NAIST-IS-MT1651028

修士論文

ガウス過程方策を用いた方策探索と ロボットによる布操作タスクへの適用

小澤 裕斗

2018 年 3 月 13 日

奈良先端科学技術大学院大学
情報科学研究科 情報科学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
修士 (工学) 授与の要件として提出した修士論文である。

小澤 裕斗

審査委員：

杉本 謙二 教授	（主指導教員）
小笠原 司 教授	（副指導教員）
松原 崇充 准教授	（副指導教員）
小蔵 正輝 助教	（副指導教員）
小林 泰介 助教	（副指導教員）

ガウス過程方策を用いた方策探索と ロボットによる布操作タスクへの適用*

小澤 裕斗

内容梗概

近年, 経験データから試行錯誤を通して行動を最適化する手法として, 強化学習が注目を浴びている. 特に, 不確実な環境下でロボットが知的かつ頑健にタスクを遂行するためには, ロボットや環境のシステムといった未知のモデルを確率モデルで表現する必要があるため, 強化学習による研究が数多く行われている. しかし, ロボットの状態は多くの場合, 連続かつ高次元であるので, 強化学習をロボットの運用に適用するのは容易ではない.

そこで, 本研究では, 高次元入力に頑健なガウス過程方策を用いた方策探索を提案する. 具体的には 2 つの手法を提案する. 1 つ目に, 学習や予測に要する計算量を低減させることを目標に, スパースガウス過程方策を用いた方策探索を提案する. 2 つ目に, ガウス過程方策の単峰性という課題を解決するために, 重複混合ガウス過程方策を用いた方策探索を提案する.

まず, 2 つの提案法それぞれに対して, 複数のシミュレーションを行い, 提案法の有効性を確認した. 次に, 実機実験として, 4 つ折りに折りたたまれた布の展開タスクを行い, 提案法のロボット制御への有効性を確認した.

キーワード

強化学習, 方策探索, ガウス過程，スパースガウス過程, 重複混合ガウス過程

*奈良先端科学技術大学院大学 情報科学研究科 修士論文, 情報科学専攻 NAIST-IS-MT1651028, 2018 年 3 月 13 日.

Gaussian Process Policy Search and Its Application to Robot Cloth Manipulation Task*

Ozawa Yuto

Abstract

Recently, reinforcement learning has been drawing much attention as a method to optimize action from experience data through trial and error. In particular, in order for a robot to perform a task intelligently and robustly under uncertain circumstances, it is necessary to express unknown models such as systems of robot and environment by probabilistic models, so many researches based on reinforcement learning are performed. However, since the state of the robot is continuous and high dimensional, the application to robot operation is often hard.

Therefore, we explore a methodology using Gaussian processes as a policy model, which can handle continuous and high dimensional input. Specifically, we propose two methods. First, we propose a policy search algorithm using Sparse pseudo-input Gaussian processes with the aim of reduction of calculation cost required for learning and prediction. Second, to tackle the limitation of unimodality in the standard Gaussian process policy search, we propose a policy search algorithm using Overlapping Mixtures of Gaussian processes as a policy model.

First, we conducted several experiments in simulation to confirm the effectiveness of both methods. Then, we conducted real robot experiments for the fabric unfolding task with a Baxter robot.

Keywords:

Reinforcement Learning, Policy Search, Gaussian Processes, Sparse Pseudo-input Gaussian Processes, Overlapping Mixtures of Gaussian Processes

*Master's Thesis, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-MT1651028, March 13, 2018.

