「深化する機械学習—技術の進展とその応用」総合特集号

論　文

プラント変動の推定に基づく近似ベイジアン強化学習とペグ・イン・ホール・タスクへの適用*

泉田　啓†・菱沼　徹†・谷　百合夏†

Approximation Bayesian Reinforcement Learning based on Estimation of Plant Variation and its Application to Peg-in-Hole Task*

Kei Senda†, Toru Hishinuma† and Yurika Tani†

In a general reinforcement learning problem, a plant, i.e. state transition probabilities, is estimated, and a learning policy for the estimated plant is applied to a real plant. If there is a difference between the estimated plant and the real plant, the obtained policy may not work well for the real plant. In this study, the real plant variation is parameterized by an interpolation of several estimated plants. This study proposes a reinforcement learning method based on estimation of parameter variation, and applies this method to 2-dimensional Peg-in-Hole Task. The effectiveness of the proposed method is demonstrated by numerical and experimental results.

1. 導入

構造物の組み立てなど，環境との力学的相互作用を伴う作業は，摩擦や計測精度などの不確定性・変動の影響を強く受ける．事前に十分正確なモデルを構築できない場合，ロボットは実環境上で適切な動作を自律的に獲得する必要がある．

ロボット自律化へのアプローチとして，強化学習が挙げられる．強化学習は，ロボットが実環境上で試行錯誤を行い，この結果に応じて行動則を学習する枠組みである [1]．ほとんどの学習アルゴリズムは，試行錯誤によるサンプリング過程から成るオンライン同定，および，同定モデルに対する方策（制御）を見いだす解法という，二つの動作から成る [2]．原理的には，環境がまったく未知である場合であっても，この方法は適用可能である．

しかし，現実のロボットに対して適用する場合には，いくつかの実際的な課題が存在する [3]．Q 学習などの従来のモデルフリー手法では，状態や行動の数が多くなった場合，状態遷移確率を推定するための現実のロボットを用いたサンプリングが膨大となり，学習には時間がかかり過ぎてしまう（実サンプルの課題）．これに対して，先見情報をもとにして作成したモデルを用いるモデルベース手法では，サンプリングを低減化することができる．しかし，モデル化における不確定性・変動の影響により，学習結果が現実では有効でない可能性が生じる（モデリングとモデル不確定性の課題）．

これらの課題を解決する方法の一つとして，モデルベースのベイジアン強化学習が考えられる．ベイジアン強化学習の枠組みでは，|状態数|$^2\times$|行動数| の状態遷移確率に比べ少数のモデルは時不変パラメータによって特徴づけられていて，このパラメータのみが未知である場合を扱う．各パラメータに対応するモデルは，先見情報に基づいてあらかじめ作成する．事前に予想されるモデル不確定性は，パラメータ形式で表現されるため，モデル不確定性へのアプローチとなる．また，少数のパラメータを推定するため，現実のロボットを用いたサンプリングを減らすことが可能で，実サンプルの課題へのアプローチとなる．しかし，ベイジアン強化学習の計算コストは非常に高いため，ロボットなどの大規模な問題に対して適用する例は少ない．

本研究では，いくつかの近似を導入したベイジアン強化学習を提案する．計算コストを低減化するため，離散有限個の未知パラメータに対してのみ，先見情報に基づきモデルを作成する．その他の連続的に存在する未知パラメータに対しては，有限個のモデルの内挿により近似表現する．内挿により生成された近似モデルに対する学習結果は，ある範囲のモデル集合に対してロバストに機

* 原稿受付　2015 年 7 月 6 日
第 59 回システム制御情報学会研究発表講演会にて発表 (2015 年 5 月)

† 京都大学　大学院　工学研究科　Graduate School of Engineering, Kyoto University; Kyotodaigaku-Katsura, Nishikyo-ku, Kyoto City, Kyoto 615-8510, JAPAN

Key Words: reinforcement learning, plant variation.

能する．この性質より，モデル不確定性と計算コスト低減のための近似誤差に対して，提案手法は有効にはたらく．提案手法を実際的なロボットタスクに適用することにより，現実のモデル不確定性に対する有効性と，実用的な時間内で学習が可能であることを示す．

2. 強化学習問題

現実のロボットを用いる場合，問題は本質的に連続時間・連続状態行動空間を伴う．プラント（ロボットと環境）の挙動が非線形であるとき，一般には最適方策が解析的には得られない．これに対してよくとられるアプローチは，価値関数近似 [2] や方策のパラメータ表現 [4] などに基づき，同定モデルの未知変数や方策の設計変数を有限個に低減化することである．

本研究では，はじめタスクの達成方法が未知である状況での学習を考える．ここでは，その中でも最も単純な方法として，状態空間を格子状に分割して扱う離散状態近似を用いて議論する．これは，関数近似法などを用いると，学習結果がその方法に強く影響を受け，学習アルゴリズムの良し悪しを判断しにくくなるためである．

2.1 標準的な強化学習

離散時間の動的システムを考える [2]．状態 s_i および行動 u_k は有限集合 $\mathcal{S}$ と $\mathcal{U}$ の元であり，離散的な変数とする．$\mathcal{S}$ の元として $s_1,s_2,\cdots,s_{N_s}$ の N_s 個と終端状態 s_0 を考える．$\mathcal{U}$ の元として $u_1,u_2,\cdots,u_K$ の K 個を考える．

時刻 t において状態が $s_i \in \mathcal{S}$ であるとき，行動 $u_k \in \mathcal{U}$ を選択すると，時刻 $t+1$ において確率的に状態 $s_j \in \mathcal{S}$ へと遷移する．このとき，1 ステップコスト $g(s_i,u_k,s_j)$ を生じる．この確率的状態遷移は，現在の時刻 t における状態と行動のみに依存し，それ以前の状態と行動の履歴に陽には依存しないとする．このようなプラントの挙動は，有限マルコフ決定過程（MDP）とよばれる．状態 s_i で行動 u_k を選択したときに状態 s_j へ遷移する確率を，状態遷移確率 $p_{ij}(u_k)$ によって表す．また，状態遷移確率は時刻に陽に依存しないとする．終端状態 s_0 においては，すべての行動 u_k に対してコストが発生せず，$p_{00}(u_k)=1$, $g(s_0,u_k,s_0)=0$ とする．

各時刻 t において，ロボットは何らかの方策に従って行動選択を行う．ここでは，各時刻における状態のみに基づいて行動選択を行うような方策を考える．とくに，$\mu(s_i)=u_k$ のように，時刻に陽に依存せず，かつ確定的に行動を選択する方策を，静的かつ確定的方策とよぶ．

