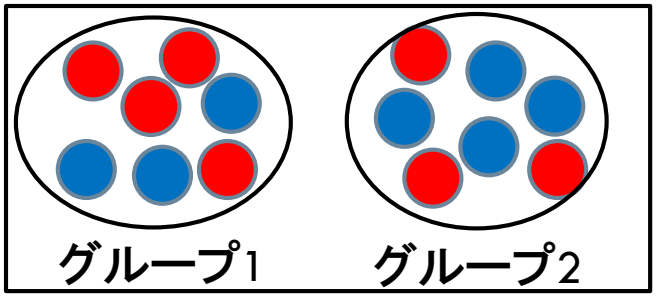
人工知能　第6回

強化学習入門

2018年5月25日　八谷　大岳

前回の課題1



箱

標本空間S

確率変数X（ボールの色）

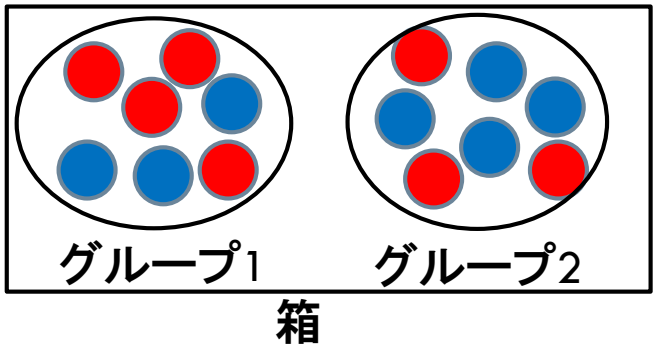
確率変数Y（グループの種類）	x_1：青	x_2：赤
y_1：グループ1	3	4
y_2：グループ2	4	3

確率変数が2つ

1. **条件付き確率$P_{X|Y}(X = x_2|Y = y_2)$を求めなさい**
2. **周辺確率$P_Y(Y = y_2)$を求めなさい**
3. **乗法定理の定義式を書き、同時確率$P_{Y|X}(Y = y_2, X = x_2)$を求めなさい**
4. 周辺確率の定義から、以下の加法定理を導出しなさい

$$P_X(x_j) = \sum_i P_{YX}(y_i, x_j)$$

課題2の解答例



標本空間S

確率変数X（ボールの色）

確率変数Y（グループの種類）	x_1：青	x_2：赤
y_1：グループ1	3	4
y_2：グループ2	4	3

1. **条件付き確率$P_{X|Y}(X = x_2|Y = y_2)$を求めなさい**

$$P_{X|Y}(X = x_2|Y = y_2) = P_{X|Y}(X = \text{赤}|Y = \text{グループ}2) = \frac{N_{22}}{N_{2\cdot}} = \frac{3}{7}$$

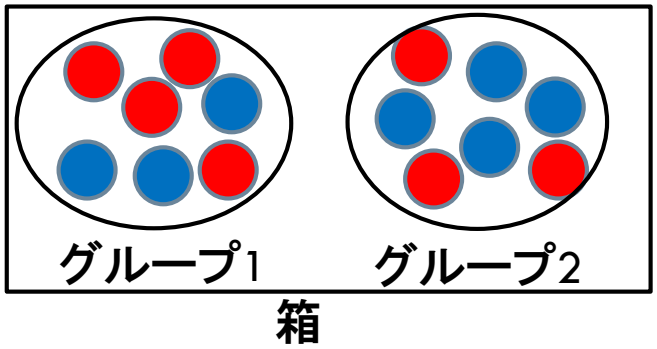
2. **周辺確率$P_Y(Y = y_2)$を求めなさい**

$$P_Y(Y = y_2) = P_Y(Y = \text{グループ}2) = \frac{N_{2\cdot}}{N} = \frac{7}{14}$$

3. **乗法定理の定義式を書き、同時確率$P_{Y|X}(Y = y_2, X = x_2)$を求めなさい**

$$P_{YX}(Y = y_2, X = x_2) = P_{X|Y}(X = x_2|Y = y_2)P_Y(Y = y_2) = \frac{3}{7}\frac{7}{14} = \frac{3}{14}$$

課題2の解答例　続き



標本空間S

確率変数X（ボールの色）

確率変数Y（グループの種類）	x_1：青	x_2：赤
y_1：グループ1	3	4
y_2：グループ2	4	3

$$N_{.2} = N_{12} + N_{22} = 4 + 3 = 7$$

4. 周辺確率の定義から、以下の**加法定理**を導出しなさい

列の場合の数の和

$$P_X(x_j) = \frac{N_{.j}}{N} = \frac{\sum_i N_{ij}}{N} = \sum_i \frac{N_{ij}}{N} = \sum_i P_{YX}(y_i, x_j)$$

前回の課題2

和の記号: $\sum_{t=0}^{2} R(X^t) = R(X^0) + R(X^1) + R(X^2)$

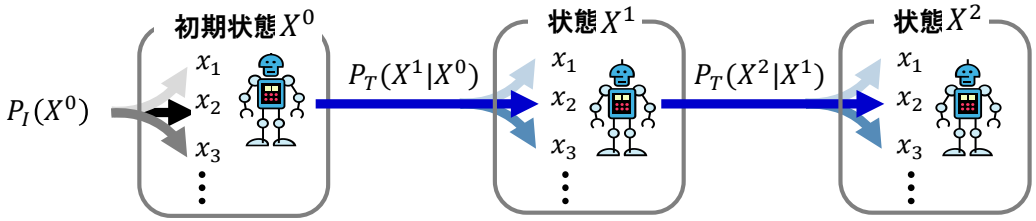
- 以下の等式が成り立つことを証明しなさい。

$$\mathop{\mathrm{E}}_{\substack{P_I(X^0)\\P_T(X^{t+1}|X^t)}}\left[\sum_{t=0}^{2} R(X^t)\right] = \mathop{\mathrm{E}}_{P_I(X^0)}\left[R(X^0) + \mathop{\mathrm{E}}_{P_T(X^{t+1}|X^t)}\left[\sum_{t=1}^{2} R(X^t)\right]\right]$$

  - ただし、確率変数は、各時刻$t = 0,1,2$ごとに存在することとし、それぞれ以下の確率分布関数に従うこととする。

$X^0 : P_I(X^0)$　　$X^1 : P_T(X^1|X^0)$　　$X^2 : P_T(X^2|X^1)$

$$\mathop{\mathrm{E}}_{\substack{P_I(X^0)\\P_T(X^{t+1}|X^t)}}\left[\sum_{t=0}^{2} R(X^t)\right] = \mathop{\mathrm{E}}_{\substack{P_I(X^0)\\P_T(X^1|X^0)\\P_T(X^2|X^1)}}\left[\sum_{t=0}^{2} R(X^t)\right]$$



$P_I(X^0)$：初期確率分布関数

$P_T(X^{t+1}|X^t)$：遷移確率関数

課題2の解答例

和の記号: $\sum_{t=0}^{2} R(X^t) = R(X^0) + R(X^1) + R(X^2)$

$$\mathop{\mathrm{E}}_{\substack{P_I(X^0)\\P_T(X^{t+1}|X^t)}}\left[\sum_{t=0}^{2} R(X^t)\right] = \mathop{\mathrm{E}}_{\substack{P_I(X^0)\\P_T(X^1|X^0)\\P_T(X^2|X^1)}}[R(X^0) + R(X^1) + R(X^2)]$$

