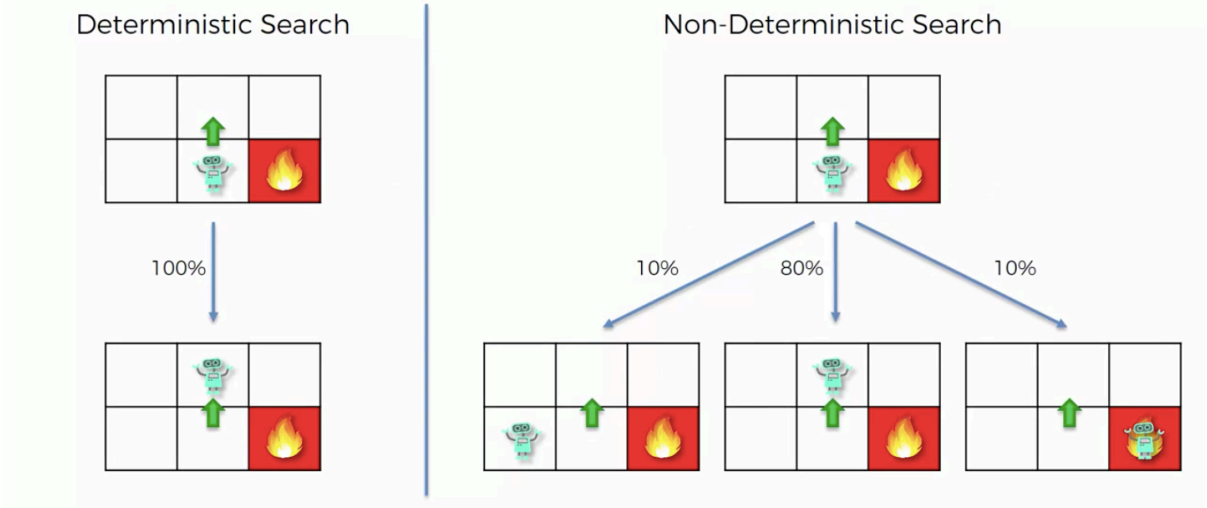
Temporal diffrence

Temporal Difference

Artificial Intelligence A-Z © SuperDataScience

Temporal Difference



Artificial Intelligence A-Z © SuperDataScience

Temporal Difference

V=0.81	V=0.9	V=1	
V=0.73		V=0.9	
V=0.66	V=0.73	V=0.81	V=0.73

Deterministic search

Temporal Difference

V=0.71	V=0.74	V=0.86	
V=0.63		V=0.39	
V=0.55	V=0.46	V=0.36	V=0.22

Non-deterministic search

Temporal Difference

$$Q(s,a) = R(s,a) + \gamma \sum_{s'} \left(P(s,a,s') \max_{a'} Q(s',a') \right)$$

Temporal Difference

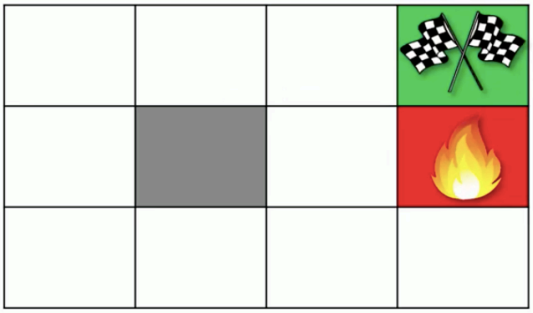
$$Q(s,a) = R(s,a) + \gamma \sum_{s'} \left(P(s,a,s') \max_{a'} Q(s',a') \right)$$

$$Q(s,a) = R(s,a) + \gamma \max_{a'} Q(s',a')$$

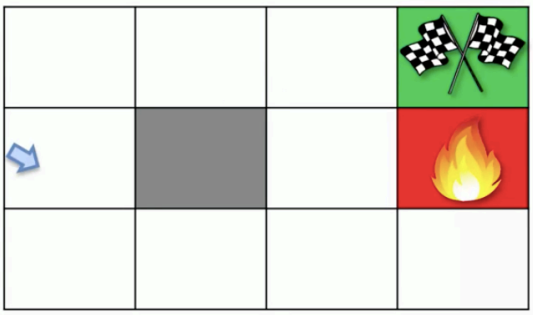
Artificial Intelligence A-Z © SuperDataScience

The second one is deterministic formula and for simplicity we use this here but we actually mean the first one.

