九州大学 工学部地球環境工学科
船舶海洋システム工学コース

システム設計工学 （担当：木村）

（13） マルコフ決定過程1

場所：船1講義室

http://sysplan.nams.kyushu-u.ac.jp/gen/index.html

マルコフ決定過程（Markov Decision Process: MDP）

●マルコフ過程に　　を付加
　各状態において「行動」を選択することで遷移確率や報酬を
　コントロールできる

●MDPで何をモデル化できるか？
　状態遷移に不確実性を伴う制御問題（意思決定問題）
　例）　在庫管理，配送計画問題，生産システム管理問題，ロボット

●MDPによるモデル化のメリットは？
　膨大な数理解析の知見を利用できる
　１）　最適性・最適解の保証
　２）　効率良く最適な意思決定の解を求める方法論がある

特に意思決定を要する行動の組合せが膨大なとき有用

マルコフ決定過程（Markov Decision Process: MDP）

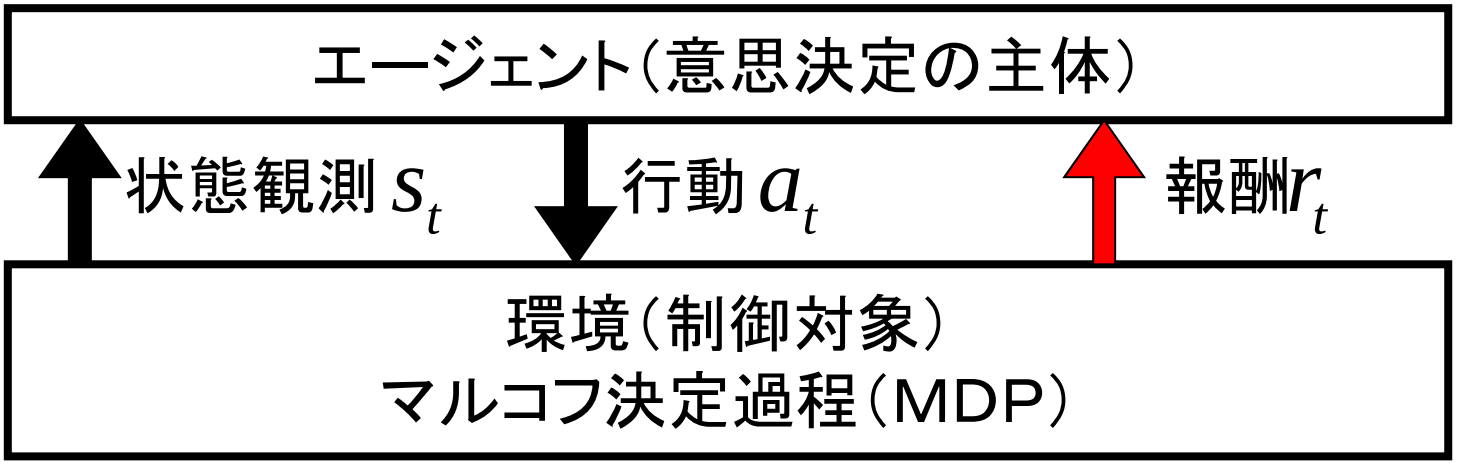
●マルコフ過程に **意思決定**（decision） を付加
各状態において「行動」を選択することで遷移確率や報酬を
コントロールできる

●MDPで何をモデル化できるか？
状態遷移に不確実性を伴う制御問題（意思決定問題）
例） 在庫管理，配送計画問題，生産システム管理問題，ロボット

●MDPによるモデル化のメリットは？
膨大な数理解析の知見を利用できる
１） 最適性・最適解の保証
２） 効率良く最適な意思決定の解を求める方法論がある

特に意思決定を要する行動の組合せが膨大なとき有用

MDPにおけるプランニングの枠組み



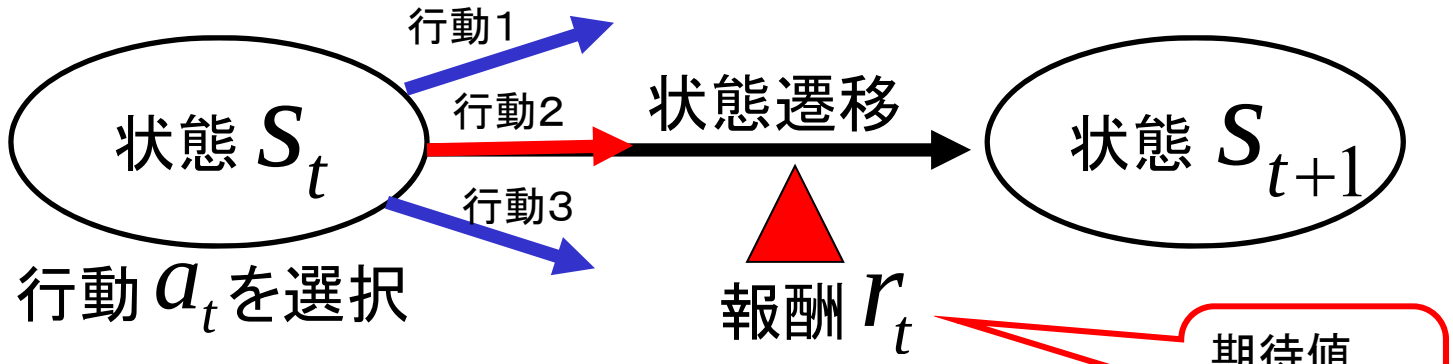
●状態観測 → 行動選択 → （状態遷移）→ 報酬　　繰返し

●何回か状態遷移した後，やっと報酬を得る
　→ 多段決定過程 （報酬に遅れ）
　　目標状態に達したら大きな報酬
　　タスクを達成したら大きな報酬
　　常にコストがかかる
　→ 報酬が状態遷移に依存する

