n-step temporal difference learning

Tuesday, 18 September 2018 9:25 AM

Discounted future rewards

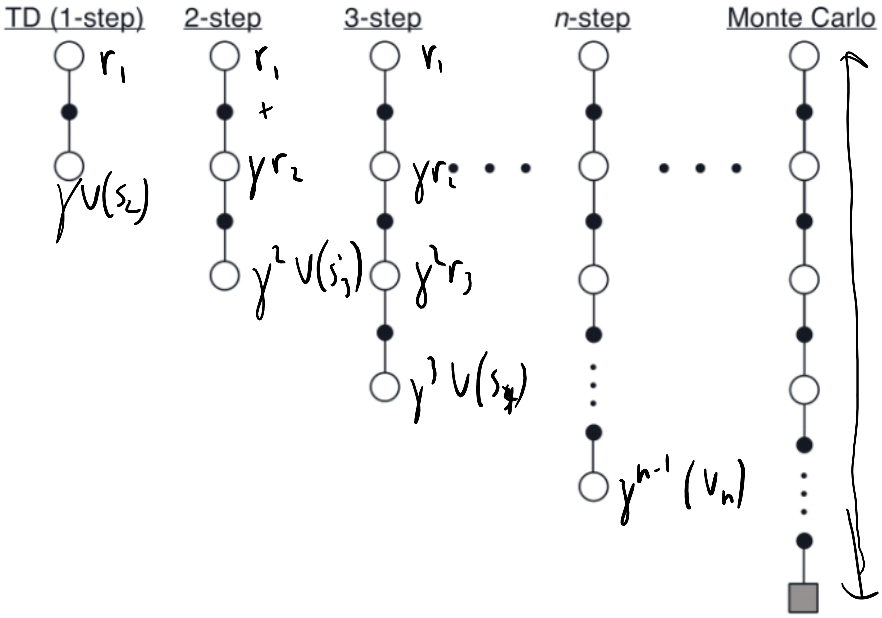
$$G = r_1 + \gamma r_2 + \gamma^2 r_3 + \gamma^3 r_4 \dots \gamma^{T-1} r_T$$

$$= r_1 + \gamma(r_2 + \gamma r_3 + \gamma^2 r_4 \dots$$

$$= r_1 + \gamma(r_2 + \gamma(r_3 + \gamma r_4)$$

$$= \underline{r_1} + \gamma \boxed{G_{t+1}} \longleftrightarrow$$ approximated using $V(s')$

Truncated future rewards



SARSA: $Q(s,a) := Q(s,a) + \alpha [r + \gamma \underline{Q(s',a')} - Q(s,a)]$

Change the update:

$$G \leftarrow \sum_{i=\boxed{\tau}+1}^{\min(\tau+n,\,T)} \gamma^{i-\tau-1} r_i$$

if $\underline{\tau + n < T}$ $\quad G \leftarrow G + \gamma^n \underline{Q(s', a')}$

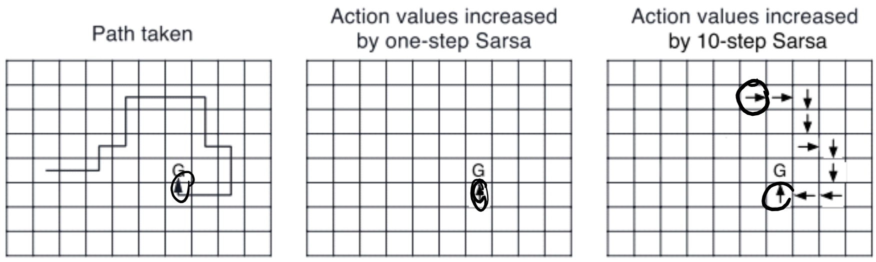
$$Q(s_\tau, a_\tau) \leftarrow Q(s_\tau, a_\tau) + \alpha [G - Q(s_\tau, a_\tau)]$$

n-step Sarsa for estimating $Q \approx q_*$, or $Q \approx q_\pi$ for a given π

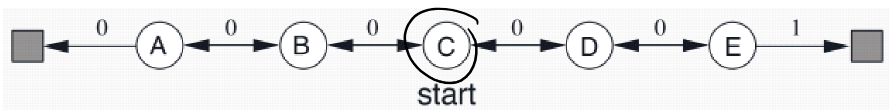
Initialize $Q(s,a)$ arbitrarily, for all $s \in \mathcal{S}, a \in \mathcal{A}$
Initialize π to be ε-greedy with respect to Q, or to a fixed given policy
Parameters: step size $\alpha \in (0,1]$, small $\varepsilon > 0$, a positive integer n
All store and access operations (for S_t, A_t, and R_t) can take their index mod n

