(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第4586129号
(P4586129)

(45)発行日　平成22年11月24日(2010.11.24)　(24)登録日　平成22年9月17日(2010.9.17)

(51)Int.Cl.　F I
*G05B 13/02 (2006.01)*　G05B 13/02 J
G05B 13/02 L

請求項の数 9　(全 35 頁)

(21)出願番号　特願2008-77671(P2008-77671)
(22)出願日　平成20年3月25日(2008.3.25)
(65)公開番号　特開2009-230645(P2009-230645A)
(43)公開日　平成21年10月8日(2009.10.8)
審査請求日　平成21年2月12日(2009.2.12)

特許法第３０条第１項適用　平成１９年９月２５日　ｈｔｔｐ：／／ｉｓｗ３．ｎａｉｓｔ．ｊｐ／ＩＳ／ＴｅｃｈＲｅｐｏｒｔ／を通じて発表、平成１９年９月２６日　ｈｔｔｐ：／／ｌｉｂｒａｒｙ．ｎａｉｓｔ．ｊｐ／ｌｉｂｒａｒｙ／ｔｒ／ｉｎｄｅｘ．ｈｔｍｌを通じて発表、平成２０年１月３０日　平成１９年度　国立大学法人奈良先端科学技術大学院大学　情報科学研究科　博士学位論文公聴会に発表

(73)特許権者　506086797
独立行政法人沖縄科学技術研究基盤整備機構
沖縄県国頭郡恩納村字谷茶１９１９－１
(74)代理人　100064746
弁理士　深見　久郎
(74)代理人　100085132
弁理士　森田　俊雄
(74)代理人　100083703
弁理士　仲村　義平
(74)代理人　100096781
弁理士　堀井　豊
(74)代理人　100098316
弁理士　野田　久登

最終頁に続く

(54)【発明の名称】制御器、制御方法および制御プログラム

(57)【特許請求の範囲】
【請求項１】
対象とするシステムの時間発展が順方向マルコフ決定過程として記述される際に、前記システムの状態に対する制御則である確率的に表現される方策を前記システムの状態量の観測により方策勾配法によって強化学習する制御器であって、
前記方策に基づいて、前記システムを制御するための制御信号を生成する制御信号生成手段と、
前記システムの前記状態量を観測する状態量検知手段と、
前記状態と前記制御信号とに予め定められた関係で依存する報酬値を獲得する報酬値獲得手段と、
確率的な前記方策を規定するパラメータである方策パラメータにより前記方策が規定されるとき、各タイムステップにおける前記状態量と前記制御信号とに基づいて、前記システムの状態の分布の定常分布の対数の前記方策パラメータについての偏微分である対数定常分布偏微分を推定することで、前記方策の勾配を推定する方策勾配推定手段と、
前記報酬値と前記方策勾配推定手段による推定結果とに基づいて、前記対数定常分布偏微分を用いて推定した平均報酬偏微分の方向に前記方策パラメータを微小変化させることで、前記方策を更新する方策更新手段とを備える、制御器。
【請求項２】
前記方策勾配推定手段は、前記状態の分布の前記状態についての和が一定であるとの条件により導かれる制約条件であって、前記対数定常分布偏微分の順方向マルコフ連鎖につ

いての期待値が０であるという制約条件の下で、逆方向マルコフ連鎖に対するＴＤ学習により、前記対数定常分布偏微分を推定する、請求項１記載の制御器。

【請求項３】

前記ＴＤ学習においては、ｉ）前記逆方向マルコフ連鎖における前記方策の対数の偏微分の１ステップ前の観測値と１ステップ前の前記対数定常分布偏微分の和と、現在の状態の前記対数定常分布偏微分との差をδとするとき、前記δの２乗の前記順方向マルコフ連鎖についての期待値と、ｉｉ）前記対数定常分布偏微分の前記順方向マルコフ連鎖についての期待値の２乗との和を最小化することにより、前記対数定常分布偏微分を推定する、請求項２記載の制御器。

【請求項４】

対象とするシステムの時間発展が順方向マルコフ決定過程として記述される際に、前記システムの状態に対する制御則である確率的に表現される方策を前記システムの状態量の観測により方策勾配法によって強化学習する制御方法であって、

前記方策に基づいて、前記システムを制御するための制御信号を生成する制御信号生成ステップと、

前記システムの前記状態量を観測する状態量検知ステップと、

前記状態と前記制御信号とに予め定められた関係で依存する報酬値を獲得する報酬値獲得ステップと、

確率的な前記方策を規定するパラメータである方策パラメータにより前記方策が規定されるとき、各タイムステップにおける前記状態量と前記制御信号とに基づいて、前記システムの状態の分布の定常分布の対数の前記方策パラメータについての偏微分である対数定常分布偏微分を推定することで、前記方策の勾配を推定する方策勾配推定ステップと、

前記報酬値と前記方策勾配推定ステップによる推定結果とに基づいて、前記対数定常分布偏微分に基づき表現される平均報酬の勾配の方向に前記方策パラメータを更新することで、前記方策を更新する方策更新ステップとを備える、制御方法。

【請求項５】

前記方策勾配推定ステップは、前記状態の分布の前記状態についての和が一定であるとの条件により導かれる制約条件であって、前記対数定常分布偏微分の順方向マルコフ連鎖についての期待値が０であるという制約条件の下で、逆方向マルコフ連鎖に対するＴＤ学習により、前記対数定常分布偏微分を推定するステップを含む、請求項４記載の制御方法。

【請求項６】

前記ＴＤ学習においては、ｉ）前記逆方向マルコフ連鎖における前記方策の対数の偏微分の１ステップ前の観測値と１ステップ前の前記対数定常分布偏微分の和と、現在の状態の前記対数定常分布偏微分との差をδとするとき、前記δの２乗の前記順方向マルコフ連鎖についての期待値と、ｉｉ）前記対数定常分布偏微分の前記順方向マルコフ連鎖についての期待値の２乗との和を最小化することにより、前記対数定常分布偏微分を推定する、請求項５記載の制御方法。

【請求項７】

対象とするシステムの時間発展が順方向マルコフ決定過程として記述される際に、前記システムの状態に対する制御則である確率的に表現される方策を前記システムの状態量の観測により方策勾配法によって強化学習する制御方法をコンピュータに実行させるためのプログラムであって、

前記方策に基づいて、前記システムを制御するための制御信号を生成する制御信号生成ステップと、

前記システムの前記状態量を観測する状態量検知ステップと、

前記状態と前記制御信号とに予め定められた関係で依存する報酬値を獲得する報酬値獲得ステップと、

前記確率的な表現を規定するパラメータである方策パラメータにより前記方策が規定されるとき、各タイムステップにおける前記状態量と前記制御信号とに基づいて、前記シス

テムの状態の分布の定常分布の対数の前記方策パラメータについての偏微分である対数定常分布偏微分を推定することで、前記方策の勾配を推定する方策勾配推定ステップと、
前記報酬値と前記方策勾配推定ステップによる推定結果とに基づいて、前記対数定常分布偏微分に基づき表現される平均報酬の勾配の方向に前記方策パラメータを更新することで、前記方策を更新する方策更新ステップとを含む、制御方法をコンピュータに実行させるための制御プログラム。
【請求項８】
前記方策勾配推定ステップは、前記状態の分布の前記状態についての和が一定であるとの条件により導かれる制約条件であって、前記対数定常分布偏微分の順方向マルコフ連鎖についての期待値が０であるという制約条件の下で、逆方向マルコフ連鎖に対するＴＤ学習により、前記対数定常分布偏微分を推定するステップを含む、請求項７記載の制御プログラム。
【請求項９】
前記ＴＤ学習においては、ｉ）前記逆方向マルコフ連鎖における前記方策の対数の偏微分の１ステップ前の観測値と１ステップ前の前記対数定常分布偏微分の和と、現在の状態の前記対数定常分布偏微分との差をδとするとき、前記δの２乗の前記順方向マルコフ連鎖についての期待値と、ｉｉ）前記対数定常分布偏微分の前記順方向マルコフ連鎖についての期待値の２乗との和を最小化することにより、前記対数定常分布偏微分を推定する、請求項８記載の制御プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、方策勾配法により制御対象を制御する制御器、制御方法および制御プログラムの構成に関する。
【背景技術】
【０００２】
「マルコフ決定過程」として定式化される制御問題は、ロボット、プラント、移動機械（電車、自動車）などの自律的制御問題として、幅広い応用を持つ重要な技術である。
【０００３】
マルコフ決定過程に対する最適制御に関する従来技術として、いわゆる「強化学習」がある。
【０００４】
「強化学習」とは、エージェントが環境と相互作用を通じて試行錯誤し、得られる累積報酬量を最大化するような「方策」と呼ばれる行動則、すなわち、制御問題に用いる場合には、「制御規則」を学習する理論的な枠組みである。この学習法は、環境やエージェント自身に関する先験的な知識をほとんど必要としない点で様々な分野から注目を集めている。
【０００５】
強化学習は大まかに２つに分類できる。価値関数を用いて間接的に方策を表現し、価値関数を更新することで方策も更新される「価値関数更新法」と、方策を明示的にもち目的関数の勾配に従って方策を更新する「直接方策更新法（方策勾配法）」である。
【０００６】
方策勾配法は、行動のランダム性を制御するパラメータも方策パラメータに含めることで確率的方策の獲得が可能であり、また連続系への適用性も高いため、特に注目を集めている。しかし一般に実タスクへ適用すると、適切な行動則を獲得するまでの時間が非現実となることがある。そこで、複数学習器の同時利用、モデルの利用、教示信号の利用等の補助機構を入れて学習時間を短縮させる研究が活発に行われ、成果も著しい。
【０００７】
ここで、方策勾配強化学習法（ＰＧＲＬ）は、方策パラメータについての平均報酬の偏微分を用いることにより、方策パラメータを改善して平均報酬を最大化するための強化学

習（ＲＬ：Reinforcemnt Learning）の一般的なアルゴリズムである。ここで、平均報酬の偏微分は、方策勾配（ＰＧ：Policy Gradient）と呼ばれる。従来の方策勾配強化学習法アルゴリズム（ＰＧアルゴリズム）［非特許文献１，非特許文献２］は、状態の定常分布の偏微分の計算が困難であったため、方策パラメータの変化によりもたらされる定常分布の変化に依存する方策勾配の項を無視している。これらは、定常分布の対数の偏微分－ＬＳＤＧ（(Log) Stationary Distribution Gradients）－と呼ばれる。このような省略による偏りは、いわゆる割引率γを1に近づければ減少するが、一方で推定された偏微分の分散は多くなってしまう。このようなトレードオフは、現実に、適切なγを見いだすことを困難にしていた。マルコフ連鎖の時間を混合する（mixing）ことは適切なγを決定するための尺度である［非特許文献１，非特許文献３］。しかしながら、時間の混合は、方策に依存し、したがって、一般には、学習が完了する前に、予測することは困難である。

【０００８】

また、平均報酬ＰＧアルゴリズム［非特許文献４，非特許文献５］は、ポアソン方程式の解としての微分コスト関数を導入することにより、割引率を使用しないものである。しかしながら、割引率が１に近い通常のＰＧと平均報酬ＰＧとのパフォーマンスには、理論的には、大きな相違が無いことが示唆されている［非特許文献６］。

【０００９】

したがって、上述したような推定される方策パラメータの偏りを減少させると分散が増大するというトレードオフを解決した学習方法が必要である。

【００１０】

なお、以下、本文中で引用することとなる方策勾配学習法に関連した先行技術文献を以下に挙げる。

【非特許文献１】Baxter, J. and P. Bartlett (2001) “ Infinite-Horizon Policy-Gradient Estimation," Journal of Artificial Intelligence Research, Vol. 15, pp. 319-350.

【非特許文献２】H. Kimura and S. Kobayashi. An analysis of actor/critic algorithms using eligibility traces: Reinforcement learning with imperfect value function. In International Conference on Machine Learning, 1998.

【非特許文献３】S. Kakade. Optimizing average reward using discounted rewards. In Annual Conference on Computational Learning Theory, volume 14. MIT Press,2001.

【非特許文献４】J. N. Tsitsiklis and B. Van Roy. Average cost temporal-difference learning. Automatica, 35(11):1799-1808, 1999.

【非特許文献５】V. S. Konda and J. N. Tsitsiklis. On actor-critic algorithms. SIAM Journal on Control and Optimization, 42(4):1143-1166, 2001.

【非特許文献６】］J. N. Tsitsiklis and B. Van Roy. On average versus discounted reward temporal-difference learning. Machine Learning, 49(2):179-191, 2002.

【非特許文献７】P. W. Glynn. Likelihood ratio gradient estimation for stochastic systems. Communications of the ACM, 33(10):75-84, 1991.

【非特許文献８】R. Y. Rubinstein. How to optimize discrete-event system from a single sample path by the score function method. Annals of Operations Research, 27(1):175-212, 1991.

【非特許文献９】A. Y. Ng, R. Parr, and D. Koller. Policy search via density estimation. In Advances in Neural Information Processing Systems. MIT Press, 2000.

【非特許文献１０】D. P. Bertsekas. Dynamic Programming and Optimal Control, Volumes 1 and 2. Athena Scientific, 1995.

