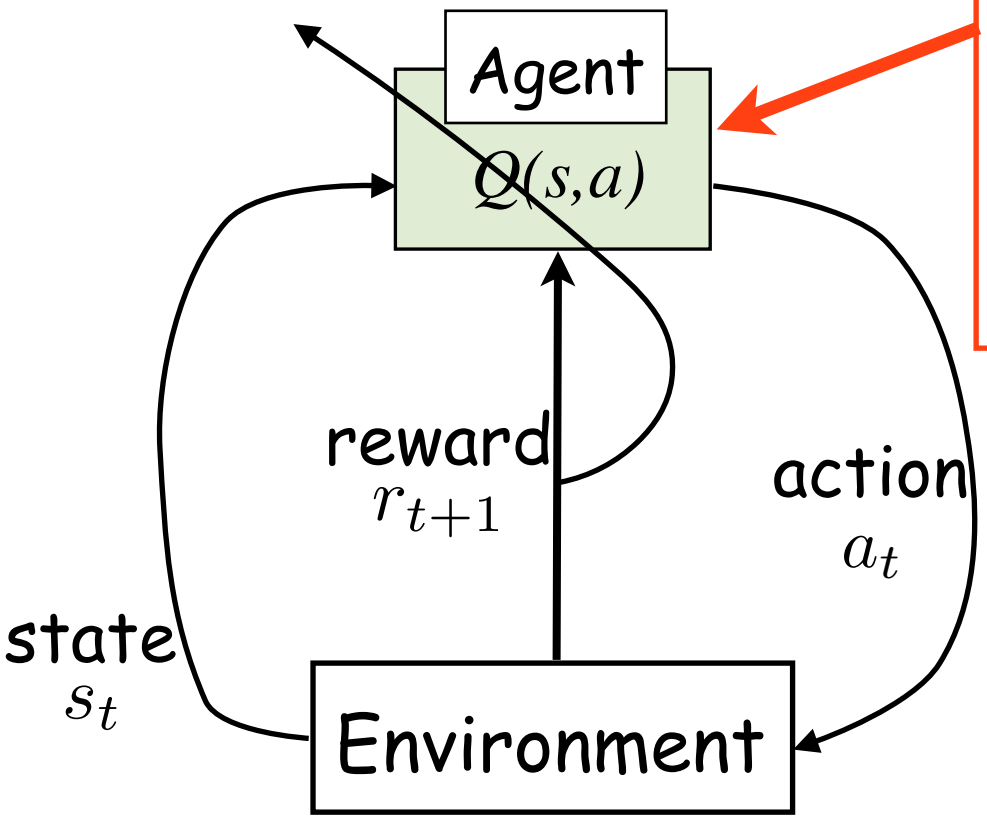
強化学習

Actor-Critic法

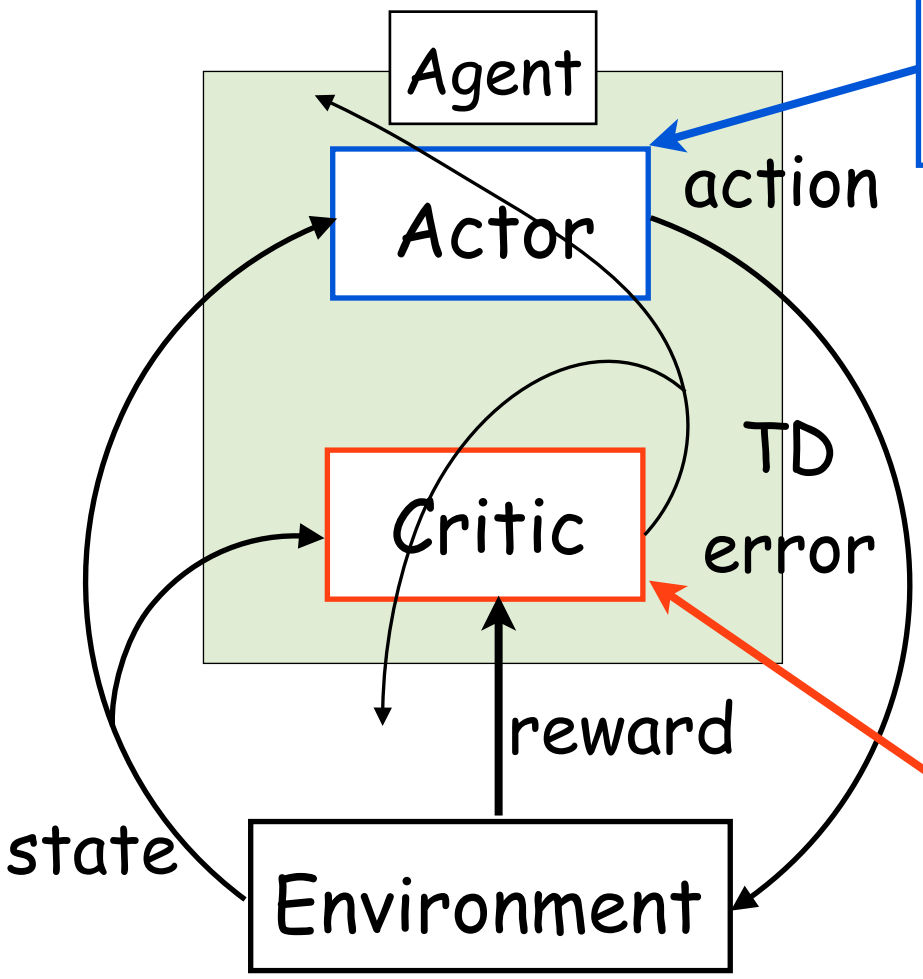
Q学習法



Agent:
状態sで行動aをとったときの、
その後の総報酬の期待値を表す
状態行動価値関数$Q(s,a)$を学習

問題点:
状態数や行動の選択肢が多いと
学習時間が非常にかかる

Actor-Critic法

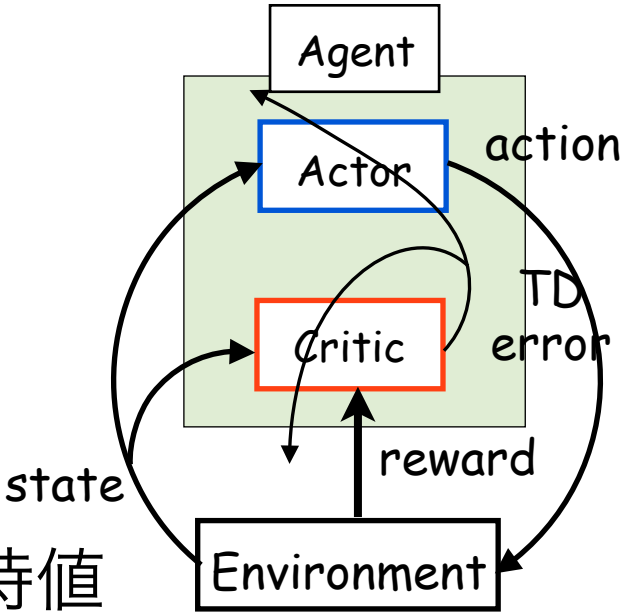


Actor:
状態に応じた行動を学習・選択

Critic:

- **状態価値関数** $V(s)$ (状態 s の後に得られる総報酬の期待値)を学習
- $V(s)$ により現在の状態の良し悪しを判断し，Actorのとった行動の結果を評価

Criticの学習

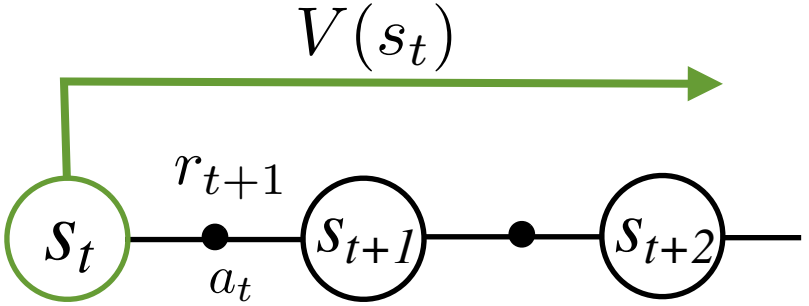


状態価値関数 $V(s)$：

状態 s になった後に得られる総報酬の期待値

$$V(s) = E\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\}$$

状態行動価値関数$Q(s,a)$よりも探索空間が狭いので学習しやすい!



