2章　マルコフ決定過程

20A3017 石川悠樹

前回と今回の問題設定の違い

バンディット問題

複数のスロットマシンを回して得られるコインの枚数を最大化する問題

何回回してもスロットマシンの
報酬設定が変わらず行動価値も一定

現実にある多くの問題は違う

前回と今回の問題設定の違い

今回の問題設定

エージェントの行動によって状況が変わる問題

例：囲碁…エージェントが石を打つ度盤面が変化する

マルコフ決定過程
Markov Decision Process(MDP)

決定過程：エージェントが（環境と相互作用しながら）行動を決定する過程

MDPについて

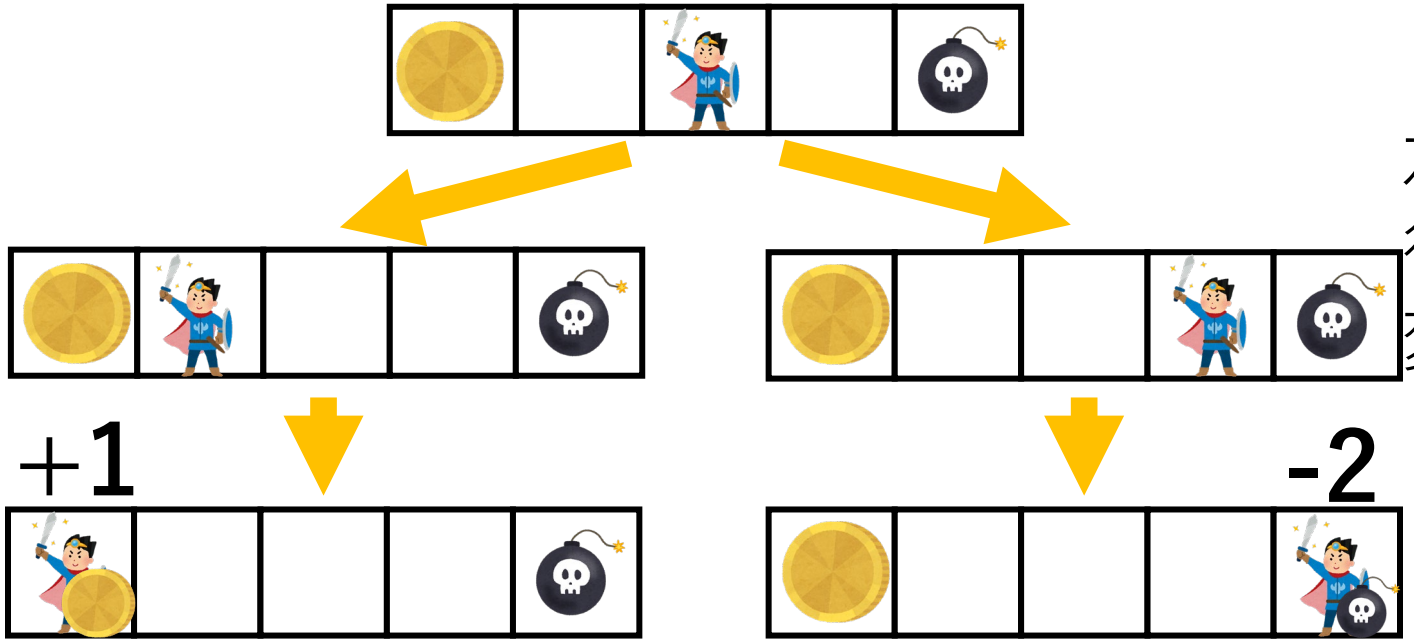
MDPの問題設定の例

報酬
コイン：＋1
爆弾：-2
空き：0

上図において勇者が「エージェント」であり
1. 右に進む
2. 左に進む
の２つの行動をとることができる

MDPについて

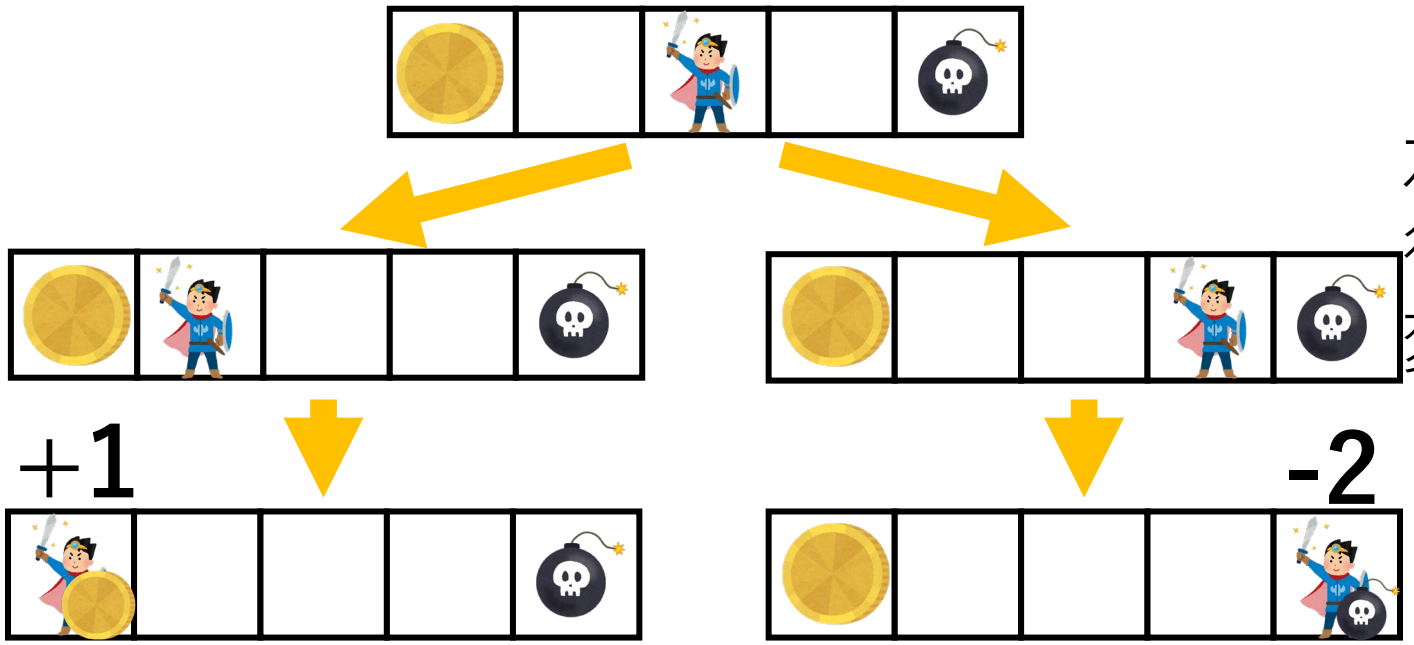
MDPの問題設定の例



左図のようにエージェントの行動によって状況が逐次的に変化する

MDPについて

MDPの問題設定の例



左図のようにエージェントの行動によって状況が逐次的に変化する

状態(state)