目 次

1. 序論 **1**

1.1 はじめに 1

1.2 ロボットの方策探索 2

1.2.1 多峰性を考慮した方策探索 3

1.2.2 画像を入力とする方策探索 4

1.2.3 画像を入力とする深層強化学習 5

1.3 研究目的 7

1.4 本論文の構成 9

2. 準備 **10**

2.1 強化学習 10

2.2 ガウス過程回帰 10

2.3 EM 方策探索 12

3. 提案法 **13**

3.1 提案法 1：スパースガウス過程方策を用いた方策探索 13

3.1.1 スパースガウス過程方策モデル 13

3.1.2 変分推論による SPGP 方策改善 14

3.1.3 SPGP 方策の予測分布 15

3.1.4 回帰問題による動作確認 16

3.2 提案法 2：重複混合ガウス過程方策を用いた多峰性方策探索 . . . 17

3.2.1 重複混合ガウス過程方策モデル 17

3.2.2 OMGP 方策モデルの方策探索 18

3.2.3 OMGP 方策の予測分布 21

3.2.4 回帰問題による動作確認 1 22

3.2.5 回帰問題による動作確認 2 24

4. 提案法 1 によるシミュレーション実験 **26**

4.1 衛星問題に対する方策探索 26

4.1.1 設定 26
4.1.2 結果 28

5. 提案法 2 によるシミュレーション実験 30
5.1 1 段階スリット通り抜け問題 30
5.1.1 設定 30
5.1.2 結果 31
5.2 3 段階スリット通り抜け問題 33
5.2.1 設定 33
5.2.2 結果 34
5.3 衛星問題に対する方策探索 40
5.3.1 設定 40
5.3.2 結果 41

6. 実機実験 43
6.1 設定 43
6.2 結果 44

7. おわりに 47
7.1 まとめ 47
7.2 今後の課題 47

謝辞 48

参考文献 49

付録 52

A. 提案法 1 における F_V の導出 52

B. 逆行列の計算に利用した行列の公式 53

C. ガウス分布の積の公式 53

D. 重複混合ガウス過程のハイパーパラメータの最適化 **54**

E. 空間的ピラミッドカーネル **54**

図 目 次

1 布操作に関する研究で用いられている特徴 2
2 方策探索によるロボットの強化学習 3
3 複数の最適な方策から自動的に 1 つ求める遷移図 4
4 パラメトリック方策の組み合わせで多峰性を表現 4
5 画像を入力とする方策探索 5
6 画像を入力とする深層強化学習 6
7 BAXTER による布展開タスク 6
8 提案法 1:スパースガウス過程方策 7
9 提案法 2:重複混合ガウス過程方策 7
10 疑似入力を 2 点選んだときの提案法 1 による予測分布. 16
11 疑似入力を 15 点選んだときの提案法 1 による予測分布 17
12 従来法 (上段) と提案法 2(下段) から得た方策の予測分布 22
13 提案法 2 から得られた確率値 $\hat{\mathbf{\Pi}}$ 23
14 一次元の入力に対して, 4 種類の異なる出力を伴う観測データ. . . 25
15 提案法 2 の方策から得られた予測分布 25
16 衛星問題の概念図 26
17 衛星問題での従来法 (緑) と提案法 1(青) の学習曲線 29
18 初期方策を用いたときの遷移図 29
19 10 回更新後の方策を用いたときの遷移図 29
20 1 段階スリット問題の概念図 30
21 1 段階スリット問題での従来法 (緑) と提案法 2(青) の学習曲線 . . 31
22 従来法による 9 回更新後の方策から得られた観測データ 32
23 提案法 2 による 9 回更新後の方策から得られた観測データ 32
24 3 段階スリット問題の概念図 33
25 3 段階スリット問題での従来法 (緑) と提案法 2(青) の学習曲線 . . 35
26 3 段階スリット問題での従来法の方策から得られた観測データ . . 36
27 3 段階スリット問題での提案法 2 の方策から得られた観測データ . 37
28 3 段階スリット問題での提案法 2 の方策 38

29 従来法の方策 (緑) と提案法 2 の方策 (赤と青) 39
30 衛星問題の概念図 39
31 衛星問題での従来法 (緑) と提案法 2(青) の学習曲線 42
32 14 回方策を更新した後の衛星の遷移図 42
33 布展開タスクの目標とする遷移図 43
34 布展開タスクの概念図 44
35 布展開タスクでの 3 回の実験結果 45
36 初期方策を用いたときの遷移図 46
37 4 回更新後の方策を用いたときの遷移図 46
38 10 回更新後の方策を用いたときの遷移図 46
39 レベル 0 のときに検出された SIFT 特徴量 56
40 レベル毎の区切り (上段) とヒストグラム (下段) 56

