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）

（11） マルコフ過程：割引報酬による評価

場所：船1講義室

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

マルコフ過程の評価

平均報酬の期待値：

$$E\{\overline{R}\} = \mathbf{a}\,\mathbf{R}$$

極限（定常）状態分布　　報酬行列

1）平均報酬：吸収状態がない場合

●極限状態分布が評価値を支配

●評価値は、初期状態や現時点での状態には依存しない

→ ただ1つの［　　］で与えられる

報酬合計の期待値：

$$\mathbf{M}\,\mathbf{R}$$

吸収的マルコフ過程の基本行列　　報酬行列

2）報酬合計：吸収状態がある場合

●各状態からスタートして吸収状態に陥るまでの報酬合計の期待値

→ 初期状態に依存する

評価値は［　　］で与えられる

3）割引報酬合計：吸収状態の有無は関係ない

マルコフ過程の評価

平均報酬の期待値：

$$E\{\overline{R}\} = \mathbf{a}\,\mathbf{R}$$

極限（定常）状態分布　　報酬行列

1）平均報酬：吸収状態がない場合

●極限状態分布が評価値を支配

●評価値は、初期状態や現時点での状態には依存しない

→ ただ1つの スカラー値 で与えられる

報酬合計の期待値：

$$\mathbf{M}\,\mathbf{R}$$

吸収的マルコフ過程の基本行列　　報酬行列

2）報酬合計：吸収状態がある場合

●各状態からスタートして吸収状態に陥るまでの報酬合計の期待値

→ 初期状態に依存する

評価値は 状態数と同数の要素の行列 で与えられる

3）割引報酬合計：吸収状態の有無は関係ない

「割引報酬合計」による評価とは？

報酬合計（の期待値）を評価する

ただし，

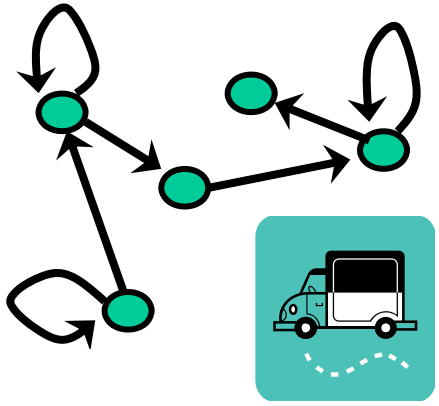
