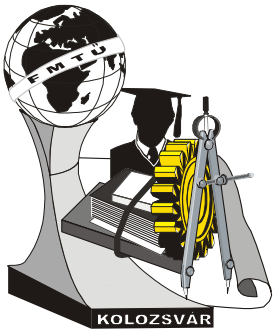
DOI: 10.36243/fmtu-2011.34



XVI. FIATAL MŰSZAKIAK TUDOMÁNYOS ÜLÉSSZAKA

Kolozsvár, 2011. március 24–25.

A LENGŐ ATWOOD GÉP VEZÉRLÉSE MEGERŐSÍTÉSES TANULÁSSAL

JAKAB Hunor Sándor

Abstract
Controlling dynamic systems using reinforcement learning is a challenging task. To facilitate testing and performance evaluation throughout the literature many toy problems have been introduced for example: pole balancing, inverted pendulum, park on the hill etc. In this paper we present a new toy problem setup for dynamic reinforcement learning control which is based on controlling the swinging Atwood's machine. To maintain stability and achieve optimal performance the control signal needs to be constrained and it needs to be drawn from a continuous interval. We present our approach to solving the learning problem with the help of approximated action-value functions and direct policy search algorithms.

Key words:
control, reinforcement learning, swinging Atwood's machine, dynamic systems

Összefoglalás
Dinamikus rendszerek megerősítéses tanulással történő vezérlése számos nehézséget rejt magában. A szakirodalomban erre a célra kidolgozott módszerek tesztelésére érdekében több klasszikus kísérleti feladatot is megfogalmaztak. Ilyenek a rúdegyensúlyozás, fordított inga, hegyre parkolás stb. Ezen dolgozatban egy új kísérleti feladatot vezetünk be amely a lengő Atwood gép vezérlésén alapszik. A szóban forgó rendszer stabilitásának megőrzése, valamint optimális teljesítmény elérése érdekében a vezérlési jel egy korlátolt de ugyanakkor folytonos tartományból kell származzon, ezért egy megszorításokkal rendelkező vezérlési feladatról van szó. Bemutatjuk a feladat megoldására kidolgozott saját megközelítésünket melyben közelített értékelőfüggvényt és egy gradiens alapú stratégia kereső algoritmust használunk.

Kulcsszavak:
irányítás, megerősítéses tanulás, lengő Atwood gép, dinamikus rendszerek

1. Vezérlés és megerősítéses tanulás

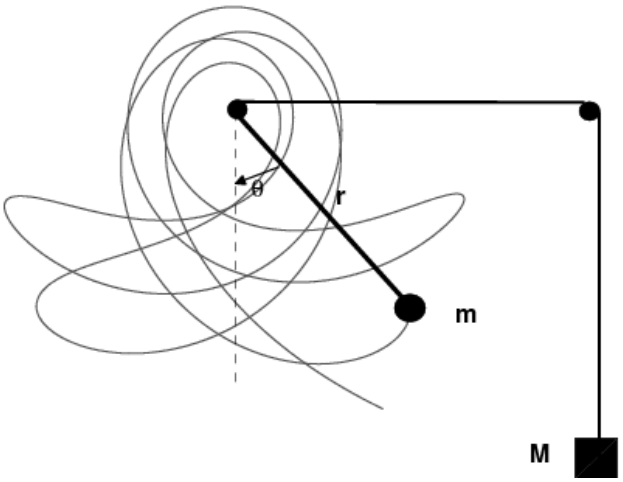
Dinamikus rendszerek vezérlése esetében a feladatot a következő képpen fogalmazhatjuk meg. Rendelkezésünkre áll egy rendszer melynek dinamikáját nem ismerjük. Ezen rendszer állapota egy adott időpontban leírható egy állapotvektor segítségével. A rendszerre külső hatást gyakorolhatunk egy vezérlési jeleket tartalmazó művelet vektor segítségével. Külső beavatkozás hatására a rendszer aktuális állapotából egy újabb állapotba kerül. Ugyanakkor rendelkezésünkre áll egy jutalom függvény amely kiértékeli az adott állapot előnyösségét az aktuális feladat végrehajtása szempontjából. Ezen elemekből álló feladatok matematikai leírását Markov Döntési folyamatokkal adhatjuk meg[3]: $M(S, A, P, R)$ amely a következő elemeket tartalmazza: S az állapotok halmaza; A a műveletek halmaza;