●報酬の合計が最大になるような制御規則（政策）を求める
　→ プランニング

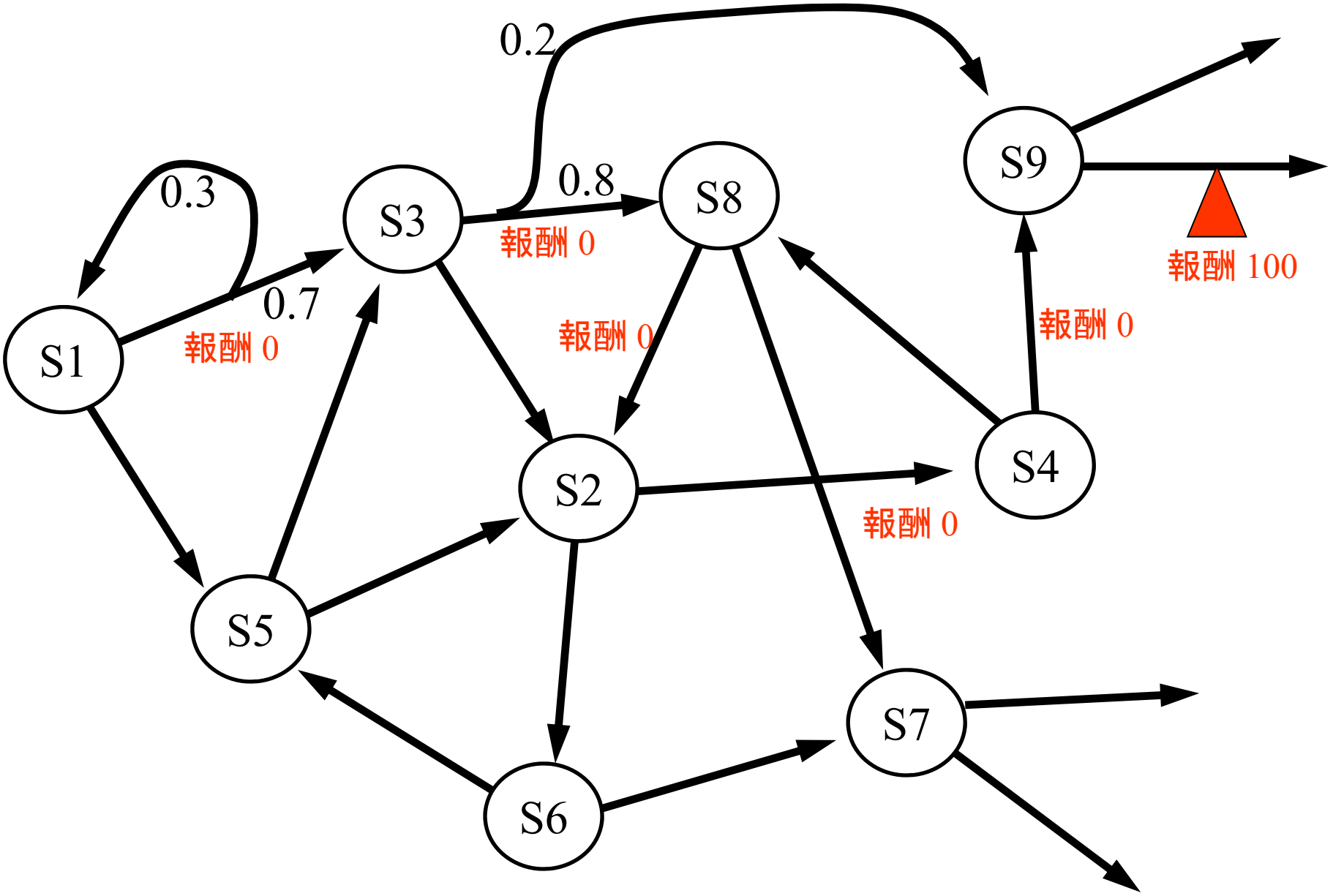
マルコフ決定過程(MDP)とは？

S	:状態の集合
A	:行動の集合
Pr(s'|s,a)	:状態sで行動aをとったときs'へ遷移する確率
$R^a(s,s')$	:状態sで行動aをとってs'へ遷移したときの報酬の期待値



状態遷移と報酬は状態S_tと行動a_tのみに依存し、
それ以前の状態や行動の履歴には依存しない → マルコフ性

マルコフ決定過程(MDP)の状態遷移



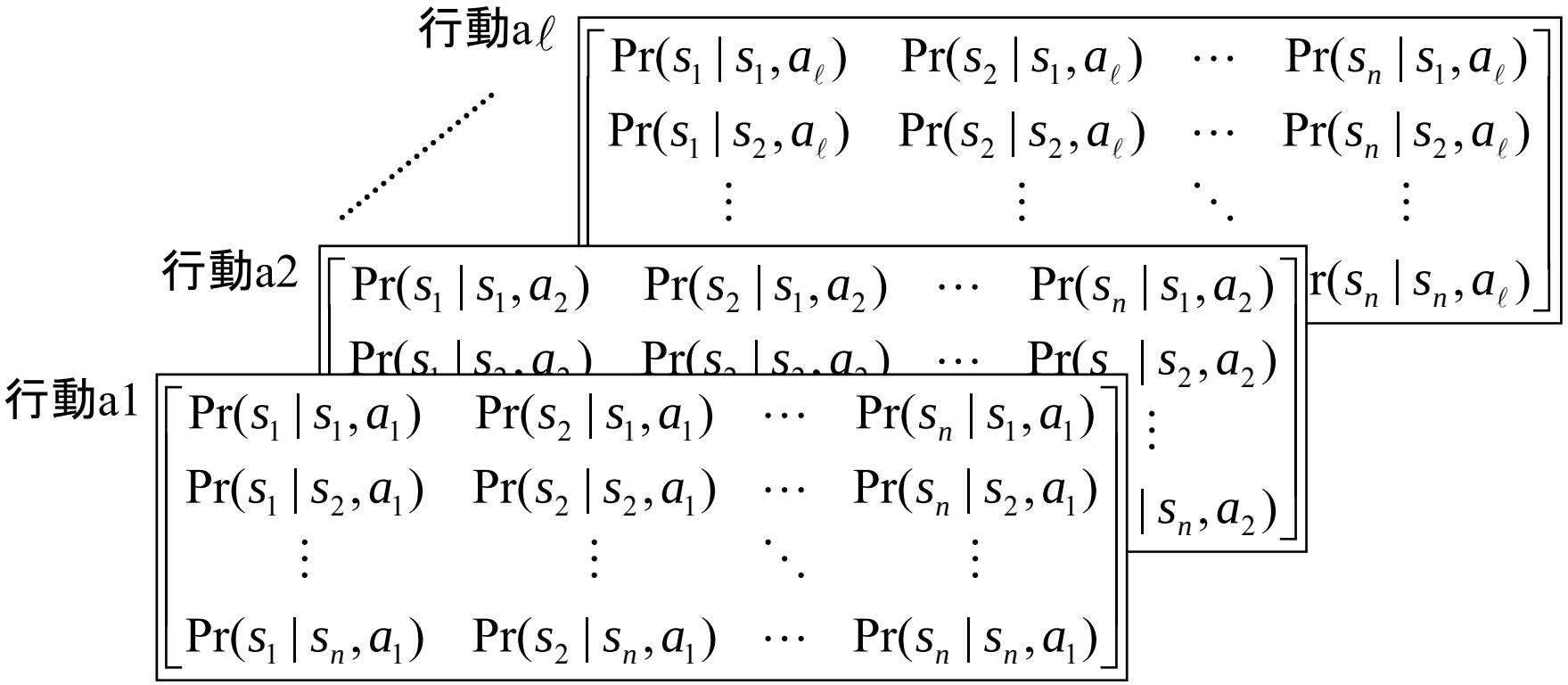
MDPの状態遷移マトリクス

ある行動 a1 を選択する場合

これだけならマルコフ過程の遷移行列と同じ

MDPの状態遷移マトリクス

状態数n，行動数ℓのとき，
マトリクスの大きさは $n \times n \times \ell$



報酬関数 $R^a(s, s')$ も同様

エージェントの制御規則：政策

- 各状態 S で選択する行動 a を規定 → $\pi(s,a)$ 政策πにおいて状態sで行動aを選ぶ確率
- ある政策πが定義されると，**状態遷移確率は n×n 正方行列**になる

