Temporal Difference Learning

Bevezetés a TD-tanulás elméletébe

Csáji Balázs Csanád

MDP szeminárium, SZTAKI

2005. december 20. – 2006. január 24.

Ajánlott irodalom

1. Dimitri P. Bertsekas, John N. Tsitsiklis: *Neuro-Dynamic Programming*; Athena Scientific, Belmont, Massachusetts (1996)

2. Richard S. Sutton, Andrew G. Barto: *Reinforcement Learning, An Introduction*; The MIT Press, Cambridge, Massachusetts (1998) http://www.cs.ualberta.ca/∼sutton/book/the-book.html

3. Dimitri P. Bertsekas: *Dynamic Programming and Optimal Control*; 2nd edition; Athena Scientific, Belmont, Massachusetts (2001)

”If one had to identify one idea as central and novel to reinforcement learning, it would undoubtedly be temporal-difference (TD) learning.” (Sutton és Barto, 1998)

A TD-tanulás néhány előnye

1. Nem szükséges hozzá a környezet dinamikájának explicit modellje, azaz az átmenetvalószínűségek és az azonnali költségek ismerete, elég ha (pl. egy programmal) szimulálni tudjuk a környezetet;
2. Nem szükséges a TD-tanuláshoz, hogy a szimuláció legvégére érjünk és az összes felhalmozott költséget megismerjük, néhány aktuális azonnali költség ismeretében is képes tanulni; „öntöltő” (bootstrap);

Jelölések - Markov Döntési Problémák

Egy (véges, diszkrét idejű, stacionárius) Markov Döntési Probléma (MDP) egy rendezett 6-os, $\mathcal{M} = \langle S, A, U, p, g, \alpha \rangle$, ahol

- $S = \{1, \ldots, n\}$ állapotok egy véges halmaza; 0 jelöli a ”nyelő” vagy ”cél” állapotot sztochasztikus legrövidebb út (SSP) problémák esetén
- A akciók egy véges halmaza
- $U : S \to \mathcal{P}(A)$ egy akciómegszorító függvény ($\mathcal{P}$ hatványhalmazt jelöl)
- $p : S \times A \to \Delta(S)$ az átmenetvalószínűségek függvénye, $p_{ij}(u)$ annak valószínűsége, hogy az i állapotból az u akcióval a j állapotba jutunk
- $g : S \times A \times S \to \mathbb{R}$ az azonnali költségek függvénye
- $\alpha \in [0, 1)$ a diszkontálási faktor

Emlékeztető - Bellman egyenlet

Politika (determinisztikus, stacionárius, Markov) $\pi : S \to A$

Egy politika értékelőfüggvénye $J^\pi : S \to \mathbb{R}$, amelynek definíciója

$$J^\pi(i) = \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t g(I_t, U_t, I_{t+1}) \;\middle|\; I_0 = i, U_t = \pi(I_t), I_{t+1} \sim p(I_t, U_t)\right]$$

$\pi_1 \leq \pi_2 \equiv \forall i : J^{\pi_1}(i) \leq J^{\pi_2}(i)$; π optimális, ha $\forall \pi' : \pi \leq \pi'$

Mindig létezik legalább egy optimális politika és az összes optimális politikának azonos és egyértelműen meghatározott az értékelőfüggvénye, amely minden i állapotra kielégíti a Bellman optimalitási egyenletet

$$J^*(i) = \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u)\Big(g(i,u,j) + \alpha J^*(j)\Big)$$

Politika kiértékelés Monte Carlo szim.-val

Tekintsünk egy sztochasztikus legrövidebb út problémát és tegyük fel, hogy adott egy π politika és keressük a hozzá tartozó J^π értékelőfüggvényt.
A rövidség kedvéért jelölje $g(i,j) := g(i,\pi(i),j)$; $p_{ij} := p_{ij}(\pi(i))$.

Sok trajektóriát generálva, amelyek mind 0-ban végződnek, inkrementálisan frissítjük az értékelőfüggvény becslést minden a trajektóriabeli i állapotra.
Legyen $\left\langle i_0^t, i_1^t, \ldots, i_{N_t}^t \right\rangle$ a t-edik trajektória ($i_{N_t}^t = 0$), legyen továbbá

$$C_t(i_k^t) = g(i_k^t, i_{k+1}^t) + \cdots + g(i_{N_t-1}^t, i_{N_t}^t)$$

Ekkor $J_0(i) = 0$ -ból indulva a következőképpen frissíthetjük a becsléseket

$$\boxed{J_{t+1}(i) = J_t(i) + \gamma_t(i)\left(C_t(i) - J_t(i)\right),} \qquad (1)$$

ahol pl. $\gamma_t(i) = 1/m(i)$ ($m(i)$ az eddigi $J(i)$-n végzett frissítések száma)

TD-tanulás - TD(1)

Jelölés: mostantól végtelennek tekinthetünk minden trajektóriát. Ha adott egy $\langle i_0, i_1, \ldots, i_N \rangle$ trajektória, akkor $\forall k > N : i_k = 0$ és $g(i_k, i_{k+1}) = 0$.

A Monte Carlo frissítést (1) átírhatjuk a következő alakba

$$J_{t+1}(i_k^t) = J_t(i_k^t) + \gamma_t(i_k^t) \cdot$$

$$\cdot \Big(g(i_k^t, i_{k+1}^t) + J_t(i_{k+1}^t) - J_t(i_k^t) +$$

$$+ g(i_{k+1}^t, i_{k+2}^t) + J_t(i_{k+2}^t) - J_t(i_{k+1}^t) +$$

$$\vdots$$

$$+ g(i_{N_t-1}^t, i_{N_t}^t) + J_t(i_{N_t}^t) - J_t(i_{N_t-1}^t) \Big),$$

ahol kihasználtuk, hogy $J_t(i_{N_t}^t) = 0$, és i_k^t előfordul az $\langle i_0^t, \ldots, i_{N_t}^t \rangle$-ben.

