Q-Learning as Failure

遡及的カルマンフィルタを利用した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．本研究では，学習の過程で失敗につながった行動と逆の行動を選択し状態を戻しながら学習（逆行動学習）することで，知識を改善できないか考える．本稿では，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のモデル化を行う．強化学習のタスクの 1 つである追跡問題で実験を行い，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 [9]．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．例えば，図 1 のように，エージェントが初期状態の真ん中の位置から，左に 2 回行動し，-100 の報酬を得たとする．(図 1 の丸は状態，数値は報酬を表している．) ここで，失敗に繋がった，左の行動と，逆の行動（右）を選択し状態を戻しながら学習（逆行動学習）することで，知識を改善できないだろうか？．以後，逆向きの行動を選択し，エージェントが受け取る報酬を絶対値の逆にし，学習する手法を逆行動学習と定義する．図 1 の例では，状態数，行動数が少ないため，全ての状態履歴や行動履歴を保存し利用することもできる．しかし，複雑なタスクでは全ての情報を記録することは実用的ではない．よって，実用的な逆行動学習手法のモデル化を行うことを本研究の目的とする．

本研究では，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のモデル化を行う．強化学習のタスクの 1 つである追跡問題で実験を行い，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

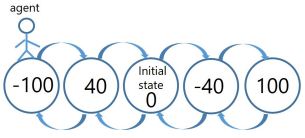

図 1　状態遷移と報酬の例

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 [7] 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 [7]．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

本研究では，学習に失敗した場合，それまでに至る行動と逆の行動が正しいという仮定を置く．この仮定を置き，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と考える．

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 [10] 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

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

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

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

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 [12] [11] 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．学習の結果が良い場合，学習を継続し，悪かった場合，逆行動学習を行う．状態を戻したい場合，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．

カルマンフィルタで事後誤差共分散行列 P_k は，事前誤差共分散行列 P_{k-1}^- とカルマンゲイン行列 G_k から求める．そのため現在の事後誤差共分散行列 P_k から事前誤差共分散行列 P_{k-1}^- を解析的に解くことはできない．事前誤差共分散行列 P_{k-1}^- は学習時に保持して起き，逆行動学習を行う場合に利用できるようにする．事前誤差共分散行列 P_{k-1}^- を保持する数はハイパーパラメータとし，逆行動学習を行うことができる回数に一致する．遡及的カルマンフィルタを以下に定義する．

遡及フィルタリングステップ

$$G_k = P_k^- C^T (C P_k^- C^T + R)^{-1} \quad (11)$$

$$\hat{X}_k^- = (I - G_k C)^{-1} (\hat{X}_k - G_k Y_k) \quad (12)$$

遡及予測ステップ

$$P_{k-1} = A^{-1}(P_k^- - BQB^T)(A^T)^{-1} \quad (13)$$

$$\hat{X}_{k-1} = A^{-1}\hat{X}_k^- \quad (14)$$

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，状態を戻す必要がある．状態を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 [8] 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．遡及的カルマンフィルタを利用し，逆行動学習を可能にした部分が他の研究と異なる．

5 実　　験

5.1 追跡問題

本研究では，強化学習の標準的なタスクである追跡問題を扱う．通常の追跡問題は，m × m の格子状の 2 次元空間のフィールドで実行するが，本研究では，m × m の連続 2 次元空間のフィールドで実行する．追跡するエージェントをハンタ，逃げるエージェントを獲物と定義する．ハンタと獲物は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，ハンタと獲物の距離が，一定の距離以上離れるように，ランダムに配置し，追跡を開始する．捕獲条件として，ハンタと獲物の距離が一定の距離以内になったとき，捕獲したと定義する．

通常の追跡問題では，ハンタのみ学習し，獲物は学習せずにランダムに行動する．本研究では，獲物は学習せず，ランダムに動く場合 (通常の場合) と，獲物も学習した場合で追跡問題を行い，提案手法の効果を調べることとする．