表 目 次

1 行動決定に要した計算時間 28

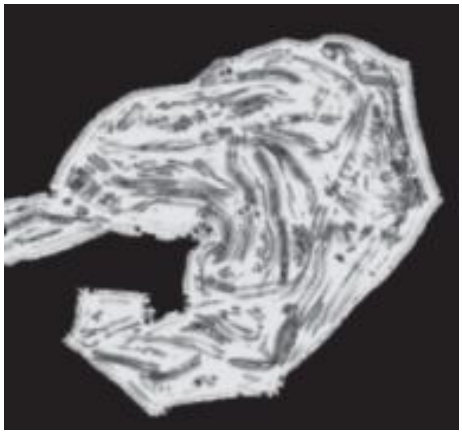
1. 序論

1.1 はじめに

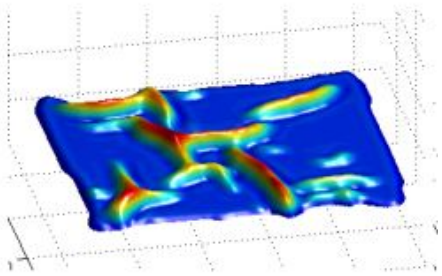
煩雑な環境下において，ロボットが環境状況を理解し，適切な行動を選択することができれば, 一般家庭や工場, 物流現場などで活躍が期待できる．そのような環境におけるロボット運用の研究では, 力を加えても形が変わらない物体である剛体を扱うことを想定したものが多かった．しかし, 実環境では力を加えると形が変わる, 紐や布のような柔軟物も多い．ロボットがこのような柔軟物を操作することができれば, 今まで以上に多くのタスクを行えるようになる．例えば紐で結び目を作る操作を応用し, 物を括ることができるようになる [1]．また, チューブやホースによる配管操作が可能になる [2]．布の操作を応用すれば, 服の折りたたみや [3], 物の包装が可能になる．このように, ロボットが柔軟物を操作できるようになれば, 生活・産業分野において大きな貢献ができる．しかし, 柔軟物操作においては柔軟物の正確なモデルの推定が困難である．そのため, 例えば布の操作であれば, 文献 [4, 5] のように, 布地の特徴や, しわの特徴に注目してロボットを動かしている (図 1)．このような特徴をもとに, 特定の柔軟物を操作することは可能であるが, 近似的な推定量のため, 柔軟物によるモデル化誤差が大きい．

そこで近年, Q 学習, 深層強化学習, 方策探索といった様々な強化学習の研究が行われている．特に, 方策探索による強化学習では, 事前に環境やロボットの状態に対して, いくつかの仮定を置くことで, 比較的少ない観測データから適切なロボットの動作を求めることができる．また, 得られるロボットの動作は連続量であるので, 連続量を扱うロボットとの相性が良い．

故に, 本研究では, 方策探索に注目した強化学習手法を提案する．方策には, 観測データの密度やばらつきに柔軟なガウス過程方策を用いた．従来のガウス過程方策では, 計算量が観測データ数に依存していたため, 多くの観測データを必要とする問題において計算量が膨大になるという課題があった．それを解決するために, 1 つ目の提案法として従来法よりも学習や予測に必要な計算量を削減する手法を提案する．また, 従来法では, ある状態に対して複数の最適な行動があるときに, 正しい行動の予測ができないという課題があった．それを解決するために,



文献 [4] の布地の特徴



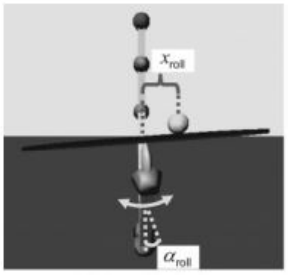
文献 [5] のしわの特徴

図 1 布操作に関する研究で用いられている特徴

2 つ目の提案法として複数のガウス過程を事前に仮定することで, 複数の最適な行動を予測可能な手法を提案する. また, 人は主に視覚情報を頼りに, 不確実な環境下においても様々なタスクを行っていると考え, 提案法では画像を入力とする方策探索を行った. 本研究の目的を述べる前に, 次節にてロボットに適用可能な方策探索の関連研究について紹介する. 複数の最適な行動を予測できる多峰性をもつ方策探索について紹介した後, 画像を入力とする方策探索について紹介する. 最後に, 近年流行りの, 深層強化学習を用いた画像入力からの行動学習について紹介する.

