F-044

# 強化学習におけるパラメータ設定に頑健な行動選択戦略

# An Action-Selection Strategy Robust to Parameter Settings in Reinforcement Learning

小野 兼嗣 † Ono Kenji　岩田 一貴 † Iwata Kazunori　林 朗 † Hayashi Akira

## 1 はじめに

強化学習 [1] が定常エルゴードマルコフ決定過程（MDP）に従うとき，状態-行動-報酬の時系列の典型集合の要素数は，MDP の確率密度関数のエントロピーによって決められる [2]．このエントロピーは MDP の確率的複雑さ (SC) と呼ばれる．エージェントの行動選択戦略のパラメータに関する SC の導関数は，探査と知識利用のジレンマ [1, Ch. 2] を調整する際に重要な指標となることが [2] において示されている．本論文では，その導関数を使って，代表的な行動選択戦略であるソフトマックス戦略のパラメータの新しい調整法を提案する．

## 2 提案手法

環境の状態の集合を $\mathcal{S} \triangleq \{s_1, \ldots, s_I\}$，エージェントの行動の集合を $\mathcal{A} \triangleq \{a_1, \ldots, a_J\}$ とする．時間ステップ $t \in \mathbb{N}$ までにエージェントが状態 $s_i \in \mathcal{S}$ に訪れた回数を $n_i(t) \in \mathbb{N}_0$，ゴール状態に到達した回数を $m(t) \in \mathbb{N}_0$ と表す．ただし，$\mathbb{N}_0 \triangleq \mathbb{N} \cup \{0\}$ である．逆に，エージェントがゴール状態に到達した回数が $m$ となったときの時間ステップを $t(m) \in \mathbb{N}$ と表す．ソフトマックス戦略の時間ステップ $t$ におけるパラメータ（温度の逆数）[1, Ch. 2][2] の値を $\beta^{(t)}$ と表記する．

任意の状態 $s_i \in \mathcal{S}$ についての $\beta_i^{(t)}$ を

$$\beta_i^{(t)} = c^{b(m(t))n_i(t)+d}, \tag{1}$$

とする．ただし，$c > 1$ および $d$ は任意の定数である．また，関数 $b : \mathbb{N}_0 \to \mathbb{R}$ は，任意の $m \in \mathbb{N}_0$ に対して

$$b(m) \triangleq \begin{cases} 0, & \text{if } m \leq 1, \\ \max\limits_{l \in \{2,\ldots,m\}} \sum\limits_{m'=2}^{l} \ln \dfrac{|\zeta(t(m'))|}{|\zeta(t(m'-1))|}, & \text{otherwise}, \end{cases} \tag{2}$$

と定義される．ただし，計算の都合上，$\zeta(t(m'-1)) = 0$ のとき，$\ln |\zeta(t(m'))|/|\zeta(t(m'-1))| = 0$ とした．関数 $\zeta : \mathbb{N} \to \mathbb{R}$ は $\beta$ に関する SC の導関数 [2] を表し，任意の $t \in \mathbb{N}$ に対して

$$\zeta(t) \triangleq -\sum_{i=1}^{I} \frac{\hat{v}_i^{(t)} \beta_i^{(t)}}{2\left(\sum_{j'' \in \mathcal{J}_i} \exp\left(\beta_i^{(t)} Q_{ij''}^{(t)}\right)\right)^2} \times \sum_{j \in \mathcal{J}_i} \sum_{j' \in \mathcal{J}_i} \left(Q_{ij}^{(t)} - Q_{ij'}^{(t)}\right)^2 \exp\left(\beta_i^{(t)} \left(Q_{ij}^{(t)} + Q_{ij'}^{(t)}\right)\right), \tag{3}$$

† 広島市立大学大学院情報科学研究科
〒731-3194 広島市安佐南区大塚東 3–4–1
Email: onokenji@robotics.im.hiroshima-cu.ac.jp

