DEIM2021 E25-4

カルマンフィルタを利用したロバストQ学習

高畑　慶†　三浦　孝夫†

† 法政大学大学院 理工学研究科 システム理工学専攻　〒 184–8584 東京都小金井市梶野町 3-7-2
E-mail: †kei.takahata.6a@stu.hosei.ac.jp, ††miurat@hosei.ac.jp

あらまし　強化学習は，環境からの報酬で知識を自動的に抽出できるが，学習に時間がかかるという問題点がある．また環境は定常とは限らないため，知識を常に更新し続ける必要がある．QLRKF では失敗した行動と逆の行動を選択し学習することで知識を改善していたが，カルマンフィルタの状態と分散を巻き戻すために事前誤差共分散行列をメモリに保存していた．本研究では解析的にカルマンフィルタの状態と分散を巻き戻すことにより，余分なメモリを使用せず，かつ観測ノイズの大きい場合にも正常に学習できるモデルを提案する．協調距離問題というタスクで実験を行い，提案手法の有用性を示す．

キーワード　強化学習，Q 学習，カルマンフィルタ，遡及的カルマンフィルタ，逆行動学習

1 前　書　き

強化学習 [1] [2] とは，動作主（エージェント）が環境との相互作用から学習を行う学習手法である．強化学習は，学習データを用いず，環境からの報酬で知識を自動的に抽出できるため，知識や戦略を獲得するための強力な学習手法として知られる [10]．しかし，学習に時間がかかるという問題点がある．ロボットなど，実世界で利用する場合，多くの経験を行うことは時間的なコストがかかる．そのため，学習の経験数を減らすことを目的とした様々な手法が提案されいる [4] [5]．これらの研究は，効率よく質の高い知識をどのように獲得するかに焦点をあてている．しかし，知識を蓄積してもその失敗を利用して，知識の改善を行う手法は少ない．Hindsight Experience Replay(HER) [7] はエージェントの失敗行動を有効に活用している手法であるが，タスクが定常だと仮定しているため，非定常なタスクに対応できない．QLRKF [14] では失敗した行動と逆の行動を選択し学習することで知識の改善を行うが，そのために事前誤差共分散行列をメモリに保存していた．またカルマンフィルタを利用する場合，観測ノイズの分散の大きさにより状態予測の精度向上に限界がある．

本研究では観測ノイズの分散が大きく，精度の低い状態予測しか行えない場合でも知識改善を行うことを目的とする．また，解析的にカルマンフィルタの状態と分散を巻き戻すことにより，余分なメモリを使用せず学習するモデルを提案する．協調距離問題というタスクで実験を行い，提案手法の有用性を示す．

第 2 章で強化学習について述べ，第 3 章ではカルマンフィルタについて述べる．第 4 章で提案手法について述べ，第 5 章では提案手法の有用性を実験で示し、第 6 章で結論とする．

2 強化学習と Q 学習

2.1 強 化 学 習

動作主（エージェント）が自身の状態の知覚，意思決定を行い，環境との相互作用から知識を得る手法を強化学習と言う．強化学習では，教師あり学習のように明示的な正解は与えず，エージェントの行動に対して環境から与えられる正負を含む報酬から学習を行う．エージェントが報酬の総和を最大にすることが強化学習の目的である．

エージェントは現状態を知覚し，その状態において行動を起こして次状態に移り報酬を得る．エージェントが時刻 t で行動 a_t を実行した時，時刻 t+1 での状態 s_{t+1} と遷移した時に得る報酬 r_{t+1} を考える．（以下，状態を state の s，行動を action の a，報酬を reward の r を用いて表す．添え字は時刻を表す．）最も一般的な場合，遷移後の状態 s_{t+1} と遷移した時に得る報酬 r_{t+1} は，時刻 t 以前の全ての状態，報酬と行動に依存する．時刻 t+1 にとり得る全ての状態と報酬をそれぞれ，s'，r と表す場合，時刻 t+1 での状態 s_{t+1} と遷移した時に得る報酬 r_{t+1} は条件付確率を使い，

$$Pr\{s_{t+1} = s', r_{t+1} = r | s_t, a_t, r_t, s_{t-1}, \cdots, r_1, s_0, a_0\} \quad (1)$$

