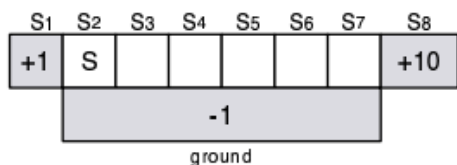


Important: Punctajul pentru exerciții se acordă doar pe baza unei rezolvări detaliate (scrieți mai întâi formulele și apoi înlocuiți cu valorile corespunzătoare)

Subiect 1: Considerăm un robot care se află pe o bârna de echilibru. Acțiunile posibile sunt: dreapta, stânga. Agentul se află în starea S2. Toate stările au recompensă 0, cu excepția stărilor terminale S1, S8, și ground. Acțiunea (stânga sau dreapta) determină o deplasare în acea direcție cu probabilitatea p . Cu probabilitatea $1 - p$, robotul cade de pe bârnă (ajunge în starea ground și primește recompensa -1).

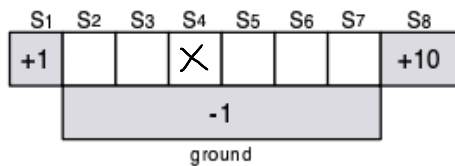


Pentru ce valori ale lui p acțiunea optimă (cea care maximizează utilitatea așteptată) din starea S2 este să te deplasezi spre dreapta și nu spre stânga atunci când $p = 1$?

Subiect 2: Plecând de la problema de mai sus, notăm acțiunea dreapta cu $\rightarrow$ și stânga cu $\leftarrow$ și presupunem ca tranzițiile pot genera alte recompense. Considerați că agentul se deplasează astfel și primește recompensele marcate cu bold: $[(s2, \rightarrow, s3, \mathbf{1}), (s3, \rightarrow, s4, \mathbf{2}), (s4, \leftarrow, s3, \mathbf{-1})]$. Valorile inițiale Q sunt setate pe 0, cu excepția: $Q(s2, \rightarrow) = 1$, $Q(s2, \leftarrow) = 2$, $Q(s3, \rightarrow) = 1.5$, $Q(s3, \leftarrow) = 0.5$, $Q(s4, \rightarrow) = 0.4$, $Q(s4, \leftarrow) = 0.6$. Aplicați detaliat pas cu pas ecuația de actualizare a valorilor Q pentru a determina noile valori Q . Rata de învățare este 0.5, iar factorul de discount 0.3.

Important: Punctajul pentru exerciții se acordă doar pe baza unei rezolvări detaliate (scrieți mai întâi formulele și apoi înlocuiți cu valorile corespunzătoare)

Subiect 1: Considerăm un robot care se află pe o bârna de echilibru. Acțiunile posibile sunt: dreapta, stânga. Agentul se află în starea S4. Toate stările au recompensă 0, cu excepția stărilor terminale S1, S8, și ground. Acțiunea (stânga sau dreapta) determină o deplasare în acea direcție cu probabilitatea p . Cu probabilitatea $1 - p$, robotul cade de pe bârnă (ajunge în starea ground și primește recompensa -1).



Pentru ce valori ale lui p acțiunea optimă (cea care maximizează utilitatea așteptată) din starea S4 este să te deplasezi spre stânga și nu spre dreapta atunci când $p = 1$?

Subiect 2: Plecând de la problema de mai sus, notăm acțiunea dreapta cu $\rightarrow$ și stânga cu $\leftarrow$ și presupunem ca tranzițiile pot genera alte recompense. Considerați că agentul se deplasează astfel și primește recompensele marcate cu bold: $[(s5, \rightarrow, s6, \mathbf{2}), (s6, \rightarrow, s7, \mathbf{-1}), (s7, \leftarrow, s6, \mathbf{1})]$. Valorile inițiale Q sunt setate pe 0, cu excepția: $Q(s5, \rightarrow) = 2$, $Q(s5, \leftarrow) = 1.5$, $Q(s6, \rightarrow) = 1$, $Q(s6, \leftarrow) = 0.6$, $Q(s7, \rightarrow) = 0.5$, $Q(s7, \leftarrow) = 0.4$. Aplicați detaliat pas cu pas ecuația de actualizare a valorilor Q pentru a determina noile valori Q . Rata de învățare este 0.3, iar factorul de discount 0.5.