1G1-1

マルチエージェント学習下における温度パラメータの調節手法

Adaptation of Temperature Parameter under Multiagent Learning Environments

野田五十樹[*1] Hyun-Tae Kim[*2]
Itsuki NODA

[*1]産業技術総合研究所 [*2]Sungkyunkwan University
AIST

In this article, I investigate influence of temperature parameters to reinforcement learning. In dynamic environments, learning agent need to adjust the temperature parameter of Boltzmann action selection to a certain positive value that is suitable for the learning target. I formalize the dynamics of the environment as a random-walking process of expected rewards of each actions, and analyze relationship among the temperature, random-walk factors, and Q values that acquired by normal Q-learning procedure.

1. はじめに

強化学習などの一般の機械学習の手法では、学習エージェントが活動する環境は不変であり、期待報酬などの統計的な量は学習を通じて変化しないと仮定している。例えば強化学習法の一種、Q 学習では、期待報酬を更新していく際に用いられるステップサイズパラメータは、学習を通じて徐々に 0 に近づけていくことが多い。これは、環境が統計的に不変であることを仮定し、個別の報酬に含まれる雑音の影響を低減しつつ、真の期待報酬を得るためである。

一方、機械学習の適用が期待されている分野は、環境が動的に変化し、事前に正しい動作や処理を決めておくことが困難な場合であることが多い。例えば、ネットワークのルーティングなどにおいては、その時々のランダムな変化以外にも、ネットワーク構成の変更や新しいネットワークサービスの追加などで、日々、ネットワーク環境が少しずつ変化している。このような環境に機械学習を適用するためには、例えば上記のステップサイズを単純に 0 にするわけにはいかない。さらに、環境が絶えず変化する状況では、ある時点で学習が終了するということはなく、学習が行われている状況が持続すると仮定することが妥当である。この場合、従来の機械学習におけるパラメータ設定法をそのまま適用するわけにはいかない。[Nod09, Nod10] はこのような問題に対し 、学習を通じて変化する環境に適応・追従してステップサイズパラメータを調整する方法を提案している。この方法では、ステップサイズの変化による期待報酬の平均予測誤差の変化を学習と並行して求め、誤差を最小化するようにステップサイズを調整している。

学習中の行動選択に用いられるボルツマン (soft-max) 選択における温度パラメータ T も学習を制御する重要なパラメータである。このパラメータが大きい場合には、エージェントの行動選択はランダムに近くなり、0 になるにしたがって最大の期待報酬を持つ行動を選択しやすくなる。Q 学習などの状態遷移を含む強化学習では学習時にあらゆる状態からのあらゆる行動選択を十分試行する必要があるため、この T を調節する。多くの場合、学習が進むに従い T を 0 に漸近させる方法がとられる。

連絡先: 野田五十樹
産業技術総合研究所 サービス工学研究センター
茨城県つくば市梅園 1-1-1 産総研中央第 2
Tel.&Fax. 029-862-6517, i.noda@aist.go.jp

しかしステップサイズパラメータ同様、動的に変化する環境では、この T の漸減的調整方法を単純に適用することができない。T がほぼ 0 の場合には、期待報酬が最大ではない行動が試行される確率はほぼ 0 となるため、環境が変化してその行動の真の期待報酬が増減しても、学習に反映させることができなくなる。つまり、動的環境では T はある程度の大きさを保っておく必要がある。一方、マルチエージェントなどで複数のエージェントが同時に学習を進める場合、あるエージェントの行動の揺らぎは、他のエージェントの学習の妨げとなる。例えば、図 1 は、2 エージェントが動的に変化する調整ゲームを学習した場合の、温度パラメータの影響の様子である。この図の (a) では温度が十分に低いためエージェントの行動選択の揺らぎが小さく、お互いの学習に悪影響をおよぼしていないが、温度の高い (b) では、お互いの行動選択の揺らぎがノイズとして相手の学習に影響してしまい、調整ゲームにおける協調行動をうまく学習できずに終わっている。

