状態遷移モデルを持つ強化学習における政策信頼性を考慮した学習効率化

氏　　　名　藤井　信治
学位の種類　博士(工学)
学位記番号　博甲第８７０号
学位授与の日付　平成１９年３月２２日
学位授与の要件　課程博士（学位規則第４条第１項）
学位授与の題目　状態遷移モデルを持つ強化学習における政策信頼性を考慮した学習効率化
論文審査委員（主査）　泉田　啓（自然科学研究科・助教授）
論文審査委員（副主査）　藤原　直史（自然科学研究科・教授），神谷　好承（自然科学研究科・教授），
関　啓明（自然科学研究科・助教授），滑川　徹（自然科学研究科・助教授）

英文要旨

This study proposes four methods to boost reinforcement learning. The first method makes a policy evaluation step more efficient. This method is based on learning rules with the Robbins-Monro estimation for the state transition probability. It enables us to determine an appropriate learning factor. Furthermore, the learning speed is accelerated by applying acceleration methods of iterative solutions for an inverse matrix. The second method recalculates the Q-factors efficiently for partial variations. This method is based on Sherman-Morrison formula, which is related to partial inverse matrix computations. It modifies the Q-factors accurately at one update without iterations. The third method gives us the required number of samples. This study uses the reliability of optimal policy as a criterion. The sampling condition is derived to guarantee the desired reliability of optimal policy. Two approximated solving methods are developed because the exact solution is difficult. The fourth method composes a suitable state space for the learning. This study considers both optimality and reliability of a policy as a criterion and defines the optimization problem in which a low-order state space is decided to optimize the criterion. A proposed method solves the problem approximately to avoid huge amount of calculation for the exact solution. Some numerical simulations and experiments show that the proposed methods are effective.

和文要旨

一般的な制御系設計では，設計者が予め用いられる環境を想定し，その環境に対して制御系を構築する．この設計法は工場などの整備された環境では有効であるが，多様な状況が存在する環境では，状況を予測し制御系を準備する必要がある．結果的に，設計者が試行錯誤的に調整あるいは再設計を繰返し，多くの労力と開発コストを費やすことになる．また，環境が複雑になるに従い多様な状況を完全に予測することは実質的に不可能なため，このアプローチは現実的でない．これに対し，機械学習やロボット工学等の分野で強化学習という方法が研究されている．強化学習は，試行錯誤を通じてコストを最小化する行動則を学習する枠組みである．設計者が模範となる行動を与えなくても，コストによりタスク達成の条件さえ与えれば，自動的に適切な行動則を生成する教師無し学習である．最も一般的には，離散的な状態と行動の空間を用い，行動とその結果得られる状態の遷移を観測することにより状態遷移確率を獲得し，それに対する制御系設計を学習により行なう．これにより，膨大な状況に対処できるとともに，実環境で学習することにより予期せぬ状況に対しても適応性を発現する．一方で，強化学習には，学習に要する時間が膨大であるという課題があり，実用化への妨げとなっている．この課題の背景には，未だ発展途上の領域であるため，学習の構造が十分に理解されておらず，アルゴリズムが非効率なことがある．現実的なタスクを達成可能にするためには，効率的に学習する学習構造とアルゴリズムが不可欠である．そのため本研究では，まず強化学習の構造を明らかにし，学習構造に注目した学習速度の加速法を４つ提案する．