【非特許文献１１】R. S. Sutton and A. G. Barto. Reinforcement Learning. MIT Press, 1998.

【非特許文献１２】R. S. Sutton. Learning to predict by the methods of temporal differences. Machine Learning, 3:9-44, 1988.

【非特許文献１３】S. J. Bradtke and A. G. Barto. Linear least-squares algorithms

for temporal difference learning. Machine Learning, 22(1-3):33-57, 1996.
【非特許文献１４】J. A. Boyan. Least-squares temporal difference learning. Machine Learning, 49(2-3):233-246, 2002.
【非特許文献１５】R. B. Schinazi. Classical and Spatial Stochastic Processes. Birkhauser, 1999.
【非特許文献１６】J. Peng and R. J. Williams. Incremental multi-step Q-learning. Machine Learning, 22:283-290, 1996.
【非特許文献１７】P. Young. Recursive Esimation and Time-series Analysis. Springer-Verlag, 1984.
【非特許文献１８】D. P. Bertsekas and J. N. Tsitsiklis. Neuro-Dynamic Programming. Athena Scientific, 1996.

【発明の開示】

【発明が解決しようとする課題】

【００１１】

すでに、（定常）状態分布の勾配を評価するための２つの方法があるが、これらは、本願発明の方法とは異なったものであり、以下のような問題がある。最初のものは、「尤度比勾配法」あるいは「スコア関数法」と呼ばれるものであり［非特許文献７，非特許文献８］、再生プロセスにしか適用できない問題がある［非特許文献１］。非特許文献９に開示された他の方法は、状態分布の直接的な評価ではなく、密度伝搬を伴う状態分布の評価を介して実行されるものである。したがって、これらの方法は、エージェントがどの状態にあるかの知識を必要とするのに対し、後に説明する本発明の方法では、ノイズを含む状態の特徴ベクトルを観測するのみでよい。

【００１２】

したがって、本発明は、上記のような問題点を解決するためになされたものであって、その目的は、方策パラメータの偏りを減少させると分散が増大するというトレードオフを解決した方策勾配学習方法を用いた制御器、制御方法または制御プログラムを提供することである。

【００１３】

本発明の他の目的は、汎用性のある方策勾配学習方法を用いた制御器、制御方法または制御プログラムを提供することである。

【課題を解決するための手段】

【００１４】

このような目的を達成するために、本発明の制御器は、対象とするシステムの時間発展が順方向マルコフ決定過程として記述される際に、システムの状態に対する制御則である確率的に表現される方策をシステムの状態量の観測により方策勾配法によって強化学習する制御器であって、方策に基づいて、システムを制御するための制御信号を生成する制御信号生成手段と、システムの状態量を観測する状態量検知手段と、状態と制御信号とに予め定められた関係で依存する報酬値を獲得する報酬値獲得手段と、確率的な方策を規定するパラメータである方策パラメータにより方策が規定されるとき、各タイムステップにおける状態量と制御信号とに基づいて、システムの状態の分布の定常分布の対数の方策パラメータについての偏微分である対数定常分布偏微分を推定することで、方策の勾配を推定する方策勾配推定手段と、報酬値と方策勾配推定手段による推定結果とに基づいて、対数定常分布偏微分を用いて推定した平均報酬偏微分の方向に方策パラメータを微小変化させることで、方策を更新する方策更新手段とを備える。

【００１５】

好ましくは、方策勾配推定手段は、状態の分布の状態についての和が一定であるとの条件により導かれる制約条件であって、対数定常分布偏微分の順方向マルコフ連鎖についての期待値が０であるという制約条件の下で、逆方向マルコフ連鎖に対するＴＤ学習により、対数定常分布偏微分を推定する。

好ましくは、ＴＤ学習においては、ｉ）逆方向マルコフ連鎖における方策の対数の偏微

分の１ステップ前の観測値と１ステップ前の対数定常分布偏微分の和と、現在の状態の対数定常分布偏微分との差をδとするとき、δの２乗の順方向マルコフ連鎖についての期待値と、ｉｉ）対数定常分布偏微分の順方向マルコフ連鎖についての期待値の２乗との和を最小化することにより、対数定常分布偏微分を推定する。

【００１６】

この発明の他の局面に従うと、対象とするシステムの時間発展が順方向マルコフ決定過程として記述される際に、システムの状態に対する制御則である確率的に表現される方策をシステムの状態量の観測により方策勾配法によって強化学習する制御方法であって、方策に基づいて、システムを制御するための制御信号を生成する制御信号生成ステップと、システムの状態量を観測する状態量検知ステップと、状態と制御信号とに予め定められた関係で依存する報酬値を獲得する報酬値獲得ステップと、確率的な方策を規定するパラメータである方策パラメータにより方策が規定されるとき、各タイムステップにおける状態量と制御信号とに基づいて、システムの状態の分布の定常分布の対数の方策パラメータについての偏微分である対数定常分布偏微分を推定することで、方策の勾配を推定する方策勾配推定ステップと、報酬値と方策勾配推定手段による推定結果とに基づいて、対数定常分布偏微分に基づき表現される平均報酬の勾配の方向に方策パラメータを更新することで、方策を更新する方策更新ステップとを備える。

【００１７】

この発明のさらに他の局面に従うと、対象とするシステムの時間発展が順方向マルコフ決定過程として記述される際に、システムの状態に対する制御則である確率的に表現される方策をシステムの状態量の観測により方策勾配法によって強化学習する制御方法をコンピュータに実行させるためのプログラムであって、方策に基づいて、システムを制御するための制御信号を生成する制御信号生成ステップと、システムの状態量を観測する状態量検知ステップと、状態と制御信号とに予め定められた関係で依存する報酬値を獲得する報酬値獲得ステップと、確率的な表現を規定するパラメータである方策パラメータにより方策が規定されるとき、各タイムステップにおける状態量と制御信号とに基づいて、システムの状態の分布の定常分布の対数の方策パラメータについての偏微分である対数定常分布偏微分を推定することで、方策の勾配を推定する方策勾配推定ステップと、報酬値と方策勾配推定ステップによる推定結果とに基づいて、対数定常分布偏微分に基づき表現される平均報酬の勾配の方向に方策パラメータを更新することで、方策を更新する方策更新ステップとを含む、制御方法をコンピュータに実行させる。

【発明を実施するための最良の形態】

【００１８】

以下の説明の構成の概要を説明すると、（１．本発明の概要）において、本発明の全体的な構成を説明し、（２．前提：方策勾配強化学習）では、従来のＰＧ法をレビューし、本発明のＬＳＤＧを評価する目的を概説する。（３．定常分布の対数の偏微分の推定）では、逆方向マルコフ連鎖の方法に基づく最小二乗ＴＤ法によるＬＳＤＧ（λ）アルゴリズムを説明する。（４．　ＬＳＤＧ推定による方策の更新）では、ＬＳＤＧ（λ）－ＰＧアルゴリズムを説明する。これは、ＬＳＤＧ（λ）を利用し、割引率γを用いない方法である。（５．数値計算の結果）では、提案する方法のパフォーマンスを確認するため、簡単なマルコフ決定過程（ＭＤＰ）における計算結果が示される。

【００１９】

（１．本発明の概要）

後に説明するように、本発明では、逆方向マルコフ連鎖の方法とＴＤ（temporal difference）学習アルゴリズムにより、定常分布の勾配としてのＬＳＤＧが導出される、新たな方策勾配法の枠組みを提案する。この枠組みにおいては、平均報酬の偏微分は、割引率γに依存せず、γ＝０とおくことにより、価値関数を学習する必要が無くなる。

【００２０】

以下、図面を参照して本発明の実施の形態について説明する。

以下の説明で明らかとなるとおり、本発明は、ロボット、プラント、移動機械（電車、

自動車）などの制御問題として、幅広い応用を持つ。

【００２１】

ただし、以下では、本発明の具体的な適用例として、特に簡単なロボットの自動制御問題を対象とするものとして説明を行う。また、数値計算の結果は、さらに簡単なモデルに対する比較を示している。しかしながら、本発明は、このような応用に限定されるものではなく、より一般的に、対象システムの時間発展が複雑な場合の対象システムの制御に適用することができる。そのようなものの例としては、巨大プラント（溶鉱炉、原子力プラント）、マルチリンクロボット（ヒューマノイドロボット）、ノンホロノームシステム（宇宙ステーション）、地下鉄ホームでの人の流れなどがある。これらは、いずれも古典的制御法での制御が困難であり、かつ重要な制御対象である。

【００２２】

（１．本発明のシステム構成）

図１は、本発明の制御方法および制御プログラムが適用される制御器を用いたシステム１０００の一例を示す概念図である。

【００２３】

図１を参照して、システム１０００は、制御対象となる被制御装置２００と、この被制御装置２００に対して制御信号を与えるためのコンピュータ１００とを備える。

【００２４】

図１を参照してこのコンピュータ１００は、ＣＤ－ＲＯＭ（Compact Disc Read-Only Memory ）上の情報を読込むためのＣＤ－ＲＯＭドライブ１０８およびフレキシブルディスク（Flexible Disk、以下ＦＤ）１１６に情報を読み書きするためのＦＤドライブ１０６を備えたコンピュータ本体１０２と、コンピュータ本体１０２に接続された表示装置としてのディスプレイ１０４と、同じくコンピュータ本体１０２に接続された入力装置としてのキーボード１１０およびマウス１１２とを含む。

【００２５】

図２は、このコンピュータ１００の構成をブロック図形式で示す図である。

図２に示されるように、このコンピュータ１００を構成するコンピュータ本体１０２は、ＣＤ－ＲＯＭドライブ１０８およびＦＤドライブ１０６に加えて、それぞれバスＢＳに接続されたＣＰＵ（Central Processing Unit ）１２０と、ＲＯＭ（Read Only Memory）およびＲＡＭ （Random Access Memory）を含むメモリ１２２と、直接アクセスメモリ装置、たとえば、ハードディスク１２４と、被制御装置２００とデータの授受を行うための通信インタフェース１２８とを含んでいる。ＣＤ－ＲＯＭドライブ１０８にはＣＤ－ＲＯＭ１１８が装着される。ＦＤドライブ１０６にはＦＤ１１６が装着される。

【００２６】

被制御装置２００からは、コンピュータ１００に対して被制御装置２００の状態を示すパラメータ（状態量）の情報、たとえば、被制御装置２００の可動部分の位置、速度、加速度、角度、角速度等の情報が与えられる。一方、コンピュータ１００からは、被制御装置２００に対して、これら状態量を制御するための制御情報が制御信号として与えられる。

【００２７】

なお、ＣＤ－ＲＯＭ１１８は、コンピュータ本体に対してインストールされるプログラム等の情報を記録可能な媒体であれば、他の媒体、たとえば、ＤＶＤ－ＲＯＭ（Digital Versatile Disc）やメモリカードなどでもよく、その場合は、コンピュータ本体１０２には、これらの媒体を読取ることが可能なドライブ装置が設けられる。

【００２８】

本発明の制御器の主要部は、コンピュータハードウェアと、ＣＰＵ１２０により実行されるソフトウェアとにより構成される。一般的にこうしたソフトウェアはＣＤ－ＲＯＭ１１８、ＦＤ１１６等の記憶媒体に格納されて流通し、ＣＤ－ＲＯＭドライブ１０８またはＦＤドライブ１０６等により記憶媒体から読取られてハードディスク１２４に一旦格納される。または、当該装置がネットワークに接続されている場合には、ネットワーク上のサ

ーバから一旦ハードディスク１２４にコピーされる。そうしてさらにハードディスク１２４からメモリ１２２中のＲＡＭに読出されてＣＰＵ１２０により実行される。なお、ネットワーク接続されている場合には、ハードディスク１２４に格納することなくＲＡＭに直接ロードして実行するようにしてもよい。

【００２９】

図１および図２に示したコンピュータのハードウェア自体およびその動作原理は一般的なものである。したがって、本発明の最も本質的な部分は、ＦＤ１１６、ＣＤ－ＲＯＭ１１８、ハードディスク１２４等の記憶媒体に記憶されたソフトウェアである。

【００３０】

なお、一般的傾向として、コンピュータのオペレーティングシステムの一部として様々なプログラムモジュールを用意しておき、アプリケーションプログラムはこれらモジュールを所定の配列で必要な時に呼び出して処理を進める方式が一般的である。そうした場合、当該制御器を実現するためのソフトウェア自体にはそうしたモジュールは含まれず、当該コンピュータでオペレーティングシステムと協働してはじめて制御器が実現することになる。しかし、一般的なプラットフォームを使用する限り、そうしたモジュールを含ませたソフトウェアを流通させる必要はなく、それらモジュールを含まないソフトウェア自体およびそれらソフトウェアを記録した記録媒体（およびそれらソフトウェアがネットワーク上を流通する場合のデータ信号）が実施の形態を構成すると考えることができる。

