exploration
vs.
exploitation
Markov decision process
MDP
i
S_t
$\mathcal{S}$
S_t
$A_t \in \mathcal{A}(S_t)$
S_{t+1}
R_{t+1}
$\mathcal{A}(s)$
s
A
$\mathcal{A}$
$\mathcal{A}: \mathcal{S} \to P(A)$[1]
$\mathcal{R}$
$S_0, A_0, R_1, S_1, A_1, R_2, S_2, ...$
p

$$p(s', r|s, a)P(S_{t+1} = s', R_{t+1} = r|S_t = s, A_t = a) \tag{1}$$

s'
r
s
a
p

$$\sum_{s',r} p(s', r|s, a) = 1 \tag{2}$$

s
$a \in \mathcal{A}(s)$
$r \in \mathcal{R}$
$s' \in \mathcal{S}$

$$P(S_t, R_t|S_0, A_0, R_1, ..., S_{t-1}, A_{t-1}) = P(S_t, R_t|S_{t-1}, A_{t-1}) \tag{3}$$

t

$$G_t = \sum_{i=0}^{\infty} \gamma^i R_{t+i+1} \tag{4}$$

γ
$\gamma = 0$[2]
γ
1
$t+1$

$$G_t = \sum_{i=0}^{\infty} \gamma^i R_{t+i+1} = R_{t+1} + \gamma \sum_{i=0}^{\infty} \gamma^i R_{(t+1)+i+1} = R_{t+1} + \gamma G_{t+1} \tag{5}$$

$(\mathcal{S}, \mathcal{A}, \mathcal{R}, p, \gamma)$
$\mathcal{S}$
$\mathcal{R}$
$\mathcal{A}(s)$
s
$\mathcal{A}(s)$
s
A
$_s$*umapretpostavljasedanizinterakcijasaokruenjem, tj.putanja, trajebeskonano, tosevidiizgornjegraniceusumi.Meuti*
1
$\gamma <$
1
R_t
$_s$*umakonvergirati.Ukolikotasumadivergira, odnosnoakojenjenavrednost*∞

$P(A)$ označava partitivni skup skupa A.
0^0 definiše