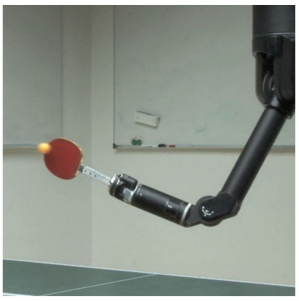
1.2 ロボットの方策探索

ロボットの行動規則を自律的に学習する手法として，強化学習の一手法である方策探索がある [6, 7]. 従来法の多くは，パラメトリックな方策モデルを仮定した上で，報酬和を最大化する方策のパラメータを探索する．例えば, 線形方策モデルを利用した方策探索として文献 [8–10] がある (図 2). 文献 [8] では, ガウス分布によるパラメトリックな方策を用いており, EM 方策探索 [11] においてパラメータの更新ごとに得られる観測データを, 効率的に再利用する手法を提案している. シーソーの上にあるボールを制御する問題の学習と, 2 リンク振り上げ安定化問



ボールバランス [8]

パンケーキ [9]



卓球 [10]

図 2 方策探索によるロボットの強化学習

題の学習によって, 提案法の有効性を示している. また, 文献 [9] でも, 同じくパラメトリックな方策を用いる. この研究では, フライパンを握ったロボットアームを用いて, パンケーキを模した厚紙を裏返す動作を学習する. 方策探索には EM 方策探索を利用している. 文献 [10] においては, パラメトリックな方策モデルを用いた方策探索によって, ロボットが卓球, ダーツ, 投球といった複数の動作を学習している. しかし, このようなパラメトリック方策を用いた方策探索は, 画像から得られる高次元特徴量を入力に使用すると次元の呪いが懸念される.

1.2.1 多峰性を考慮した方策探索

図 3 に示すように, 多峰性を考慮した単峰方策の研究もある [12]. この文献では, 求める単峰ガウス分布方策を, 複数の最適な方策から自動的に 1 つ求める手法である. しかし, 計算が複雑になるという問題がある. また, 図 4 に示すように, 方策を多峰に拡張することで, 多峰性のある問題に対応した研究もある [13]. この文献では階層表現を用いることで多峰な方策を提案している. しかし, パラメトリック方策の組み合わせによって多峰性を表現しているため, 高次元入力に対して次元の呪いが懸念される. 次元の呪いを解決しつつ多峰性を考慮したものに, 混合ガウス過程 [14] を用いた方策探索がある [15]. これは無限次元の混合ガウス過程を用いて, 方策のパラメータと報酬の関係を理解し, 多峰な方策を表現している.

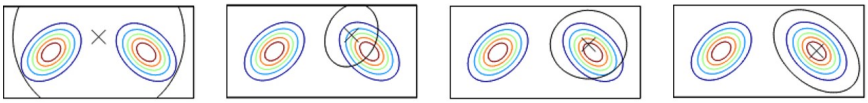


図 3 複数の解から自動的に 1 つの解を求める遷移図 [12]

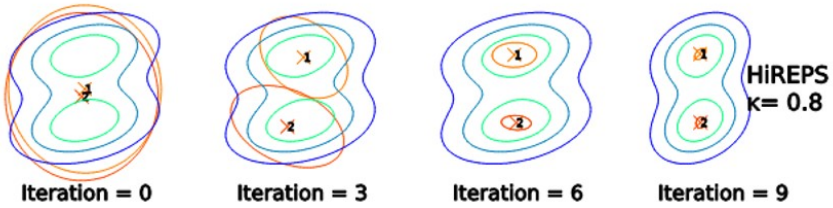


図 4 パラメトリック方策の組み合わせで多峰性を表現 [13]

1.2.2 画像を入力とする方策探索

次元の呪いを解決したものとして, 近年, ガウス過程やカーネル法などを用いたノンパラメトリック方策モデルの利用が提案されている [16,17]. これらの文献では, 従来のガウス過程方策を改良することで, 高次元入力に対して頑健な方策探索を提案しており, リーチングタスクや振り上げタスクのシミュレーションにより提案法の有効性を示している (図 5). ガウス過程を用いることで, カーネルトリックにより, 高次元特徴空間に基づく方策をリーズナブルな計算量で求めることができる. しかし, サンプル数に応じて計算量が増加するため, ロボットの制御周期の制限から, 大規模な問題への適用は困難となる. また, 方策モデルにガウス分布が仮定されるため, 最適解が複数存在する問題において, 単峰な方策モデルによる方策探索を行うと, 分散が大きくなり, 適切な方策モデルが求まらないといった問題がある.