$P(s' \mid s,a) : S \times S \times A \to [0,1]$ az állapot átmeneti valószínűségek és $R : S \times A \to \mathsf{R}$, $R(s,a)$ a jutalom függvény. A Markov döntési folyamaton kívül a tanuló rendszerünk a következő elemekkel rendelkezik: *paraméterezett stratégia:* $\pi_\theta : S \times A \to \mathsf{R}$, amely egy valószínűségi eloszlás állapot-művelet téren, és minden $s \in S$ állapot és $a \in A$ művelet esetében visszatéríti az adott művelet választásának valószínűségét: $\pi(a \mid s, \theta)$. Ez tölti be tulajdonképpen a sztochasztikus vezérlő szerepét. A feladat megoldását a stratégia paramétereinek oly módon történő változtatásával érhetjük el, melyeknek segítségével maximalizáljuk a hosszú távú várható jutalmat:

$$\pi^* = \pi_{\theta^*} = arg\max_{\pi}\left(J^\pi\right), \quad \text{ahol} \qquad J^\pi = E_\pi\left(\sum_{t=0}^{\infty}\gamma^t r_{t+1}\right) \tag{1}$$

2. A lengő Atwood gép vezérlése

A lengő Atwood gép egy dinamikus rendszer amely két előre megadott tömeggel rendelkező testből és az ezeket összekötő huzalból valamint két csigából áll. Egyike a súlyoknak (m) tetszőlegesen kilenghet a kétdimenziós térben, a második súly (M) viszont csak függőleges irányban mozdulhat el. A rendszer gravitáció hatása alatt működik, szerkezete az 1. ábrán látható.



1 ábra *A lengő Atwood gép szerkezete. A négyzettel jelölt súly függőleges pályán mozdulhat el, a megjelenített görbe a szabadon elmozduló súly pályáját szemlélteti külső behatás hiányában.*

Az Atwood gép állapotát egy adott pillanatban egy állapot vektor jellemzi állapotvektor $s_t = [\theta \quad r \quad \dot{\theta} \quad \dot{r}]^T$ ahol θ a második súly kilengése által eredményezett függőlegessel bezárt szög, r a kötél hossza a második csigától számítva, az utolsó két elem pedig ezek időbeli változása (1. ábra). A rendszer energiájának leírása Hamiltoni dinakmikával [2] a következő képlettel adható meg:

$$H(r,t) = \frac{1}{2}M\dot{r}^2 + \frac{1}{2}\left(\dot{r}^2 + r\dot{\theta}^2\right) + gr\left(M - m\cos(\theta)\right) \tag{2}$$

ahol g a gravitációs együttható, a surlódást pedig figyelmen kívül hagytuk. Amint látható a lengő Atwood gép két szabadságfokkal és egy négy dimenziós fázistérrel rendelkezik. A vezérlési feladat célja olyan erőkkel hatni a függőleges irányban elmozduló súlyra a gép működése során, amelyek hatására a szabadon elmozduló súly egy konstans sugarú körhöz lehető legjobban közelítő pályát írjon le. Ennek érdekében definiálnunk kell az 1. részben említett jutalom függvényt, mely három különböző komponensből tevődik össze:

$$R(s_t, a_t) = \exp - \frac{|\dot{\theta} - \omega|}{2\sigma^2} - |r - R| - |\dot{r}| \tag{3}$$

A második és harmadik komponens az r változónak a kívánt sugár R-től való eltérését valamint az r időbeli változását penalizálja. Ideális esetben $r = R$ és $\dot{r} = 0$. Az első komponens esetében ω a kívánt tartandó szögsebesség, ezen komponens enyhébben befolyásolja a jutalomfüggvény végső értékét. Mivel a paraméterezett vezérlők tanítása esetében viszonylag sok kísérlet végrehajtására van szükség, ezért a rendszer szimulációját használjuk módszereink kiértékelésére. A lengő Atwood gép alapbeállítása esetén a tömegek aránya $\frac{M}{n} = 3$, ebben az esetben a rendszer integrálható ami fontos a számítógépes szimulációk stabilitása szempontjából, viszont ez a tulajdonság könnyen érvényét veszti amennyiben a külső behatások nagy mértékben felborítják a tömegek közti arányt. Ennek elkerülése érdekében a műveletvektor értéktartományát a $[-1,1]$ tartományra korlátozzuk.