このように、温度パラメータの調整は単純ではなく、環境にあわせて適切に設定されなければならない。本稿では、この温度パラメータ T について、動的環境にあわせて自動的に調整する方法の確立に向け、温度パラメータが動的環境の学習に与える影響を分析する。

2. 温度パラメータと動的環境

2.1 ボルツマン選択

ボルツマン選択では、以下の確率に従って各行動 a を選択する。

$$\mathcal{P}(a|s) \quad = \quad \frac{e^{Q(s,a)/T}}{\sum_{a'} e^{Q(s,a)/T}}$$

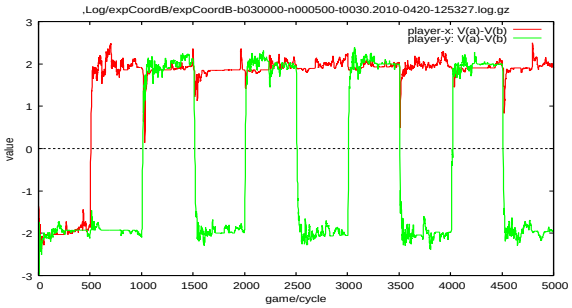
ただし、s はエージェントのいる状態、$Q(s,a)$ は状態 s における行動 a の期待報酬の推定値である。

また、Q 学習では、この期待報酬の推定値を以下の式で更新していく。

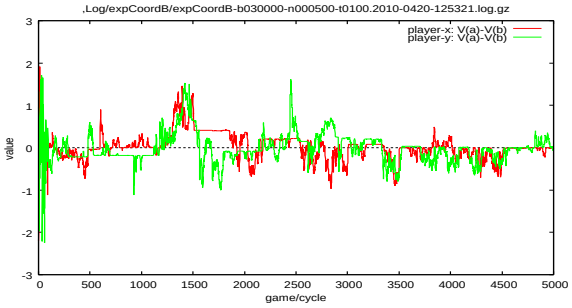
$$Q(s,a) \quad \leftarrow \quad (1-\alpha)Q(s,a) + \alpha(r + \max_{a'} Q(s',a')) \quad (1)$$

ただし、r は状態 s で行動 a を選択した際に実際に得られた報酬を表す。

以下で行う議論では状態遷移が本質に関わらないので、ここでは展開を容易にするため、状態 s を無視し、(1) 式の右辺第



(a) $T = 0.3$ の場合: 2 エージェントの行動がうまく分化し、最適行動を学習できている。



(a) $T = 1.0$ の場合: 相互の行動選択が定まらず、学習が影響しあい、最適行動を学習できていない。

図 1: 2 エージェントが同時に調整ゲームを学習した場合の Q 値の挙動

2 項も αr と簡単化する[*1]。また、行動についても 2 種の行動 (a,b) のみであると仮定する。よって、

$$\Delta Q \quad = \quad Q(a) - Q(b)$$

とすると、行動 a,b を選択する確率は、

$$\begin{aligned} \mathcal{P}(a) &= \mathcal{S}(\Delta Q; T) \\ \mathcal{P}(b) &= 1 - \mathcal{S}(\Delta Q; T) \end{aligned} \tag{2}$$

となる。ただし、

$$\mathcal{S}(x; T) \quad = \quad \frac{1}{1 + e^{-x/T}}$$

すなわちシグモイド関数である。また、以下の議論では $Q(a) \geq Q(b)$ としておく。

温度パラメータ T は、すべての行動系列を十分な回数試行できるように設定されなければならない。そのためには T の値は十分に大きな値である必要がある。しかし、T が大きすぎる場合、エージェントの行動選択は完全にランダムになってしまうため、状態数や行動の種類が多い場合には効率的な探索ができない。さらに、前節で述べているように、複数のエージェントが同時に学習する環境では、あるエージェントの行動選択のランダムさが他のエージェントにとっての外乱となるため、T はできるだけ 0 に近づけておく必要がある。