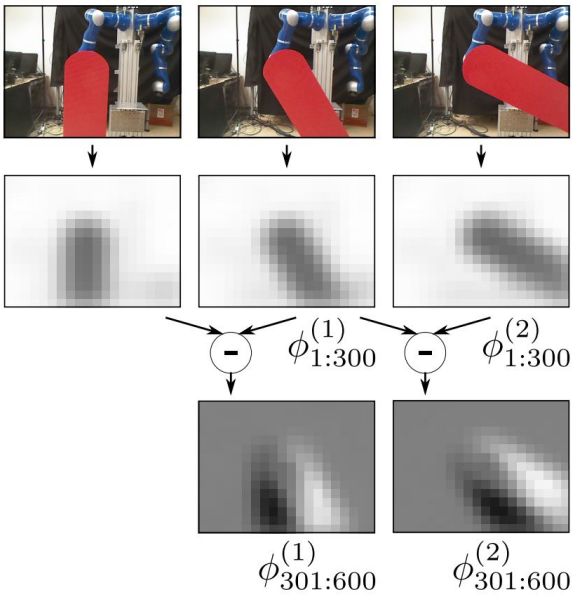
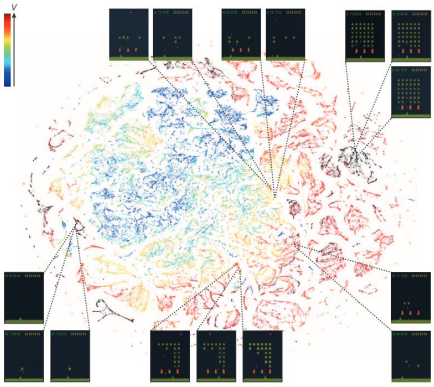


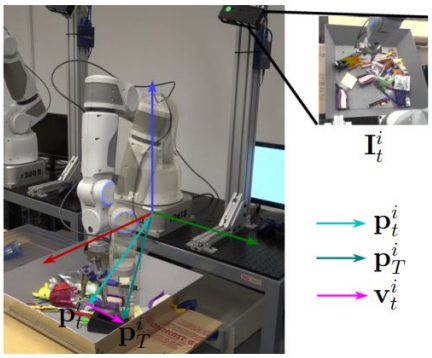
図 5　画像入力を用いた方策探索による研究 [16, 17]

1.2.3　画像を入力とする深層強化学習

方策探索以外の研究で, 高次元入力に頑健な学習としては, 深層強化学習に関する研究がある. 例えば, 文献 [18] では深層強化学習を用いて 49 種類の TV ゲームの操作を学習している. この研究では入力が画像のピクセルとゲームスコアであり, 画像から行動を学習している. 図 6 の左図にて, 深層強化学習によって学習したネットワークを可視化したものを示す. 29 種類の TV ゲームで人間と同等以上の操作を学習できたことが報告されている. しかし, 学習には 3〜5 日かけて 470 万回以上の操作経験データが必要なため, 学習にかかるコストが大きい. また, 文献 [19] ではロボットアームに備わっている単眼カメラ画像からロボットの把持動作を深層強化学習によって学習している. 図 6 の右図にてロボットが把持動作を学習している様子を示す. 6〜14 台のロボットが 3ヶ月学習することで, 把持動作を学習することができたが, 膨大な学習時間と資金が必要である.

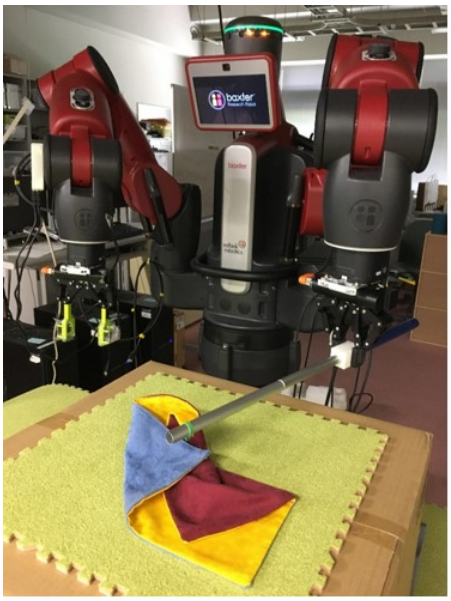


学習したネットワークの可視化 [18]



