# 強化学習を用いた横断歩道での自動運転車の人身事故回避

M2022SC016 芳野茂樹
指導教員：坂本登

## 1 はじめに

現在，自動運転技術の開発，特に，レベル 4（限定条件下での完全自動運転）に向けた取り組みが盛んに行われている [1]．そのためにクリアすべき大きな課題の 1 つが人身事故の回避であり，歩行者の行動振る舞いを考慮した自動車の意思決定が求められている．そこで本研究では，横断歩道での歩行者との接触回避を実現する意思決定法の構築を目指す．

本稿ではまず，マルコフ決定過程の概要と強化学習の基礎について説明する．次に，本研究のシミュレーションで扱う横断歩道の問題設定について説明する．設定した問題に対して，危険に関する搭乗者の感性を考慮しない報酬と考慮した報酬を用いた強化学習アルゴリズムを提案する．最後に提案手法の数値シミュレーションを行い，結果に関する比較・考察する．

## 2 マルコフ決定過程

### 2.1 マルコフ決定過程の定義 [2]

マルコフ決定過程は $(S, \{A(s)|s \in S\}, P_0, P, r)$ で記述される．ただし，$S$ は状態の有限集合．$A(s)$ は状態 $s \in S$ において選択できる行動の集合，$P_0 : S \to [0, 1]$ は初期時刻における初期状態の分布確率であり，次式を満たす．

$$\sum_{s \in S} P(s) = 1$$

また，$P : S \times S \times A \to [0, 1]$ は状態遷移関数で，状態 $s \in S$ で行動 $a \in A(s)$ を選択して状態 $s'$ に遷移する確率が $P(s'|s, a)$ であり，次式を満たす．任意の $s \in S$ と $a \in A(s)$ に対して

$$\sum_{s' \in S} P(s'|s, a) = 1$$

$r : S \times A \times S \to R$ は報酬関数で，状態 $s \in S$ で行動 $a \in A(s)$ を選択して状態 $s' \in S$ に遷移したときの報酬が $r(s, a, s')$ である．

### 2.2 活用と探索のトレードオフ

まず強化学習の行動を選択する際には以下の二つの目的がある．詳しくは [3] を参照されたい．

- 探索 (exploration)：方策の精度
- 活用 (exploitation)：報酬の最大化する方策の利用

「探索」はよりより良い方策を学習することを目的に，あえて最適でない行動を選択することである．一方,「活用」は今までに得られて観測データのもとで最適である方策を使って行動を決定することである，活用のみを行うと最適方策を学習できない場合がある．

活用と探索は相反する目的となるので，強化学習を行うときに両者のバランスをどのようにとるかは重要な問題であり，この問題を「活用と探索のトレードオフ」と呼ぶ．

### 2.3 greedy アルゴリズム

強化学習では探索を行うときの方法がいくつかあるが，ここでは，最も単純な手法である greedy アルゴリズムについて説明する．各状態 $s \in S$ において行動価値関数が最大になる行動を選択する決定的方策 $\pi_{greedy}$ である．詳しくは [3] を参照されたい．

$$\pi_{greedy}(s; q) \stackrel{\triangle}{=} \arg\max_{a \in S} \ q(s, a) \tag{1}$$

式 (1) で用いている $q$ は行動価値関数である．

greedy アルゴリズムは行動価値関数が最大の行動を選択するので，行動価値関数が既知の場合は行動価値関数が未知でも学習が十分で「活用」のみを行いたい場合は最適なアルゴリズムといえる．

### 2.4 $\epsilon$ - greedy 方策

後のシミュレーション用いる $\epsilon$ - greedy 方策について説明する．詳しくは [3] を参照されたい．
$\epsilon$ - greedy 方策とは以下の式にしたがって行動を選択する方法である．

$$\pi(a|s; q.\epsilon) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{|A|} & (a = argmax q(s, b)) \\ \frac{\epsilon}{|A|} & (\text{それ以外}) \end{cases} \tag{2}$$

式 (2) での $\pi$ は今までは行動を示していたが，ここでは各行動に対する選択確率になっている．$\epsilon$-greedy 方策は式 (2) の 1 行目では「活用」を 2 行目では「探索」を行っており，この活用と探索を確率 $\epsilon$ で行うという方策になっている．ここで $\epsilon$ はハイパーパラメータで，$\epsilon$ が 1 に近いほど「探索」し，0 に近いほど「活用」する．また，$q$ は行動価値関数である．

