

学習進度を反映した割引率の調整

尾川 順子[†] 並木 明夫^{††,†} 石川 正俊[†]

[†] 東京大学 大学院情報理工学系研究科 〒113-8656 東京都文京区本郷 7-3-1

^{††} 科学技術振興事業団 戦略的基礎推進事業 〒101-0032 東京都千代田区岩本町 2-8-12

E-mail: †{naoko,namik,ishikawa}@k2.t.u-tokyo.ac.jp

あらまし 強化学習における割引率を学習進度によって調整することの有用性を示す。学習進度が浅いときには割引率を下げて即時報酬を重視し、学習が進むにつれて次第に割引率を大きくして、将来の報酬も考慮していくという戦略を提案する。また、学習進度の調整法として、指数的調整、TD 誤差による調整、信頼度による調整を提案する。これを windy gridworld 課題により検証する。

キーワード 強化学習, 割引率, 学習進度, 信頼度, windy gridworld 課題

Adjustment of Discount Rate Using Index for Progress of Learning

Naoko OGAWA[†], Akio NAMIKI^{††,†}, and Masatoshi ISHIKAWA[†]

[†] Graduate School of Information Science and Technology, Univ. of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

^{††} CREST, JST 2-8-12 Iwamoto-cho, Chiyoda-ku, Tokyo 101-0032, Japan
E-mail: †{naoko,namik,ishikawa}@k2.t.u-tokyo.ac.jp

Abstract We show that it can be effective to adjust the discount rate using an index for progress of learning. In the strategy that we propose, the discount rate is small when the learning does not progress enough, and is increased as the learning advances. We also propose three methods for its adjustment; exponential, by TD error, and by reliability, which are veriflicated by numerical experiments for a windy gridworld task.

Key words Reinforcement Learning, Discount Rate, Progress of Learning, Reliability, Windy Gridworld Task

1. はじめに

強化学習は、環境から与えられる累積報酬の最大化を目標とする試行錯誤的な学習方法である [1]。通常の強化学習では、時刻 t において、エージェント（学習主体）は次式で与えられる累積報酬 R_t を最大化するように学習を行う：

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots + \gamma^T r_{t+T+1} \\ = \sum_{k=0}^T \gamma^k r_{t+k+1}, \quad (1)$$

ここで、 r_k はエージェントがある時刻 k に環境から受け取る報酬の大きさ、 T はタスクの終了時刻である ($T = \infty$ の場合も多い)。 γ は割引率と呼ばれる $0 \leq \gamma \leq 1$ のメタパラメータで、遠い将来の報酬ほど割引いて考えることを表している。式 (1) より、 γ の値が大きければ遠い将来の報酬まで考慮することになり、小さければ即時的な報酬を優先することになる。

一般的な強化学習では、強化信号をできるだけ大きくするために、割引率 γ を 1 に近い固定値にすることが多い [2]。しか

し、割引率を 1 に近づけすぎると収束が遅くなるため [3], [4], 割引率のチューニングは重要かつ困難な問題である。

この問題に対し、割引率自体を廃止する試みはいくつか行われているが [4], 割引率の調整を正面から扱い、その明確な指針を示した研究は少ない。銅谷らは、強化学習におけるメタパラメータに脳内の神経修飾物質系の働きを関連づける仮説を提案しているが [5], 割引率については現時点では実験的証拠が十分とは言えず、具体的なパラメータ調整指針についても明らかになっていない。阪口らは「内部モデルの信頼度」という指標によって学習率や逆温度を制御するアルゴリズムを提案しているが、割引率については触れられていない [6]~[8]。吉田らは価値関数の分散を用いて逆温度を制御しているが、やはり割引率は扱っていない [9]。岡田らは報酬性強化信号と嫌悪性強化信号の 2 元評価による強化学習を提案しており、報酬性強化信号に対する割引率を高く、嫌悪性強化信号に対する割引率を低く設定しているが [10], 割引率の調整を対象とした研究ではない。尾形らは危険度に相当する「自己保存評価値」という指標によって割引率などのメタパラメータを調整し、適応能力の向上を示

したが、詳細は未発表であり、今後の報告が待たれるところである [11].