状態S → 状態S’

$$
\begin{array}{c} \\ S1 \\ S2 \\ \cdot \\ Sn \end{array}
\begin{array}{c}
\begin{array}{cccc} S1 & S2 & \cdots & Sn \end{array} \\
\begin{bmatrix}
P^{\pi}(s_1,s_1) & P^{\pi}(s_1,s_2) & \cdots & P^{\pi}(s_1,s_n) \\
P^{\pi}(s_2,s_1) & P^{\pi}(s_2,s_2) & \cdots & P^{\pi}(s_2,s_n) \\
\vdots & \vdots & \ddots & \vdots \\
P^{\pi}(s_n,s_1) & P^{\pi}(s_n,s_2) & \cdots & P^{\pi}(s_n,s_n)
\end{bmatrix}
\end{array}
= \mathbf{P}^{\pi}
$$

$$
= \begin{bmatrix}
\sum_a \pi(s_1,a)\Pr(s_1 \mid s_1,a) & \sum_a \pi(s_1,a)\Pr(s_2 \mid s_1,a) & \cdots & \sum_a \pi(s_1,a)\Pr(s_n \mid s_1,a) \\
\sum_a \pi(s_2,a)\Pr(s_1 \mid s_2,a) & \sum_a \pi(s_2,a)\Pr(s_2 \mid s_2,a) & \cdots & \sum_a \pi(s_2,a)\Pr(s_n \mid s_2,a) \\
\vdots & \vdots & \ddots & \vdots \\
\sum_a \pi(s_n,a)\Pr(s_1 \mid s_n,a) & \sum_a \pi(s_n,a)\Pr(s_2 \mid s_n,a) & \cdots & \sum_a \pi(s_n,a)\Pr(s_n \mid s_n,a)
\end{bmatrix}
$$

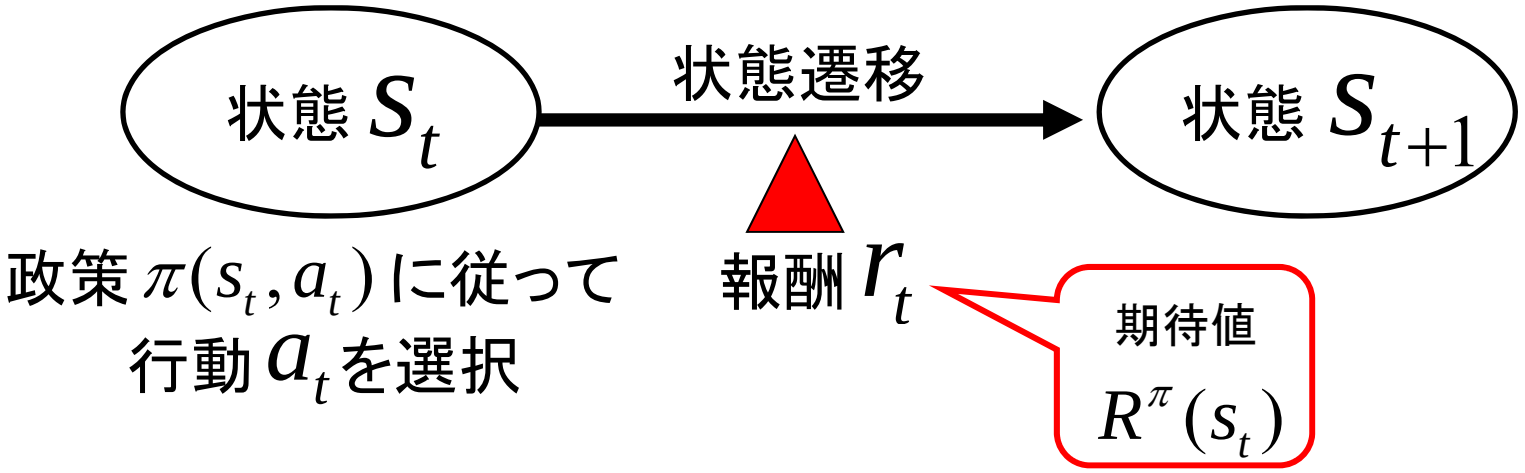
エージェントの制御規則：政策

$\pi(s,a)$ 政策πにおいて状態sで行動aを選ぶ確率

- 各状態 S で選択する行動 a を規定
- ある政策πが定義されると，**報酬の期待値は 1×n の行列**になる

$$\mathbf{R}^{\pi} = \begin{bmatrix} \sum_{s'}\sum_{a}\pi(s_1,a)\Pr(s'|s_1,a)R^{a}(s_1,s') \\ \sum_{s'}\sum_{a}\pi(s_2,a)\Pr(s'|s_2,a)R^{a}(s_2,s') \\ \vdots \\ \sum_{s'}\sum_{a'}\pi(s_n,a)\Pr(s'|s_n,a)R^{a}(s_n,s') \end{bmatrix} = \begin{bmatrix} R^{\pi}(s_1) \\ R^{\pi}(s_2) \\ \vdots \\ R^{\pi}(s_n) \end{bmatrix}$$

←状態S1の報酬の期待値
←状態S2の報酬の期待値
←状態Snの報酬の期待値



マルコフ決定過程における「政策」の性質

- 政策は「状態」から「行動」へのマッピング
- 政策を定めると、マルコフ決定過程は単なる [　　] になる
- 平均報酬（あるいは報酬合計）を最大化する**最適政策**は、[　　] 場合がある。
- 最適な政策には、各状態において、ある行動をとる確率が1であるような「**決定論的な政策**」が必ず存在する。
- マルコフ決定過程の遷移行列と報酬行列から最適政策を求めることは [　　] と呼ばれる。

