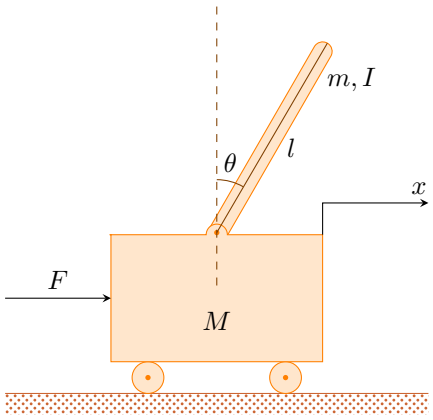
Završni domaći zadatak: Cart-pole

Samoobučavajći i adaptivni algoritmi

16. jun 2025.

Opis problema

Problem inverznog klatna (eng. *cart-pole problem*) često se koristi kao jedan od osnovih primera u učenju potkrepljenjem. To je šipka je pričvršćena nepokretnim zglobom za kolica koja se kreću pravolinijski duž pruge, bez trenja. Centar mase šipke nalazi se iznad njene pivot tačke. Cilj je održavati sistem balansiranim, i to primenom adekvatnih sila na sama kolica. Ovakav fizički sistem, zbog svoje nelinearne dinamike i nestabilne radne tačke, predstavlja izazov u domenu upravljanja. S tim u vezi, ideja ovog završnog domaćeg zadatka je da se primeni neka od metoda učenja potkrepljenjem za upravljanje ovakvim fizičkim sistemom. Shema ovako postavljenog problema data je na slici 1.



Slika 1: Prikaz modela inverznog klatna.

Inverzno klatno može da se modeluje sistemom nelinearnih jednačina

$$\begin{aligned} \ddot{\theta} &= \frac{Mg\sin\theta - \cos\theta(F + ml\dot{\theta}^2\sin\theta)}{(1+k)Ml - ml\cos^2\theta} \\ \ddot{x} &= \frac{mg\sin\theta\cos\theta - (1+k)(F + ml\dot{\theta}^2\sin\theta)}{m\cos^2\theta - (1+k)M}\,, \end{aligned} \tag{1}$$

Jednačine su preuzete sa `https://sharpneat.sourceforge.io/research/cart-pole/cart-pole-equations.html`, pri čemu je:

- m - masa štapa. $(100g)$
- M - ukupna masa štapa i kolica $(1.1kg)$
- l - dužina šipke $(0.5m)$
- g - gravitaciona konstanta (nema predložene vrednosti)
- θ - ugao između šipke i vertikalne ose ($\theta = 0°$ odgovara ravnotežnom stanju)
- x - pozicija
- F - sila $(F \in [-10N, 10N])$
- k - konstanta vezana za moment inercije štapa $(k = 0; k = 1/3; k = 1)$

U zagradi stoje predložene vrednosti koje možete iskoristiti pri implementaciji. U sažetoj formi, model možemo zapisati i u obliku

$$\begin{aligned} \ddot{\theta} &= f_\theta(x, \dot{x}, \theta, \dot{\theta}, F) \\ \ddot{x} &= f_x(x, \dot{x}, \theta, \dot{\theta}, F) \ , \end{aligned} \tag{2}$$

pri čemu su f_θ i f_x nelinearne funkcije.

Prevođenje matematičkog modela u MPO

Da bismo ovakvim modelom mogli da upravljamo nekim od metoda sa kojima smo se upoznali na ovom predmetu, potrebno je da ga prevedemo u pogodnu formu. Premda se Markovljevi procesi odlučivanja mogu uopštiti i na slučajeve kada vreme teče kontinualno, mi se time ovde nećemo baviti. Umesto toga, dati model ćemo prevesti u formu koja je ekvivalentna Markovljevom procesu odlučivanja: naredno stanje i nagrada su funkcije tekućeg stanja i primenjene akcije, odnosno

$$\begin{aligned} s^+ &= f(s, a) \\ r &= h(s, a) \ . \end{aligned} \tag{3}$$

Iako to možda nije očigledno, prvi korak je da snizimo red sistema koji je dat jednačinama 2. Uvođenjem smena

$$\begin{aligned} z_1 &= x \\ z_2 &= \dot{x} \\ z_3 &= \theta \\ z_4 &= \dot{\theta} \ , \end{aligned} \tag{4}$$

i diferenciranjem po vremenu leve strane svake jednačine, dobijamo matematički model u prostoru stanja u obliku

