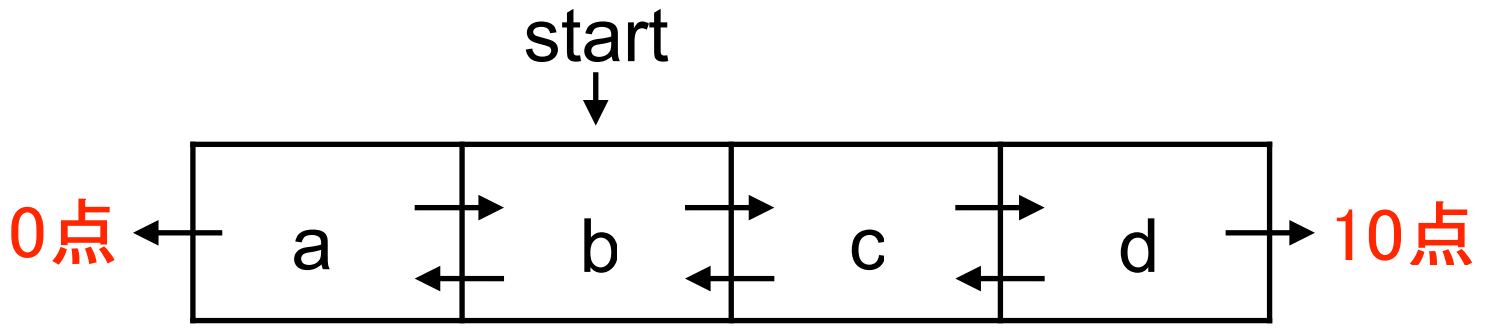
Q学習

総報酬を最大にする行動を学習する!

Q学習法（Watkins, 1989）



行動価値関数 $Q(s,a)$

状態 s_t=sで行動 a_t=a をとった時の総報酬 R_t の期待値

$$Q(s,a) = E_\pi\{R_t | s_t = s, a_t = a\}$$

ポリシーπに基づいて行動した時の期待値

$Q(s,a)$ がわかれば望ましいactionは

$$a = \boxed{?}\ Q(s,a')$$

$\arg\max_{a'} Q(s,a')$

Q学習法（Watkins, 1989）

$$Q(s,a) \leftarrow Q(s,a) + \alpha\{r + \gamma \max_{a'} Q(s',a') - Q(s,a)\}$$

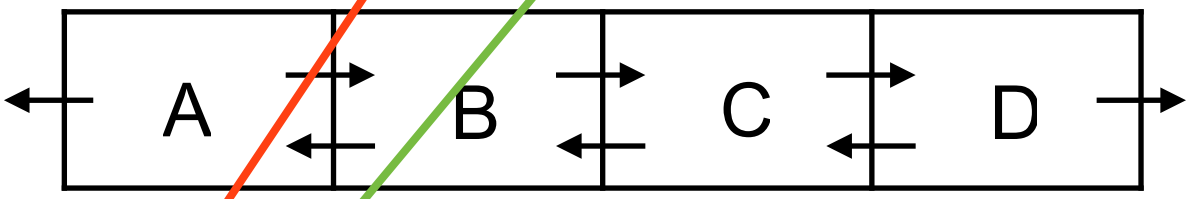
(s,a) の後に得られた状態

問: この式の意味は?