一方、T を 0 に近づけるということは、最適行動ではない b を選択する確率を下げることであり、それは $Q(b)$ の値を更新するためのサンプリングの間隔を長くすることに相当する。

*1 s が無視される場合、全ての a について max の項は一定となるため、無視できる。

もしエージェントの学習環境の統計的な性質が不変であれば、サンプリング間隔が長くなっても、十分に長い時間学習を行なえば、$Q(b)$ の値も真の期待報酬 $R(b)$ に近づくであろう。しかし、動的に変化する環境では、サンプリング間隔が長くなると、$Q(b)$ の値が $R(b)$ の時間変化に追従できなくなる。これを避けるためには、T を十分に大きく保ち、最適でない行動についても適切なサンプリング間隔を確保する必要がある。

以下では、この T に関するトレードオフを定量的に解消する指標の導入を考える。

2.2 動的環境と真の期待報酬推定

ここでは動的に変化する環境として、各行動の真の期待報酬 $R(a)$, $R(b)$ が徐々に変化していく状況を考える。以下では簡単のため、2 つの真の期待報酬の差 $\Delta = R(a) - R(b)$ を考え、この値がランダムウォークで徐々に変化していくものを考える。すなわち、ある時刻 t に真の期待報酬の差が Δ であるとき、次の時刻 $t+1$ にはその値が以下の式で変化するとする。

$$\Delta R \quad \leftarrow \quad \Delta R + \delta_R$$

ただし、δ_R は平均 0 、標準偏差 σ_{δ_R} の正規分布 $\mathcal{G}(\delta_R; 0, \sigma_{\delta_R})$ に従う乱数とする。

この ΔR の変化は単位時刻毎であるので、報酬のサンプリングの間隔が長くなれば、ランダムウォークの乱数の標準偏差もそれに応じて大きくなる。ここで平均のサンプリング間隔を τ とすると、サンプリング毎の ΔR の変化は、正規分布 $\mathcal{G}(\delta'_R; 0, \sqrt{\tau}\sigma_{\delta_R})$ に従う乱数 δ'_R を用いて、

$$\Delta R \quad \leftarrow \quad \Delta R + \delta'_R \tag{3}$$

で表される。一方、平均のサンプリング間隔 τ は各行動の選択確率に依存する。ここではより稀にしかサンプリングされない行動 b のサンプリング間隔のみを考慮すると、

$$\tau \quad = \quad \frac{1}{\mathcal{P}(b)} = 1 + e^{\Delta Q/T}$$

すなわち、温度 T を下げるとサンプリング間隔が長くなり、その分、(3) 式で表されるランダムウォークにより ΔR の不確かさが増すことになる。この不確かさの度合いを定量的に求めていくことにする。

上記では、真の期待報酬 R の変化について考察したが、これとその値の推定値である Q の変化について議論する。エージェントが実際に行動を行なって得られる報酬 r は、真の期待報酬 R に雑音が重畳したものとみなす。(この雑音の標準偏差を σ_ϵ としておく。) よって、(1) 式で更新される Q 値は R そのものではなく、R の最尤値とみなせる。ここで、R の尤度分布が平均 Q である正規分布で表されると仮定する。同様に、2 つの行動に対する R の差 ΔR の尤度分布も、平均 ΔQ、標準偏差 $\sigma_{\Delta R}$ の正規分布 $\mathcal{G}(\Delta R; \Delta Q, \sigma_{\Delta R})$ で表されるとする。

仮に、この尤度の標準偏差 $\sigma_{\Delta R}$ が行動 b のサンプリング直後であったとし、次に b が選択される直前での尤度を考えると、以下のようになる。

$$\mathcal{L}(\Delta R) \quad \propto \quad \mathcal{G}(\Delta R; \Delta Q, \sqrt{\sigma^2_{\Delta R} + \tau\sigma^2_{\delta_R}}) \tag{4}$$