【００３１】

［制御方法の一般的説明］

以下、本発明の構成について、その理論的な構成をまず説明する。

【００３２】

（制御器の構成）

（２．前提：方策勾配強化学習）

以下では、マルコフ決定過程（ＭＤＰ）について考えることにし、制御対象（制御器の環境）は状態遷移確率（時刻ｔにおいて、状態$x_t$であるときに行動（制御）$u_t$を実行することで状態が$x_{t+1}$となる確率）と報酬関数$r_{t+1}=r(x_t, u_t)$（なお、$r_{t+1}=r(x_{t+1}, x_t, u_t)$の場合にも同様に議論できる）によって特徴づけられるものとする。なお、この報酬関数については、制御対象の制御目標に応じて予め定められているものとする。状態入力$x \in X$から行動出力$u \in U$への写像を方策と呼び、以下で説明するように確率的に表現する。方策は、パラメータθで、規定される。

【００３３】

ここでは、従来のＰＧＲＬアルゴリズムをレビューし、本発明の新しいアルゴリズムの主たるアイデアを提示する。有限な状態$X \ni x$の組と行動$U \ni u$とを有する離散時間ＭＤＰは、以下の状態遷移確率ｐと報酬関数$r_{+1}$によって規定される。

【００３４】

【数１】

$$p(x_{+1} \mid x, u)$$

$$r_{+1} = r(x_{+1}, x, u) \in [R_{\min}, R_{\max}]$$

【００３５】

ここで、記載の簡単のために、$x_{+1}$は、状態ｘにおいて、行動ｕにより与えられる次の状態であり、$r_{+1}$は、$x_{+1}$において観測された即時報酬である［非特許文献１０，非特許文献１１］。$x_{+k}$および$u_{+k}$は、それぞれ、状態ｘからｋ時間ステップ先の状態および行動であり、添え字が－ｋとなっていれば、その反対である。（ＭＤＰにおける）決定は、$\theta \in R^d$によりパラメータ化された以下の確率的方策πにしたがってなされる。

【００３６】

【数２】

$$\pi_\theta(x,u) \equiv p(\mathbf{u} \mid \mathbf{x};\theta)$$

【００３７】

以下の方策πは、全てのｘ∈Ｘおよびｕ∈Ｕに対して、θについて微分可能であると仮定する。

【００３８】

【数３】

$$\pi_\theta(x,u)$$

【００３９】

さらに、以下のような仮定をおく。

（仮定１）

以下のような状態遷移確率ｐと確率的方策πとを有するマルコフ連鎖Ｍ（θ）は、全ての方策パラメータについてエルゴート的（既約で非周期的）である。

【００４０】

【数４】

$$p(x_{+1} \mid x,u)$$

$$\pi_\theta(x,u) \equiv p(\mathbf{u} \mid \mathbf{x};\theta)$$

【００４１】

したがって、以下のただ１つの定常分布が存在する。

【００４２】

【数５】

$$d^\pi(\mathbf{x}) = \lim_{k\to\infty} p(x_{+k} = x \mid \pi_\theta) > 0$$

【００４３】

この定常分布は、初期状態には独立であって、以下の式を満たす。

【００４４】

【数６】

$$d^\pi(\mathbf{x}) = \sum_{x_{-1},u_{-1}} p_{M(\theta)}(x,u_{-1} \mid x_{-1}) d^\pi(x_{-1}) \qquad (1)$$

【００４５】

ここで、以下の式が成り立つ。

【００４６】

【数７】

$$p_{M(\theta)}(x,u_{-1} \mid x_{-1}) \equiv \pi_\theta(x_{-1},u_{-1}) p(x \mid x_{-1},u_{-1})$$

【００４７】

ＰＲＧＬの目的は、以下の「平均報酬」と呼ばれる即時報酬の平均を最大化する方策パラメータθ＊を見いだすことである。

【００４８】

【数８】

$$R(\theta) \equiv \lim_{K\to\infty} \frac{1}{K} E_{M(\theta)} \left\{ \sum_{k=1}^{K} r_{+k} \mid x \right\}$$

ここで、$E_{M(\theta)}$はマルコフ連鎖Ｍ（θ）についての期待値である。

【００４９】

仮定１の下では、平均報酬は、初期状態ｘには独立で、以下の式に等しいことが示される［非特許文献１０］：

【００５０】

【数９】

$$R(\theta) = \sum_{x_{+1},x,u} d^{\pi}(x)\pi_{\theta}(x,u)p(x_{+1} \mid x,u)r(x_{+1},x,u) \quad (2)$$

【００５１】

方策勾配ＲＬアルゴリズムは、方策パラメータθを、以下の式に示されるθについての平均報酬Ｒ（θ）の勾配の方向に更新する。

【００５２】

【数１０】

$$\nabla_{\theta} R(\theta) \equiv \left[ \frac{\partial}{\partial \theta_1} R(\boldsymbol{\theta}), \cdots, \frac{\partial}{\partial \theta_d} R(\boldsymbol{\theta}) \right]^T$$

【００５３】

以下、単に方策勾配（ＰＧ）と、しばしば呼ばれる。この方策勾配は、以下のように与えられる。

【００５４】

【数１１】

$$\nabla_{\theta} R(\theta) = \sum_{x_{+1},x,u} d^{\pi}(x)\pi_{\theta}(x,u)(\nabla_{\theta} \ln \pi_{\theta}(x,u) + \nabla_{\theta} \ln d^{\pi}(x))p(x_{+1} \mid x,u)r(x_{+1},x,u) \quad (3)$$

【００５５】

以下の式に示される定常分布の対数の偏微分の導出は簡単ではない。

【００５６】

【数１２】

$$\nabla_{\theta} \ln d^{\pi}(x)$$

【００５７】

そこで、従来のＰＧアルゴリズム［非特許文献１，非特許文献２］は、ＰＧのもう一つの表現を利用している。

【００５８】

【数１３】

$$\nabla_\theta R(\theta) = \sum_{x,u} d^\pi(x)\pi_\theta(x,u)\nabla_\theta \ln \pi_\theta(x,u) Q_\gamma^\pi(x,u) + (1-\gamma)\sum_x d^\pi(x)\nabla_\theta \ln d^\pi(x) V_\gamma^\pi(x) \quad (4)$$

【００５９】

ここで、割引率γ∈［０，１）で、それぞれ、行動価値関数Ｑと状態価値関数Ｖとは以下のように表される［非特許文献１１］。

【００６０】

【数１４】

$$Q_\gamma^\pi(x,u) \equiv \lim_{K\to\infty} E_{M(\theta)}\left\{\sum_{k=1}^K \gamma^{k-1} r_{+k} \mid x,u\right\}$$

$$V_\gamma^\pi(x,u) \equiv \lim_{K\to\infty} E_{M(\theta)}\left\{\sum_{k=1}^K \gamma^{k-1} r_{+k} \mid x\right\}$$

【００６１】

式（４）の第２項の寄与は、γが１に近づくにつれて小さくなるので［非特許文献１］、従来のアルゴリズム［非特許文献１，非特許文献２］は、γ～１とすることで、第１項のみからＰＧを近似している。このような省略による偏りは、割引率γを1に近づければ小さくなるが、推定の分散は多くなってしまう。

【００６２】

ここで、本発明では、もう１つのアプローチを提案する。そこでは、以下の式の定常分布の対数の偏微分（ＬＳＤＧ）を推定し、ＰＧを導出するために式（３）を用いる。

【００６３】

【数１５】

$$\nabla_\theta \ln d^\pi(x)$$

【００６４】

著しい特徴は、価値関数を学習する必要がなく、したがって、そのアルゴリズムは、割引率γの選択において、偏りと分散のトレードオフと関係がないことである。

（３．定常分布の対数の偏微分の推定）

以下では、最小二乗法に基づくＬＳＤＧ推定アルゴリズム、ＬＳＤＧ（λ）を提案する。この目的のために、エルゴート的なマルコフ連鎖Ｍ（θ）の逆過程を定式化し、ＬＳＤＧは、ＴＤ法［非特許文献１２，非特許文献１３，非特許文献１４］の枠組みで推定できることを示す。

（３．１　順方向および逆方向マルコフ連鎖の性質）

ベイズの理論を用いれば、現在の状態から過去の状態および行動の対への逆方向の確率は、以下の式で表される。

【００６５】

【数１６】

$$q(x_{-1},u_{-1} \mid x) = \frac{p(x \mid x_{-1},u_{-1})p(x_{-1},u_{-1})}{\sum_{x_{-1},u_{-1}} p(x \mid x_{-1},u_{-1})p(x_{-1},u_{-1})}$$

【００６６】

以下の事後確率ｑは事前分布ｐに依存する。

【００６７】

【数１７】

$$q(x_{-1},u_{-1} \mid x)$$

$$p(x_{-1},u_{-1})$$

【００６８】
以下のように、事前分布ｐが定常分布ｄと方策πに従うとき、事後分布ｑは、定常逆方向確率と呼ばれ、下付添え字Ｂ（θ）が加えられる。
【００６９】
【数１８】

$$p(x_{-1},u_{-1})=\pi_\theta(x_{-1},u_{-1})d^\pi(x_{-1})$$

ここで、$q_{B(\theta)}(x_{-1},u_{-1}\mid x)$は、以下のように表される：

$$q_{B(\theta)}(x_{-1},u_{-1}\mid x)=\frac{p(x\mid x_{-1},u_{-1})\pi_\theta(x_{-1},u_{-1})d^\pi(x_{-1})}{d^\pi(x)} \qquad (5)$$
$$=\frac{p_{M(\theta)}(x,u_{-1}\mid x_{-1})d^\pi(x_{-1})}{d^\pi(x)}$$

もしも、マルコフ連鎖が$q_{B(\theta)}(x_{-1},u_{-1}\mid x)$に従うならば、それを$p_{M(\theta)}(x,u_{-1}\mid x_{-1})$に従う順方向マルコフ連鎖Ｍ（θ）に関連する逆方向マルコフ連鎖Ｂ（θ）と呼ぶ。

【００７０】
マルコフ連鎖－Ｍ（θ）とＢ（θ）の両方－は、以下の２つの定理において記述されるように密接に関連している。
（定理１）
【００７１】
【数１９】
遷移確率$p_{M(\theta)}(x\mid x_{-1})\equiv\sum_{u_{-1}}p_{M(\theta)}(x,u_{-1}\mid x_{-1})$により特徴づけられるマルコフ連鎖Ｍ（θ）が既約でエルゴート的であるとき、$p_{M(\theta)}(x\mid x_{-1})$に対する逆方向（定常）遷移確率$q_{B(\theta)}(x_{-1}\mid x)\equiv\sum_{u_{-1}}q_{B(\theta)}(x_{-1},u_{-1}\mid x)$により特徴づけられる逆方向マルコフ連鎖Ｂ（θ）もまた、エルゴート的であり、Ｍ（θ）と同一のただ一つの定常分布を有する：

$$d_{M(\theta)}(x)=d_{B(\theta)}(x) \qquad (6)$$

ここで、$d_{M(\theta)}(x)\equiv d^\pi(x)$と$d_{B(\theta)}(x)$は、それぞれ、Ｍ（θ）とＢ（θ）の定常分布である。

【００７２】
（証明）
式（５）の両辺に以下の定常分布をかける。
【００７３】

【数２０】

$$d^{\pi}(x)$$

【００７４】

すると、全ての可能な行動$u_{-1} \in U$について総和をとると、以下の式が得られる：

【００７５】

【数２１】

$$q_{B(\theta)}(x_{-1} \mid x) d^{\pi}(x) = P_{M(\theta)}(x \mid x_{-1}) d^{\pi}(x_{-1}) \qquad (7)$$

【００７６】

そして、式（７）の両辺を可能な状態$x \in X$について総和をとると、以下の式が成り立つ。

【００７７】

【数２２】

$$\sum_x q_{B(\theta)}(x_{-1} \mid x) d^{\pi}(x) = d^{\pi}(x_{-1})$$

【００７８】

このことは、以下の２点を成立させる．（ｉ）Ｂ（θ）は、Ｍ（θ）と同一の定常分布を有すること、（ｉｉ）Ｂ（θ）はＭ（θ）と同じ既約な性質を有すること。

【００７９】

【数２３】

式（７）は、遷移確率$p_{M(\theta)}(x \mid x_{-1})$または$p_{B(\theta)}(x_{-1} \mid x)$により、行列表現$\mathbf{P}_{M(\theta)}$または$\mathbf{Q}_{B(\theta)}$にそれぞれまとめ、定常分布をベクトル表現$\mathbf{d}^{\pi}$にまとめると：

$$\mathbf{Q}_{B(\theta)} = diag(\mathbf{d}^{\pi})^{-1} \mathbf{P}_{M(\theta)}^{T} diag(\mathbf{d}^{\pi})$$

容易に、任意の自然数ｎに対して$(\mathbf{P}_{M(\theta)})^{n}$の対角成分は、$(\mathbf{Q}_{B(\theta)})^{n}$のそれに等しいことがわかる。

【００８０】

このことは、（ｉｉｉ）Ｂ（θ）がＭ（θ）と同じ非周期的な性質をもっていることを示唆する。式（６）は、（ｉ）－（ｉｉｉ）により直接証明される［非特許文献１５］。

（定理２）

【００８１】

【数２４】

$x_{-K}$の分布が$d^{\pi}(x)$にしたがうとき、ｋ∈[０，Ｋ]にわたる任意の関数$f(x_{-k},u_{-k})$

の和に関する順方向逆方向のマルコフ連鎖の期待値は、相互に等しい：

$$\mathbf{E}_{B(\theta)}\left\{\sum_{k=0}^{K} f(x_{-k},u_{-k})\,|\,x\right\}=\mathbf{E}_{M(\theta)}\left\{\sum_{k=0}^{K} f(x_{-k},u_{-k})\,|\,x,d^{\pi}(x_{-K})\right\}$$