システムが**1－γの確率で停止する**場合を考える

 を導入

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



●吸収的マルコフ過程モデルでモデル化もできるが、
割引率γを導入したほうがモデル表現が単純

例）　同じ配送路を故障率の異なるトラックで配送する場合

→　配送路における遷移確率はそのままで、故障率1－γのみ変えて計算するだけで済む

「割引報酬合計」による評価とは？

報酬合計（の期待値）を評価する

ただし，

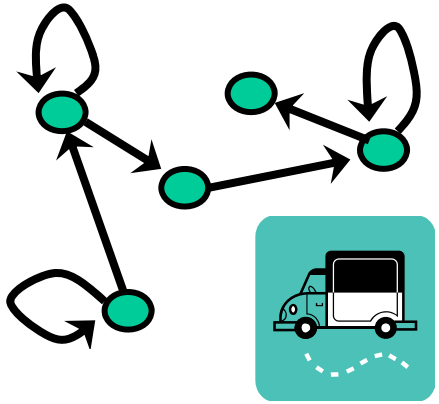
システムが**1－γの確率で停止する**場合を考える

割引率　γ を導入

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



●吸収的マルコフ過程モデルでモデル化もできるが、
割引率γを導入したほうがモデル表現が単純

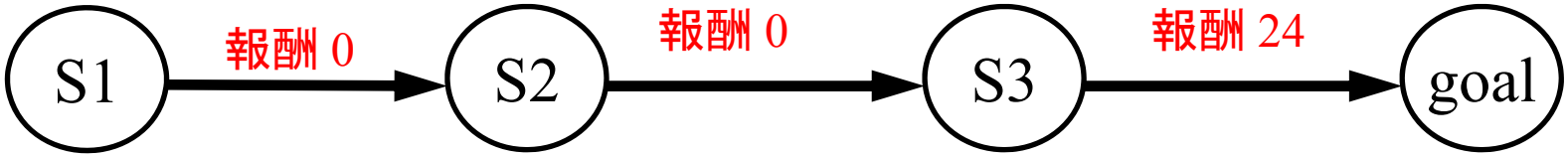
例）　同じ配送路を故障率の異なるトラックで配送する場合

→　配送路における遷移確率はそのままで、故障率1－γのみ変えて計算するだけで済む

【練習問題】

以下のプランA・Bの報酬合計の期待値を計算せよ。
ただし遷移に不確実性は無いが、遷移のたびに確率１－γ= 0.5 で
故障が発生し、この場合遷移先の状態で停止して動けなくなる。
＝　状態S1における割引率γ＝0.5 の割引報酬合計の期待値

【プランA】



報酬の平均および合計はABとも同じ

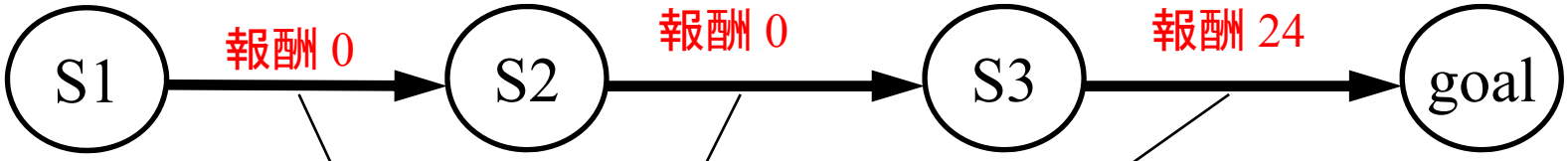
【プランB】

割引報酬では，プランBの方が高い評価値

【練習問題】

以下のプランA・Bの報酬合計の期待値を計算せよ。
ただし遷移に不確実性は無いが、遷移のたびに確率１－γ= 0.5 で
故障が発生し、この場合遷移先の状態で停止して動けなくなる。
＝　状態S1における割引率γ＝0.5 の割引報酬合計の期待値