TD-tanulás - TD(1)

Vezessük be a következő jelölést (”temporal difference”)

$$d_{k,t} = g(i_k^t, i_{k+1}^t) + J_t(i_{k+1}^t) - J_t(i_k^t),$$

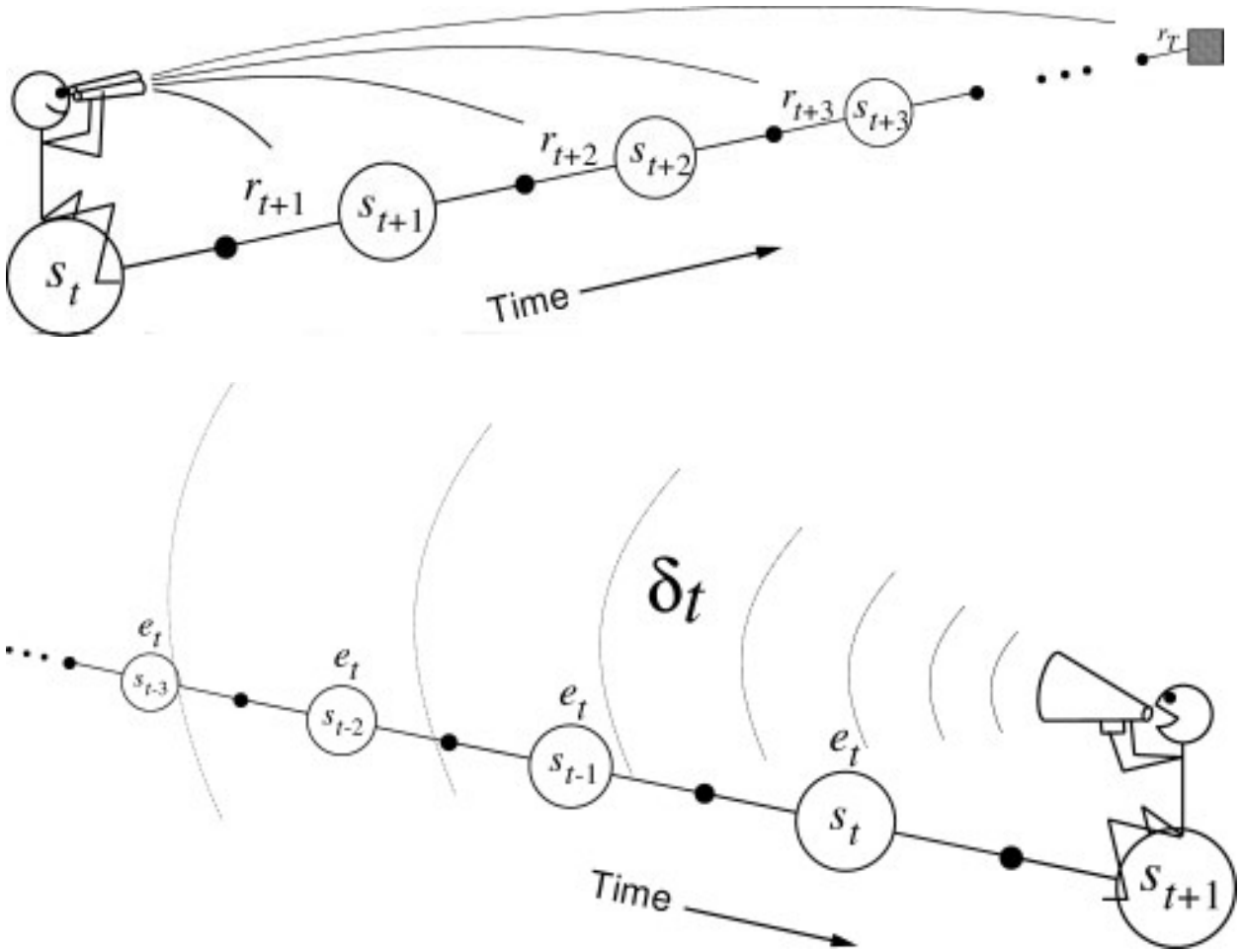
ekkor – nyilván – az (1) frissítési képlet úgy is írható, hogy

$$J_{t+1}(i_k^t) = J_t(i_k^t) + \gamma_t(i_k^t)\,(d_{k,t} + d_{k+1,t} + \cdots + d_{N_t-1,t}) \qquad (2)$$

Vegyük észre, hogy $d_{m,t}$ a t-edik szimuláció $m+1$-edik lépése után már kiszámolható, valamint $d_{m,t}$ előfordul az összes $J(i_k^t)$ frissítésénél, ahol $k \leq m$. Tehát, azonnal frissíthetünk az $m+1$-edik lépés után, és a változás

$$\Delta J(i_k^t) = \gamma_t(i_k^t)\, d_{m,t} \qquad \forall k \in \{1, \ldots, m\} \qquad (3)$$

Figyelem: az off-line (2) és az on-line (3) változat között van egy kismértékű különbség, ha a trajektóriában van olyan állapot, amely többször is előfordul.


r_T
r_{t+3}
s_{t+3}
r_{t+2}
s_{t+2}
r_{t+1}
s_{t+1}
s_t
Time
δt
e_t
s_{t-3}
s_{t-2}
s_{t-1}
s_t
s_{t+1}
Time