本研究では、学習進度に応じて割引率を調整する手法を提案し、その結果を報告する。以下、2. では割引率の意味を学習進度に関連づけて解釈することにより、関連づけの大きな方向性に関する仮説を提起する。そして 3. では調整に用いる学習進度の指標を 3 種類提案する。さらに 4. で数値実験によりこれらを検証し、5. で考察を加える。

2. 割引率の調整戦略の提起

学習は、内部モデル（いわゆる世界像）を学習者の内部に作り上げていく過程とみなせるが、学習初期段階ではこの内部モデルの構築が不十分である。したがって、将来の報酬の予測は困難であり、予測値の信頼性は著しく低下する。そのような状況下で遠い将来の報酬を目先の報酬と同等に評価することは無意味だと思われる。すなわち、従来法のように割引率を 1 に近い値に固定し続けることは必ずしも適切な戦略とはいえない。むしろ、とりあえず目先の報酬を確保しておくような予測戦略の方が望ましい可能性がある。

よって、

学習進度が浅いときには割引率を下げ即時報酬を重視し、学習が進むにつれて次第に割引率を大きくして、将来の報酬も考慮していく

という戦略が有効でないかと考えられる。

これに対し、学習が進んでしまえば将来の情報は必要ない、むしろ学習初期にこそ、遠い将来に対する道標を示す必要があるという、全く逆の考え方もあるかも知れない。そこで、上記の仮説とは逆に割引率を下げていく手法についても 4. の実験により比較検討する。

3. 割引率の調整方法

本節では、割引率調整のための具体的な方法について述べる。まず考慮すべきは「学習進度をどのように把握するか」ということである。以下では、3 種類の方法を提案する。

3.1 指数的調整

一般的な学習過程においては、大まかに言って学習進度が時間発展とともに単調増加すると仮定できる。そこで、時間方向に単調増加するような学習進度曲線をあらかじめ「決め打ち」で仮定し、それに従って割引率を制御する方法が最も簡単である。

例えば、学習進度曲線として、時間 $t \rightarrow \infty$ で 1 に漸近するような曲線

$$f(t) = 1 - ae^{-bt}$$

を仮定する。このとき、割引率 $\gamma(t)$ を

$$\gamma(t) = \gamma_0 f(t) = \gamma_0 (1 - ae^{-bt}) \quad (2)$$

と調整する方法が考えられる。ここで、 γ_0 は $\gamma(t)$ の終端値で、1 に近い正の実数を指定する。また、 a , b は学習進度を調整す

る正值のパラメータである。

この手法は適用が容易である反面、仮定した曲線が必ずしも現実を反映しているとは限らないことが欠点である。そこで、残る 2 つの手法では、実際の学習進度を推定することにより、より現実に即した調整を試みる。

3.2 TD 誤差による調整

学習進度の指標として、もっとも単純なものは恐らく TD 誤差であろう。すなわち、TD 誤差が大きいときには学習が進んでいないとみなし、小さくなれば学習が進んだとみなすのである。

時刻 t における TD 誤差 δ_t は、例えば Q 学習 [1] では、報酬 r_t 、価値関数 $Q(s, a)$ 、割引率 γ に対して

$$\delta_t = r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a)$$

であり、Actor-Critic 法 [1] では報酬 r_t 、価値関数 $V(s)$ 、割引率 γ に対して

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$$

となる。この TD 誤差を用いた割引率の制御方法としては、例えば

$$\gamma(t) = \gamma_0 \min \left(1, \frac{1}{|\delta_t|} \right) \quad (3)$$

などを考えることができる。

この方法は簡便ではあるが、TD 誤差は各試行ごとでのばらつきが大きいため、学習が不安定になる可能性がある。そこで、3 番目の指標として次節では「信頼度」という尺度の導入を試みる。

3.3 信頼度による調整

学習進度の推定法として、ここでは、個体が内部モデルをどの程度信頼できるかを主観的に評価した尺度である「信頼度」[6]~[8], [12], [13] と呼ばれる指標を導入する。