ハンタと獲物が学習する場合，ハンタは捕獲するように，獲物は逃げられるように学習させるために，報酬の設定を行う．具体的には，捕獲した場合，ハンタに正の報酬，獲物に負の報酬を与え，捕獲できなかった場合はハンタに負の報酬，獲物に正の報酬を与える．

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

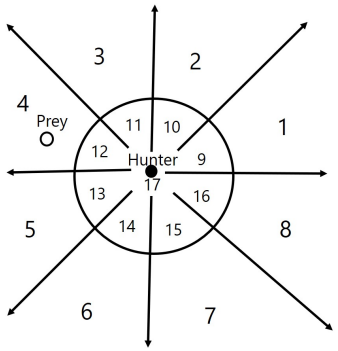

図 2　Hunter's state

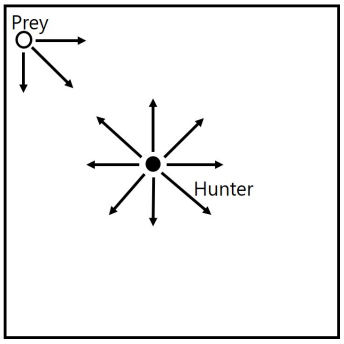

図 3　Examples of hunter and prey behavior

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 2 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 3 のようにハンタがフィールドの中央付近，獲物がフィールドの左上にいる場合，ハンタは全方向と停止の行動が可能である．しかし，獲物はこれ以上，上や左に進むとフィールド外に出てしまう．このような場合，前述のとおり，獲物は右，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，獲物とハンタは交互に，相手のエリアの知覚，行動，学習を繰り返す．具体的に，ハンタと獲物が学習する場合の例を示す.（獲物が学習しない場合は，獲物は“ハンタがいるエリア”の知覚や学習は行わず，ランダムに行動を選択する）.

通常の学習時，以下の a から g を繰り返す.

a. 獲物とハンタを配置
b. 獲物が“ハンタがいるエリア”を知覚し，行動
c. ハンタが“獲物がいるエリア”を知覚し，行動
d. 獲物とハンタが，捕獲状況に応じた報酬を受容
e. 獲物が“ハンタがいるエリア”を知覚し Q 学習
f. ハンタが“獲物がいるエリア”を知覚し Q 学習
g. 捕獲していれば a に戻り，捕獲していなければ b に戻る

提案手法では，ハンタが一定回数連続で捕獲できなかった場合，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. ハンタが“獲物がいるエリア”を知覚し，行動
b. 獲物が“ハンタがいるエリア”を知覚し，行動
c. 獲物とハンタが，報酬を受容
d. ハンタが“獲物がいるエリア”を知覚し Q 学習
e. 獲物が“ハンタがいるエリア”を知覚し Q 学習

5.2 実験準備

ハンタ 1 体，獲物 1 体で追跡問題を行う．提案手法（QLRKF）の学習の効率性を示すために，2 パターンの実験を行う．第一に，ハンタのみが学習を行い，獲物が学習を行わずランダムに行動する一般的な場合である．第二に，ハンタと獲物が両者とも学習を行う場合である．それぞれの場合でハンタが ε グリーディで学習した場合（比較手法 1），QLKF で学習した場合（比較手法 2）と QLRKF で学習した場合（提案手法）で学習を行い，評価する．評価には捕獲までのステップ数を用いる．捕獲までのステップ数は，知識の質の良さを表し，ステップ数が少ないほど良い学習ができていると判断する．

ハンタが QLKF（比較手法 2），QLRKF（提案手法）で学習する場合，ハンタは獲物の位置を予測するカルマンフィルタを使用し，確率 ε で予測した位置に一番近づく行動を選択する．

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，捕獲条件の距離は，0.1 以内とする．捕獲時のハンタの報酬を 100，捕獲できなかった場合の報酬を-1 とする．提案手法の逆行動学習時の報酬は，1 とする.（捕獲できなかった場合の逆の数値とする.）捕獲できなかったことを前提として，逆行動学習を行うので，捕獲時の逆の報酬は考えないこととする．獲物が学習する場合，捕獲された場合の報酬を-80，捕獲されなかった場合の報酬を 1 とする．ハンタと同様，提案手法の逆行動時の報酬は，-1 とする．

