Vol.2011-ICS-163 No.4
2011/3/23

マルチエージェント環境における強化学習パラメータの調整

松下直樹[†]　　原田拓[†]

強化学習においてメタパラメータを適切に調整することは，効果的な学習を行うために重要である．しかし，対象問題ごとにメタパラメータを適切に調整することは，学習システムの設計者にとって負担となる．特にマルチエージェント環境では，他のエージェントの行動などによる環境の変化に対して適切に対応することが要求されるため，メタパラメータの調整は難しい．そこで，本研究では，マルチエージェント環境において強化学習のメタパラメータを調整するための手法を提案する．そして，実験を行うことによって，その有効性を検証する．

Adjustment of Reinforcement Learning Parameters in Multi-agent Environment

Naoki Matsushita[†] and Taku Harada[†]

In reinforcement learning, adjusting the meta-parameters is important for effective learning. The adjustment forces the designer of learning system to coordinate the parameters appropriately. In Multi-agent environment, the adjustment for dynamic change of environment is difficult. In this paper, we propose a method to adjust meta-parameters in Multi-agent environment and verify its effectiveness.

1. はじめに

強化学習とは，数値化された報酬信号を最大化するために，何をすべきかを学習する機械学習の 1 つである [1]．強化学習アルゴリズムには，エージェントがどのように学習するかを決定するメタパラメータがある．メタパラメータには学習率や割引率などがあり，これらの調整が適切でないと学習が破綻してしまうこともあり，強化学習においてメタパラメータの調整は非常に重要である．しかし，メタパラメータの調整は設計者が試行錯誤により設定する必要があるため，適切に設定することは非常に困難である．そのため，メタパラメータの調整を自動的に行う研究が注目されている．

シングルエージェント環境におけるメタパラメータの調整に関する研究が行われている．例えば，文献 2)では，TD 誤差を用いたメタパラメータ学習法を提案している．この手法は TD 誤差の絶対値に依存する変数に基づき，メタパラメータである学習率，割引率，温度定数をステップの実行ごとに更新していくもので，TD 誤差に応じて適切に各メタパラメータが調整されることを示している．文献 3)では，強化学習の並列モデルを効率的に組み合わせることによって，環境が変化した場合においても適応的に学習率を自動調節する並列型強化学習手法を提案している．

マルチエージェント環境における強化学習では，他エージェントの行動や環境の変化への追従と，気まぐれな行動などを原因とする雑音成分への耐性の両立が重要になる [4]．一方で，通常の強化学習では環境の変化に対する追従は行わず，雑音成分に対する耐性を強化するという方策をとっていることが多い [5]．しかし，実際の応用場面においては，環境の変化や他エージェントの行動を無視することはできない [4]．文献 4)では，Newton 法を用いて 2 乗誤差の指数時間平均を最小化することで，最適なステップサイズパラメータを求める手法を提案し，提案手法が迅速に最適な学習率に到達できると同時に，マルチエージェント環境においても環境の変化に追従しつつ，相手エージェントの挙動による外乱や雑音にロバストに期待報酬を学習できることを示している．しかし，文献 4)では，調整するメタパラメータはステップサイズパラメータのみであり，他のメタパラメータについては設計者自身が設定しなければならない．また，2 乗誤差の指数時間平均を求めるための新たなパラメータが導入されている．文献 6)では，囚人のジレンマゲームにおける学習率に関する定理をもとに，2 人 2 行動対称ゲームのための学習率調整 Q 学習を提案している．しかし，調整するパラメータは学習率のみである．

これに対して本研究では，マルチエージェント環境のもとで，強化学習のメタパラメータである学習率，割引率，探査率の調整を可能とするアルゴリズムを提案する．そして，学習途中で環境が変化する状況に対して，その有効性を実証する．本研究で

