Graphical user interface, text, application

Description automatically generated

A picture containing text

Description automatically generated

A picture containing text

Description automatically generated

A picture containing text

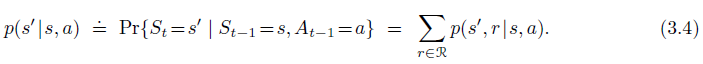
Description automatically generated

A picture containing text

Description automatically generated

A picture containing text

Description automatically generated





Text

Description automatically generated

Since the condition works for policy iteration according to the policy change to get an optimal one, but more to concern about is the iteration times or the value iteration result. Once we got an optimal value, that should be okay to stop our loop as a policy stable.

Text

Description automatically generated

Text

Description automatically generated

Value iteration does not have analogous bug. The value iteration is stable or not is not depends on the policy function but the value function. The policy is optimal as long as the result is optimal. Therefore, such bug does not exist.

Text

Description automatically generated

Text, letter

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Diagram

Description automatically generated

Text

Description automatically generated

We wish to arrive at state z as soon as possible as a result. The only action that reaches state z is through action b succeeds with low probability, so the agent should minimize the in the process. This suggests that the agent should try action b in state x and try action a to get to state x rather than directly to state in state y.

Text

Description automatically generated

Text, letter

Description automatically generated

Text

Description automatically generated

Therefore, in state y, the preferred action is b to move to state x. And state x is preferred with action a to move to state z.

Text

Description automatically generated

Text, letter

Description automatically generated

Text, letter

Description automatically generated

If the policy has action b in both state, it is unsolvable. With the discounting factor, the result is changed. The value range is from negative infinite to -2. By changing the gamma value, we van control the optimal policy for state y and x. So optimal policy is depended on the discount factor.

A picture containing text

Description automatically generated

Graphical user interface, text

Description automatically generated

A picture containing text, keyboard, electronics

Description automatically generated

A picture containing text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generatedText

Description automatically generated

A picture containing text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Diagram, schematic

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text, letter

Description automatically generated

Text

Description automatically generated

Text, letter

Description automatically generated

Text, letter

Description automatically generated

Text, letter

Description automatically generated