【プランA】

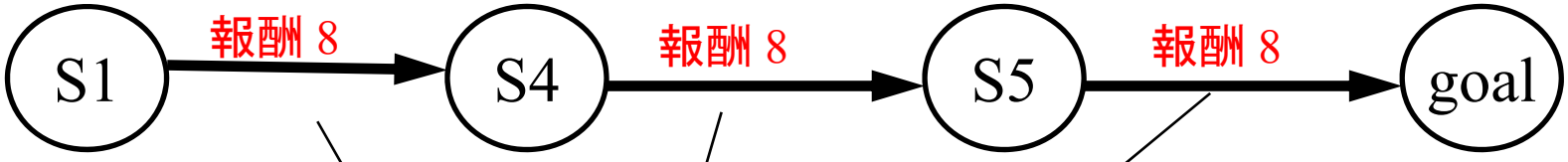


報酬の平均および合計はABとも同じ 24

$$V(S_1) = 0 + 0.5 \times 0 + 0.5^2 \times 24 = 6$$

割引報酬合計　　γ　　γ^2

【プランB】



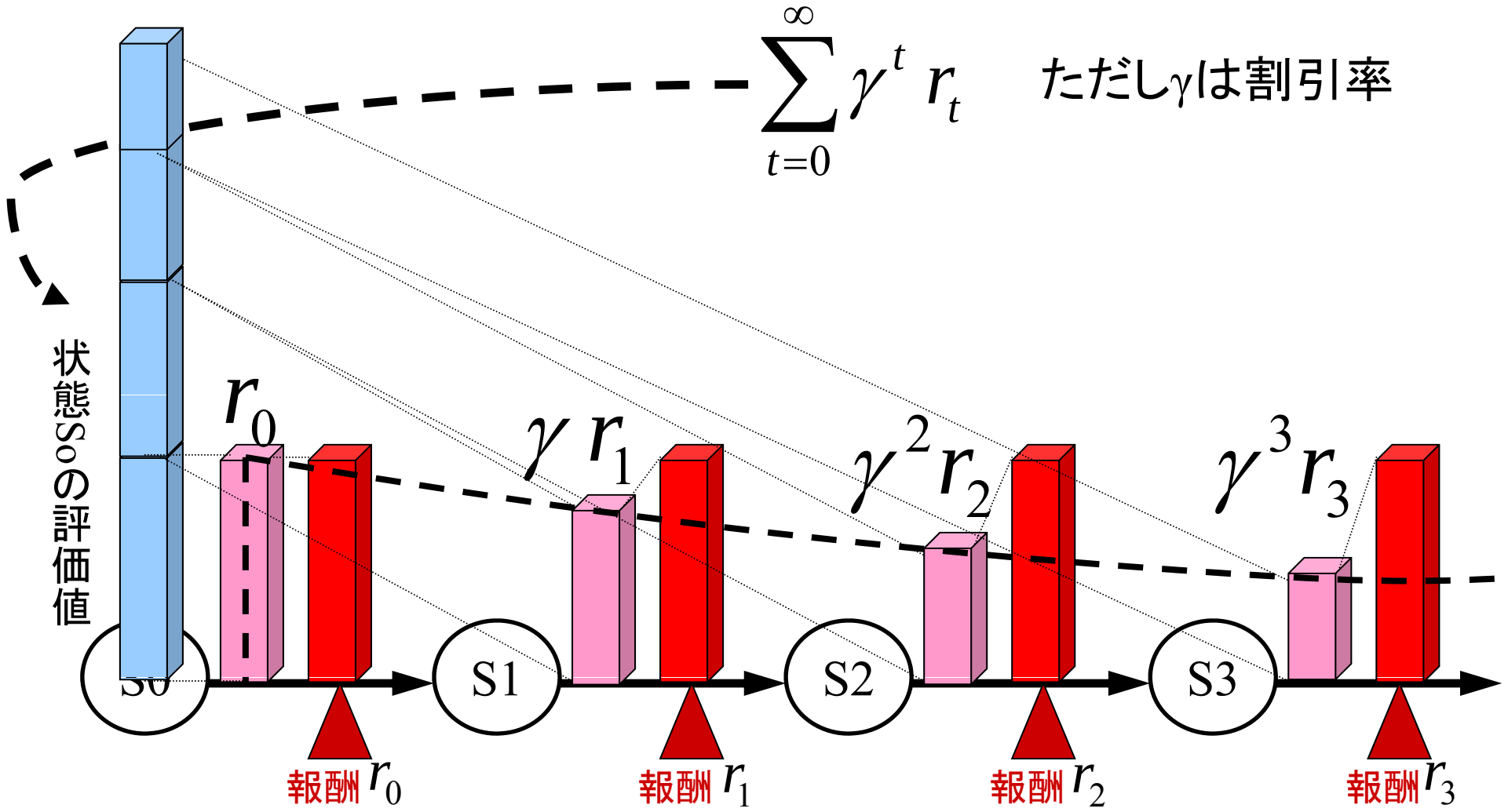
割引報酬では，プランBの方が高い評価値

$$V(S_1) = 8 + 0.5 \times 8 + 0.5^2 \times 8 = 14$$

割引報酬合計　　γ　　γ^2

「割引報酬合計」による評価

未来の報酬を割引いて足し合わせる
→ 現在の**状態の評価値**とする



報酬を割引く理由：

1）未来の報酬はあてにならない

（環境の変化などによるモデルの誤差があるため）

→ 未来に得る報酬を、遠い未来ほど割引いて評価

$\gamma = 1$のとき：　単なる報酬合計

$\gamma < 1$のとき：　$1-\gamma$の確率で停止する場合の報酬合計の期待値

2）計算の利便性

・モデルの吸収状態の有無に関係なく定義可能

・評価値に上界・下界が存在する（計算機への実装が容易）

・割引率$\gamma \to 1$に近づけると、

（割引報酬合計）×（$1-\gamma$）　は　□　に漸近

報酬を割引く理由：

1）未来の報酬はあてにならない

（環境の変化などによるモデルの誤差があるため）

→ 未来に得る報酬を、遠い未来ほど割引いて評価

$\gamma = 1$のとき：　単なる報酬合計

$\gamma < 1$のとき：　$1-\gamma$の確率で停止する場合の報酬合計の期待値

2）計算の利便性

・モデルの吸収状態の有無に関係なく定義可能

・評価値に上界・下界が存在する（計算機への実装が容易）