[†] 東京理科大学
Tokyo University of Science

Vol.2011-ICS-163 No.4
2011/3/23

は強化学習の手法の 1 つである Q-Learning をその対象とする．

2. メタパラメータ

Q-Learning では，実際に得られた報酬と予測された報酬の差分である TD 誤差を用いて最適な政策を求める．状態と行動を 1 組として表す状態行動価値関数Q値（$Q(s,a)$）を持ち，このQ値を基に試行錯誤を通じて最適な政策を得るように学習する．αを学習率，γを割引率，sを状態，aを行動，rを報酬，tを時刻とすると，状態行動価値関数は式(1)で表される．

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)] \quad (1)$$

学習率α（$0 < \alpha \leq 1$）は，状態行動価値関数の推定値に対して時刻tに生じた TD 誤差をどの程度反映させるかを決定するメタパラメータである．すなわち，学習速度を決定するメタパラメータであるといえる．一般的に，学習率αが小さければ学習は安定するが学習速度が遅くなってしまい，逆に大きければ学習速度は速くなるが学習が不安定になる傾向がある．つまり，学習率αは学習の速度と安定性のトレードオフをとる働きをもつメタパラメータである．

割引率γ（$0 \leq \gamma \leq 1$）について説明する．強化学習の目標は，得られる累積報酬を最大化することであり，強化学習における学習とは，累積報酬の最大化を実現するための評価関数として割引報酬和の期待値を推定することである．割引率γは即時報酬の重みを 1 としたときの将来得られる報酬への重みづけの割合を決めており，遠い将来の報酬ほど割り引いて考えることを表している．すなわち，割引率γは報酬予測の時間スケールを定義するメタパラメータであるといえる．

強化学習では，全ての行動を十分な回数選択すれば，行動選択方法にかかわらず最適な行動を選択するように学習できる．しかし，常に最適な行動を選択し続けたり，ランダムに行動選択し続けると局所的な行動に陥るなどして学習が遅くなる場合がある．速く学習させるためには学習途中でなるべく多くの報酬を与えるような行動選択が必要とされる．そこで，本研究では行動選択方法としてεグリーディ行動選択手法を利用する．εグリーディ行動選択手法は探査率ε（$0 \leq \varepsilon \leq 1$）の確率でランダムに行動を選択し，確率$1-\varepsilon$である状態の中で最も大きな$Q$値を持つ行動を選択する．$\varepsilon$グリーディ行動選択手法において，エージェントは基本的にはQ値が最大となる行動を選択するが，εの確率でランダムに選択し，探索を行う．つまり，εの値が大きければランダム選択に近づき，小さければ最も大きなQ値を持つ行動を選択するグリーディ選択に近づく．探査率εは，局所的な最適戦略に陥らないために，確率的に行動探索を行う割合を決める強化学習メタパラメータである．

3. 提案手法

本研究では，強化学習におけるメタパラメータの調整に，文献 7)で提案されているQ値の評価値を利用する．Q値の評価値とは，群強化学習において最良のQ値を持つエージェントを調べるために，Q値を評価する際に用いた指標であり，Lをそのエピソードでの全行動回数，r_kをk回目の行動で得られた報酬，dを報酬を割り引く割合として，式(2)で表される．

$$E = \sum_{k=1}^{L} d^{L-k} r_k \quad (2)$$

Q値を評価する最も適した指標は収益であるが，収益は時間ステップtの後に受け取った報酬の合計であるため，収益を正確に算出するには多くのシミュレーションを実行しなければならず，現実的でない．そこで，文献 7)では収益を近似する方法として，そのエピソードで得られた報酬の割引和を評価値とする方法を用いている．

Q値の評価値は学習の進捗状況によって変動の大きさが異なる．学習初期段階，もしくは環境に変化があった場合は，エージェントの行動が不安定なため，Q値の評価値の変動は大きくなる．しかし，学習が十分に進行すると，エージェントは極端に悪い行動はとらなくなり，特定の行動に収束することによって，Q値の評価値はほぼ変動がなくなる．よって，Q値の評価値の変動をみることによって，学習の進捗状況を見分けることが可能になる．以上の要因から，この指標を用いてメタパラメータの調整を行うことが可能になると考えられる．