ハンタの学習率は 0.1，割引率は 0.9 に設定する．獲物が学習する場合，獲物の学習率も 0.1，割引率も 0.9 に設定する．それぞれの行動選択時のハンタの ε，獲物が学習する場合の獲物の ε は共に ε=0.1 に設定する．提案手法の P_{k-1}^{-} を保持する数は 50 に設定する．

QLKF と QLRKF のカルマンフィルタの初期値は，誤差の共分散行列を $10^4 I$，プロセスノイズの共分散行列を $0.05I$，観測ノイズの共分散行列を $0.9 \times 0.999^{\text{学習回数}} I$ とする．観測ノイズの共分散行列を $0.9 \times 0.999^{\text{学習回数}} I$ とすることで，学習回数が増えるほど誤差が小さくなるように設計する．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 15，観測方程式を式 16 のように設定する．

$$\begin{pmatrix} x_{t+1} \\ y_{t+1} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_t \\ y_t \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} hv_{xt} \\ hv_{yt} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} V_t \tag{15}$$

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_t \\ y_t \end{pmatrix} + W_t \tag{16}$$

5.3 捕獲までのステップ数の評価方法

獲物とハンタがそれぞれ行動をしたとき，1 ステップと定義する．ハンタが 100 回学習（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，ハンタと獲物を配置しなおし，ハンタが獲物を捕獲するまで追跡させる．（捕獲ステップ数の評価実験）1 インターバルごとに捕獲ステップ数の評価実験を 10 回行い，捕獲までのステップ数の平均をそのインターバル時の捕獲までのステップ数として記録する．1000 ステップに達しても捕獲できていない場合，追跡を停止し，捕獲までのステップ数を 1000 とし，平均を求めるのに用いる．100 インターバル（1 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，10 エピソード実行する．各インターバルごとに 10 エピソード分の平均を求めた結果を，そのインターバル時の捕獲までのステップ数として評価に用いる．捕獲ステップ数の評価実験時の行動選択手法は，学習時と同じである．つまり，学習時に ε グリーディで学習した場合，評価時も ε グリーディである．しかし，提案手法の評価実験時は，逆行動学習は行わないため，QLKF で行動選択を行う．獲物は，学習してない場合はランダムな行動選択，ε グリーディで学習している場合は ε グリーディで行動選択を行うように設定し，捕獲ステップ数の評価実験を行う．（以後，行動選択として ε グリーディを用いたものを QL と表記する．）獲物は 2 パターンに分けられるので，全部で $3 \times 2 = 6$ パターンの実験を行う．

一般的に学習回数が多いほど，Q 値の更新が行われるので捕獲までのステップ数が少なくなると考えられる．逆に，学習回数が少ないほど，捕獲までのステップ数は多くなると考えられる．そこで，前述した 6 パターンの各インターバルの結果ごとに学習回数と捕獲までのステップ数の調和平均を求める．学習回数と捕獲までのステップ数は少ないほど良いので，調和平均も値が小さいほど良い．最後に，各パターンごとに調和平均の算術平均と標準偏差を求め，各パターンの総合指標として用いることにする．

5.4 実 験 結 果

学習回数と捕獲までのステップ数の結果を比較する．ハンタのみが学習した場合の結果を表 2，ハンタと獲物が学習した場合の結果を表 3 に示す．（評価実験は学習 100 回ごとに行ったが，表にまとめるため，学習 500 回ごとにまとめなおしている．）捕獲までのステップ数の結果をハンタのみが学習した場合でグラフ化したものを図 4，ハンタと獲物が学習した場合でグラフ化したものを図 5 に示す． 全体的に学習時に QLRKF