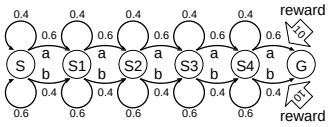



図 1　shortcut 環境

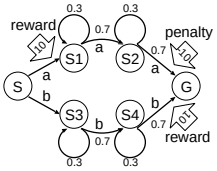



図 2　misleading 環境

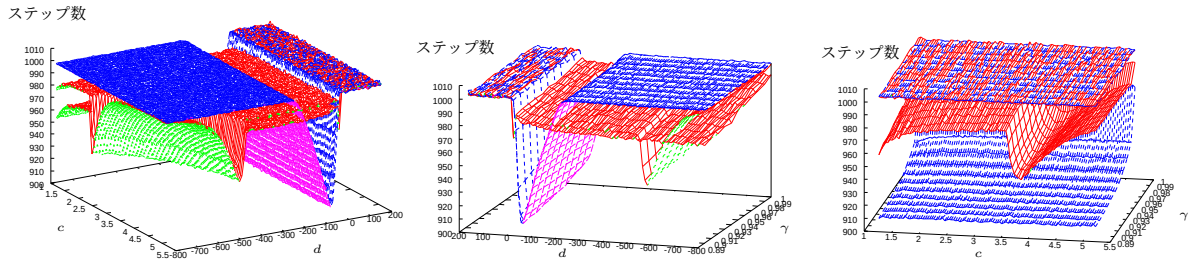
表 1　提案手法の方が良い結果となるパラメータの割合

| 環境 | $\left(\frac{\text{提案手法の方が良い結果となるパラメータの数}}{\text{サンプリングしたパラメータの総数}}\right) \times 100[\%]$ |
|---|---|
| shortcut | 74.48 |
| misleading | 88.33 |

によって近似される．ただし，$\hat{v}_i^{(t)} \triangleq n_i(t)/t$ とし，$\mathcal{J}_i$ は状態 $s_i$ で取りうる行動のインデックス集合，$Q_{ij}^{(t)}$ は時間ステップ $t$ における状態-行動 $(s_i, a_j) \in \mathcal{S} \times \mathcal{A}$ についての行動価値関数の推定値を示す．式 (1) において，関数 $b(m)$ の値を $m$ に依存しない定数とする方法が従来のソフトマックス戦略に相当する．

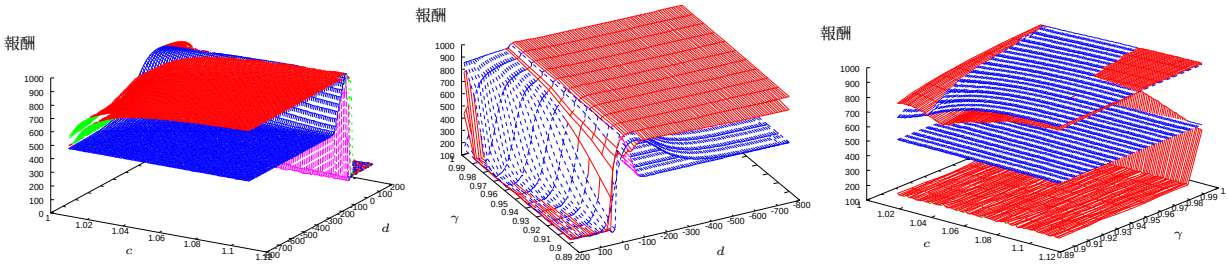
## 3 計算機実験

実験環境を図 1，2 に示す．図 1 の shortcut 環境 [3] の状態集合，行動集合はそれぞれ $\mathcal{S} = \{S, S1, \ldots, S4, G\}, \mathcal{A} = \{a, b\}$ であり，図 2 の misleading 環境 [3] では $\mathcal{S} = \{S, S1, \ldots, S4, G\}, \mathcal{A} = \{a, b\}$ である．図中の細い矢印は，丸で示す各状態で行動 a もしくは b を選択したときの状態遷移を表し，数字は状態遷移確率を表す．太い矢印の数字は，状態遷移のときにエージェントが受け取る報酬もしくは罰を示す．エージェントが初期状態 $S$ からゴール状態 $G$ に到達するまでを 1 エピソードとし，1 試行は 100 エピソードから成る．行動価値関数の推定には $Q$ 学習 [1, Ch. 6] を使い，学習率を $\alpha(n_{ij}(t)) = 50/(100 + n_{ij}(t))$ とした．ただし，$n_{ij}(t) \in \mathbb{N}_0$ は時間ステップ $t$ までに $s_i$ で $a_j$ が選択された回数を表す．また，$\gamma$ は $Q$ 学習の割引率を表す．行動価値関数の推定値は，すべての $i$，$j$ に対して，$Q_{ij} = 0$ で初期化した．図 1 の shortcut 環境では，1 エピソードあたりに得られる報酬が変わらないので，最