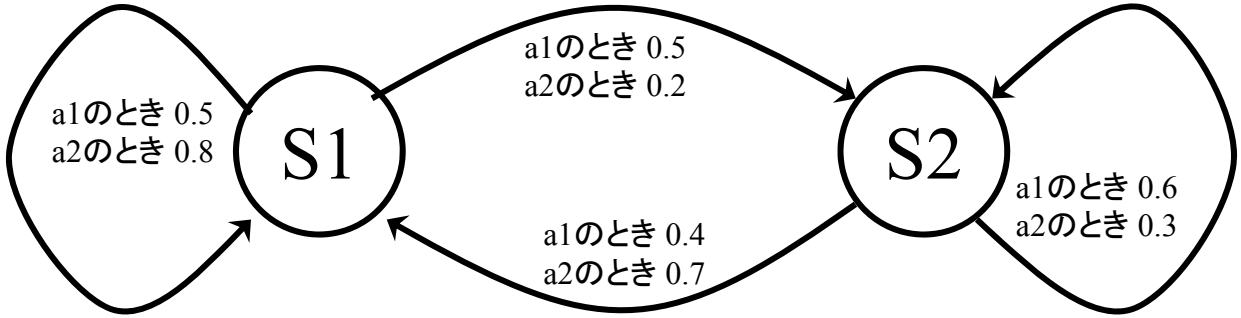
マルコフ決定過程における「政策」の性質

- 政策は「状態」から「行動」へのマッピング
- 政策を定めると、マルコフ決定過程は単なる マルコフ過程 になる
- 平均報酬（あるいは報酬合計）を最大化する**最適政策**は、
  複数存在する 場合がある。
- 最適な政策には、各状態において、ある行動をとる確率が1で
  あるような「**決定論的な政策**」が必ず存在する。
- マルコフ決定過程の遷移行列と報酬行列から最適政策を求める
  ことは プランニング と呼ばれる。

例題： 製品製造業者の意思決定問題

ある製品の製造業者が、市場における製品の人気の状態を観測し、各状態に応じて適切な決定を下す。

- ●状態 S1：市場において製品の人気がある
- ●状態 S2：市場において製品の人気がない
- ●状態 S1 における行動 a1 ：広告をしない、行動 a2：広告を出す
- ●状態 S2 における行動 a1 ：何もしない、行動 a2：新製品を開発
- ●報酬：各期の売り上げから経費を差し引いた金額

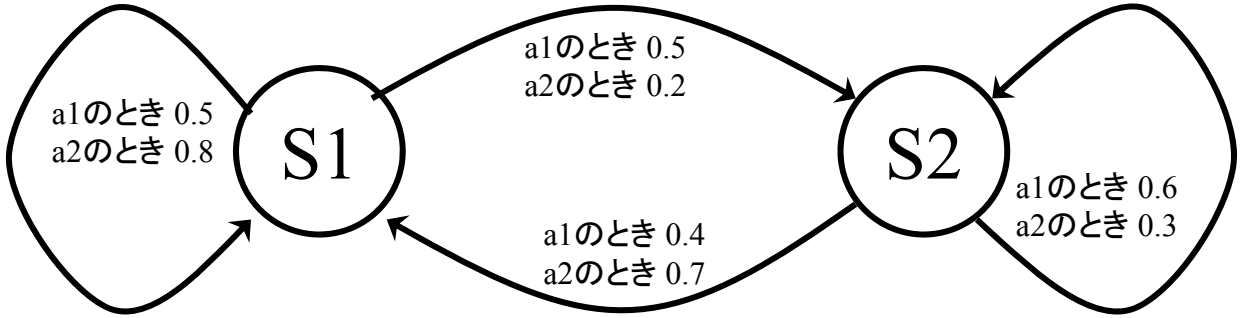


状態 s	行動 a	推移確率		報酬		直接報酬の期待値
		$\Pr(s_1\|s,a)$	$\Pr(s_2\|s,a)$	$R^a(s,s_1)$	$R^a(s,s_2)$	
s_1：人気あり	a_1：広告なし	0.5	0.5	9	3	
	a_2：広告あり	0.8	0.2	4	4	
s_2：人気なし	a_1：研究なし	0.4	0.6	3	−7	
	a_2：研究あり	0.7	0.3	1	−19	

例題： 製品製造業者の意思決定問題

ある製品の製造業者が、市場における製品の人気の状態を観測し、各状態に応じて適切な決定を下す。

- 状態 S1：市場において製品の人気がある
- 状態 S2：市場において製品の人気がない
- 状態 S1 における行動 a1 ：広告をしない、行動 a2：広告を出す
- 状態 S2 における行動 a1 ：何もしない、行動 a2：新製品を開発
- 報酬：各期の売り上げから経費を差し引いた金額



状態 s	行動 a	推移確率		報酬		直接報酬の期待値
		$\Pr(s_1\|s,a)$	$\Pr(s_2\|s,a)$	$R^a(s,s_1)$	$R^a(s,s_2)$	
s_1：人気あり	a_1：広告なし	0.5	0.5	9	3	$6 = 0.5 \times 9 + 0.5 \times 3$
	a_2：広告あり	0.8	0.2	4	4	$4 = 0.8 \times 4 + 0.2 \times 4$
s_2：人気なし	a_1：研究なし	0.4	0.6	3	-7	$-3 = 0.4 \times 3 + 0.6 \times (-7)$
	a_2：研究あり	0.7	0.3	1	-19	$-5 = 0.7 \times 1 + 0.3 \times (-19)$

