124

セーリング戦略の強化学習シミュレーション

橘 完太[*1]

Reinforcement Learning Simulation of Sailing Strategy

Kanta TACHIBANA[*1]

[*1] Kogakuin Univ., Dept. of Information Design
Nishishinjuku 1-24-2, Shinjuku, Tokyo, 163-8677, Japan

The only energy source of sailboat is wind. Decision making is need in partially observable and unsteady environment. Sailing is a reinforcement learning problem. State and action spaces consist of continuous variables such as appearance wind, target position and sail direction. Some of them are vectors. This paper shows reinforcement learning simulation with particle filter which aims to learn human sailor's knowledge acquisition of predicting and utilizing wind.

Key Words : Uncertain reward, continuous state space, particle filter

1. 緒　　言

洋上のチェスとも呼ばれる帆走は，自ら動力源を持たず，環境に散在する風や潮のエネルギーを集約して利用する移動手段である．目標位置まで移動するためには，環境の中で現在置かれている状態と目標達成までの状態の系列を正しく予測及び評価し，行動の系列を決定しなければならない．帆走における環境，特に主要なエネルギー源である風は時々刻々変化し，また，たとえある地点の風向風力が同じであってもその地点に進入する艇の速度によって推進力が変わるため，変化に柔軟に対応でき，かつ，将来の状態を見越した戦略が必要である．

強化学習(1)は，不確定な環境の中でエージェントが意思決定する問題に対し，状態に応じて適切な行動を選択できるよう学習させる手法である。強化学習は，エージェントが獲得する報酬が大きくなるように進む．強化学習において，Actor-Critic モデルが広く用いられている．Actor-Critic モデルでは，Critic は状態の価値（その状態から将来予想される報酬）を評価し，Actor は現在の状態から価値の高い次の状態へ遷移させるような行動を選択する．

帆走は、状態も行動も連続な強化学習問題ととらえることができる．人間が帆を操作する場合，風によって生じる推進力が大きくなるよう帆の向きを調整する．調整方法は，失敗・成功を含めた試行を多く経験して獲得される．推進力が得られない失敗時には負の報酬を，大きな推進力を得る成功時には正の報酬を獲得することで，現在の状態とこれまで経験した試行を比較し，より多くの報酬が期待される行動を選択する．森本と銅谷(2)は，起き上がり動作という連続な状態空間での強化学習にガウス分布モデルで評価関数を近似する手法を提案した．柏村ら(3)は，行動空間が連続である強化学習問題に対する粒子フィルタの有効性を指摘した．いずれの問題も状態空間や行動空間が連続で，強化学習手法を拡張する必要があったが，環境の揺らぎが存在せず，状態価値を確定的に決めることで十分に学習が達成できる問題である．Sterne(4)は帆走の問題に強化学習手法を導入したが，こちらも環境に揺らぎを想定せず，確定的な状態価値によって学習が達成できる．これらの研究に対して，筆者(5)は，同じ状態においても環境の揺らぎによって将来の報酬，つまり，状態価値に変動があることを想定する強化学習法を提案した．Critic が評価する状態価値を確率変数とし，その分布と，Actor が選ぶ行動を粒子で離散近似表現する強化学習手法を検討する．本稿では，状態価値を確定的とする場合との比較を行う．

[*1] 工学院大学（〒163-8677　東京都新宿区西新宿 1-24-2）
E-mail: kanta@cc.kogakuin.ac.jp

2. 帆船操縦の強化学習方法

本研究では，状態空間と行動空間がどちらも連続な帆船操縦の問題に対し，Actor-Critic モデルでの強化学習へ粒子フィルタを導入する．2.1 節～2.3 節では Critic モデルについて述べる．従来の Critic モデルでは，将来得られる報酬に遠い未来ほど大きく割り引くような割引率を乗じ，報酬の期待値を価値関数として用いることが多い．本研究では，将来の報酬（状態価値）の期待値のみを考慮するのではなく報酬自体を確率変数とする．つまり，状態価値の確率分布を考慮する．本稿の Critic モデルは，連続な状態空間を離散化するため，状態中心を持つ．また，各状態中心は価値の確定的な値ではなく価値の分布を粒子表現して保持する．2.4 節で述べる Actor モデルでは連続な行動空間の中で，粒子で表現する行動候補を生成し，Critic が評価した価値に応じて行動を選択する．2.5 節では Actor の選択した行動の結果として新たに生じた経験に基づき状態中心を更新する手法について述べる．2.6 節では，本稿で考察する帆走の条件を記述する．