本論文は８章で構成され，各章の概要は以下の通りである．

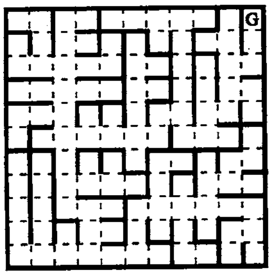

Fig. 1: 11×11 maze

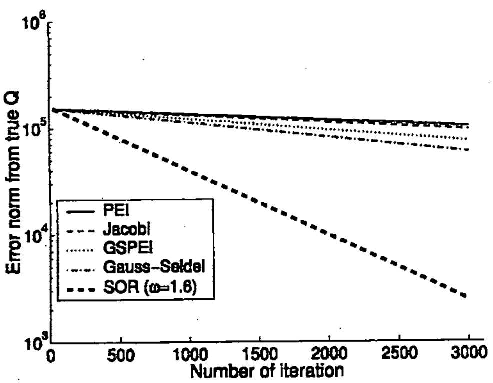

Fig. 2: Comparison of covergence speed

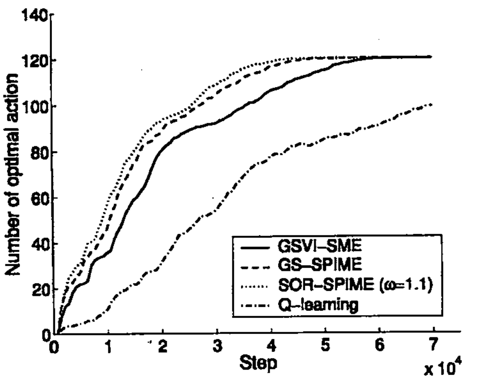

Fig. 3: Number of states to obtain optimal action

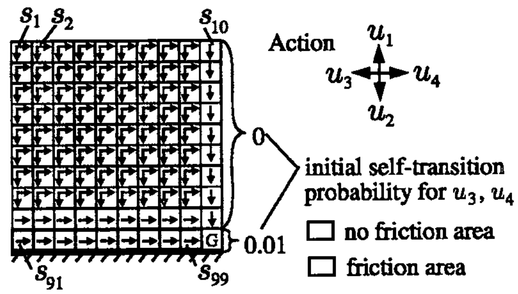

Fig. 4: Friction problem

第1章では，本研究の背景と目的，各章の概要を述べる．

第2章では，強化学習問題と代表的解法について説明する．一般的な強化学習問題では，離散的な状態と行動の対に，将来にわたるコスト期待値であるQ値という評価値を付け，Q値を最小化する政策すなわち各状態で選択する行動を決定する．行動はQ値に従い選択するので，最適政策を導く最適Q値を獲得することが学習目的である．その解法は，状態遷移確率推定，政策評価，政策改善の3動作からなる．状態遷移確率推定は，未知あるいは不確定な系の状態遷移確率を推定する動作で，確率的サンプリングを用いることが多い．政策評価は，行動のコスト，状態遷移確率，政策から代数方程式の解としてQ値を求める動作であり，その多くは逆行列の反復法に基づく．政策改善は，求められたQ値に基づき各状態の最適な行動として政策を改善する動作である．代表的解法には，価値反復やQ学習などがある．価値反復は推定された状態遷移確率を用いて残りの2動作を逐次行なう方法である．Q学習は3動作を逐次行なう方法で，価値反復の近似解法として導くことができるが，価値反復に比べて学習の収束が遅い．

第3章では，政策評価を効率化する方法を考える．本研究では，価値反復などにRobbins-Monro確率近似アルゴリズムに基づく状態遷移確率の推定則を組込む効率的な学習則を提案する．提案法は，Q学習に比べ同じサンプル数に対するQ値改善速度が速く，既存の方法では定まらない学習率に対応するパラメータも適切に決定できる．さらに，政策評価の構造に注目し，反復法の加速法であるGauss-Seidel法，SOR法などを応用した更新則を示す．

Fig. 1に示す11×11の迷路を用いて検証する．行動は上下左右の4つとする．何れの行動も確率10%で現在の状態に留まり，コストとして1を生じる．まず，真の状態遷移確率とランダム政策に対する政策評価で得られたQ値の誤差ノルムの変化をFig. 2に示す．図の横軸の1反復は全状態行動対について1度ずつQ値を更新することを意味する．政策評価で一般的に用いられるQ値更新則であるPEI (Policy Evaluation Iteration)に比べ，加速法のQ値改善速度が速いことが分かる．つぎに，状態遷移確率と政策を逐次更新する場合に同じQ値更新則を適用した結果をFig. 3に示す．Q学習に比べ，Gauss-Seidel法，SOR法を応用したGS-SPIME，GSOR-SPIMEは，3.3倍，3.8倍早く全状態の最適行動を獲得する．

第4章では，現実と予測モデル間に誤差や変動があり，遷移確率を逐次部分的に変更する場合，あるいは行動のコストや行動選択確率を部分的に変更する場合に，部分的変更がQ値や政策に与える影響を効