MDPの解法：最適性原理とダイナミックプログラミング

MDPの最適政策は、**決定論的政策**（すなわち各状態においてある行動を確率1で選択する）の中に存在する

→ **決定論的政策**は有限個なので、全ての政策を評価すれば良いが
状態・行動の個数が増えると非効率的

状態数 n，行動数ℓのとき，
決定論的な政策の個数は 

大規模問題では、全ての政策を計算／比較するのは無理

【MDPの最適性原理】
「最適な政策」とは、一連の意思決定において、
各状態で常に最適な行動を選択すること

→ 局所的な最適化の繰り返しによって全体が最適化できる
評価する解候補を有望なものに絞り込むことで効率良く探索

ダイナミックプログラミング（動的計画法）

MDPの解法：最適性原理とダイナミックプログラミング

MDPの最適政策は、**決定論的政策**（すなわち各状態においてある行動を確率1で選択する）の中に存在する

→ **決定論的政策**は有限個なので、全ての政策を評価すれば良いが
状態・行動の個数が増えると非効率的

状態数 n，行動数ℓのとき，
決定論的な政策の個数は ℓ^n → 大規模問題では、全ての政策を計算／比較するのは無理

例）状態数20，行動数2
→ $2^{20} > 10^{13}$ =10兆

【MDPの最適性原理】
「最適な政策」とは、一連の意思決定において、
各状態で常に最適な行動を選択すること

→ 局所的な最適化の繰り返しによって全体が最適化できる
評価する解候補を有望なものに絞り込むことで効率良く探索

ダイナミックプログラミング（動的計画法）

【復習】「割引報酬」による評価

報酬合計の期待値を最大化する政策を見つける

ただし，

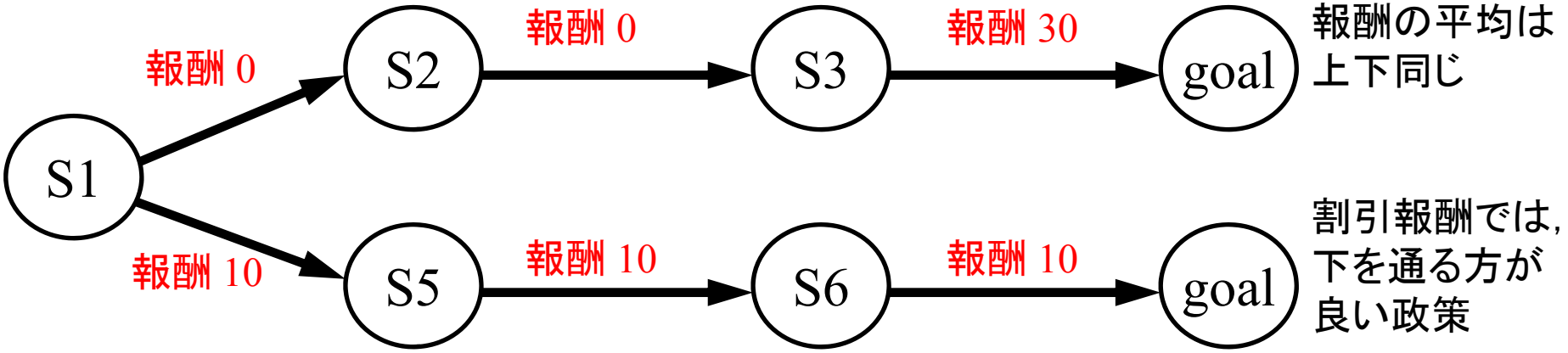
１－γの確率で停止する場合の報酬合計＝**「割引報酬」**

割引率γ

１ステップあたりγの確率で活動を続ける

γ→１　長期的な利益最大化

γ→０　目先の利益最大化



「割引報酬」による最適性の定義

政策π：各状態における行動選択確率
以下の**割引報酬の合計**を最大化する政策を見つける

$$\sum_{t=0}^{\infty} \gamma^t r_t$$ ただしγは割引率

報酬を割引く理由：
1）未来の報酬はあてにならない（環境の変化や誤差等）
　→未来に得る報酬を、遠い未来ほど割引いて評価
　γ＝1のとき：　単なる報酬合計
　γ＜1のとき：　1－γの確率で停止する場合の報酬合計
2）計算の利便性

未来におけるリスクを考慮

$1-\gamma$の確率で停止する場合の報酬合計評価は状態依存

$$\mathbf{V}^{\pi} = \begin{bmatrix} V^{\pi}(s_1) \\ V^{\pi}(s_2) \\ \vdots \\ V^{\pi}(s_n) \end{bmatrix}$$

←S1からスタートした場合の報酬合計の期待値

←S2からスタートした場合の報酬合計の期待値

←Snからスタートした場合の報酬合計の期待値

$$= \mathbf{R}^{\pi} + \gamma \mathbf{P}^{\pi}\mathbf{R}^{\pi} + \gamma^2 (\mathbf{P}^{\pi})^2 \mathbf{R}^{\pi} + \cdots$$

最初の遷移で得る報酬の期待値

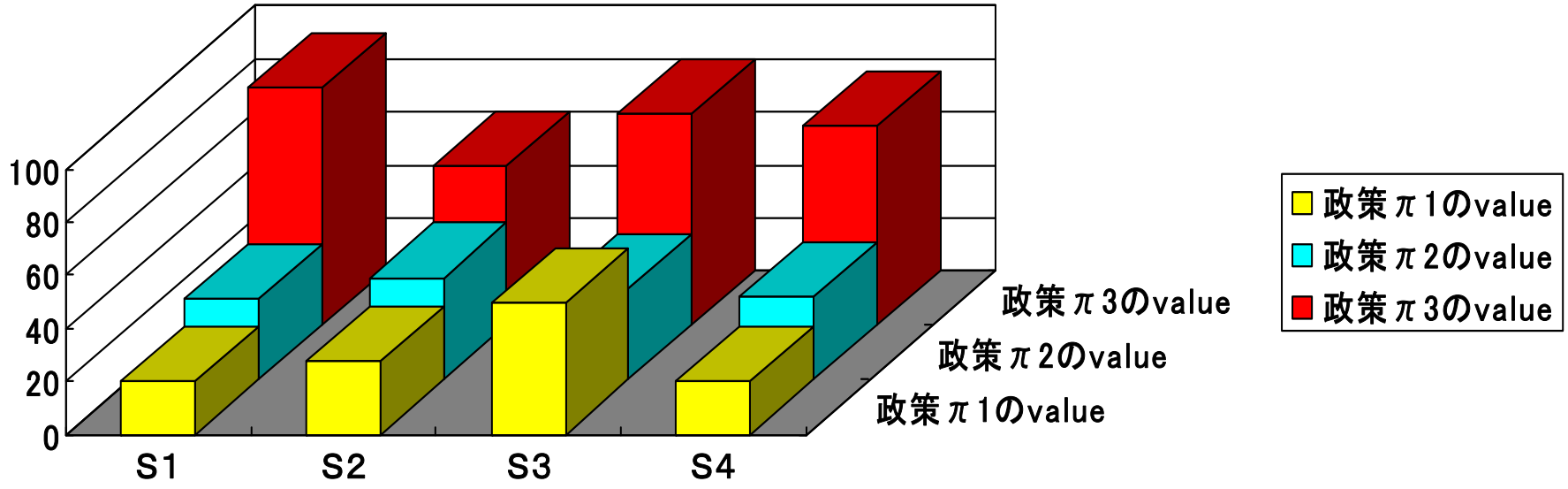
2回目の遷移で得る報酬の期待値

3回目の遷移で得る報酬の期待値

$$= \mathbf{R}^{\pi} + \gamma \mathbf{P}^{\pi}\mathbf{V}^{\pi}$$

$$= (\mathbf{I} - \gamma \mathbf{P}^{\pi})^{-1} \mathbf{R}^{\pi}$$

評価値(value)が状態依存の場合の政策の比較方法



●政策π1とπ2は状態によって評価値の大小関係が入れ替わる→政策の良し悪しははっきり決められない

●政策π3は全てのvalueの要素が他の政策の値をドミネートしている→π3が最も良い政策

●MDPでは他の政策のvalueをドミネートする最適政策が必ず存在

MDPの割引報酬評価における最適性

MDPにおいて定常政策πをとるとき，
割引報酬合計の期待値(value)は状態Sの関数になる：

$$V^{\pi}(s) = E\left\{\sum_{t=0}^{\infty} \gamma^{t} r_t \mid s_{t=0} = s\right\}$$

すなわち，**政策の評価値は状態依存**

$$\left(V^{\pi}(s_1), V^{\pi}(s_2), \cdots V^{\pi}(s_n)\right)$$

この全要素を最大化する政策が最適政策 π^*
そのときのvalueが最適value関数 V^*

最適value関数 V^* は［　　］だが、
［　　］することがある

このとき、全ての最適政策は同じ最適value関数を共有する
→ 最適value関数を求めよう！

MDPの割引報酬評価における最適性

MDPにおいて定常政策πをとるとき，
割引報酬合計の期待値(value)は状態Sの関数になる：

$$V^{\pi}(s) = E\left\{\sum_{t=0}^{\infty} \gamma^{t} r_t \mid s_{t=0} = s\right\}$$