さらに、実際の行動により観測される報酬 (の差)Δr はこの尤度分布に σ_ϵ の標準偏差を持つ雑音が重畳されると見なせるので、その確率分布は以下のようになる。

$$\mathcal{P}(\Delta r) \quad = \quad \mathcal{G}(r; \Delta Q, \sqrt{\sigma^2_{\Delta R} + \tau\sigma^2_{\delta_R} + \sigma^2_\epsilon})$$

ここで、実際に得られた報酬が $\Delta r'$ だったとすると、この時点での真の期待報酬の差 $\Delta R'$ の尤度分布は次のようになる。

$$\begin{aligned}\mathcal{L}(\Delta R') &\propto \mathcal{P}(\Delta R')\cdot\mathcal{G}(\Delta r'-\Delta R';0,\sigma_\epsilon)\\ &\propto e^{-\frac{(\Delta R'-\overline{\Delta R})^2}{2K}}\\ &\propto \mathcal{G}(\Delta R';\overline{\Delta R},\sqrt{K})\end{aligned}$$

ただし、

$$\begin{aligned}\overline{\Delta R} &= \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2+\sigma_{\Delta R}^2+\tau\sigma_{\delta_R}^2}\Delta Q+\frac{\sigma_{\Delta R}^2+\tau\sigma_{\delta_R}^2}{\sigma_\epsilon^2+\sigma_{\Delta R}^2+\tau\sigma_{\delta_R}^2}\Delta r'\\ K &= \frac{\sigma_\epsilon^2(\sigma_{\Delta R}^2+\tau\sigma_{\delta_R}^2)}{\sigma_\epsilon^2+\sigma_{\Delta R}^2+\tau\sigma_{\delta_R}^2}\end{aligned}$$

である。つまり、$\overline{\Delta R}$ の式はサンプリング直後の ΔR の最尤推定値であり、

$$\alpha = \frac{\sigma_{\Delta R}^2+\tau\sigma_{\delta_R}^2}{\sigma_\epsilon^2+\sigma_{\Delta R}^2+\tau\sigma_{\delta_R}^2}$$

とすれば、Q 学習の更新式 (1) 式 に対応することになる。

一方、K は ΔR の尤度分布を正規分布と見なしたときの標準偏差の自乗に当たるが、ここで、サンプリングによる ΔR の尤度分布推定が定常状態にあると仮定する。この場合、$\sigma_{\Delta R}^2=K$ と見なせるので、

$$\sigma_{\Delta R}^2 = \frac{-\tau\sigma_{\delta_R}^2+\sqrt{\tau^2\sigma_{\delta_R}^4+4\tau\sigma_{\delta_R}^2\sigma_\epsilon^2}}{2}$$

と求めることができる。

2.3 正答確率

次に、ボルツマン選択でエージェントが行動選択を行なう場合に、選ばれた行動が真の最適行動である確率、正答確率 $\mathcal{P}_{ok}$ を以下のように定義する。

$$\begin{aligned}\mathcal{P}_{ok} &= (\text{行動 } a \text{ が選ばれる確率})\times\\ &\quad (\text{行動 } a \text{ の真の期待報酬 } R(a) \text{ 最大である確率})\\ &= \mathcal{P}(a)\cdot\mathcal{P}(\Delta R>0) \qquad (5)\end{aligned}$$

ただし、行動 a は Q 値が最大の行動、すなわち $\Delta Q>0$ とする。このうち、$\mathcal{P}(a)$ は (2) 式で与えられているので、$\mathcal{P}(\Delta R>0)$ に注目する。