期待値を和に変形

$$= \sum_{X^0}\sum_{X^1}\sum_{X^2}[R(X^0) + R(X^1) + R(X^2)]\, P_I(X^0)P_T(X^1|X^0)P_T(X^2|X^1)$$

$\sum_{X^1}\sum_{X^2}$ を移動

$$= \sum_{X^0}\Big[R(X^0) + \sum_{X^1}\sum_{X^2}[R(X^1) + R(X^2)]\, P_T(X^1|X^0)P_T(X^2|X^1)\Big]\, P_I(X^0)$$

$R(X^1) + R(X^2)$を$\sum_{t=1}^{2} R(X^t)$に変形

$$= \sum_{X^0}\Big[R(X^0) + \sum_{X^1}\sum_{X^2}\Big[\sum_{t=1}^{2} R(X^t)\Big]\, P_T(X^1|X^0)P_T(X^2|X^1)\Big]\, P_I(X^0)$$

$$= \mathop{\mathrm{E}}_{P_I(X^0)}\left[R(X^0) + \mathop{\mathrm{E}}_{\substack{P_T(X^1|X^0)\\P_T(X^2|X^1)}}\left[\sum_{t=1}^{2} R(X^t)\right]\right]$$

$$= \mathop{\mathrm{E}}_{P_I(X^0)}\left[R(X^0) + \mathop{\mathrm{E}}_{P_T(X^{t+1}|X^t)}\left[\sum_{t=1}^{2} R(X^t)\right]\right]$$

講義内容

回	内容
1	知識処理
2	計算モデル
3	探索モデル
4	判別モデル
5	機械学習の基礎
6	強化学習
7	強化学習演習
8	ニューラルネットワーク
9	深層学習
10	プランニング
11	推論
12	推論演習
13	知識表現
14	グループ討論
15	総合演習

機械学習のアルゴリズム

内容

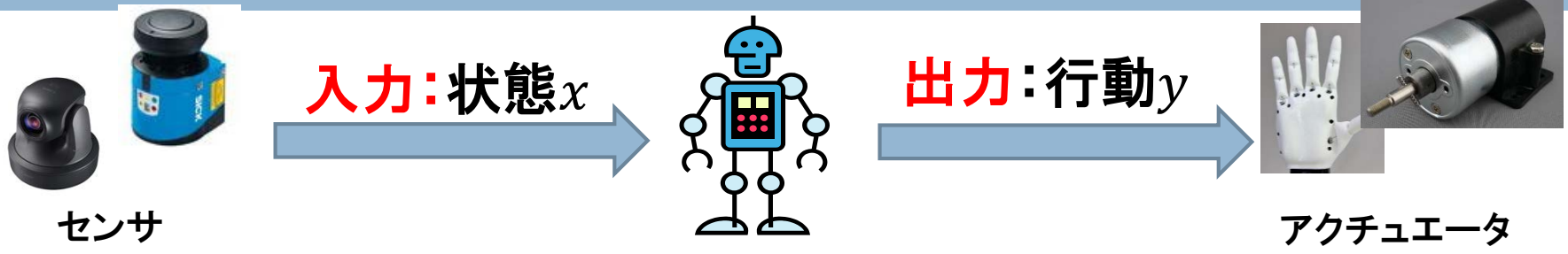
- 教師あり学習
  - 教師あり学習の例と問題点
- 強化学習
  - 強化学習の目的と成功例
  - 動物の行動学習
  - 強化学習の定式化
  - 動的計画法による最適化

機械学習手法の種類

問題	定義	代表的な方法	応用例
教師あり学習	入力と出力のデータに基づき、入力を出力に変換する関数を学習	SVM, 最小二乗法、決定木、ランダムフォーレストなど	スパム分類、顔検出、一般物体認識、将棋局面の評価など
教師なし学習	入力のみの事例に基づき、入力の特性 （パターン、構造）を学習	PCA, LDA、HMMなど	データの可視化（クラスタリング、次元圧縮）
半教師学習	入力のうち部分的に付与された出力の事例に基づき入力を出力に変換する関数を学習	transductiveSVM, Laplacian SVMなど	画像、音声、Webログなどの大量データで、コストの問題で一部のデータのみしか出力（答え）が付与されていない場合
強化学習	入力と、出力に対する報酬（評価）のデータに基づき、入力を出力に変換する関数を学習	Q-learning、policy iteration, policy gradient	ロボット制御、Web広告選択、マーケティング