・割引率$\gamma \to 1$に近づけると、

（割引報酬合計）×（$1-\gamma$）　は　平均報酬　に漸近

1−γの確率で停止する場合の報酬合計評価

（割引報酬合計）の値は状態依存

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

←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

$$= \mathbf{R} + \gamma \mathbf{PR} + \gamma^2 \mathbf{P}^2\mathbf{R} + \cdots$$

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

$=$

$=$

大規模数値計算をする場合

ここに注目

各要素について式を書く

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

普通はこれを解く

1−γの確率で停止する場合の報酬合計評価

（割引報酬合計）の値は状態依存

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

←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

$$= \mathbf{R} + \gamma \mathbf{PR} + \gamma^2 \mathbf{P}^2\mathbf{R} + \cdots$$

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

$$= \mathbf{R} + \gamma \mathbf{PV}$$

大規模数値計算をする場合

ここに注目

各要素について式を書く

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

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

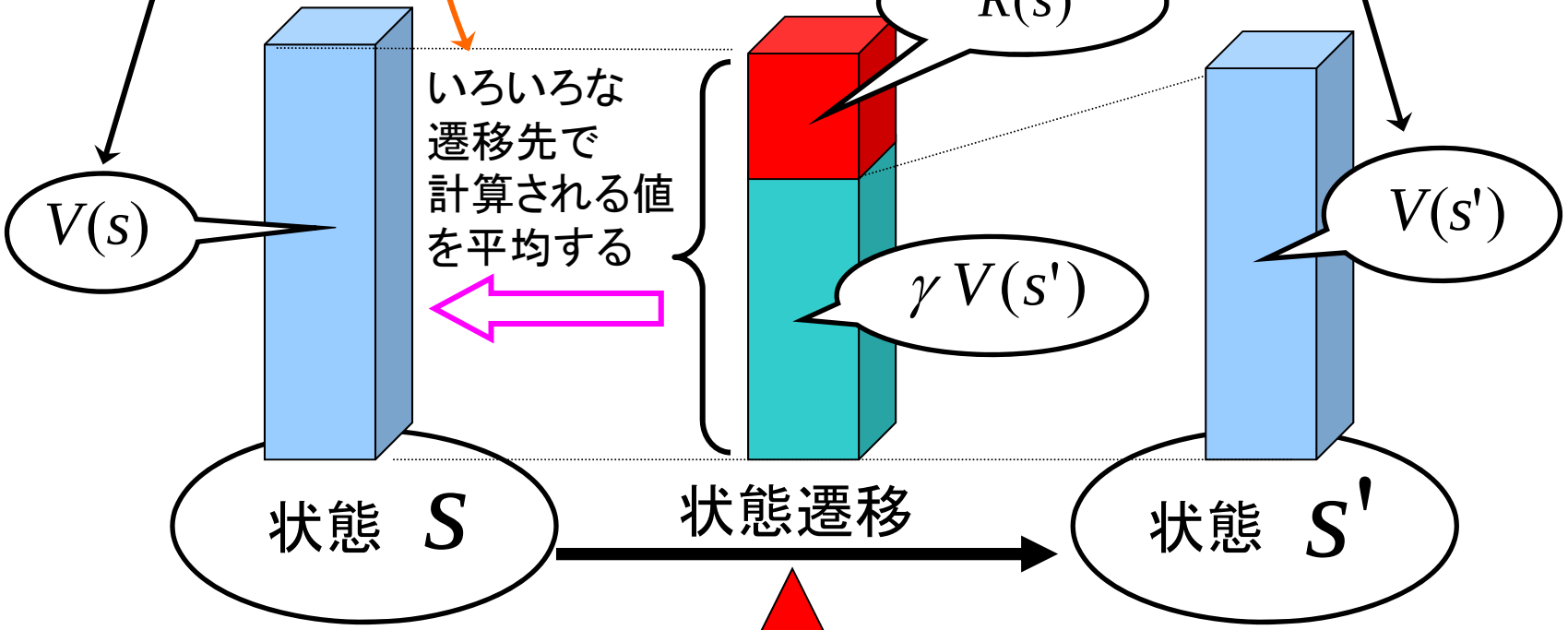
普通はこれを解く

イメージによる式の理解

状態遷移確率　　割引率

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

直接報酬の期待値　　遷移先での状態価値



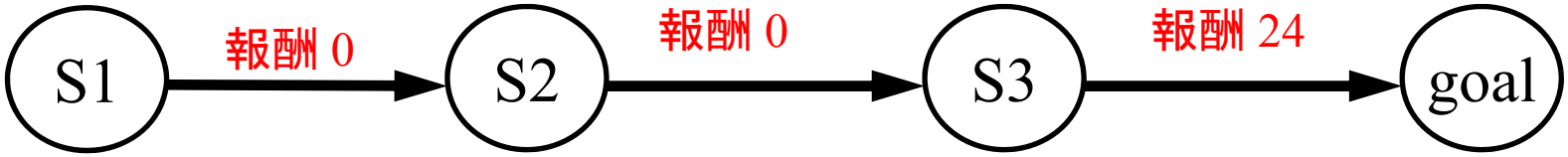
ある状態の割引報酬期待値は
直接報酬と他の状態の割引報酬
期待値で与えられる

報酬の期待値 $R(s)$

【練習問題】

以下のプランA・Bの報酬合計の期待値を計算せよ。
ただし遷移に不確実性は無いが、遷移のたびに確率１－γ= 0.5 で
故障が発生し、この場合遷移先の状態で停止して動けなくなる。
＝　状態S1における割引率γ＝0.5 の割引報酬合計の期待値

【プランA】



報酬の平均および合計はABとも同じ

$V(S_1) =$

$=$

$=$

$V(S_2) =$

【プランB】

割引報酬では，プランBの方が高い評価値

$V(S_1) =$

$=$

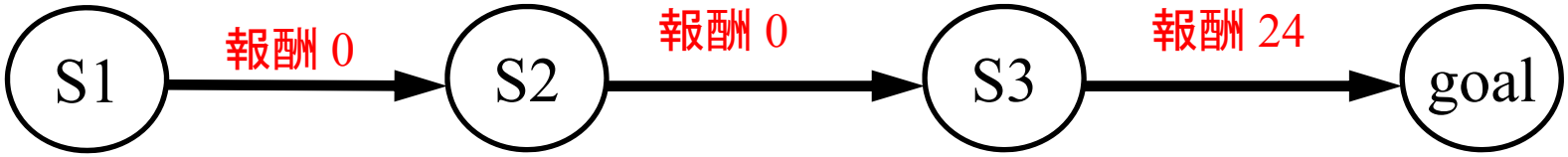
$=$

$V(S_4) =$

【練習問題】

以下のプランA・Bの報酬合計の期待値を計算せよ。
ただし遷移に不確実性は無いが、遷移のたびに確率１－γ= 0.5 で
故障が発生し、この場合遷移先の状態で停止して動けなくなる。
＝　状態S1における割引率γ＝0.5 の割引報酬合計の期待値