ロボットアームの把持動作学習 [19]

図 6　画像を入力とする深層強化学習



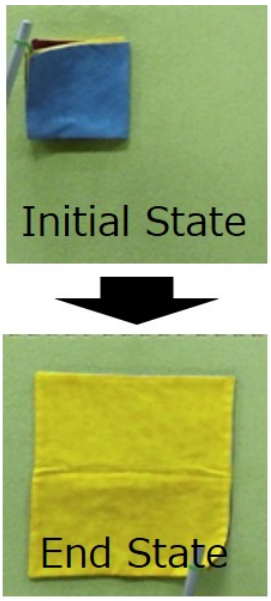


図 7　BAXTER による布展開タスク

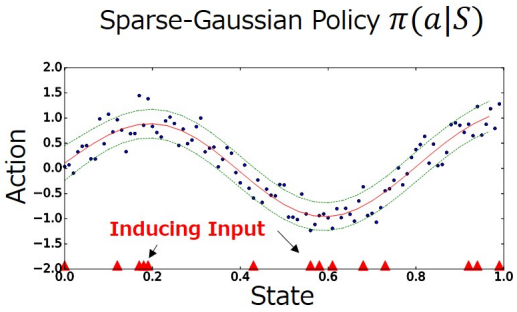


図 8 提案法 1:スパースガウス過程方策

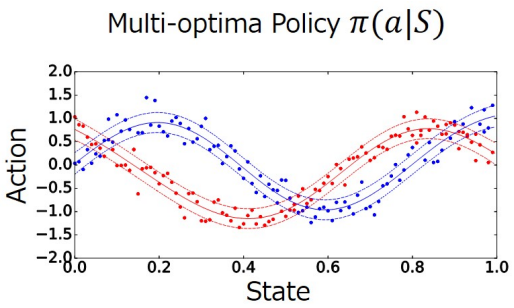


図 9 提案法 2:重複混合ガウス過程方策

1.3 研究目的

1.1 節で述べたように, 観測データの密度やばらつきに柔軟なガウス過程方策を用いた方策探索を提案する. ガウス過程方策はノンパラメトリック方策であるので, 入力の状態が高次元でも頑健な学習ができる. そこで, 図 7 に示すようなロボットを用いた, 方策探索による布操作タスクの実現を目標に, 提案法では画像を入力とする方策探索について考える. これは, 人が布を操作するときに, 視覚情報に基づいて操作していると考えたからである. 1.2 節で示したように方策探索には様々な従来研究がある. しかし, 従来法を用いてロボットによる布操作タスクの学習を行うと以下に示す 2 つの問題が主に考えられる.

1 つ目に, 得られる観測データを全て用いて方策探索を行うと, 計算量が観測データ数に大きく依存するという問題がある. 布の状態遷移は複雑なため, 学習には多くのデータが必要である. また, 画像を入力とする強化学習では, 入力が高次元なため, 通常の学習問題よりも多くの観測データが必要である. そのため, 計算量が観測データ数に大きく依存する従来法では, ロボットの行動の予測にかかる時間が長いために, ロボットの制御周期の制限で実用に適さない可能性がある.

2 つ目に, ある状態に対する複数の最適な行動候補が存在する状況で, 正しい行動の予測ができないという問題がある. ガウス過程方策では, 平均と分散を用いて方策モデルを作成する. これは得られる観測データが, ある 1 つの潜在関数からノイズを伴って得られたという仮定を事前においているからである. 例えば布を展開する操作においては, 同一の状態から複数の展開操作が考えられるため,

従来法では上手く学習することができない.

以上の 2 つの問題を解決するために図 8, 図 9 に示すような方策を用いた方策探索を提案する. 計算量を削減するために, 1 つ目の提案法として, 事前分布にスパースガウス過程 [20] を仮定した方策に対して, 変分推論 [21] に基づく近似 EM 方策探索を行う手法を提案する. 以降, このような方策をスパースガウス過程方策と表記する. スパースガウス過程によって, 得られた観測データから重要度の高い疑似入力を選ぶことができる. この疑似入力を用いることで, ロボットの行動の予測に必要な時間を短縮できる. また, スパースガウス過程方策は明示的なパラメータを持たないため, 画像入力のような入力次元数が大きい状況においても, 次元の呪いを回避することができる. 方策改善においては, 観測することができない全ての潜在変数を考慮した期待報酬値の最大化を行う必要がある. そこで本研究では, 変分推論に基づく近似 EM 方策探索によって方策探索を行った.