- 問題（データの条件など）と目的に合わせて適切な機械学習方法を選択する必要がある

ロボット制御での教師あり学習



- **方策関数（policy）**：状態xを行動yに変換する関数

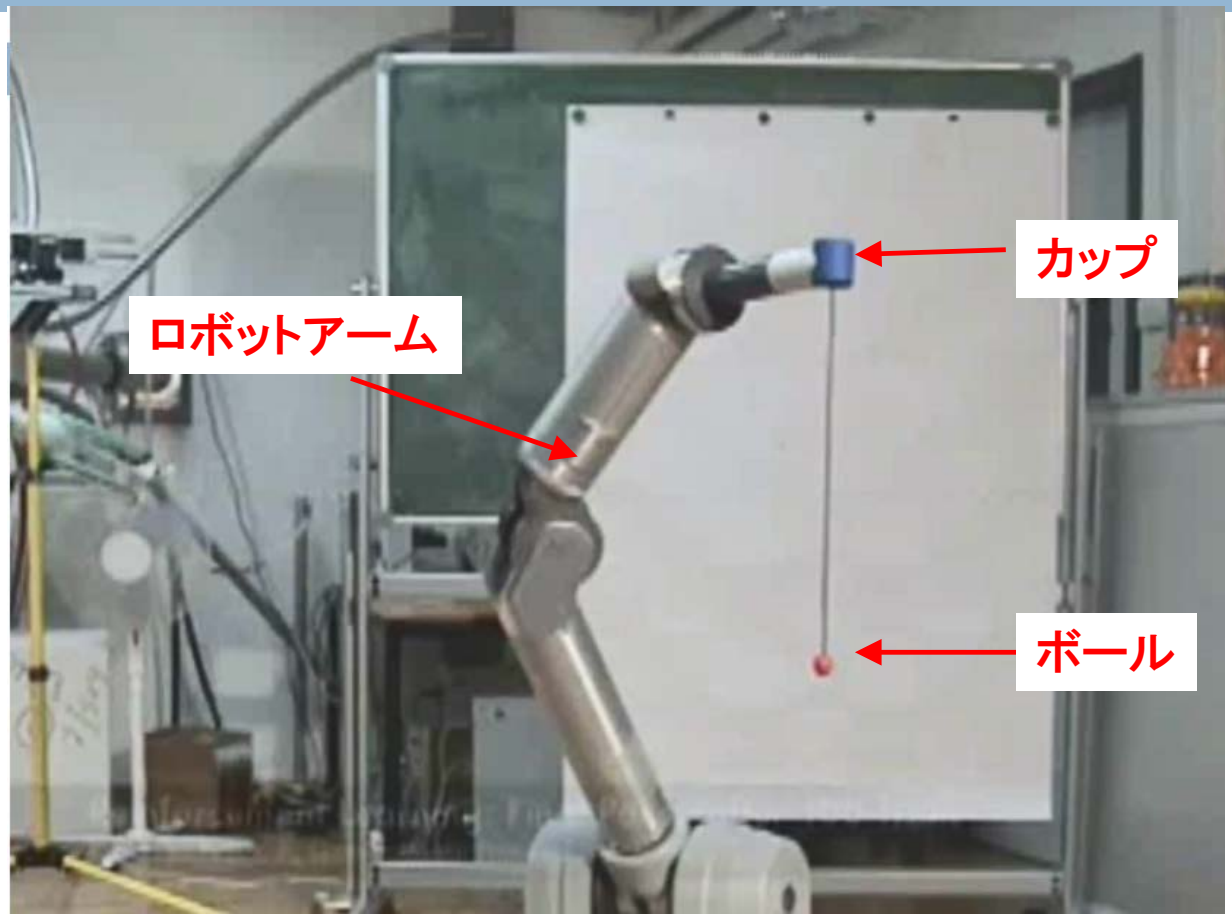
$$y = \pi(x)$$

- **教師あり学習**：各状態xで**取るべき正しい行動**yのデータを用意

$$(x^0, y^0), (x^1, y^1), \ldots, (x^N, y^N) = \{(x^t, y^t)\}_{t=0}^{N}$$

- **目標**：各状態で取るべき正しい行動を再現する**方策**$\pi(x)$を獲得

制御の例：Ball in a cup（けん玉）



- **目標**：ロボットにボールをカップの中に入れる方策$\pi(s)$を獲得

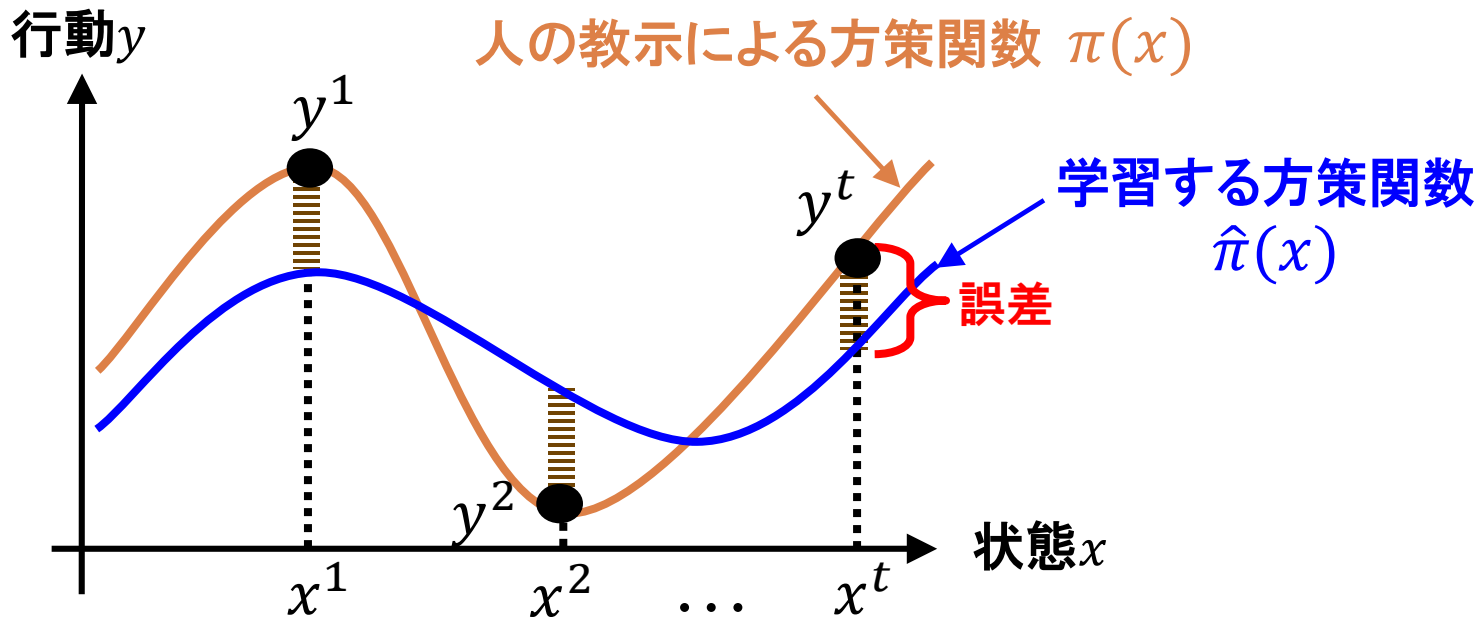
教師データの収集

ドイツ・マックスプランク研究所

- 状態と行動の組のデータ：$\{(x^t, y^t)\}_{t=0}^{N}$を収集
- 状態x：各関節の角度・角速度、 行動y：各関節に加えるトルク量

教師あり学習（回帰）

学習データ:$\{(x^t, x^t)\}_{t=1}^{N}$



- 学習データを用いて**二乗誤差和を最小化**する方策関数$\hat{\pi}(s)$を学習

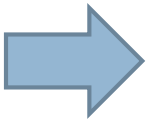
$$\min \sum_{t=1}^{N} (y^t - \hat{\pi}(x^t))^2$$

教師あり学習の結果の例

- 動作は似ているが、ボールをカップに入れることができない

教師あり学習の問題点

- 学習に十分な教師データを収集するのは困難
  - 教師データ数は最低でも学習パラメータの十数倍は必要
  - 教師データを用意するのは人なので、時間および人件費が大
- 教師データが不正確（特にロボットでは）
  - 人間とロボットでは、骨格、筋肉配置が異なるため、ロボットにとって最適なダイナミックな行動（関節のトルク）を教えるのが困難
- 人間の知識の限界
  - 人間が必ずしも最適な方法を知っているとは限らない



学習した方策関数：
良くて人間の模倣、コンピュータにとって最適とは限らない

内容

- 教師あり学習
  - 教師あり学習の例と問題点
- 強化学習
  - 強化学習の目的と成功例
  - 動物の行動学習
  - 強化学習の定式化
  - 動的計画法による最適化

機械学習手法の種類

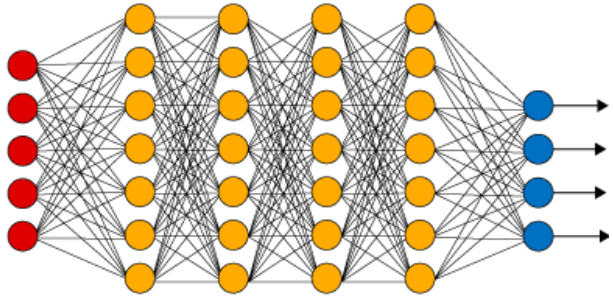
問題	定義	代表的な方法	応用例
教師あり学習	入力と出力のデータに基づき、入力を出力に変換する関数を学習	SVM, 最小二乗法、決定木、ランダムフォーレストなど	スパム分類、顔検出、一般物体認識、将棋局面の評価など
教師なし学習	入力のみの事例に基づき、入力の特性 （パターン、構造）を学習	PCA, LDA、HMMなど	データの可視化（クラスタリング、次元圧縮）
半教師学習	入力のうち部分的に付与された出力の事例に基づき入力を出力に変換する関数を学習	transductiveSVM, Laplacian SVMなど	画像、音声、Webログなどの大量データで、コストの問題で一部のデータのみしか出力（答え）が付与されていない場合
強化学習	入力と、出力に対する報酬（評価）のデータに基づき、入力を出力に変換する関数を学習	Q-learning、policy iteration, policy gradient	ロボット制御、Web広告選択、マーケティング