時間(エージェントの意思決定の間隔)が進む度、新しい状態に遷移する

MDPについて

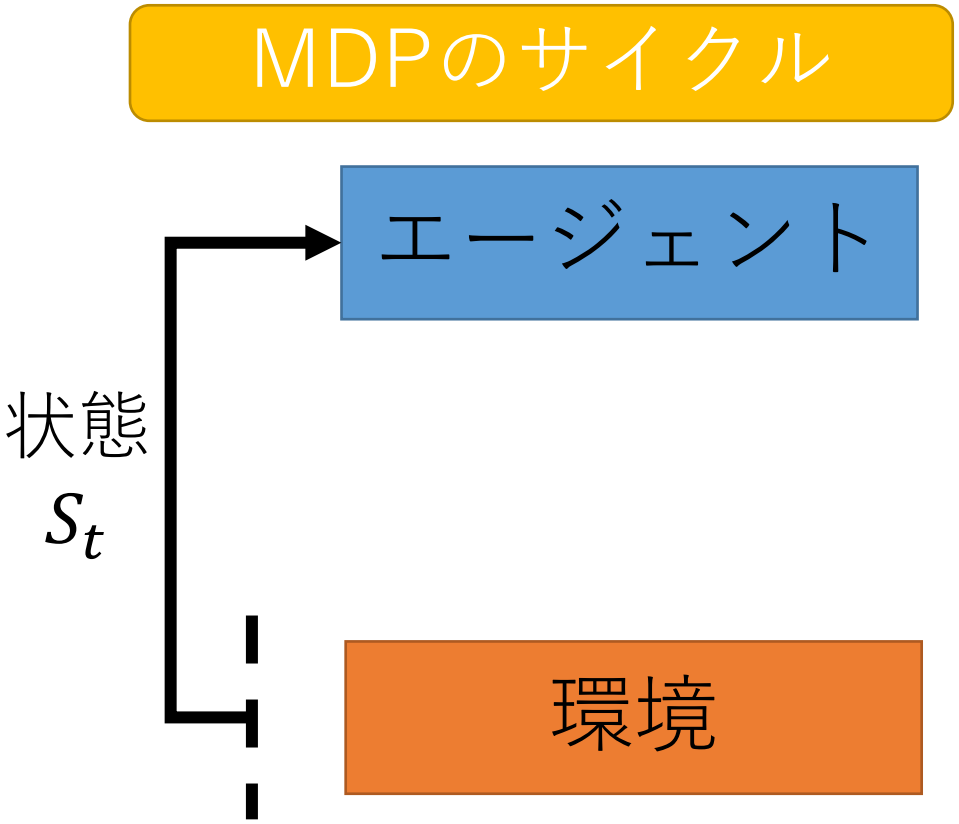
MDPの問題設定の例

報酬
コインの山：+6
コイン：+1
爆弾：-2
空き：0

- 右に2回進んだ場合　⇒　-2 + 6 = 4
- 左に2回進んだ場合　⇒　0 + 1 = 1

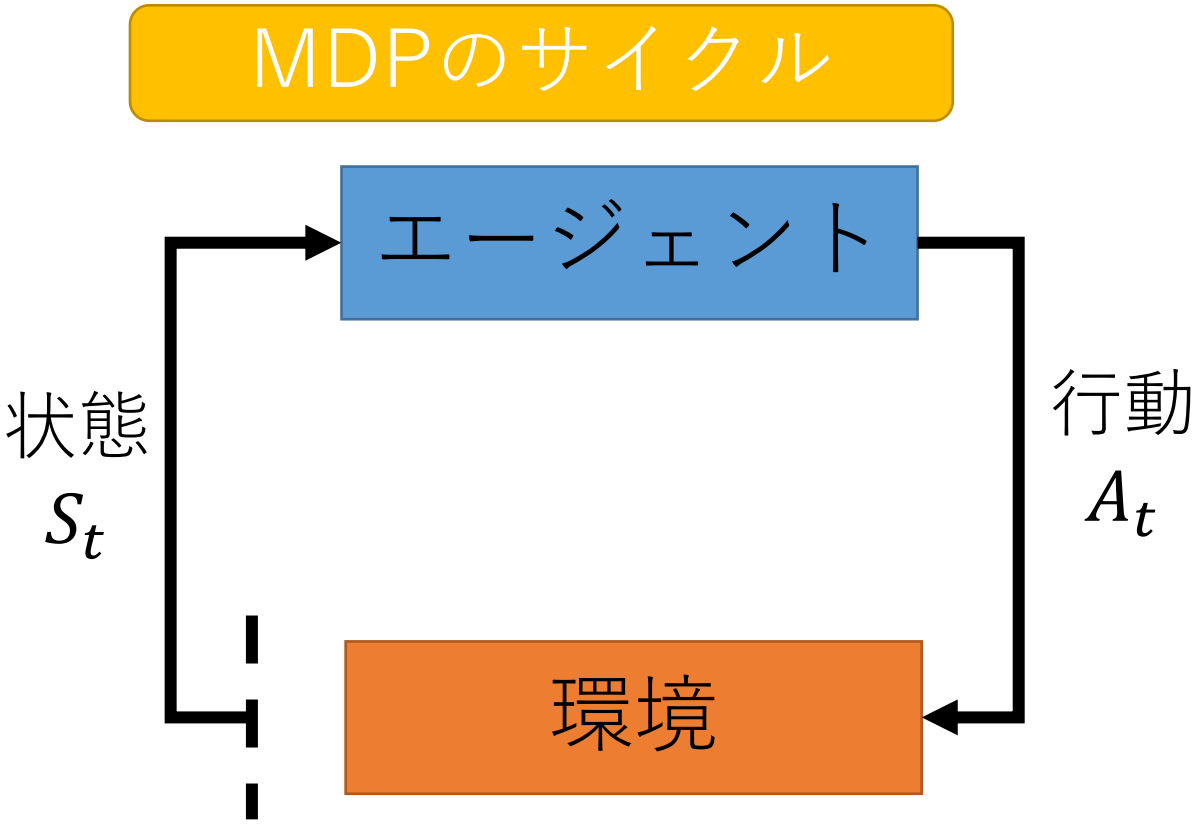
報酬の総和を最大化することが目的

エージェントと環境



1. 状態S_tが与えられる