と表せる．状態と報酬がマルコフ性（次の状態と報酬が，現在の状態と行動のみに依存する性質）を満たす場合は，

$$Pr\{s_{t+1} = s', r_{t+1} = r | s_t, a_t\} \quad (2)$$

と表せる．環境がマルコフ性を満たし，現在の状態と行動が与えられている場合，式 (2) より次の状態と報酬を予測することができる．更に式 (2) の反復計算を行う事で，現在の状態のみから将来の状態と期待される全ての報酬を予測することができる．強化学習では，エージェントの行動と価値関数（後述）が現在の状態のみに依存した関数であると仮定している．

全ての状態 S のそれぞれの状態 s において，行える行動 a と，その行動をする確率をまとめたものを政策と言う．エージェントの目的は，与えられた問題を効果的に解く政策 π，あるいは，報酬の総和を最大にする政策 π を獲得することである．エージェントが適切な政策を獲得するために価値関数を用いる．本研究で使用する Q 学習 [8] では価値関数として行動価値関数（Q 値）を使用する．Q 値は，政策 π のもとで状態 s において

行動 a を行ったときの報酬の総和の期待値を表す．ここで，状態 s で行動 a をとった時の Q 値を $Q(s,a)$ と書く．Q 値は適切に更新されれば，期待報酬値に近づいていく．

報酬の総和を最大にするために，エージェントは今までの経験から得た“知識の利用”と今より良い政策を見つけるための“探査”が必要となる．“知識の利用”と“探査”は互いにトレードオフの関係にある．エージェントは両者をバランスよく行い，報酬の総和を最大にする知識の獲得をしなければならない．エージェントは政策をもとに行動を行うので，価値関数を政策に変換し，その政策から行動を決定する必要がある．しかし，政策は価値関数の結果を確率として解釈したものなので，大小関係を考えるうえでは，価値関数の値のみを見ればよい．代表的な行動選択手法として，最も価値関数の値が大きいものを選ぶ「グリーディ手法」，確率 $1-\varepsilon$ でグリーディ手法を行い，確率 ε でランダムな行動を選択する「ε グリーディ手法」，価値関数の比を計算し，比の大きいものを高い確率で選択する「ソフトマックス手法」が知られている．「グリーディ手法」は“知識の利用”のみを行う手法で，「ε グリーディ手法」，「ソフトマックス手法」は“知識の利用”と“探査”を行う．「ε グリーディ手法」，「ソフトマックス手法」でエージェントの学習を行い，「グリーディ手法」で評価を行うのが一般的である．

2.2 Q 学 習

強化学習の代表的な学習手法として，Q 学習が知られている．Q 学習は，マルコフ決定過程の環境では，学習率が更新回数ごとに小さくなるなど適切に調整されれば，無限時間での最適解の収束が証明されている [8]．状態 s で行動 a を実行するときの Q 値を $Q(s,a)$，実行したときに得られる報酬を r，実行後の状態 s' において最大の Q 値となる行動 a' を実行するときの Q 値を $\max_{a' \in A(s')} Q(s',a')$，1 回の学習での更新の割合を表す学習率を α，将来獲得予定の報酬を考慮する割合を表す割引率を γ とすると，Q 値の更新式は

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a' \in A(s')} Q(s',a') - Q(s,a)] \quad (3)$$

と表せる．$Q(s,a)$ を，$r + \gamma \max_{a' \in A(s')} Q(s',a')$ に近づくように更新している．Q 値の更新は，エージェントが 1 つの行動を実行し，次の状態に移るごとに行われる．

2.3 逆行動学習

学習に失敗した場合，それまでに至る行動と逆の行動の方がベターという仮定を置き，学習する方法を逆行動学習と言う．逆行動学習を行うことで Q 値を効率よく学習できる．失敗した場合，通常の学習で失敗に至る行動の価値は下がる．逆行動学習では，逆の行動を選択し，絶対値の逆の報酬を受け取るため，逆の行動の価値を上げることができる．失敗に至る行動の価値を下げ，逆の行動の価値を上げることで，効率よく学習できる．