（８）

ここで、$\mathbf{E}_{B(\theta)}$および$\mathbf{E}_{M(\theta)}$は、順方向および逆方向マルコフ連鎖Ｂ（θ）とＭ（θ）

とのそれぞれの期待値を示し、

$$\mathbf{E}\{\cdot\,|\,d^{\pi}(x_{-K})\}\equiv\mathbf{E}\{\cdot\,|\,p(x_{-K})=d^{\pi}(x_{-K})\}$$

である。式（８）は、Ｋ→∞の極限においても成り立つ。

【００８２】

（証明）

マルコフ連鎖の特性と式（５）を代入することにより、以下の関係が得られる。

【００８３】

【数２５】

$$\begin{aligned}&q_{B(\theta)}(x_{-1},u_{-1},\cdots,x_{-K},u_{-K}\,|\,x)\\&=q_{B(\theta)}(x_{-1},u_{-1}\,|\,x)\cdots q_{B(\theta)}(x_{-K},u_{-K}\,|\,x_{-K+1})\\&\propto p_{M(\theta)}(x,u_{-1}\,|\,x_{-1})\cdots p_{M(\theta)}(x_{-K+1},u_{-K}\,|\,x_{-K})d^{\pi}(x_{-K})\end{aligned}$$

【００８４】

このことは、有限のＫの場合において式（８）が成立することを証明する。定理１から以下の式が導かれるので、式（８）のＫ→∞の極限の場合も成立することが、すぐさま証明される。

【００８５】

【数２６】

$$\begin{aligned}\lim_{K\to\infty}\mathbf{E}_{B(\theta)}\{f(x_{-k},u_{-k})\,|\,x\}&=\lim_{K\to\infty}\mathbf{E}_{M(\theta)}\{f(x_{-k},u_{-k})\,|\,x,d^{\pi}(x_{-K})\}\\&=\sum_{x,u}\pi_{\theta}(x,u)d^{\pi}(x)f(x,u)\end{aligned}$$

【００８６】

定理１および定理２は、これらが、定常分布に収束する状態分布の下で、順方向マルコフ連鎖Ｍ（θ）からのサンプルは、そのまま、逆方向マルコフ連鎖Ｂ（θ）に関する推定に使用できることになるので、重要である。そして、これらは、後に利用されうるものである。

（３．２　逆方向から順方向のマルコフ連鎖のＬＳＤＧのためのＴＤ（Temporal Difference）学習法）

ＬＳＤＧは式（５）を用いて、以下のように分解される。

【００８７】

【数２７】

$$\nabla_\theta \ln d^\pi(x) = \frac{1}{d^\pi(x)} \sum_{x_{-1},u_{-1}} p(x \mid x_{-1},u_{-1}) \pi_\theta(x_{-1},u_{-1}) d^\pi(x_{-1})$$
$$\left\{ \nabla_\theta \ln \pi_\theta(x_{-1},u_{-1}) + \nabla_\theta \ln d^\pi(x_{-1}) \right\}$$
$$= \sum_{x_{-1},u_{-1}} q_{B(\theta)}(x_{-1},u_{-1} \mid x) \left\{ \nabla_\theta \ln \pi_\theta(x_{-1},u_{-1}) + \nabla_\theta \ln d^\pi(x_{-1}) \right\}$$
$$= \mathbf{E}_{B(\theta)} \left\{ \nabla_\theta \ln \pi_\theta(x_{-1},u_{-1}) + \nabla_\theta \ln d^\pi(x_{-1}) \mid x \right\}$$

… （９）

ここで、式（９）において$\nabla_\theta \ln d^\pi(x)$および$\nabla_\theta \ln d^\pi(x_{-1})$が存在することに注意して、式（９）の反復により、以下の式が得られる。

$$\nabla_\theta \ln d^\pi(x) = \lim_{K\to\infty} \mathbf{E}_{B(\theta)} \left\{ \sum_{k=1}^{K} \nabla_\theta \ln \pi_\theta(x_{-k},u_{-k}) + \nabla_\theta \ln d^\pi(x_{-K}) \mid x \right\}$$

… （１０）

【００８８】

式（１０）は、状態ｘのＬＳＤＧは、以下の式で表される方策の対数の偏微分の状態ｘから逆方向マルコフ連鎖Ｂ（θ）の無限区間の集積であることを暗示している。

【００８９】

【数２８】

$$\nabla_\theta \ln \pi_\theta(x,u)$$

【００９０】

式（９）および（１０）から、ＬＳＤＧは、Ｍ（θ）よりもむしろ逆方向マルコフ連鎖Ｂ（θ）についての、以下のような逆方向ＴＤ　δに関するＴＤ学習[非特許文献１２]により推定されうる。

【００９１】

【数２９】

$$\delta(x) \equiv \nabla_\theta \ln \pi_\theta(x_{-1},u_{-1}) + \nabla_\theta \ln d^\pi(x_{-1}) - \nabla_\theta \ln d^\pi(x)$$

【００９２】

ここで、最初の２つの項は、Ｂ（θ）における方策の対数の偏微分の１ステップ前の実際の観測値と１ステップ前のＬＳＤＧであり、現在の状態のＬＳＤＧが最後の項である。

【００９３】

【数３０】

δ（ｘ）はランダム変数である一方、$\mathbf{E}_{B(\theta)}\{\delta(x)\mid x\}=0$が成り立つ。ＬＳＤＧの推定のために、逆方向ＴＤエラーの二乗の平均$\mathbf{E}_{B(\theta)}\left\{\hat{\delta}(x)^2\right\}$を最小化することになる。ここで、$\hat{\delta}(x)$は、ＬＳＤＧ$\nabla_\theta \ln d^\pi(x)$よりもむしろＬＳＤＧ推定値$\hat{\nabla}_\theta \ln d^\pi(x)$によって構成される。$\hat{\delta}(x)^2$は、$\delta(x)^T\delta(x)$を簡単に示すものである。

【００９４】

適格度減衰率λ∈［０，１］と逆方向追跡時間ステップＫ∈Ｎを用いて、式（１０）は、以下のように一般化される。

【００９５】

【数３１】

$$\nabla_\theta \ln d^\pi(x)$$
$$=\mathbf{E}_{B(\theta)}\left\{\sum_{k=1}^{K}\lambda^{k-1}\left\{\nabla_\theta \ln \pi_\theta(x_{-k},u_{-k})+(1-\lambda)\nabla_\theta \ln d^\pi(x_{-K})\right\}+\lambda^k\nabla_\theta \ln d^\pi(x_{-K})\mid x\right\}$$

このような変形に従って、逆方向ＴＤは、逆方向ＴＤ$(\lambda),\delta_{\lambda,K}(x)$に変形される。

$$\delta_{\lambda,K}(x)\equiv\sum_{k=1}^{K}\lambda^{k-1}\left\{\nabla_\theta \ln \pi_\theta(x_{-k},u_{-k})+(1-\lambda)\nabla_\theta \ln d^\pi(x_{-K})\right\}$$
$$+\lambda^k\nabla_\theta \ln d^\pi(x_{-K})-\nabla_\theta \ln d^\pi(x)$$

ここで、偏り無しの特性、$\mathbf{E}_{B(\theta)}\{\delta_{\lambda,K}(x)\mid x\}=0$は、依然、保持される。λ＝１かつＫ→∞の極限において、$\mathbf{E}_{B(\theta)}\left\{\hat{\delta}_{\lambda,K}(x)^2\right\}$を最小化することは、ウィドロ－ホッフ（Widrow-Hoff）教師付き学習とみなされる。

【００９６】

上記のような設定でなくとも、大きなλやＫを用いたならば、このような最小化は従来の価値関数に対するＴＤ（λ）学習の場合のように、非マルコフ効果に対しては、より敏感ではない［非特許文献１６］。

【００９７】

【数３２】

ＬＳＤＧの推定として$\mathbf{E}_{B(\theta)}\left\{\hat{\delta}_{\lambda,K}(x)^2\right\}$を最小化するために、逆方向マルコフ連鎖Ｂ（θ）から取り出された多くのサンプルを集める必要がある。幸いなことに、定理１および定理２を利用すると、以下のように交換可能な特性を用いることができる：

$$\mathbf{E}_{B(\theta)}\left\{\hat{\delta}_{\lambda,K}(x)^2\right\}=\sum_x d_{B(\theta)}(x)\mathbf{E}_{B(\theta)}\left\{\hat{\delta}_{\lambda,K}(x)^2\mid x\right\}$$
$$=\sum_x d^{\pi}(x)\mathbf{E}_{M(\theta)}\left\{\hat{\delta}_{\lambda,K}(x)^2\mid x,d^{\pi}(x_{-K})\right\}$$
$$=\mathbf{E}_{M(\theta)}\left\{\hat{\delta}_{\lambda,K}(x)^2\mid d^{\pi}(x_{-K})\right\}$$

…（１１）

すなわち、$x_{-K}\sim d^{\pi}(x)$であれば、$\mathbf{E}_{B(\theta)}\left\{\hat{\delta}_{\lambda,K}(x)^2\right\}$を最小化するために、実際のサンプルを再利用できる。現実の問題では、しかしながら、初期状態は、定常状態$d^{\pi}(x)$から取り出されることはまれである。

【００９８】

理論的な仮定と現実への適用との間のギャップを埋めるために、以下の２つのうちのいずれかの戦略をとる必要がある。（ｉ）λ～１ならば、Ｋは、あまり大きな整数に設定しない、（ｉｉ））Ｋ～ｔならば、λは、１に設定しない、ここで、ｔは、現実の順方向マルコフ連鎖の現在のタイムステップである。

（３．３　ＬＳＤＧ推定アルゴリズム：制限付き逆方向ＴＤ（λ）の最小二乗法）

【００９９】

【数３３】

３．２では、ＬＳＤＧの推定がマルコフ連鎖Ｍ（θ）における$\hat{\delta}_{\lambda,K}(x)^2$の平均二乗$\mathbf{E}_{M(\theta)}\left\{\hat{\delta}_{\lambda,K}(x)^2\mid d^{\pi}(x_{-K})\right\}$を最小化することにより実行できるという理論を紹介した。しかしながら、ＬＳＤＧには、また、$\sum_x d^{\pi}(x)=1$との条件から導かれる以下のような制約条件がある。

$$\mathbf{E}_{M(\theta)}\left\{\nabla_\theta \ln d^{\pi}(x)\right\}=\sum_x d^{\pi}(x)\nabla_\theta \ln d^{\pi}(x)=\nabla_\theta\sum_x d^{\pi}(x)=0$$

…（１２）

【０１００】
この３．３では、最小二乗法に基づく［非特許文献１７，非特許文献１３，非特許文献１４］、ＬＳＤＧ推定アルゴリズム、ＬＳＤＧ（λ）を提案する。これは、同時に、平均二乗を減少させるとともに、制約条件を満足することを達成しようとするものである。
【０１０１】
【数３４】

ＬＳＤＧ推定$\hat{\nabla}_\theta \ln d^\pi(x)$が、線形ベクトル関数近似器$f(x;\Omega)\equiv\Omega\phi(x)$で表される。ここで、$\phi(x)\in R^e$は基底関数であり、$\Omega\equiv[\omega_1,\cdots,\omega_d]^T\in R^{d\times e}$は、調整可能なパラメータ行列であり、最適パラメータ$\Omega^*$は、$\nabla_\theta \ln d^\pi(x)=\Omega^*\phi(x)$を満たす。簡単のために、方策パラメータθのうち、ｉ番目の要素である$\theta_i$のみに注意を向けることにする。そして、$f(x;\omega_i)\equiv\omega_i^T\phi(x)$とし、$\nabla_{\theta_i}\ln\pi_\theta(x,u)\equiv\partial\ln\pi_\theta(x,u)/\partial\theta_i$とし、$\hat{\delta}_{\lambda,K}(x,\omega_i)$とは、$\hat{\delta}_{\lambda,K}(x)$のｉ番目の要素であるものとする。

【０１０２】
したがって、最小化すべき対象となる関数は、以下の式（１３）となる。
【０１０３】
【数３５】

$$\varepsilon(\omega_i)=\frac{1}{2}\mathbf{E}_{M(\theta)}\left\{\hat{\delta}_{\lambda,K}(x;\omega_i)^2\mid d^\pi(x_{-K})\right\}+\frac{1}{2}\mathbf{E}_{M(\theta)}\left\{f(x;\omega_i)\right\}^2 \quad \cdots(13)$$

【０１０４】
ここで、右辺の第２項は、式（１２）の制約条件のためのものである。したがって、式（１３）の偏微分は、以下のようになる。
【０１０５】

【数３６】

$$\nabla_{\omega_i}\varepsilon(\omega_i)=$$

$$\mathbf{E}_{M(\theta)}\left\{\hat{\delta}_{\lambda,K}(x;\omega_i)\nabla_{\omega_i}\hat{\delta}_{\lambda,K}(x;\omega_i)\,|\,d^{\pi}(x_{-K})\right\}+\frac{1}{2}\nabla_{\omega_i}\mathbf{E}_{M(\theta)}\left\{f(x;\omega_i)\right\}^2$$