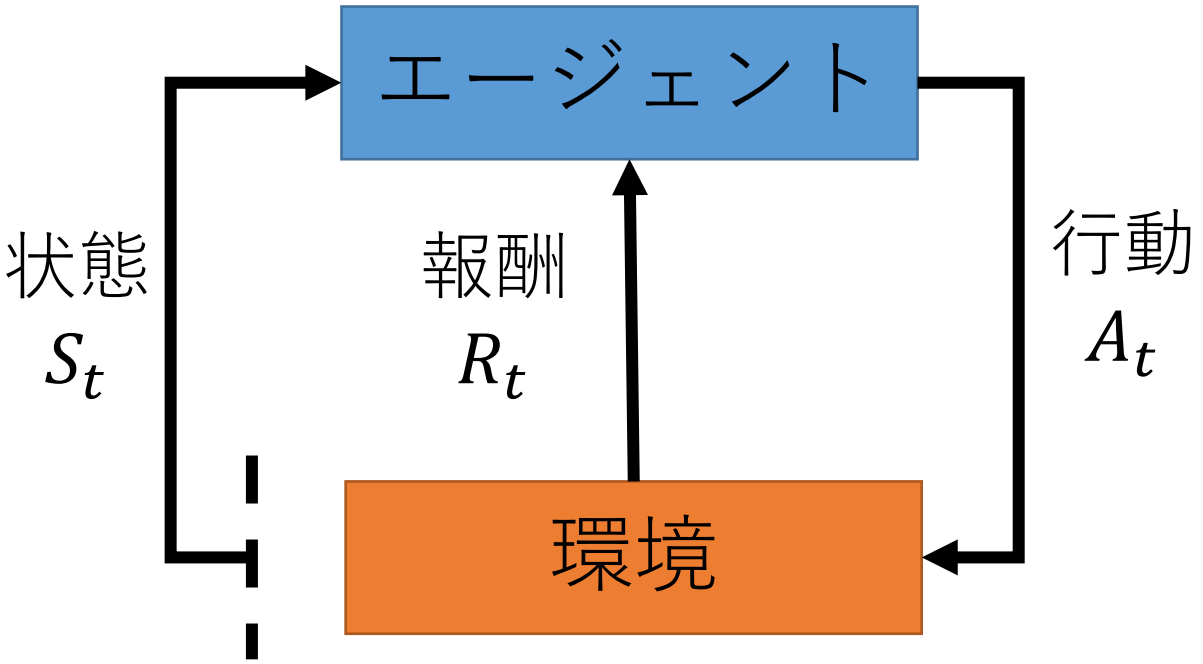
エージェントと環境



1. 状態S_tが与えられる
2. 状態S_tに応じて行動A_tを行う

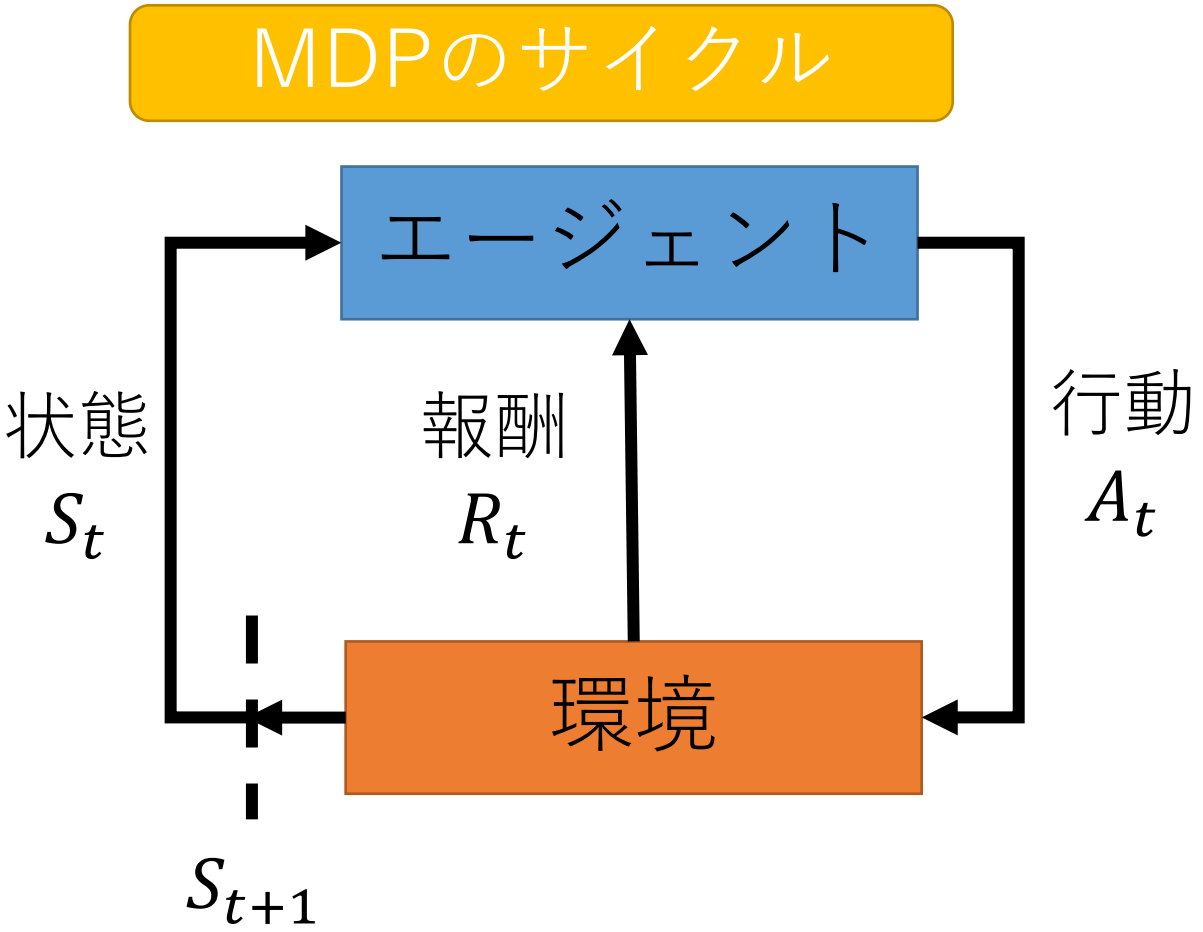
エージェントと環境

MDPのサイクル



1. 状態S_tが与えられる
2. 状態S_tに応じて行動A_tを行う
3. 行動A_tに対して報酬R_tを得る

エージェントと環境



1. 状態S_tが与えられる
2. 状態S_tに応じて行動A_tを行う