3 カルマンフィルタ

観測値から状態を推定するフィルタリング手法としてカルマンフィルタ [11] が知られている．カルマンフィルタは，観測値と状態に分け現象をモデル化する状態空間モデルを用いてフィルタリングを行う．状態空間モデルは，1 ステップ後の状態と現在の状態を関連付ける状態方程式，観測値と状態を関連付ける観測方程式の 2 つの式から表現される．現在の状態を X_k，1 ステップ後の状態を X_{k+1}，現在の状態と 1 ステップ後の状態を関連付ける係数行列を A，時刻 k から k+1 に移る過程で生じる誤差（プロセスノイズ）を V_k，プロセスノイズの係数行列を B とすると状態方程式は

$$X_{k+1} = AX_k + BV_k \quad (4)$$

と表せる．また時刻 k での観測値を Y_k，観測値と状態を関連付ける係数行列を C，観測時に生じる誤差（観測ノイズ）を W_k とすると観測方程式は，

$$Y_k = CX_k + W_k \quad (5)$$

と表せる．プロセスノイズ V_k の共分散行列を Q，観測ノイズ W_k の共分散行列を R とすると，各ノイズは $p(V) \sim N(0,Q)$, $p(W) \sim N(0,R)$ に従う．ここで，$p(x) \sim N(0,w)$ は，平均 0，分散 w の正規分布を示し，互いに関連しないことから白色雑音と呼ばれる．

カルマンフィルタでは，新たな時系列データが入るたびに逐次的に状態推定値を更新する．現時刻を k とし，1 時刻前 $k-1$ までに利用可能なデータに基づき時刻 k の状態 X を推定したものを事前状態推定値と言い，$\hat{X}_k^-$ と表す．また，時刻 k での観測値 Y_k を用いてフィルタリングを行った後の状態推定値を事後状態推定値と言い，$\hat{X}_k$ と表す．状態の誤差の共分散行列を P で表記し，事前誤差共分散行列を P_{k-1}^-，事後誤差共分散行列を P_k と表記する．フィルタリングにより，事後誤差共分散行列の要素を小さくすることで，予測の精度を上げることができる．時刻 k での更新の割合を表すカルマンゲイン行列を G_k とすると，カルマンフィルタによるフィルタリングの流れは以下のように表せる．

予測ステップ

$$\hat{X}_k^- = A\hat{X}_{k-1} \quad (6)$$

$$P_k^- = AP_{k-1}A^T + BQB^T \quad (7)$$

フィルタリングステップ

$$G_k = P_k^- C^T (CP_k^- C^T + R)^{-1} \quad (8)$$

$$\hat{X}_k = \hat{X}_k^- + G_k(Y_k - C\hat{X}_k^-) \quad (9)$$

$$P_k = (I - G_k C)P_k^- \quad (10)$$

初期値として，状態の初期値 $\hat{X}_{k-1}$ と誤差共分散行列の初期値 P_{k-1}，ノイズの共分散行列 Q, R を設定する必要がある．

カルマンフィルタは誤差共分散を小さくすることで，状態予測の精度を上げる．カルマンゲインは観測値から状態を更新する割合を表している．具体的に，事前状態の予測誤差の分散が大きい（事前状態が信頼できない）場合と観測ノイズが小さい（観測値が信頼できる）場合，観測値の方が信頼できるため，事

前状態を大きく更新するためにカルマンゲインも大きくなる．逆に，予測誤差の分散が小さい場合と観測ノイズが大きい場合，観測値よりも状態遷移の方が信頼できるため，カルマンゲインは小さくなる．誤差共分散が最小になるように式 (8) でカルマンゲインを計算している．

カルマンフィルタと強化学習を組み合わせたもので，KTD [13] [12] が提案されている．KTD では，連続値強化学習（状態が連続値の場合の強化学習）の重みパラメータを推定している．KTD では，初期パラメタ依存性問題が生じる．提案手法では，エージェントの行動選択にカルマンフィルタを利用するので，KTD とは根本的に異なる．

4 提案手法

4.1 遡及的カルマンフィルタ

本研究では，Q 値更新の結果によって逆行動学習を行うことで，効率的に学習する手法を提案する．エージェントに失敗条件を定義し，失敗条件を満たさない場合は学習を継続する．失敗条件を満たす場合，逆行動学習を行う．状態を巻き戻したい場合，すべての状態遷移の履歴，すべての行動履歴を保持していれば，状態をもとに戻すことが可能である．しかし，状態数や行動数，経験数が多くなるにつれ，保持しなければならないデータ量が増え，現実的ではない．過去の状態を持たず，逆行動学習を行うことを目的として，現在の状態から 1 ステップ前の状態を予測する遡及的カルマンフィルタを利用する．カルマンフィルタでは 1 ステップ前の事後状態推定値 $\hat{X}_{k-1}$，1 ステップ前の事後誤差共分散行列 P_{k-1} から，現在の事後状態推定値 $\hat{X}_k$，現在の事後誤差共分散行列 P_k を推定する．遡及的カルマンフィルタでは現在の事後状態推定値 $\hat{X}_k$，現在の事後誤差共分散行列 P_k から 1 ステップ前の事後状態推定値 $\hat{X}_{k-1}$，1 ステップ前の事後誤差共分散行列 P_{k-1} を推定する．