前節で議論したように、2 つの行動 a、b の Q 値の差が ΔQ であるとき、真の期待報酬の差 ΔR の尤度は (4) 式で与えられる。よって、この ΔR が正である確率は、以下の正規累積分布関数で表される。

$$\begin{aligned}\mathcal{P}(\Delta R>0) &= \int_0^\infty \mathcal{G}(\Delta R;\Delta Q,\sqrt{\sigma_{\Delta R}^2+\tau\sigma_{\delta_R}^2})\,d\Delta R\\ &= \int_{-\infty}^{\frac{\Delta Q}{\sqrt{\sigma_{\Delta R}^2+\tau\sigma_{\delta_R}^2}}}\mathcal{G}(\Delta R;0,1)d\Delta R\\ &= \Phi(\frac{\Delta Q}{\sqrt{\sigma_{\Delta R}^2+\tau\sigma_{\delta_R}^2}};0,1)\end{aligned}$$

ただし、Φ は正規累積分布関数である。

これらを (5) 式に代入すると、

$$\mathcal{P}_{ok} = \mathcal{S}(\Delta Q;T)\cdot\Phi(\frac{\Delta Q}{\sqrt{\sigma_{\Delta R}^2+\tau\sigma_{\delta_R}^2}};0,1) \qquad (6)$$

3. グラフによる考察

(6) 式で求められた正答確率について、温度パラメータとの関係をプロットすると図 2 となる。このグラフの各曲線は、$\Delta Q=1.0$、$\sigma_{\delta_R}=0.1$ と固定し、σ_ϵ の各々の値に対し、温度パラメータ T に対する正答確率 $\mathcal{P}_{ok}$ の変化を示している。また、グラフ中の赤の破線は、正答確率が最大となる点を結んだものである。

また、図 3 は、σ_ϵ を 0.1 に固定し、異なる ΔQ について温度パラメータ T に対する正答確率 $\mathcal{P}_{ok}$ の変化をプロットしたものである。

これらの図から、以下のことを読み取ることができる。

- 報酬に含まれる雑音成分が比較的小さい場合は、温度パラメータは高めとすべきである。特に雑音成分が Q 値の差の 1 倍〜10 倍程度の時には、雑音効果を低減させるために最良行動以外の行動も比較的まんべんなく選択すべきである。
- 雑音成分が非常に大きい場合は逆に、温度を下げる必要がある。これは、いずれの行動を選んで Q 値を更新しても、R の大小の不確かさを十分に低減できないため、まずは正答と思われる方 (Q 値が最大のもの) を選択した方が良いためである。

4. おわりに

本稿では、強化学習の行動選択における温度パラメータ T に注目し、真の期待報酬が徐々に変化する動的環境において、選択した行動が真の最適行動となる確率を最大化するという指標を用いて考察を行なった。

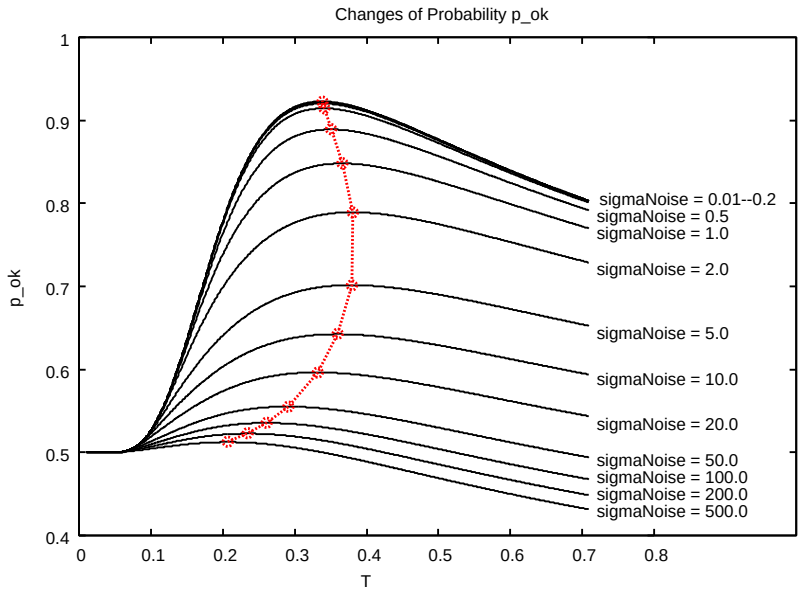
その結果、真の期待報酬がランダムウォークするという仮定の下では、行動における期待報酬の差 ΔQ、報酬に含まれる雑音の標準偏差 σ_ϵ、真の期待報酬の変化の度合いを示す σ_{δ_R} に依存して最適な温度パラメータが決まることが分かった。