2・1 状態価値の確率分布

本稿で考える帆走の場合，状態 s の価値 $V(s)$ は目標達成までの所要時間 $T(s)$ について単調減少とすべきである．所要時間は目標までの経路上の風向風力によって揺らぐため確定的な値でなく， $T(s)$ を確率変数と考える．確率変数としての状態価値を $V(s) = \exp\{-T(s)/\tau\}$ と設定する．ここで， τ は定数である．$V(s)$ を確率変数 $V(s) = \{v,\ f(v \mid s)\}$ とみなす．ここで，価値 v がとりうる範囲は $0 \le v \le 1$ であり， $f(v \mid s)$ は状態 s で条件付けた確率密度関数である．ただし，目標に到達した状態 s_G については一定の確率 $1-\epsilon$ で $v = 1$ とし，また，望ましくない状態 s_B については一定の確率 $1-\varepsilon$ で $v = 0$ とする．

2・2 連続な状態空間の離散化

連続な状態空間中に K 個の状態中心 c_k を置き，任意の状態 s を c_k を用いて表す．各状態中心 $c_k,\ k \in \{1, \dots, K\}$ には，状態空間での広がりを表す分散パラメータ σ_k^2 を持たせる．任意の状態 s について中心 c_k ，分散 σ_k^2 の正規分布の確率密度 $\alpha_k = N(s; c_k, \sigma_k^2)$ を算出する．確率密度 α_k は k 番目の状態中心 c_k から状態 s が生成される度合いを表す．また，状態中心は価値の確率分布を保持する．本研究では，状態価値の分布を I 個の粒子で離散表現する． i 番目の粒子の価値は $v_i \sim V(c_k)$ に従って生成され，保持される．

2・3 状態価値の実現値の抽出

従来のモデルと同じく Critic は状態価値を評価するが，本研究で導入する Critic は，状態価値を確率変数とする．状態 s の価値 $V(s)$ の確率密度を $f(v \mid s) = f(v \mid c)\, P(c \mid s)$ と分解する．まず，状態 s について計算した K 個の α_k をもとに，状態中心をひとつ $\tilde{c}(s) \sim P(c \mid s) = \alpha_k / Z_\alpha$ に従って抽出する．ここで $Z_\alpha = \sum_k \alpha_k$ である．次に，抽出された状態中心 $\tilde{c}(s)$ に保持されている価値の分布から実現値 $\tilde{v} \sim f\big(v \mid \tilde{c}(s)\big)$ を抽出する．

2・4 行動の選択

Actor は、次時刻における状態について Critic が評価する価値 $V(s')$ に応じて s' へ遷移させる行動 a を選択する．まず， J 個の行動候補を $a_j \sim f(a \mid s)$ に従って生成する．次に，各行動候補 a_j について次状態 s_j' を確定的に $s_j' = g(a_j, s)$ と計算する．そして，次状態の価値 $V(s_j')$ が高い行動 a_j がより高い確率で選択されるよう，次状態の実現値を $\tilde{s}' \sim f(s', a \mid s) \propto V(s') f(a \mid s)$ に従って抽出する。

2・5 状態中心の価値の更新

Actor が選択した次状態の価値 $V(s')$ は，目標達成までの時間 $T(s')$ の関数となる．行動選択前の状態 s から目標達成までの時間は $T(s) = T(s') + 1$ であり，行動選択前の状態の価値は $V(s) = V(s') \exp(\tau^{-1})$ と評価される．よって，次状態の価値の実現値 $\tilde{v} \sim V(s')$ をひとつ抽出する一方で，行動選択前の状態 s について状態中心をひとつ $\tilde{c} \sim P(c \mid s)$ に従って抽出し，抽出した状態中心 $\tilde{c}$ の価値と状態空間内での位置および分散を更新する．

価値の確率分布の更新では， $V(\tilde{c})$ を表現する粒子を I 個の中からひとつ選択し，その粒子の値を $\tilde{v}'\exp(\tau^{-1})$ に置き換える．状態空間内での位置と分散については，予め $\tilde{c}$ に慣性パラメータ w を持たせておき，分散を $\sigma^2 \coloneqq \big(w\sigma^2 + d^2(\tilde{c},s)\big)/(w+1)$ と更新する．ここで， $d(\tilde{c},s)$ は状態中心 $\tilde{c}$ と状態 s との距離である．その後，位置を $\tilde{c} \coloneqq (w\tilde{c}+s)/(w+1)$ と更新する．