本研究では，割引なしかつ期間不定でコストが蓄積する infinite horizon 問題を扱う．初期時刻 t_0 で初期状態 s_i から方策 μ を用いたときの全コストの期待値（J 値）は次式で定義される．

$$J^{\mu}(s_i)=E\left[\sum_{t=t_0}^{\infty} g(s^t,u^t,s^{t+1}) \,\middle|\, s^{t_0}=s_i,\mu\right] \qquad (1)$$

ただし，$E[\cdot]$ は期待値を表す．最適 J 値は，$J^*(s_i)=\min_{\mu} J^{\mu}(s_i)$ で定義される．すべての状態に対して $J^{\mu}(s_i)=J^*(s_i)$ なら，方策 μ は最適である．最適方策 μ^* を求めることが強化学習問題の目的である．この問題では，静的かつ確定的方策が，最適方策として得られることが知られている [2]．

$J^{\mu}(s_i)<\infty, \forall i$ であるような静的方策を，プロパー方策とよぶ．また，プロパーでない方策はインプロパー方策とよばれる．インプロパー方策を用いた場合，有限ステップで終端状態 s_0 に到達できないような状態 s_i が少なくとも一つ存在する．

モデルフリー強化学習では，状態遷移確率 $p_{ij}(u_k)$ が未知である場合を考える．実プラント上での試行錯誤の結果に基づいて $p_{ij}(u_k)$ を推定して，価値反復法などの動的計画法を適用することで，最適方策を得ることができる [2]．Q 学習などのアルゴリズムは，$p_{ij}(u_k)$ を陽に推定して保持しないが，価値反復法を確率的に近似した解法であるため，本質的な動作としては同じである．これらの学習方法が収束する速度は，最終的には状態遷移確率 $p_{ij}(u_k)$ を推定する速度に規定される [5]．実際のロボットに対する適用を考える場合，$p_{ij}(u_k)$ を推定するための状態遷移回数が膨大となってしまうため，現実的な時間内で学習が収束しない（実サンプルの課題）．

モデルベース強化学習では，状態遷移確率 $p_{ij}(u_k)$ が既知，あるいは数値シミュレーション上で再現できるような場合を考える．シミュレーション上では試行錯誤過程を高速化できるため，モデルベース強化学習はロボットなどの大規模な問題に対しても適用可能である．しかし，シミュレーション上で得られた学習結果を現実のロボットに対して適用する場合，状態遷移確率 $p_{ij}(u_k)$ が十分正確に得られていなければ学習結果が有効に機能しない（モデリングとモデル不確定性の課題）．

2.2 ベイジアン強化学習

標準的な有限 MDP の強化学習問題では，未知パラメータ $p_{ij}(u_k)$ の個数は N_s^2K であり，これらが完全に未知であると考える．一方でロボットなどの物理系においては，プラントの挙動は連続時間・空間の支配方程式に従って得られると考えられるため，$p_{ij}(u_k)$ が完全に未知であるとまではいえない．ベイジアン強化学習の枠組みでは，プラントの状態遷移確率が時不変パラメータ $\boldsymbol{\gamma}\in\boldsymbol{R}^N$ によって特徴付けられ，はじめロボットにとってパラメータ $\boldsymbol{\gamma}$ のみが未知であるような場合を考える．

パラメータが $\boldsymbol{\gamma}$ であるようなプラントを，$\eta(\boldsymbol{\gamma})$ によって表す．プラント $\eta(\boldsymbol{\gamma})$ において，状態 s_i で行動 u_k を選択したときに状態 s_j へ遷移する確率を $p_{ij}(u_k;\eta(\boldsymbol{\gamma}))$ によって表す．また，このときに生じる 1 ステップコストを $g(s_i,u_k,s_j;\eta(\boldsymbol{\gamma}))$ によって表す．

ロボットは，パラメータ $\boldsymbol{\gamma}$ を直接観測することはできないが，先見情報として $\boldsymbol{\gamma}$ の事前確率分布 $f^{t_0}(\cdot)$ が与え

られる．時刻 $t \to t+1$ において遷移 (s_i,u_k,s_j) が生起するとき，観測できない $\boldsymbol{\gamma}$ の事後確率分布は，ベイズの法則よりつぎのように得られる．

$$f^{t+1}(\boldsymbol{\gamma})|_{(s_i,u_k,s_j)} = \frac{p_{ij}(u_k;\eta(\boldsymbol{\gamma}))\ f^t(\boldsymbol{\gamma})}{\int p_{ij}(u_k;\eta(\boldsymbol{\gamma}))\ f^t(\boldsymbol{\gamma})\ d\boldsymbol{\gamma}} \quad (2)$$

初期時刻 t_0 で，事前分布 f^{t_0} が与えられていて，初期状態 s_i から方策 μ を用いたときの全コストの期待値（J 値）は次式で定義される．

$$J^{\mu}(s_i,f^{t_0}) = E\left[\sum_{t=t_0}^{\infty} g(s^t,u^t,s^{t+1};\eta) \,\middle|\, s^{t_0}=s_i,f^{t_0},\mu\right] \quad (3)$$

標準的な強化学習問題と同様に，最適 J 値は $J^*(s_i) = \min_{\mu} J^{\mu}(s_i)$ で，最適方策はこれを満たす μ として定義される．ベイジアン強化学習では，観測できる s_i と観測できない $\boldsymbol{\gamma}$ の事後確率分布 $f^t(\cdot)$ に基づいて行動選択を行う方策が，最適方策として得られることが知られている．

ベイジアン強化学習では，少ない未知パラメータに対して構造が与えられるため，少ない実プラント状態遷移回数によって推定を行うことができる．得られる最適解は，実プラントを推定する動作とタスクを達成する動作の間のトレードオフを最適化している．プラントの未知パラメータを適切に設定できれば，現実的な時間内でプラント推定とタスク達成を実現できる．しかし，以下に述べるような課題が存在するため，ロボットの問題に対する適用はむずかしい．

ベイジアン強化学習の厳密解を直接求める計算量は，問題の horizon に対して指数的に増加する．計算コストの観点から現実的ではないため，近似的に解く方法が一般的である．