遡及的カルマンフィルタを以下に定義する．事前誤差共分散行列を解析的に求めている点が既存のものと異なる．解析的に求めることで，余分なメモリを使用せずに演算することが可能になる．

遡及フィルタリングステップ

$$P_k^- = (I - P_k C^T R^{-1} C)^{-1} P_k \tag{11}$$

$$G_k = P_k^- C^T (C P_k^- C^T + R)^{-1} \tag{12}$$

$$\hat{X}_k^- = (I - G_k C)^{-1} (\hat{X}_k - G_k Y_k) \tag{13}$$

遡及予測ステップ

$$P_{k-1} = A^{-1}(P_k^- - BQB^T)(A^T)^{-1} \tag{14}$$

$$\hat{X}_{k-1} = A^{-1}\hat{X}_k^- \tag{15}$$

4.2 QLRKF

遡及的カルマンフィルタを利用し，逆行動学習を行えるようにした学習手法を QLRKF と定義する．逆行動学習を行う場合，状態と事後誤差共分散行列を巻き戻す必要がある．状態と事後誤差共分散行列を巻き戻すことを目的として，遡及的カルマンフィルタを利用する．

逆行動学習を行うため，QLKF [9] を利用する．QLKF では，確率 ε でカルマンフィルタの状態予測の結果を利用した行動選択，確率 $(1-\varepsilon)$ でグリーディな行動を選択する．

QLRKF は通常の学習時は QLKF と同じように学習する．学習に失敗し，逆行動学習を行う場合，確率 ε で 1 ステップ前の状態予測を行い，通常学習時に選択すべき行動と逆の行動選択，確率 $(1-\varepsilon)$ でグリーディな行動と逆の行動を選択する．

5 実　　験

5.1 協調距離問題

本研究では，協調距離問題というタスクを扱う．この問題は 2 体のエージェント（A, B）がある一定の範囲内（協調距離範囲）にいる場合，両者に正の報酬を与えるタスクである．m × m の連続 2 次元空間のフィールドで実行する．エージェントを 2 体用意し，フィールド外には移動できないように設定する．初期条件として，A と B の距離が，一定の距離以上離れるようにランダムに配置し，開始する．協調距離範囲の上限より離れすぎている場合は 2 体のエージェントに負の報酬を与え，下限より近づきすぎた場合，一方のエージェントに正の報酬，他方に負の報酬を与え，エージェントの位置の初期化を行う．

各エージェントは，自分を中心とした相手の相対位置が分かるものとする．強化学習は，状態と行動を離散値で扱うので，連続空間を離散化する必要がある．本研究では，自分との相対位置で状態の離散化を行う．具体的には，自分を中心として，45 度ずつに区切った計 8 つの角度方向と，各方向で自分との距離が一定数以内か，否かで離散化を行う．(以後，強化学習のために離散化した状態をエリアと表現する.）つまり，8 つの方向と指定距離より遠いか近いかで，16 のエリアに離散化する．さらに確率としては少ないが同じ座標にいる場合も考えられるので，同じ座標いる場合の計 17 個のエリアに離散化する．具体的に，図 1 の場合，ハンタから見た獲物の位置は 4 の範囲の中に含まれるので，ハンタが知覚したエリアは 4 となる． 行動は，上下左右と各斜め方向の計 8 つと，停止の計 9 つに設定し，学習を行う．図 2 のように A がフィールドの中央付近，B がフィールドの左上にいる場合，A は全方向と停止の行動が可能である．しかし，B はこれ以上，上や左に進むとフィールド

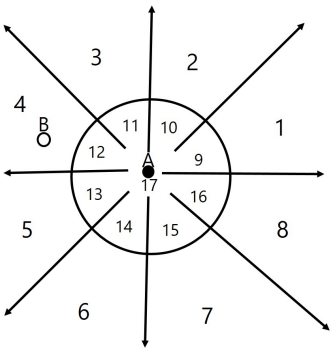

図 1　A's　state

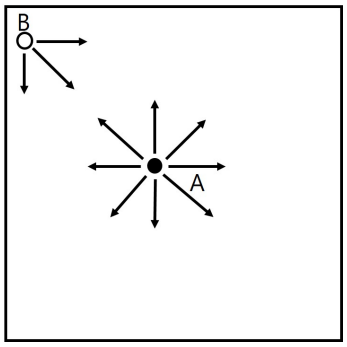

図 2　Examples of A and B behavior

外に出てしまう．このような場合，前述のとおり，B は右，右下，下，停止の行動しかできないように設定する．初期位置から，B と A は交互に，相手のエリアの知覚，行動，学習を繰り返す．具体的に，A と B が学習する場合の例を示す．