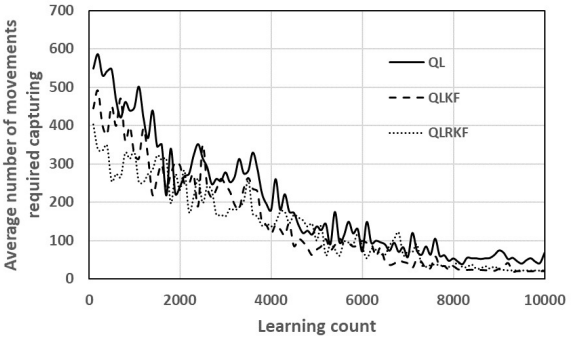

図 4 Relationship between learning count and capture step number (When only hunter is learning)

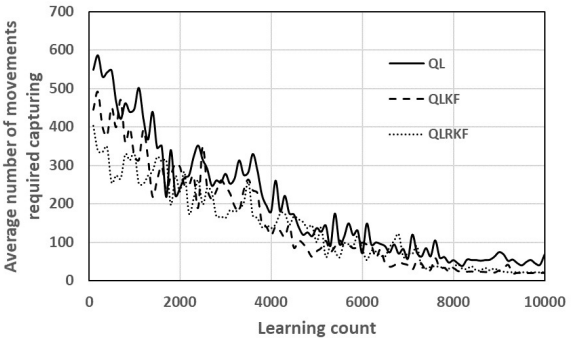

図 5 Relationship between learning count and capture step number (When both hunters and prey are learning)

を用いたものの捕獲までのステップ数が少なくなっているので期待通りの結果が得られている．具体的に，学習時に QL で追跡したものを基準にして，捕獲までのステップ数の総和が何倍になったかをまとめたものを表 1 に示す．獲物が学習しない場合，学習時に QLRKF を利用することで QL で学習する場合の 6 割 5 分，QLKF で学習する場合の 8 割 5 分に改善していることがわかる．また獲物が学習する場合，QL で学習する場合の 4 割以下，QLKF で学習する場合の 8 割 4 分に改善していることがわかる．

学習回数と捕獲までのステップ数との調和平均を 500 回ごとに表記し，調和平均の算術平均と標準偏差をまとめなおしたものを，表 4，表 5 に示す．ハンタのみが学習した場合の結果が表 4，ハンタと獲物が学習した場合の結果が表 5 である．表 4，表 5 の調和平均の算術平均が減っていることから，QLRKF で学習することで，少ない学習回数で捕獲するまでのステップ数が減ることがわかる．具体的に，調和平均の算術平均は獲物が学習しない場合，QL で学習する場合の 7 割以下，QLKF で学習する場合の 9 割以下に改善している．獲物が学習する場合，QL で学習する場合の約 4 割，QLKF で学習する場合の約 8 割

表 1 捕獲までのステップ数の比較

学習時	QL	QLKF	QLRKF
ハンタのみ学習	19467	14916(0.77)	12695(0.65)
ハンタと獲物が学習	19868	9308(0.47)	7806(0.39)

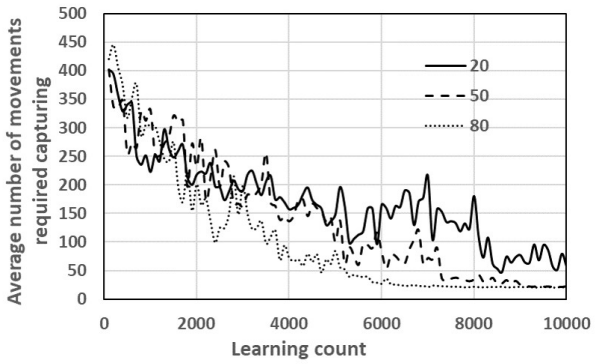

図 6　逆行動学習の回数の変化に伴う捕獲までのステップ数の変化 (When only hunter is learning)