ベイジアン強化学習では，パラメータ $\boldsymbol{\gamma}$ に対応する状態遷移確率 $p_{ij}(u_k;\eta(\boldsymbol{\gamma}))$ を，先見情報として事前に与えなければならない．ロボットのような連続系の問題に対して適用する場合，支配方程式モデルに基づき時間・状態空間についての離散化を行うなどして，支配方程式中のパラメータ $\boldsymbol{\gamma}$ に対応する状態遷移確率 $p_{ij}(u_k;\eta(\boldsymbol{\gamma}))$ を得る必要がある．このとき，一般には $\boldsymbol{\gamma}$ に対して $p_{ij}(u_k;\eta(\boldsymbol{\gamma}))$ が解析的に与えられないため，シミュレーションなどを通じて数値的に得なければならない．この先見モデルが適切に与えられていない場合，プラントの推定あるいはタスク達成がうまくいかない可能性がある（モデリングとモデル不確定性の課題）．また，連続的に存在するパラメータ $\boldsymbol{\gamma}$ のすべてに対して数値的に状態遷移確率を保持することは計算コストの面からも困難である．この理由から，ロボットの問題に対してベイジアン強化学習を適用する際には，事前に与える先見情報の近似誤差が避けられない．

3. 提案手法

本研究では，ベイジアン強化学習をロボットの問題へ適用可能にするためのアプローチを提案する．上述の課題を踏まえると，先見情報の近似誤差をある程度許容しつつ現実的な計算コストで実行可能な，問題設定と近似解法が必要である．有限個のプラントの内挿による近似ベイジアン強化学習問題の設定を **3.1** 節に，実プラント上でのパラメータ点推定と逐次学習に基づく近似解法を **3.2** 節に示す．

3.1 近似ベイジアン強化学習問題の設定

本研究では，有限個のプラントを用いた内挿により，連続的に存在する $\boldsymbol{\gamma}$ に対応するプラント $\eta(\boldsymbol{\gamma})$ を近似表現する．まず，変動パラメータの有限集合 $\{\boldsymbol{\gamma}_0,\cdots,\boldsymbol{\gamma}_{N-1}\}$ を作り，この元に対応するプラントの有限集合 $\mathcal{H} = \{\eta_0,\cdots\eta_{N-1}\}$ を生成する（$\eta_n = \eta(\boldsymbol{\gamma}_n)$ とした）．プラントの有限集合 $\mathcal{H}$ に対する内挿プラント $\eta_{itp(\boldsymbol{\alpha})}$ を，つぎの状態遷移確率と 1 ステップコストをもつとして定義する．

$$p_{ij}(u_k;\eta_{itp(\boldsymbol{\alpha})}) \equiv \sum_{n=0}^{N-1} \alpha_n p_{ij}(u_k;\eta_n) \quad (4)$$

$$g(s_i,u_k,s_j;\eta_{itp(\boldsymbol{\alpha})}) \equiv \sum_{n=0}^{N-1} \alpha_n g(s_i,u_k,s_j;\eta_n) \quad (5)$$

$$0 \le \alpha_n \le 1,\ \sum_{n=0}^{N-1} \alpha_n = 1 \quad (6)$$

ここで，$\boldsymbol{\alpha} = [\alpha_0,\cdots,\alpha_{N-1}]$ は時不変の内挿重みであり，内挿プラントは $\boldsymbol{\alpha}$ によって特徴づけられる．$p_{ij}(u_k;\eta_{itp(\boldsymbol{\alpha})})$ が確率測度の定義を満たすために，拘束 (6) をおいている．任意の内挿プラントを元にもつ無限集合を $\mathcal{H}_{itp}$ によって表す．この近似ベイジアン強化学習問題において，ロボットにとって未知なパラメータは内挿の重み $\boldsymbol{\alpha}$ である．観測できない $\boldsymbol{\alpha}$ の事前分布を $\bar{f}^{t_0}(\boldsymbol{\alpha})$ によって表す．

内挿プラントの集合 $\mathcal{H}_{itp}$ において，初期時刻 t_0 で，事前分布 $\bar{f}^{t_0}$ が与えられていて，初期状態 s_i から方策 μ を用いたときの全コストの期待値を次式で定義する．

$$J^{\mu}(s_i,\bar{f}^{t_0}) = E\left[\sum_{t=t_0}^{\infty} g(s^t,u^t,s^{t+1};\eta_{itp(\boldsymbol{\alpha})}) \,\middle|\, s^{t_0}=s_i,\bar{f}^{t_0},\mu\right] \quad (7)$$

付録に示す定理より，ある内挿プラント $\eta_{itp(\boldsymbol{\alpha})}$ に対してプロパーな方策は，別のプラント $\eta_{itp(\boldsymbol{\alpha}')}$ に対してもプロパーである．このように内挿プラントに対する学習自体結果にある程度のロバスト性が付加されるため，先見情報の近似誤差に対して内挿によるプラント表現は有効なアプローチであると期待できる．

3.2 近似解法

一般に，ベイジアン強化学習問題の厳密解を得ることは計算的に困難であり，近似解法が必要となる．本研究

では，Bayesian DP[6] をもとに近似解法を考える．

Bayesian DP では，パラメータ $\boldsymbol{\alpha}$ を事前分布に基づいて一つサンプルし，プラント $\eta_{itp(\boldsymbol{\alpha})}$ の最適方策を動的計画法によって計算する．その方策で数ステップ行動したら，その経験から得られた事後分布を使って新たな $\boldsymbol{\alpha}$ を抽出し，以降このプロセス全体を繰り返す．このように，「パラメータの適応的サンプリング」と「サンプルに対する逐次学習」の二つの動作を組み合わせたアルゴリズムにより，もとのベイジアン強化学習を近似的に解く．

本研究では，現実のロボット問題への拡張を考える．有限個のプラントの内挿による近似表現によりベイジアン強化学習の先見情報を与えているが，実プラントのモデル化誤差のためにベイズ推定が失敗する可能性がある．そのため，経験から得られる事後分布からパラメータ $\boldsymbol{\alpha}$ を一つ抽出する動作を，以下で述べるように置き換える．

まず，事後分布からパラメータ点推定量を得る方法として代表的である，MAP（Maximum a posteriori）推定 [7] について述べる．時刻 t における事後分布を $\bar{f}^t(\cdot)$，t までに状態行動対 (s_i,u_k) が選択された回数を z_{ik}，遷移 (s_i,u_k,s_j) が観測された回数を z_{ikj}，状態遷移確率の推定 $\tilde{p}_{ij}(u_k)=z_{ikj}/z_{ik}$ として，次式のように書ける．

$$
\begin{aligned}
&\tilde{\boldsymbol{\alpha}}_{MAP}\\
&=\arg\max_{\alpha}\log\bar{f}^t(\boldsymbol{\alpha})\\
&=\arg\min_{\alpha}\left\{\sum_{i,k}\left(z_{ik}\sum_j\tilde{p}_{ij}(u_k)\log\frac{\tilde{p}_{ij}(u_k)}{p_{ij}(u_k;\eta_{itp(\boldsymbol{\alpha})})}\right)\right.\\
&\qquad\left.-\log\bar{f}^{t_0}(\boldsymbol{\alpha})\right\}
\end{aligned}
$$