## 3 問題設定と研究の概要

### 3.1 取り扱うシナリオと研究目的

はじめに，本研究で取り扱う問題設定について説明する．本研究で対象とする横断歩道の状況を図 1 に示す．自車が図 1 に示す位置で前方の横断歩道を渡ろうとする歩行者を発見したとする．このとき，自車はブレーキをかけて自車を停止線の前で停止させようとするか，ブレーキをかけずに現在の速度で運転して横断歩道を通過しようとするか，いずれの行動が最適な行動であるかという問題を考える．簡単のために車両の長さを無視し，自車の側面に歩行者が衝突することはないと仮定する．本研究では，歩行者を発見した際の一回のみで自車両の意思

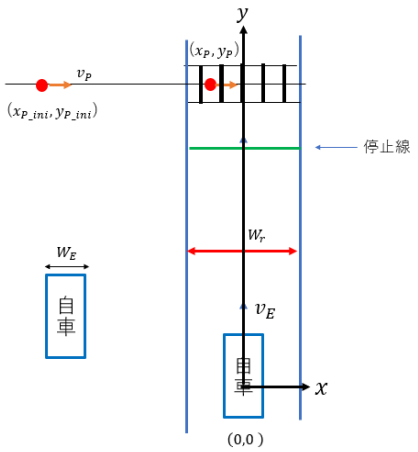


図 1　問題設定

決定を行うことを考える．具体的には，発見時に得られる歩行者の情報（位置・速度），および，報酬を用いて，自車両の行動である「止まる（ブレーキを踏む）」「進む（ブレーキを踏まない）」を選択する方策を獲得することを目指す．

### 3.2　扱った強化学習の大枠

本論文で扱う強化学習の大枠について図 2 を用いて説明する．「物理シミュレータ」は自車両と歩行者の連続

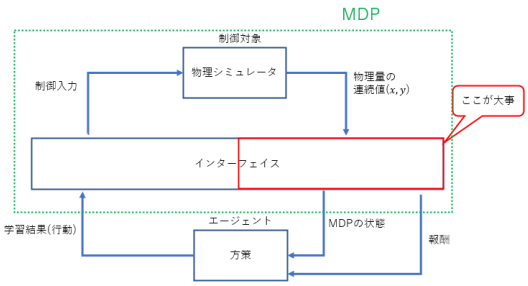


図 2　扱った強化学習の大枠

的な運動を模擬するものであり，「インターフェース」と結合することで，強化学習の対象である MDP として振る舞うものであると考える．具体的には，制御入力は自車両の加速度であり，「インターフェース」において，本研究の学習対象である行動の「進む」「止まる」により特徴づけされている．一方，出力である自車両と歩行者の連続値は「インターフェース」においてサンプリングされ，衝突の有無，その際の相対位置・速度が抽出される．この情報が MDP の状態や報酬として扱われ，最適方策が学習されることになる．

### 3.3　パラメータと定数

本研究のシミュレーションで扱うパラメータと定数を表 1 に示す．ただし，定数記号についてはその値を示し，値が示されていない記号がパラメータである．

表 1　パラメータと定数 (値が示されていない記号がパラメータである)

| 記号 | 名称 | 値 | 単位 |
|---|---|---|---|
| $(x_E, y_E)$ | 車両の位置 | | m |
| $(x_{P_{ini}}, y_{P_{ini}})$ | 人の初期位置 | $(-5, 50)$ | m |
| $(x_P, y_P)$ | 人の位置 | | m |
| $(vE_x, vE_y)$ | 車両の速度 | $(0, 40)$ | km/h |
| $(vP_x, vP_y)$ | 人の速度 | $(4, 0)$ | km/h |
| $W_E$ | 車両の横幅 | 2 | m |
| $W_r$ | 道路の幅 | 4 | m |
| $L_{P_x}$ | 自車と他車の $x$ 座標の相対距離 | | |

## 4　衝突回避を目的とする強化学習

### 4.1　マルコフ決定過程を用いたモデル化

3 章で述べた問題について，歩行者と自車の間の関係をマルコフ決定過程でモデル化する．この問題では，自車の行動決定は，歩行者を発見した時に行われ，以降はその行動に従って自車を操作する．そこで，歩行者発見時を初期状態とし，「止まる」を選択して自車が停止したときに取り得る状態および「進む」を選択して自車が横断歩道を通過したときに取り得る状態を最終状態とする．また，歩行者発見時の自車と歩行者の位置と速度を自車基準で差を取り，相対位置と相対速度として定義する．具体的には，以下のように状態を定義する．状態 $\{s^1, s^2, s^3, s^4, s^5, s^6, s^7\} \in S$ とする．