2・6　帆船の状態と操作する行動

帆走では，環境中に吹く絶対風と航行によって生じる進行風の合成風を帆に受け，帆の操作によって推進力へ転換する．絶対風自体が揺らぐことが一般的であり，また，同じ絶対風でも進行速度によって，推進力を増大させる帆の使い方が変わる．実際の航行では，目標状態へ到達するために他の艇などの障害物を回避する必要があるが，本稿では簡単にするため，障害物は考慮せず，前方へ移動することを目標とする．また，適切なタイミングで方向転換してコース取りを決定することが帆走で解決すべき主要な問題であるが，本稿では簡単のため，艇は方向転換せず，様々な方向から風を受ける環境中で推進力を得る帆の使い方を強化学習させるものとする．各地点での風ベクトルは風向風力とも時々刻々確率的に変化するものとする．

このような条件の帆走は，状態空間も行動空間もともに連続な強化学習問題ととらえることができる．状態 s を，帆に受ける力 f_s ，および，帆の角度 θ_s と定義する．帆の角度は鉛直に立つマストを軸とする水平面内の一自由度の回転角度とする．艇首方向を $\theta_s = 0$ ，左舷方向を $\theta_s = \pi/2$ ，艇尾方向を $\theta_s = \pi$ とする．帆に受ける力は， $\theta_s = \pi$ のときに右舷を向く側で受ける力を正，その裏側から受ける力を負とする．このように定義すると，推進力が $p_s = f_s \sin\theta_s$ と計算できる．推進力は，合成風の方向と帆の角度によっては負の値，つまり，減速力となる．

目標は前方への移動であるので，推進力がある正の値 $p_G(>0)$ へ達した状態を目標達成とみなし，一定の確率 $1-\epsilon$ で状態価値を 1 とする．また，推進力がある負の値 $p_B(<0)$ を下回った状態の状態価値を一定の確率 $1-\varepsilon$ で 0 とする．状態中心 c と状態 s との距離は $d(c,s) = |p_c - p_s|$ のように，推進力の差とする．

帆走における行動を，帆の角度の変化 $\Delta\theta$ と定義する．行動候補は，フォン・ミーゼス分布に従って生成する．

3.　シミュレーション

平面の原点を目標点とし，12 方向から原点へ向けて帆走する艇速獲得実験を行う．風は平面内のどこでも確率的に同一で，X 軸方向へ吹く風速 5 の風を基本とし， $w_x \sim N(5,\sigma_w^2)$, $w_y \sim N(0,\sigma_w^2)$ により各地点で毎時刻風ベクトル $\vec{w} = \big(w_x, w_y\big)^t$ を生成する．

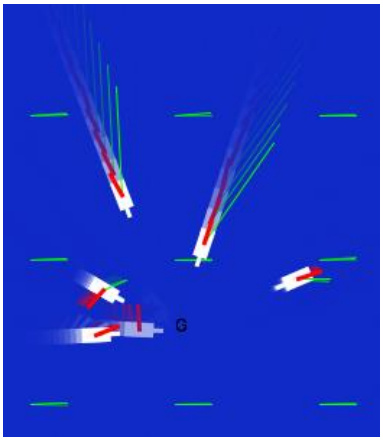


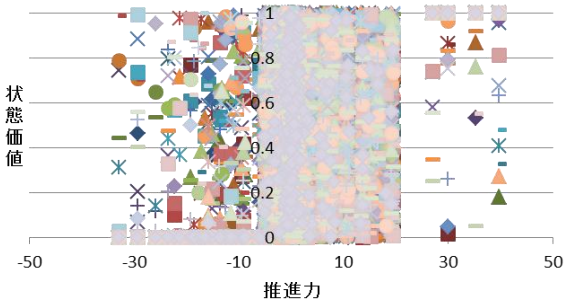
図 1: 帆走のシミュレーション画像
緑の線分で表す風を、赤い線分が表す帆で受け推進力を得る。白い船から出ている緑の線分は合成風を示す。

帆船の位置 $\vec{x}$ から艇首を $-\vec{x}$ へ向け，前方，つまり，原点の方向へ艇速を獲得することを目標とする．艇首方向を X 軸, 左舷方向を Y 軸とする艇の座標系で表した風 $\vec{w}_B$ と現在の艇の速さ v から，合成風は $\vec{w}_A = \vec{w}_B - (v,0)^t$ となる．帆に受ける力は，合成風の速さの自乗に比例するものとし，艇の座標系で合成風の方向を φ としたとき $f_s = |\vec{w}_A|^2 \sin(\theta_s - \varphi)$ により求める．前述の推進力と，艇速の自乗に比例する抵抗力から，次の時刻における艇速を求める．なお，合成風からは艇首の方向を変える力や艇自体を横流れさせる力も受けるが，方向転換しないという単純化のため，それらはすべて打ち消されるものとする．