信頼度の概念は、運動計画の学習における内部モデルに関する考察から生まれた [12]。学習という営みには、学習主体が学習対象の内部モデルを何らかの形で自らの中に構築していくことが不可欠である。仮に内部モデルが学習対象を正しく反映していれば、学習主体は内部モデルによる予測を信頼して行動すればよい。

しかし、未知の環境や変動する環境に適応する過程では、内部モデルは必ずしも完全でない。その予測を盲目的に信じて行動を定めることは必ずしも良いことではない。そこで、内部モデルがどの程度正確に学習対象を反映しているか、言い換えればどの程度内部モデルを信用できるか、を示す尺度が必要となってくる。その尺度として導入されたのが信頼度である。

信頼度は、内部モデルの予測誤差に基づき更新される。信頼度に比べて誤差が小さければ信頼度は向上し、大きければ低下する。信頼度は確率モデルとしての内部モデルの分散の逆数と見なすことができ [12]、また報酬の期待値としての量を担っているとも見なせる [13]。

そこで、以下ではこの信頼度を学習進度の指標として用いる。

すなわち、信頼度が低いことは内部モデルが信頼できず、まだ学習が必要であることを示し、信頼度が高いことは十分に学習した状態であることを示す、と考えるのである。

信頼度の更新にはさまざまな手法が考えられるが、ここでは以下のような手法 [8] を導入する。この手法では学習対象の内部モデルとして価値関数を想定し、信頼度はこれに対して定義した。以下では、各状態 s について信頼度を考え、その指標を $R(s)$ と置く。 $R(s)$ が大きいことは信頼度が高いことを意味する。なお、状態 s と行動 a について信頼度を定義することも可能であるが [7]、ここでは割愛し、4.1 節で数値実験用にあらためて定義する。

以降の実際の実アルゴリズムにおいては、

$$R^2(s) \stackrel{\text{def}}{=} \frac{1}{R(s)}$$

で定義される信頼度の逆数 $R^2(s)$ を使用する。 $R(s)$ は報酬の次元を持つ。信頼度が高いほど $R^2(s)$ が小さくなることに注意が必要である。

$R^2(s)$ の更新則は以下の通り、TD 学習ライクな方法で行う：

$$R_{t+1}^2(s_t) = R_t^2(s_t) + \alpha_R R \delta_t, \quad (4)$$

$$\text{where } R \delta_t \stackrel{\text{def}}{=} \delta_t^2 + \gamma_R R_t^2(s_{t+1}) - R_t^2(s_t). \quad (5)$$

ここで α_R , γ_R , $R \delta$ はそれぞれ信頼度の学習率、信頼度の割引率、信頼度の TD 誤差である（以下では簡単のため $\gamma_R = 0$ とする）。また、 δ_t は TD 学習自体の TD 誤差である。

式 (5) は、期待した誤差 R^2 よりも実際の誤差 δ が小さければ R^2 を小さくして信頼度を上げ、大きければ R^2 を大きくして信頼度を下げることによって、 R^2 を実際の誤差に近づけていく過程であるとみなすことができる。

さらに式 (5) から、信頼度は二乗 TD 誤差の累積値と見なすことができ、TD 誤差の時間積分、あるいはローパスフィルタリングされた TD 誤差という解釈もできる。この操作により TD 誤差の試行ごとのばらつきが吸収され、学習が安定になると予想される。

この信頼度の概念を用いて割引率の制御を行う方法として、例えば以下のような手法を提案する：

$$\gamma_t = \gamma_0 \min \left(1, \frac{1}{R_t} \right). \quad (6)$$

ここで R_t は、例えば各状態に対する $R(s)$ の平均値でもよいし、代表的な、あるいは本質的なある一状態に対する R でもよい。

4. 数値実験

4.1 問題設定

今回は例題として、windy gridworld（風が吹く格子世界）課題 [1] を取り上げる。Gridworld 課題は非常に単純化された迷路課題であり、エージェントがゴールに向かって正方格子状上を移動するというものである。格子のセルは環境の状態 $s = (s_x, s_y)$ に相当し、別のセルへの移動が行動 a に相当する。

今回の実験では、各セルからは 8-近傍のセルのいずれかへの

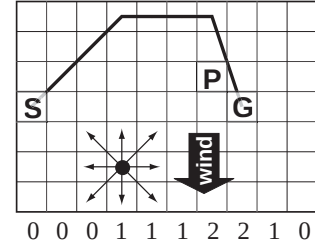


図 1 Windy gridworld. 文献 [1] より改変して引用。図示した道筋は最短経路の一例。

Fig. 1 Windy gridworld. Cited from reference [1] (modified). The shown path is an example of the shortest paths.