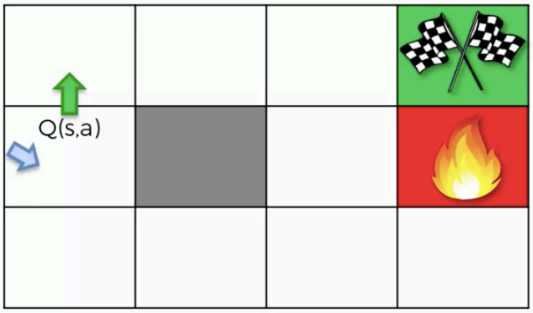
Temporal Difference



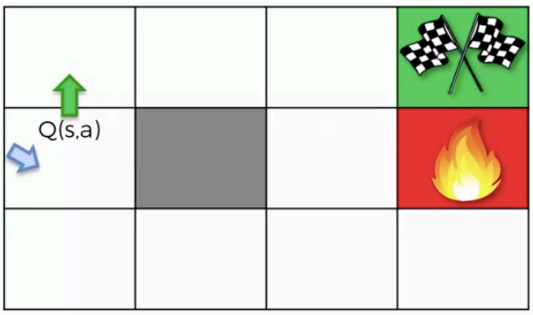
Temporal Difference



Temporal Difference



Temporal Difference

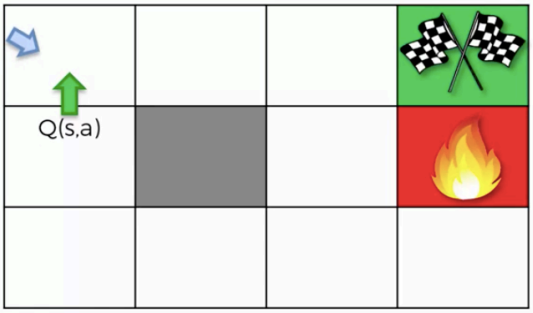


Before:

$Q(s, a)$

before (the agent takes action)

Temporal Difference

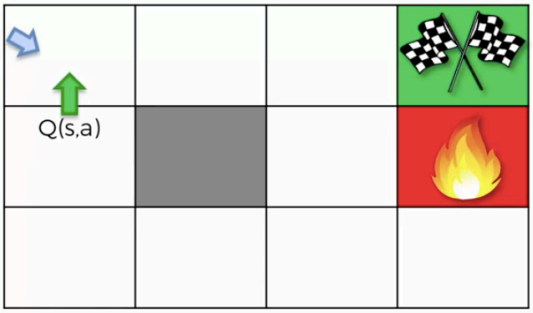


Before:

$Q(s,a)$

Now the agent take action

Temporal Difference



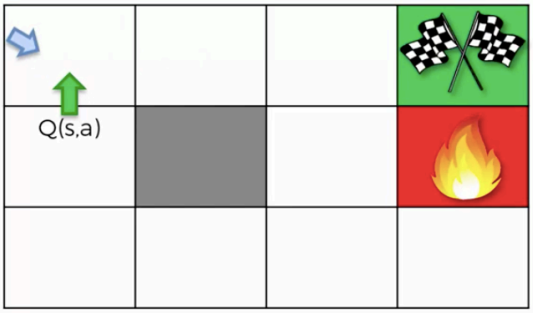
Before:

$Q(s,a)$

After:

$R(s,a) + \gamma \max_{a'} Q(s',a')$

Temporal Difference



Before:

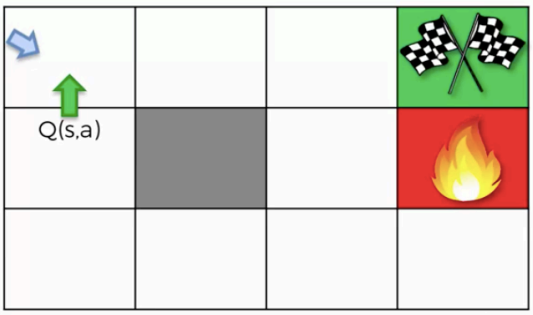
$Q(s,a)$

After:

$R(s,a) + \gamma \max_{a'} Q(s',a')$

$$TD(a,s) = R(s,a) + \gamma \max_{a'} Q(s',a') - Q(s,a)$$

Temporal Difference



Before:

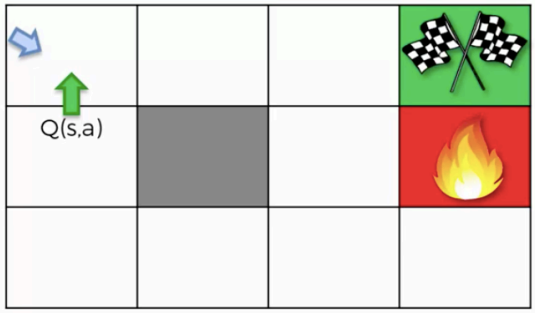
$Q(s,a)$

After:

$R(s,a) + \gamma \max_{a'} Q(s',a')$

$$TD(a,s) = R(s,a) + \gamma \max_{a'} Q(s',a') - Q(s,a)$$

Temporal Difference



Before:

$Q(s,a)$

After:

$R(s,a) + \gamma \max_{a'} Q(s',a')$

$$TD(a,s) = R(s,a) + \gamma \max_{a'} Q(s',a') - Q(s,a)$$

This is temporal difference and Ideally before and after should be same.
Remember that in here Q is what we did before after lots of iteration but after is what we've done now.

Temporal Difference

$$TD(a,s) = R(s,a) + \gamma \max_{a'} Q(s',a') - Q(s,a)$$

Temporal Difference

$$TD(a,s) = R(s,a) + \gamma \max_{a'} Q(s',a') - Q(s,a)$$

$$Q(s,a) = Q(s,a) + \alpha TD(a,s)$$

alfa is our learning rate

Temporal Difference

$$TD(a,s) = R(s,a) + \gamma \max_{a'} Q(s',a') - Q_{t-1}(s,a)$$

$$Q_t(s,a) = Q_{t-1}(s,a) + \alpha TD_t(a,s)$$

We are adding time in here for making more sense

Temporal Difference

$$TD(a,s) = R(s,a) + \gamma \max_{a'} Q(s',a') - Q_{t-1}(s,a)$$

$$Q_t(s,a) = Q_{t-1}(s,a) + \alpha TD_t(a,s)$$

$$Q_t(s,a) = Q_{t-1}(s,a) + \alpha \left(R(s,a) + \gamma \max_{a'} Q(s',a') - Q_{t-1}(s,a) \right)$$

If alpha in here is equal to 1 the two Q in t-1 will be eliminated, and if it's equal to 0 then the two Q in t and t-1 will be equal which it's not so good because we're not gonna learning anything. So, it's better that the alpha not be exactly 0 or 1. In here we want our TD to converge which that means we want the difference to be equal to 0.
The reason for using this the third formula is when our environment is constantly changing and modifying so we continuously need to learn because it's not possible to learn everything and come up with optimal policy. Because policy also changes over time. That's why we need constantly update the temporal difference and calculating the Q values

Additional Reading

Additional Reading:

Learning to Predict by the Methods of Temporal Differences

By Richard Sutton (1988)

Link:

https://link.springer.com/article/10.1007/BF00115009