$$\begin{aligned} \dot{z}_1 &= z_2 \\ \dot{z}_2 &= f_x(\underline{z}, F) \\ \dot{z}_3 &= z_4 \\ \dot{z}_4 &= f_\theta(\underline{z}, F) \ , \end{aligned} \tag{5}$$

pri čemu je $\underline{z} = (z_1, z_2, z_3, z_4)$. Drugim rečima, u $\underline{z}$ su pobrojane sve promenljive stanja. Drugi korak je diskretizacija. Za potrebe ovog zadatka, izabraćemo *Euler* 1 diskretizaciju, tj. numeričku integraciju levim pravougaonicima, jer za nelinearne modele nekada prosto nije moguće primeniti ništa drugo. Kada izraz za aproksimiranje prvog izvoda po vremenu (diferenciranje unapred), uvrstimo u 5, dobijamo

$$\begin{aligned} z_1^+ &= z_1 + Tz_2 \\ z_2^+ &= z_2 + Tf_x(\underline{z}, F) \\ z_3^+ &= z_3 + Tz_4 \\ z_4^+ &= z_4 + Tf_\theta(\underline{z}, F) \ . \end{aligned} \tag{6}$$

Fizički sistem je sada preveden u formu u MPO, sa detaljem da sada imamo vektor stanja $\underline{z} = (z_1, z_2, z_3, z_4)$. S druge strane akcija a je ništa drugo nego sila F koja se u datom trenutku vremena primenjuje na kolica. Akcija je takođe diskretna i ograničena veličina. Vreme odabiranja je T.

Formiranje nagrade

Neizostavno je i pitanje načina formiranja nagrade, odnosno funkcije h u jednačini $r = h(s, a)$ determinističkog MPO. U ovom delu nudimo nekoliko predloga za način formiranja funkcije nagrade:

1. Ukoliko odaberemo graničnu vrednost ugla θ_{th} i faktor $\alpha > 0$, funkciju nagrade moguće je formirati kao

$$r = \begin{cases} -|\theta|^\alpha & , |\theta| < \theta_{th} \\ -1000 & , |\theta| \geq \theta_{th} \end{cases} \ . \tag{7}$$

   Ideja je da se svako stanje pri kojem je vrednost ugla izvan predefinisanog opsega značajno "kažnjava". U suprotnom, nagrada raste sa opadanjem apsolutne vrednost ugla.

2. Alternatvno, može se izabrati dovoljno mala vrednost δ, a funkcija nagrade definisati u obliku

$$r = \begin{cases} 0 & , \theta \in [-\delta, \delta]^\circ \\ -100 & , \text{inače} \end{cases} \ , \tag{8}$$

U ovom slučaju, agent se nagrađuje samo ukoliko je vrednost ugla u bliskoj okolini ravnotežne vrednosti.

3. Treći predlog je dat po uzoru na predefinisano *Gymnasium* okruženje za inverzno klatno. S obzirom na cilj zadatka, nagrada sa vrednošću +1 dodeljuje se za svaki korak simulacije. Što više koraka u epizodi protekne pre nastupanja terminalnog stanja, nagrada za tu epizodu biće veća.

Na sledećim hiperlinkovima možete pronaći opšta uputstva za korišćenje *Gymnasium* okruženja, treniranje agenata, kao i za korišćenje samog *cart-pole* okruženja.

- Korišćenje okruženja
- Primer treniranja agenta
- Cart-pole okruženje

Napominjemo da korišćenje predefinisanog okruženja nije obavezno, ovde ga samo predlažemo kao opciju. Iako bismo u idealnom slučaju izuzetno cenili da sami implementirate okruženje i rešenje, *Gymnasium* nudimo kao (potencijalno) olakšavajuću okolnost. S tim u vezi, odluka o tome da li ga koristite neće uticati na broj ostvarenih bodova na ovom završnom zadatku.

Zadaci za samostalan rad

1. Izvršiti diskretizaciju matematičkog modela u prostoru stanja. u slučaju da koristite *Gymnasium*, ovaj korak nije potreban.
2. Definisati prostor akcija. Napominjemo da prostor akcija treba da bude diskretan i ograničen. U slučaju da koristite *Gymnasium*, prostor akcija je već predefinisan.
3. Izabrati način formiranja nagrade. Nije obavezno da izaberete neki od datih predloga, slobodno osmislite originalan način formiranja nagrade.
4. Trenirati agenta pomoću Q-Learning ili SARSA.
5. Simulirati okruženje i prikazati dobijene rezultate. Ukoliko se ne služite *Gymnasium* okruženjem, nije obavezno da pravite vizualizaciju.