今後は、この最適な温度パラメータを学習過程において推定する方法を構築数必要がある。

参考文献

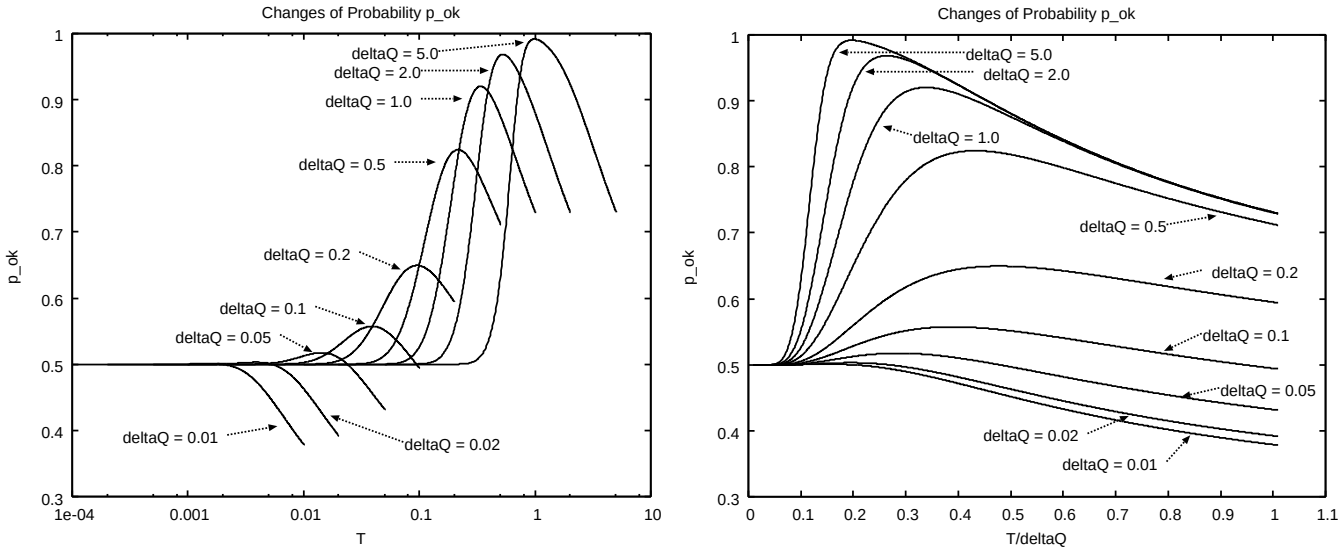
[Nod09] Itsuki Noda. Recursive adaptation of stepsize parameter for unstable environments. In ???, editor, *Proc. of ALA-2009*, pages ???–??? Springer???, May 2009.

[Nod10] Itsuki Noda. Adaptation of stepsize parameter to minimize exponential moving average of square error by newton's method. In Marek Grzes and Matthew Taylor, editors, *Proc. of Adaptive and Learning Agents Workshop*, page ??? AAMAS, May 2010.



図中の “sigmaNoise” は 本文中の σ_ϵ に対応する。また、$\Delta Q = 1.0$、$\sigma_{\delta_R} = 0.1$ である。

図 2: 温度 T に対する正答確率 $\mathcal{P}_{ok}$ の変化 (ΔQ 固定)



図中の “deltaQ” は 本文中の ΔQ に対応する。また、$\sigma_\epsilon = 0.1$、$\sigma_{\delta_R} = 0.1$ である。

図 3: 温度 T に対する正答確率 $\mathcal{P}_{ok}$ の変化 (σ_ϵ 固定)