Criticの学習

状態価値関数 $V(s)$の学習則：

$$V(s_t) \leftarrow V(s_t) + \alpha\delta_t \quad (\ \alpha\text{: 学習定数})$$

TD誤差：$V(s)$の精度を示す誤差

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$$

問: この式の意味は?

Actor-Critic法

TD誤差: $\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$

Actor: 状態に応じた行動を学習・選択

ex) softmax法をポリシーとする例

p: (s, a)の良さを表す関数

$$\pi(s,a) = \frac{e^{p(s,a)}}{\sum_{b \in A} e^{p(s,b)}}$$

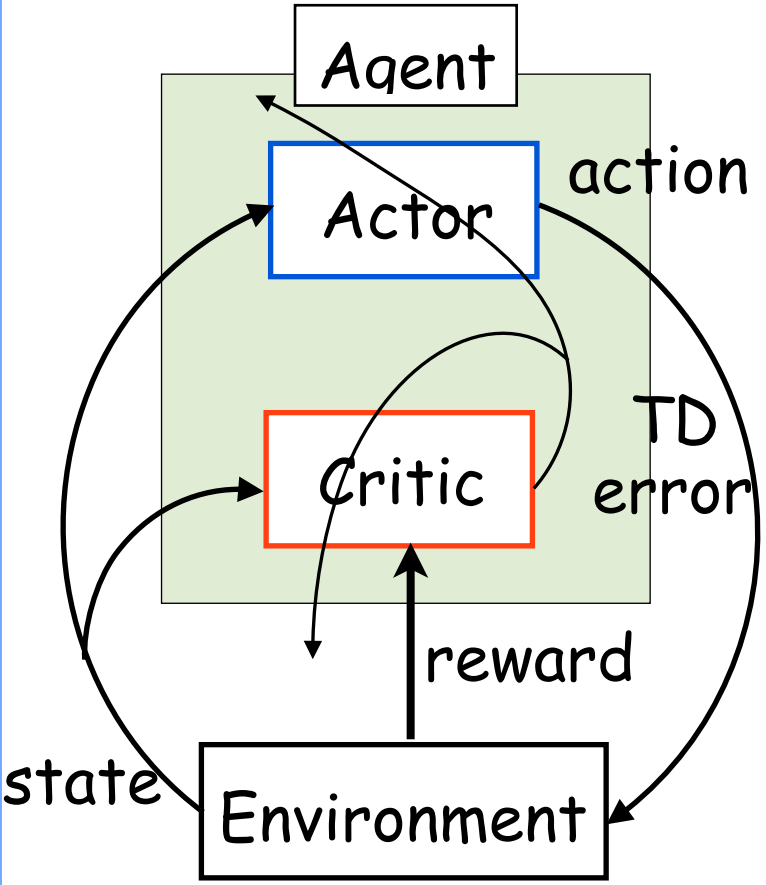
π : 状態sで行動aをとる確率

A: とりうる行動の集合

行動の学習

$$p(s_t, a_t) \leftarrow p(s_t, a_t) + \beta\delta_t$$

(β: 学習定数)



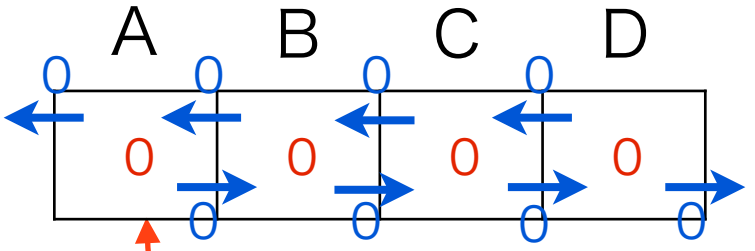
Actor-Critic法の学習例

学習則

Critic: $\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$ … TD誤差

$V(s_t) \leftarrow V(s_t) + \alpha\delta_t$ … 状態価値関数

Actor: $p(s_t, a_t) \leftarrow p(s_t, a_t) + \beta\delta_t$



$s \in \{A, B, C, D\}$

$a \in \{R, L\}$

どの部屋がどの位良いかを学習！
(Q学習と同じ例題)