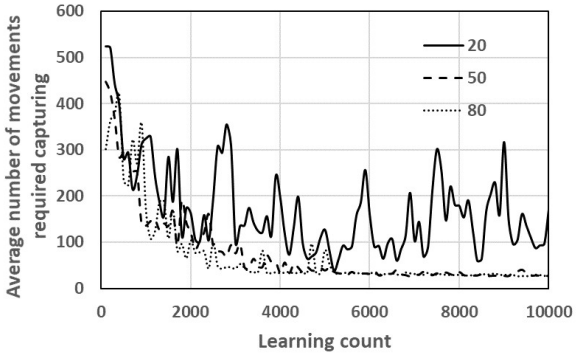

図 7　逆行動学習の回数の変化に伴う捕獲までのステップ数の変化 (When both hunters and prey are learning)

5 分に改善している．調和平均の標準偏差は獲物が学習しない場合，QL で学習する場合の 7 割以下，QLKF で学習する場合の約 7 割に改善している．獲物が学習する場合，QL で学習する場合の約 5 割，QLKF で学習する場合の 7 割 5 分に改善している．このことから，学習が安定して進むことがわかる．

5.5 考　　察

QLRKF で逆行動学習を行うことができる回数を 20，50，80 と変化させ，追跡実験を行った．ハンタのみが学習した場合の結果が表 6，ハンタと獲物が学習した場合の結果が表 7 である．また，それぞれをグラフ化したものを図 6，ハンタと獲物が学習した場合でグラフ化したものを図 7 に示す．表 6，表 7 より，逆行動学習の回数を多くすることで捕獲までのステップ数が減ることがわかる．具体的に，捕獲ステップ数の総和の比率は 50 回を基準として比較する．獲物が学習しない場合，20 回逆学習を行うことができる場合は 1.2 倍に悪化し，80 回の場合は 0.75 倍に改善している．獲物が学習する場合，20 回の場合は 2.2 倍に悪化し，80 回の場合は 0.93 倍に改善している．この結果から，逆行動学習を行うことができる回数が結果に大きく影響することがわかる．

実験結果と同様に，学習回数と捕獲までのステップ数の調和平均，調和平均の算術平均と標準偏差をまとめたものを，表 8，表 9 に示す．ハンタのみが学習した場合の結果が表 8，ハンタと獲物が学習した場合の結果が表 9 である．逆行動学習の回数を多くすることで，少ない学習回数で捕獲するまでのステップ数が減ることがわかる．捕獲ステップ数の総和と同様に，50 回を基準に比較を行う．調和平均の算術平均は獲物が学習しない場合，20 回逆学習を行うことができる場合は 1.26 倍に悪化し，80 回の場合は 0.7 倍に改善している．獲物が学習する場合，20 回の場合は 2.35 倍に悪化し，80 回の場合は 0.92 倍に改善している．この結果からも，逆行動学習を行うことができる回数が結果に大きく影響することがわかる．調和平均の標準偏差は獲物が学習しない場合，20 回の場合は 0.66 倍，80 回の場合は 0.94 倍になっている．獲物が学習する場合，20 回の場合は 1.34 倍，80 回の場合は 1.06 倍になっている．よって逆行動学習ができる回数によって影響があるか断言できない結果となった．

6 結　　論

本研究では，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．学習時に QLRKF を用いることにより，テスト全体の総ステップ数は獲物が学習しない場合，QL で学習する場合の 6 割 5 分，QLKF で学習する場合の 8 割 5 分になることがわかった．獲物が学習する場合，QL で学習する場合の 4 割，QLKF で学習する場合の 8 割 4 分になることがわかった．また，学習回数と捕獲までのステップ数の調和平均の算術平均は，獲物が学習しない場合，QL で学習する場合の 7 割以下，QLKF で学習する場合の 9 割以下になることがわかった．獲物が学習する場合，QL で学習する場合の約 4 割，QLKF で学習する場合の約 8 割 5 分になることがわかった．