…（１４）

ここで、以下の式が成り立つ。

$$\hat{\delta}_{\lambda,K}(x;\omega_i)=\sum_{k=1}^{K}\lambda^{k-1}\nabla_{\theta_i}\log\pi_{\theta}(x_{-k},u_{-k})+\omega_i^{T}\nabla_{\omega_i}\hat{\delta}_{\lambda,K}(x;\omega_i)$$

$$\nabla_{\omega_i}\hat{\delta}_{\lambda,K}(x;\omega_i)=(1-\lambda)\sum_{k=1}^{K}\lambda^{k-1}\phi(x_{-k})+\lambda^{K}\phi(x_{-K})-\phi(x)$$

従来の最小二乗法は、$\nabla_{\omega_i}\varepsilon(\omega_i)=\mathbf{0}$を満たすパラメータを真のパラメータ$\omega_i^{*}$として見つけることを目的とするものであるが、式（１３）の右辺第１項に関連する誤差$\hat{\delta}_{\lambda,K}(x,\omega_i^{*})$とその微分$\nabla_{\omega_i}\hat{\delta}_{\lambda,K}(x,\omega_i^{*})$との間に相関が存在すると、推定の偏りを生んでしまう。

すなわち、もしも、以下の式が成り立つとする。

$$\mathbf{E}_{M(\theta)}\left\{\hat{\delta}_{\lambda,K}(x;\omega_i^{*})\nabla_{\omega_i}\hat{\delta}_{\lambda,K}(x;\omega_i^{*})\,|\,d^{\pi}(x_{-K})\right\}\neq 0$$

すると、$\nabla_{\omega_i}\varepsilon(\omega_i^{*})\neq\mathbf{0}$となる。

【０１０６】

一般的なＲＬ問題では、このような相関が存在するので、このような偏りを除くために、操作変数法（instrumental variable method）を適用する［非特許文献１７，非特許文献１３］。

【０１０７】

【数３７】

そのためには、$\nabla_{\omega_i}\hat{\delta}_{\lambda,K}(x,\omega_i)$は、$\nabla_{\omega_i}\hat{\delta}_{\lambda,K}(x,\omega_i^*)$とは相関があるものの$\hat{\delta}_{\lambda,K}(x,\omega_i^*)$とは相関のない操作変数$\iota(x)$により置き換えられる。このような条件は、明らかに、ＬＳＴＤ（λ）［１３，１４］と同様に、$\iota(x)=\phi(x)$のときに満たされる。式（１４）の代わりに、真のパラメータ$\omega_i^*$、すなわち、$\tilde{\nabla}_{\omega_i}\varepsilon(\omega_i^*)=\mathbf{0}$を計算するために、以下の式を０とするパラメータを見いだすことを試みる。

$$\tilde{\nabla}_{\omega_i}\varepsilon(\omega_i)=\mathbf{E}_{M(\theta)}\left\{\hat{\delta}_{\lambda,K}(x;\omega_i)\phi(x)\,|\,d^{\pi}(x_{-K})\right\}+\mathbf{E}_{M(\theta)}\{\phi(x)\}\mathbf{E}_{M(\theta)}\{\phi(x)\}^{T}\omega_i$$

…（１５）

【０１０８】

以下では、現実のマルコフ連鎖Ｍ（θ）における時間ステップｔの状態を示すために、ノーテーションを$x_t$に変更する。提案するＬＳＤＧ推定アルゴリズム、ＬＳＤＧ（λ）は、適格性減衰率λ∈［０，１）の下で、逆方向にさかのぼる時間ステップＫを現在の状態$x_t$のタイムステップｔと同じにする。すなわち、以下が成り立つ。

【０１０９】

【数３８】

$$\hat{\delta}_{\lambda,K}(x_t;\omega_i) = g_{\lambda,i}(x_t) - (z_\lambda(x_{t-1}) - \phi(x_t))^T \omega_i$$

ここで、

$$g_{\lambda,i}(x_t) = \sum_{s=0}^{t} \lambda^{t-s} \nabla_{\theta_i} \ln \pi_\theta(x,u)$$

であり、

$$z_\lambda(x_t) = (1-\lambda)\sum_{s=1}^{t} \lambda^{t-s}\phi(x_s) + \lambda^t \phi(x_0)$$ である。

式（１５）における期待値は、以下のように推定される：

$$\lim_{K\to\infty} \mathbf{E}_{M(\theta)}\left\{ \hat{\delta}_{\lambda,K}(x;\omega_i)\phi(x) \mid d^\pi(x_{-K}) \right\}$$

$$\cong \frac{1}{T}\sum_{t=1}^{T} \phi(x_t)\left\{ g_{\lambda,i}(x_{t-1}) - (\phi(x_t) - z_\lambda(x_{t-1}))^T \omega_i \right\}$$

$$= \mathbf{b}_T - \mathbf{A}_T \omega_i$$

ここで、以下の式が成り立つ。

$$\mathbf{b}_T \equiv \frac{1}{T}\sum_{t=1}^{T} \phi(x_t) g_{\lambda,i}(x_{t-1})$$

$$\mathbf{A}_T \equiv \frac{1}{T}\sum_{t=1}^{T} \phi(x_t)(\phi(x_t) - z_\lambda(x_{t-1}))^T$$

$$\mathbf{E}_{M(\theta)}\{\phi(x)\} \cong \frac{1}{T+1}\sum_{t=0}^{T} \phi(x_t) \equiv \mathbf{c}_T$$

したがって、これらの予測子を式（１５）に代入すると、時間ステップＴにおける推定値 $\hat{\omega}_i^*$ は、以下のとおり計算される。

$$\mathbf{b}_T - \mathbf{A}_T \hat{\omega}_i^* + \mathbf{c}_t \mathbf{c}_T^T \hat{\omega}_i^* = 0$$

$$\Leftrightarrow \hat{\omega}_i^* = (\mathbf{A}_T - \mathbf{c}_t \mathbf{c}_T^T)^{-1} \mathbf{b}_T$$

【０１１０】

図３は、ＬＳＤＧ（λ）を求める手順をアルゴリズム１として示す図である。

【０１１１】

【数３９】

図３では、$\hat{\omega}_i^*$ よりも行列パラメータ $\Omega^*$ の場合についてのＬＳＤＧ（λ）をアルゴリズム１として示す。

【０１１２】
また、図４は、アルゴリズム１を示すフローチャートである。
図４を参照して、まず、ステップ１００において、処理の前提として、以下の設定がなされる。
【０１１３】
【数４０】

$$\text{方策を}\pi_\theta(x,u)\text{とする（パラメータ}\theta\text{は固定）}$$

$$\text{状態の特徴ベクトル関数を}\phi\text{（ｘ）とする}$$

【０１１４】
続いて、初期化処理として、以下の処理が行われる（ステップＳ１０２）。
【０１１５】
【数４１】

$$\text{初期化：}\lambda\in[0,\ 1),\ \mathbf{c}:=\mathbf{0};\mathbf{z}=\mathbf{0};\mathbf{g}=\mathbf{0};\mathbf{A}:=\mathbf{0};\mathbf{B}:=\mathbf{0}$$

【０１１６】
時間ステップｔがｔ＝０とされ（ステップＳ１０４）、以下の処理が、ｔ＝０からｔ＝Ｔ－１まで繰り返される（ステップＳ１０６～Ｓ１１６）。
【０１１７】
まず、ステップＳ１０６においてｔ＝０であれば、初期状態が観測され（ステップＳ１０８）、続いて、以下の設定が行われる（Ｓ１１０）。
【０１１８】
【数４２】

$$\mathbf{z}:=\phi(\mathbf{x}_0);\quad \mathbf{c}:=\phi(\mathbf{x}_0)$$

【０１１９】
一方、ステップＳ１０６において、ｔが０でなければ、以下の処理が行われる（ステップＳ１１２）。
【０１２０】
【数４３】

$$\mathbf{z}:=\lambda\mathbf{z}+(1-\lambda)\phi(\mathbf{x}_t)$$

【０１２１】
ステップＳ１１０またはＳ１１２に続いて、以下の計算が行われる（ステップＳ１１４）。
【０１２２】
【数４４】

$$\mathbf{c}:=\mathbf{c}+\phi(\mathbf{x}_{t+1});$$

$$\mathbf{g}:=\lambda\mathbf{g}+\nabla_\theta \ln \pi_\theta(x_t,u_t);$$

$$\mathbf{A}:=\mathbf{A}+\phi(\mathbf{x}_{t+1})(\phi(\mathbf{x}_{t+1})-\mathbf{z})^T;$$

$$\mathbf{B}:=\mathbf{B}+\phi(\mathbf{x}_{t+1})\mathbf{g}^T$$

【０１２３】
ステップＳ１１６にて、ｔがＴよりも小さければ処理はステップＳ１０６に復帰し、ｔがＴ以上であれば、処理はステップＳ１１８に移行して、以下の計算を行う。
【０１２４】

【数４５】

$$\Omega := (\mathbf{A} - \mathbf{c}\mathbf{c}^T / t)^{-1}\mathbf{B};$$

【０１２５】

続いて、以下の計算によりＬＳＤＧの推定値を得る。

【０１２６】

【数４６】

$$\hat{\nabla}_\theta \ln d^\pi(x) = \Omega\phi(\mathbf{x})$$

【０１２７】

（４．　ＬＳＤＧ推定による方策の更新）

ここでは、上述したＬＳＤＧ推定に基づくＰＧＲＬアルゴリズムを定義する。

【０１２８】

【数４７】

ＬＳＤＧ（λ）により$\nabla_\theta \ln d^\pi(x)$に対する推定を実現すると、ただちに、割引率γとは独立に、ＰＧに対する以下のような推定が導かれる。

$$\hat{\nabla}_\theta R(\theta) = \frac{1}{T}\sum_{t=0}^{T}(\nabla_\theta \ln \pi_\theta(x_{x_t}, u_{x_t}) + \hat{\nabla}_\theta \ln d^\pi(x_{x_t}))r_{t+1}$$

【０１２９】

方策パラメータは、適切なステップサイズαの確率的勾配法により更新される。

【０１３０】

【数４８】

$$\theta := \theta + \alpha(\nabla_\theta \ln \pi_\theta(x_{x_t}, u_{x_t}) + \hat{\nabla}_\theta \ln d^\pi(x_{x_t}))r_{t+1}$$

【０１３１】

ここで、：＝は、右辺を左辺に代入することを示す。

図５は、ＬＳＤＧ（λ）-ＰＧを、ＬＳＤＧ（λ）を利用した、ＰＧアルゴリズムについての最も簡単な実現法の１つとして、アルゴリズム２として示す図である。ここで、減衰率パラメータβ∈［０，１）は、古いθの値により与えられる過去の推定を捨てていくために導入されている。

【０１３２】

関数近似器によるＬＳＤＧ推定は他の重要な内容を構成する。すなわち、特に、連続状態問題において、基底関数φ（ｘ）をいかにして設定するか、ということである。

【０１３３】

【数４９】

ＰＧ推定に対しては、ＬＳＤＧに関する目的は、正確にＬＳＤＧを推定することでなく、単に、$\sum_{x,u} d^\pi(x)\pi_\theta(x,u)\nabla_\theta \ln d^\pi(x) r(x,u)$を近似することである。

【０１３４】

したがって、以下のような定理が有用である。

（定理３）

【０１３５】

【数５０】

ＬＳＤＧの推定器における基底関数を

$\phi(x)=\sum_u \pi_\theta(x,u)r(x,u)$とするとき、

関数推定器$f(x;\omega)=\omega\sum_u \pi_\theta(x,u)r(x,u)$は、

ＰＧ $\sum_{x,u} d^\pi(x)\pi_\theta(x,u)\nabla_\theta \ln d^\pi(x)r(x,u)$の第２項を表すことができる。

ここで、調整可能なパラメータ$\omega$は、$R^{d\times 1}$のベクトルである：

$$\sum_{x,u} d^\pi(x)\pi_\theta(x,u)r(x,u)\nabla_\theta \ln d^\pi(x)=\sum_{x,u} d^\pi(x)\pi_\theta(x,u)r(x,u)f(x;\omega^*)$$

ここで、$\omega^*$は、平均誤差$\varepsilon(\omega)=\frac{1}{2}\sum_x d^\pi(x)\left\{\nabla_\theta \ln d^\pi(x)-f(x;\omega)\right\}^2$を最小化するものである。

【０１３６】

（証明）

以下の式により証明される。

【０１３７】

【数５１】

$$\nabla_\omega \varepsilon(\omega^*)=\sum_{x,u} d^\pi(x)\pi_\theta(x,u)r(x,u)\left\{\nabla_\theta \ln d^\pi(x)-f(x;\omega^*)\right\}=0$$