3. 行動A_tに対して報酬R_tを得る
4. 状態がS_{t+1}に遷移する

環境とエージェントの定式化

MDPのサイクルは数式によって定式化される

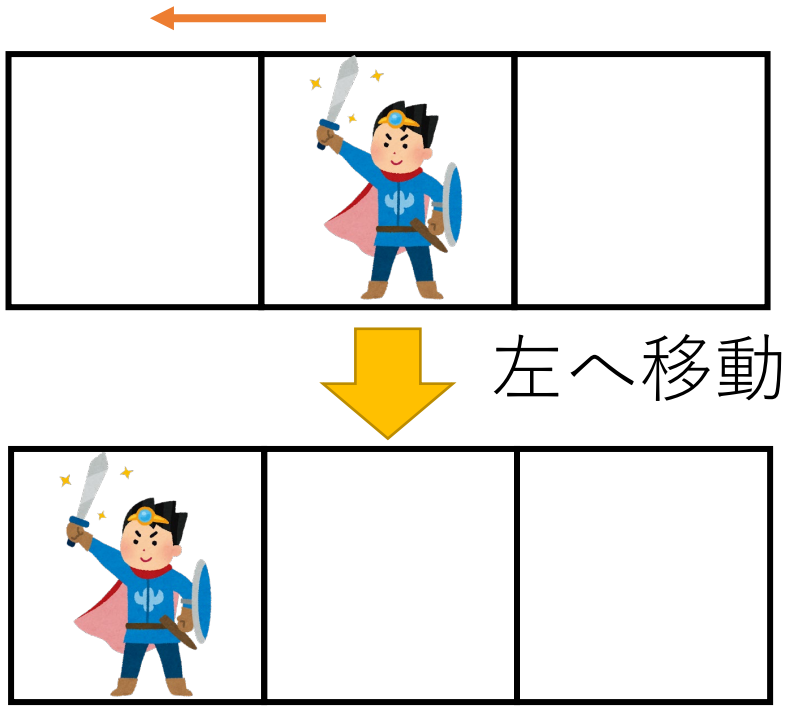
定式化に必要な3つの要素

- 状態遷移：状態がどのように遷移するか
- 報酬：報酬がどのように与えられるか
- 方策：エージェントがどのように行動を決定するか

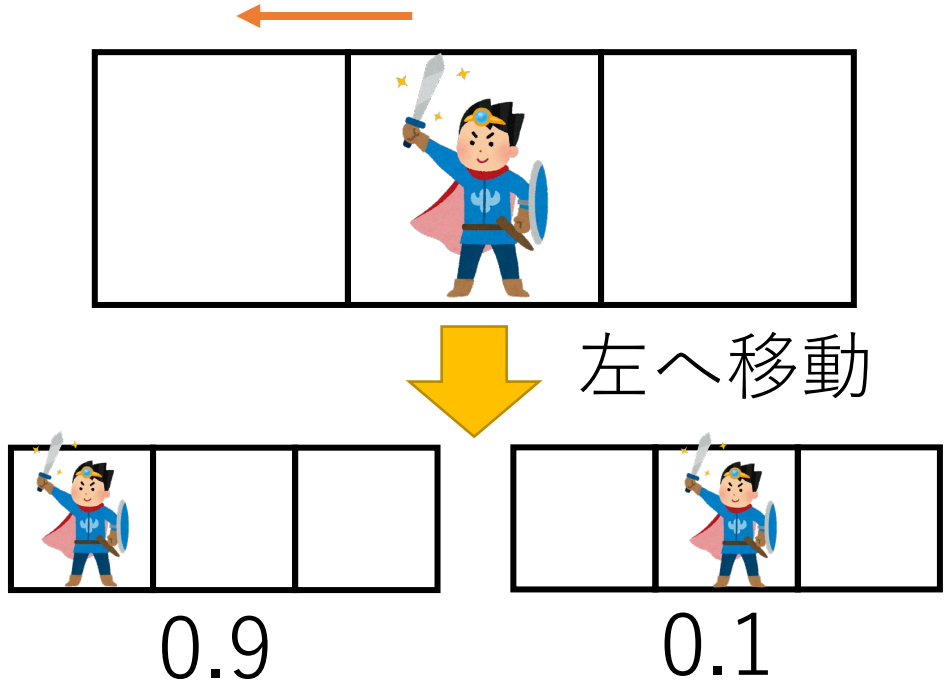
環境とエージェントの定式化

状態遷移

決定論的