(a) $x$ 軸：$c$，$y$ 軸：$d$　(b) $x$ 軸：$d$，$y$ 軸：$\gamma$　(c) $x$ 軸：$\gamma$，$y$ 軸：$c$

図 3　shortcut 環境における各パラメータについての実験結果 (青：従来手法，赤：提案手法)




(a) $x$ 軸：$c$，$y$ 軸：$d$　(b) $x$ 軸：$d$，$y$ 軸：$\gamma$　(c) $x$ 軸：$\gamma$，$y$ 軸：$c$

図 4　misleading 環境における各パラメータについての実験結果 (青：従来手法，赤：提案手法)

適政策を最小ステップ数でゴール状態 $G$ に到達する政策，すなわち各状態で行動 $a$ を選択する政策とする．図 2 の misleading 環境では，1 エピソードあたりに得られる報酬が最も高い政策を最適政策とする．このような最適政策は，状態 $S$ で行動 $b$ を選択する政策である．実験では，3 つのパラメータ $c$，$d$，$\gamma$ の値を shortcut 環境では $[1.2, 5.2]$，$[-800, 200]$，$[0.9, 1]$ の範囲において，misleading 環境では $[1.01, 1.11]$，$[-800, 200]$，$[0.9, 1]$ の範囲においてある一定の間隔でサンプリングし，サンプリングした各パラメータに対して 10000 回の試行を行った．実験結果の評価として，shortcut 環境では 1 試行あたりの平均ステップ数，misleading 環境では 1 試行あたりの平均報酬を用いた．提案手法におけるパラメータ $\beta$ は (1) とし，従来手法においては (1) で $b(m) = 1$ と固定した式により与えた．

表 1 はサンプリングした各パラメータを用いて，従来手法と提案手法を shortcut 環境および misleading 環境で実行した結果，提案手法の方が良い結果となったパラメータの割合を表している．図 3，4 は各パラメータについての実験結果を図示したもので，青は従来手法，赤は提案手法による結果を示す．図 3 の $z$ 軸は 1 試行あたりの平均ステップ数，図 4 の $z$ 軸は 1 試行あたりの平均報酬を表す．図 3(a)，4(a) は $x, y$ 軸を $c, d$ とし，図 3(b)，4(b) は $x, y$ 軸を $d, \gamma$ とし，図 3(c)，4(c) は $x, y$ 軸を $\gamma, c$ とし，$x, y$ 軸にとっていない残り 1 つのパラメータを変化させたときの最大・最小値を表す．例えば，図 3(a) の結果は，従来手法および提案手法において $\gamma$ を変化させたときの最大・最小平均ステップ数を表し，図 4(a) おいては最大・最小平均報酬を表す．

表 1 から，サンプリングしたパラメータの大部分において提案手法の方が良い結果となることが確認できる．よって，提案手法は従来手法よりもパラメータ設定に頑強であることがわかる．次に，各パラメータの影響について詳しく調べる．図 3(a)，4(a) から，パラメータ $\gamma$ を変化させたときの最大・最小値の間に差がほとんど見られないことから，従来手法・提案手法ともにパラメータ $\gamma$ にほとんど依存しないことがわかる．図 3(b)，4(b) から，従来手法ではパラメータ $c$ を変化させたときの最大・最小値の間に差はあまり見られないものの，提案手法においては若干差が見られる．また，図 3(c)，4(c) から，従来手法・提案手法ともにパラメータ $d$ を変化させたときの最大・最小値の間に大きな差が見られる．従って，従来手法・提案手法ともに $\beta_i$ の初期値を決めるパラメータ $d$ の設定が重要であることがわかる．そこで，図 3，4 においてパラメータ $d$ に関する実験結果の変化に注目する．従来手法では良い結果となるパラメータ $d$ の範囲は非常に狭く，それ以外の範囲ではかなり悪い結果となることから，従来手法はパラメータ $d$ の変化に敏感である．一方，提案手法ではパラメータ $d$ の変化により頑強であることが確認できる．

## 4　まとめ

ソフトマックス戦略のパラメータ設定に頑強なパラメータ調整法を提案し，その有効性を計算機実験によって確認した．

## 参考文献

[1] R. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, The MIT Press, 1998.
[2] K. Iwata et al., *Neural Networks*, pp. 62–75, 2006.
[3] K. Iwata et al., *IEEE Trans. Neural Networks*, pp. 792–799, 2004.