TD-tanulás - TD(0)

A Monte Carlo alapú politika kiértékelés (1) tekinthető egy Robbins-Monro típusú sztochasztikus approximációs eljárásnak, amely a

$$J^{\pi}(i) = \mathbb{E}\left[\sum_{m=0}^{\infty} g(I_m, I_{m+1}) \;\middle|\; I_0 = i\right]$$

egyenleteteket oldja meg; i az állapottér elemeit futja be. Ennek mintájára egy sor hasonló közelítő eljárást lehet definiálni, például a

$$J^{\pi}(i) = \mathbb{E}\left[g(I_0, I_1) + J^{\pi}(I_1) \mid I_0 = i\right],$$

amely egyenletekre épülő sztochasztikus approximácós eljárás a következő

$$J_{t+1}(i_k^t) = J_t(i_k^t) + \gamma_t(i_k^t)\left(g(i_k^t, i_{k+1}^t) + J_t(i_{k+1}^t) - J_t(i_k^t)\right),$$

amelyet szokás TD(0)-nak nevezni, később ismertetett okokból.

TD-tanulás - TD(λ)

Az előzőekhez hasonlóan kiindulhatnánk a következő egyenletből

$$J^{\pi}(i) = \mathbb{E}\left[\sum_{m=0}^{l} g(I_m, I_{m+1}) + J^{\pi}(I_{l+1}) \;\middle|\; I_0 = i\right] \qquad (4)$$

Mivel nem tudjuk, hogy melyik l-et kellene preferálnuk tegyük a következőt.

Vegyük a súlyozott összegét az összes többlépésű Bellman egyenletnek.

Pontosabban, rögzítsünk egy $\lambda < 1$ számot és szorozzuk meg (4)-et $(1-\lambda)\lambda^l$-el és összegezzünk az összes nemnegatív l-re

$$J^{\pi}(i) = (1-\lambda)\,\mathbb{E}\left[\sum_{l=0}^{\infty} \lambda^l \left(\sum_{m=0}^{l} g(I_m, I_{m+1}) + J^{\pi}(I_{l+1})\right) \;\middle|\; I_0 = i\right]$$

TD-tanulás - TD(λ)

$$J^\pi(i) = (1-\lambda)\,\mathbb{E}\left[\sum_{l=0}^{\infty}\lambda^l\left(\sum_{m=0}^{l} g(I_m, I_{m+1}) + J^\pi(I_{l+1})\right)\,\middle|\, I_0 = i\right] =$$

átrendezéssel és felhasználva, hogy $(1-\lambda)\sum_{l=m}^{\infty}\lambda^l = \lambda^m$ kapjuk, hogy

$$= \mathbb{E}\left[(1-\lambda)\sum_{m=0}^{\infty} g(I_m, I_{m+1})\sum_{l=m}^{\infty}\lambda^l + \sum_{l=0}^{\infty} J^\pi(I_{l+1})(\lambda^l - \lambda^{l+1})\,\middle|\, I_0 = i\right] =$$

$$= \mathbb{E}\left[\sum_{m=0}^{\infty}\lambda^m\left(g(I_m, I_{m+1}) + J^\pi(I_{m+1}) - J^\pi(I_m)\right)\,\middle|\, I_0 = i\right] + J^\pi(i)$$

Véges $\langle i_0, \ldots, i_N\rangle$ trajektóriáknál $\forall k > N : i_k = 0$ és $g(i_k, i_{k+1}) = 0$.

TD-tanulás - TD(λ)

Alkalmazzuk azt a (”temporal difference”) jelölést, hogy

$$D_m^\pi = g(I_m, I_{m+1}) + J^\pi(I_{m+1}) - J^\pi(I_m),$$

ekkor ahhoz a súlyozott Bellman-típusú egyenlethez jutunk, hogy

$$J^\pi(i) = \mathbb{E}\left[\sum_{m=0}^{\infty} \lambda^m D_m^\pi \;\middle|\; I_0 = i\right] + J^\pi(i)$$

Az erre alapuló Robbins-Monro típusú sztochasztikus approximációs eljárás

$$\boxed{J_{t+1}(i_k^t) = J_t(i_k^t) + \gamma_t(i_k^t) \sum_{m=k}^{\infty} \lambda^{m-k} d_{m,t},}$$

ahol $\gamma_t(i_k^t)$ egy lépésköz paraméter, és $d_{m,t}$ ugyanaz, mint korábban.

TD-tanulás - TD(λ), TD(1), TD(0)

Megjegyzés: a TD(λ) tanulási formula $\lambda = 0$ illetve $\lambda = 1$-ben vett határértékeiben a TD(0)-át illetve a TD(1)-et adja eredményül.

$$J_{t+1}(i_k^t) = J_t(i_k^t) + \gamma_t(i_k^t) \sum_{m=k}^{\infty} \lambda^{m-k} d_{m,t} \qquad \text{TD}(\lambda)$$

$$J_{t+1}(i_k^t) = J_t(i_k^t) + \gamma_t(i_k^t) \sum_{m=k}^{\infty} d_{m,t} \qquad \text{TD(1)}$$

$$J_{t+1}(i_k^t) = J_t(i_k^t) + \gamma_t(i_k^t)\, d_{k,t} \qquad \text{TD(0)}$$

TD-tanulás - ”Every-Visit” változat

Mi történik, ha egy trajektória folyamán egy állapot többször is előfordul?

