0.1 Razširitve modela

Potencialne nadaljne raziskave in razširitve modela:

0.1.1 Izognitveno obnašanje (*angl. aversive behaviour*)

V večini del, ki se ukvarjajo s spodbujevanim učenjem se izognitev stanj, za katere želimo, da se jih agent izogiba, doseže s pomočjo negativne nagrade. Negativna nagrada tako v enačbi sinapse RSTDP obrne predznak posodobitve in so tako sinapse, ki so odgovorne za vstop v neželeno stanje najbolj negativno posodobljene. V človeških možganih negativnega dopamina ni. Porodi se ideja, da je izogibanje negativnim stanjem prav tako posledica učenja, kjer je nivo dopamina > 0. Dopamin namreč predstavlja učenje, ne nujno nagrade. Negativna nagrada je predstavljena s posebnim vhodom, ki prestavlja abstraktno "bolečino", ki jo tekom učenja želimo zmanjšati. Tako lahko prav tako uporabimo načela STDP in TD učenja, kjer zmanjšanje nivoja bolečine prestavlja nagrado. Trenutnemu akter-kritik sistemu bi dodali še eno kopijo kritika, ki računa temporalno razliko nivoja bolečine in deluje na dopaminergične nevrone, ki so skupni obema kritikoma. Ob prehodu iz stanja z visokim nivojem bolečine v stanje z nizkim dopaminergične nevrone ekscitiramo, v obratnem primeru inhibiramo, v primeru enakega nivoja dovedene bolečine pa kritik negativne nagrade ne vpliva na dopaminergične nevrone. Sistem se v tem primeru obnaša kratkovidno, kljub računanju temporalne razlike. Nagrajene bodo samo povezave, ki so nas vodile stran od bolečine, ker pa je vhod bolčine vedno odvisen samo od zunanje stimulacije, za razliko od striatuma, ki nagrado napoveduje sam, bodo nagrajene povezave samo v stanja neposredno ob negativnem stanju. Če želimo okrog negativnega stanja negativno označiti tudi stanja, ki nas potencialno vodijo vanj, pa moramo v sistem dodati skupino nevronov, ki stanja asociirajo z bolečinskim vhodom in jo tako napovedujejo.
Pričakujemo, da tako oba kritika med seboj tekmujeta za nagrajevanje akcij, ki vodijo bližje nagradi in sinaps, ki vodijo stran od negativnega stanja.

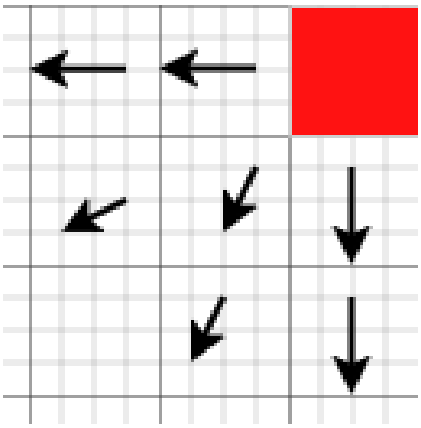

Figure 1: Pričakovana politika ob kritiku negativnih stanj (brez kritika nagrajenih stanj)

0.1.2 Rekurenčne povezave

Velika predpostavka trenutnega sistema je ta, da rekurenčnih povezav ni. Tako so stanja časovno med seboj skoraj popolnoma neodvisna (med prehodi stanj se sinapse še vedno lahko križno asociirajo, zaradi česar smo v našem modelu uvedli 50ms pavzo stimulacije pred prehodom v naslednje stanje. To je opisano v poglavju **TD učenje in model actor-critic**). V kolikor dodamo več vmesnih nivojev in rekurenčne povezave pa bodo stanja med seboj časovno odvisna. Pravzaprav stanja ne moremo več definirati samo z aktivnostjo vhodnih nevronov, saj v vsakem trenutku stanje vsebuje tudi vplive rekurenčnih povezav, ki nosijo informacijo iz stanj arbitrarno v preteklost. V primeru našega akter-kritik sistema bi tako v vsakem trenutku t kritik računal temoralno razliko med dvema neskončno kratkima stanjema s_t in s_{t-d}, kjer je d zakasnitev direktne povezave. Kljub temu pričakujemo, da rezultat ne bi bil drugačen, saj je to samo miselna prilagoditev. 200ms stimulacija, ki je do zdaj predstavljala stanje, bi vseeno v tem intervalu vodila do nevronske aktivnosti, ki je v glavnem pogojena s to vhodno stimulacijo in bi tako trenutki (oziroma neskončno kratka stanja) vseeno bili v naboru trenutkov značilnih za trenutno vhodno stimulacijo.
V eksperimentih izvedenih do zdaj, pravilna akcija določenega stanja ni bila odvisna od akcij, ki so nas privedle v to stanje oziroma zgodovine stanj. V primeru sprehajanja po mreži bomo končno stanje nagradili neglede na to iz katerega stanja vstopimo v nagrajeno stanje. Rekurenčne povezave bi tako predvideno predstavljale prednost pri nalogah, kjer je zgodovina stanj pomembna, oziroma kjer je nagrada stanja odvisna od prejšnjih stanj. Če bi

v primeru sprehajanja po mreži premik v končno stanje iz stanja nad njim pripeljalo do nagrade, prehod iz stanja levo pa ne, bi lahko tako končno stanje obravnavali kot dva različna stanja, glede na prehod. Zanimivo bi bilo realna stanja neke naloge tako razviti v drevo abstraktnih stanj in označiti pričakovane nagrade in preferirane akcije, ki jih sistem napoveduje. Sistem je še vedno stimuliran glede na realna stanja naloge, vendar bi s pomočjo rekurenčnih povezav interno predstavljal nabor abstraktnih stanj.
Rekurenčne povezave privedejo tudi do določenih situacij, kjer bi bila potrebna redefinicija trenutnega načina izbire stanj. Dva izhodna nevrona sta namreč lahko povezana med sabo in se bosta vedno prožila skupaj. V tem primeru moramo v sistemu dopuščati izbiro večih akcij hkrati in takšno situacijo na primer "kaznovati".

0.2 Nevronska vezja

1. Modeliranje človeških dopaminskih mehanizmov, dopaminskih receptorjev in odzivov na uspešno izvedene akcije.

   - **Trajanje: 2 tedna**
   - Predpogoji: aktivnosti poglavja **??** in 7.1-7.4.

2. Nov pristop k implementaciji nevronskih vezij.

   - **Trajanje: 4 tedni**
   - Predpogoji: aktivnost 7.5:1.