A と B がともに Q 学習で学習する場合，以下の a から g を繰り返す．

a. A と B を配置

b. B が“A がいるエリア”を知覚し，行動

c. A が“B がいるエリア”を知覚し，行動

d. A と B が，距離状況に応じた報酬を受容

e. B が“A がいるエリア”を知覚し Q 学習

f. A が“B がいるエリア”を知覚し Q 学習

g. 協調距離以内であれば a に戻り，それ以外であれば b に戻る

提案手法では，A が一定回数連続で協調距離にいない場合，逆行動学習を行う．逆行動学習では，以下の a から e を一定回数繰り返す．

a. A が“B がいるエリア”を知覚し，行動

b. B が“A がいるエリア”を知覚し，行動

c. A と B が，報酬を受容

d. A が“B がいるエリア”を知覚し Q 学習

e. B が“A がいるエリア”を知覚し Q 学習

5.2 実験準備

エージェント 2 体 (A, B) で協調距離問題を行う．本研究では，以下の 4 パターンで実験を行い，提案手法の効果を調べる事とする．

・A と B がともにランダムに行動（比較手法 1）
・A と B がともに Q 学習（比較手法 2）
・A が QLKF で B が Q 学習（比較手法 3）
・A が QLRKF で B が Q 学習（提案手法）

評価には協調距離範囲内のステップ数を用いる．協調距離範囲内のステップ数が多いほど，良い学習ができていると判断する．

A が QLKF（比較手法 3），QLRKF（提案手法）で学習する場合，A は B の位置を予測するカルマンフィルタを使用し，確率 ε で予測した位置を利用した行動を選択する．具体的に，予測した位置と自分の位置との距離を計算し，協調距離範囲外であれば予測した位置に一番近づく行動，協調距離範囲内であれば停止を選択する．

フィールドは 2 次元座標空間の $(x, y) \in [0, 1]^2$ を使用する．初期条件として離す距離は 0.8，協調距離範囲 d を $0.1 < d \leqq 0.4$ とする．エージェントの距離が協調距離範囲内の場合，A と B の報酬を+5，協調距離範囲の上限より大きい場合，A と B の報酬を-10，下限より小さい場合，A の報酬+50，B の報酬-50 とする．提案手法の逆行動学習時の A と B の報酬は+10 とする．（通常と絶対値の逆の数値とする．）逆行動学習時に協調距離範囲に入った場合，逆行動プロセスを終了する．

A，B 共に学習時の学習率は 0.1，割引率は 0.9 に設定する．また行動選択時の ε も共に ε=0.1 に設定する．

提案手法では 2 体のエージェントの距離が 0.4 より上 (協調距離範囲の上限より上) の状況が 50 回連続した場合を失敗条件とする．カルマンフィルタの状態と分散を巻き戻すため，逆行動プロセスの上限は 50 回である．

QLKF と QLRKF のカルマンフィルタの初期値は，誤差の共分散行列の対角成分を $10^4 I$，プロセスノイズの共分散行列の対角成分を $0.05I$，観測ノイズの共分散行列の対角成分を 1 とする．観測ノイズの分散を 1 にすることで状態予測の誤差が大きい場合での結果を調べることが可能となる．QLKF と QLRKF で学習するハンタは，カルマンフィルタで獲物の相対位置を予測するために，獲物の相対位置（相対座標）を観測する．ハンタは，前回のフィルタリング後の獲物の相対位置を用いて，現在の相対位置を予測する．時刻 t でのハンタから見た獲物の相対 x 座標を x_t，相対 y 座標を y_t，ハンタの x 座標の速度を hv_{xt}，y 座標の速度を hv_{yt}，対角成分に時刻 t のプロセスノイズがまとめてある対角行列を V_t，対角成分に時刻 t の観測ノイズがまとめてある対角行列を W_t と定義し，状態方程式を式 16，観測方程式を式 17 のように設定する．

$$\begin{pmatrix} x_{t+1} \\ y_{t+1} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_t \\ y_t \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} hv_{xt} \\ hv_{yt} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} V_t \tag{16}$$

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_t \\ y_t \end{pmatrix} + W_t \tag{17}$$

