	ごとう　たかくに
氏　　名	後　藤　太　邦
授　与　学　位	博士（工学）
学位授与年月日	平成１８年９月１３日
学位授与の根拠法規	学位規則第４条第１項
研究科、専攻の名称	東北大学大学院工学研究科（博士課程）電気・通信工学専攻
学位論文題目	Reinforcement Learning Model for Manually Controlling Nonholonomic Systems （非ホロノミック手動制御系の強化学習モデルに関する研究）
指　導　教　員	東北大学教授　吉澤　誠
論文審査委員	主査　東北大学教授　吉澤　誠　　東北大学教授　阿曽　弘具 東北大学教授　川又　政征

論文内容要旨

現存する機械システムの多くは、複雑で非線形な特性をもつ。このようなシステムの一つである非ホロノミックシステムの多くは、目標値までの軌道を任意に選ぶことができないという特徴をもつ。そのため、非ホロノミックシステムの制御では、目標値に対する偏差のフィードバックループに、目標値に到達可能な軌道の計画を、何らかの形で導入する必要がある。この問題に対して統一された制御系の設計手法は今のところ確立されていない。したがって、個々のシステムに応じた制御法の開発が必要であり、その際システムに関する詳細な知識が求められる。しかし複雑なシステムの場合は、システムの詳細情報が未知の場合も多い。

このような複雑な未知システムの場合に、学習によって制御手法を獲得する枠組みがある。そのひとつが強化学習である。強化学習は、行動や制御入力を決定する「エージェント」が、訪れた状態や取った行動によって設計者より与えられる、「報酬」と呼ばれる強化信号の重みつき期待総和である「価値関数」を最大化するような行動規則「方策」を、試行錯誤によって探索・改善する学習の枠組みである。システムに関する詳細な知識を必要とせず、自動的に最適、もしくは準最適制御則を獲得できることが強化学習の主な利点であり、複雑化する機械システムの制御器として期待されている。しかし環境が複雑化してくると、素早い学習時間で適切な制御則を獲得する事が困難になる、いわゆる探索と知識利用のジレンマの問題に対処しなければならない。

一方、人間は複雑な機械システムの制御を、教示なしの場合においても試行錯誤によって効率よく探索し、獲得可能である事が知られている。このような人間オペレータの効率的探索戦略を解明し、その機構を学習アルゴリズムに組込むことで、上記ジレンマの解決が期待できる。本論文では、複雑なシス

テム、とりわけ非ホロノミックシステムにおけるこのような人間の探索戦略の調査・解析、ならびに強化学習における探索手法への応用に関しての研究をまとめており、全６章からなる。各章の概要を以下に列挙する。

第１章では、研究の背景と目的、本論文の構成について述べた。

第２章ではまず、強化学習の枠組みにおける基礎事項であるマルコフ決定過程、報酬、価値関数、方策についての諸定義を述べた。次に、基本的学習アルゴリズムである TD 学習やQ学習、Actor-Critic 法について説明した。最後に複雑な環境に対して基本的な探索手法を用いた場合に直面する探索と知識利用のジレンマの問題と、その問題に対して学習の効率化を図った従来の手法である楽観的初期化手法および shaping 手法の概念について紹介し、双方の問題点を指摘した。

第３章の前半では、試行錯誤による人間の探索戦略を解析する方法として、強化学習の枠組みにおける価値関数を用いた新しい解析手法を提案した。強化学習における価値関数は、予め設計された報酬と、その報酬環境において取られるエージェントの方策に従って形成される。一方、本解析手法における価値関数は、制御成績として人間オペレータに提示される評価関数を基に設計された報酬と、人間オペレータの取った方策、つまり制御入力および出力軌道からオンラインで形成される。そのため、形成された価値関数の形状の変化を評価することで、人間オペレータの制御則の変化を各学習段階に応じて検出可能である。