上式では，遷移結果から推定される遷移確率 $\tilde{p}_{ij}(u_k)$ に対する内挿プラントの遷移確率 $p_{ij}(u_k;\eta_{itp(\boldsymbol{\alpha})})$ のカルバック情報量を，状態行動対 (s_i,u_k) が選ばれた回数 z_{ik} で重み付けして，最小化する $\boldsymbol{\alpha}$ を求めている．

しかし，実プラントとのモデル化誤差のために MAP 推定が失敗する場合がある．プラントの有限集合 $\mathcal{H}=\{\eta_0,\cdots\eta_{N-1}\}$ のすべての元において生じない遷移 (s_i,u_k,s_j) が実プラント上で観測された場合，すべての $\boldsymbol{\alpha}$ に対してカルバック情報量が無限大となり，推定が発散する．

このように，ベイズ推定は先見情報の誤差に対するロバスト性が低いため，各 (s_i,u_k) についての状態遷移確率の差異を表す量として，代わりにユークリッド距離を用いることを考える．さらに，事前分布一様を仮定して，つぎの評価関数を用いる．

$$
\begin{aligned}
&\tilde{\boldsymbol{\alpha}}_{LS}\\
&=\arg\min_{\alpha}\sum_{i,k}\left(z_{ik}\sum_j\left[p_{ij}(u_k;\eta_{itp(\boldsymbol{\alpha})})-\tilde{p}_{ij}(u_k)\right]^2\right)
\end{aligned}
\tag{8}
$$

上の最小化問題は，$\boldsymbol{\alpha}$ についての制約付き 2 次計画問題であり，比較的簡単に解を得ることができる．

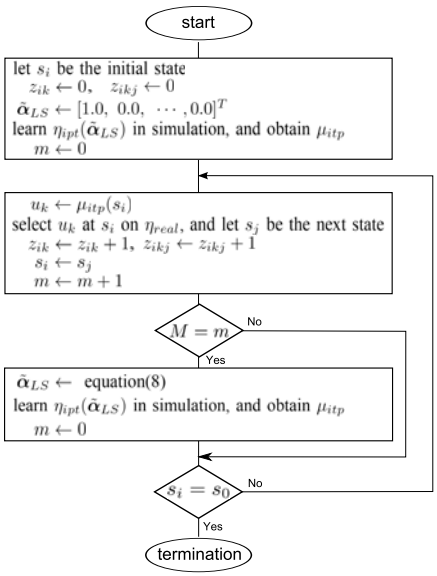

Fig. 1　提案手法 μ_{ABDP} のアルゴリズム

提案手法のアルゴリズムを Fig. 1 に示す．実プラントを η_{real} で表す．パラメータ $\boldsymbol{\alpha}$ をもつ内挿プラントを $\eta_{itp(\boldsymbol{\alpha})}$，これに対する最適方策を μ_{itp} によって表す．状態行動対の回数 z_{ik} と遷移の回数 z_{ikj} から，未知パラメータ $\boldsymbol{\alpha}$ の事後確率分布が得られる．この近似解法では，上述の理由で，Bayesian DP の中で事後確率分布からサンプルを抽出する動作を (8) 式で置き換えている．M は，実プラント上で用いる方策 η_{itp} を更新するタイミングを表し，必要とされる探索深さを指示する [6].

もともとの Bayesian DP がもつ性質として，M が大き過ぎる場合には新しい情報が反映されるのが遅いため，学習速度が低下することがある．一方で，M が小さ過ぎる場合には方策を急速に切り替えてしまい，結果的に終端状態にたどり着くのが遅くなる場合がある [8]．本研究の提案手法は，Bayesian DP を拡張した方法であるため，同様の性質をもつ可能性がある．

4.　ペグ・イン・ホール・タスクへの適用

提案手法を実際的なロボットタスクに適用し，有効性を検証する．本研究では，組立作業の本質を捉えたタスクの例として，ペグ・イン・ホール・タスクを対象に，数値シミュレーションと実機実験を行う．

ペグ・イン・ホール・タスクでは，ホールの位置・姿勢を計測し，ロボットが把持したペグをほぼ同サイズのホールへと挿入する．環境との相互作用を伴い，ビジョンシステムによるホールの位置姿勢計測誤差など，変動の影響を強く受ける．本研究では，実験装置に合わせて 2 次元平面問題を扱う．実験には，Fig. 2 に示す 3 関節 SCARA 型マニピュレータを用いる．各関節にはトルクセンサとサーボ制御器を備え，高精度のトルク制御を実現する [9]．また，数値シミュレーションには，Russell Smith 氏により開発された多関節剛体動力学シミュレータ・ライブラリである Open Dynamics Engine

Fig. 2 2次元ペグ・イン・ホール・タスク

を用いる.

4.1 問題設定

Fig. 3 に示すように，慣性座標系 Σ_0，手先座標系 Σ_E，ホール固定座標系 Σ_{hl} とする．タスク中では Σ_0 と Σ_{hl} の相対関係は固定される．Fig. 4 に示すように，ペグ幅は 74.0 [mm]，ホール幅は 74.25 [mm] と，ほぼ同サイズである．ロボットの手先が持つペグの位置は，ペグの左端から 37.0 [mm]，下端から 50.0 [mm] の位置である．

手先作業変数 $\boldsymbol{y}$ は，Σ_0 に関する Σ_P の位置 (y_x,y_y)，$\boldsymbol{k}_0$ 軸周りの回転 y_θ とする．手先への制御入力 $\boldsymbol{f}$ は，Σ_0 に関する手先が環境に加える力 (f_x,f_y)，トルク f_θ とする．参照作業変数 $\boldsymbol{y}_d$ が与えられたとき，つぎの制御則にしたがって関節トルク $\boldsymbol{\tau}$ を与える．

$$\boldsymbol{\tau} = -\boldsymbol{J}^T\boldsymbol{K}_P(\boldsymbol{y}-\boldsymbol{y}_d)-\boldsymbol{K}_D\dot{\boldsymbol{q}} \tag{9}$$

ただし，$\boldsymbol{J}$ はヤコビ行列，$\boldsymbol{q}$ は関節角度，$\boldsymbol{K}_P$, $\boldsymbol{K}_D$ はフィードバックゲインである．