Q学習法（Watkins, 1989）

$$Q(s,a) \leftarrow Q(s,a) + \alpha\{r + \gamma \max_{a'} Q(s',a') - Q(s,a)\}$$

実行例

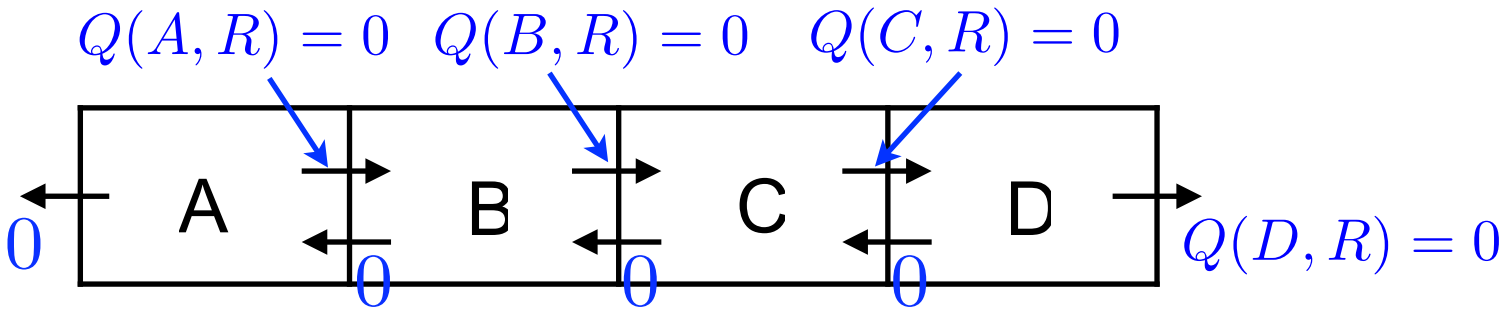


1) α, γ を決める

今回は $\alpha = 0.5, \gamma = 0.5$ にする

Q学習法（Watkins, 1989）

$$Q(s,a) \leftarrow Q(s,a) + \alpha\{r + \gamma \max_{a'} Q(s',a') - Q(s,a)\}$$

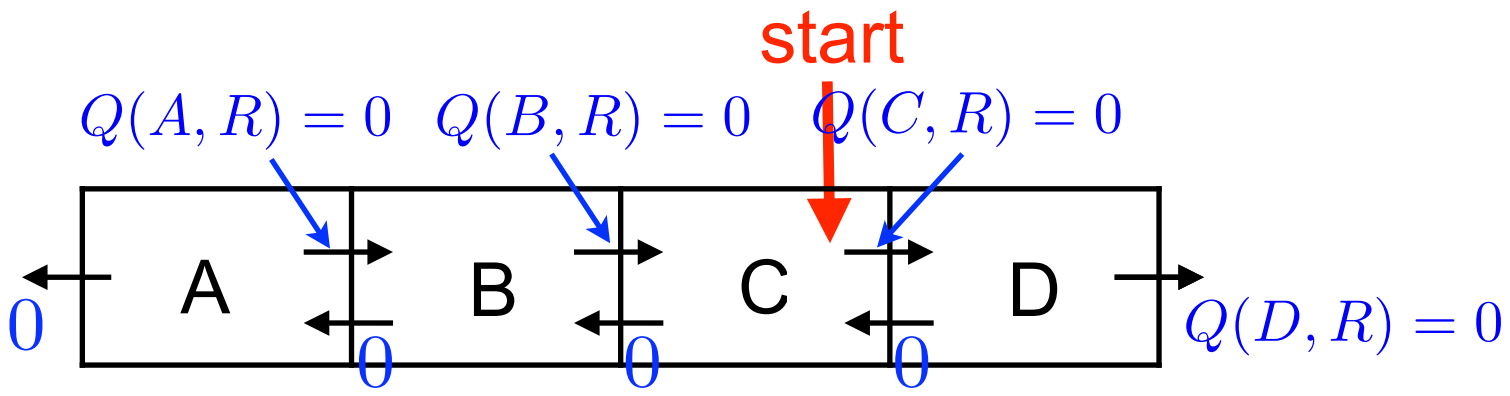


2) 価値$Q(s,a)$の初期値

ex) $Q(s,a) = 0$ for all s, a

Q学習法（Watkins, 1989）

$$Q(s,a) \leftarrow Q(s,a) + \alpha\{r + \gamma \max_{a'} Q(s',a') - Q(s,a)\}$$

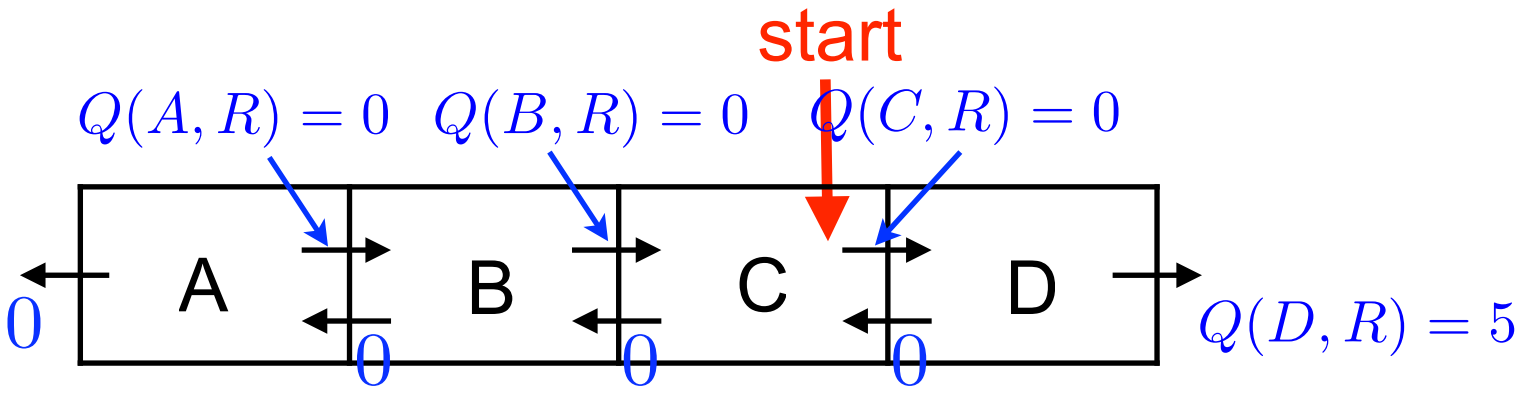


3) policyを決めて探索 … 今回はまずrandom policyにする

問1) C → D と進んだらQ(C,R)はどのように更新されるか?
問2) さらにD → 外 と進んだらQ(D,R)はどのように更新されるか?