- 問題（データの条件など）と目的に合わせて適切な機械学習方法を選択する必要がある

強化学習の目標

- 人間が十分なデータを用意できない問題を解決
  - ロボット制御
  - Web広告の選択

- 人間を超える新しい戦略の発見
  - 商品の販売戦略
  - 新薬の開発
  - 機械学習アルゴリズムの開発
  - 将棋、囲碁、ゲームの戦略



強化学習のゲームにおける成功例

- 将棋：2013年の電王戦（現役プロとコンピュータとの対局）にて**コンピュータが3勝1敗1分けで勝ち越し**

- 囲碁：2016年3月、Google DeepMind社の「Alpha Go」が世界No.2の韓国の李セドル氏に**4勝1敗の大差で勝利**

  - **強化学習アルゴリズム同士の対戦を繰り返すことにより、人間が数千年かけても見つけられなかった「定石」を発見**

動画：https://newswebeasy.github.io/ja201710/news/web/movie/2017/10/19/k10011182291_201710190529_201710190530.mp4

内容

- 教師あり学習
  - 教師あり学習の例と問題点
- 強化学習
  - 強化学習の目的と成功例
  - 動物の行動学習
  - 強化学習の定式化
  - 動的計画法による最適化

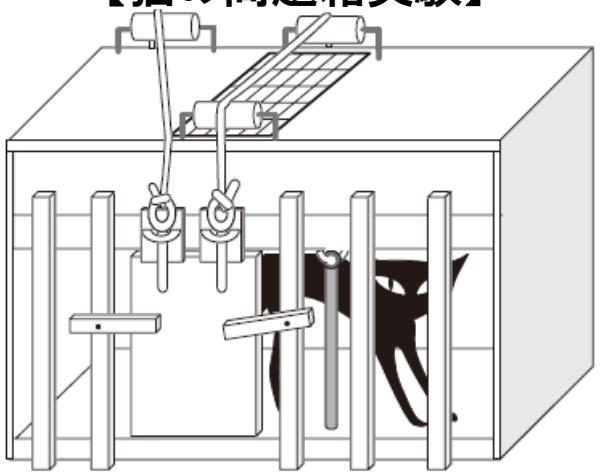
動物の行動学習：試行錯誤学習

- 心理学の行動主義者の学習の定義
  - 行動や反応の変化として表れ、外部から観察できる現象

- 試行錯誤学習

Edward Thorndike（1874年–1949年）
迷路のような箱の中で，猫は試行錯誤的に様々な行動を取る。
偶然にでも外に出る行動を取り，それを何度か繰り返すと，
やがて同じ行動が出現する頻度が高くなることを実験的に確認

【猫の問題箱実験】



猫が様々な行動をとる → 偶然に外に出られた → また箱にいれる

試行錯誤の過程を経て何度か繰り返すと、やがて同じ行動が出現する頻度が高くなる

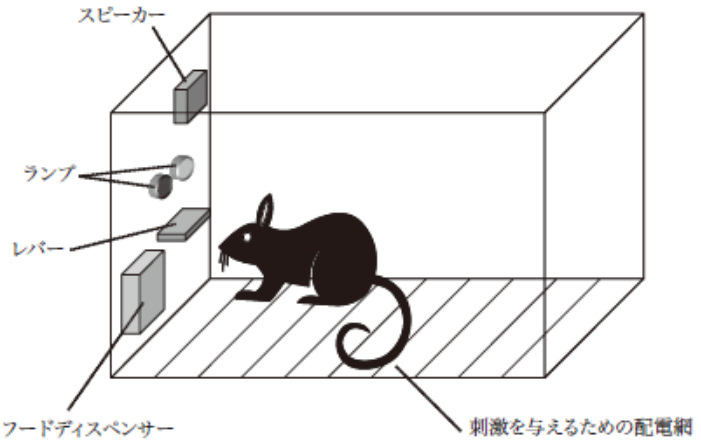
行動後の「満足」または「不快」の度合いに応じて、
行動の出現頻度が学習の過程を経て変化

動物の行動学習：報酬学習

□ 報酬学習

Burrhus Frederic Skinner（1904年–1990年）
レバーを押すと餌が出る仕組みになっている箱の中で，ラットが偶然にでもレバーを押し，餌を得ることを何度か繰り返すと，ラットはレバーの近くにいることが多くなり，やがてレバーを押す行動を取る頻度が高くなるのを実験的に確認

【スキナーの箱実験】



ラットが様々な行動をとる → レバーを押す → 餌が出る

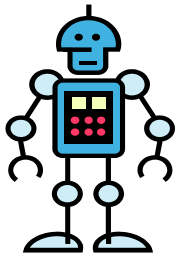
何度か繰り返すと、レバーを押す頻度が高くなる

満足を「餌」という「報酬」で明示的に与えた「試行錯誤学習」。
報酬に応じて行動の自発頻度が変化していく：報酬学習

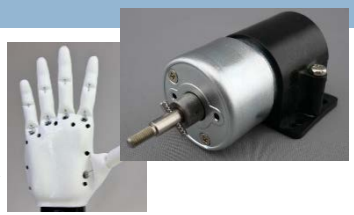
ロボット制御での強化学習

センサ

入力：状態x



出力：行動y



アクチュエータ

環境

報酬 r

- **試行錯誤学習**の要素：**方策関数を条件付き確率分布関数に拡張**

$$y = \pi(x) \quad \Rightarrow \quad \pi(Y = y|X = x)$$

条件付き確率に従いランダムに行動を選択

- **報酬学習**の要素：**報酬rの総和を最大化する方策を獲得**
  - 報酬関数$R(x, y, x')$は人間が設計：**教師データを集めるより断然簡単**

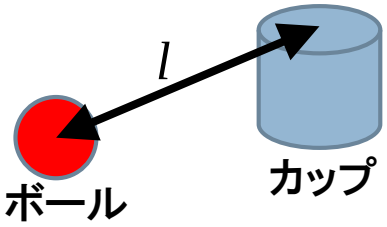
$$r = R(x, y, x')$$

x'：次の状態

強化学習の結果の例

報酬の例：

$$r = \exp(-l)$$



l :ボールとカップの距離

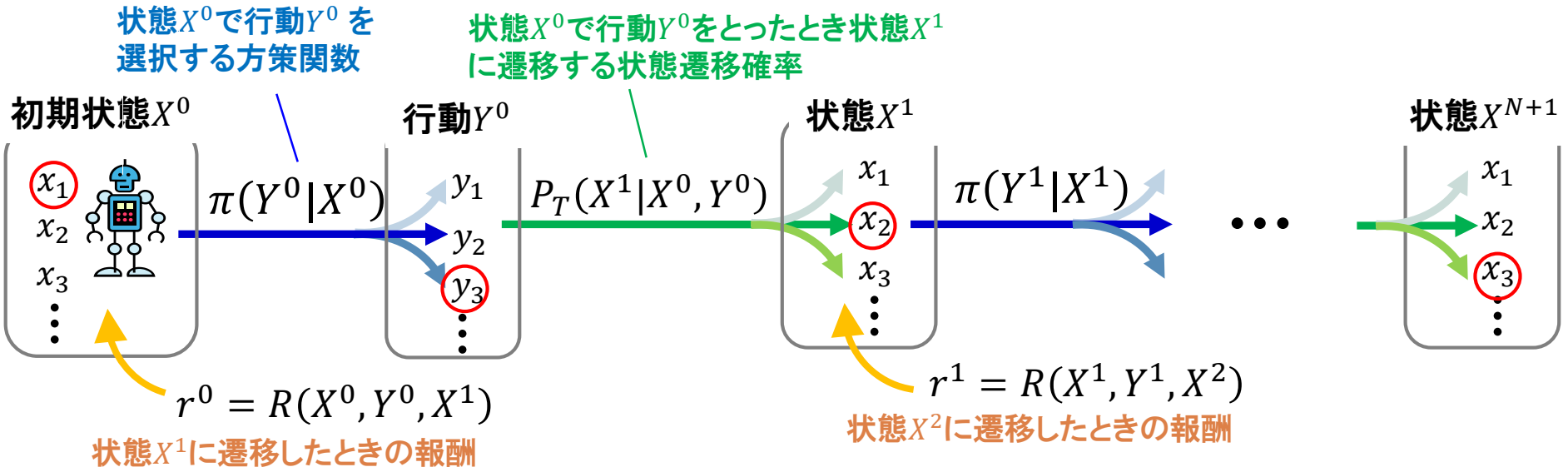
- 方策は報酬に基づき試行錯誤的に改善され、ボールをカップに入れることができるようになる。

