論 文 内 容 の 要 旨

博士論文題目　方策勾配法に基づく強化学習法に関する研究

氏　名　森 健

（論文内容の要旨）

近年、方策勾配法と呼ばれる強化学習法が提案された。方策勾配法では、方策は、パラメータを持つ確率分布として与えられ、そのパラメータは、価値関数のパラメータに関する微分(方策勾配)から学習される。方策勾配は、方策のパラメータから決まる関数(互換関数)と価値関数の内積の期待値により与えられるので、価値関数を互換関数の空間へ射影した線形モデルを真の価値関数の代りに用いても、方策勾配に偏りは生じない。これにより、価値関数の漸近的な近似誤差が方策の改善に影響を与えなくなり、一般的な収束証明や分析を行うことができるようになるため、工学的な問題に対しても安心して適用できる強化学習法として期待されている。本論文では、まだあまり性質が知られていない方策勾配法を実際的な問題に適用して検証すると共に、新たな方策勾配法を提案することで、方策勾配法の可能性を探った。

従来の強化学習法では、状態行動空間上に与えた価値関数を近似推定する問題は、状態行動空間が複雑になるほど困難になる。一方、方策勾配法では、低次元の互換関数を用いて価値関数を近似できるため、状態行動空間が複雑になっても価値関数の近似推定は比較的容易であると考えられる。しかしながら、複雑な状態行動空間を持つ問題で実際に学習が容易になることは、これまで検証されてこなかった。本論文ではまず、方策勾配法を複雑な状態行動空間を持つ二足歩行運動シミュレータの制御則の獲得課題に対して適用し、方策勾配法が動的計画法に基づく強化学習法よりも速く安定して学習できることを示した。

方策勾配法では、方策を改善するごとに新たなサンプル系列を生成し方策勾配を推定する必要があり、学習が完了するまでに多くのサンプル系列の生成が必要になる。サンプル系列を生成するためには、ロボットなどの制御対象を実際に動かす必要があるが、手間や故障などのコストがかかるため問題である。本論文では、過去の方策の下で生成したサンプル系列が重点サンプリングと呼ばれる手法を用いて再利用できることに着目し、この問題を解決した。方策オフ型 Natural Actor-Critic 法 (Off-NAC 法) と呼ぶ新たな方策勾配法を提案し、ヘビ型運動シミュレータの制御則の獲得課題において、従来の方策勾配法よりも速く安定に学習できることを示した。また、提案手法をさらに高速化するための最適化法を提案した。

方策勾配法では、パラメータの数が多い複雑な方策についての価値関数の推定は、推定すべきパラメータの数が多くなることから困難であり、逆に単純な方策の価値関数の推定は容易である。本論文では、単純な方策と複雑な方策を組み合わせた階層的な方策の学習法として、階層型 off-NAC 法と呼ぶ方策勾配法を提案した。この手法では、サンプル系列の数が少ない学習初期には単純で抽象的な上層の方策を学習し、サンプル系列が増加するに従い、複雑で具体的な下層の方策を学習する。ヘビ型運動シミュレータの制御則の獲得課題において、階層型 off-NAC 法がOff-NAC 法よりも速く学習できることを示した。

氏　名	森　健

（論文審査結果の要旨）

強化学習法は、ロボットやプラントの自動制御など、明示的な教師信号が与えられない多くの複雑な工学的問題に対して、潜在的に有効であると考えられている。しかし、一般的な条件の下での収束が保証されておらず、学習過程も不安定であることから、現実的な工学的適用は困難とされてきた。近年、方策勾配法と呼ばれる強化学習法が提案された。方策勾配法では、収束は保証されており、学習過程も安定であると考えられることから、安心して適用できる強化学習法として期待されている。しかし、必要な分析や拡張がこれまでほとんどなされておらず、研究は進展していない。本論文では、方策勾配法の有効性が検証されるとともに、新たな二つの方策勾配法が提案された。本論文の主な成果は以下のように要約される。

1．方策勾配法では、従来の強化学習法では学習が困難であった複雑な工学的問題に対しても、学習が容易になると考えられる。しかし、これまで検証はあまりされてこなかった。本論文では、方策勾配法を複雑な工学的問題の一つである二足歩行運動シミュレータの制御則の獲得課題に対して適用し、方策勾配法が従来の強化学習法よりも速く安定して学習できることを示した。
2．方策勾配法では、方策を改善するごとに新たなサンプル系列を生成する必要があり、学習が完了するまでには多くのサンプル系列の生成が必要になる。サンプル系列の生成には、ロボットやプラントなどの制御対象を実際に動かす必要があるため、手間や故障などのコストがかかることが問題であった。本論文では、過去のサンプル系列を再利用できる、方策オフ型 Natural Actor-Critic 法（Off-NAC 法）と呼ぶ新たな方策勾配法を提案し、従来の方策勾配法よりも速く安定的に学習できることを示した。
3．方策勾配法では、複雑な方策の学習は困難であり、単純な方策の学習は容易である。本論文では、単純な方策と複雑な方策を組み合わせた階層的な方策の学習法として、階層型 off-NAC 法と呼ぶ新たな階層型強化学習法を提案し、階層型 off-NAC 法は Off-NAC 法よりもさらに速い学習を可能にすることを示した。

本論文では、これまで曖昧であった方策勾配法の有効性が証明された。さらに、方策勾配法を工学的な問題に適用する際の障害となっていた「学習に多くのサンプル系列が必要」、「学習すべき方策が複雑」という二つの問題に対する新たな解決法が提案され、その有効性が実証された。本論文の成果は、新しい強化学習法として期待されている方策勾配法の研究を著しく進展させるものであり、強化学習法の工学的な問題への適用可能性を広げることにも大きく貢献している。よって、博士（工学）の学位論文として価値のあるものと認める。