Q学習法（Watkins, 1989）

$$Q(s,a) \leftarrow Q(s,a) + \alpha\{r + \gamma \max_{a'} Q(s',a') - Q(s,a)\}$$

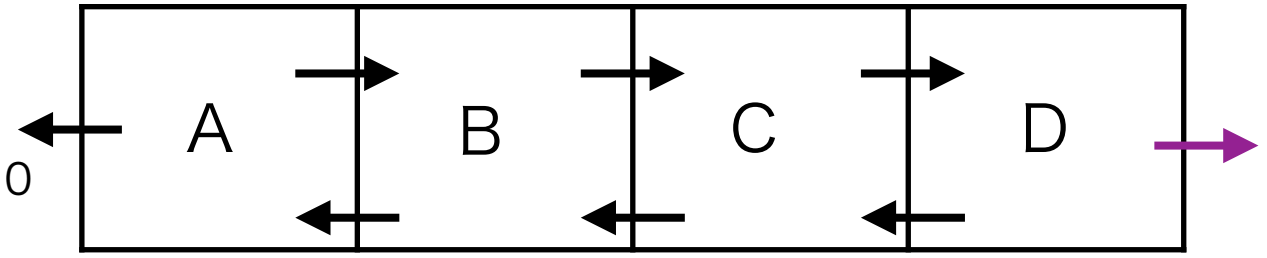


問3) 次もC → D と進んだらQ(C,R)はどのように更新される?

問4) さらにD → 外 と進んだらQ(D,R)はどのように更新される?

Q学習法（Watkins, 1989）

問1) 無限回学習した時の各Q値を予測しなさい
問2) ポリシーをgreedy法にした場合, 上記で得られたQ値に従うとどのような行動選択を行うことになるか?



ヒント)十分学習が進めば $Q(s,a)$ の変化はなくなる!

$$Q(s,a) = Q(s,a) + \alpha\{r + \gamma \max_{a'} Q(s',a') - Q(s,a)\}$$

$$\therefore \; Q(s,a) = r + \gamma \max_{a'} Q(s',a')$$

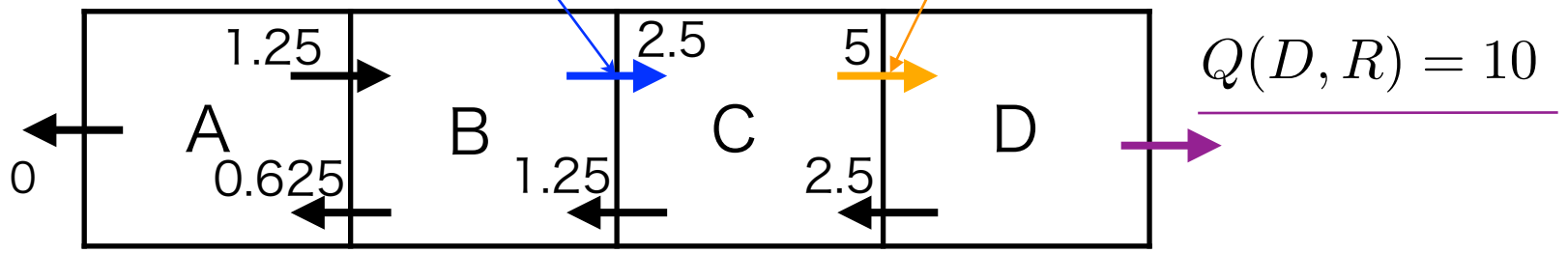
誤差項→0

Q学習法（Watkins, 1989）

1) 無限回学習した時の各Q値

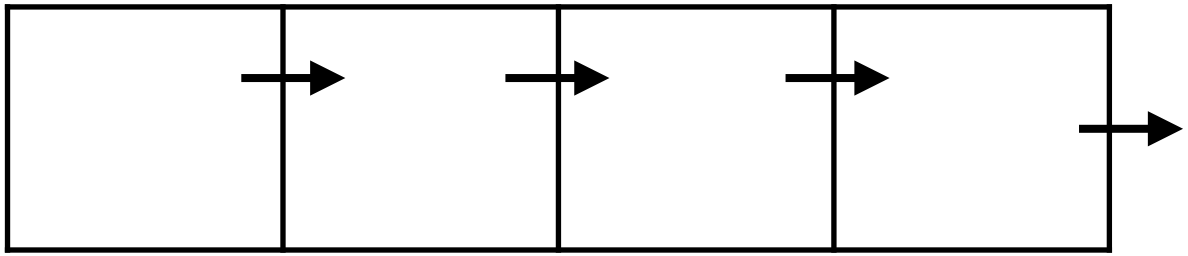
$$Q(s,a) = r + \gamma \max_{a'} Q(s',a')$$

$Q(B,R) = 0 + \gamma Q(C,R) = 2.5$　　$Q(C,R) = 0 + \gamma Q(D,R) = 5$



$Q(D,R) = 10$

2）ポリシーをgreedy法にした場合の行動選択



explorationからexploitationへ

Q学習法（Watkins, 1989）

理論値は求まらないことが多い。

その場合は学習で求める。