(i) α, β, γ 決定 … ex) $\alpha = 0.5, \beta = 0.5, \gamma = 0.4$

(ii) $V(s)$を初期化 … ex) $V(s)=0$ for all s

(iii) $p(s,a)$を初期化 … ex) $p(s,a)=0$ for all s, a

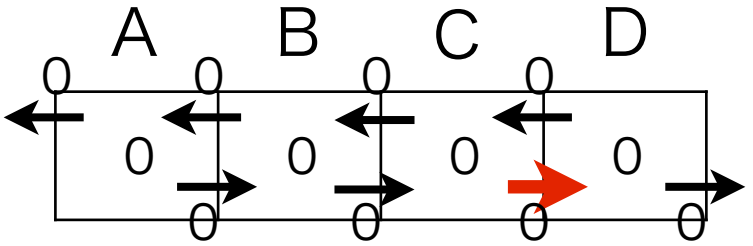
Actor-Critic法の学習例

学習則

Critic: $\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$ … TD誤差

$V(s_t) \leftarrow V(s_t) + \alpha\delta_t$ … 状態価値関数

Actor: $p(s_t, a_t) \leftarrow p(s_t, a_t) + \beta\delta_t$ $(\alpha = 0.5, \beta = 0.5, \gamma = 0.4)$



(iv)　Actorの決定に従って行動！

問: 以下のsoftmax法に従うとCから右(R方向)に進む確率は？

$$\pi(s,a) = \frac{e^{p(s,a)}}{\sum_{b \in A} e^{p(s,b)}}$$

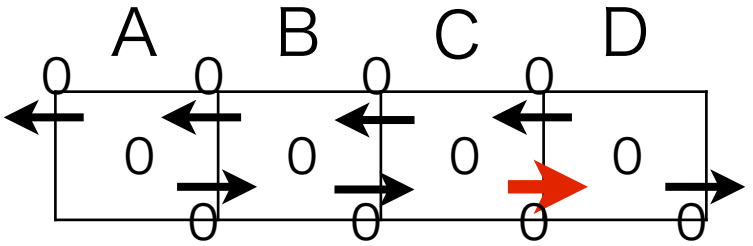
Actor-Critic法の学習例

学習則

Critic: $\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$ … TD誤差

$V(s_t) \leftarrow V(s_t) + \alpha\delta_t$ … 状態価値関数

Actor: $p(s_t, a_t) \leftarrow p(s_t, a_t) + \beta\delta_t$ $(\alpha = 0.5, \beta = 0.5, \gamma = 0.4)$



(iv)　Actorの決定に従って行動！

①　エージェントがCから右(*R*方向)に進んだとする

問1: ＴＤ誤差を求めなさい

問2: $V(C)$, $p(C,R)$を更新しなさい

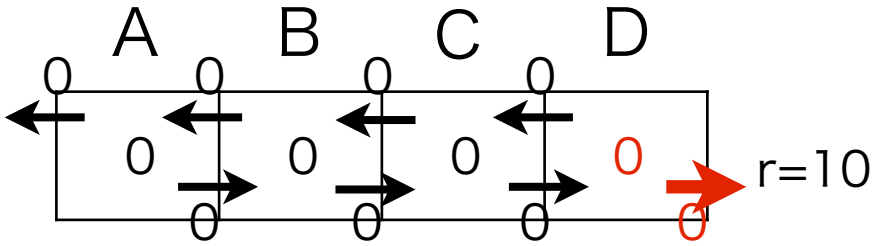
Actor-Critic法の学習例

学習則

Critic: $\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$ … TD誤差

$V(s_t) \leftarrow V(s_t) + \alpha\delta_t$ … 状態価値関数

Actor: $p(s_t, a_t) \leftarrow p(s_t, a_t) + \beta\delta_t$ $(\alpha = 0.5, \beta = 0.5, \gamma = 0.4)$



② エージェントがDから右(*R*)に進んで外に出た!