【プランA】



報酬の平均および合計はABとも同じ

$$V(S_1) = 0 + 0.5 \times 0 + 0.5^2 \times 24 = 0 + 0.5 \times V(S_2) = 6$$

$$V(S_2) = 0 + 0.5 \times 24 = 12$$

【プランB】

割引報酬では，プランBの方が高い評価値

$$V(S_1) = 8 + 0.5 \times 8 + 0.5^2 \times 8 = 8 + 0.5 \times V(S_4) = 14$$

$$V(S_4) = 8 + 0.5 \times 8 = 12$$

割引報酬合計期待値の解法(2):反復解法

求める未知数

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

(1) まず未知数Vの各要素に適当な値(ゼロなど)を入れて初期化
このときの値を $\mathbf{V}_0$ とおく。

$$\mathbf{V}_0 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

(2)  を計算

(3) 十分大きな n まで計算を繰り返し、収束したら打ち切る
このとき得られた $\mathbf{V}_n$ を答えとする。

0次　1次　2次

$$\mathbf{V} = \mathbf{R} + \gamma \mathbf{P}\mathbf{R} + \gamma^2 \mathbf{P}^2\mathbf{R} + \cdots$$

- この計算における $\mathbf{V}_n$ は、$\mathbf{V} = \mathbf{R} + \gamma \mathbf{P}\mathbf{R} + \gamma^2 \mathbf{P}^2\mathbf{R} + \cdots$ の n 次近似と等価
- **行列が大きくて sparse な(ゼロを多く含む)場合**、逆行列計算よりもこの反復解法のほうが計算コストが少なくて済む。

割引報酬合計期待値の解法(2):反復解法

求める未知数

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

(1) まず未知数Vの各要素に適当な値(ゼロなど)を入れて初期化
このときの値を $\mathbf{V}_0$ とおく。

$$\mathbf{V}_0 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

(2) $\mathbf{V}_{n+1} = \mathbf{R} + \gamma \mathbf{P}\mathbf{V}_n$ を計算

(3) 十分大きな n まで計算を繰り返し、収束したら打ち切る
このとき得られた $\mathbf{V}_n$ を答えとする。

- この計算における $\mathbf{V}_n$ は、$\mathbf{V} = \mathbf{R} + \gamma \mathbf{P}\mathbf{R} + \gamma^2 \mathbf{P}^2\mathbf{R} + \cdots$ の n 次近似と等価
- 行列が大きくて sparse な(ゼロを多く含む)場合、逆行列計算よりもこの反復解法のほうが計算コストが少なくて済む。

【参考】 連立方程式 $\mathbf{b} = \mathbf{A}\mathbf{x}$ の反復解法 （Jacobiの反復法）

$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$
を計算したいが
行列が大きすぎて
逆行列が計算
できない場合 →

$$\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{U}$ に分解すると、方程式は

$$\mathbf{D}\mathbf{x} = (\mathbf{L} + \mathbf{U})\mathbf{x} + \mathbf{b}$$ と書ける。 よって

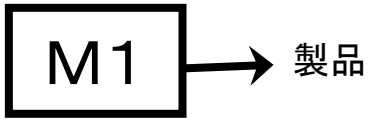
$$\mathbf{x} = \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\mathbf{x} + \mathbf{D}^{-1}\mathbf{b}$$

対角行列なので全対角要素を単なる逆数にするだけ

$$\mathbf{x}_{new} = \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\mathbf{x}_{old} + \mathbf{D}^{-1}\mathbf{b}$$

反復計算

【演習問題】2017.01.13



学籍番号
氏　名

●機械M1は「稼働」と「休止」の2状態を有する.
●機械の状態遷移は，一定時間間隔で離散的に起きる.
●機械が「稼動」状態にあるとき、次ステップでは確率 0.5 で稼動状態を継続
●機械が「休止」状態にあるとき、次ステップでは確率 0.4 で稼動状態へ復帰
●機械が「稼動」状態にあり、次のステップでも稼動しているとき報酬9を得る
●機械が「稼動」状態にあり、次のステップで休止するとき報酬3を得る
●機械が「休止」状態にあり、次のステップでも休止しているとき報酬－7を得る
●機械が「休止」状態にあり、次のステップで稼動するとき報酬3を得る