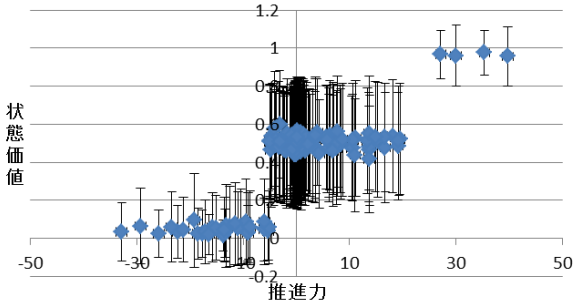
シミュレーションでは，12 艇の帆船が $\pi/6$ ずつ扇形の領域を担当し，強化学習は領域内のランダムな位置に艇速 0 で居る状態から開始する．図 1 にシミュレーションの画像を示す．艇が原点に十分近付いた場合 $(|\vec{x}| < 50)$ ，または，艇が原点から離れた場合 $(|\vec{x}| > 550)$ には，もう一度，担当する扇形領域内のランダムな位置から再開

する．その時，艇の速さは継続するものとする．20,000 時刻のシミュレーションを行い，全ての艇の状態を使って学習を行うため，状態中心の価値，位置および分散の更新は 240,000 回行う。

図 2 に，学習前における各状態中心の推進力と状態価値の分布との関係を示す． $p_G = 25$，$p_B = -5$ と設定した．この際， $\epsilon = \varepsilon = 0.1$ とした．



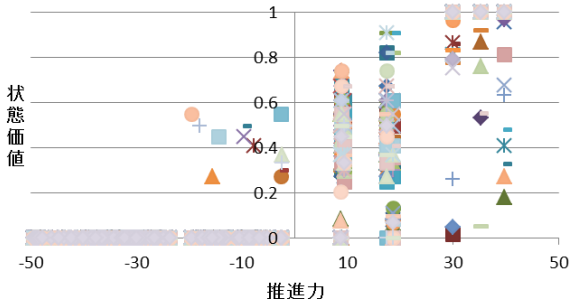
(a) 粒子で表現する確率分布



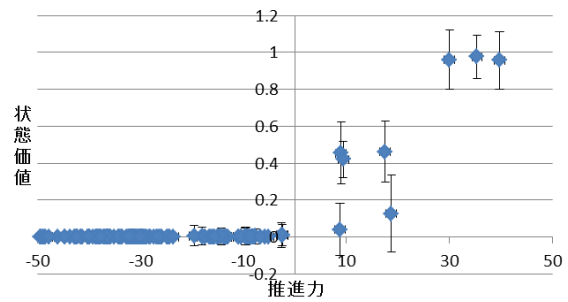
(b) 価値の期待値と標準偏差

図 2: 艇速獲得学習前の状態価値

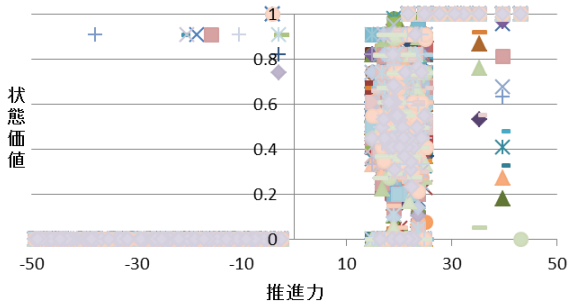
図 3 に，風が安定して吹いている場合 $(\sigma_w = 0.1)$ および風の揺らぎが大きな場合 $(\sigma_w = 1)$ での学習後の状態価値を示す．各状態中心の推進力を横軸に，状態中心が保持する価値の確率分布を縦軸に示す．設計通り，推進力が $p_G = 25$ より大きい目標達成状態では価値が 1 に近く，推進力が $p_B = -5$ より小さい状態では価値が 0 に近くなっている．それらの中間の推進力に相当する状態中心の価値は，大きく広がっている．揺らぎの大小に注目して比較すると，中間の推進力に相当する状態中心の数が多く，それぞれの価値の分布の分散が大きい．これは，同じ状態であっても，環境の揺らぎのために，その後に受ける報酬が大きくばらつくためである．