すなわち，**政策の評価値は状態依存**

$$\left(V^{\pi}(s_1), V^{\pi}(s_2), \cdots V^{\pi}(s_n)\right)$$

この**全要素を最大化**する政策が**最適政策** π^*
そのときのvalueが最適value関数 V^*

最適value関数 V^* は **ただ1つ** だが、
最適政策 π^* は 複数存在 することがある

このとき、**全ての最適政策は同じ最適value関数を共有**する
→ **最適value関数を求めよう！**

ダイナミックプログラミングの準備

$$\mathbf{V}^{\boldsymbol{\pi}} = \begin{bmatrix} V^{\pi}(s_1) \\ V^{\pi}(s_2) \\ \vdots \\ V^{\pi}(s_n) \end{bmatrix}$$

←S1からスタートした場合の報酬合計の期待値

←S2からスタートした場合の報酬合計の期待値

←Snからスタートした場合の報酬合計の期待値

$$= \mathbf{R}^{\boldsymbol{\pi}} + \gamma \mathbf{P}^{\boldsymbol{\pi}} \mathbf{R}^{\boldsymbol{\pi}} + \gamma^2 (\mathbf{P}^{\boldsymbol{\pi}})^2 \mathbf{R}^{\boldsymbol{\pi}} + \cdots$$

$$= \mathbf{R}^{\boldsymbol{\pi}} + \gamma \mathbf{P}^{\boldsymbol{\pi}} \mathbf{V}^{\boldsymbol{\pi}}$$

$$= (\mathbf{I} - \gamma \mathbf{P}^{\boldsymbol{\pi}})^{-1} \mathbf{R}^{\boldsymbol{\pi}}$$

ここに注目
各要素について式を書く

$$V^{\pi}(s) = R^{\pi}(s) + \sum_{s' \in S} P^{\pi}(s' | s, a) \gamma V^{\pi}(s')$$

イメージによる式の理解

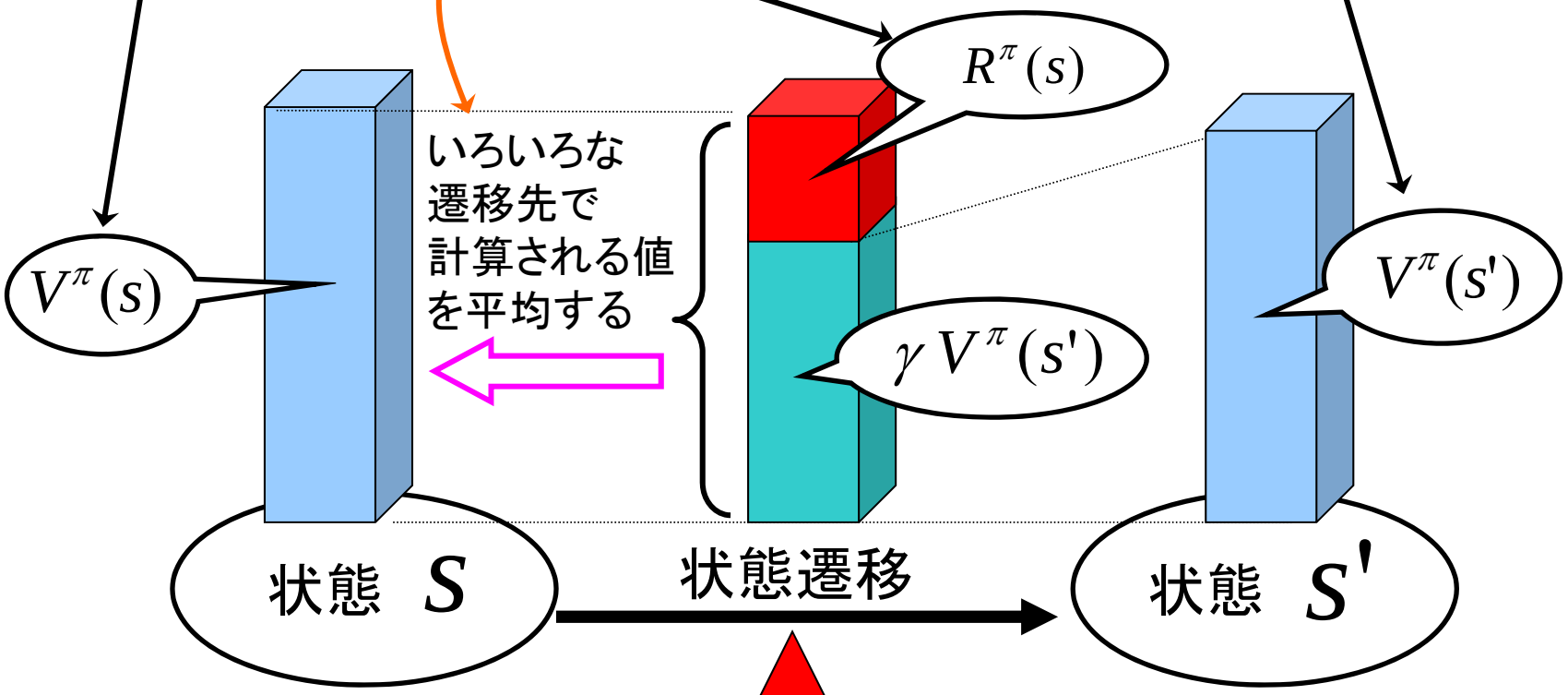
政策πのときの状態遷移確率

割引率

$$V^{\pi}(s) = R^{\pi}(s) + \sum_{s' \in S} P^{\pi}(s' \mid s, a)\, \gamma V^{\pi}(s')$$

直接報酬の期待値

遷移先での状態価値



【重要】ある状態の評価値は
直接報酬と他の状態の評価値
で与えられる＝**Bellmanの最適性原理**

MDPの割引報酬評価における最適性原理

「最適な政策」とは、一連の意思決定において、
各状態で常に最適な行動を選択すること
→　局所的な最適化の集まりによって全体が最適化、すなわち

最適な価値関数 $\mathbf{V}^* = \begin{bmatrix} V^*(s_1) \\ V^*(s_2) \\ \vdots \\ V^*(s_n) \end{bmatrix}$ とおくと、全ての状態 s について

以下のn個の非線形連立方程式が成り立つ

Bellman方程式

割引率

$$V^*(s) = \max_{a \in A} \sum_{s' \in S} P(s'|s,a)\left(R(s'|s,a) + \gamma V^*(s')\right)$$