確率的



環境とエージェントの定式化

状態遷移

状態遷移は決定論的な場合と確率的な場合で次のように表せる

決定論的

$$s' = f(s, a)$$

状態遷移関数

s：状態
a：行動

環境とエージェントの定式化

状態遷移

状態遷移は決定論的な場合と確率的な場合で次のように表せる

決定論的

$$s' = f(s, a)$$

状態遷移関数

確率的

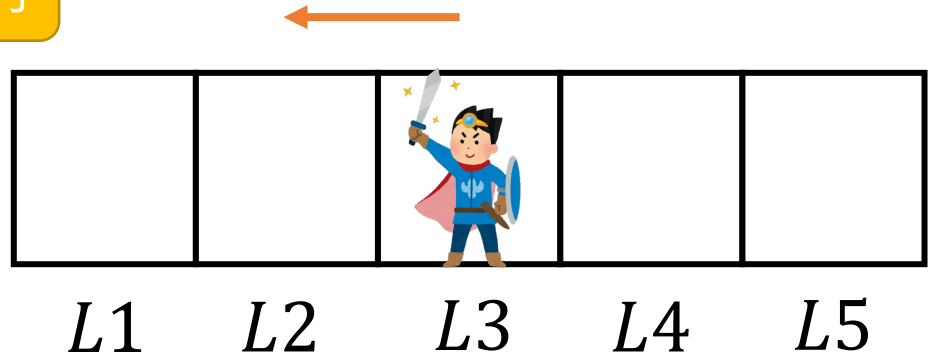
$$p(s'|s, a)$$

状態遷移確率

s：状態
a：行動

