

要 旨

本報告では、現在取るべき最適な行動が、過去の動作の系列(時系列)に依存するようなモデルを対象として、その時系列データを行動獲得の手掛かりとした強化学習を導入することで、与えられた行動目的を達成する手法を考案する。対象モデルとして周期的な運動、かつ安定性の保証が必要とするような代表的問題として簡単な二脚歩行問題を取り上げ、提案した手法に基づくその問題解決の一手法を示す。

1. 緒言

我々は、日常、実に多様で複雑な動作を特に意識することなく行っている。このことは我々人間に限らず、生物全体に関することであり、とても興味深い。「歩く」という動作ひとつをとってみても、路面変化(不整地面)への対応、歩行速度の調整、障害物からなどの危険回避などの判断を瞬時にやり、動作に移すことができる。これらの諸技能を材料的に定義することを目的として現在までに多くの制御手法が提案されてきたが、それらの多くは厳密な制御対象のモデルが必要であり、パラメータ変化や外乱の重視により、ロバスト性に欠け、生物の持つ柔軟な運動獲得能力を発揮するまでには至っていない。

本報告では、過去の情報を用いた時系列学習型の強化学習法を導入することで、柔軟な周期的運動のダイナミクスを獲得させる一手法を提案し、簡単な二脚歩行問題への適用を試みる。

2. 強化学習による時系列信号の処理

強化学習は一般に、行動戦略が明確なものに対して、連続的な行動決定を行うのに非常に有効な方法であるとされる。しかし、ある状態の要素が状態表現から欠けている(隠れている)とき、環境を完全に同定することは難しく、状態表現からそのときの環境での最適な行動決定はできないことが指摘できる。この問題は一般にhidden state problemとして知られている。歩行などの過去の行動が影響する運動の獲得には、この状態の隠れた要素(hidden states)の問題解決が不可避であると考えられる。

本報告では、この問題の解決法の一つとして、基本構造にQ-learningを用い、過去の情報(history information)を時系列処理をすることで隠れた状態を学習し、行動を決定する手法を提案する。

これまで、Q-learningは関数Q-functionに観測した状態を与え、学習することで、評価値が最大の行動を選ぶことにより与えられた状態内での最適な行動を得ることができるとされてきた[2]。しかし、先に述べたような隠れた状態要素が存在するときには最適な評価をする保証はない。このような認識の下にFig.1に示すような隠れた状態

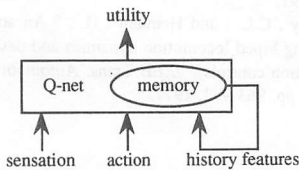


Fig.1 A memory architecture (recurrent-Q)

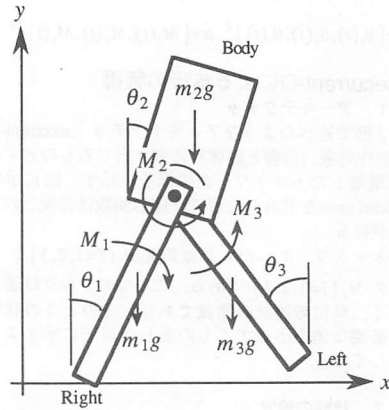


Fig.2 A model of biped locomotion

の可学習性をも目的とした新しいQ-learningのメモリーアーキテクチャが提案されている。図中のQ-netは、Q-functionをネットワークで表現したものであり、その詳細は参考文献[1]に示されている。

本報告では、このアーキテクチャを複数用いてネットワークを構築し、学習を行う。このことにより単一構造のネットワークによって学習するよりも、柔軟に非線形関数を内部に構築する能力の付与が期待される。

3. 二脚歩行のモデル

移動機構において、歩行は極めて動的であり、このような系に対する強化学習の手法の適用は困難とされてきた。逆に、このことは多数エージェントによる制御学習の有効性の検証として、この二脚歩行は興味深い対象である。

本報告では、二脚歩行における両脚と胴体が状態に応じた協調した動きをするように学習させるために、文献[3]に做った簡単なモデルを用いる(Fig.2)。このモデルは動作を矢状面に限定し、歩行モデル自身は $\theta_1, \theta_2, \theta_3$ の3自由度を持つ。右脚、左脚の駆動トルクを M_1, M_3 、胴体の角度調整トルクを M_2 とする。またここでは足首の動作はこれを無視するものとする。

二脚歩行での1サイクルの歩行運動中には、右脚支持相から左脚支持相に変化する相、左脚支持相、左脚支持相から右脚支持相に変化する相の4つの相が含まれる。本モデルでは、支持脚の変換相(両脚支持期)は一瞬で行われるとする。また、次の仮定条件を設定する。

- (1) 接地面は十分に堅く、大きい静止摩擦力が生じる。
- (2) 上方向に関する動作を無視する。
- (3) 全ての関節での摩擦力を無視する。

以上の条件でラグランジュの運動方程式を用いて床との間に拘束がないときの脚の運動方程式の一般形は次のように書ける[3].