【０１３８】

図６は、図５に対応するフローチャートである。

図６を参照して、まず、ステップ２００において、処理の前提として、以下の設定がなされる。

【０１３９】

【数５２】

方策を$\pi_\theta(x,u)$とする（パラメータθは調整可能）

状態の特徴ベクトル関数をφ（ｘ）とする

【０１４０】

続いて、初期化処理として、以下の処理が行われる（ステップＳ２０２）。

【０１４１】

【数５３】

初期化：$\lambda\in[0,1)$，$\beta\in[0,1)$，$\alpha$，

$$\mathbf{c}:=\mathbf{0};\ \mathbf{z}=\mathbf{0};\ \mathbf{g}=\mathbf{0};\ \mathbf{A}:=\mathbf{0};\ \mathbf{B}:=\mathbf{0}$$

【０１４２】

時間ステップｔがｔ＝０とされ（ステップＳ２０４）、以下の処理が、ｔ＝０からｔ＝

Ｔ－１まで繰り返される（ステップＳ２０６～Ｓ２１６）。

【０１４３】

まず、ステップＳ２０６においてｔ＝０であれば、初期状態が観測され（ステップＳ２０８）、続いて、以下の設定が行われる（Ｓ２１０）。

【０１４４】

【数５４】

$$\mathbf{z} := \phi(\mathbf{x}_0); \quad \mathbf{c} := \phi(\mathbf{x}_0)$$

【０１４５】

一方、ステップＳ２０６において、ｔが０でなければ、以下の処理が行われる（ステップＳ２１２）。

【０１４６】

【数５５】

$$\mathbf{z} := \lambda\mathbf{z} + (1-\lambda)\phi(\mathbf{x}_t)$$

【０１４７】

ステップＳ２１０またはＳ２１２に続いて、以下の計算が行われる（ステップＳ２１４）。

【０１４８】

【数５６】

$$\mathbf{c} := \beta\mathbf{c} + \phi(\mathbf{x}_{t+1});$$

$$\mathbf{g} := \beta\lambda\mathbf{g} + \nabla_\theta \ln \pi_\theta(x_t, u_t);$$

$$\mathbf{A} := \beta\mathbf{A} + \phi(\mathbf{x}_{t+1})(\phi(\mathbf{x}_{t+1}) - \mathbf{z})^T;$$

$$\mathbf{B} := \beta\mathbf{B} + \phi(\mathbf{x}_{t+1})\mathbf{g}^T;$$

$$\Omega := (\mathbf{A} - \mathbf{c}\mathbf{c}^T / \|\mathbf{c}\|)^{-1}\mathbf{B};$$

$$\theta := \theta + \alpha\mathbf{r}_{t+1}\{\nabla_\theta \ln \pi_\theta(x_t, u_t) + \Omega^T\phi(\mathbf{x}_t)\};$$

【０１４９】

ステップＳ１１６にて、ｔがＴよりも小さければ処理はステップＳ２０６に復帰し、ｔがＴ以上であれば、処理はステップＳ２２０に移行して、以下の計算を行うことで、方策がアップデートされる。

【０１５０】

【数５７】

$$p(u \mid x; \theta) = \pi_\theta(x_t, u_t)$$

【０１５１】

図７は、本発明の制御器の構成の概念図である。

本発明の制御器は、行動、すなわち、制御信号を制御対象に与える処理を行って、制御対象の状態量を観測器（たとえば、位置センサ、角度センサ、加速度センサ、角加速度センサなど）で観測し、この観測結果により「定常分布の対数の偏微分」（ＬＳＤＧ）を推定し、方策パラメータを更新し、これにより方策を更新する。そして、更新された方策により、さらに、制御対象が制御される。

（５．数値計算の結果）

有限のグリッドの組Ｘ＝｛１，…、ｍ｝と２つの可能な行動Ｕ＝｛Ｌ，Ｒ｝（左（Ｌ）または右（Ｒ）への１グリッド分の運動）を有している「１次元トーラス状グリッド空間」において、われわれが提案したアルゴリズムのパフォーマンスを検証した。これは、典

型的なｍ状態ＭＤＰタスクであり、状態の遷移確率は以下のように与えられる。

【０１５２】

【数５８】

$$p(x|x_{-1}=i,u_{-1}=L)=\begin{cases}p_i & \text{if } x=i-1\\ \frac{1-p_i}{2} & \text{if } x=i \text{ or } i+1\\ 0 & \text{otherwise,}\end{cases}$$

$$p(x|x_{-1}=i,u_{-1}=R)=\begin{cases}p_i & \text{if } x=i+1\\ \frac{1-p_i}{2} & \text{if } x=i \text{ or } i-1\\ 0 & \text{otherwise,}\end{cases}$$

【０１５３】

ここで、ｘ＝０およびｘ＝ｍ（ｘ＝１およびｘ＝ｍ＋１）とは、同じ状態であり、$p_i \in [0, 1]$（ｉ＝１，…，ｍ）は、タスクに依存する定数である。われわれの数値計算では、確率的方策は、以下のようなシグモイダル関数で表される：

【０１５４】

【数５９】

$$\pi_\theta(x,u=L)=1-\pi_\theta(x,u=R)=1/(1+\exp(\theta^T\phi(x)))$$

【０１５５】

ここで、状態特徴ベクトルφ（１），…，φ（ｍ）$\in R^m$の全ての要素は、定常正規分布Ｎ（０，$1^2$）からシミュレーションごとに独立に取り出された。これは、確率的方策のパラメータ化がいかにして、われわれのアルゴリズムのパフォーマンスに影響を与えるかを検証するためであった。状態特徴ベクトルφ（ｘ）は、ＬＳＤＧ推定のための基底関数としても使用された。各シミュレーションは、$10^5$タイムステップ以上からなる１つのエピソードを実行した。

【０１５６】

まず、われわれは、問題設定や方策パラメータθに関わりなく、ＬＳＤＧ（λ）がどれぐらい正確に以下の「定常分布の対数の偏微分」（ＬＳＤＧ）を推定しているのかを検証した。

【０１５７】

【数６０】

$$\nabla_\theta \ln d^\pi(x)$$

【０１５８】

この目的を達成するために、タスク依存の定数$p_1$，…，$p_m$は、区間［０．７，１］の均一な分布から独立に選ばれ、各シミュレーションでは固定された。方策パラメータθは、正規分布Ｎ（０，$0.5^2$）に従ってランダムに選択され、各シミュレーションでは固定された。

【０１５９】

図８は、ｍ＝３のときに、推定されたＬＳＤＧの典型的な時間経過を示す図である。ここでは、９の異なった色が、ＬＳＤＧの異なった要素を示している。実線は、ＬＳＤＧ（０）により推定された値を示し、点線は、ＬＳＤＧの解析的な解を示している。

【０１６０】

図８に示すように、ＬＳＤＧ（０）による推定は、ｍ＝３の場合は約１００回のタイムステップで解析的な解に収束している。

【０１６１】
７状態タスクを用いて、適格性減衰率λの影響を調べた。異なった設定についての平均的なパフォーマンスを評価するために、以下で定義される「相対誤差」基準を採用した。
【０１６２】
【数６１】

$$\mathbf{E}_{M(\theta)}\left\{\left(f(x;\Omega^*)-f(x;\Omega)\right)^2\right\}/\mathbf{E}_{M(\theta)}\left\{\left(f(x;\Omega^*)\right)^2\right\}$$

【０１６３】
ここで、Ω＊は定理３において定義された最適パラメータであり、解析的に計算された。図９および図１０は、λ＝０，０．３，０．９および１について、相対誤差の２００シミュレーションについての平均の時間推移を示している。これら２つの図の間の相違は、ただ、特徴ベクトルφ（ｘ）の要素の数だけである。
【０１６４】
図９において使用される特徴ベクトルφ（ｘ）∈Ｒ$^{7}$は、適切なものであり、異なった状態を区別するのに十分であった一方、図１０において使用される特徴ベクトルφ（ｘ）∈Ｒ$^{6}$は、適切でない。これらの結果は、理論的な予想と合致するものである。つまり、もしも基底関数が適切であれば（図９）、λ＝１以外の任意の値にλを設定できるのに対して、そうでないときは、λ＝１以外の大きな値にλを設定する必要がある（図１０）。
【０１６５】
最後に、ＬＳＤＧ（λ）－ＰＧを他の従来からのＰＧ法とを、３状態タスクにおいて比較した。ここで、状態遷移確率は、全てのｉ∈｛１，２，３｝についてｐ$_{i}$＝１に設定された。
【０１６６】
図１１は、このタスクにおける報酬の設定を示す図である。ここでは、２種類の報酬がある。定数“ｒ＝（±）２”と変数“ｒ＝（±）ｓ”である。変数ｓは、各シミュレーションにおいて、区間［０．８，１）での均一分布からランダムに設定された。報酬ｓは、最適方策を見いだすためのγの最小値を以下のように規定していることに注意されたい：
【０１６７】
【数６２】

$$\gamma^2+\gamma>2s/(2-s)$$

【０１６８】
したがって、γの設定は重要であり、このタスクにおいては困難である。従来のＰＧ法のパフォーマンスのベースラインとして、２つのアルゴリズムを採用した：ＧＰＯＭＤＰ［非特許文献１］とKondaのactor-critic法とである［非特許文献５］。
【０１６９】
図１２は、１００回のシミュレーションについての３つの方法の平均のパフォーマンスを示す図である。エラーバーは１００回についての標準偏差を示す。ここでは、パフォーマンスは、３Ｒ（θ）／（２－２ｓ）により評価された。すなわち、平均の報酬がその上限値である（２－２ｓ）／３により正規化されている。結果はＬＳＤＧ（λ）－ＰＧ法は他の方法よりもパフォーマンスが優れていることを示している。
【０１７０】
以上説明したように、本発明では、現実の順方向および逆方向のマルコフ連鎖は密接に関連しており、定理において共通の性質を有することを利用している。これらを用いて、定常分布の対数の偏微分（ＬＳＤＧ）を推定するアルゴリズムとしてＬＳＤＧ（λ）を、ＬＳＤＧ推定を用いたＰＧアルゴリズムとしてＬＳＤＧ（λ）－ＰＧを提案した。実験結果はＬＳＤＧ（λ）は、適格性減衰率λ∈［０，１）で動作することができ、かつ、ＬＳＤＧ（λ）－ＰＧは、割引率γとは独立に学習をすることができる。
【０１７１】

今回開示された実施の形態はすべての点で例示であって制限的なものではないと考えられるべきである。本発明の範囲は上記した説明ではなくて特許請求の範囲によって示され、特許請求の範囲と均等の意味および範囲内でのすべての変更が含まれることが意図される。

【図面の簡単な説明】

【０１７２】

【図１】本発明の制御方法および制御プログラムが適用される制御器を用いたシステム１０００の一例を示す概念図である。

【図２】コンピュータ１００の構成をブロック図形式で示す図である。

【図３】ＬＳＤＧ（λ）を求める手順をアルゴリズム１として示す図である。

【図４】アルゴリズム１を示すフローチャートである。

【図５】ＬＳＤＧ（λ）-ＰＧを示す図である。

【図６】図５に対応するフローチャートである。

【図７】本発明の制御器の構成の概念図である。

【図８】ｍ＝３のときに、推定されたＬＳＤＧの典型的な時間経過を示す図である。

【図９】ｍ＝７のときに、十分な特徴ベクトルを用いて推定されたＬＳＤＧの相対誤差の時間経過を示す図である。

【図１０】ｍ＝７のときに、不十分な特徴ベクトルを用いて推定されたＬＳＤＧの相対誤差の時間経過を示す図である。

【図１１】タスクにおける報酬の設定を示す図である。

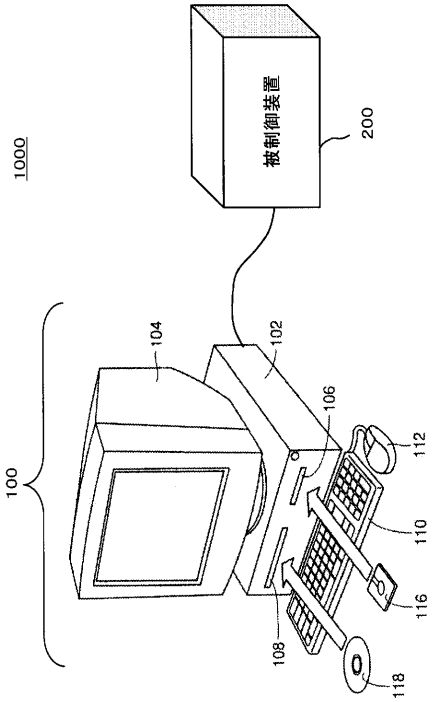
【図１２】１００回のシミュレーションについての３つの方法の平均のパフォーマンスを示す図である。