問1: ＴＤ誤差を求めなさい

問2: *V(D)*, *p(D,R)*を更新しなさい

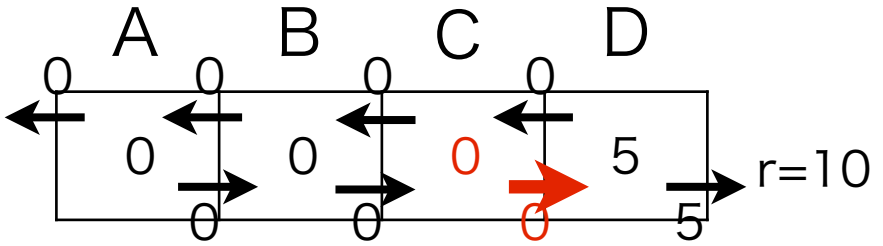
Actor-Critic法の学習例

学習則

Critic: $\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$ … TD誤差

$V(s_t) \leftarrow V(s_t) + \alpha\delta_t$ … 状態価値関数

Actor: $p(s_t, a_t) \leftarrow p(s_t, a_t) + \beta\delta_t$ $(\alpha = 0.5, \beta = 0.5, \gamma = 0.4)$



③ エージェントが再びCからスタートし、右(*R*方向)に進んだとする

問1: ＴＤ誤差を求めなさい

問2: *V(C)*, *p(C,R)*を更新しなさい

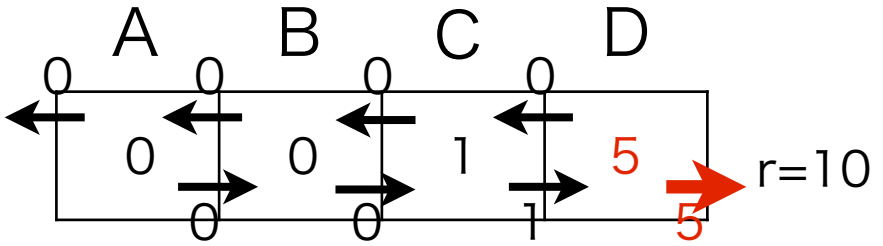
Actor-Critic法の学習例

学習則

Critic: $\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$ … TD誤差

$V(s_t) \leftarrow V(s_t) + \alpha\delta_t$ … 状態価値関数

Actor: $p(s_t, a_t) \leftarrow p(s_t, a_t) + \beta\delta_t$ $(\alpha = 0.5, \beta = 0.5, \gamma = 0.4)$



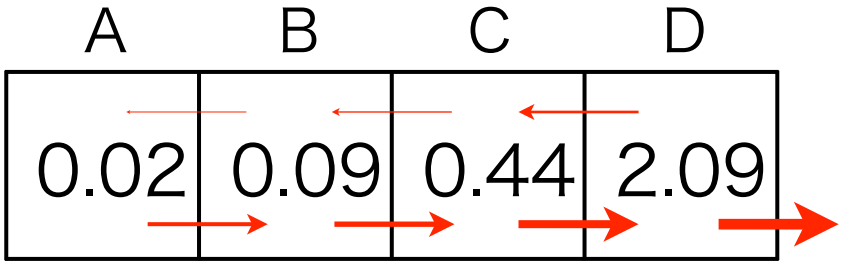
④ エージェントがDから今回も右(*R*方向)に進んで外に出たとする

問1: ＴＤ誤差を求めなさい

問2: *V(D)*, *p(D,R)*を更新しなさい

Actor-Critic法の学習例

学習が進むと...



(数値はV(s)の理論値)

Criticは状態の良さを, Actorは各状態でとるべき行動を獲得