

states encapsulate (yards to go, downs left, yards to go to 1st down, kicks left, hashmark, score margin, time outs, ^{own} time outs, ^{opp} time outs, possession)

(for TD) yards to go, downs left, yards to go to 1st down, kicks left, hashmark, score margin, time outs, ^{own} time outs, ^{opp} time outs, possession

yards to go: gems to collect (47, 51), or wait until Busby (4 winks), check around gems needed (10), or call rings left (20, 19)

left/right/center

want some generalization

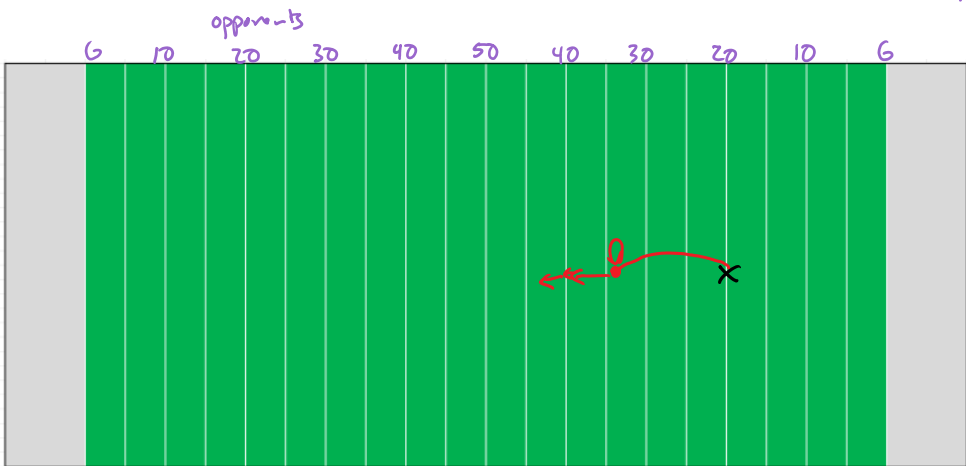
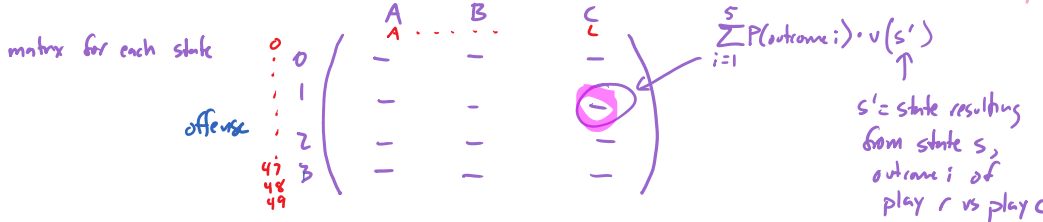
possible values: 99, 4, 99, 361, 3, 201, 4, 4, 2

total # of states: ~273 billion states

~63 years @ 135 mb/s/sec

you behind 100
i
you ahead 100

$$v(s) = P(\text{offense wins in state } s)$$



Home	quarter	time	down	distance	Away
<u>13</u>	4	2:00	4	41	<u>17</u>
		1:50			
		1:45			
		1:38			
		1:25			

Function Approximators

input: state/action output: $\hat{q}(s,a)$
(approximation of exact $q(s,a)$)

Linear Approximator

Define features of state/action pairs

$$f_1(s,a) = \begin{cases} 1 & \text{if action } a \text{ moves ball to closest sideline} \\ 0 & \text{otherwise} \end{cases}$$

$$f_2(s,a) = \begin{cases} 1 & \text{if short yardage } (\leq 2 \text{ yards to get 1st down}) \\ 0 & \text{otherwise} \end{cases}$$

$$f_3(s,a) \dots$$

\vdots

$$f_n(s,a)$$

fn of only s (not a)

make a copy $f_{i,a}$ for each action a

$$f_{i,a}(s,a) = \begin{cases} f_i(s,a) & \text{if } a'=a \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{q}(s,a) = w_1 \cdot f_1(s,a) + \dots + w_n f_n(s,a)$$

for nonterminal s

learn weights w / q-learning

for a bunch of episodes

$S \leftarrow S_0$

while s not terminal

In state s

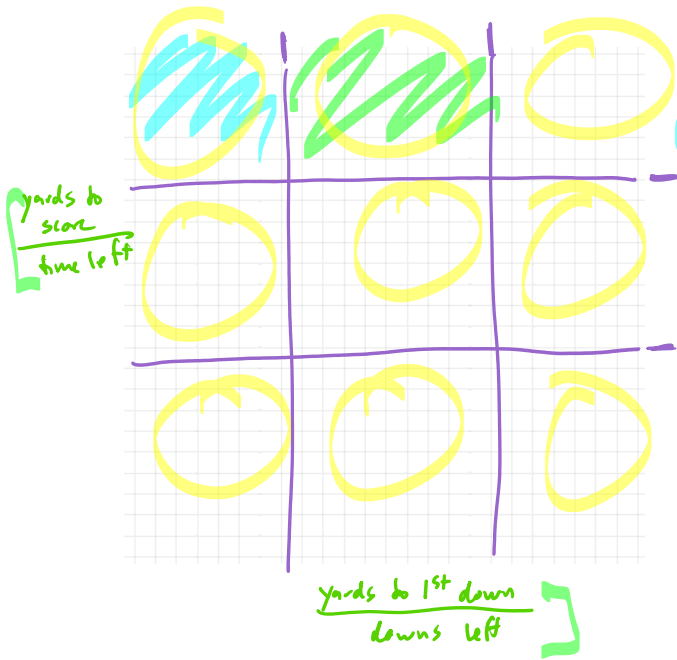
Choose action a (ϵ -greedy)