「稼動」状態をS1,「休止」状態をS2として以下の問に答えよ:

【1】 状態遷移行列と報酬行列を求めよ。

分数のままで計算せよ

【2】 定常分布および平均報酬を求めよ。

計算結果が分数の場合、約分して解答せよ

【3】 割引率$\gamma=0.8$ のときの割引報酬の期待値を求めよ。

2x2行列の逆行列の公式

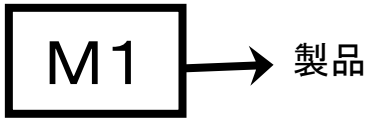
$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

単位行列

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1}\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

【演習問題】



学籍番号
氏　名

●機械M1は「稼働」と「休止」の2状態を有する.
●機械の状態遷移は，一定時間間隔で離散的に起きる.
●機械が「稼動」状態にあるとき、次ステップでは確率 0.5 で稼動状態を継続
●機械が「休止」状態にあるとき、次ステップでは確率 0.4 で稼動状態へ復帰
●機械が「稼動」状態にあり、次のステップでも稼動しているとき報酬9を得る
●機械が「稼動」状態にあり、次のステップで休止するとき報酬3を得る
●機械が「休止」状態にあり、次のステップでも休止しているとき報酬－7を得る
●機械が「休止」状態にあり、次のステップで稼動するとき報酬3を得る

「稼動」状態をS1,「休止」状態をS2として以下の問に答えよ:

【1】 状態遷移行列と報酬行列を求めよ。

$$\mathbf{P}=\begin{bmatrix} P(S_1\mid S_1) & P(S_2\mid S_1)\\ P(S_1\mid S_2) & P(S_2\mid S_2)\end{bmatrix}=\begin{bmatrix} 0.5 & 0.5\\ 0.4 & 0.6\end{bmatrix}$$

$$\mathbf{R}=\begin{bmatrix} R(S_1)\\ R(S_2)\end{bmatrix}=\begin{bmatrix} 0.5\times 9+0.5\times 3\\ 0.4\times 3+0.6\times(-7)\end{bmatrix}=\begin{bmatrix} 6\\ -3\end{bmatrix}$$

【2】 定常分布および平均報酬を求めよ。

a1, a2をそれぞれ定常分布における状態1，状態2の確率とすると、

$\mathbf{aP}=\mathbf{a}$ より

$$\begin{bmatrix} a_1 & a_2\end{bmatrix}\begin{bmatrix} 0.5 & 0.5\\ 0.4 & 0.6\end{bmatrix}=\begin{bmatrix} a_1 & a_2\end{bmatrix}$$

よって $\mathbf{a}=\begin{bmatrix} a_1 & a_2\end{bmatrix}=\begin{bmatrix} \frac{4}{9} & \frac{5}{9}\end{bmatrix}$

平均報酬は

$$\mathbf{aR}=\begin{bmatrix} \frac{4}{9} & \frac{5}{9}\end{bmatrix}\begin{bmatrix} 6\\ -3\end{bmatrix}=1$$

【3】 割引率$\gamma=0.8$ のときの割引報酬の期待値を求めよ。

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

$$=\left(\begin{bmatrix} 1 & 0\\ 0 & 1\end{bmatrix}-0.8\begin{bmatrix} 0.5 & 0.5\\ 0.4 & 0.6\end{bmatrix}\right)^{-1}\begin{bmatrix} 6\\ -3\end{bmatrix}=\begin{bmatrix} \frac{65}{23} & \frac{50}{23}\\ \frac{40}{23} & \frac{75}{23}\end{bmatrix}\begin{bmatrix} 6\\ -3\end{bmatrix}=\begin{bmatrix} \frac{240}{23}\\ \frac{15}{23}\end{bmatrix}=\begin{bmatrix} 10.434\\ 0.6521\end{bmatrix}$$