”Every-Visit” változat: rögzítsünk egy i állapotot és egy trajektóriát. Tegyük fel, hogy i összesen M-szer fordul elő a t-edik trajektóriában, konkrétan a szimuláció $m_1, m_2, \ldots, m_M$ lépéseiben. Ekkor a TD(λ) eljárás

$$J_{t+1}(i) = J_t(i) + \gamma_t(i) \sum_{j=1}^{M} \sum_{m=m_j}^{\infty} \lambda^{m-m_j} d_{m,t},$$

ez olyan mintha az általános képletben több mintát használtunk volna a jobboldalon. Azonban ezek a minták korreláltak, mivel ugyanaz a $d_{m,t}$ több j-re is előfordulhat. Ez azonban nem rontja el az eljárást (indoklás később).

TD-tanulás - ”First-Visit” változat

”First-Visit” változat: ha i először az m_1 lépésben fordult elő

$$J_{t+1}(i) = J_t(i) + \gamma_t(i) \sum_{m=m_1}^{\infty} \lambda^{m-m_1} d_{m,t},$$

$\lambda = 1$ esetén, TD(1) két változata megegyezik a Monte Carlo politika kiértékelés kétfajta változatával. Mindkét változat garantáltan konvergál, de a ”first-visit” változatnak kisebb átlagos négyzetes hibája, így az az ajánlott.

$\lambda < 1$ esetén nem ez a helyzet. Ha nagyon hosszú trajektóriákkal dolgozunk, például ha $n \ll N$, akkor a szimuláció során kapott legtöbb információt nem használjuk a ”first-visit”-nél, ami arra utal, hogy talán nem ez a jó választás.

TD-tanulás - Off-line és On-line változat

Off-line változat: az összes állapot egyidejű frissítése egy trajektóriában

On-line változat: a szimuláció minden lépése után frissítünk, azaz a t-edik szimuláció $\left\langle i_m^t, i_{m+1}^t \right\rangle$ lépése után azonnal frissítjük $J_t(i_k^t)$-t ($k \leq m$)

$$J_{0,t}(i) = J_t(i) \qquad \forall i$$

$$d_{m,t} = g(i_m^t, i_{m+1}^t) + J_{m,t}(i_{m+1}^t) - J_{m,t}(i_m^t)$$

$$J_{m+1,t}(i_k^t) = J_{m,t}(i_k^t) + \gamma_t(i_k^t)\lambda^{m-k} d_{m,t} \qquad \forall k \leq m$$

$$J_{t+1}(i) = J_{N_t,t}(i) \qquad \forall i$$

Az off-line és az on-line változat közötti különbség $O(\gamma^2)$ nagyságrendű.

TD(λ) konvergencia tétel

Állítás. Sztochasztikus legrövidebb út problémák esetén, ha a π politika megfelelő, akkor $J_0 \equiv 0$ -ból az off-line first-visit TD(λ)-val kapott J_t sorozat egy valószínűséggel konvergál J^π-hez, feltéve, hogy minden állapot végtelen sok trajektóriában szerepel és a lépésközre teljesülnek a szokásos feltételek.

Bizonyítás (vázlat). Emlékezzünk vissza (NDP könyv 2.2.1-es fejezet), hogy létezik egy olyan pozitív ξ vektor, hogy $\|PJ\|_\xi \leq \beta \|J\|_\xi$, ahol $\beta < 1$ és P egy $n \times n$-es mátrix, ahol $P(i,j) = p_{ij}$. A TD(λ) kiinduló egyenlete ekkor

$$J^\pi = G + (1-\lambda) \sum_{l=0}^{\infty} \lambda^l P^{l+1} J^\pi, \tag{5}$$

ahol G egy konstans vektor (a várható azonnali költségekből).

TD(λ) konvergencia tétel

Kihasználva P kontrakciós tulajdonságát kapjuk, hogy

$$\left\| (1-\lambda) \sum_{l=0}^{\infty} \lambda^l P^{l+1} J \right\|_\xi \leq (1-\lambda) \sum_{l=0}^{\infty} \lambda^l \left\| P^{l+1} J \right\|_\xi \leq$$

$$\leq \beta(1-\lambda) \sum_{l=0}^{\infty} \lambda^l \left\| J \right\|_\xi \leq \beta \left\| J \right\|_\xi \qquad \forall J\text{-re}$$

Tehát (5) egyenlet $J^\pi = H J^\pi$ alakú, ahol H kontrakció a súlyozott maximum normában. Tudjuk, hogy TD(λ) az (5)-re épülő Robbins-Monro típusú sztochasztikus approximációs eljárás. A H operátor J^π fixpontjához való konvergenciája egyszerű következménye a már ismert iteratív sztochasztikus approximációs eljárások konvergenciájának (NDP könyv 4.3.2 fejezet). Q.E.D.

TD(λ) diszkontált problémákra

Diszkontált problémák esetén két út kínálkozik számunkra:

(1) Az ismert módon átalakítjuk a diszkontált MDP-t egy nem-diszkontált sztochasztikus legrövidebb út problémává, és alkalmazzuk az eddigi eredményeket. Az átalakítás során az α diszkontálási faktort úgy tekintjük, mint egy folytatási valószínűséget és $1-\alpha$ valószínűséggel mindig leállunk.