3. A vezérlő tanítása

Mivel megerősítéses tanulást használva szeretnénk az előbbiekben leírt vezérlési feladat megoldására jó stratégiát kialakítani, ezért csupán a konkrét kisérletek végrehajtása során szerzett információkat használhatjuk fel információ forrásként. Ezen adatokból szeretnénk következtetéseket levonni, ennek érdekében gradiens alapú stratégia optimalizációs módszerekhez fogunk folyamodni. Ennek értelmében a stratégiánk θ *paramétereit* a következő sztochasztikus gradines szabály alapján változtatjuk:

$$\theta_{t+1} = \theta_t + \alpha \nabla_\theta J(\theta) \tag{4}$$

ahol J a (1)-ban definiált hosszú távú várható jutalom, alpha a tanulási együttható, $\nabla_\theta J(\theta)$ pedig a stratégia paramétereinek parciális deriváltjait tartalmazza a várható jutalom függvényében.

$$\nabla_\theta J(\theta) = E_\tau\left[\sum_{t=0}^{H-1} \nabla_\theta \log \pi(a_t \mid s_t) R(\tau)\right] \tag{5}$$

Látható, hogy a gradiens kifejezhető egy várható értékként ahol τ egy H lépésből álló kísérlet során bejárt állapotokat és az ezekhez tartozó azonnali jutalmakat tartalmazza $\tau = \{(s_1, a_1, R(a_1, s_1))(s_H, a_H, R(a_H, s_H))\}$. Szükségünk van tehát egy pontos és lehetőleg kis szórással rendelkező gradiens becslésre, melyenk segítségével lépésenként finomíthatjuk a stratégiánkat. Williams REINFORCE [4] módszere egyike az első olyan algoritmusoknak amely megoldást kínál erre a problémára. A gradienst olya módon becsüli, hogy felcseréli az (5) képletbeli $R(\tau)$ tagot az egyes állapotokból elérhető várható jutalmak Monte Carlo mintáival, amelyek pontos de zajos gradiens becslést eredményeznek. Ezen a módszeren nagy mértékben javítani lehet állapot-művelet téren értelmezett függvényközelítők használatával. Ez megvalósítható Gauss folyamatok segítségével is, ahogyan azt bemutattuk [1]-ben. A Gauss folyamat segítségével közelített állapot-művelet értékelőfüggvény használata csökkenti a becsült gradiens szórását, ezáltal gyorsabb konvergenciát eredményez. A módszer jól alkalmazható a lengő Atwood gép vezérlésére, hiszen lehetővé teszi a folytonos állapot-műveletterek lekezelését, ugyanakkor a stratégiánk paraméterein végrehajtott változtatás mértékét megfelelően alacsonyan tartja ahhoz, hogy a rendszer stabilitása megmaradjon.

Irodalom

[1] Hunor Jakab, Lehel Csató: *Using Gaussian processes for variance reduction in policy gradient algorithms*, In ICAI2010: *Proceedings of the 8th International Conference on Applied Informatics*, 2010

[2] Radford M. Neal.: *MCMC using Hamiltonian dynamics,* In Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng, editors: *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC Press, 2010

[3] Richard S. Sutton and Andrew G. Barto: *Reinforcement Learning: An Introduction,* MIT Press, 1998

[4] Ronald J. Williams: *Simple statistical gradient-following algorithms for connectionist reinforcement learning*, Journal of Machine Learning, 8:229-256, 1992

Jakab Hunor Sándor, doktorandus

Munkahely: Babes-Bolyai Tudományegyetem, Matematika Informatika Kar
Cím: 400084, Románia, Kolozsvár, Kogălniceanu, 1
Tel: +40-747-253 683
E-mail: jakabh@cs.ubbcluj.ro