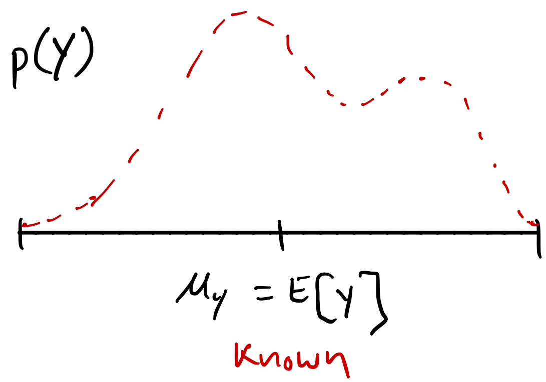
Targets:

n-step SARSA : $R_t + \gamma R_{t+1} + \cdots + \gamma^{n-1} R_{t+n-1} + \gamma^n Q(S_{t+n}, a_{t+n})$

n-step off policy SARSA : $\rho_{t:t+n-1}\left[R_t + \gamma R_{t+1} + \cdots + \gamma^{n-1} R_{t+n-1} + \gamma^n Q(S_{t+n}, a_{t+n})\right]$

# Control Variates



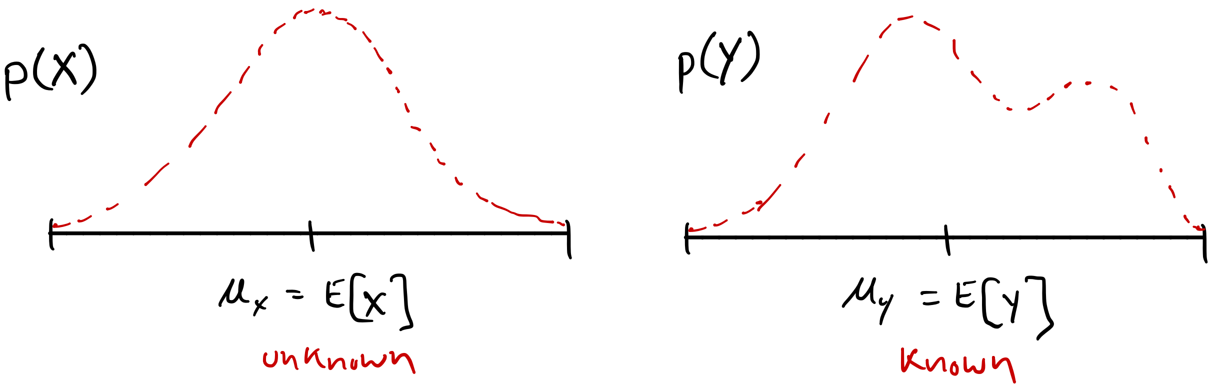




Assumption: $X$ is positively correlated with $Y$

i.e. when $X > E[X]$, it is likely that $Y > E[Y]$

ex: rain and traffic

# Control Variates




Assumption: $X$ is positively correlated with $Y$

i.e. when $X > E[X]$, it is likely that $Y > E[Y]$

ex: rain and traffic

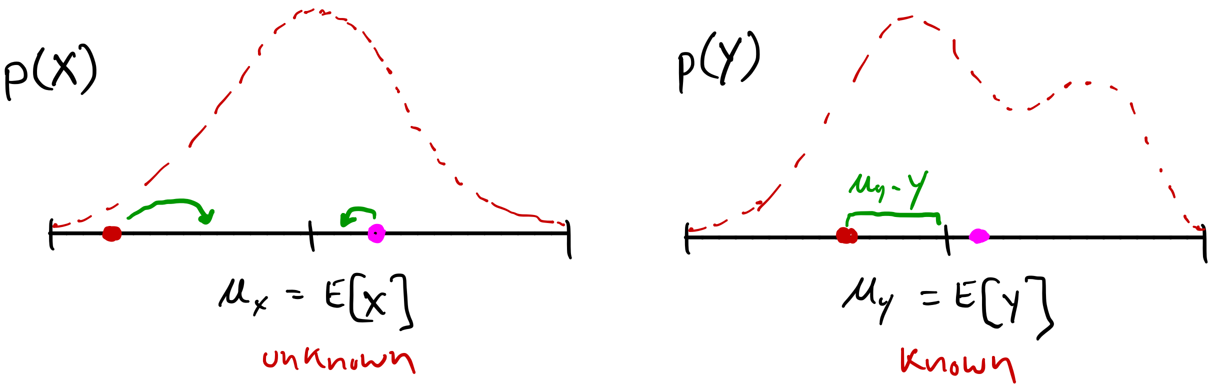
Sample $x \sim p(x)$ and $y \sim p(y)$

Consider two different estimators:

$$\hat{\mu}_X = X$$

$$\hat{\mu}_X = X + [\mu_Y - Y]$$

# Control Variates




Assumption: $X$ is positively correlated with $Y$

i.e. when $X > E[X]$, it is likely that $y > E[y]$

ex: rain and traffic

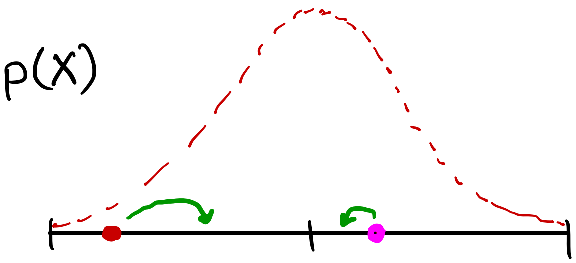
Sample $x \sim p(x)$ and $y \sim p(y)$

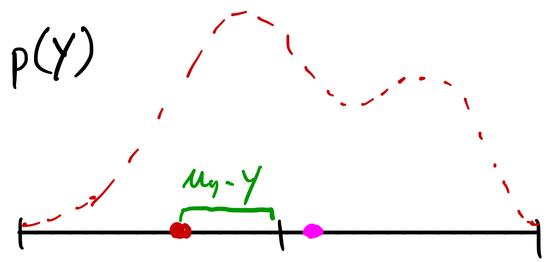
Consider two different estimators:

$\hat{\mu}_x = X$

$$\hat{\mu}_x = X + \underbrace{\left[\mu_y - y\right]}_{\text{Expectation zero!}}$$

## Control Variates




Assumption: X is positively correlated with Y

i.e. when $X > E[X]$, it is likely that $Y > E[Y]$

ex: rain and traffic

Sample $x \sim p(x)$ and $y \sim p(y)$

Consider two different estimators:

$$\hat{\mu}_X = X$$

$$\hat{\mu}_X = X + \underbrace{[\mu_Y - Y]}_{\text{Expectation zero!}}$$

## n-Step PDIS with CV

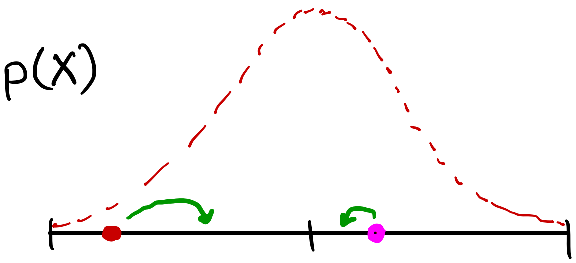
$$G_{t:h} = \rho_t (R_{t+1} + \gamma G_{t+1:h}) + \underbrace{(1-\rho_t) V_{h-1}(s_t)}_{\text{CV}}$$

$E[\rho_t] = 1$, thus;

$E[1 - \rho_t] = 0$

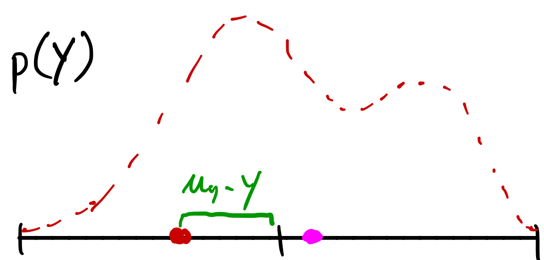
$$\rightarrow E[(1-\rho_t) V_{h-1}(s_t)] = 0$$

## Control Variates




$\mu_x = E[x]$

unknown




$\mu_y = E[y]$

known

Assumption: $X$ is positively correlated with $Y$

i.e. when $x > E[X]$, it is likely that $y > E[y]$

ex: rain and traffic

Sample $x \sim p(x)$ and $y \sim p(y)$

Consider two different estimators:

$$\hat{\mu}_x = x$$

$$\hat{\mu}_x = x + \underbrace{[\mu_y - y]}_{\text{Expectation zero!}}$$

## n-Step PDIS with CV

$$G_{t:h} = \rho_t (R_{t+1} + \gamma G_{t+1:h}) + \underbrace{(1-\rho_t) V_{h-1}(s_t)}_{\text{CV}}$$

$E[\rho_t] = 1$, thus:

$E[1 - \rho_t] = 0$

$\rightarrow E[(1-\rho_t) V_{h-1}(s_t)] = 0$

Compare to without CV:

$$G_{t:h} = \rho_t \left[ R_{t+1} + \gamma G_{t+1:h} \right]$$

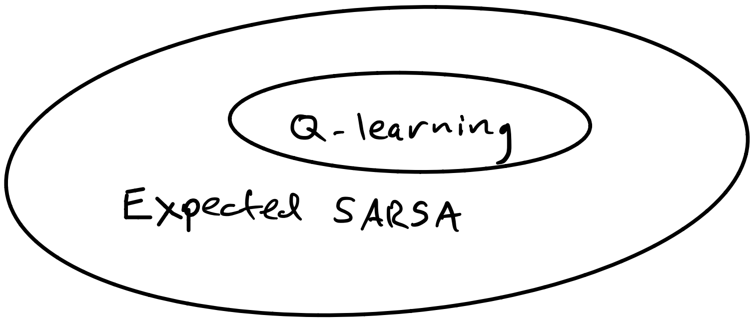
if $\rho_t = 0$:

CV: $G_{t:h} = V_{h-1}(s_t)$

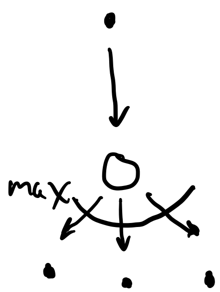
No CV: $G_{t:h} = 0$

"how much better or worse than average is this sample?"

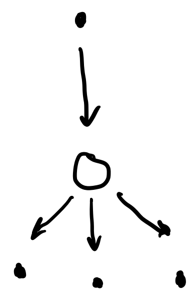
# one step off policy




Same when $\pi_{Eval} = \pi_\star$



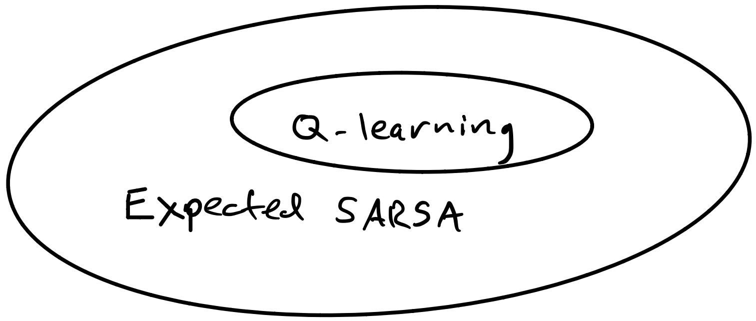

Q-learning



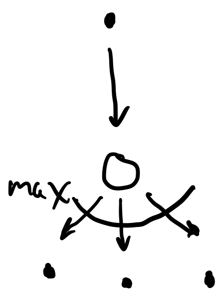
Expected SARSA

- Both off policy
- No importance sampling !

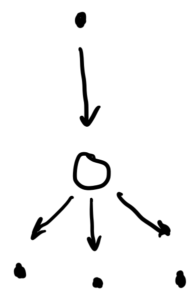
## one step off policy




Same when $\pi_{Eval} = \pi_*$



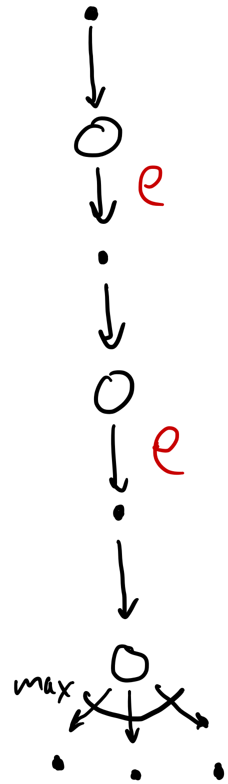

Q-learning



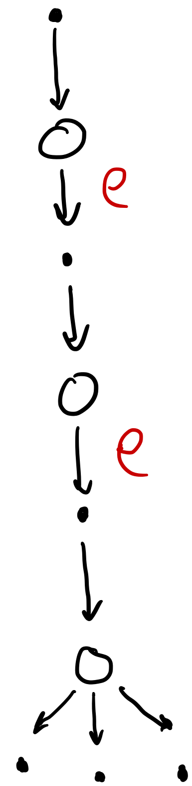
Expected SARSA

- Both off policy
- No importance sampling!

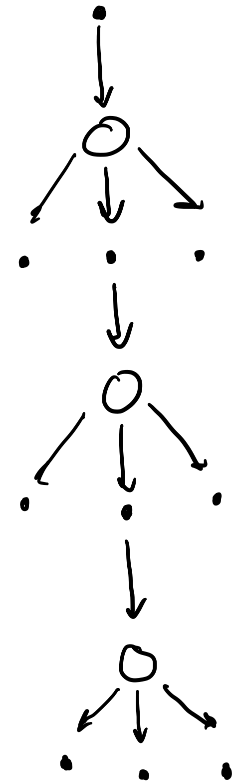
## n-step off policy




n-step Q-learning




n-step Expected SARSA



Tree Backups

Tree Backups Target:

$$G_{t:t+n} \doteq R_{t+1} + \gamma \sum_{a \neq A_{t+1}} \pi(a|S_{t+1}) Q_{t+n-1}(S_{t+1}, a) \ + \ \gamma \pi(A_{t+1}|S_{t+1}) G_{t+1:t+n}$$