(2) Sokkal hatékonyabb megoldást ad, amikor a diszkontált problémák l lépéses Bellman egyenleteiből indulunk ki és alkalmazzunk a λ hatványaival való súlyozott összeg „trükköt” a TD(λ) levezetésénél.

$$J^{\pi}(i) = (1-\lambda)\,\mathbb{E}\left[\sum_{l=0}^{\infty}\lambda^{l}\left(\sum_{m=0}^{l}\alpha^{m}g(I_m,I_{m+1})+\alpha^{l+1}J^{\pi}(I_{l+1})\right)\;\middle|\;I_0=i\right]$$

TD(λ) diszkontált problémákra

Ekkor – a sztochasztikus legrövidebb út problémáknál ismertetett átalakításokkal analóg módon – a következő egyenlethez jutunk

$$J^\pi(i) = \mathbb{E}\left[\sum_{m=0}^{\infty} (\alpha\lambda)^m D^\pi_{\alpha,m} \;\middle|\; I_0 = i\right] + J^\pi(i),$$

ahol $D^\pi_{\alpha,m}$ ismét a ”temporal difference” együthatókat jelöli

$$D^\pi_{\alpha,m} = g(I_m, I_{m+1}) + \alpha J^\pi(I_{m+1}) - J^\pi(I_m)$$

Az ezen alapuló sztochasztikus approximációs eljárás

$$\boxed{J_{t+1}(i^t_k) = J_t(i^t_k) + \gamma_t(i^t_k) \sum_{m=k}^{\infty} (\alpha\lambda)^{m-k} d_{\alpha,m,t}}$$

Példa: az összköltség varianciája

Tekintsünk egy olyan MDP-t, amelynek csak egy állapota van, $i = 1$.

A diszkont faktor $\alpha < 1$. Az eddigiektől eltérően legyen g valószínűségi változó 0 várható értékkel és σ^2 varianciával (nyilván $J^\pi(1) = 0$).

Alakítsuk át az MDP-t egy sztochasztikus legrövidebb út problémává.

Ekkor két állapot lesz $\{0, 1\}$ és α megállási valószínűségnek tekinthető.

Mekkora lesz a összes költség varianciája off-line first-visit TD(1)-esetén?

Egy trajektóriánál jelölje N v.v. az 1-be vivő átmenetek számát a megállásig.

$$\mathbb{E}\left[(g_1 + g_2 + \cdots + g_N)^2\right] = \sum_{k=1}^{\infty} \mathbb{E}\left[(g_1 + g_2 + \cdots + g_k)^2\right] \mathbb{P}(N = k) =$$

$$= \sum_{k=1}^{\infty} k\sigma^2 \, \mathbb{P}(N = k) = \sigma^2 \, \mathbb{E}\left[N\right] = \frac{\sigma^2}{1 - \alpha} \qquad (6)$$

Példa: az összköltség varianciája

Tekintsünk az előző egy állapotú MDP-t, de most ne alakítsuk át SSP problémává, hanem az off-line first-visit TD(1) azt a változatát használjuk, amikor α megjelenik a TD együtthatókban, azaz amikor

$$d_{\alpha,m,t} = g(i_m^t, i_{m+1}^t) + \alpha J_t(i_{m+1}^t) - J_t(i_m^t)$$

Ekkor egy trajektória során összegyűjtött költségek $g_1 + \alpha g_2 + \alpha^2 g_3 + \ldots$ alakúak, ahol g_k a k-adik átmenet során elszenvedett költség. Ismét az a kérdés, hogy mekkora lesz a összes költség varianciája egy trajektóriában?

$$\mathbb{E}\left[\sum_{k=0}^{\infty} (\alpha^k g_{k+1})^2\right] = \sigma^2 \sum_{k=0}^{\infty} \alpha^{2k} = \frac{\sigma^2}{1-\alpha^2} = \frac{\sigma^2}{(1-\alpha)(1+\alpha)}$$

Ha α közel van 1-hez, akkor egy kettes szorzót nyertünk (6)-hoz képest.

Végtelen trajektóriák – diszkontált eset

(1) Bevezethetünk egy 1 valószínűséggel véges N_t megállási paramétert és csak N_t-ig szimuláljuk a t-edik trajektóriát. Az előző példa jelöléseivel ez azt jelenti, hogy $g_1 + \alpha g_2 + \alpha^2 g_3 + \dots$ helyett azt használjuk, hogy

$$g_{1,t} + \alpha g_{2,t} + \alpha^2 g_{3,t} + \dots + \alpha^{N_t-1} g_{N_t,t} + \alpha^{N_t} J_t(i_{N_t}),$$

ahol $\alpha^{N_t} J_t(i_{N_t})$ az N_t-edik átmenet után keletkezett TD-együtthatóból maradt. Mivel ekkor minden trajektóriát egy valószínűséggel elég véges ideig szimulálni, használhatjuk az off-line TD-tanulást.

(2) Nem állítjuk le a trajektóriát és on-line TD-tanulást használunk. Mivel így csak egy trajektóriánk van, a γ lépésköz paraméternek átmenetről átmenetre kell változnia és nem trajektóriáról trajektóriára, mint eddig.

Végtelen trajektóriák – diszkontált eset

Az on-line TD-tanulás végtelen trajektória esetén

$$J_{m+1}(i) = J_m(i) + \gamma_m(i) z_m(i) d_{\alpha,m}(i),$$

ahol $\gamma_m(i)$ egy nem-negatív lépésköz paraméter, és $z_m(i)$ pedig

$$z_m(i) = \begin{cases} \alpha\lambda z_{m-1}(i) & \text{ha } i_m \neq i \\ \alpha\lambda z_{m-1}(i) + 1 & \text{ha } i_m = i \end{cases}$$

amely az every-visit TD(λ)-nak felel meg. A TD(λ) „újrakezdő” változata

$$z_m(i) = \begin{cases} \alpha\lambda z_{m-1}(i) & \text{ha } i_m \neq i \\ 1 & \text{ha } i_m = i \end{cases}$$

Feltettük, hogy $z_{-1}(i) = 0$. Mindkét változat konvergál, lásd később.