内容

- 教師あり学習
  - 教師あり学習の例と問題点
- 強化学習
  - 強化学習の目的と成功例
  - 動物の行動学習
  - 強化学習の定式化
  - 動的計画法による最適化

強化学習の定式化

- $\{\pi, R, P_T\}$に従う確率過程(**マルコフ決定過程**)を考える



- **価値関数**:ステップτから始まる**報酬和の期待値(平均)**

$$V_\pi^\tau(x) = \mathop{\mathrm{E}}_{\substack{\pi(Y^t|X^t) \\ P_T(X^{t+1}|X^t,Y^t)}} \left[\sum_{t=\tau}^{N} \gamma^{t-\tau} R(X^t, Y^t, X^{t+1}) \middle| X^\tau = x \right]$$

$\gamma \in (0,1]$:**割引率**

遠い先の報酬ほど割り引く

価値関数の例

- 各ステップの価値関数

ステップ0:
$$V_\pi^0(x) = \mathop{\mathrm{E}}_{\substack{\pi(Y^t|X^t) \\ P_T(X^{t+1}|X^t,Y^t)}} \left[\sum_{t=0}^{N} \gamma^{t-0} R(X^t, Y^t, X^{t+1}) \middle| X^0 = x \right]$$

ステップ1:
$$V_\pi^1(x) = \mathop{\mathrm{E}}_{\substack{\pi(Y^t|X^t) \\ P_T(X^{t+1}|X^t,Y^t)}} \left[\sum_{t=1}^{N} \gamma^{t-1} R(X^t, Y^t, X^{t+1}) \middle| X^1 = x \right]$$

ステップ2:
$$V_\pi^2(x) = \mathop{\mathrm{E}}_{\substack{\pi(Y^t|X^t) \\ P_T(X^{t+1}|X^t,Y^t)}} \left[\sum_{t=2}^{N} \gamma^{t-2} R(X^t, Y^t, X^{t+1}) \middle| X^2 = x \right]$$

⋮

- 開始ステップが異なるが関数の形は同じ

価値関数の漸化式表現

- 価値関数は**漸化式**で表現可能

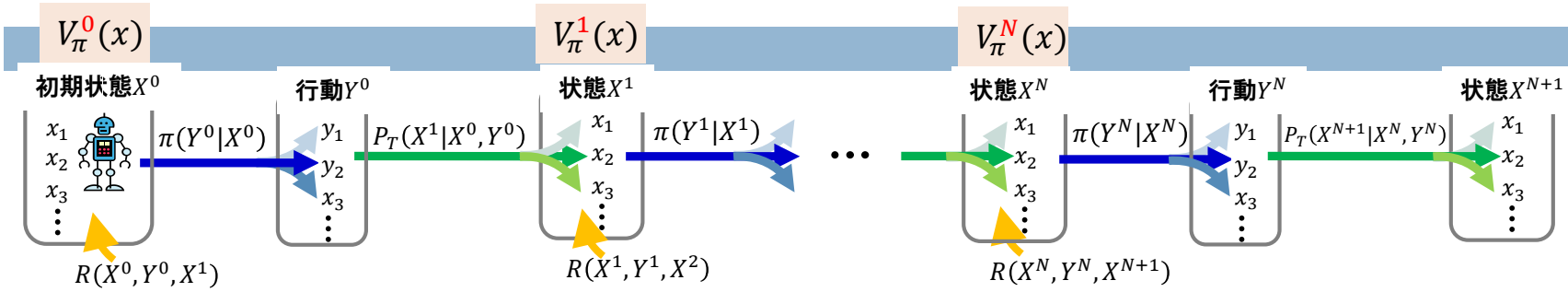
$$V^0(x) = \mathop{\mathrm{E}}_{\substack{\pi(Y^t|X^t) \\ P_T(X^{t+1}|X^t,Y^t)}} \left[\sum_{t=0}^{N} \gamma^t R(X^t, Y^t, X^{t+1}) \middle| X^0 = x \right]$$

$$= \mathop{\mathrm{E}}_{\substack{\pi(Y^0|X^0) \\ P_T(X^1|X^0,Y^0)}} \left[R(X^0, Y^0, X^1) + \mathop{\mathrm{E}}_{\substack{\pi(Y^t|X^t) \\ P_T(X^{t+1}|X^t,Y^t)}} \left[\sum_{t=1}^{N} \gamma^t R(X^t, Y^t, X^{t+1}) \right] \middle| X^0 = x \right]$$

$$= \mathop{\mathrm{E}}_{\substack{\pi(Y^0|X^0) \\ P_T(X^1|X^0,Y^0)}} \left[R(X^0, Y^0, X^1) + \gamma \underbrace{\mathop{\mathrm{E}}_{\substack{\pi(Y^t|X^t) \\ P_T(X^{t+1}|X^t,Y^t)}} \left[\sum_{t=1}^{N} \gamma^{t-1} R(X^t, Y^t, X^{t+1}) \middle| X^1 \right]}_{V^1(X^1)} \middle| X^0 = x \right]$$

$$= \mathop{\mathrm{E}}_{\substack{\pi(Y^0|X^0) \\ P_T(X^1|X^0,Y^0)}} [R(X^0, Y^0, X^1) + \gamma V^1(X^1) | X^0 = x]$$

各ステップの価値関数



ステップ 0

$$V_\pi^0(x) = \mathop{\mathrm{E}}_{\substack{\pi(Y^0|X^0) \\ P_T(X^1|X^0,Y^0)}} [R(X^0,Y^0,X^1) + \gamma V_\pi^1(X^1)|X^0 = x]$$

ステップ 1

$$V_\pi^1(x) = \mathop{\mathrm{E}}_{\substack{\pi(Y^1|X^1) \\ P_T(X^2|X^1,Y^1)}} [R(X^1,Y^1,X^2) + \gamma V_\pi^2(X^2)|X^1 = x]$$