5.3 協調距離範囲内のステップ数の評価方法

2 体のエージェントがそれぞれ行動をしたとき，1 ステップと定義する．エージェントが 100 回行動（Q 値の更新）した時，1 インターバルと定義する．1 インターバルごとに，協調距離範囲内の累積ステップ数を記録する．1000 インターバル（10 万回の学習）を 1 エピソードと定義し，1 エピソードごとに Q 値をリセットさせ，100 エピソード実行する．各インターバルごとに 100 エピソード分の中央値を求めた結果を，そのインターバル時の協調距離範囲内の累積ステップ数として評価に用いる．

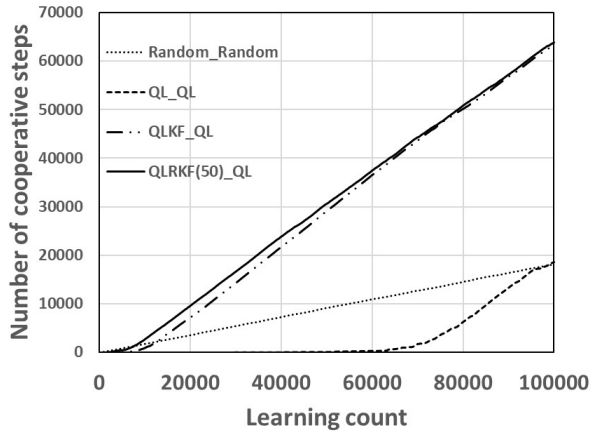

図 3　Cooperative steps

5.4 実験結果

学習回数と協調距離範囲内の累積ステップ数の結果を比較する．結果を表 1 に示す.（データの取得は学習 100 回ごとに行ったが，表にまとめるため，学習 2000 回ごとにまとめなおしている.）協調距離範囲内の累積ステップ数の結果をグラフ化したものを図 3 に示す．表の項目は A の学習方法だけを記載している．

協調距離範囲内の累積ステップ数が多いほど良い結果である．学習時にカルマンフィルタの予測結果を利用するものとそれ以外で結果が大きく異なることが分かる．具体的に学習 10 万回時点での提案手法の改善率は比較手法 1 の 353%，比較手法 2 の 342%，比較手法 3 の 7% となっている．比較手法 3 と提案手法では学習初期に結果が大きく異なっている．学習 1 万回時点では提案手法の改善率は比較手法 3 の 286% の改善率になっている．以上の結果より，提案手法は学習初期に結果が優れていることがわかる．

5.5 考　　察

提案手法と比較手法 3 の結果が学習初期に大きく異なった理由として，学習初期ほど逆行動プロセスによる Q 値更新の影響を大きく受けるためだと考えられる．ロボットなどで応用する場合，学習回数や時間が物理的に限られているので，学習初期に改善されている提案手法は有用だと考える．

提案手法の A の失敗条件を 30 回, 80 回に変更し協調距離問題で実験を行った.（B の学習手法は QL である.）結果の違いが分かりやすい 1 万回までの結果を表 2 に示す.（データの取得は学習 100 回ごとに行ったが，表にまとめるため，学習 200 回ごとにまとめなおしている.）表 2 の項目には A の学習手法のみを記載している．表 2 の結果をグラフ化したものを図 4 に示す．表 2，図 4 より，失敗条件の回数を大きくした方が結果が良くなることがわかる．具体的に，失敗条件が 50 の時と比較して 30 回の場合は 62% に悪化し，80 回の場合は 124% に改善している．