連続系の問題であるペグ・イン・ホール・タスクを離散時間有限 MDP の枠組みで扱うため，つぎのような制限を加えることにより問題の単純化を行う．ペグが動き始めてから静止するまでを 1 ステップの状態遷移とみなす．そのため，プラントの状態は $[y_x,y_y,y_\theta,f_x,f_y,f_\theta]$ を用いて表現される．これら連続値をとる状態変数をそれぞれ，1.0 [mm]，1.0 [mm]，0.5π/180 [rad]，1.5 [N]，1.0 [N]，0.3 [Nm] の幅で量子化する．各状態変数を量子化する個数は [10,7,11,3,5,5] とし，計 57750 個の離散状態が存在する．また，離散化された状態変数 y_x が最小値の場合を，ペグが挿入された終端状態とみなす．

ロボットは，Fig. 3 に示す離散行動 $u_1\cdots,u_6$ を通じて，目標とする参照作業変数 $\boldsymbol{y}_d$ を与える．各離散行動を選択することは，$\boldsymbol{i}_0$, $\boldsymbol{j}_0$ 軸方向の位置または $\boldsymbol{k}_0$ 軸回りの姿勢に関して，$\boldsymbol{y}_d$ の連続値を 1 離散状態幅だけ増減させることに対応する．ロボットは，参照作業変数 $\boldsymbol{y}_d$ に対して，(9) 式に基づいてペグが静止するまで制御を行う．ペグを加速させてから静止するまでの一連の離散行動に要する時間は，おおよそ 4 秒である．

学習を行う際の 1 ステップコストは，すべての状態遷移に対して等しく 1 を与える．また，離散状態空間の外側に遷移した場合には，特別に 10^7 のコストを追加する．

なお，50 ステップを超過したエピソードをタスク失敗として扱う．これは，実験装置が故障しないように，実験装置の連続稼働時間を考慮して設定されている．本実験では，実サンプルの課題で言及されている実際的な学習時間の許容される上限を，50 ステップとみなす．

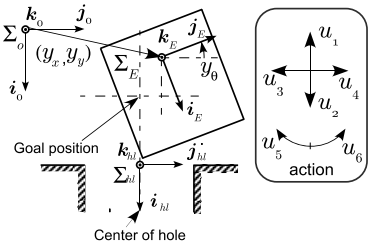

Fig. 3 問題定義

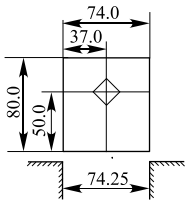

Fig. 4 ペグ･ホール

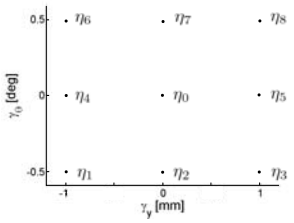

Fig. 5 プラントの有限集合 $\mathcal{H}$

4.2 先見情報として与える変動モデル

変動モデルとして，ビジョン・システムによるホールの計測誤差を想定する．計測の $\boldsymbol{j}_0$ 方向の位置誤差 γ_y は，平均 $m_y=0.0$ [mm]，標準偏差 $\sigma_y=0.75$ [mm]，$\boldsymbol{k}_0$ 軸周りの姿勢誤差 γ_θ は平均 $m_\theta=0.0$ [rad]，標準偏差 $\sigma_\theta=0.5\pi/180$ [rad] の正規分布に従うとする．したがって，環境を特徴づける未知パラメータは，$\boldsymbol{\gamma}=(\gamma_y,\gamma_\theta)$ である．未知パラメータ $\boldsymbol{\gamma}$ は，エピソードごとに変化する．

パラメータ $\boldsymbol{\gamma}$ に対応する状態遷移確率 $p_{ij}(u_k;\eta(\boldsymbol{\gamma}))$ は，以下の手順で生成される．動力学シミュレータ上で，プラント $\eta(\boldsymbol{\gamma})$ に対して実験装置と同等の制御 (9) を適用し，状態遷移 (s_i,u_k,s_j) のサンプルを得る．これらのサンプルを各状態行動対 (s_i,u_k) について収集し，状態遷移確率 $p_{ij}(u_k;\eta(\boldsymbol{\gamma}))$ を導く．なお，シミュレータ上で設定するペグとホール間の摩擦係数は，別途行った実験装置の同定試験結果に基づいて与えている．

内挿に用いるプラントの有限集合 $\mathcal{H}$ の要素として，$\gamma_y\in[-1.0,\ 0.0,\ 1.0]$ [mm]，$\gamma_\theta\in[-0.5\pi/180,\ 0.0,\ 0.5\pi/180]$ [rad] の誤差 $\boldsymbol{\gamma}$ をもつ，9 個のプラントを考える．ここで考慮する誤差パラメータ $\boldsymbol{\gamma}$ の間隔は，状態変数 $\boldsymbol{y}$ の離散化幅と対応して選ばれている．$\gamma_y=0$, $\gamma_\theta=0$ の場合を，ノミナルプラント η_0 とする．また，変動プラント $\eta_1,\cdots,\eta_8$ は Fig. 5 に示す通りとする．これら 9 個のプラント η_n は，上述の手順で状態遷移確率 $p_{ij}(u_k;\eta_n)$ を生成する．その他の連続的に存在する変動プラント $\eta(\boldsymbol{\gamma})$ は，有限プラント集合 $\mathcal{H}$ の内挿により近似表現する．

Table 1　シミュレーション結果 (s^{t_0} 固定・10000 エピソード)

方策	タスク達成	状態空間外	超過	平均	標準偏差
μ^*_{nom}	42.1%	0.04%	57.9%	35.3	18.4
μ_{ABDP}	93.0%	0.00%	7.0%	18.6	10.7

Table 2　実機実験結果 (s^{t_0} 固定・50 エピソード)

方策	タスク達成	状態空間外	超過	平均	標準偏差
μ^*_{nom}	40%	0%	60%	34.5	19.2
μ_{ABDP}	96%	0%	4%	17.2	10.2

Table 3　M を変更した場合のシミュレーション結果（s^{t_0} ランダム・10000 エピソード）

頻度	タスク達成	状態空間外	超過	平均	標準偏差
$M=1$	99.1%	0.00%	0.9%	13.7	5.3
$M=5$	98.3%	0.22%	1.5%	15.2	7.1
$M=10$	96.8%	0.73%	2.5%	16.7	8.6
$M=15$	93.5%	2.60%	3.9%	19.8	10.7

4.3　結果

ノミナルプラントに対する最適方策 μ^*_{nom} と，提案手法 μ_{ABDP} を比較する．ある初期状態 s^{t_0} に固定して，プラントのパラメータ $\boldsymbol{\gamma}$ が **4.2** 節のように変化する場合に，μ^*_{nom}, μ_{ABDP} を適用した時のシミュレーション結果を Table 1 に，実機実験結果を Table 2 に示す．なお，表中の平均と分散は，タスク失敗時のステップ数を 50 として，成功・失敗時の両方のステップ数を考慮して算出したものである．