Általánosított TD-tanulás

A TD(λ) egy általánosításához tegyük fel, hogy a t-edik szimuláció során egy $\langle i_0^t, i_1^t, \dots \rangle$ trajektóriát generáltunk és $d_{m,t}$ jelölje a hozzá tartozó ”temporal difference” együtthatókat. Tekintsük ekkor a következő iteratív eljárást

$$J_{t+1}(i) = J_t(i) + \gamma_t(i) \sum_{m=0}^{\infty} z_m^t(i) d_{m,t},$$

ahol $z_m^t(i) \geq 0$ az un. alkalmazhatósági együtthatók (”eligibility coefficients”).

Egy általánosított konvergenciatétel kimondása előtt néhány példát ismertetünk. Ezekben feltesszük, hogy az i rögzített állapotot $m_1, \dots, m_M$ alkalmakkor látogattuk meg egy rögzített trajektóriában. Alkalmazzuk $m_{M+1} = \infty$ jelölést. Mivel TD(λ) csak akkor frissít egy állapotot, ha az az állapot már látogatva volt, ezért feltesszük, hogy $z_m(i) = 0$ ha $m < m_1$.

Általánosított TD-tanulás - példák

$$J_{t+1}(i) = J_t(i) + \gamma_t(i) \sum_{m=0}^{\infty} z_m^t(i) d_{m,t},$$

Rögzítünk egy trajektóriát és egy $m_1, \ldots, m_M$ -kor érintett i állapotot.

(1) "First-Visit" TD(λ): $z_m(i) = \lambda^{m-m_1}$ ha $m \geq m_1$

(2) "Every-Visit" TD(λ): $z_m(i) = \sum_{\{j|m_j \leq m\}} \lambda^{m-m_j}$

(3) "Restarting" TD(λ): $z_m(i) = \lambda^{m-m_j}$ ha $m_j \leq m \leq m_{j+1}$

(Számítógépes kísérletek alapján hatékonyabb, mint az "every-visit" változat.)

(4) "Stopping" TD(λ): $z_m(i) = \lambda^{m-m_1}$ ha $m < \tau$ és $z_m(i) = 0$ ha $m \geq \tau$

(Az eljárás megállását a τ valószínűségi változó adja meg, amelyet teljes egészében meghatároz a szimuláció története az aktuális állapotig.)

(5) "Stopping in X" TD(λ): lásd előző, de $\tau = \min\{k \geq 1 \mid i_k \in X\}$

Általánosított TD-tanulás - A1 feltételek

Az ”eligibility” együtthatóktól megkövetelt tulajdonságok (A1):

1. $z_m^t(i) \geq 0$
2. $z_{-1}^t(i) = 0$
3. $z_m^t(i) \leq z_{m-1}^t(i)$ ha $i_m^t \neq i$
4. $z_m^t(i) \leq z_{m-1}^t(i) + 1$ ha $i_m^t = i$
5. $z_m^t(i)$-t teljes mértékben meghatározza $\mathcal{F}_t$ és $i_0^t, \ldots, i_m^t$

Ahol t jelöli a szimulált trajektória indexét, $\mathcal{F}_t$ pedig a folyamat múltját a t-edik trajektóriáig. A (2) egyszerű konvenció a kezdeti kinullázáshoz, a (3)-al együtt biztosítja, hogy $z_m^t(i) = 0$ ha i nem volt az m-edik lépésig meglátogatva a t-edik szimuláció során. A (4) feltétel főleg az ”every-visit” változatokhoz kell.

Általánosított TD-tanulás - A2 feltételek

A megkövetelt tulajdonságok második csoportja (A2):

1. Minden i-re és t-re, $z_m^t(i) = 1$ az első olyan m-re, amelyre $z_m^t(i) > 0$.
2. Létezik egy $\delta > 0$ konstans, amelyre $q_t(i) \geq \delta$ minden $t \in T^i$-re és i-re.
3. Minden $t \in T^i$-re $\gamma_t(i) \geq 0$ és minden $t \notin T^i$-re $\gamma_t(i) = 0$.
4. $\sum_{t \in T^i} \gamma_t(i) = \infty$ minden i-re.
5. $\sum_{t \in T^i} \gamma_t^2(i) < \infty$ minden i-re.

Ahol $q_t(i) = \mathbb{P}(\exists m : z_m^t(i) > 0 \mid \mathcal{F}_t)$ és $T^i = \{t \mid q_t(i) > 0\}$.

T_i azoknak a trajektóriáknak a halmaza, amelyek megváltoztatják $J(i)$-t.

A $\gamma_t(i)$ a t-edik trajektória alatt az i állapotra érvényes lépésköz (”stepsize”).

A (2) feltétel nélkül is igaz marad a konvergenciát kimondó állítás.

Általánosított TD-tanulás - konvergencia

Állítás. Tekintsük a következő (általánosított off-line TD-tanulás) eljárást

$$J_{t+1}(i) = J_t(i) + \gamma_t(i) \sum_{m=0}^{N_t-1} z_m^t(i) d_{m,t},$$

$$d_{m,t} = g(i_m^t, i_{m+1}^t) + J_t(i_{m+1}^t) - J_t(i_m^t)$$

Ha a π politika megfelelő (proper) és fennállnak az A1 és A2 tulajdonságok, akkor $\forall i$-re $J_0(i) = 0$ -ból kiindulva a $J_t(i)$ sorozat egy valószínűséggel konvergál $J^\pi(i)$-hez. (Bizonyítás az NDP könyv 5.3.2 fejezetében. Az alapötlete, mint a TD(λ)-nál, hogy megmutatjuk a definiált sztochasztikus approximációs operátorról, hogy pszeudó-kontrakció a súlyozott maximum normában. A $z_m^t(i)$-ket a lépésközbe integráljuk. Nehézséget okoz, hogy $z_m^t(i)$-k trajektóriánként változhatnak, az iterációs operátor idő-függő.)