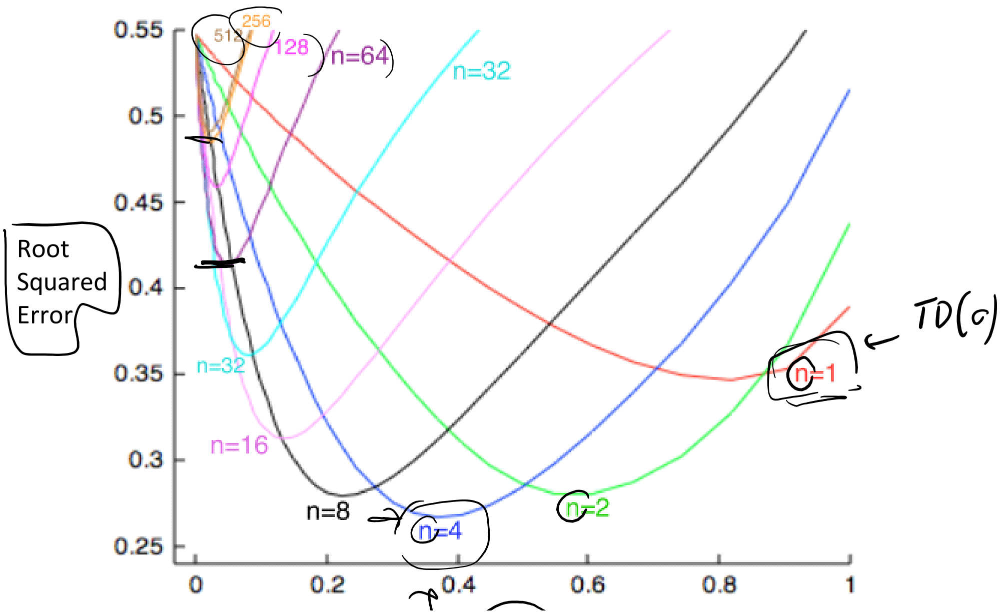
Repeat (for each episode):
  Initialize and store $S_0 \neq$ terminal
  Select and store an action $A_0 \sim \pi(\cdot|S_0)$
  $T \leftarrow \infty$
  For $t = 0, 1, 2, \ldots$:
    If $t < T$, then:
      Take action A_t
      Observe and store the next reward as R_{t+1} and the next state as S_{t+1}
      If S_{t+1} is terminal, then:
        $T \leftarrow t + 1$
      else:
        Select and store an action $A_{t+1} \sim \pi(\cdot|S_{t+1})$
    $\tau \leftarrow t - n + 1$ (τ is the time whose estimate is being updated)
    If $\tau \geq 0$:
      $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n,T)} \gamma^{i-\tau-1} R_i$
      If $\tau + n < T$, then $G \leftarrow G + \gamma^n Q(S_{\tau+n}, A_{\tau+n})$ $(G_{\tau:\tau+n})$
      $Q(S_\tau, A_\tau) \leftarrow Q(S_\tau, A_\tau) + \alpha\,[G - Q(S_\tau, A_\tau)]$
      If π is being learned, then ensure that $\pi(\cdot|S_\tau)$ is ε-greedy wrt Q
  Until $\tau = T - 1$

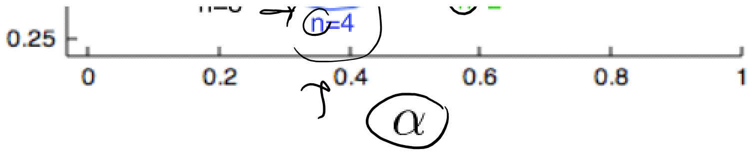
TD updates:



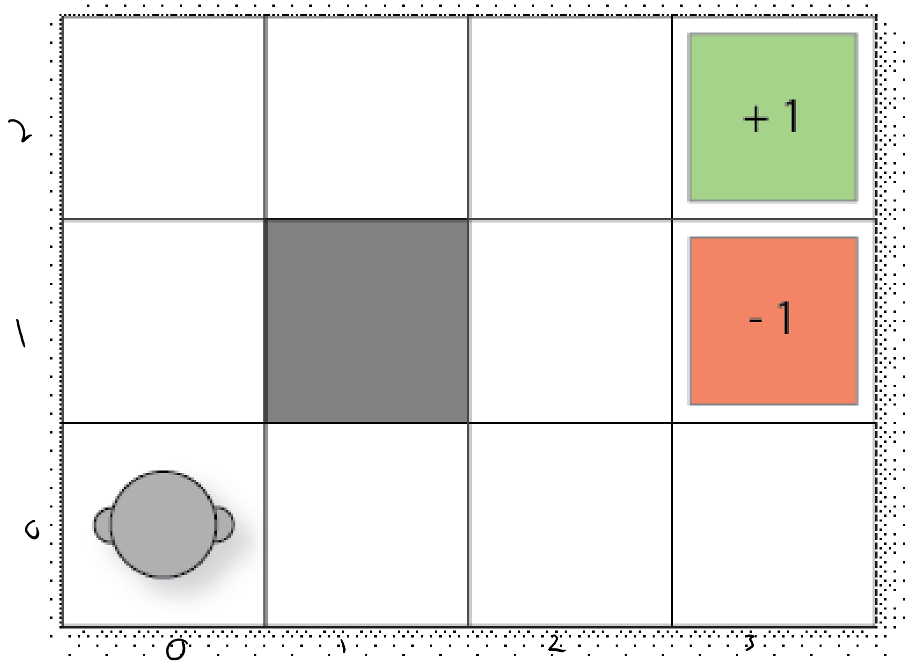
Example: Random walk







Exercise: Grid World



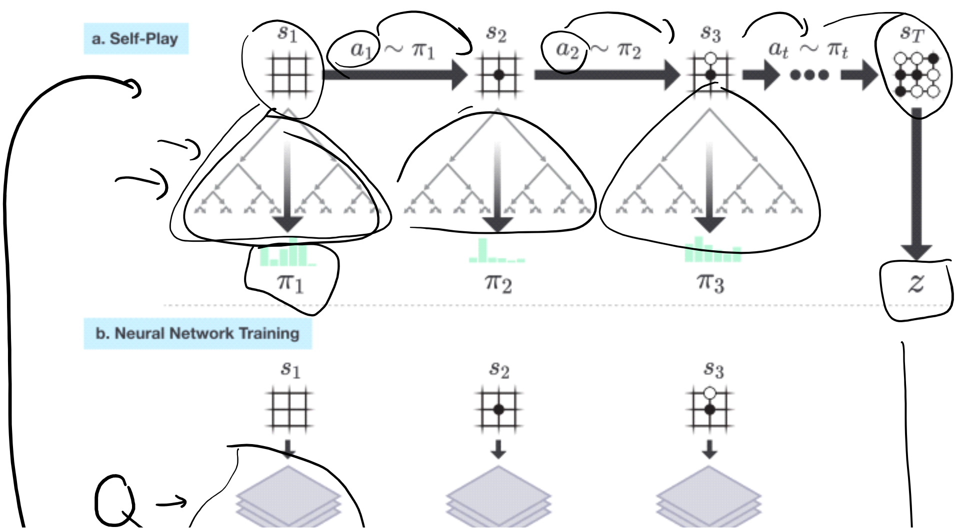
Compute 2-step SARSA update,
$\alpha = 0.5$
$\gamma = 0.99$

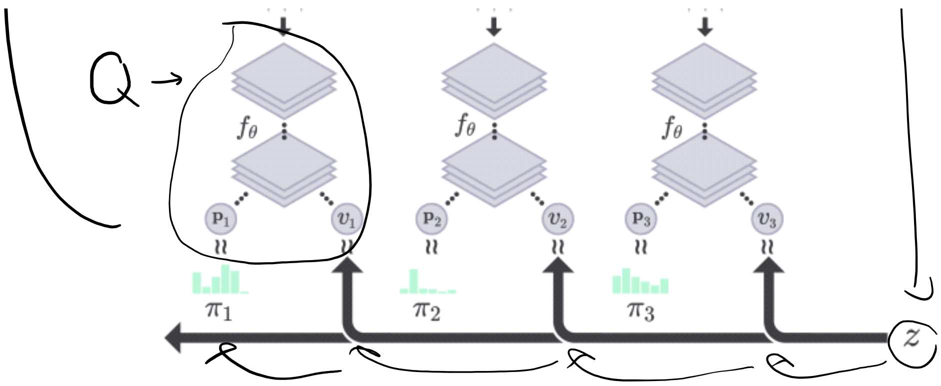
Gt =

$Q(s_{(1,2)}, E)$ =

MCTS + Reinforcement learning

AlphaGoZero:




Q
f_θ
f_θ
f_θ
$\mathbf{p}_1$
v_1
$\mathbf{p}_2$
v_2
$\mathbf{p}_3$
v_3
π_1
π_2
π_3
z