相対距離：$\{-5[m], 50[m]\}$

相対速度：$\{-4[km/h], 40[km/h]\}$

ここで $s^1$〜$s^7$ のそれぞれの状態について説明する．$s^2$〜$s^7$ の状態を簡単に図 3 に表す．$s^1$ は相対距離と相対速度の組み合わせを考える．今回の組み合わせは一つになるので以下のようになる．

$$s^1 = \{(-5, 50), (-4, 40)\}$$

$s^2$ =「進む」を選択，人が自車の左側にいる状態
$s^3$ =「進む」を選択，人が自車の正面にいる状態
$s^4$ =「進む」を選択，人が自車の左側にいる状態
$s^5$ =「止まる」を選択，人が自車の左側にいる状態
$s^6$ =「止まる」を選択，人が自車の正面にいる状態
$s^7$ =「止まる」を選択，人が自車の右側にいる状態

次に行動についても以下のようにモデル化する．行動集合 $A = \{a^1, a^2\}$ である．ただし，

$a^1$ =「進む」
$a^2$ =「止まる」

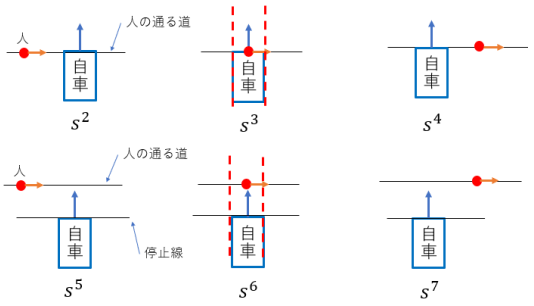


図 3　定義した状態図

### 4.2 報酬関数の方針

自車が歩行者と衝突したときに最も大きい負の報酬を与え. 自車が止まった時，歩行者が自車の前にいた場合は止まって正解ということで最も大きい正の報酬を与える. また自車が進んだ場合と止まった場合での歩行者の位置によってどのように感じるかを報酬の設定で考慮する.

### 4.3 報酬関数の設計

4.1 節で定義したモデルに対して強化学習を行うための報酬関数をどう設計したのかについて説明する.
本章では定義した遷移先の各状態に対して自車が停止もしくは横断歩道を通過時の歩行者と自車との距離によらず定数を返す報酬 $R(s')$ を用いた. 与えた値は以下の通りである.

$$R(s') = \begin{cases} 10 & (s' = s^2) \\ -100 & (s' = s^3) \\ 70 & (s' = s^4) \\ 50 & (s' = s^5) \\ 100 & (s' = s^6) \\ 30 & (s' = s^7) \end{cases}$$

## 5 危険に対する感性を考慮した報酬設計

4 章では，歩行者との衝突の有無のみで報酬を設計した．しかし，実際には，衝突せずに横断歩道を通過したとしても歩行者にかなり接近した場合には歩行者と衝突するのではないかと搭乗者が感じ，ヒアッとする．本章では，この危険に対する感性を考慮した報酬を提案し，その報酬の下で強化学習による最適行動について，シミュレーションにより検討する.

### 5.1 マルコフ決定過程でのモデル化

本節では搭乗者の感性を反映した報酬をもつマルコフ決定過程を提案する．マルコフ決定過程では状態と行動をモデル化する必要がある．シミュレータの開始時と終了時の他車の相対距離と相対速度を 人 − 車 として考えて状態としてサンプリングする．$s^1$ は車線変更をするかどうかの意思決定をする状態で，初期状態である．状態 $\{s^1, s^2, s^3\} \in S$ とする.
相対距離：$\{-5[m], 50[m]\}$
相対速度：$\{-4[km/h], 40[km/h]\}$　ここで $s^1, s^2, s^3$ のそれぞれの状態について説明する．$s^1$ は相対距離と相対速度の組み合わせを考える．今回の組み合わせは一つになるので以下のようになる.

$$s^1 = \{(-5, 50), (-4, 40)\}$$
$$s^2 =「進む」を選択した状態$$
$$s^3 =「止まる」を選択した状態$$

次に行動についても以下のようにモデル化する．行動の集合 $A$ は $A = \{a^1, a^2\}$ である．ただし，各行動の意味は以下のとおりである.

$$a^1 =「進む」を選択$$
$$a^2 =「止まる」を選択$$

「止まる」の時の最低報酬を-50,「進む」の最低報酬を-100 とした．理由は,「止まる」の時はどんな状況でも衝突はしないため,「進む」を選択して衝突した時の報酬-100 より小さくすべきと判断した.