移動が可能であるとした。すなわち、エージェントはたて、よこ、ななめ方向へ 1 マスずつ動くことができる。Gridworld の大きさはたて 7 マス、よこ 10 マスとし、左上の角を原点として、図 1 のようにスタート $S(1,4)$ 、ゴール $G(8,4)$ を置いた。1 回のエピソードは、エージェントがゴールに移動した時点で終了し、エージェントはスタートに戻る。

さらに、gridworld には「風」が吹いており、その影響でエージェントの遷移先が上下方向にずれるようになっている。風の強さは各列で異なるものとし、その強さ（風の向きにずれるセルの個数）は図 1 の下に数値で示したとおりである。例えばゴールの右隣のセルにいる場合、左方向に 1 マス進む行動によって、ゴールの真下のセルに移動することになる。

また、各セルに移動するたびに報酬 $r(t)$ をエージェントに与えることとした。報酬の大きさは、ゴール時に 1、範囲外に出た時に -1 、その他のセルでは 0 とした。

具体的な設定は以下の通りである。学習アルゴリズムとしては Actor-Critic 法 [1] を用い、価値関数の更新則は、

$$\Delta V(s_t) = \alpha_V \delta_t,$$

$$\Delta p(s_t, a_t) = \alpha_p \delta_t,$$

$$\text{where } \delta_t = r_{t+1} + \gamma_t V(s_{t+1}) - V(s_t)$$

とした。また、行動選択には softmax 法 [1],

$$\Pr(a) = \frac{\exp(\beta p(s_t, a))}{\sum_{a'} \exp(\beta p(s_t, a'))}$$

を用いた。ここで V , p はそれぞれ状態価値関数および政策、 α_V , α_p , β はそれぞれ V の学習率、 p のステップサイズパラメータ、softmax 法の逆温度パラメータである。

以下は、実験に用いた割引率の調整方法である。

a) 割引率調整なし

割引率を調整しない従来の Actor-Critic 法を用いて実験を行った。

b) 割引率の指数的調整

指数関数を用いた割引率の事前調整について実験を行った。この場合、割引率 γ の調整則としては、

$$\gamma_t = \gamma_0 (1 - 0.7e^{-0.05t}) \quad (7)$$

とした。また、仮説の検証のため、時間の経過に従い割引率を

$$\gamma(t) = 0.7\gamma_0 e^{-0.05t} \quad (8)$$

のように指数関数的に減少させた場合についても実験を行った。

c) TD 誤差による割引率調整

TD 誤差については、式 (3) と同様、

$$\gamma(t) = \gamma_0 \min \left(1, \frac{1}{|\delta_t|} \right) \quad (9)$$

とした。

d) 信頼度による割引率調整調整

信頼度については、3.3 節で提案したような状態に対する信頼度ではなく、状態及び行動に対する信頼度 $R(s_t, a_t)$ を考え [7], その更新則は

$$R_{t+1}^2(s_t, a_t) = R_t^2(s_t, a_t) + \alpha_R {}^R\delta_t,$$

$$\text{where } {}^R\delta_t \stackrel{\text{def}}{=} \delta_t^2 + \gamma_R R_t^2(s_{t+1}, a_t) - R_t^2(s_t, a_t)$$

とした。

割引率 γ の調整則としては、式 (6) と同様、

$$\gamma_t = \gamma_0 \min \left(1, \frac{1}{R_t(s^*, a^*)} \right) \quad (10)$$

とした。ただし、 $s^* = (7, 3)$ はゴール地点 $G(8, 4)$ の斜め左上の地点 P の座標であり、 $a^* = 3$ は、右隣へマス進む行動である。よって $R_t(s^*, a^*)$ は、ゴールの手前地点 $P(7, 3)$ から右へ進もうとする行動の信頼度の指標に相当する。