環境とエージェントの定式化

状態遷移確率の例



s'	$L1$	$L2$	$L3$	$L4$	$L5$
$p(s'\|s,a)$	0	0.9	0.1	0	0

$L3$から左$(Left)$に行く行動を選択した場合、
状態遷移確率$p(s'|s, = L3\ a = Left)$は上表のように表せる

環境とエージェントの定式化

マルコフ性

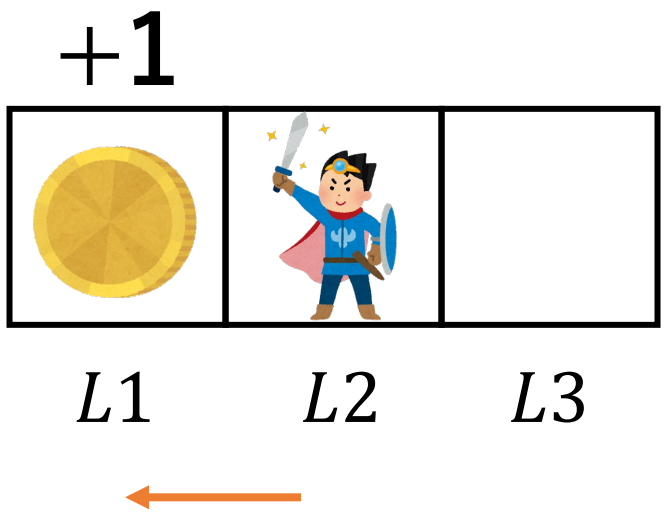
$p(s'|s,a)$は現在の情報(状態、行動)のみに依存している
過去の情報は必要ない