パラメータ $\boldsymbol{\gamma}$ が変動することをまったく考慮しないノミナルプラント最適方策 μ^*_{nom} と比べて，提案手法 μ_{ABDP} ではタスク達成率が向上した．また，初期状態 s^{t_0} を変えて行った場合についても，同様の結果が得られた．

タスク失敗の条件を 50 ステップから 5000 ステップに変更してシミュレーションを行ったところ，提案手法 μ_{ABDP} ではタスク達成率 100%を達成した．装置の制約のため同条件での実験を行うことはできなかったが，その場合についても同様の結果が得られると考えられる．

また，初期状態をランダムにとり，M の値を変えて提案手法を適用したシミュレーション結果を，Table 3 に示す．M が大きくなるにつれて，情報が反映されるのが遅くなり，コストが高い状態空間外への遷移をする割合が増加している．一方で，M が小さすぎる場合の方策の急速な切り替えによる悪影響は，本タスクでは見られなかった．

4.4　挙動

変動 $\boldsymbol{\gamma}_y = 1.0$ [mm]，$\boldsymbol{\gamma}_\theta = 0.0$ [rad] のプラントに対して提案手法を適用した場合の，実機実験の一例を示す．

ペグの挙動を Fig. 6 に図示する．また，内挿の重み $\boldsymbol{\alpha}$

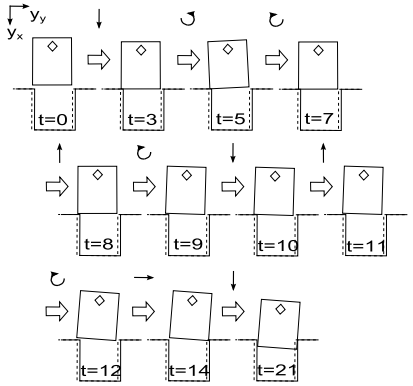

Fig. 6　提案手法 μ_{ABDP}；ペグの挙動

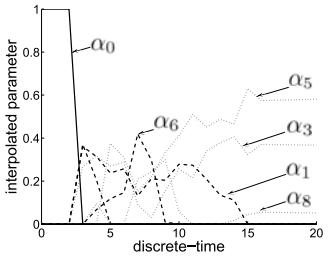

Fig. 7　提案手法 μ_{ABDP}；推定の履歴

の推定の履歴を Fig. 7 に示す．Fig. 7 中では，内挿重み成分 α_n について，y_y 方向の変動の違いを線の形式で区別している．実線が変動なし $\gamma_y = 0.0$ [mm]，破線が変動 $\gamma_y = -1.0$ [mm]，点線が変動 $\gamma_y = 1.0$ [mm] を表す．

タスクを開始した時点 $t=0$ では，プラント変動について情報が得られていないため，ノミナルプラント（点線）のホールに向かってペグを差し入れようとする．

$t=3$ では，実プラント（実線）のホールのずれにより，ペグがホールから y_x 方向反力を受けて止まる．この情報をもとに，η_1, η_3, η_4 に関する内挿の重みが大きいプラントが，現実をよく近似すると推定される．この内挿プラントにおける最適行動は y_θ 正方向の回転である．

$t=5$ では，実プラントのホールの y_y 正方向のずれにより，ペグがホールから受ける y_x 方向反力と y_θ 方向トルクが大きくなる．この情報をもとに，η_5（実プラントと同じ変動を考慮）に関する内挿の重みが大きいプラントが，現実をよく近似すると推定される．この内挿プラントにおける最適行動は y_θ 負方向の回転であり，以前に生じた y_θ 正方向の回転をもとに戻そうとする行動である．

$t=7$ では，$t=3$ とほぼ同じような状況に戻る．これまでの情報をもとにすると，η_6 に関する内挿の重みを大きくすることが，現実をよく近似すると推定される．内挿プラントに対して学習した結果，y_x 負方向の移動を行い，ホールからの反力が加わらないようにする．

$t=8$ では，離散状態空間上で見ればペグには力が働いていないような状況となる．これまでの情報をもとにすると，η_1, η_5, η_6, η_8 に関する内挿の重みを大きくすることが，現実をよく近似すると推定される．それぞれの

プラントについての最適行動は，η_1 については u_5，η_5 については u_4，η_6 については u_3，η_8 については u_6 と分かれるが，この内挿プラントでは u_6 が選択される．

$t=9$ では，ホールに拘束されることなく，ペグが y_θ 負方向に回転している．この状態遷移は η_1 ではほとんど起こらないため，η_6 の重みが小さくなり，一方で η_3 が大きくなることが，現実をよく近似すると推定される．内挿プラントに対して学習した結果，y_x 正方向の移動を行い，ペグを差し入れようとする．

$t=10$ では，ペグがホールに接触し，反力を受けて止まる．これまでの情報より，η_5 に関する内挿の重みを大きくすることが，現実をよく近似すると推定される．内挿プラントに対して学習した結果，y_x 負方向の移動を行い，ペグをホールから離す．

$t=11$ では，ペグを y_θ 負方向に回転させる．$t=12$ では，ペグを y_y 正方向に並進させる．$t=14$ では，ペグを y_x 正方向に並進させ，ペグを突き入れる．

本実験では，必要とされる探索深さは未知であるため，$M=1$ として提案手法を適用した．事後分布から $\tilde{\boldsymbol{\alpha}}$ の抽出，およびオンラインで η_{itp} の最適方策を計算するのに要する時間は，おおよそ 5 秒である（CPU に AMD Athlon(TM) XP 1900+を使用した場合）．そのため，ロボットアーム制御と方策更新を交互に行ったとしても 1 ステップ状態遷移は 9 秒程度であり，提案手法は計算コストの点でも実用的であると考える．

このように提案手法 μ_{ABDP} では，実プラント上で生じた状態遷移から内挿の重み $\boldsymbol{\alpha}$ を通じてホールの誤差を推定し，推定された内挿プラント $\eta_{itp(\tilde{\boldsymbol{\alpha}})}$ に対する学習結果を適用する．ロバスト性をもつ方策を，実プラントの推定に基づいて変化させながら用いるため，提案手法は高い確率で効率的にタスクを達成できる．