本研究で提案する手法では，Q値の評価値のエピソード間の変化に合わせてメタパラメータを調整していく．具体的には，エピソード間のQ値の評価値の差の絶対値に依存して変化する変数$\delta(t)$をとり，それに基づいて各メタパラメータを更新する．$\delta(t)$は文献 2)における TD 誤差の絶対値に依存して変化する変数を参考にし，τを時定数として，式(3)のように定義する．

$$\delta(t) = \left(1 - \frac{1}{\tau}\right)\delta(t-1) + \frac{1}{\tau}|E(t) - E(t-1)| \quad (3)$$

ただし，$\delta(0) = 0$

学習率α，探査率εは，学習初期の段階や環境が変化して再学習の必要が出た場合には探索を行うために高くすることが望ましく，学習が十分進行した場合には過学習を防ぐため，および，学習を安定させるために低くすることが望ましい．逆に割引率γは，学習初期の段階や環境が変化して再学習の必要が出た場合には低くすることが望ましく，学習が十分進行した場合には高くすることが望ましい．これに対して$\delta(t)$は，学習初期の段階や環境が変化して再学習の必要が出た場合には高くなり，学習が進むと

ほぼ 0 に収束する．そこで，時刻tにおける$\delta(t)$を時刻tまでの$\delta(t)$の最大値$\max\ \delta(t)$で割った$\delta(t)/\max\ \delta(t)$を基に標準シグモイド関数を利用して各メタパラメータを更新する．$\delta(t)/\max\ \delta(t)$は学習初期の段階や環境が変化した場合には 1 に近い値になり，学習が十分進行した場合には 0 に近づく．シグモイド関数は式(4)で表される関数で，$a=1$とした標準シグモイド関数を用いる．各メタパラメータの調整に用いる範囲は異なり，学習の速度，精度，安定性を考慮した予備実験の結果，学習率αは$-5 \le x \le 5$，割引率γは$0 \le x \le 5$，探査率εは$-10 \le x \le -5$とした．

$$f(x)=\frac{1}{1+e^{-ax}} \tag{4}$$

$\delta(t)/\max\ \delta(t)$に基づいて標準シグモイド関数を利用し，各メタパラメータを適切に調整するように定義した更新式を式(5)～式(7)に示す．

$$\alpha(t)=\frac{1}{1+e^{-(10\frac{\delta(t)}{\max\ \delta(t)}-5)}} \tag{5}$$

$$\gamma(t)=\frac{1}{1+e^{-(-5\frac{\delta(t)}{\max\ \delta(t)}+5)}} \tag{6}$$

$$\varepsilon(t)=\frac{1}{1+e^{-(5\frac{\delta(t)}{\max\ \delta(t)}-10)}} \tag{7}$$

この更新式により，$\delta(t)$の減少に合わせてα，εは低く，逆にγは高くなり，一方増加した場合はその逆になる．このように，$\delta(t)$の変化に合わせて各メタパラメータの値はエピソードが終了するごとに適切に更新される．