Table 1: Computational effort for $n_{pv}=2$

	×, ÷	time [s]	ratio to SPI
PMA	5.3×10^{2}	6.4×10^{-3}	0.53
API	1.4×10^{3}	8.1×10^{-3}	0.67
SPI	3.2×10^{4}	12.1×10^{-3}	1.0

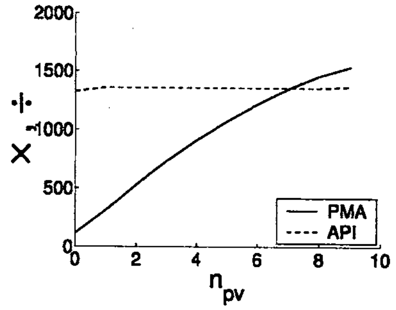

Fig. 5: Computational effort for n_{pv}

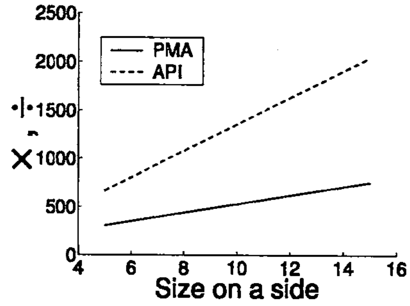

Fig. 6: Computational effort for the size of state

率よく計算する方法を考える．本研究では，Q 値の効率的な再計算法として PMA (Partial Modification Algorithm) を提案する．一般的な強化学習法は反復法に基づくが，提案法は逆行列を部分的修正する直接法である Sherman-Morrison 公式に基づくため，一度で正確に Q 値を修正することが可能である．そのため，部分的な Q 値再計算において計算量を低減化できる．

Fig. 4 に示す状態空間を用いて検証する．一辺が 10 状態の正方形の空間で，壁沿いの状態 $s_{91}\sim s_{99}$ は左右方向に生じる摩擦を同じ状態に留まる確率 0.01 としてモデル化する．行動は上下左右の 4 つとする．最適 Q 値及び最適政策 (Fig. 4) が予め計算されている．しかし，実際には (s_{99},u_4) に対する摩擦が想定よりも大きいとする．この差を状態遷移確率の部分的変化とし，生じる政策改善も考慮した学習問題に PMA を適用した結果を Table 1 と Fig. 5 に示す．表および図中の n_{pv} は最適行動が変化した状態数であり，乗除算の計算量を示す．加減算についても同様のオーダーとなる．SPI は一般的な政策反復であり，

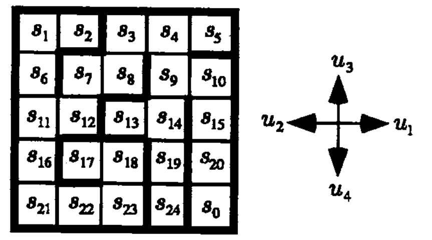

Fig. 7: 5×5 maze problem

API は非同期政策反復である．部分的に修正を行なうため SPI に比べ計算量が少なく，n_{pv} が小さい場合に PMA は API に比べ計算量が少ない．また，同様の正方形領域を考え，一辺の大きさを変化させた場合の計算量を Fig. 6 に示す．ただし，状態空間の大きさによらず，$n_{pv}=2$ とする．n_{pv} が一定であれば，状態空間が大きくなるほど PMA が有効であるといえる．

第 5 章では，必要なサンプル数の適切な算出法を考える．状態遷移確率の推定は，それから導かれる政策に影響するため，推定精度が問題になる．十分な推定精度に到らなければ，得られる最適政策の信頼性向上のため，サンプル数を増やす必要がある．しかし，その最適政策の信頼度がわからなければ，いつサンプリングを止めて良いか決まらず，必要以上に多くのサンプル数まで続けざるを得ない．本研究では，所望の信頼度で最適政策を保証する明確なサンプリング条件を導く．