⋮

ステップ N-1

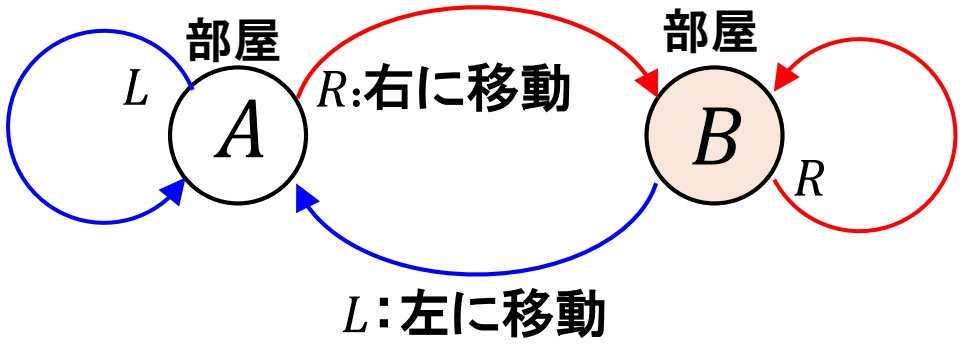
$$V_\pi^{N-1}(x) = \mathop{\mathrm{E}}_{\substack{\pi(Y^{N-1}|X^{N-1}) \\ P_T(X^N|X^{N-1},Y^{N-1})}} [R(X^{N-1},Y^{N-1},X^N) + \gamma V_\pi^N(X^N)|X^{N-1} = x]$$

次のステップの価値関数を参照

ステップ N

$$V_\pi^N(x) = \mathop{\mathrm{E}}_{\substack{\pi(Y^N|X^N) \\ P_T(X^{N+1}|X^N,Y^N)}} [R(X^N,Y^N,X^{N+1})|X^N = x]$$

2状態チェーンウォーク問題



状態遷移確率

	次の状態 A	次の状態 B
$P_T(A,L,\cdot)$	1	0
$P_T(A,R,\cdot)$	0	1
$P_T(B,L,\cdot)$	1	0
$P_T(B,R,\cdot)$	0	1

必ず行動を取った方向に移動

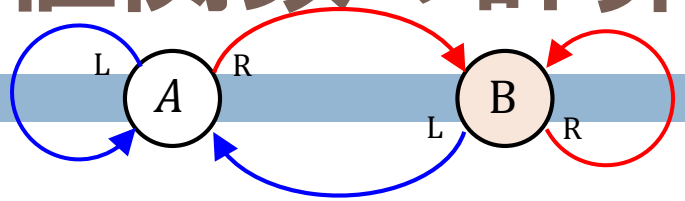
状態：$X \in \{A,B\}$　　行動：$A \in \{L,R\}$　　割引率：$\gamma = 0.9$

ステップ数：$N = 2$　　報酬関数：$R(B,R,B) = 1$　　初期状態確率：$P_I(\mathrm{A}) = 1$

- **決定的な方策関数**
  - 確率「1」でどちらかの行動（LまたはR）を選択
  - 各状態に行動が2種類あるので、方策は以下の4種類

$\pi1(Y = L\|X = A) = 1$	$\pi1(Y = R\|X = B) = 1$
$\pi2(Y = L\|X = A) = 1$	$\pi2(Y = L\|X = B) = 1$
$\pi3(Y = R\|X = A) = 1$	$\pi3(Y = L\|X = B) = 1$
$\pi4(Y = R\|X = A) = 1$	$\pi4(Y = R\|X = B) = 1$

価値関数の計算



状態：$S \in \{\mathrm{A}, \mathrm{B}\}$　行動：$A \in \{\mathrm{L}, \mathrm{R}\}$　割引率：$\gamma = 0.9$

ステップ数：$N = 2$　報酬関数：$R(B, R, \mathrm{B}) = 1$　初期状態確率：$P_I(\mathrm{A}) = 1$

状態遷移確率	次の状態 A	次の状態 B
$P_T(\mathrm{A}, \mathrm{L}, \cdot)$	1	0
$P_T(\mathrm{A}, \mathrm{R}, \cdot)$	0	1
$P_T(\mathrm{B}, \mathrm{L}, \cdot)$	1	0
$P_T(\mathrm{B}, \mathrm{R}, \cdot)$	0	1

- **方策$\pi 1$の場合**：$\pi(L|A) = 1$、$\pi(R|B) = 1$

ステップ2

$$V_{\pi 1}^{2}(A) = \mathop{\mathrm{E}}_{\substack{\pi 1(Y^2|X^2) \\ P_T(X^3|X^2,Y^2)}} [R(X^2, Y^2, X^3)|X^2 = A]$$

行動Rをとる場合　行動Lをとる場合

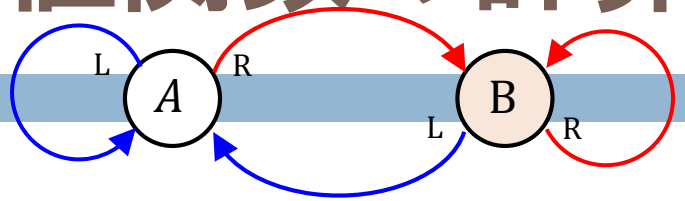
$$= \underbrace{\pi 1(R|\mathrm{A})}_{=0}\underbrace{R(A,R,B)}_{=0}\underbrace{P_T(B|A,R)}_{=0} + \underbrace{\pi 1(L|A)}_{=1}\underbrace{R(A,L,\mathrm{A})}_{=0}\underbrace{P_T(A|A,L)}_{=1} = 0$$

$$V_{\pi 1}^{2}(B) = \mathop{\mathrm{E}}_{\substack{\pi 1(Y^2|X^2) \\ P_T(X^3|X^2,Y^2)}} [R(X^2, Y^2, X^3)|X^2 = B]$$

行動Rをとる場合　行動Lをとる場合

$$= \underbrace{\pi 1(R|B)}_{=1}\underbrace{R(B,R,B)}_{=1}\underbrace{P_T(B|B,R)}_{=1} + \underbrace{\pi 1(L|B)}_{=0}\underbrace{R(B,L,B)}_{=0}\underbrace{P_T(B|B,L)}_{=1} = 1$$

価値関数の計算 2



状態：$S \in \{\mathrm{A}, \mathrm{B}\}$　行動：$A \in \{\mathrm{L}, \mathrm{R}\}$　割引率：$\gamma = 0.9$

ステップ数：$N = 2$　報酬関数：$R(B, R, \mathrm{B}) = 1$　初期状態確率：$P_I(\mathrm{A}) = 1$

状態遷移確率

	次の状態 A	次の状態 B
$P_T(\mathrm{A}, \mathrm{L}, \cdot)$	1	0
$P_T(\mathrm{A}, \mathrm{R}, \cdot)$	0	1
$P_T(\mathrm{B}, \mathrm{L}, \cdot)$	1	0
$P_T(\mathrm{B}, \mathrm{R}, \cdot)$	0	1

- **方策$\pi 1$の場合**：$\pi(L|A) = 1$、$\pi(R|B) = 1$

ステップ1

$$V_{\pi 1}^{1}(A) = \mathop{\mathrm{E}}_{\substack{\pi 1(Y^1|X^1) \\ P_T(X^2|X^1,Y^1)}} \left[R(X^1, Y^1, X^2) + \gamma V_{\pi 1}^{2}(X^2) \middle| X^1 = A \right]$$

行動Lをとる場合

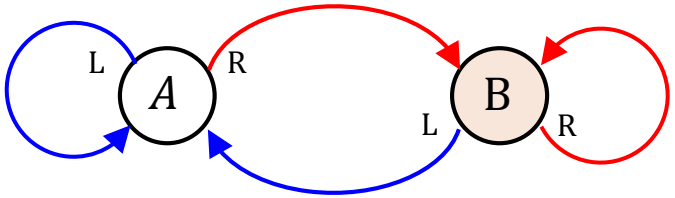
$$= \pi 1(L|A) \left(R(A, L, \mathrm{A}) + \gamma V_{\pi 1}^{2}(A) \right) P_T(A|A, L) = 1 * (0 + 0.9 * 0) * 0 = 0$$

$$V_{\pi 1}^{1}(B) = \mathop{\mathrm{E}}_{\substack{\pi 1(Y^1|X^1) \\ P_T(X^2|X^1,Y^1)}} \left[R(X^1, Y^1, X^2) + \gamma V_{\pi 1}^{2}(X^2) \middle| X^1 = B \right]$$