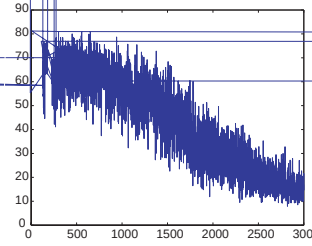
各種パラメータは、Actor のステップサイズパラメータ $\alpha_p = 0.1$, Critic の学習率 $\alpha_V = 0.1$, 信頼度の学習率 $\alpha_R = 0.1$, 逆温度パラメータ $\beta = 1$, 割引率の初期値 $\gamma_0 = 0.99$, 信頼度の割引率 $\gamma_R = 0$, 1 エピソードあたりの最大ステップ数 $n = 80$, エピソード数 $N = 1,800$ である。

4.2 実験結果

図 2 は割引率を (a) 調整しない, (b) 事前に調整をする, (c) TD 誤差を用いた調整をする, (d) 信頼度を用いた調整をする, のそれぞれに従わせた場合について、横軸にエピソード数、縦軸にスタートからゴールまでにかかったステップ数をプロットしたものである。図 3 は同様に、横軸にエピソードを、縦軸に 20 エピソードあたりの成功率をプロットしたものである。

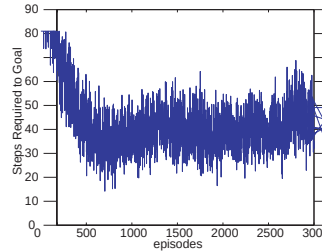
提案手法はどれも、従来手法（調整なし）と遜色ない、あるいはそれを上回る成績を示している。また、TD 誤差や信頼度を用いた場合に学習が加速されていることがわかる。特に信頼度を用いた場合の学習は非常に高速であり、安定している。TD 誤差を用いた場合は学習後期でやや成績が悪化しているが、これは前述した試行ごとのばらつきが原因と考えられる。

また、信頼度の指標 $R^2(s)$ を式 (5) に沿って変化させたときの、地点 P における $R^2(s)$ の時間推移を図 4 (a) に示す。またこの信頼度を用いて調節した割引率 γ の推移を図 4 (b) に示す。学習初期に信頼度が下がる（ $R^2(s)$ が上がる）ことで割引率が下がり、学習が進むにつれて信頼度が上がり、割引率も回復していくことがわかる。これにより、信頼度は学習進度の指標として適切であるということが出来る。



(c) TD 誤差を用いた調整

(c) Adjustment by TD error.



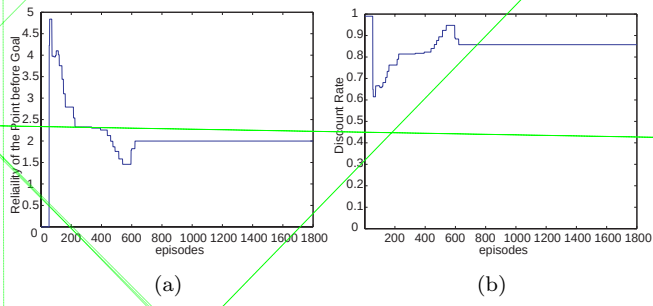
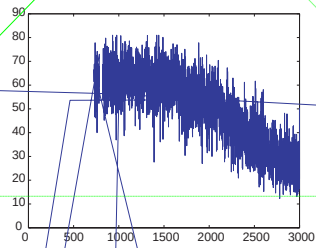


図 4 (a) 地点 P における信頼度の指標 $R^2(s)$ の推移. (b) 地点 P における割引率 γ の推移.
 Fig. 4 (a) Transition of index of reliability $R^2(s)$ at the point P.
 (b) Transition of discount rate γ at the point P.



れているのが、大脳基底核と強化学習とを関連させた一連の流れである [14], [15].