5. 結論

本研究では，有限個のモデルの内挿補間によって未知パラメータを近似表現するベイジアン強化学習を提案した．内挿モデルに対する学習結果は，あるモデル集合に対してロバストに機能するという性質をもつ．この性質を利用した提案手法は，少ない実サンプル数と低い計算コストで用いることが可能であり，先見情報の近似誤差に対しても有効にはたらく．

提案手法を実際的なロボット問題に対して適用したところ，数値シミュレーションと実機実験の両方で，実際的な時間内でタスクを達成した．これらの結果より，提案手法において導入した近似方法が，現実に存在するモデル化誤差に対して有効であることを示した．また，計算コストの点でも利用可能な方法であることを実証した．

提案手法の今後の課題として，つぎのことが挙げられる．まず，内挿プラントの最適方策をオンラインで計算する必要があるため，タスクのサイズが大きい場合には，オンライン近似解で代用する必要がある．また，連続無限個の変動プラントを有限個の変動プラントで内挿近似しているため，近似誤差が大きい場合には有効ではない可能性がある．扱う変動の幅が大きい場合には，内挿に用いる変動プラントの個数 N を増やさなくてはならないが，これは計算コストに影響する．動的なタスクなどでは，状態遷移とオンライン計算を同時に行う必要がある．本研究で扱ったペグ・イン・ホール・タスクのように，状態遷移とオンライン計算を分離できない場合には，方策の更新間隔 M を任意に設定できない．

謝　辞

本研究は，文部科学省の科学研究費補助金に関連してなされたことを付記する．

参 考 文 献

[1] R. S. Sutton and A. G. Barto: *Reinforcement Learning*, MIT Press (1998)
[2] D. P. Bertsekas and J. N. Tsitsiklis: *Neuro-Dynamic Programming*, Athena Scientific (1996)
[3] J. Kober, D. Bagnell and J. Peters: Reinforcement learning in robotics: A survey; *Int. J. Robotics Research*, Vol. 32, No. 11, pp. 1238–1274 (2013)
[4] R. S. Sutton, et al.: Policy gradient methods for reinforcement learning with function approximation; *Proc. NIPS 12*, pp. 1057–1063 (1999)
[5] 藤井, 泉田, 真野：状態遷移モデルを逐次推定する強化学習の学習加速；計測自動制御学会論文集, Vol. 42, No. 1, pp. 47–53 (2006)
[6] M. Strens: A Bayesian framework for reinforcement learning; *Proc. ICML*, pp. 943–950 (2000)
[7] C. M. Bishop: *Pattern Recognition and Machine Learning*, Springer (2006)
[8] J. Asmuth, et al.: A Bayesian sampling approach to exploration in reinforcement learning; *Proc. UAI*, pp. 19–26 (2009)
[9] K. Senda and Y. Tani: Autonomous robust skill generation using reinforcement learning with plant variation; *Advances in Mechanical Engineering*, Vol. 2014, 276264, 12 pp (2014)

付　録

定理

プラント $\eta_\nu \in \mathcal{H}$ の内挿によって生成される変動プラント η^ν_{itp} は，次の状態遷移確率をもつ．

$$p_{ij}(u_k;\eta^\nu_{itp}) \equiv \sum_{n=0}^{N-1} \alpha_n p_{ij}(u_k;\eta_n) \tag{A1}$$

$$\begin{cases} 0 < \alpha_n \leq 1 \, (n=\nu) \\ 0 \leq \alpha_n < 1 \, (n \neq \nu) \\ \sum_{n=0}^{N-1} \alpha_n = 1, \, \forall i, \, \forall k \end{cases}$$

また，内挿によって生成される変動プラント η^{ν}_{itp} の無限集合 $\mathcal{H}^{\nu}_{itp}$ をつぎのように定義する．

$$\mathcal{H}^{\nu}_{itp} \equiv \left\{ \eta'_{itp(\boldsymbol{\alpha})} \left| \begin{array}{l} 0<\alpha_n \le 1\,(n=\nu) \\ 0\le \alpha_n <1\,(n\neq\nu) \\ \sum_{n=0}^{N-1}\alpha_n = 1 \end{array} \right. \right\}$$

このとき，プラント $\eta_{\nu}\in\mathcal{H}$ に対してプロパーな静的方策 μ_{ν} は，プラント η_{ν} の内挿によって生成される変動プラント $\eta'_{itp}\in\mathcal{H}^{\nu}_{itp}$ に対してプロパーである．

記号の定義

静的方策 μ をプラント η_n に適用するとき，状態 s_i から s_j への遷移が起こる確率を次式で表す．

$$p^{\mu,\eta_n}_{ij} \equiv \sum_{k=1}^{K} \pi(s_i,u_k) p_{ij}(u_k;\eta_n) \tag{A2}$$

ただし，$\pi(s_i,u_k)$ は，状態 s_i において行動 u_k を選択する確率を表す．(A2) 式を i 行 j 列要素にもつ $(N_s+1)\times(N_s+1)$ 行列を，遷移確率行列 $\boldsymbol{P}^{\mu}_{\eta_n}$ とする．N_s+1 番目の要素は終端状態 s_0 に対応する．行列 $\boldsymbol{P}^{\mu}_{\eta_n}$ に対応する隣接行列を $\boldsymbol{R}^{\mu}_{\eta_n}$ によって表す．$\boldsymbol{R}^{\mu}_{\eta_n}$ は $(N_s+1)\times(N_s+1)$ 行列で，ij 要素は，$p^{\mu,\eta_n}_{ij}=0$ のとき 0，$p^{\mu,\eta_n}_{ij}>0$ のとき 1 である．すなわち，$\boldsymbol{R}^{\mu}_{\eta_n}$ の ij 要素は，状態 s_i から s_j へ 1 ステップで遷移する確率が 0 であれば 0，同確率が正であれば 1 となる．終端状態 s_0 の定義より，行列 $\boldsymbol{P}^{\mu}_{\eta_n}$ と $\boldsymbol{R}^{\mu}_{\eta_n}$ の (N_s+1) 行は，ともに $[0\ \cdots\ 0\ 1]$ である．

隣接行列 $\boldsymbol{R}^{\mu}_{\eta_n}$ の行列演算における和積演算に，つぎのブール演算を用いる．

$$\begin{array}{llll} 0+0=0 & 0+1=1+0=1 & 1+1=1 \\ 0\cdot 0=0 & 0\cdot 1=1\cdot 0=0 & 1\cdot 1=1 \end{array} \tag{A3}$$