Általánosított TD-tanulás - on-line változat

A t-edik trajektória elején rendelkezésünkre áll az aktuális J_t vektor, kiválasztunk egy i_0^t kezdőállapotot és szimulálunk egy trajektóriát i_0^t-ből.

$$J_{0,t}(i) = J_t(i) \qquad \forall i$$

$$d_{m,t} = g(i_m^t, i_{m+1}^t) + J_{m,t}(i_{m+1}^t) - J_{m,t}(i_m^t)$$

$$J_{m+1,t}(i) = J_{m,t}(i) + \gamma_t(i) z_m^t(i) d_{m,t} \qquad \forall i$$

$$J_{t+1}(i) = J_{N_t,t}(i) \qquad \forall i$$

Az on-line változat konvergenciájához kell, hogy $z_m^t(i)$ felülről korlátos, amely a klasszikus változatok közül egyedül az every-visit TD(1)-re nem teljesül.
Az off-line és az on-line változatok közötti különbség $\forall i : O(\gamma_t^2(i))$.

Általánosított TD-tanulás - diszkontálás

Tekintsük azt a változatot, amikor a t-edik trajektóriát leállítjuk N_t lépés után. Ekkor $z_m^t(i) = 0$ ha $m \geq N_t$. A diszkontált problémáknál jelentkező főbb különbségek az előzőkhöz képest, hogy α most is bejön a TD-együtthatókba

$$d_{\alpha,m,t} = g(i_m^t, i_{m+1}^t) + \alpha J_t(i_{m+1}^t) - J_t(i_m^t)$$

Továbbá a konvergenciához A1-es 3. feltétele helyett azt kell feltenni, hogy

$$z_m^t(i) \leq \alpha z_{m-1}^t(i) \qquad \text{ha } i_m^t \neq i$$

Valamint kell még egy további tulajdonság

$$\mathbb{P}(N_t \geq k \mid \mathcal{F}_t) \leq A\rho^k \qquad \forall k, t,$$

ahol A és $\rho < 1$ nem-negatív konstansok. Ezekkel a módosításokkal az általánosított TD-tanulás konvergál, mind az off-line mind az on-line esetben.

Optimista Politika Iteráció

Eddig egy rögzített π politika kiértékelésével foglalkoztuk.

Hogyan találhatjuk meg TD-tanulással J^*-ot vagy egy optimális politikát?

(1) Kiértékeljük a politikát TD-tanulással, majd vesszük a kapott értékelésre mohó politikát, majd azt is kiértékeljük, stb. Tehát, miután „elég sokáig” alkalmaztuk a TD-tanulást a π_k politika kiértékelésére, vehetjük a kapott $\hat{J}^{\pi_k}$ állapotértékelésre mohó politikát, azaz egy olyan π_{k+1} politikát, amelyre

$$\pi_{k+1}(i) \in \arg\min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u) \Big(g(i,u,j) + \alpha \hat{J}^{\pi_k}(j) \Big)$$

Ha π_{k+1} mohó J^{π_k}-re, akkor az eljárás tekinthető egy „sima” politika iterációnak, ahol a politika kiértékelést TD-tanulással végeztük. A politikai iterációról már tudjuk, hogy konvergál J^*-hoz ill. egy optimális politikához.

Optimista Politika Iteráció

(2) Optimista módon, már néhány (pl. egy) frissítés után azt a politikát használjuk, amelyik mohó az aktuális állapotértékelésre. Példa: TD(0)

A „sima” TD(0) tanulás frissítési szabálya (tekintsük a diszkontált esetet)

$$J_{t+1}(i) = J_t(i) + \gamma_t(i)\Big(g(i,\pi(i),j) + \alpha J_t(j) - J_t(i)\Big)$$

ahol j-t a rögzített π politika és az átmenetvalószínűségek segítségével generáltuk, azaz a j-be menés valószínűsége $p_{ij}(\pi(i))$. Ez átírható

$$J_{t+1}(i) = (1-\gamma_t(i))J_t(i) + \gamma_t(i)(T_\pi J_t)(i) + \gamma_t(i)w_t,$$

ahol w_t egy 0 várhatóértékű zaj, és T_π a π-hez tartozó kiértékelő operátor

$$(T_\pi J)(i) = \sum_{j=1}^{n} p_{ij}(\pi(i))\Big(g(i,\pi(i),j) + \alpha J(j)\Big)$$

Optimista Politika Iteráció

Az TD(0) optimista változatát akkor kapjuk, ha minden frissítési lépés után azonnal megváltoztatjuk a π politikát az aktuális értékelőfüggvényre mohó politikára. Azaz mindig fennáll, hogy $T_{\pi_t} J_t = T J_t$, ekkor a frissítés

$$J_{t+1}(i) = (1 - \gamma_t(i)) J_t(i) + \gamma_t(i)(T J_t)(i) + \gamma_t(i) w_t,$$

ahol w_t ismét egy 0 várhatóértékű zajt jelöl. A zajtól eltekintve ez egy aszinkron érték iteráció kis lépsközzel. Ekkor a konvergencia egyszerű következménye az NDP könyv 4.3 fejezet tételeinek; azon feltételek mellet, hogy minden állapotra végtelen sok frissítés történik, valamint, hogy a lépésközre teljesüljenek a szokásos feltételek. A zajról megmutatható, hogy a varianciája csak kvadratikusan nő J-vel. Optimista TD(1)-re szinkron frissítés esetén van konvergencia bizonyítás, és az ”initial-state-only” aszinkron esetre.

