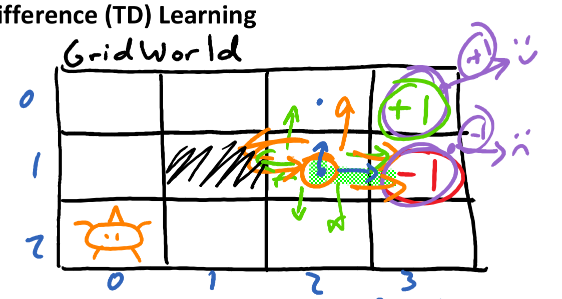
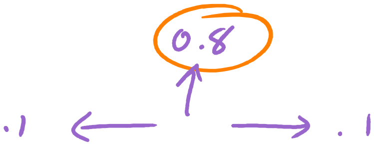
Temporal Difference (TD) Learning



at each step go $\uparrow, \downarrow, \leftarrow, \rightarrow$



$p((0,2) \mid (1,2), up) =$

$p((1,3) \mid (1,2), up) = \frac{1}{2}.$

$p((1,2) \mid (1,2), up) = \frac{1}{2}.$

with some prob, go in intended dir

w/ some prob, go in perpendicular dir

model
$p(s' \mid s, a) =$ prob that from state s, choosing action a puts you in s'

$r(s, a, s') =$ reward for going $s \xrightarrow{a} s'$

$r(s,a,(0,3)) = 1$
$r(s,a,(1,3)) = -1$
$r(s,a,s') = 0$ for other s'

policy : function π : states $\rightarrow$ action

$$v_\pi(s) = \begin{cases} 0 & \text{for terminal states} \\ \sum_{s'} p(s' \mid s, \pi(s)) \cdot (r(s,a,s') + \gamma \cdot v_\pi(s')) \end{cases}$$

π^* : policy that maximizes $v_\pi(s)$ for each s

$v^*(s) = v_{\pi^*}(s)$

Value iteration –

initialize $v(s)$ to 0 for each s

until v has converged

for each state s

$$v(s) = \sum_{s'} P(s' \mid s, \pi(s)) \cdot (r(s,a,s') + \gamma \cdot v_\pi(s'))$$

$v^*(s) = \max_a q^*(s,a)$

$q(s,a)$ the value of taking action a from s

$$\pi(s) = \arg\max_a \sum_{s'} P(s' \mid s, a) \cdot (r(s,a,s') + \gamma \cdot v_\pi(s'))$$

TD Value Learning (TD(0)) – learn $v_\pi(s)$ w/o knowing model (model-free)

TD → temporal difference

converges if for all states s,

$\sum_{t=1}^{\infty} \alpha_{s,t} = \infty$ $\quad \sum_{t=1}^{\infty} \alpha_{s,t}^2 < \infty$

α applied to t^{th} update to state s

$\alpha_{s,t} = \frac{1}{t}$

until done

episode
$s \leftarrow s_0$ initial state

while s not terminal

$a \leftarrow \pi(s)$ q-learning: choose an action (ε-greedy) episode

$a \leftarrow \pi(s)$ q-learning: choose an action (ϵ-greedy)

$\alpha_{s,t} = \frac{1}{t}$

observe reward r, new state s'

$$v(s) = (1-\alpha)\cdot v(s) + \alpha\cdot(r + \gamma \cdot v(s'))$$

$$v(s) + \underbrace{\alpha}_{\text{learning rate}} \underbrace{(r + \gamma\cdot v(s') - v(s))}_{\text{surprise}}$$

$v(s') = \max_{a'} q(s', a')$

observe $s \xrightarrow{a} s'$

$$q(s,a) += q(s,a) + \alpha\Big(\underbrace{r + \gamma \cdot \max_{a'} q(s',a')}_{\text{sample}} - q(s,a)\Big)$$

Q - learning

sampled value of $V(s)$ =

update estimate $V(s) \leftarrow V(s)$

$$= (1-\alpha)\ V(s) + \alpha\ R(s,s',a) + \gamma V(s')$$