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] Watkins, Christopher JCH, and Peter Dayan. ”Q-learning.” Machine learning 8.3-4 (1992): 279-292.
[8] Kei Takahata, Takao Miura. “Reinforcement Learning using Kalman Filters.” IEEE International Conference on Cognitive Informatics and Cognitive Computing(ICCICC) 2019
[9] 高玉 圭樹.” マルチエージェント学習.” コロナ社, 2004
[10] 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．

表 2　捕獲までのステップ数（ハンタのみが学習している場合）

学習回数	QL	QLKF	QLRKF
500	546.4	451.1	342.4
1000	446.3	326.1	312.6
1500	345.2	262.6	290.2
2000	240.6	296.8	192.6
2500	314.8	350.1	109.4
3000	277.9	245.8	178.9
3500	281.9	263.8	139.4
4000	179.8	118.5	154.9
4500	172.4	86.6	147.9
5000	137.1	77.1	118.8
5500	96.0	105.7	86.9
6000	71.8	99.8	92.7
6500	90.2	46.4	93.9
7000	57.4	41.3	43.6
7500	63.8	29.4	51.3
8000	53.7	32.9	38.0
8500	54.1	23.3	28.4
9000	74.5	27.2	39.6
9500	40.3	20.8	26.0
10000	68.1	20.8	40.5

[11] 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,

[12] Matthieu Geist, Olivier Pietquin: “Kalman Temporal Differences.” J. Artif. Intell. Res. 39: 483-532 (2010)

表 3　捕獲までのステップ数（ハンタと獲物が学習している場合）

学習回数	QL	QLKF	QLRKF
500	581.7	419.4	295.6
1000	340.4	271.3	136.2
1500	488.5	173.9	135.3
2000	415.7	64.3	121.0
2500	328.2	149.8	92.0
3000	274.7	124.2	75.9
3500	160.7	71.2	49.4
4000	125.2	68.9	34.0
4500	107.4	35.8	35.2
5000	119.1	68.1	38.7
5500	96.8	31.5	29.2
6000	88.6	30.8	31.7
6500	71.1	30.5	32.1
7000	83.6	30.7	29.9
7500	59.5	30.8	32.5
8000	134.0	29.7	35.2
8500	155.8	29.4	30.1
9000	148.5	30.5	27.2
9500	91.0	31.9	33.2
10000	42.1	30.4	27.3

表 4　学習回数と捕獲までのステップ数の調和平均（ハンタのみが学習している場合）

学習回数	QL	QLKF	QLRKF
100	169.1	163.3	162.9
500	522.2	474.3	406.5
1000	617.2	491.8	476.3
1500	561.3	446.9	486.4
2000	429.5	516.9	351.4
2500	559.2	614.2	209.7
3000	508.7	454.3	337.7
3500	521.7	490.6	268.1
4000	344.1	230.2	298.3
4500	332.1	169.9	286.4
5000	266.8	152.0	232.1
5500	188.7	207.4	171.1
6000	141.8	196.4	182.7
6500	177.9	92.1	185.1
7000	113.8	82.1	86.6
7500	126.5	58.6	101.9
8000	106.8	65.6	75.6
8500	107.6	46.5	56.5
9000	147.7	54.3	78.9
9500	80.2	41.5	51.8
10000	135.3	41.6	80.6
平均	310.2	240.6	213.9
標準偏差	184.5	178.0	127.0

表 5　学習回数と捕獲までのステップ数の調和平均（ハンタと獲物が学習している場合）

学習回数	QL	QLKF	QLRKF
500	537.7	456.2	371.6
1000	507.9	426.8	239.8
1500	737.0	311.7	248.2
2000	688.4	124.6	228.2
2500	580.1	282.6	177.6
3000	503.3	238.4	148.0
3500	307.3	139.5	97.5
4000	242.7	135.6	67.3
4500	209.8	71.1	69.8
5000	232.6	134.3	76.8
5500	190.3	62.6	58.2
6000	174.6	61.3	63.0
6500	140.6	60.8	63.8
7000	165.2	61.2	59.4
7500	118.1	61.4	64.7
8000	263.6	59.2	70.1
8500	305.9	58.6	60.0
9000	292.2	60.8	54.2
9500	180.2	63.5	66.1
10000	83.8	60.6	54.4
平均	311.2	148.5	125.6
標準偏差	185.0	126.4	94.2