$$A(\theta)\ddot{\theta} = H(\theta, \dot{\theta}) + Bu \quad (1)$$

where

$$\theta = [\theta_1(t), \theta_2(t), \theta_3(t)]^T, u = [M_1(t), M_2(t), M_3(t)]^T$$

4 Recurrent-Qによる歩行の獲得

4.1 アーキテクチャ

さて2節で述べたようなアーキテクチャ (recurrent-Q) を、各動作対象 (両脚と胴体) に割り当てるものとする。Fig.3に構築したネットワークの構造を示す。図に示すようにhidden unitsを共有した形で非線形関数は表現されることが判る。

なおネットワークへの入力は角度 $\theta_i (i=1,2,3)$, 出力はトルク $M_i (i=1,2,3)$ である。ここでのトルクは連続値ではなく、単位離散値の増減であり、そのときの状態に応じて必要な値を出力するものとし、以下に示すように定式化しておく。

4.2 状態の設定

制御対象から出力される状態変数をQ-learningで制御可能にするために、状態変数の空間を分割し、各状態に設定したQ-functionの値 (Q値) の最大な行動 (最適な行動) を選択する。モデルの各脚の角度 θ_1, θ_2 は一定角度 ($\pm 30^\circ$) 以上にはならないようにする。このことはモデルの転倒による歩行不能の状態を想定することである。同様に、胴体の角度 θ_3 が $\pm 30^\circ$ を越えた時も転倒すると判断する。この歩行可能な状態内で、脚の状態分割は、脚の支持状態 (支持脚か遊脚か) と位置 (角度) の二つの条件でコーディングを行う。角度は等分割を基本とする。また、歩行モデルの胴体は、常に地面に対して垂直になると仮定する。このことにより胴体の状態は $\theta_3 = 0$ を境界として進行方向とその逆方向の二値の状態と設定できる。例えば、 $\theta_1, \theta_2, \theta_3$ を各10分割にすると状態 S は以下ようになる。

$$S = (10 \times 2) \times (10 \times 2) \times 2 = 800$$

この状態の離散化は、強化学習の能力に大きく影響する。分割単位が大きい場合、各状態で選択した行動がその状態の最適な行動にならない。一方、分割単位を小さくしすぎた場合、状態数が多くなり、学習の収束性が悪くなり行動選択が最適に行われないことが予想される。このようなことを考慮しながら状態分割を行わなければならない。

4.3 報酬信号

この歩行モデルの達成すべき一種の動的目標状態は、モデルが転倒しないように各脚の支持状態が支持脚から遊脚、遊脚から支持脚を繰り返すことである。これは、まず第1に各関節角が、

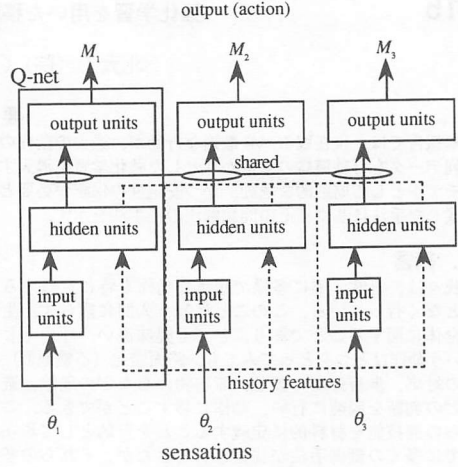


Fig.3 An architecture for recurrent Q-net

$$-\frac{\pi}{6} < \theta_1 < \frac{\pi}{6}, -\frac{\pi}{6} < \theta_2 < \frac{\pi}{6} \text{ and } -\frac{\pi}{6} < \theta_3 < \frac{\pi}{6} \quad (2)$$

で制御されるものとし、第2にそのときに全リンク (両脚と胴体) の支点が直前のサイクルに比較して進行方向に移動していなければならないという二つの条件と等価と見做すことができる。よってrecurrent-Qにはこれらの条件を満たした行動にrewardを与え、それ以外のものにはpenaltyを与えることになる。

5 結言

時系列の処理が必要なモデルに対して強化学習 (recurrent-Q) の再帰的な形式を用いることで、特に歩行に限定した運動の周期的行動獲得のための一手法を提案した。

参考文献

- [1] Lin, L.J. and Mitchel, T.M. " Reinforcement Learning With Hidden States," *From animals to animats 2* .pp.271-280, The MIT Press, 1993.
- [2] Whitehead, S.D. " Complexity and Cooperation in Q-learning," *Machine Learning*, Morgan Kaufmann, pp.363-367, 1991.
- [3] Golliday, C.L., and Hemami, H., " An approach to analyzing biped locomotion dynamics and designing robot locomotion controls," *IEEE Trans. Automatic Contr.* vol. AC-42, pp. 963-972, 1977.