この流れの中で、割引率がセロトニンによって modulate されているとする仮説が近年提案された [16]. この仮説の検証は今後待たれるところであるが、臨床学的知見によれば、セロトニンはうつ病や衝動的行動に関連しているとされる. 例えば選択的セロトニン再取込み阻害剤 (Selective Serotonin-Reuptake Inhibitor, SSRI) は、さまざまな精神的障害における衝動的行動を減少させることが知られている [17]. このことは、前節でみた価値関数の即時的反映のような「気が早い」振舞いと類似しており、臨床的に同様の結果が出るとすれば興味深いことである.

6. む す び

強化学習における割引率を学習進度によって調整することの有用性を示し、学習初期に割引率を下げることによって学習性能が向上することを示した. 特に、信頼度を学習進度の指標として用いた場合に性能の向上が顕著であることがわかった.

今後は、本手法を適用可能な問題の範囲を明らかにする必要がある. Gridworld の次元や規模による影響なども調査していきたい.

謝辞 本研究にあたり貴重なご意見を頂きました電気通信大学の阪口豊助教授に心より感謝いたします.

文 献

- [1] R.S. Sutton, and A.G. Barto, Reinforcement Learning: An Introduction, The MIT Press, 1998.
- [2] T. Dean, K. Basye, and J. Shewchuk, "Reinforcement learning for planning and control," in Machine Learning Methods for Planning and Scheduling, ed. S. Minton, chapter 1, pp.1–20, Morgan Kaufmann, 1992.
- [3] A. Schwartz, "A reinforcement learning method for maximizing undiscounted rewards," Proc. 10th Int. Conf. Machine Learning (ICML93), pp.298–305, 1993.
- [4] S. Mahadevan, "Average reward reinforcement learning: Foundation, algorithms, and empirical results," Machine Learning, vol.22, pp.159–196, 1996.
- [5] K. Doya, "Metalearning and neuromodulation," Neural Networks, vol.15, no.4–6, pp.495–506, June/July 2002.
- [6] Y. Sakaguchi, and M. Takano, "Learning to switch behaviors for different environments: A computational model for incremental modular learning," Proc. 2001 Int. Symp. Non-linear Theory and its Applications (NOLTA2001), pp.383–386, Oct. 2001.
- [7] 阪口豊, 高野光雄, "環境変化への適応と文脈切り替え," 第 16 回生体・生理工学シンポジウム予稿集, Aug. 2001.
- [8] 阪口豊, 高野光雄, "内部モデルの信頼度に基づく強化学習のアルゴリズム," 日本神経回路学会第 11 回全国大会予稿集, Sept. 2001.
- [9] 吉田和子, 石井信, "強化学習における exploration と exploitation の制御," 信学技報 NC2001-28, pp.41–48, June 2001.
- [10] 岡田浩之, 山川宏, 大森隆司, "環境同定と報酬獲得のトレードオフを解消する報酬・嫌悪の二次元評価強化学習の提案," 日本ロボット学会誌, vol.19, no.2, pp.96–103, March 2001.
- [11] 尾形哲也, 松本典剛, 菅野重樹, "内分泌系モデルによる強化学習パラメータの動的調整," 第 19 回日本ロボット学会学術講演会講演論文集, pp.1239–1240, Sept. 2001.
- [12] 阪口豊, "内部モデルの信頼度に基づく運動計画のアルゴリズム," 信学論 D-II, vol.J79-D-II, no.2, pp.248–256, Feb. 1996.
- [13] 阪口豊, 高野光雄, "強化学習と教師あり学習を組み合わせたプリズム適応のモデル," 信学技報 NC2000-169, pp.99–106, March 2001.
- [14] K. Doya, "What are the computations of the cerebellum, the basal ganglia and the cerebral cortex?," Neural Networks, vol.12, pp.961–974, 1999.
- [15] W. Schultz, P. Dayan, and P.R. Montague, "A neural substrate of prediction and reward," Science, vol.275, pp.1593–1599, March 1997.
- [16] K. Doya, "Metalearning, neuromodulation, and emotion," in Affective Minds, eds. G. Hatano, N. Okada, and H. Tanabe, pp.101–104, Elsevier Science B. V., 2000.
- [17] E. Hollander, and M. Evers, "New developments in impulsivity," Lancet, vol.358, pp.949–950, Sept. 2001.