マルコフ性

計算量を抑え、問題を解きやすくすることができる

環境とエージェントの定式化

報酬



報酬が決定論的に与えられることを
前提とする
現在の状態がsのとき、エージェントが
行動aを起こしたとすると得られる報酬は

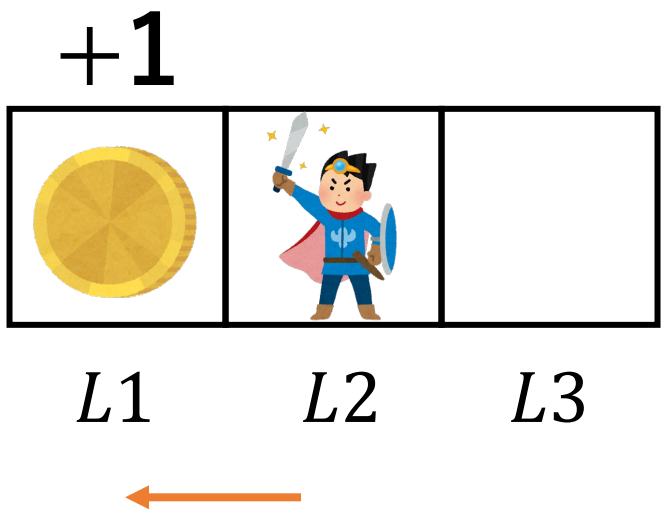
$$r(s, a, s')$$

で表せる

$$r(s = L2, a = Left, s' = L1) = 1$$

環境とエージェントの定式化

報酬



報酬が決定論的に与えられることを
前提とする
現在の状態がsのとき、エージェントが
行動aを起こしたとすると得られる報酬は

$$r(s, a, s')$$

で表せる

$$r(s = L2, a = Left, s' = L1) = 1$$

報酬関数

環境とエージェントの定式化

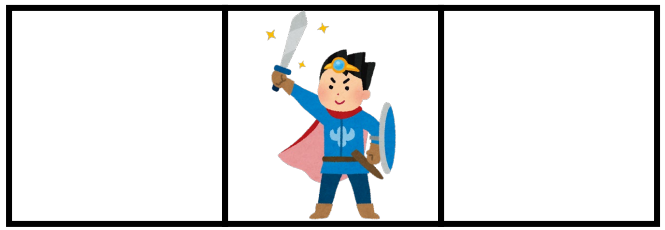
方策

エージェントがどのように行動を決めるかを表す

MDPはマルコフ性を持っているため、エージェントは現在の情報sのみで最適な選択を行うことができる

決定論的

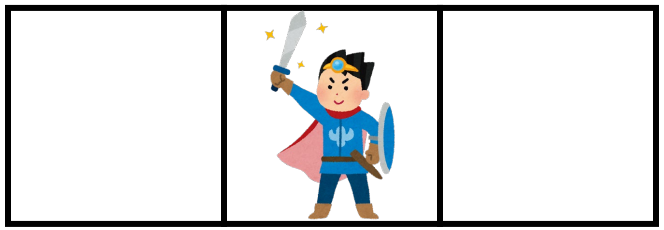
$Left$



$L1$ $L2$ $L3$

確率的

$Left(0.4)$ $Right(0.6)$



$L1$ $L2$ $L3$

環境とエージェントの定式化

方策

エージェントがどのように行動を決めるかを表す

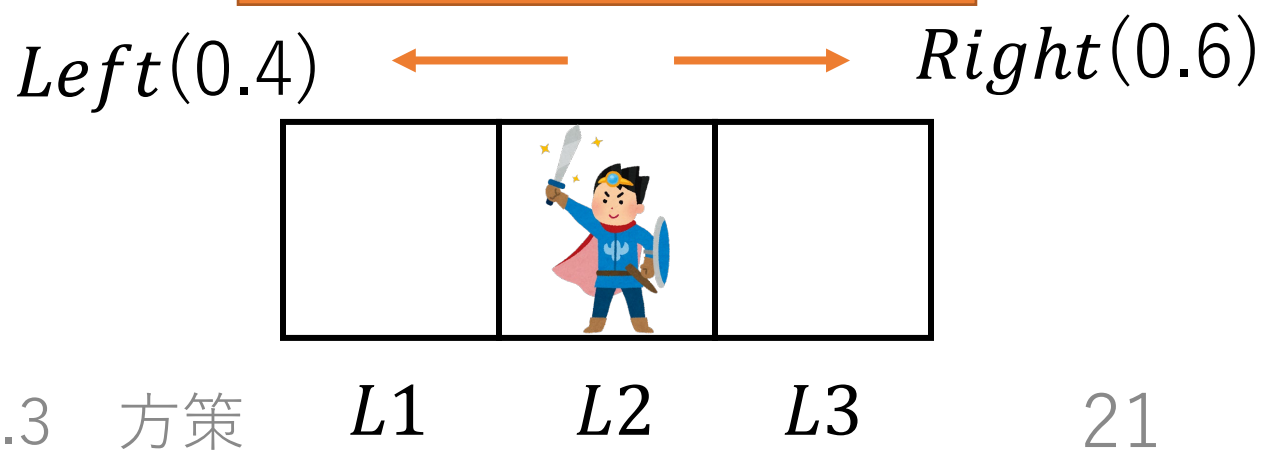
MDPはマルコフ性を持っているため、エージェントは現在の情報sのみで最適な選択を行うことができる

決定論的

$$a = \mu(s)$$

$$\mu(s = L2)$$

確率的



環境とエージェントの定式化

方策

エージェントがどのように行動を決めるかを表す

MDPはマルコフ性を持っているため、エージェントは現在の情報sのみで最適な選択を行うことができる

決定論的

$$a = \mu(s)$$

$$\mu(s = L2)$$

確率的

$$\pi(a|s)$$

$$\pi(a = Left|s = L3) = 0.4$$

$$\pi(a = Right|s = L3) = 0.6$$

まとめ

- MDPは現在の状態をもとに行動を決定する過程のこと
- MDPでは環境とエージェントが相互に作用しあう
- MDPの定式化に必要な3つの要素
  - 状態遷移
    - $s' = f(s, a)$
    - $p(s'|s, a)$
  - 報酬
    - $r(s, a, s')$
  - 方策
    - $a = \mu(s)$
    - $\pi(a|s)$