この演算によれば，$(\boldsymbol{R}^{\mu}_{\eta_n})^m$ の ij 要素は，状態 s_i から s_j へ m ステップで遷移する確率が正であれば 1，同確率が 0 であれば 0 となる．したがって，つぎの隣接行列

$$(\boldsymbol{R}^{\mu}_{\eta_n})^{\sim N_s} \equiv \boldsymbol{I} + \boldsymbol{R}^{\mu}_{\eta_n} + (\boldsymbol{R}^{\mu}_{\eta_n})^2 + \cdots + (\boldsymbol{R}^{\mu}_{\eta_n})^{N_s} \tag{A4}$$

の ij 要素は，状態 s_i から s_j へ N_s ステップ以内で遷移する確率が正であれば 1，同確率が 0 であれば 0 となる．

証明

プロパー性の定義より，静的方策 μ_{ν} がプラント η_{ν} に対してプロパーである場合，任意の状態 s_i から N_s ステップ以内に終端状態に到達する確率は正である．(A4) 式より，μ_{ν} が η_{ν} に対してプロパーであることは，$(\boldsymbol{R}^{\mu_{\nu}}_{\eta_{\nu}})^{\sim N_s}$ の N_s+1 列の要素がすべて 1 であることと等価である．

プラント η_{ν} に対してプロパーな静的方策 μ_{ν} を，内挿プラント $\eta'_{itp(\boldsymbol{\alpha})}$ に対して適用する．状態 s_i から s_j への遷移確率は，(A1) 式の定義よりつぎのように得られる．

$$p^{\mu_{\nu},\eta'_{itp}}_{ij} \equiv \sum_{k=1}^{K} \pi_{\nu}(s_i,u_k) p_{ij}(u_k;\eta_{itp}) = \sum_{n=1}^{N-1} \alpha_n p^{\mu_{\nu},\eta_n}_{ij}$$

上式を ij 要素にもつ行列を遷移行列 $\boldsymbol{P}^{\mu_{\nu}}_{\eta_{itp}}$，その隣接行列を $\boldsymbol{R}^{\mu_{\nu}}_{\eta_{itp}}$ とする．プラント η_{ν} に対して $0<\alpha_{\nu ij}\le 1, \forall i,j$ である．(A1) 式より，$p^{\mu_{\nu},\eta_{\nu}}_{ij}>0$ ならば $p^{\mu_{\nu},\eta^{\nu}_{itp}}_{ij}>0$，$p^{\mu_{\nu},\eta_{\nu}}_{ij}=0$ ならば $p^{\mu_{\nu},\eta^{\nu}_{itp}}_{ij}\ge 0$ であり，つぎが成り立つ．

$$\boldsymbol{R}^{\mu_{\nu}}_{\eta_{\nu}} \le \boldsymbol{R}^{\mu_{\nu}}_{\eta^{\nu}_{itp}}$$

上式は $\boldsymbol{R}^{\mu_{\nu}}_{\eta_{\nu}}(i,j) \le \boldsymbol{R}^{\mu_{\nu}}_{\eta^{\nu}_{itp}}(i,j), \forall i,j$ を意味する．すなわち，隣接行列 $\boldsymbol{R}^{\mu_{\nu}}_{\eta_{\nu}}$，$\boldsymbol{R}^{\mu_{\nu}}_{\eta^{\nu}_{itp}}$ のすべての ij 要素について，つぎのいずれかの関係のみが成り立つ．

$$\begin{array}{l} \boldsymbol{R}^{\mu_{\nu}}_{\eta_{\nu}}(i,j)=0 \text{ かつ } \boldsymbol{R}^{\mu_{\nu}}_{\eta^{\nu}_{itp}}(i,j)=0 \\ \boldsymbol{R}^{\mu_{\nu}}_{\eta_{\nu}}(i,j)=0 \text{ かつ } \boldsymbol{R}^{\mu_{\nu}}_{\eta^{\nu}_{itp}}(i,j)=1 \\ \boldsymbol{R}^{\mu_{\nu}}_{\eta_{\nu}}(i,j)=1 \text{ かつ } \boldsymbol{R}^{\mu_{\nu}}_{\eta^{\nu}_{itp}}(i,j)=1 \end{array}$$

規則 (A3) より，$\boldsymbol{R}^{\mu_{\nu}}_{\eta^{\nu}_{itp}}$ を用いる演算結果は，$\boldsymbol{R}^{\mu_{\nu}}_{\eta_{\nu}}$ を用いる演算結果より小さくならない．そのため，$\boldsymbol{R}^{\mu_{\nu}}_{\eta_{\nu}} \le \boldsymbol{R}^{\mu_{\nu}}_{\eta^{\nu}_{itp}}$ ならば，$(\boldsymbol{R}^{\mu_{\nu}}_{\eta_{\nu}})^m \le (\boldsymbol{R}^{\mu_{\nu}}_{\eta^{\nu}_{itp}})^m$ であり，つぎが成り立つ．

$$(\boldsymbol{R}^{\mu_{\nu}}_{\eta_{\nu}})^{\sim N_s} \le (\boldsymbol{R}^{\mu_{\nu}}_{\eta^{\nu}_{itp}})^{\sim N_s} \tag{A5}$$

静的方策 μ_{ν} がプラント η_{ν} に対してプロパーであるならば，$(\boldsymbol{R}^{\mu_{\nu}}_{\eta_{\nu}})^{\sim N_s}$ の N_s+1 列の全要素が 1 であるため，$(\boldsymbol{R}^{\mu_{\nu}}_{\eta^{\nu}_{itp}})^{\sim N_s}$ の N_s+1 列の全要素が 1 となる．したがって，プラント η_{ν} に対してプロパーな静的方策 μ_{ν} は，内挿によって生成される変動プラント η^{ν}_{itp} の無限集合 $\mathcal{H}^{\nu}_{itp}$ のすべてに対してプロパーである．(証明終了)

著 者 略 歴

泉田（せんだ）　啓（けい）(正会員)

1988 年大阪府立大学大学院博士前期課程修了．同年同大学工学部助手，1994 年同助教授，2002 年金沢大学助教授，2007 年同教授，2008 年京都大学工学研究科教授．1996-97 年ミシガン州立大学客員教授など兼任．ロボットや動物の運動知能などの研究に従事．博士（工学）．1994 年システム制御情報学会賞論文賞など受賞．日本ロボット学会などの会員．

菱沼（ひしぬま）　徹（とおる）(学生会員)

2015 年 3 月京都大学大学院工学研究科航空宇宙工学専攻修士課程修了，同年 4 月同専攻博士課程に進学，現在に至る．ロボットの知能化や強化学習などの研究に従事．

谷（たに）　百合夏（ゆりか）

2012 年 3 月京都大学大学院工学研究科航空宇宙工学専攻修士課程修了．ロボットの知能化や強化学習などの研究に従事．