サンプリング条件を求める流れを概説する．まず，唯一の正しい最適政策 μ^* を導く遷移確率の集合 $\mathcal{P}^{\bar{\mu}^*}$ を定める．この集合はある広がりをもち，その元である遷移確率はいずれも μ^* を最適解とする．つぎに，サンプリングにより推定した遷移確率について，信頼度 ℓ で真の確率が存在する集合 $\mathcal{P}^{\ell z}$ が定まる．ゆえに，$\mathcal{P}^{\ell z}\subseteq\mathcal{P}^{\bar{\mu}^*}$ なら，少なくとも信頼度 ℓ で正しい μ^* が導かれる．$\mathcal{P}^{\ell z}$ の大きさはサンプル数の増加とともに減少するので，上の条件を満たすサンプル数を求めることになる．

サンプリング条件を正確に求めることは困難であるため，2 つの近似解法を提案する．1 つは最大特異値に基づく方法である．利点として，簡単なアルゴリズムとなる点，現在の推定確率の精度が不十分な場合でも比較的安定に条件が求まる点があり，欠点として，十分条件を用いるためサンプル数が多くなる点，特異値分解による計算コストが大きい点がある．もう 1 つは現在の推定確率の精度に基づく方法である．利点として，最大特異値に基づく方法に比べ計算量を少なくできる点があり，欠点として，推

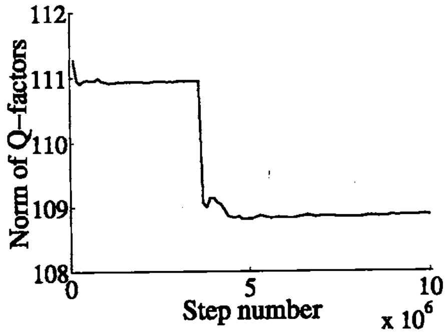

Fig. 8: Norm of Q-factors calcurated based on current estimate probability

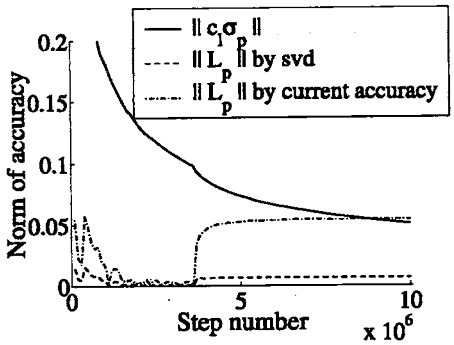

Fig. 9: History of $||c_\ell \sigma_p||$ and $||L_p||$

定確率の一部の精度が悪い場合に，十分な精度が得られるまでサンプリング条件が変動し易い点がある．

Fig. 7 に示す 5×5 迷路問題で，提案法によるサンプリング終了条件と一般的に用いられる Q 値に基づくサンプリング終了条件を比較検証する．行動は上下左右の 4 つとし，壁への移動では同じ状態に留まる．いずれの行動も確率 10[%] で同じ状態に留まり，コスト 1 を生じる．信頼度は ℓ=0.954 とする．100 ステップに 1 度の割合で推定確率 $\bar{p}$ を更新し，最適 Q 値と最適政策を再計算する．挙動政策は $\epsilon = 0.1$ の ϵ-greedy 政策とする．Q 値に基づく終了条件は，各状態行動対に対して過去 10 回分の Q 値履歴から標準偏差を算出し，すべて閾値以下となれば終了と見なす．シミュレーションを 100 回行なった結果，Q 値に基づく終了判定では 20 回誤った最適政策を導いたが，どちらの提案法でも 0 回であった．Fig. 8, 9 に示す結果を用いて，その理由を説明する．ステップ数が 4.0×10^6 程度で遷移確率の推定精度 $c_\ell \sigma_p$ の向上により greedy 政策が変化し，それに伴ない Q 値のノルムが大きく変化している．そのステップ数に到るまで長い間，政策が変化せず，現在の推定確率の変化が僅かである場合，推定確率や Q 値の変化から終了条件を決定することは困難である．一方，提案法では，greedy 政策に応じて要求精度 L_p と要求

