Szakdolgozat kivonat

Sztochasztikus approximáció időben változó Markov döntési folyamatokban

Molnár Anna Enikő

Témavezető: Csáji Balázs Csanád

A megerősítéses tanulás (RL) egyike a gépi tanulás és a mesterséges intelligencia módszereinek, mely pozitív és negatív visszacsatolások alapján tanul a környezetből annak érdekében, hogy megtalálja az optimális politikát (stratégiát), azaz maximalizálni tudja a várható kumulatív nyereséget. A környezetet általában Markov döntési folyamattal (MDP) modellezik. Számos megoldási módszer ismert a megerősítéses tanulás területéről, amely kiszámítja vagy közelíti egy MDP optimális irányítási politikáját. Ezeknek a módszereknek a segítségével már számtalan problémát oldottak meg, mint például a szállítás és a készletgazdálkodás, csatorna elhelyezés, robot vezérlés, termelési ütemezés. Sok logikai játéknál – mint például az elmúlt évek egyik legnagyobb áttörését elérő AlphaGo – szintén a megerősítéses tanulást használják a stratégiák javítására.

Szakdolgozatomban Markov döntési folyamatokat vizsgálok időben változó környezetben. A változó környezet alatt azt értjük, hogy az átmenetvalószínűség és az azonnali költségek függvénye valamilyen korláton belül változhat. Először ismertetem a témához kapcsolódó legfontosabb alapfogalmakat. Ezután áttekintem, hogy különböző környezetbeli változások hatására egy diszkontált MDP optimális értékelőfüggvénye hogyan becsülhető, és bebizonyítom, hogy Lipschitz-folytonosan függ az átmenetvalószínűségtől, illetve az azonnali költségek függvényétől, azonban a diszkontálási faktortól nem. Emellett azt is megmutatom, hogy ezek a korlátokat az akció-értékelő függvényekre is ugyanúgy érvényesek. A továbbiakban tanulmányozom a sztochasztikus approximáció legfontosabb elemeit, bevezetem a Ljapunov-függvény fogalmát, mely az algoritmus egy valószínűségű konvergenciájának megállapítására ad módszert. Ezután bemutatok egy konvergencia eredményt, ami a Ljapunov-függvény tulajdonságaira támaszkodik, és felvázolom a tétel két fontos alkalmazását: a sztochasztikus gradiens algoritmust és az euklideszi norma pszeudo-kontrakciót. Ezt követően azt vizsgálom, hogy az értékelőfüggvényen alapuló tanulási módszerek hogyan viselkednek időben változó környezetekben. Ehhez először bevezetek egy olyan értékelőfüggvény operátort, amely időben változhat és az ilyen sztochasztikus approximációs algoritmusokra felírom a relaxált konvergencia tételt, melynek következményeként kapok egy approximációs tételt az értékelőfüggvényen alapuló RL algoritmusokra az (ε, δ)-MDP-kben. Végül felvázolok három algoritmust – az aszinkron értékiterációt, a Q-tanulást és a SARSA-t –, melyeken tanulva szemléltetem a környezeti változások hatását két problémán keresztül.