比較手法 2(QL 同士) より比較手法 1(ランダム同士) の方がよくなった理由として，QL 同士の場合，学習回数が 10 万回で

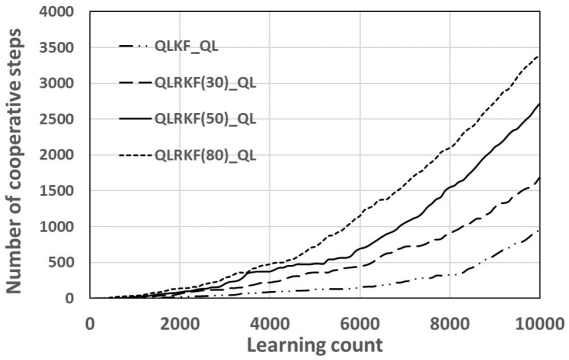

図 4　Cooperative Steps with respect to the number of Failure Condition

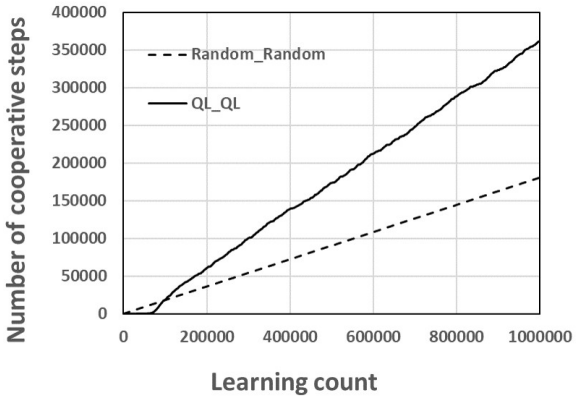

図 5　Cooperative Steps(Both Random and QL)

も学習途中であるのでこのような結果になったと考えられる．学習を 100 万回まで行った結果を表 3 に示す.（データの取得は学習 100 回ごとに行ったが，表にまとめるため，学習 20000 回ごとにまとめなおしている.）表 3 の結果をグラフ化したものを図 5 に示す．表 3，図 5 より，学習回数が多くなると QL の結果が良くなることがわかる．学習回数 100 万回時点では，QL 同士は Random 同士の 2 倍に改善している．しかし，学習回数が足りないと，Random 同士の結果よりも悪くなることがわかる．

6 結　　論

本研究では，事前誤差の共分散行列を保持せず，1 ステップ前の状態予測を行う遡及的カルマンフィルタを利用することで，逆行動学習を行う手法（QLRKF）を提案した．学習 10 万回時点で QL の 342%，QLKF の 7% の改善率になることがわかった．また学習中の 1 万回時点で QLKF の 286% の改善率になることがわかった．QLKF との学習初期の比較より，観測ノイズの分散が大きく，精度の低い予測しか行えない場合でも知識改善を行う事ができるとわかった．

文　　献

[1] Sutton, Richard S., and Andrew G. Barto. Reinforcement learning: An introduction. Vol. 1. No. 1. Cambridge: MIT

press, 1998.
[2] Leslie Pack Kaelbling, Michael L. Littman, Andrew W. Moore. “Reinforcement Learning: A Survey.” CoRR cs.AI/9605103 (1996)
[3] Hado van Hasselt. “Double Q-learning.” NIPS 2010: 2613-2621
[4] Marco A, Wiering, and Hado van Hasselt. “Ensemble Algorithms in Reinforcement Learning.” IEEE Trans. Systems, Man, and Cybernetics, Part B 38(4): 930-936 (2008)
[5] Vukosi Ntsakisi Marivate, Michael L. Littman. “An Ensemble of Linearly Combined Reinforcement-Learning Agents.” AAAI (Late-Breaking Developments) 2013
[6] Benjamin Eysenbach, Shixiang Gu, Julian Ibarz, Sergey Levine. “Leave no Trace: Learning to Reset for Safe and Autonomous Reinforcement Learning.” CoRR abs/1711.06782 (2017)
[7] Marcin Andrychowicz, Dwight Crow, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, Wojciech Zaremba: “Hindsight Experience Replay.” NIPS 2017
[8] Watkins, Christopher JCH, and Peter Dayan. ”Q-learning.” Machine learning 8.3-4 (1992): 279-292.
[9] Kei Takahata, Takao Miura. “Reinforcement Learning using Kalman Filters.” IEEE International Conference on Cognitive Informatics and Cognitive Computing(ICCICC) 2019
[10] 高玉 圭樹.” マルチエージェント学習.” コロナ社, 2004
[11] 足立修一，丸田一郎．”カルマンフィルタの基礎．”東京電機大学出版局．
[12] 北尾 健大，三浦 孝夫：マルチエージェント環境における政策推定, 第 10 回データ工学と情報マネジメントに関するフォーラム (DEIM)，2018, 福井,
[13] Matthieu Geist, Olivier Pietquin: “Kalman Temporal Differences.” J. Artif. Intell. Res. 39: 483-532 (2010)
[14] Kei Takahata Takao Miura “Q-Learning as Failure.” EJC 2020