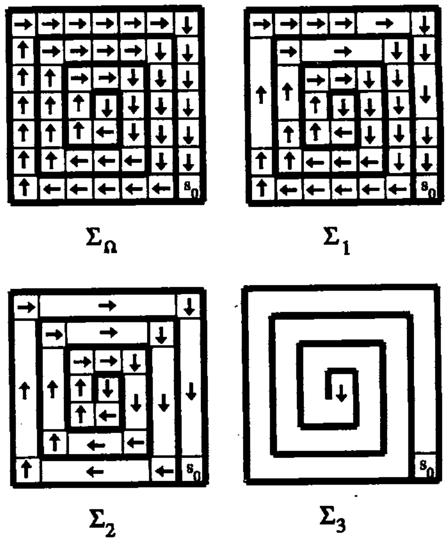

Fig. 10: Optimal policy on the each state space

サンプル数を求めるため，政策や Q 値が一旦定常になるだけでは終了しない．従って，提案法は適切な終了条件を与えるといえる．

第 6 章では，学習に適した状態空間を構成する．限られたサンプルで推定確率の信頼性を向上させるには，状態空間を粗くして 1 状態あたりのサンプル数を増やす方法がある．それにより，推定確率の信頼性が向上し，導かれる最適政策の信頼性が向上する．一方，状態空間を粗くし過ぎると，状態空間の解像度が不足するため，政策の最適性が低下する．これに対し，本研究では最適政策の最適性と信頼性を両立する低次状態空間決定問題を考える．

この問題を完全に解くためには，候補となり得る全ての状態空間を用いて最適政策を求めて比較する必要があり，莫大な計算時間を要す．そこで，低次状態空間の候補を少数構成し，その中から最良の低次状態空間を選択する近似解法を提案する．また，状態を統合するアルゴリズムにより低次状態空間の候補を構成する方法を示す．

迷路問題を用いて検証する．Fig. 10 の Σ_Ω の迷路上でランダムウォークによりサンプルを得た後，提案した状態統合アルゴリズムを適用し，Σ_1，Σ_2，Σ_3 の低次状態空間候補を得た．その候補に対して提案法により低次状態空間決定問題を解くと，状態空間 Σ_Ω，Σ_1，Σ_2 を用いて導かれる最適政策が正しい最適政策となり，所望の信頼度 $\ell = 0.954$ を与える状態空間は Σ_2 と Σ_3 となる．それゆえ，適切な状態空間として Σ_2 を選択できる．また，Σ_Ω の 65.6[%] 程度のサンプル数で信頼性を保証し，学習を効率化できる．

第７章では，総合的な検証として，組立作業の代表的タスクであるペグ・イン・ホールで想定される２つの問題に，第３章から第６章で提案した方法を組合せて適用する．まず，ホールに残っていた想定外のバリに対し，遷移確率の変化を捉えるために適切な再サンプリングの終了条件を得る．この変化に対し，僅か 2.7[msec] で正確に Q 値を再計算し，適応的に行動を変更できることを確認する．また，Q 学習に対して，第３章の方法により 6.3 倍，第６章の方法と組合せることにより 116 倍の学習の加速効果が得られる．

第８章を本論文の結論とする．

学位論文審査結果の要旨

当該学位論文に関し，平成１９年１月３０日に第１回審査委員会を開き，面接調査を行った後，論文内容を詳細に検討した．さらに，平成１９年２月５日に行われた口頭発表の後に第２回審査委員会を開き，協議の結果以下のように判定した．

本論文では，必要なデータや計算コストが膨大という課題をもつ強化学習について，未だ十分に解明されていない学習構造を整理し，それに基づいて学習加速法を４つ提案する．一般的解法は，状態遷移の確率推定，政策評価，政策改善の３動作からなるが，まず政策評価が代数方程式の求解と同じ構造であると理解される．それに基づき系統的に政策評価の加速法を提案する．第二の方法では問題が部分的に変化する場合を扱い，先の理解に基づく検討の結果，一般に反復解法による政策評価を Sherman-Morrison 公式に基づく直接解法とする方が効率的であることを示す．次に試行錯誤的なサンプリングによる確率推定精度が，最適政策にどう影響するか整理する．それに基づいて，第三の方法では最適政策を所望の信頼度で求めるために必要な条件を求め，第四の方法では政策の信頼度と最適性を高める状態空間の構成方法を示す．これらにより，必要サンプリング数を減らして確率推定を効率化できる．さらに，ロボットによる組立作業に提案手法を総合的に適用し，100 倍を超える加速効果を得ている．これらは強化学習を実用化する上で意義ある成果と結論できる．よって本論文は博士（工学）論文に値する．