【符号の説明】

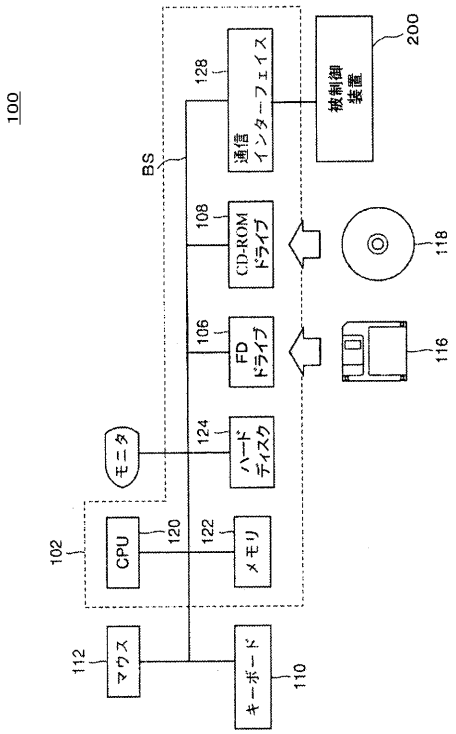
【０１７３】

１００　コンピュータ、１０２　コンピュータ本体、１０４　ディスプレイ、１０６　ＦＤドライブ、１０８　ＣＤ－ＲＯＭドライブ、１１０　キーボード、１１２　マウス、１１８　ＣＤ－ＲＯＭ、１２０　ＣＰＵ、１２２　メモリ、１２４　ハードディスク、１２８　通信インタフェース、２００　被制御装置、１０００　システム。

【図１】




【図２】




【図３】

アルゴリズム1

LSDG( $\lambda$ ): $\nabla_\theta \ln d^\pi(x)$ に対する推定

所与の条件

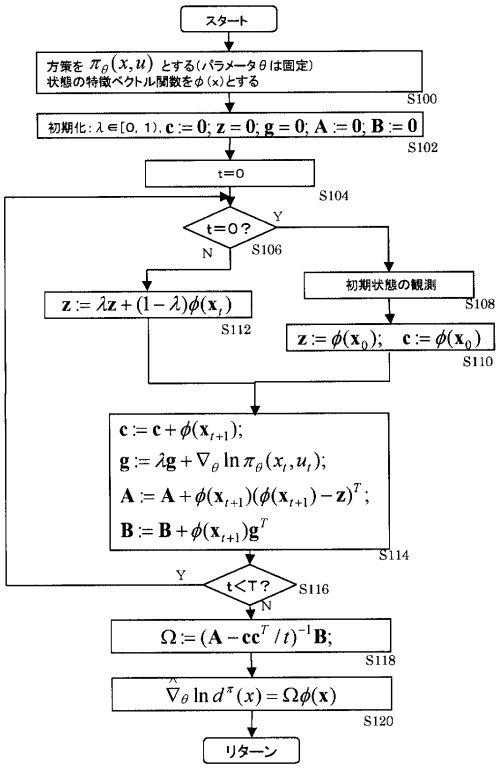
- a policy $\pi_\theta(x,u)$ with a fixed $\theta$.
- a feature vector function of state $\phi(x)$.

初期化 ：$\lambda \in [0,1)$.

Set: $c := 0$; $z = 0$; $g := 0$; $A := 0$; $B := 0$.
for $t = 0$ to $T-1$ do
if $t = 0$ then
$z := \phi(x_0)$; $c := \phi(x_0)$;
else
$z := \lambda z + (1-\lambda)\phi(x_t)$;
end if
$c := c + \phi(x_{t+1})$;
$g := \lambda g + \nabla_\theta \ln \pi_\theta(x_t, u_t)$;
$A := A + \phi(x_{t+1})(\phi(x_{t-1}) - z)^\top$;
$B := B + \phi(x_{t+1}) g^\top$;
end for
$\Omega := (A - cc^\top / t)^{-1} B$;
Return: $\widehat{\nabla}_\theta \ln d^\pi(x) = \Omega\, \phi(x)$.

【図４】

アルゴリズム1



【図５】

アルゴリズム2

LSDG-PG(λ): 方策に対する最適化

所与の条件
- a policy $\pi_\theta(x_t, u_t)$ with an adjustable $\boldsymbol{\theta}$,
- a feature vector function of state $\boldsymbol{\phi}(x)$.

初期化 ： $\boldsymbol{\theta}$, $\lambda \in [0,1)$, $\beta \in [0,1)$, $\alpha$.

**Set:** $\boldsymbol{c} := \mathbf{0}$; $z = \mathbf{0}$; $\boldsymbol{g} := \mathbf{0}$; $\boldsymbol{A} := \mathbf{0}$; $\boldsymbol{B} := \mathbf{0}$.

**for** $t = 0$ **to** $T-1$ **do**

**if** $t = 0$ **then**

$z := \phi(x_0)$; $\boldsymbol{c} := \boldsymbol{\phi}(x_0)$;

**else**

$z := \lambda z + (1-\lambda)\boldsymbol{\phi}(x_t)$;

**end if**

$\boldsymbol{c} := \beta \boldsymbol{c} + \boldsymbol{\phi}(x_{t+1})$;

$\boldsymbol{g} := \beta\lambda \boldsymbol{g} + \nabla_{\boldsymbol{\theta}} \ln \pi_\theta(x_t, u_t)$;

$\boldsymbol{A} := \beta \boldsymbol{A} + \boldsymbol{\phi}(x_{t+1})(\boldsymbol{\phi}(x_{t+1}) - z)^\top$;

$\boldsymbol{B} := \beta \boldsymbol{B} + \boldsymbol{\phi}(x_{t+1})]\boldsymbol{g}^\top$;

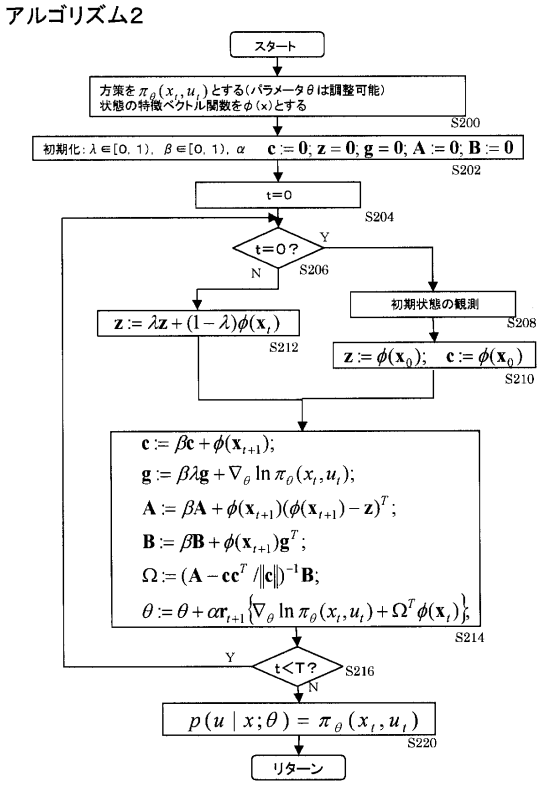
$\boldsymbol{\Omega} := (\boldsymbol{A} - \boldsymbol{c}\boldsymbol{c}^\top / \|\boldsymbol{c}\|)^{-1}\boldsymbol{B}$;

$\boldsymbol{\theta} := \boldsymbol{\theta} + \alpha r_{t+1}\{\nabla_{\boldsymbol{\theta}} \ln \pi_\theta(x_t, u_t) + \boldsymbol{\Omega}^\top \boldsymbol{\phi}(x_t)\}$;