なお，提案手法において設計者が設定すべきパラメータは，式(2)におけるdと式(3)におけるτである．

4. 実験

4.1 固定パラメータとの比較

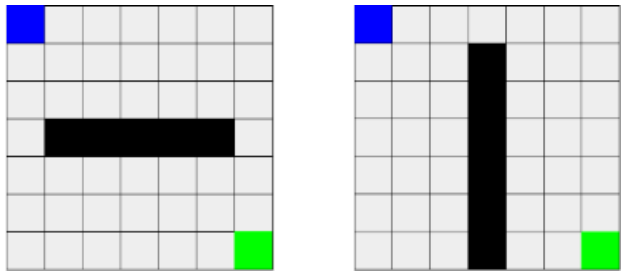
マルチエージェント問題の 1 つである追跡問題を用いて実験を行う．追跡問題とは，複数のハンターが獲物を追跡するという問題である．ここでは，格子状のマップにハンターと獲物を配置し，ハンターが獲物を取り囲めば捕獲（目標達成）とした．本研究では，7×7 マスのマップ上の左上に 2 体のハンター，右下に 1 体の獲物をそれぞれ配置した．各端側は壁とみなし，各ハンターが同マスに存在せず，かつ獲物の周囲 8 マスのどこかに存在した場合に捕獲とする．ハンターと獲物はそれぞれ［上，右上，右，右下，下，左下，左，左上］に 1 マス移動，または停滞の計 9 種類の行動が可能である．獲物は各ハンターとの距離の総和が最大となるよう行動し，獲物と両ハンターは同じタイミングで行動を行う．これを 1 ステップとし，ハンターが獲物を捕獲する，または 1000000 ステップが経過した時点でエピソードを終了とする．捕獲時に正の報酬 1000 を，それ以外は 1 ステップごとに負の報酬－1 を与える．またマップ上には障害物が存在し，2001 エピソードでその位置および大きさを変化させる．これらの様子を図 1 に示す．

また，式(2)において$d=0.999$，式(3)において$\tau=80$とした．これらの値は，精度と速度が共に優れた結果になるように設定したものである．精度については 1501～2000 エピソード，4501～5000 エピソードまでのステップ数の平均から判断し，速度については実行結果のグラフでの収束の速さから判断した．

メタパラメータの値を固定した場合については，提案手法の学習の速度と精度を比較するために，その値として 2 パターンの組み合わせを用意した．速度を比較するためのメタパラメータの値は学習率$\alpha=0.4$，割引率$\gamma=0.9$，探査率$\varepsilon=0.01$とし，精度を比較するためのメタパラメータの値は学習率$\alpha=0.1$，割引率$\gamma=0.45$，探査率$\varepsilon=0.01$とした．これらの値は，図 1(a)のマップで 5000 エピソードを 20 回実行し，最後の 500 エピソードの平均から精度を，実行結果のグラフから収束の速さを判断した結果として設定したものである．これらの値を用いて評価実験を行う．

4.2 提案手法における設定パラメータの影響

本研究の提案手法において設計者が設定すべきパラメータは，式(2)の報酬を割り引く割合dと式(3)の時定数τである．そこで，これらの値の違いによる影響を調べる．パラメータdについては，0.7～1.0 までの範囲で 5 つの値を設定し，その影響を調べる．また，パラメータτについては，1～1000 までの範囲で 7 つの値を設定し，その影響を調べる．これらの調査において，dやτ以外の数値，マップ，環境を変化させるエピソード数などは前述した比較実験と同様である．



(a)変化前　　(b)変化後

図 1　追跡問題における障害物の位置の変化

Vol.2011-ICS-163 No.4
2011/3/23

5. 実験結果および考察

5.1 固定パラメータとの比較

提案手法とメタパラメータを固定した場合の各エピソードにおける捕獲までに要したステップ数の変化を図 2 に示す．また，提案手法におけるQ値の評価値$E(t)$，変数$\delta(t)$の時間推移をそれぞれ図 3，図 4 に，提案手法によって学習した各メタパラメータの時間推移をそれぞれ図 5，図 6，図 7 に示す．なお，この実験結果は 20 回試行した結果の平均である．

まず提案手法とメタパラメータを固定した場合を比較する．図 2 より，学習の精度については，精度が良くなるように定めたメタパラメータと比較して，提案手法の方が若干劣るものの，ある程度の精度を得ることができていることがわかる．また，学習の速度については提案手法の方が速く，特に環境に変化があったときに，メタパラメータを固定した場合よりも学習が高速になることがわかる．さらに，収束後の挙動については，精度，速度が良くなるように定めた場合のどちらと比較しても，提案手法の方が安定している．