Q-learning - SSP problémák esetén

A Q-learning algoritmus az érték iteráció egy átalánosítása ”model-free” esetre, amikor csak szimulálni tudjuk a környezetet. Bevezetjük az akció-értékelő Q függvény fogalmát. Az optimális Q^* függvény definíciója

$$Q^*(i,u) = \sum_{j=1}^{n} p_{ij}(u)\Big(g(i,u,j) + J^*(j)\Big), \tag{7}$$

minden i-re és u-ra. Ekkor a Bellman egyenlet úgy is írható, hogy

$$J^*(i) = \min_{u \in U(i)} Q^*(i,u) \tag{8}$$

Kombinálva a (7) és a (8) egyenleteket, Q^*-ra azt kapjuk, hogy

$$Q^*(i,u) = \sum_{j=1}^{n} p_{ij}(u)\Big(g(i,u,j) + \min_{v \in U(j)} Q^*(j,v)\Big)$$

Q-learning - SSP problémák esetén

SSP problémáknál a célállapotra feltesszük, hogy $Q(0,u)=0$ minden u-ra.

Az érték iteráció algoritmus Q függvényekkel felírva

$$Q_{t+1}(i,u)=\sum_{j=1}^{n}p_{ij}(u)\Big(g(i,u,j)+\min_{v\in U(j)}Q_t(j,v)\Big)$$

Ennek egy általánosabb, γ lépésköz paraméterrel ellátott változata

$$Q_{t+1}(i,u)=\big(1-\gamma_t(i,u)\big)Q_t(i,u)+$$

$$+\gamma_t(i,u)\left(\sum_{j=1}^{n}p_{ij}(u)\Big(g(i,u,j)+\min_{v\in U(j)}Q_t(j,v)\Big)\right)$$

A Q-learning algoritmus ennek egy közelítő változata, amikor a j-re vett várható érték egyetlen szimulációval generált mintával van közelítve.

Q-learning - SSP problémák esetén

A Q-learning algoritmus formális definíciója

$$Q_{t+1}(i,u) = \big(1-\gamma_t(i,u)\big)Q_t(i,u) + \gamma_t(i,u)\Big(g(i,u,j) + \min_{v\in U(j)} Q_t(j,v)\Big),$$

ahol j véletlenszerűen van választva $p_{ij}(u)$ eloszlással, valamint a lépésközre teljesül, hogy $\gamma_t(i,u) = 0$ a nem aktuális állapot-akció párokra.

Észrevehető a kapcsolat (pl. a ”min” figyelmen kívül hagyásával) a Q-learning és a TD(0) között. A Q-learning-et is lehet általánosítani λ együtthatókkal, de ez nem egyértelmű, többféle Q(λ) kiterjesztés létezik.

Megjegyzés: a Q-learning egy ”off-policy” módszer, tehát az optimális értékelőfüggvényhez való konvergenciája nem függ a használt politikától.

Q-learning - konvergencia tétel

Állítás. Tegyük fel, hogy a lépésközre teljesülnek az alábbi feltételek

$$\sum_{i=0}^{\infty} \gamma_t(i,u) = \infty, \qquad \sum_{i=0}^{\infty} \gamma_t^2(i,u) < \infty, \qquad \forall i,u \in U(i)$$

Ekkor, minden i-re és $u \in U(i)$-ra $Q_t(i,u)$ egy valószínűséggel konvergál $Q^*(i,u)$-hoz feltéve, hogy valamelyik fennáll az alábbi esetek közül

(a) minden politika megfelelő (proper), azaz egy valószínűséggel eljut a 0-ba

(b) létezik legalább egy megfelelő politika és minden nem-megfelelő μ politikára van olyan i kezdeti állapot, amelyikben $J^\mu(i) = \infty$, továbbá $Q_t(i,u)$ egy valószínűséggel korlátos. (Ez teljesül, ha $g(i,u,j) \geq 0$ minden i, u, j-ra, valamint minden $\gamma_t(i,u) \leq 1$ és $Q_0(i,u) \geq 0$ minden i, u-ra.)

Q-learning - diszkontáltálás, felfedezés

A Q-learning algoritmus definíciója diszkontált esetben

$$Q_{t+1}(i,u) = \big(1-\gamma_t(i,u)\big)Q_t(i,u)+\gamma_t(i,u)\Big(g(i,u,j)+\alpha \min_{v\in U(j)} Q_t(j,v)\Big)$$

A konvergeciához szükséges feltételek ugyanazok, mint az SSP esetben.

A kihasználás/felfedezés probléma egy gyakori megoldása a Bolztmann (vagy Gibbs) formula használata (figyelem, ekkor a politika véletlenített lesz!)

$$\pi(i,u) = \frac{\exp(-Q(i,u)/\tau)}{\sum_{v\in U(i)} \exp(-Q(i,v)/\tau)},$$

ahol $\tau > 0$ az un. ”hőmérséklet” paraméter. Ahogy τ tart 0-hoz, a politika tart az aktuális Q akció-értéklelő függvényre mohó politikához. Ha azonban τ a ∞-be tart, a π tart az akciókat egyenletes eloszlással választó politikához.

Köszönöm a figyelmet!

http://www.sztaki.hu/∼csaji/td-tanulas.pdf