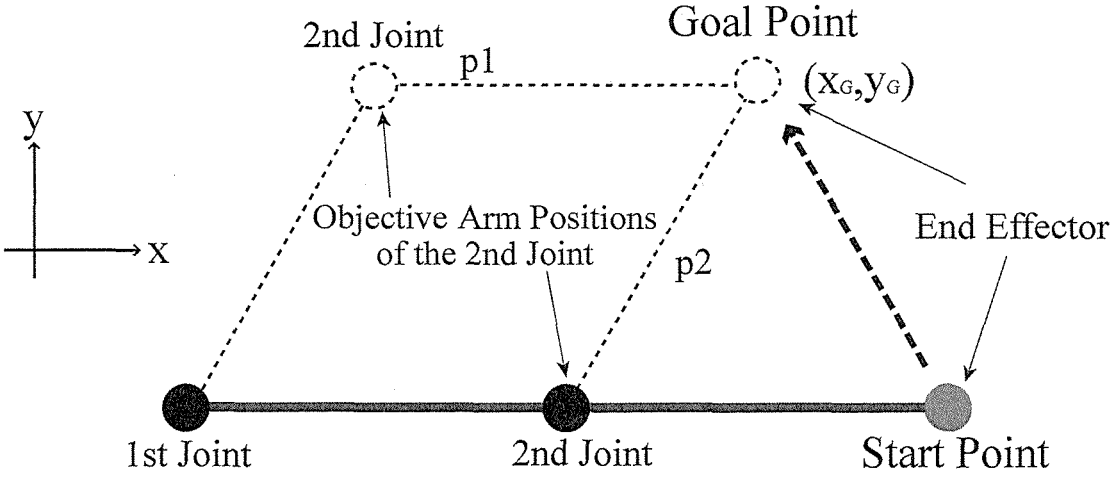

図 1. ２リンク平面型劣駆動マニピュレータ（２ＰＵＡＭ）を用いた制御タスク

後半では、試行錯誤による人間の学習過程を調査するために行った手動制御実験について述べた。実験環境は、図１に示される２リンク平面型劣駆動マニピュレータ（２ＰＵＡＭ）を用い、制御タスクは２ＰＵＡＭの先端をゴール位置で停止させる位置決め制御タスクとした。２ＰＵＡＭは第一関節が駆動関節、第二関節が非駆動となっており、目標位置で停止させることが困難な非ホロノミックシステムの１つである。人間オペレータの学習過程の解析には、各被験者の制御成績を表す評価値の推移と、提案した価値関数による解析手法の２つを用いた。解析結果から人間の学習過程について、次の２つの示唆が得られた。すなわち、（１）人間は教示がなくても、目標軌道の探索から目標軌道通過速度の向上へ至る段階的学習をおこなっていること、また、（２）速度向上段階では、発見した空間的な経路上に探索領域を拘束することで効率的に学習していることである。そして、このような段階的学習により、探索と知識利用のジレンマの問題に対処している可能性があることが示唆された。

第４章の前半では、第３章で得られた非ホロノミック手動制御系における人間の探索戦略に関する知見を基にした、強化学習における新たな探索手法を提案した。本手法は、探索空間を見込みのある空間領域内に合理的に拘束する働きをもつ。この働きは、探索領域を発見経路上に合理的に拘束するための shaping 関数と、価値関数の楽観的初期化によって実現される。提案した shaping 関数は、学習中にエージェントが目標状態へ到達した時の情報を利用して自動生成される点が従来手法と異なる。また、shaping 報酬の枠組みにおいて、shaping 関数は、元の報酬に追加する副報酬として用いられる。この副報酬を追加した場合に、元の報酬環境における最適方策と副報酬追加後の環境における最適方策が一致しているかどうかが重要な問題である。本論文で提案している shaping 関数は、この最適方策一致の条件を満たすことを第４章の後半で証明した。

第５章では、第４章で提案したアルゴリズムを評価するためにシミュレーション実験を行った。シミュレーション実験では、単純な環境における質点の位置決めタスクと、手動制御実験と同様の２ＰＵＡＭにおける位置決めタスクの２種類の物理シミュレーション環境を用いた。結果の比較は探索手法以外を同じ条件としたＱ学習や Actor-Critic 手法を用いて行った。本実