次に，提案手法における各メタパラメータの値について考察する．図 3 より，学習の初期段階または環境に変化があった時に，Q値の評価値の変動が大きくなり，学習が収束するとほぼ変動しなくなることがわかる．また，図 4 より，$\delta(t)$はQ値の評価値の変動が大きいときには増加し，逆に小さいときには減少していることがわかる．また，図 5，図 6，図 7 より，学習初期段階には，$\delta(t)$の変化に合わせて，各メタパラメータは大きく変化しているが，学習の収束が進むと，学習率α，探査率εは学習が収束したことを示す小さい値に，割引率γはほぼ 1 となり，ほとんど変化していないことがわかる．さらに，障害物の位置が変化した時（2001 エピソード時）には，ほぼ 0 に収束していた$\delta(t)$が再び増加し，それによって再探索のために学習率α，探査率εが高い値へ，割引率γが低い値へと調整されていることがわかる．以上より，学習の進捗状況に合わせて，各メタパラメータが適切に調整されていることがわかる．

5.2 提案手法における設定パラメータの影響

提案手法におけるパラメータdによる影響を図 8 に，パラメータτによる影響を図 9 に示す．なお，この実験結果は 20 回試行した結果の平均である．

まず，パラメータdの値の違いによる影響について考察する．図 8 より，障害物の位置が変化する前までは，dの値の違いによる影響は少ない．しかし，障害物の位置が変化すると，その影響は大きくなる．dが 0.999 以外の値では，環境の変化に追従できてはいるものの，学習の速度および精度はともに$d = 0.999$の場合と比べると劣っている．これはdの値が 0.999 以外では，Q値の評価値を正確に表現できていないためであると考えられる．また，文献 7)での数値実験においても$d = 0.999$と設定している．以上のことから，Q値の評価値を正確に表現し，その値を用いて，メタパラメータを動的な環境においても適切に調整するためには$d = 0.999$に設定することが良いと考えられる．

次に，パラメータτの値の違いによる影響について考察する．図 9 より，障害物の位置が変化すると，τが 1，10，1000 のときに学習が遅くなっていることがわかる．これは式(3)の構造から説明することができる．式(3)より，τの値が大きいとQ値の評価値の変動を$\delta(t)$に反映させる割合が小さくなる．その結果，環境の変化に対する$\delta(t)$の反応が遅れ，各メタパラメータの調整も遅くなり，環境の変化に対する追従が遅れてしまう．逆にτの値が小さいとQ値の評価値の変動を$\delta(t)$に反映させる割合が大きくなる．そのため，エージェントの気まぐれな行動などにより，1 エピソード間でもQ値の評価値の変動が大きくなると，それが$\delta(t)$の値に大きく影響してしまう．その結果，各メタパラメータが頻繁に変動してしまい，収束が遅くなる．よって，パラメータτについては，適用する問題により，ある程度試行錯誤により設定する必要があることがわかる．