表 1　協調距離範囲の累積ステップ数

学習回数	Random	QL	QLKF	QLRKF(50)
2000	339	0	22	86
4000	692	0	90	375
6000	1046	0	162	694
8000	1403	0	329	1551
10000	1784	0	951	2718
12000	2113	0	1765	4076
14000	2480	0	3117	5441
16000	2842	0	4504	6793
18000	3182	8	5908	8225
20000	3559	9	7211	9626
22000	3964	16	8549	11046
24000	4324	19	9956	12409
26000	4694	26	11436	13872
28000	5080	28	12889	15280
30000	5424	39	14310	16658
32000	5825	41	15807	18069
34000	6135	43	17285	19528
36000	6502	43	18805	21024
38000	6884	49	20253	22473
40000	7255	73	21740	23877
42000	7670	91	23187	25230
44000	8003	109	24593	26428
46000	8344	109	26095	27760
48000	8735	109	27640	29233
50000	9134	135	29131	30638
52000	9522	195	30571	31970
54000	9802	235	32083	33301
56000	10182	263	33544	34663
58000	10528	317	35026	36027
60000	10900	337	36560	37481
62000	11242	373	37940	38827
64000	11576	654	39421	40180
66000	11924	865	40880	41502
68000	12297	1133	42279	42974
70000	12669	1782	43652	44333
72000	13017	2128	45104	45581
74000	13328	3002	46368	46774
76000	13670	4068	47764	48048
78000	14089	5093	49032	49456
80000	14470	6328	50158	50882
82000	14857	7562	51424	52153
84000	15268	9011	52697	53406
86000	15611	10543	54026	54717
88000	15910	11972	55437	55950
90000	16286	13375	56784	57265
92000	16625	14613	58057	58603
94000	17031	16032	59488	60037
96000	17351	17185	60919	61463
98000	17707	17750	62134	62774
100000	18085	18657	63418	63865

表 2　協調距離範囲の累積ステップ数 (失敗条件を変更した場合)

学習回数	QLKF	QLRKF(30)	QLRKF(50)	QLRKF(80)
200	0	0	0	0
400	0	0	0	6
600	0	0	2	23
800	0	0	11	31
1000	0	0	22	37
1200	0	13	32	52
1400	4	25	47	78
1600	8	38	61	94
1800	15	40	73	127
2000	22	64	86	140
2200	27	87	100	146
2400	29	100	127	162
2600	34	112	137	206
2800	42	119	155	231
3000	45	120	204	288
3200	49	142	237	334
3400	70	152	291	360
3600	72	183	364	412
3800	82	214	375	455
4000	90	220	375	478
4200	98	239	427	502
4400	98	284	453	531
4600	108	310	476	576
4800	110	343	476	654
5000	126	361	485	721
5200	129	361	485	815
5400	132	387	543	882
5600	132	418	567	948
5800	137	430	591	1054
6000	162	442	694	1150
6200	166	476	739	1263
6400	180	556	820	1361
6600	193	626	899	1383
6800	220	677	957	1489
7000	237	720	1054	1589
7200	267	727	1118	1707
7400	284	742	1197	1787
7600	288	812	1322	1904
7800	319	826	1414	2036
8000	329	905	1551	2101
8200	339	968	1625	2220
8400	408	1026	1718	2357
8600	461	1111	1847	2458
8800	544	1163	1988	2615
9000	603	1269	2111	2740
9200	681	1329	2207	2889
9400	746	1415	2339	2995
9600	788	1499	2435	3168
9800	875	1552	2557	3284
10000	951	1688	2718	3394

表 3　協調距離範囲の累積ステップ数 (ランダム同士と QL 同士)

学習回数	Random	QL
20000	3559	9
40000	7255	73
60000	10900	337
80000	14470	6328
100000	18085	18657
120000	21772	29335
140000	25399	38170
160000	28969	45349
180000	32564	52828
200000	36155	60687
220000	39777	69237
240000	43300	75468
260000	46933	84629
280000	50699	91069
300000	54344	100317
320000	57773	107496
340000	61418	116191
360000	64911	122927
380000	68625	130768
400000	72264	139227
420000	75871	143633
440000	79599	150365
460000	83250	157829
480000	86808	166223
500000	90256	173657
520000	93912	181468
540000	97451	188035
560000	101137	195846
580000	104766	204546
600000	108396	212526
620000	112024	218239
640000	115656	224932
660000	119406	232248
680000	123045	238542
700000	126557	248440
720000	130165	257331
740000	133540	264048
760000	137330	270789
780000	140878	280250
800000	144518	288651
820000	148101	295569
840000	151902	301855
860000	155381	306143
880000	159145	316303
900000	162721	323515
920000	166338	329157
940000	169943	339470
960000	173676	347228
980000	177107	354610
1000000	180714	361953