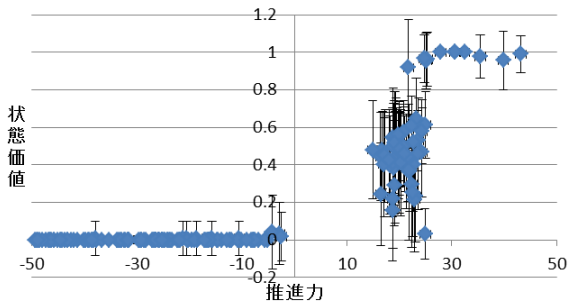
(a) 揺らぎの小さな学習後（粒子）



(b) 揺らぎの小さな学習後（期待値と標準偏差）



(c) 揺らぎの大きな学習後（粒子）

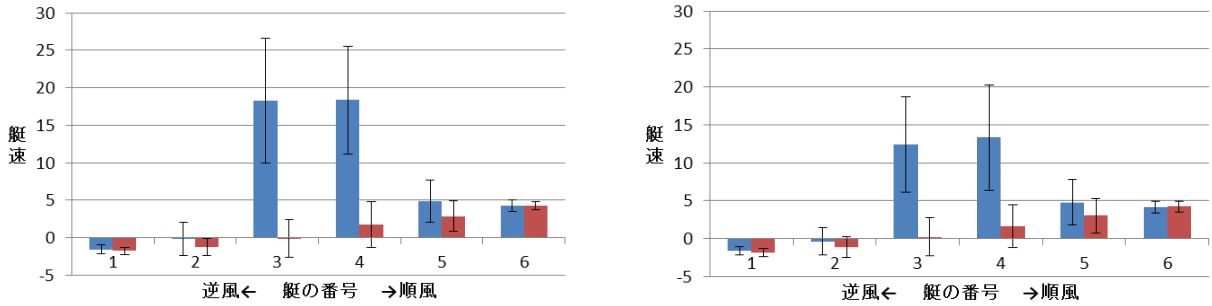


(d) 揺らぎの大きな学習後（期待値と標準偏差）

図 3: 学習後の状態価値 (a) 粒子で表現する確率分布(揺らぎ小); (b) 価値の期待値と標準偏差(揺らぎ小); (c) 粒子で表現する確率分布(揺らぎ大); (d) 価値の期待値と標準偏差(揺らぎ大)

図 4 には，学習シミュレーション後半における各艇の速さの平均値と標準偏差を示す．推進力を考える場合には左右対称となるので，6 つの場合に分けて番号を付けて示す．帆船の番号は，原点へ向かうと逆風となる領域を担当する船が 1，順風となる領域を担当する船が 6 となる．青い棒グラフは粒子数 I =100 とした場合，つまり，

状態価値を確率分布として保持した場合を示す．一方、赤い棒グラフは粒子数 $I = 1$ として状態価値を確定的に学習した場合を示す．確率的に状態価値を学習する手法は，確定的に学習する手法に比べ，揺らぎの小さい場合でも大きい場合でも強い推進力を得て，艇速を維持した．



(a) 揺らぎの小さな学習後半　　(b) 揺らぎの大きな学習後半

図 4: 揺らぎの大きい環境下での学習後半の艇速。青: 提案手法; 赤: 価値を確定的に学習

4. 結　　語

帆走の艇速獲得シミュレーションを行い, 状態価値を確率変数とする強化学習によって価値の形成がなされることを確認した．環境の揺らぎの小さい場合には，学習後の状態中心が目標達成状態か望ましくない状態のどちらかにはっきりと分かれ，中間の状態でも状態中心ごとの価値の分散が小さくなる方向へ学習が進んだ．一方，環境の揺らぎを大きく設定した場合には，保持する価値の分布が大きく広がる状態中心も存在し，価値を確率変数とする提案の効果が発揮された．

文　　献

(1) R.S. Sutton, A.G. Barto: 強化学習，森北出版，三上貞芳，皆川雅章(訳), 2000
(2) 森本淳，銅谷賢治: 強化学習による起き上がり運動パターンの獲得，信学技報, NC97-28, 1997-07, pp. 25 – 32, 1997
(3) 柏村洋平，上野敦志，辰巳昭治: 強化学習のための Particle Filter を用いた連続行動空間表現，人工知能学会全国大会論文集, JSAI08, pp.118 – 121, 2008
(4) P.J. Sterne, Reinforcement Sailing, Master of Science Artificial Intelligence School of Informatics, University of Edinburgh, 2004
(5) 橘完太，粒子フィルタを用いた強化学習による無人帆走の検討，第 29 回ファジィシステムシンポジウム予稿集, 2013