行動Rをとる場合

$$= \pi 1(\mathrm{R}|\mathrm{B}) \left(R(B, R, B) + \gamma V_{\pi 1}^{2}(B) \right) * P_T(B|B, R) = 1 * (1 + 0.9 * 1) * 1 = 1.9$$

演習1



状態：$S \in \{\mathrm{A}, \mathrm{B}\}$　行動：$A \in \{\mathrm{L}, \mathrm{R}\}$

ステップ数：$N = 2$　報酬関数：$R(B, R, \mathrm{B}) = 1$

割引率：$\gamma = 0.9$

初期状態確率：$P_I(\mathrm{A}) = 1$

状態遷移確率	次の状態 A	次の状態 B
$P_T(\mathrm{A}, \mathrm{L}, \cdot)$	1	0
$P_T(\mathrm{A}, \mathrm{R}, \cdot)$	0	1
$P_T(\mathrm{B}, \mathrm{L}, \cdot)$	1	0
$P_T(\mathrm{B}, \mathrm{R}, \cdot)$	0	1

- 方策$\pi 1$のステップ0の価値関数を求め、方策$\pi 1$について考察しなさい。
  - $\pi(L|A) = 1$、$\pi(R|B) = 1$

ステップ0

$$V_\pi^0(x) = \mathop{\mathrm{E}}_{\substack{\pi(Y^0|X^0) \\ P_T(X^1|X^0,Y^0)}} [R(X^0, Y^0, X^1) + \gamma V_\pi^1(X^1)|X^0 = x]$$

- タイトル「演習レポート」、日付、学生番号、氏名を用紙の一番上に記載

内容

- 教師あり学習
  - 教師あり学習の例と問題点
- 強化学習
  - 強化学習の目的と成功例
  - 動物の行動学習
  - 強化学習の定式化
  - 動的計画法による最適化

強化学習の目的

- 価値関数を最大化する最適な方策π^*を求める。

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, V_\pi^0(x)$$

- 単純なアプローチ：
  - 各方策に対する価値を計算し、最大価値の方策を選択

$$V_{\pi 1}^0(x), V_{\pi 2}^0(x), V_{\pi 3}^0(x), V_{\pi 4}^0(x), \ldots$$

  - 必要な演算数（価値関数の計算回数）

（状態数×行動数） × 状態数 × ステップ数

（状態数×行動数）：方策の数　　状態数 × ステップ数：各方策で計算する価値の数

  - 3ステップ2状態チェインウォークの場合：（2×2）×2×3＝24回
  - 実問題では状態数が膨大

ゲーム	局面の数
将棋	10の226乗
囲碁	10の360乗

- 効率よく最適な価値を求めたい

動的計画法

- **最適化問題を、複数の部分問題に分割して解く**
  1. 強化学習の場合、Nステップの経路を2ステップずつに分割
  2. Nステップ目の価値$V_\pi^N(x)$を計算

$$V_{\pi 1}^N(x), V_{\pi 2}^N(x), V_{\pi 3}^N(x), V_{\pi 4}^N(x), \ldots$$

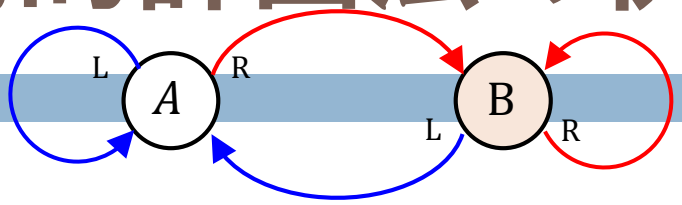
  3. **最大の価値$V_{\pi *}^N(x)$を選択**
  4. **最大の価値$V_{\pi *}^N(x)$を用いて$N-1$ステップ目の価値$V_\pi^{N-1}(x)$を計算**

$$V_\pi^{N-1}(x) = \mathop{\mathrm{E}}_{\substack{\pi(Y^{N-1}|X^{N-1}) \\ P_T(X^N|X^{N-1},Y^{N-1})}} [R(X^{N-1}, Y^{N-1}, X^N) + \gamma V_{\pi *}^N(X^N) | X^{N-1} = x]$$

$$V_{\pi 1}^{N-1}(x), V_{\pi 2}^{N-1}(x), V_{\pi 3}^{N-1}(x), V_{\pi 4}^{N-1}(x), \ldots$$

  5. **最大の価値$V_{\pi *}^{N-1}(x)$を選択**
  6. 4-5を、$V_{\pi *}^0(x)$を獲得するまで繰り返す

動的計画法の例



状態：$S \in \{\mathrm{A}, \mathrm{B}\}$　行動：$A \in \{\mathrm{L}, \mathrm{R}\}$

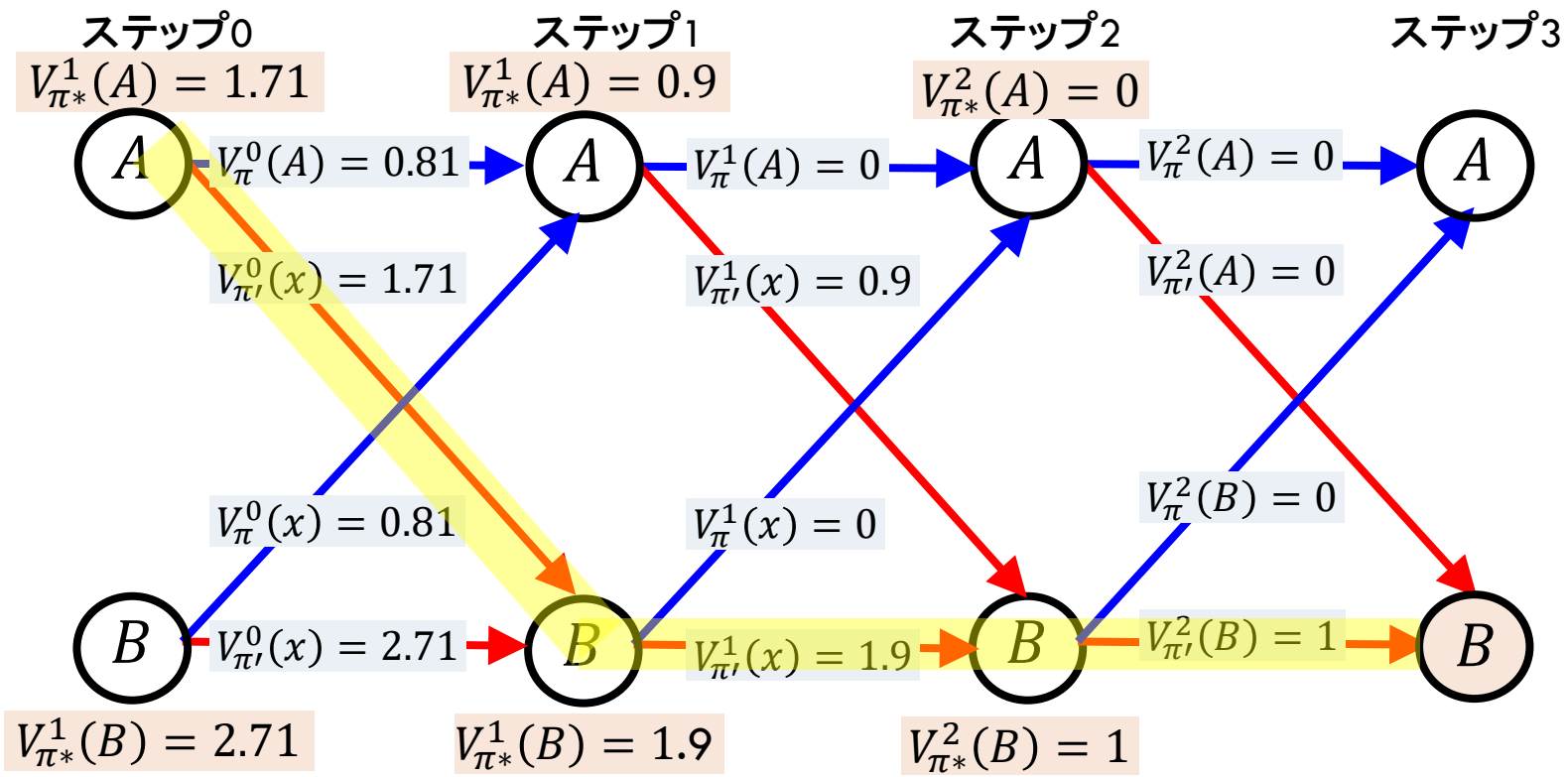
ステップ数：$N = 2$　報酬関数：$R(B, R, \mathrm{B}) = 1$