Observe transition (s, a, r, s')

Update for each feature i $w_i \leftarrow w_i + \alpha \cdot f_i(s,a) \cdot (r + \gamma \max_{a'} \hat{q}(s',a') - \hat{q}(s,a))$

$$\hat{q}(s',a_1) = \hat{q}(s',a_2) \dots = \hat{q}(s',a_n)$$

State Aggregation (Coarse Coding/Buckets/Grid)



$$f_{11}(s,a) = \begin{cases} 1 & \text{if state } s \text{ belongs in } l,1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_{12}(s,a) = \begin{cases} 1 & \text{if state } s \text{ belongs in } l,2 \\ 0 & \text{otherwise} \end{cases}$$

$$f_{11,1}(s,a) = \begin{cases} 1 & \text{if state } s \text{ belongs in } l,1 \text{ and } a=1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_{11,2}(s,a) = \begin{cases} 1 & \text{if state } s \text{ belongs in } l,1 \text{ and } a=2 \\ 0 & \text{otherwise} \end{cases}$$

while not done

$s \leftarrow s_0$

while s not terminal

choose action a

observe transition (s,a,r,s')

update $q(s,a) \leftarrow q(s,a) + \alpha (r + \gamma \cdot \max_{a'} q(s',a') - q(s,a))$

$s \leftarrow s'$

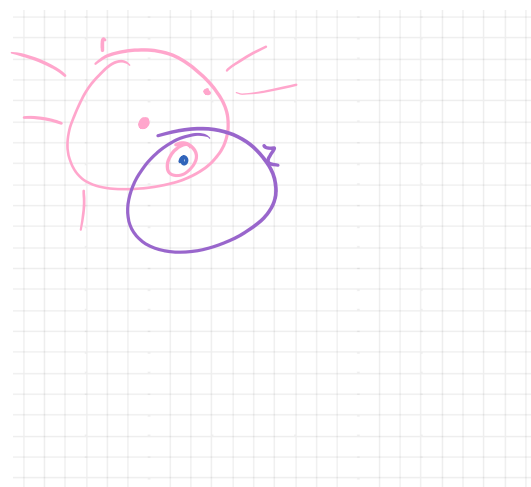
$$\hat{q}(s,a) = w_1 \cdot f_1(s,a) + \dots + w_n f_n(s,a)$$

$n=36$ (9 grid cells in partition 4 actions)

exactly 1 is 1, rest are 0
 so only updating $f_{r,c,a'}$ for s goes in cell r,c and $a=a'$

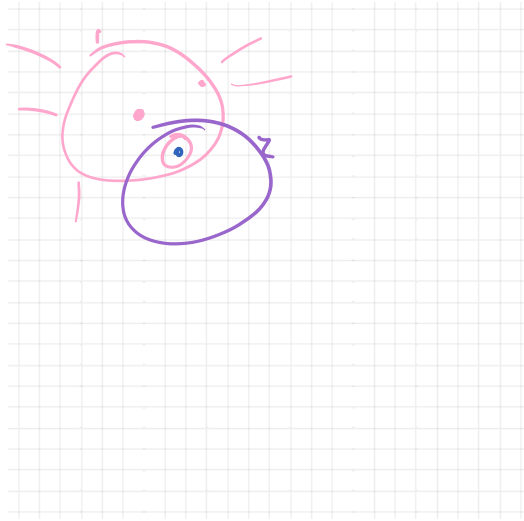
$$w_i \leftarrow w_i + \alpha f_i(s,a) (r + \gamma \max_{a'} \hat{q}(s',a') - \hat{q}(s,a))$$

$w_{r,c,a}$ is the only one updated \rightarrow call that $\hat{q}(\hat{s},a)$



$$f_1 = \begin{cases} 1 & \text{if } s \text{ in region 1} \\ 0 & \text{otherwise} \end{cases}$$

$$f_2 = \begin{cases} 1 & \text{if } s \text{ in region 2} \\ 0 & \text{otherwise} \end{cases}$$



$$f_1 = \begin{cases} 1 & \text{if } s \text{ in region 1} \\ 0 & \text{otherwise} \end{cases}$$

$$f_2 = \begin{cases} 1 & \text{if } s \text{ in region 2} \\ 0 & \text{otherwise} \end{cases}$$