状態遷移確率　直接報酬　遷移先での状態価値

MDPの割引報酬評価では、最適政策π*は　　が
最適価値関数V*は　　である

よって、この方程式を解いて最適価値関数を求めてから最適政策を得ればよい

MDPの割引報酬評価における最適性原理

「最適な政策」とは、一連の意思決定において、
各状態で常に最適な行動を選択すること
→　局所的な最適化の集まりによって全体が最適化、すなわち

最適な価値関数 $\mathbf{V}^* = \begin{bmatrix} V^*(s_1) \\ V^*(s_2) \\ \vdots \\ V^*(s_n) \end{bmatrix}$ とおくと、全ての状態 s について

以下のn個の非線形連立方程式が成り立つ

Bellman方程式

$$V^*(s) = \max_{a \in A} \sum_{s' \in S} P(s'|s,a)\left(R(s'|s,a) + \gamma V^*(s')\right)$$

割引率

状態遷移確率　　直接報酬　　遷移先での状態価値

MDPの割引報酬評価では、最適政策π*は複数存在できるが
最適価値関数V*はただ一つである

よって、この方程式を解いて最適価値関数を求めてから最適政策を得ればよい

Bellman方程式の意味は？

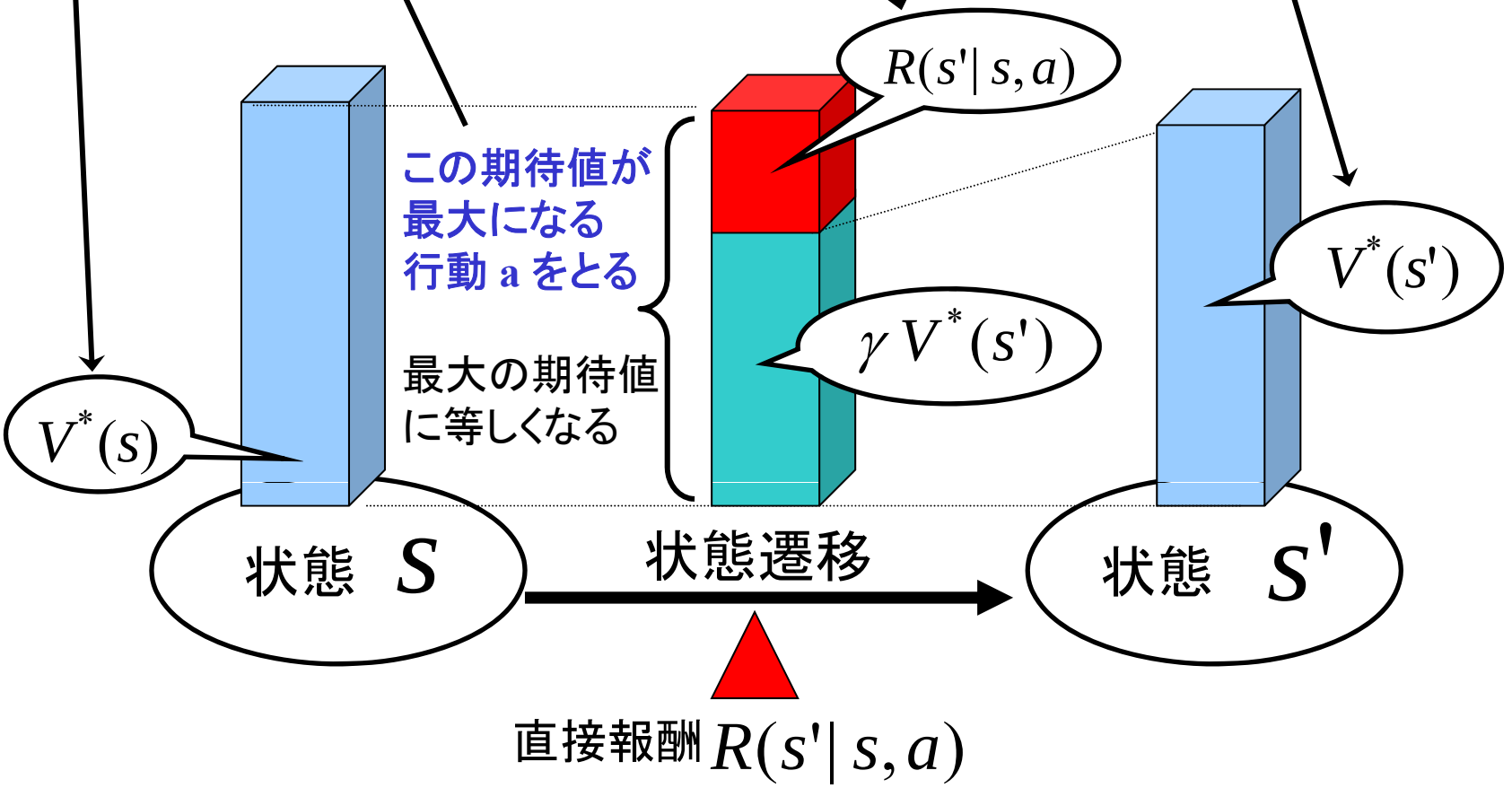
$$V^*(s) = \max_{a \in A} \sum_{s' \in S} P(s'|s,a)\left(R(s'|s,a) + \gamma V^*(s')\right)$$

割引率

状態遷移確率

直接報酬

遷移先での状態価値



ダイナミックプログラミングによるMDPの解法：

Bellman方程式は**非線形**のため、**単純な行列演算では解けない**

解法1 → Howard の政策反復法（Policy Iteration Method）

- 「**政策の改善**」と「**価値関数の計算**」を交互に繰返す
- 政策の改善ができなくなると、それが最適政策
  そのときの価値関数が最適価値関数

解法2 → 価値反復法（Value Iteration Method）

- 「**行動価値関数（Q関数）**」の非線形なBellman方程式を
  反復法（数値計算）によって直接解く
- 同じ状態で行動価値関数が最大の行動が最適行動
  →「最適行動価値関数」から最適行動を求める

詳細は次回

まとめ

１）「マルコフ決定過程」とは？
　　各状態で行動を選ぶと遷移確率が変わるマルコフ過程

２）「政策」とは？
　　各状態での行動の取り方

３）　政策を決めると、単なるマルコフ過程と同じ

４）　最適な政策は複数ありうるが、最適な価値関数は1つ

５）　最適価値関数は**Bellman方程式**が成り立つ

６）　最適政策や最適価値関数を求める方法は2種類

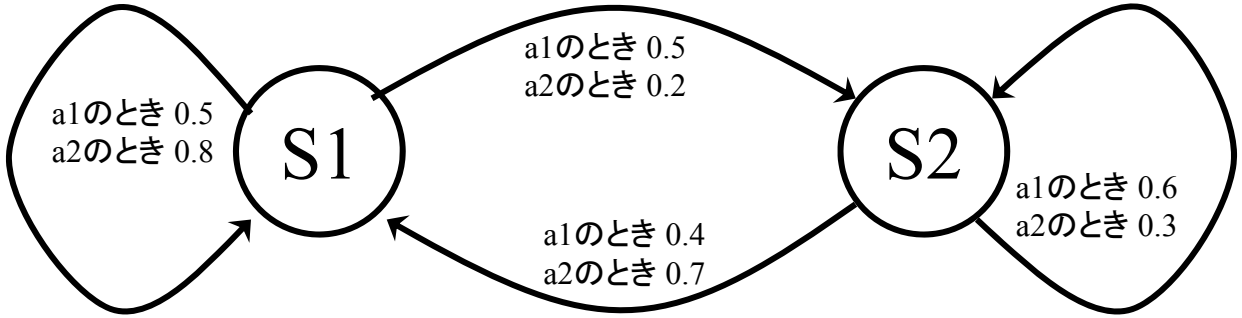
【演習問題】　2017.01.27　学籍番号＿＿＿＿＿＿＿＿＿＿氏名＿＿＿＿＿＿＿＿＿＿