割引率：$\gamma = 0.9$

初期状態確率：$P_I(\mathrm{A}) = 1$

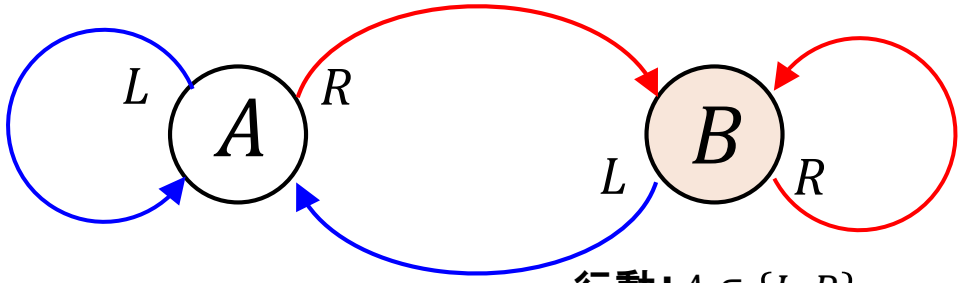
状態遷移確率

	次の状態 A	次の状態 B
$P_T(\mathrm{A}, \mathrm{L}, \cdot)$	1	0
$P_T(\mathrm{A}, \mathrm{R}, \cdot)$	0	1
$P_T(\mathrm{B}, \mathrm{L}, \cdot)$	1	0
$P_T(\mathrm{B}, \mathrm{R}, \cdot)$	0	1



- 価値の計算回数は12回、最適な方策は$\pi(\mathrm{R}|A) = 1$、$\pi(R|B) = 1$

演習2



状態：$X \in \{A, B\}$

ステップ数：$N = 2$

行動：$A \in \{L, R\}$

報酬関数：$R(A, R, B) = 1$

$R(B, R, B) = 0.5$

状態遷移確率

	次の状態 A	次の状態 B
$P_T(A, L, \cdot)$	1	0
$P_T(A, R, \cdot)$	0	1
$P_T(B, L, \cdot)$	1	0
$P_T(B, R, \cdot)$	0	1

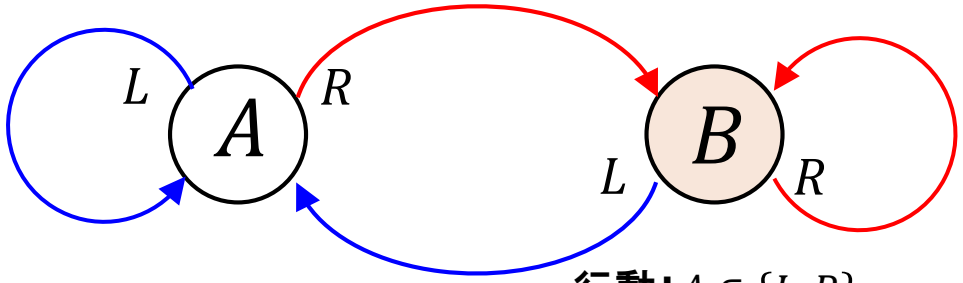
割引率：$\gamma = 0.9$

初期状態確率：$P_I(B) = 1$

1. 動的計画法を用いて、各ステップ最大価値を求めなさい。
2. 最大価値に基づき、最適な方策を求めなさい。

- タイトル「演習レポート」、日付、学生番号、氏名を用紙の一番上に記載

課題



状態：$X \in \{A, B\}$

ステップ数：$N = 2$

行動：$A \in \{L, R\}$

報酬関数：$R(A, R, B) = 1$

$R(B, R, B) = 0.3$

次の状態

状態遷移確率	A	B
$P_T(A, L, \cdot)$	1	0
$P_T(A, R, \cdot)$	0	1
$P_T(B, L, \cdot)$	1	0
$P_T(B, R, \cdot)$	0	1

割引率：$\gamma = 0.9$

初期状態確率：$P_I(B) = 1$

1. 動的計画法を用いて、各ステップ最大価値を求めなさい。
2. 最適な経路を求めなさい。

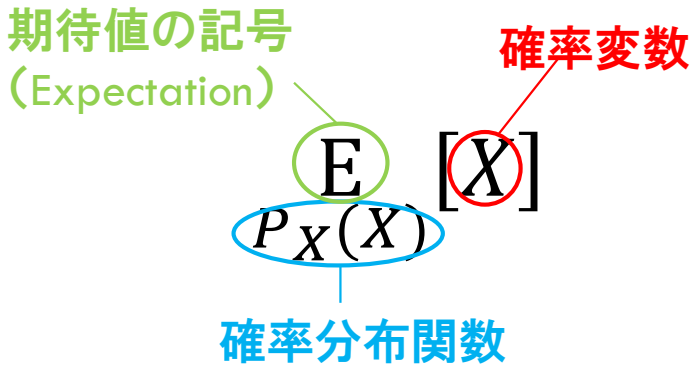
レポートの提出方法

- **演習レポート：**
  - タイトル「演習レポート」、日付・学生番号・氏名を用紙の一番上に記載
- **課題レポート：**
  - タイトル「課題レポート」、出題日・学生番号・氏名を用紙の一番上に記載
  - 2ページ以上になる場合は、ホッチキス留め
  - A4サイズの用紙を使用
  - 一度に複数の課題レポートを提出する場合出題日ごとに別々に綴じる

期待値の定義

- 1変数Xの場合：

$$\underset{P_X(X)}{\mathrm{E}}[X] = \sum_j x_j \, P_X(X = x_j)$$



- 2変数XとYの和の場合：

$$\underset{\substack{P_X(Y) \\ P_{X|Y}(X|Y)}}{\mathrm{E}}[Y + X] = \Sigma_i \Sigma_j (y_i + x_j) \, P_{X|Y}(x_j|y_i) P_Y(y_i)$$

  - 変数Yがyと観測された場合：

$$\underset{P_{X|Y}(X|Y)}{\mathrm{E}}[Y + X|Y = y] = \Sigma_j (y + x_j) \, P_{X|Y}(x_j|y)$$