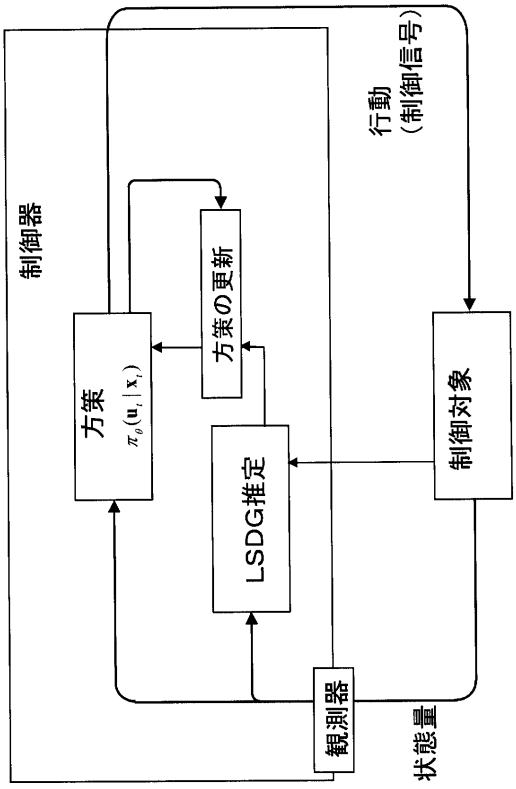
**end for**

**Return:** $p(u|x;\theta) = \pi_\theta(x, u)$.

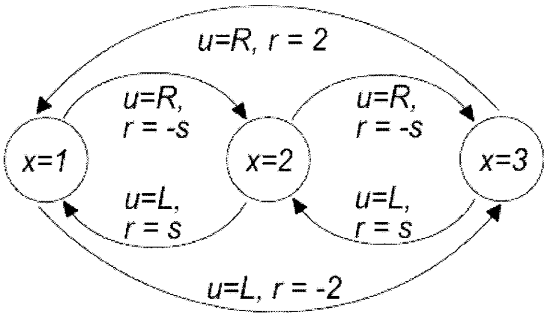
【図６】




【図７】

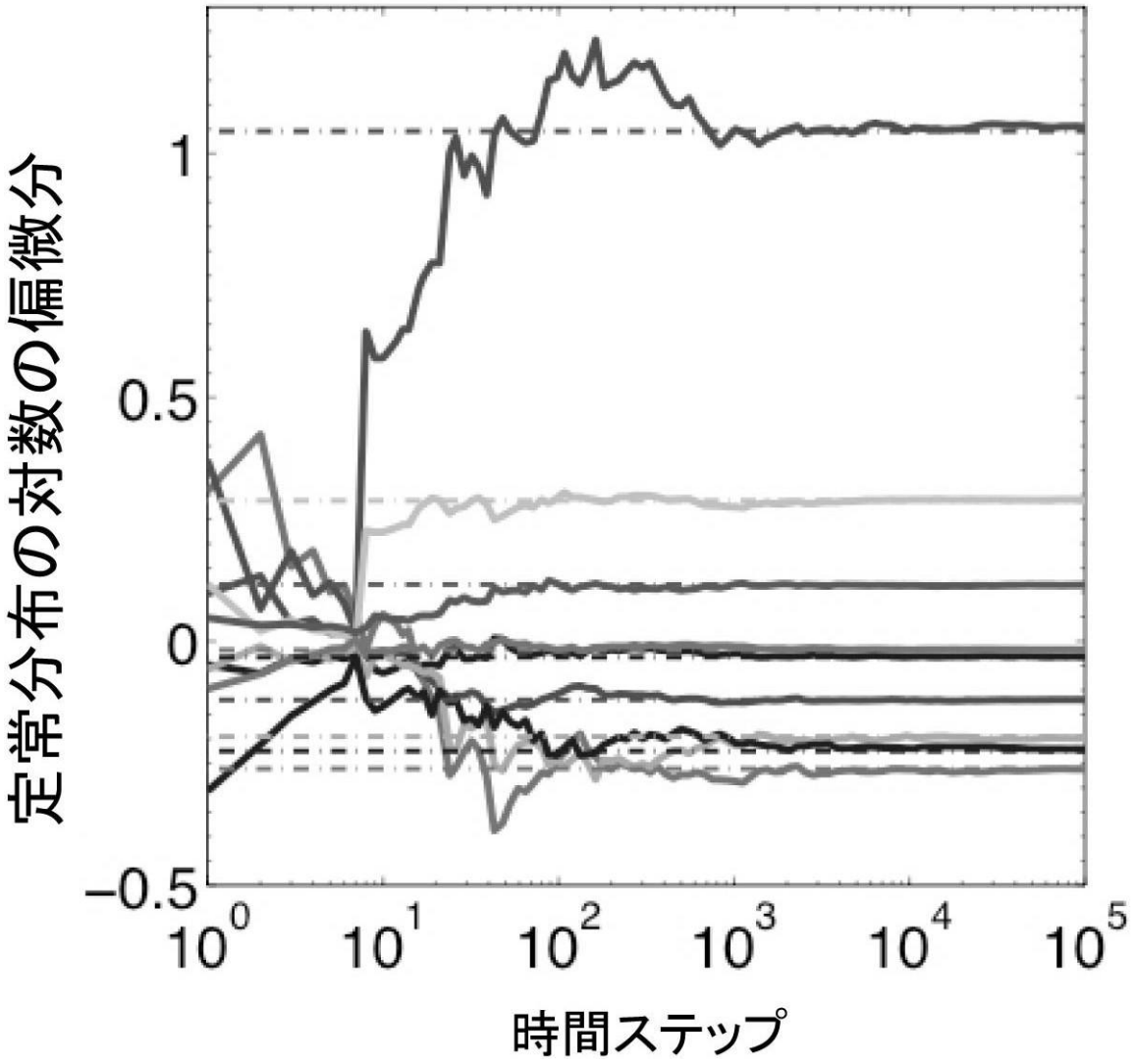



【図１１】

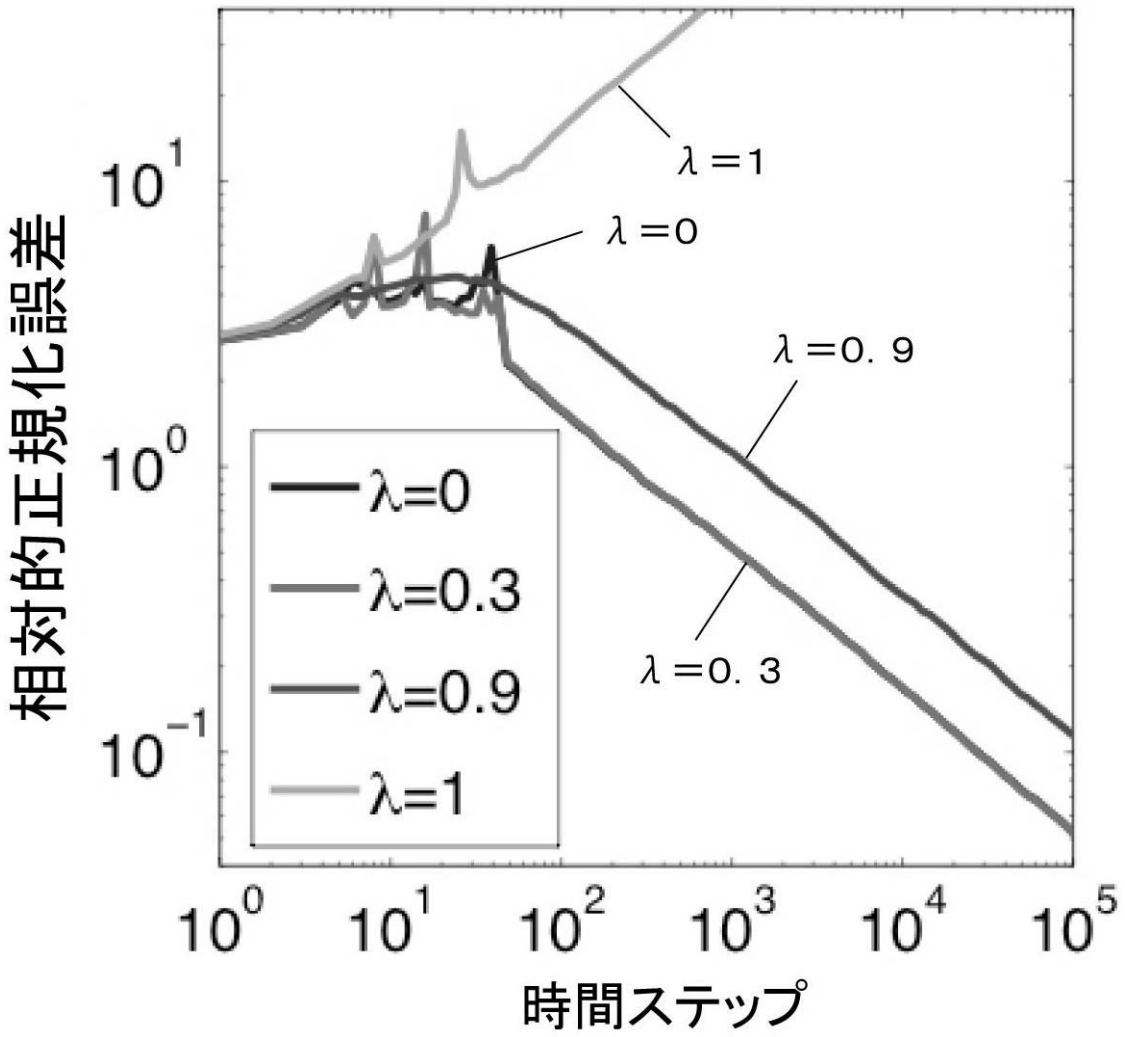



3状態マルコフ決定過程問題

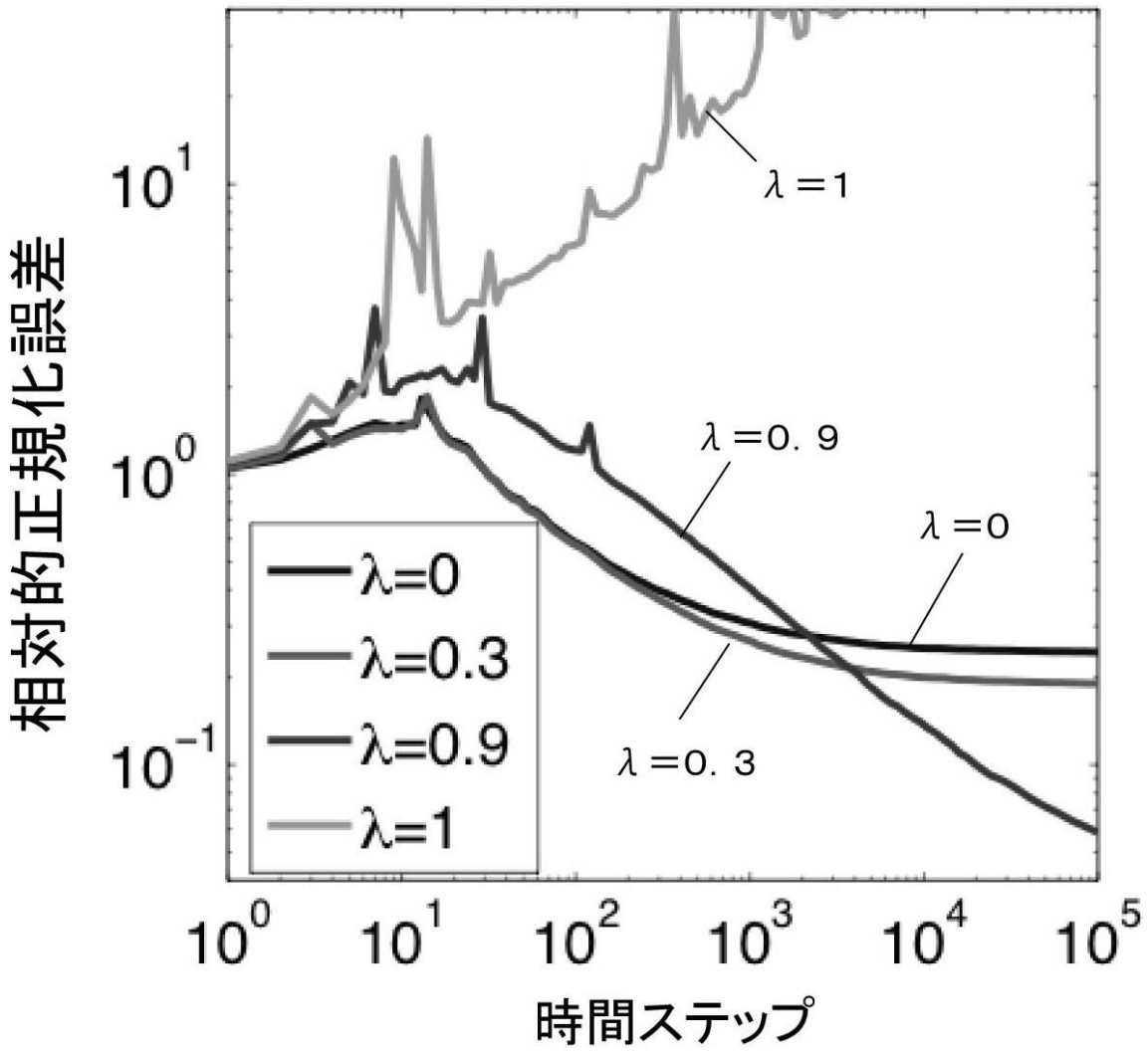
【図８】



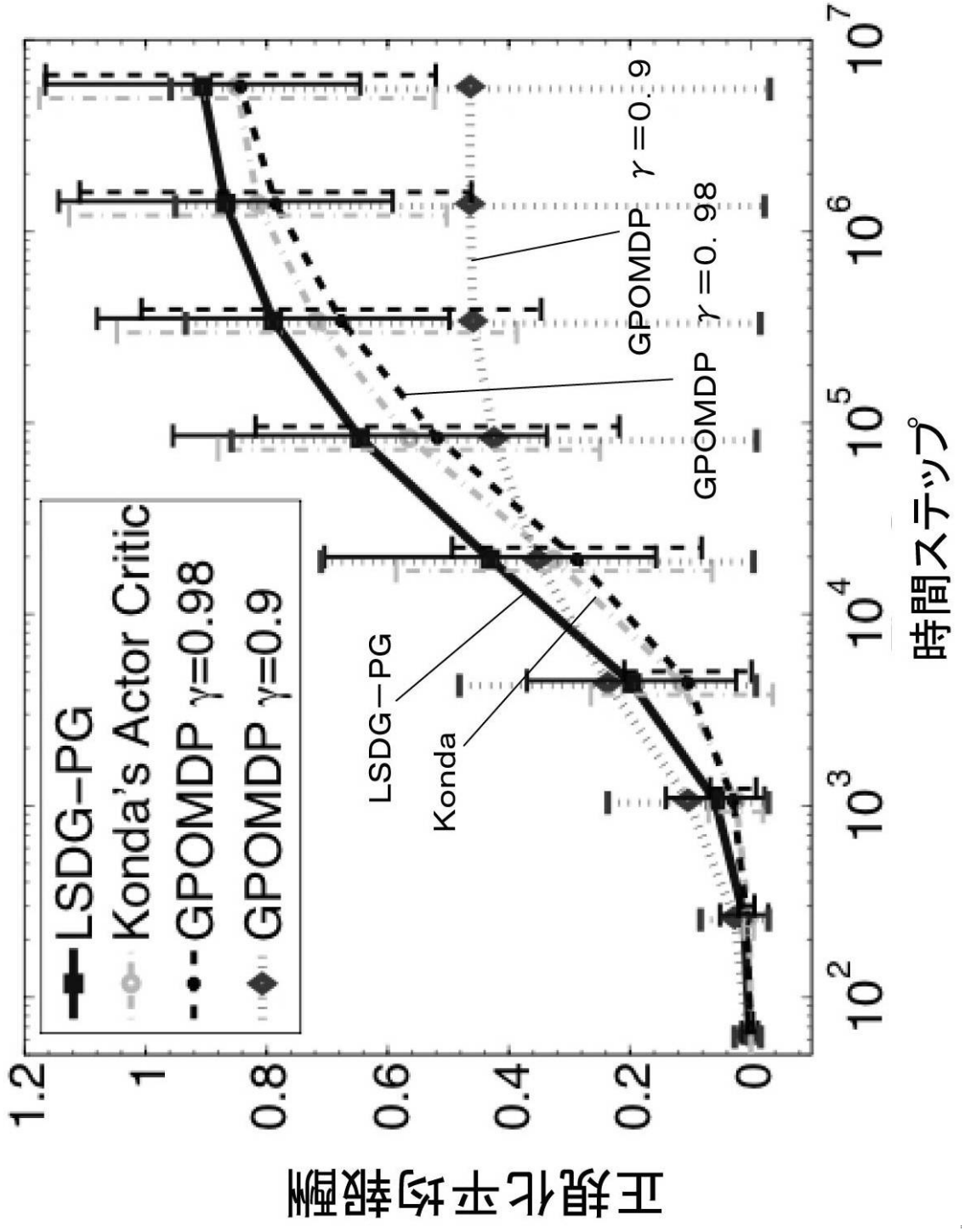
【図９】


相対的正規化誤差
$10^{1}$
$10^{0}$
$10^{-1}$
λ＝1
λ＝0
λ＝0．９
λ＝0．３
λ=0
λ=0.3
λ=0.9
λ=1
$10^{0}$
$10^{1}$
$10^{2}$
$10^{3}$
$10^{4}$
$10^{5}$
時間ステップ

【図１０】


相対的正規化誤差
$10^1$
$10^0$
$10^{-1}$
λ＝1
λ＝0．９
λ＝0
λ＝0．３
λ=0
λ=0.3
λ=0.9
λ=1
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
$10^5$
時間ステップ

【図１２】


LSDG−PG
Konda's Actor Critic
GPOMDP γ=0.98
GPOMDP γ=0.9
LSDG−PG
Konda
GPOMDP γ=0.98
GPOMDP γ=0.9
正規化平均報酬
時間ステップ
0
0.2
0.4
0.6
0.8
1
1.2
10^2
10^3
10^4
10^5
10^6
10^7

フロントページの続き

(74)代理人　100109162
　　弁理士　酒井　將行
(74)代理人　100111246
　　弁理士　荒川　伸夫
(72)発明者　森村　哲郎
　　沖縄県国頭郡恩納村字恩納７５４２番地　独立行政法人沖縄科学技術研究基盤整備機構内
(72)発明者　内部　英治
　　沖縄県国頭郡恩納村字恩納７５４２番地　独立行政法人沖縄科学技術研究基盤整備機構内
(72)発明者　吉本　潤一郎
　　沖縄県国頭郡恩納村字恩納７５４２番地　独立行政法人沖縄科学技術研究基盤整備機構内
(72)発明者　銅谷　賢治
　　沖縄県国頭郡恩納村字恩納７５４２番地　独立行政法人沖縄科学技術研究基盤整備機構内

審査官　柿崎　拓

(56)参考文献　特開２００７－０６５９２９（ＪＰ，Ａ）
　　特開２００５－０８４８３４（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
　　Ｇ０５Ｂ　　１１／００－１３／０４