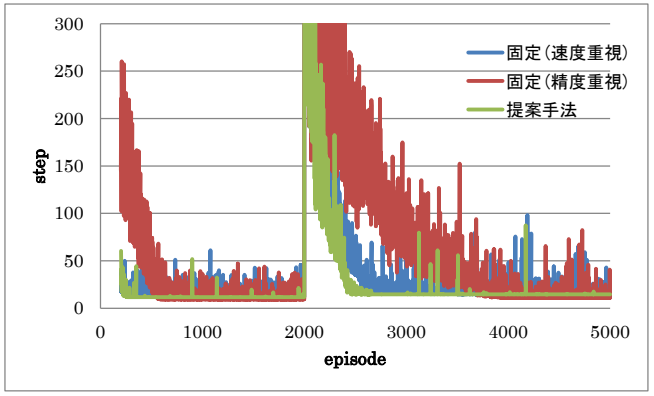


図 2　捕獲までのステップ数

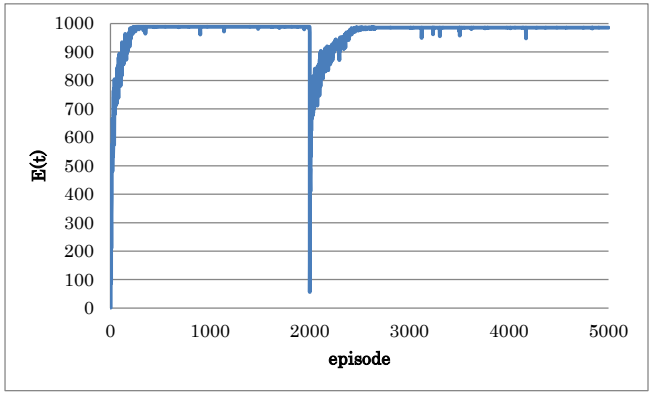


図 3　Q値の評価値$E(t)$の推移

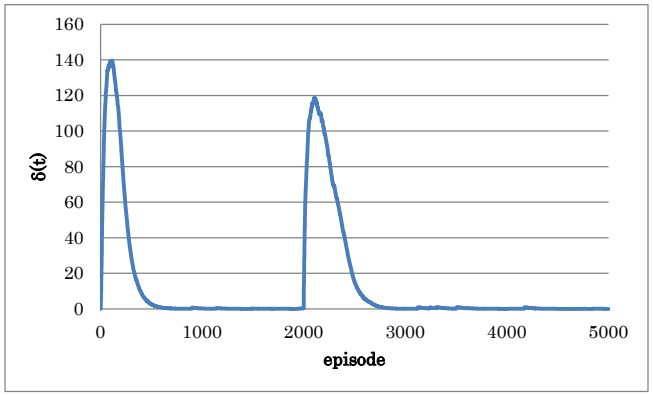


図 4　$\delta(t)$の推移

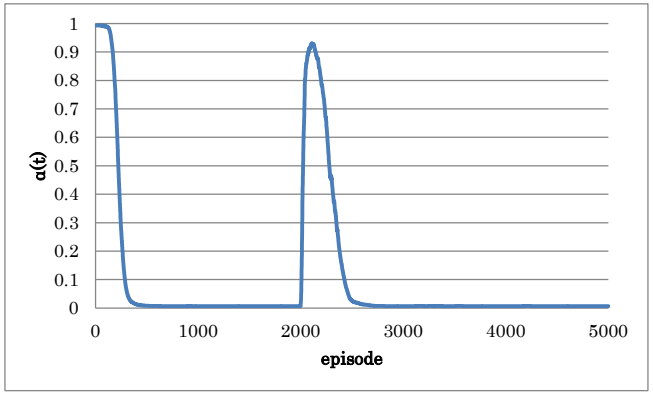


図 5　学習率の推移

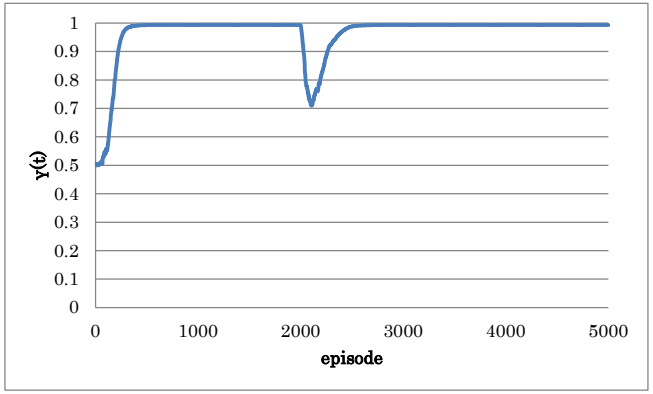


図 6　割引率の推移

Vol.2011-ICS-163 No.4
2011/3/23

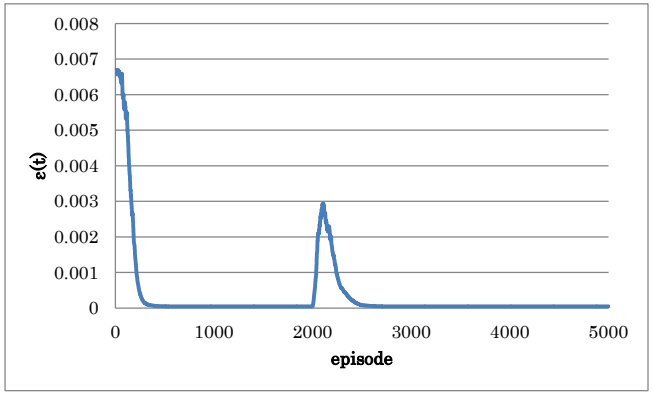


図 7　探査率の推移

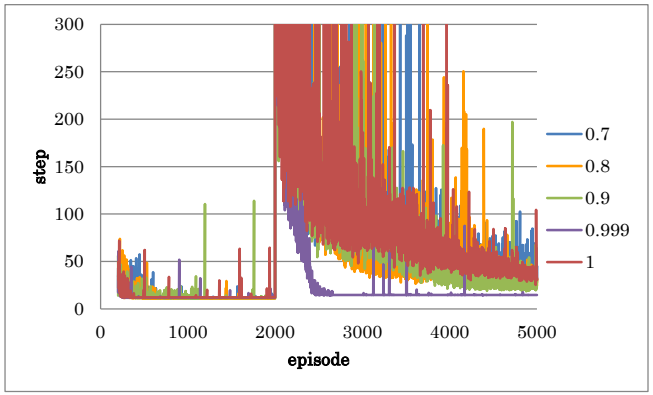


図 8　提案手法におけるdの影響

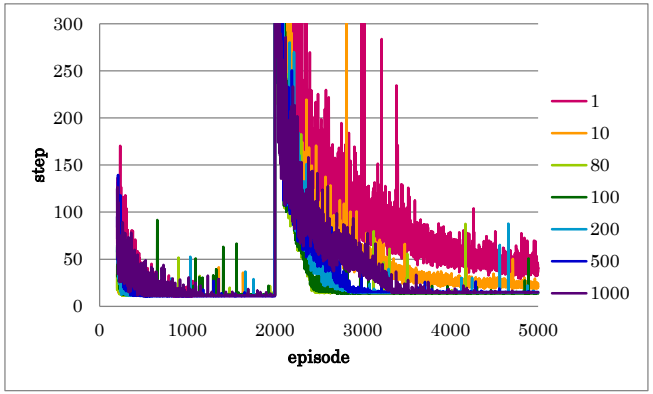


図 9　提案手法におけるτの影響

6.　おわりに

本研究では，マルチエージェント環境における Q-Learning 強化学習のメタパラメータの学習手法を提案した．提案手法ではQ値の評価値を用いてメタパラメータ学習を行った．そして，評価実験を行い，メタパラメータを固定した場合と比較して，ある程度の学習の精度を保ち，学習の速度が速くなり，収束後の挙動も安定しているという結果を得た．これにより，学習の進捗状況や環境の変化に応じて適切に各メタパラメータを調整することができることを示した．