#### 5.1.1 「進む」を選択時の報酬関数

まず,「進む」を選択した時の報酬関数について説明する．$d_{left}$ と $d_{right}$ が搭乗者の感性に依存して定まるパラメータで以下の意味をもつ.
・$d_{left}$：搭乗者が危険と感じ始める相対位置
・$d_{right}$：搭乗者が安全と感じ始める相対位置

報酬 $R^2(L_x)$ は次式で記述される.

$$R^2(L_x) = \begin{cases} 0 & (L_x < d_{left}) \\ \frac{-100}{-1-d_{left}}(L_x - d_{left}) & (d_{left} \leq L_x < -1) \\ -100 & (-1 \leq L_x \leq 1) \\ \frac{100}{d_{right}-1}(L_x - d_{right}) & (1 < L_x \leq d_{right}) \\ 0 & (L_x < d_{right}) \end{cases}$$

#### 5.1.2 報酬関数 $R^2$ の説明

$L_x$ について場合分けして説明する．衝突する場合の報酬を-100 とし，まったく危険を感じない場合の報酬を 0 として，自車と歩行者との距離に依存して以下の方針で報酬関数を設計する.
$(L_x < d_{left})$：最大報酬となる 0 を与える.
$(d_{left} \leq L_x < -1)$：危険と感じ始める位置 $(L_x = d_{left})$ で報酬 0，衝突 $(L_x = -1)$ の時に報酬-100 とし，この間を直線で補間する.
$(-1 \leq L_x \leq 1)$：衝突しているため，報酬は-100 である.
$(1 < L_x \leq d_{right})$：衝突 $(L_x = 1)$ の時に報酬-100，安全と感じ始める位置 $(L_x = d_{right})$ の時に報酬 0 とし，この間を直線で補間する.
$(L_x > d_{right})$：報酬 0 を与える.

#### 5.1.3 「止まる」を選択時の報酬関数

次に「止まる」 を選択した時の報酬関数について説明する.「止まる」を選択した場合，搭乗者目線では衝突しないので危険と思う事はなくなるので，$d_{left}$ と $d_{right}$ の意味が以下のようになる.
・$d_{left}$：搭乗者が止まってよかったと思い始める相対位置
・$d_{right}$：搭乗者が行けたかもしれないと思い始める相対位置

報酬 $R^3(L_x)$ の与え方は以下のようにした.

$$R^3(L_x) = \begin{cases} -50 & (L_x < d_{left}) \\ \frac{-50}{-1-d_{left}}(L_x - d_{left}) - 50 & (d_{left} \leq L_x < -1) \\ 0 & (-1 \leq L_x \leq 1) \\ \frac{-50}{d_{right}-1}(L_x - d_{right}) - 50 & (1 < L_x \leq d_{right}) \\ -50 & (L_x < d_{right}) \end{cases}$$

$L_x$ について場合分けして説明する.
・$(L_x < d_{left})$
最小の報酬となる-50 を与える.
・$(d_{left} \leq L_x < -1)$

止まってよかったと感じ始める位置 ($L_x = d_{left}$) の時に報酬-50，止まることで衝突を避けられた位置 ($L_x = -1$) の時に報酬 0 とし，この間を直線で補間する．

・$(d_{left} \leq L_x < -1)$

止まってよかったと思うので報酬は 0 である．

・$(1 < L_x \leq d_{right})$

止まって正解と感じ始める位置 ($L_x = 1$) で報酬 0，後悔を感じ始める位置 ($L_x = d_{right}$) の時に報酬-50 とし，その間を直線で補間する．

・$(d_{right} < L_x)$

報酬-50 を与える．

### 5.2 モデルの特徴と理由

4 章で説明したモデル化では状態が $s^1$〜$s^7$ の 4 個であるが，本章のモデル化では $s^1$〜$s^3$ の 3 個になり，状態数を減らしているという特徴がある．ここで，状態 $s^2, s^3$ と行動 $a^1, a^2$ が同じように見え必要とは思えないかもしれないが,「$s^2, s^3$ にはそれぞれの状態に対して Q 値を計算する」,「各状態 $s^2, s^3$ には距離の概念も入っている」という理由からその状態 $s^2, s^3$ を導入する．

