Üzleti Intelligencia

6. Előadás: Mély Q-tanulási architektúrák

Kuknyó Dániel
Budapesti Gazdasági Egyetem

2023/24
1.félév

A Q-tanulás alapjai

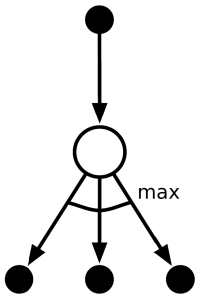
A megerősítéses tanulásban az egyik nagy áttörést egy politikafüggetlen TD algoritmus kifejlesztése hozta el.

Ebben az esetben a becsült **állapot-cselekvés minőség függvény**, Q, ami megadja, hogy mennyire jövedelmező az ügynöknek s állapotban a cselekvést végrehajtani.

Q-tanulás

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \underset{a}{max} Q(s_{t+1}, a) - Q(s_t, a_t) \right]$$

A Q-tanulás ezáltal egy teljesen online tanulási algoritmus, ami a követett **politikától függetlenül** garantáltan konvergálni fog a valós Q értékekhez.



Dupla Q-tanulás

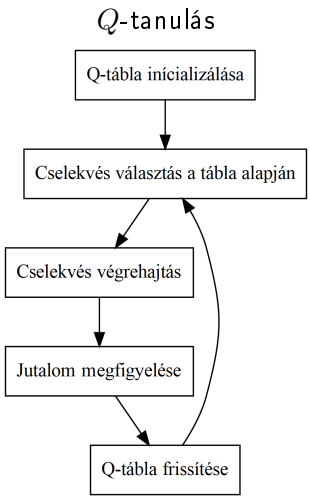
A kettős tanulás ötlete természetesen kiterjed a teljes MDP algoritmusaira. A Q-tanulásban a becsült Q-értékek torzítottak lehetnek, ha alacsony a minta számossága, vagy zaj van a rendszerben. Egy módja a Q-tanulás regularizálásának, ha egy helyet **két Q-táblát tart nyilván az algoritmus**, Q_1-et és Q_2-t.

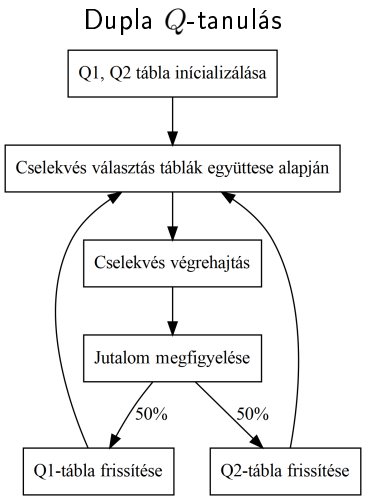
A Q-tanulással analóg dupla Q-tanulás nevű kettős tanulási algoritmus két részre osztja az időlépéseket, **minden lépésnél egy érmét feldobva**. Ha az érme fejre esik, a frissítés a következő:

$$Q_1(s_t, a_t) \leftarrow Q_1(s_t, a_t) + \alpha \left[r_{t+1} + \gamma Q_2(s_t + 1, \underset{a}{argmax} Q_1(s_{t+1}, a)) - Q_1(s_t, a_t) \right]$$

Ha az érme pedig írásra esik, akkor ugyanez a frissítés Q_1 és Q_2 felcserélésével történik, így Q_2 frissül. A két közelítő értékfüggvényt teljesen szimmetrikusan kezeli az algoritmus. Például egy ε-mohó politika a dupla tanulás esetében az egyes cselekvési értékbecslések **átlagára vagy összegére épülhet**.

Alapvető Q-tanulási eljárások



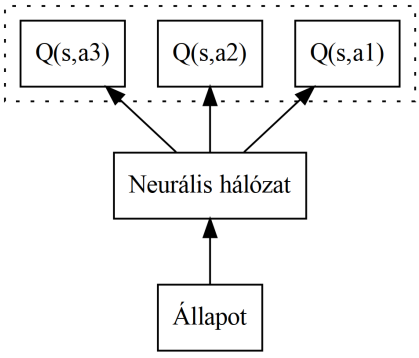


A Q-hálózat (DQN)

A Q-hálózat egy természetes kiterjesztése a hagyományos Q-tanulásnak. A naív Q-hálózat inputja a **környezetet leíró változók** vektora vagy mátrixa, és az outputja pedig az ügynök számára elérhető **cselekvések** $Q(s, a)$ **értéke** minden $a_1, a_2, .., a_n$ cselekvéshez tartozóan.

A cselekvés választáshoz az ügynök kiválasztja a **legnagyobb becsült** Q **értéket**, és az ahhoz tartozó cselekvést fogja végrehajtani.

A Q-hálózat költségfüggvénye az **átlagos négyzetes Bellman hiba**, a paraméter frissítése pedig a költségfüggvény gradiense és a lépésméret szerint történik.



A DQN költségfüggvénye

$$J(\theta) = E_{s,a,s' \sim D_{replay}} \left[\left(r + \gamma \underset{a'}{max} Q_\theta(s', a') - Q_\theta(s, a) \right)^2 \right]$$