参考文献

1) R.S.Sutton and A.G.Barto（著），三上貞芳，皆川雅章（共訳）：強化学習，森北出版，2000.
2) 溝上裕之，小林邦和，呉本尭，大林正直：TD 誤差に基づく強化学習のメタパラメータ学習法，電気学会論文誌 C，Vol.129，No.9，pp.1730-1736，2009.
3) 阿知波健，渡辺亮平，田中昭雄，大家淳二：強化学習の並列型メタ学習：学習率の調整，電子情報通信学会論文誌，D-I，Vol.J88-D-I，No.12，pp.1773-1784，2005.
4) 野田五十樹：マルチエージェント環境下における強化学習のステップサイズパラメータの適応，人工知能学会第 24 回全国大会論文集，2010.
5) Eyal Even-Dar，Yishay Mansour: Learning Rates for Q-learning, *Journal of Machine Learning Research*，Vol.5，pp.1-25，2003.
6) 森山甲一：2 人 2 行動対称ゲームのための学習率調整 Q 学習，電子情報通信学会論文誌 D，Vol.J92-D，No.11，pp.1891-1826，2009.
7) 飯間等，黒江康明：エージェント間の情報交換に基づく群強化学習法，計測自動制御学会論文集，Vol.42，No.11，pp.1244-1251，2006.