例題の製造業者の意思決定問題について、割引率を γ とした割引報酬のBellman方程式を書け。
ただし $x < y$ のとき関数 $\max(x, y) = y$ と表されるものとする。

例題： 製品製造業者の意思決定問題

ある製品の製造業者が、市場における製品の人気の状態を観測し、各状態に応じて適切な決定を下す。

- 状態 S1：市場において製品の人気がある
- 状態 S2：市場において製品の人気がない
- 状態 S1 における行動 a1 ：広告をしない、行動 a2：広告を出す
- 状態 S2 における行動 a2：何もしない、行動 a2：新製品を開発
- 報酬：各期の売り上げから経費を差し引いた金額



状態 s	行動 a	推移確率		報酬		直接報酬の期待値
		$\Pr(s_1\|s,a)$	$\Pr(s_2\|s,a)$	$R^a(s,s_1)$	$R^a(s,s_2)$	
s_1：人気あり	a_1：広告なし	0.5	0.5	9	3	$6 = 0.5 \times 9 + 0.5 \times 3$
	a_2：広告あり	0.8	0.2	4	4	$4 = 0.8 \times 4 + 0.2 \times 4$
s_2：人気なし	a_1：研究なし	0.4	0.6	3	-7	$-3 = 0.4 \times 3 + 0.6 \times (-7)$
	a_2：研究あり	0.7	0.3	1	-19	$-5 = 0.7 \times 1 + 0.3 \times (-19)$

【演習問題】

学籍番号＿＿＿＿＿＿＿＿＿＿氏名＿＿＿＿＿＿＿＿＿＿

例題の製造業者の意思決定問題について、割引率をγとした割引報酬のBellman方程式を書け。
ただし $x < y$ のとき関数 $\max(x, y) = y$ と表されるものとする。

$$V^*(s) = \max_{a \in A} \sum_{s' \in S} P(s'|s,a)\left(R(s'|s,a) + \gamma V^*(s')\right)$$

状態は2つ　行動は2つ　状態遷移確率　直接報酬　割引率　遷移先での状態価値

$$\begin{cases} V^*(s_1) = \max_{a \in A} \sum_{s' \in S} P(s'|s_1,a)\left(R(s'|s_1,a) + \gamma V^*(s')\right) \\ V^*(s_2) = \max_{a \in A} \sum_{s' \in S} P(s'|s_2,a)\left(R(s'|s_2,a) + \gamma V^*(s')\right) \end{cases}$$

∑を展開　同時にmaxも展開

$$\begin{cases} V^*(s_1) = \max\Big(\big(P(s_1|s_1,a_1)\big(R(s_1|s_1,a_1) + \gamma V^*(s_1)\big) + P(s_2|s_1,a_1)\big(R(s_2|s_1,a_1) + \gamma V^*(s_2)\big)\big), \\ \qquad \big(P(s_1|s_1,a_2)\big(R(s_1|s_1,a_2) + \gamma V^*(s_1)\big) + P(s_2|s_1,a_2)\big(R(s_2|s_1,a_2) + \gamma V^*(s_2)\big)\big)\Big) \\ V^*(s_2) = \max\Big(\big(P(s_1|s_2,a_1)\big(R(s_1|s_2,a_1) + \gamma V^*(s_1)\big) + P(s_2|s_2,a_1)\big(R(s_2|s_2,a_1) + \gamma V^*(s_2)\big)\big), \\ \qquad \big(P(s_1|s_2,a_2)\big(R(s_1|s_2,a_2) + \gamma V^*(s_1)\big) + P(s_2|s_2,a_2)\big(R(s_2|s_2,a_2) + \gamma V^*(s_2)\big)\big)\Big) \end{cases}$$

ここで例題の推移確率と報酬の値を入れると…

$$\begin{cases} V^*(s_1) = \max\Big(\big(P(s_1 \mid s_1, a_1)\big(R(s_1 \mid s_1, a_1) + \gamma V^*(s_1)\big) + P(s_2 \mid s_1, a_1)\big(R(s_2 \mid s_1, a_1) + \gamma V^*(s_2)\big)\big), \\ \qquad \big(P(s_1 \mid s_1, a_2)\big(R(s_1 \mid s_1, a_2) + \gamma V^*(s_1)\big) + P(s_2 \mid s_1, a_2)\big(R(s_2 \mid s_1, a_2) + \gamma V^*(s_2)\big)\big)\Big) \\ V^*(s_2) = \max\Big(\big(P(s_1 \mid s_2, a_1)\big(R(s_1 \mid s_2, a_1) + \gamma V^*(s_1)\big) + P(s_2 \mid s_2, a_1)\big(R(s_2 \mid s_2, a_1) + \gamma V^*(s_2)\big)\big), \\ \qquad \big(P(s_1 \mid s_2, a_2)\big(R(s_1 \mid s_2, a_2) + \gamma V^*(s_1)\big) + P(s_2 \mid s_2, a_2)\big(R(s_2 \mid s_2, a_2) + \gamma V^*(s_2)\big)\big)\Big) \end{cases}$$

ここで例題の推移確率と報酬の値を入れると

$$\begin{cases} V^*(s_1) = \max\Big(\big(0.5\big(9 + \gamma V^*(s_1)\big) + 0.5\big(3 + \gamma V^*(s_2)\big)\big), \\ \qquad \big(0.8\big(4 + \gamma V^*(s_1)\big) + 0.2\big(4 + \gamma V^*(s_2)\big)\big)\Big) \\ V^*(s_2) = \max\Big(\big(0.4\big(3 + \gamma V^*(s_1)\big) + 0.6\big(-7 + \gamma V^*(s_2)\big)\big), \\ \qquad \big(0.7\big(1 + \gamma V^*(s_1)\big) + 0.3\big(-19 + \gamma V^*(s_2)\big)\big)\Big) \end{cases}$$

← **行動a1**をとった場合の状態価値

← **行動a2**をとった場合の状態価値

← **行動a1**をとった場合の状態価値

← **行動a2**をとった場合の状態価値