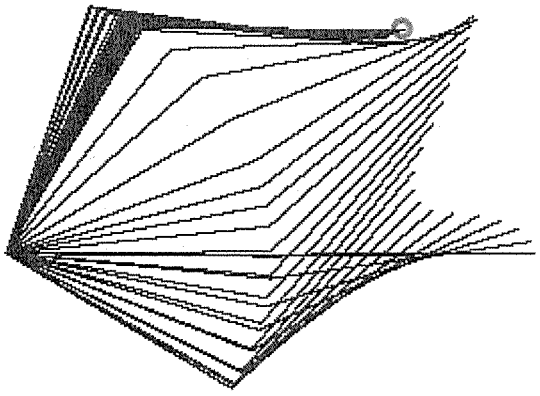

図２．提案手法が獲得した最良軌道

験により、単純な環境においては学習速度、２ＰＵＡＭにおいては学習速度に加え、獲得した制御則の両面で提案手法が優れていることを明らかにした。特に、２ＰＵＡＭの環境においては、図２に示されるように、目標位置でほぼ停止する軌道を獲得できた。これは手動制御実験において人間がとった最良の制御則と同等のものであり、提案手法が適切な制御則を効率的に獲得できることを示す結果が得られた。

第６章は結論であり、本研究で得られた主な結果を各章ごとにとりまとめ、今後の課題について論じた。

論文審査結果の要旨

未知なシステムの制御則を自動的に獲得する自律的探索手法の一つとして強化学習が挙げられる。強化学習を非ホロノミック系などの複雑な機械システムの制御に適用する場合，素早い学習時間で適切な制御則を獲得するという相反する２つの要求を同時に満たすことが困難になる，いわゆる探索と知識利用のジレンマという問題に直面する。一方，人間は複雑な環境においても学習によって適切な制御則を比較的効率よく獲得している。本論文は，このような人間の探索戦略の原理の調査および強化学習における探索手法への応用に関する研究を取りまとめたもので，全編６章からなる。

第１章は序論であり，本研究の背景と目的を述べている。

第２章では，本研究における基本事項であるマルコフ決定過程，報酬，価値関数，方策およびQ学習やActor-Critic法について述べている。続いて複雑な環境に対する強化学習の従来の探索手法である，楽観的初期化およびshaping関数の概念について紹介し，これまでの問題点を明らかにしている。

第３章では，非ホロノミック手動制御系における人間の探索戦略を調査する方法として，価値関数を用いた解析手法を提案している。本解析手法は，制御入力と出力軌道から人間の制御則に従った価値関数を作成する新しい手法であり，形成された価値関数を評価することで，人間の制御則の変化を視覚的に検出可能とするものである。次に，本解析手法を非ホロノミック系の１つである２リンク平面型劣駆動マニピュレータの手動制御実験に適用した結果について述べている。解析結果より，人間は教示がなくても，目標軌道の探索から目標軌道通過速度の向上へ至る段階的学習を行っていること，および，速度向上段階では，発見した空間的な経路上に探索領域を拘束することで効率的に学習していることを明らかにしている。これらは従来にない知見である。

第４章では，強化学習における探索を，見込みのある領域内に合理的に拘束するための新たな手法を提案している。本探索手法は，第３章で得られた人間の探索戦略に基づくものであり，探索領域を発見経路上に合理的に拘束するためのshaping関数を自動的に形成する特徴をもつ。一般に，shaping関数は強化学習における追加報酬としての働きをもつが，本論文で提案したshaping関数は，これを用いても元の報酬環境と比べて最適方策に変化がないことを明らかにしている。この性質は，実用上極めて重要であり，人間の学習・探索機構としても興味深い。

第５章では，第４章で提案したアルゴリズムを，２種類の物理制御タスクに適用したシミュレーション実験について述べている。実験結果より，探索手法以外を同条件とした従来のQ学習やActor-Critic法と比較して，学習速度および獲得した制御則の両面で提案方法が優れていることを明らかにしている。特に，手動制御実験と同じ制御タスクでは，人間がとった最良の制御則と同等の制御則を効率的に獲得することに成功している。このような最良制御則の獲得は従来法では非常に難しく，提案探索手法の妥当性ならびに有用性を示すものである。

第６章は結論である。

以上要するに本論文は，非ホロノミック系のような複雑な環境に対する人間の学習過程，特に探索戦略の特徴を解明し，この知見に基づく探索手法を提案することで，強化学習の主問題である探索と知識利用のジレンマに対する一解法を提案したものであり，機械学習ならびに制御工学の発展に寄与するところが少なくない。

よって，本論文は博士(工学)の学位論文として合格と認める。