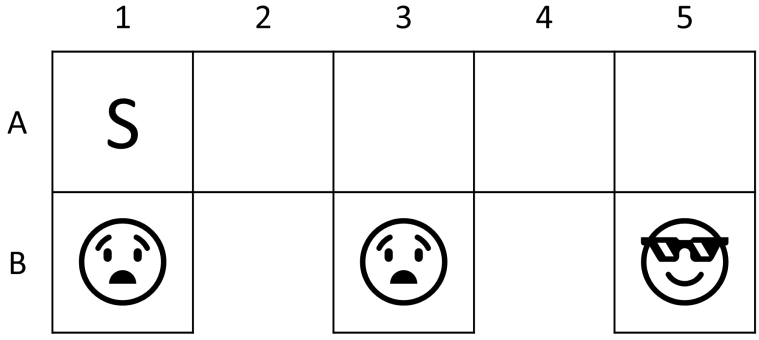
# 13M051MU, 5. domaći zadatak 2025/26
# Učenje podsticanjem

Treba da implementirate:
1) simulator,
2) algoritam Q-učenja,
3) algoritam REINFORCE,

za jednostavno okruženje, prema uputstvima datim u nastavku.

## Okruženje

Agent živi u jednostavnom okruženju prikazanom na slici 1.




Slika 1: Okruženje.

Početno stanje je A1 (označeno sa S na slici 1). Stanja B1, B3 i B5 su terminalna; kada se agent nađe u bilo kom od njih, kakvu god akciju da

preduzme, dobija nagradu $-1$ (u stanjima B1 i B3) ili $+3$ (u stanju B5), nakon čega se epizoda završava i agent se vraća u početno stanje. U svim ostalim slučajevima, agent ne dobija nikakvu nagradu. Smatrajte da je okruženje potpuno opservabilno i da agent uvek zna u kom stanju se nalazi.

U svakom od 8 stanja agent ima na raspolaganju 4 akcije: gore, dole, levo, desno. Okruženje je stohastično, u sledećem smislu: kada agent odabere akciju, pomera se u izabranom smeru sa verovatnoćom 0.7; u protivnom se pomera u nekom od preostala 3 smera, sa podjednakim verovatnoćama. Ukoliko udari u zid, ostaje u istom stanju. Na primer, ako izabere akciju "desno" u početnom stanju, sa verovatnoćom 0.7 prelazi u polje A2, sa verovatnoćom 0.2 udara u zid i ostaje u polju A1 i sa verovatnoćom 0.1 prelazi u polje B1.

## Simulator

Simulatoru se zadaje akcija na osnovu koje on ažurira interno stanje u skladu sa prethodnim opisom okruženja, i vraća:

- osvojenu nagradu u tom koraku,
- novo stanje,
- informaciju o tome da li je epizoda završena.

## Q-učenje

Faktor umanjenja budućih nagrada je $\gamma = 0.9$, ako drugačije nije rečeno. Koristite $\epsilon$-gramzivo istraživanje. Adekvatnu vrednost za $\epsilon$ odredite kroz eksperimente. Teorijski, stopa učenja bi trebalo da se smanjuje tokom vremena, ali ne prebrzo. Isprobajte strategiju koja uzima $\alpha_e = \ln(e+1)/(e+1)$, gde je $e$ redni broj epizode. U praksi se često usvaja i konstantna stopa učenja (ista u svim epizodama). Kroz eksperimente probajte da pronađete neku adekvatnu vrednost za konstantnu stopu učenja, pa uporedite brzinu konvergencije sa prethodnim slučajem kada je stopa promenljiva.

Da biste pratili kako učenje napreduje i da li je konvergiralo, prikažite kako se sa iteracijama $t$ menjaju V-vrednosti $V_t(s) = \max_a Q_t(s, a)$. Testirajte konačnu naučenu politiku kroz 10 epizoda interakcije sa okruženjem i izračunajte prosečnu ukupnu nagradu koju agent osvaja tokom jedne epizode

prateći ovu politiku. Zatim ponovite ovo za $\gamma = 0.999$. Ima li razlika u odnosu na slučaj sa $\gamma = 0.9$? Kako ih tumačite?

## REINFORCE

Pratite kako učenje napreduje tako što ćete povremeno "zamrzavati" do tada naučenu politiku, ponavljati 10 epizoda interakcije agenta sa okruženjem, i računati prosečnu ukupnu nagradu koju agent osvaja tokom jedne epizode. Grafički prikažite kako se tokom učenja menjaju

- pomenuta nagrada u 10 uzastopnih epizoda,
- vrednosti parametara politike u neterminalnim stanjima.

Eksperimentišite sa stopama učenja, kao kod Q-učenja. Usvojite $\gamma = 0.9$.

## Napomene

Python kôd i izveštaj sa traženim graficima u pdf ili html formatu predaje se putem MS Teams-a. U izveštaju navedite ili grafički prikažite konačnu politiku koju su naučili algoritmi. *Ne zaboravite da kliknete na Turn In!*