表 6　逆行動学習の回数の変化させた場合の捕獲までのステップ数（ハンタのみが学習している場合）

学習回数	20	50	80
500	340.87	255.06	317.13
1000	222.79	331.42	302.41
1500	248.24	321.37	274.67
2000	217.9	231.54	200.91
2500	195.94	197.54	124.38
3000	189.51	162.41	198.83
3500	205.96	255.78	97.06
4000	157.18	135.94	71.78
4500	173.24	166.76	64.27
5000	147.86	99.51	83.88
5500	112.45	59.87	39.15
6000	165.12	82.91	26.41
6500	190.25	62.12	23.03
7000	215.91	71.73	20.62
7500	136.16	35.29	21.14
8000	180.38	34.06	21.15
8500	52.85	26.38	20.57
9000	76.67	24.82	21
9500	94.36	21.37	20.4
10000	56.83	22.29	19.39

表 7　逆行動学習の回数の変化させた場合の捕獲までのステップ数（ハンタと獲物が学習している場合）

学習回数	20	50	80
500	280.8	295.63	234.62
1000	324.7	136.22	165.73
1500	284.99	135.28	107.87
2000	162.72	120.99	104.12
2500	195.17	92.05	100.71
3000	98.18	75.89	42.75
3500	124.92	49.44	34.61
4000	195.77	33.96	32.73
4500	98.25	35.18	35.22
5000	126.89	38.71	83.36
5500	84.8	29.23	34.4
6000	158.09	31.69	33.55
6500	106.94	32.08	32.6
7000	102.99	29.85	31.17
7500	301.55	32.48	29.73
8000	178.92	35.21	29.42
8500	64.19	30.1	31.41
9000	316.93	27.18	30.14
9500	135.52	33.18	26.17
10000	172.26	27.26	25.55

表 8　逆行動学習の回数を変化させた場合の調和平均（ハンタのみが学習している場合）

学習回数	20	50	80
500	405.4	337.8	388.1
1000	364.4	497.8	464.4
1500	426.0	529.3	464.3
2000	393.0	415.0	365.1
2500	363.4	366.1	237.0
3000	356.5	308.1	372.9
3500	389.0	476.7	188.9
4000	302.5	262.9	141.0
4500	333.6	321.6	126.7
5000	287.2	195.1	165.0
5500	220.4	118.5	77.7
6000	321.4	163.6	52.6
6500	369.7	123.1	45.9
7000	418.9	142.0	41.1
7500	267.5	70.2	42.2
8000	352.8	67.8	42.2
8500	105.0	52.6	41.0
9000	152.0	49.5	41.9
9500	186.9	42.6	40.7
10000	113.0	44.5	38.7
平均	297.1	235.0	165.7
標準偏差	100.1	151.4	143.5

表 9　逆行動学習の回数を変化させた場合の調和平均（ハンタと獲物が学習している場合）

学習回数	20	50	80
500	359.6	371.6	319.4
1000	490.2	239.8	284.3
1500	479.0	248.2	201.3
2000	301.0	228.2	197.9
2500	362.1	177.6	193.6
3000	190.1	148.0	84.3
3500	241.2	97.5	68.5
4000	373.3	67.3	64.9
4500	192.3	69.8	69.9
5000	247.5	76.8	164.0
5500	167.0	58.2	68.4
6000	308.1	63.0	66.7
6500	210.4	63.8	64.9
7000	203.0	59.4	62.1
7500	579.8	64.7	59.2
8000	350.0	70.1	58.6
8500	127.4	60.0	62.6
9000	612.3	54.2	60.1
9500	267.2	66.1	52.2
10000	338.7	54.4	51.0
平均	296.0	125.6	116.4
標準偏差	126.4	94.2	100.4