人の危険への感性を報酬として表現する場合，この報酬は自車と歩行者との距離に依存する．そこで，本章では歩行者と車両の距離に対する関数として表現する方法を用いる．その方法を用いることで,「自車と歩行者の相対距離による搭乗者の感性を考慮できる点」と「状態数を減らすことができる点」という利点がある．

### 5.3 強化学習シミュレーション

最後にこの報酬関数を使ってシミュレーションを行った結果を示す．以降のシミュレーションでは，greedy アルゴリズムを用いる．また，搭乗者の感性を表す $d_{left}$ と $d_{right}$ は報酬関数 $R^2$ と $R^3$ のどちらでも同じ値を用いる．

#### 5.3.1 学習結果 (基準とした搭乗者の感性)

まず，搭乗者の感性を表す $d_{left}$ と $d_{right}$ の値を決める．搭乗者の感性は $d_{left} = -6$ と $d_{right=6}$ を基準とし，学習を行った．学習設定を $\epsilon$=0.05，学習回数：1000 回とした．基準の人での，学習結果を表 2 に示す．

表 2 最後の学習の Q 値 (基準)

| 行動 | 進む | 止まる |
|---|---|---|
| Q 値 | -50.0489 | -31.5855 |

学習 1 回ごとの Q 値の変化を図 4 に示す．図 4 より,「止まる」の Q 値は-30 付近で収束している．しかし,「進む」の Q 値に関しては収束しているとは言えない．理由としては，$\epsilon$-greedy 方策で Q 値の高い行動を選択し，Q 学習法で Q 値を更新していくので Q 値の低い方である「進む」は選択回数が少ないため更新回数が減り，収束が遅れていると考えられる．

#### 5.3.2 学習結果 (せっかちな人)

$d_{left} = -3$ と $d_{right} = 3$ とし，学習を行った．学習設定を $\epsilon$=0.05，学習回数：1000 回とした．
せっかちな人での，学習結果を表 3 に示す．

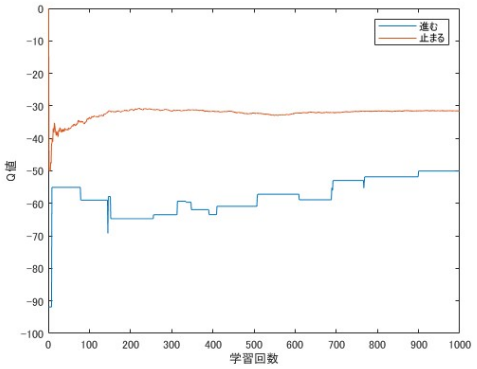


図 4 学習 1 回ごとの Q 値の変化 (基準)

表 3 最後の学習の Q 値 (せっかち)

| 行動 | 進む | 止まる |
|---|---|---|
| Q 値 | -33.1069 | -38.9124 |

先程の基準の人と比較すると,「止まる」の Q 値と「進む」の Q 値の差が狭まっている．結果として,「進む」の Q 値が「止まる」の Q 値よりも高くなっており，最適な行動として「進む」を選択している．したがって，せっかちな人を表現できている．

#### 5.3.3 学習結果 (慎重な人)

$d_{left} = -10$ と $d_{right} = 10$ とし，学習を行った. 学習設定を $\epsilon$=0.05，学習回数：1000 回とした．
慎重な人での，学習結果を表 4 に示す．

表 4 最後の学習の Q 値 (慎重)

| 行動 | 進む | 止まる |
|---|---|---|
| Q 値 | -69.3348 | -23.5457 |

基準の人と比べると,「止まる」の Q 値と「進む」の Q 値の差が大きく広がっているため，より「止まる」という行動を重要視している．したがって，慎重な人を表す結果になっている．

## 6 おわりに

本研究では，マルコフ決定過程でモデルを作成し，そのモデルに強化学習を適応した．greedy アルゴリズムを用いて強化学習を行い，結果を比較した．また，搭乗者の感性を考慮することで最適な行動を選択する結果が変わることが確認できた．

## 参考文献

[1] 日本ロボット学会　監修，香月 理絵　編著，荒井 幸代 ほか　共著：『自動運転技術入門』．オーム社，東京，2021.

[2] 前田　康成：『マルコフ決定過程：モデル化の基礎と応用事例』．森北出版，東京，2021.

[3] 森村哲郎：『強化学習 (機械学習プロフェッショナルシリーズ)』．講談社，東京，2019.