最適解が複数存在する問題において, 適切な方策モデルが求まらないという課題を解決するために, 2 つ目の提案法として, 事前分布に重複混合ガウス過程 [22] を仮定した方策に対して, 変分推論に基づく近似 EM 方策探索を行う手法を提案する. 以降, このような方策を重複混合ガウス過程方策と表記する. 重複混合ガウス過程を用いることで, 複数の潜在関数と観測データの関連性を, ハイパーパラメータ及び潜在関数を考慮した周辺尤度の最大化によって, 一意に求めることができる. これにより, 従来法で困難だった, 同じ状態に対して複数の行動をもつ方策モデルを表現することができ, 多峰性のある問題に対しても方策探索によって学習できる. 方策改善に関しては, スパースガウス過程方策同様, 潜在変数に対して周辺化した期待報酬値を最大化する必要があったため, 変分推論に基づく近似 EM 方策探索法を用いた.

本研究では, 2 つの提案法に対して以下の問題に対して実験を行い, それぞれの提案法の有効性を確認した.

- シミュレーション実験
  - 提案法 1：スパースガウス過程方策を用いた方策探索
    - 衛星画像入力によるリーチング問題
  - 提案法 2：重複混合ガウス過程方策を用いた多峰性方策探索
    - 1 段階スリット通り抜け問題
    - 3 段階スリット通り抜け問題
    - 衛星画像入力によるリーチング問題
- 実機実験
  - 提案法 1 を用いた布操作タスク実験

1.4 本論文の構成

本論文の構成について述べる．第 2 章では, 本研究の基礎知識となる強化学習について述べた後, 提案法に関連する手法について述べる．第 3 章では 2 つの提案法について詳しく述べる．第 4 章で提案法 1 によるシミュレーション実験の結果を示す．第 5 章で提案法 2 によるシミュレーション実験の結果を示す．第 6 章で提案法 1 による実機実験の結果を示す．

2. 準備

2.1 強化学習

強化学習とは状態, 行動, 方策, 報酬の 4 つの要素を用いて, 報酬を最大化する方策を求める. 本稿では, 状態 $\mathbf{s}_n$, 行動 a_n, 方策 $\pi(a_n|\mathbf{s}_n;\boldsymbol{\theta})$, 状態遷移モデルを $P(\mathbf{s}_{n+1}|\mathbf{s}_n, a_n)$ とおき, 方策を求める. $\boldsymbol{\theta}$ は方策のパラメータである. 状態 $\mathbf{s}_n$ と行動 a_n の時系列で表されるエピソード d と, d の生成確率 $P(d;\boldsymbol{\theta})$ はステップ数を T としたとき, 式 (1),(2) で表せる

$$d \equiv (\mathbf{s}_1, a_1, \mathbf{s}_2, a_2, \cdots, \mathbf{s}_T, a_T, \mathbf{s}_{T+1}) \tag{1}$$

$$P(d;\boldsymbol{\theta}) \equiv P(\mathbf{s}_1)\prod_{n=1}^{T}\pi(a_n|\mathbf{s}_n;\boldsymbol{\theta})P(\mathbf{s}_{n+1}|\mathbf{s}_n, a_n) \tag{2}$$

このとき, エピソード d での報酬和 $R(d)$ と, 期待報酬 $J(\boldsymbol{\theta})$ は割引率 γ を用いると次のように表せる.

$$R(d) \equiv \sum_{n=1}^{T}\gamma^{n-1}R(\mathbf{s}_n, a_n, \mathbf{s}_{n+1}) \tag{3}$$

$$J(\boldsymbol{\theta}) = \int R(d)P(d;\boldsymbol{\theta})dd \tag{4}$$

方策に注目した強化学習である方策探索では, 式 (5) に示すように, 期待報酬値 $J(\boldsymbol{\theta})$ を最大化するパラメータ $\boldsymbol{\theta}^*$ を求めることを目標とする.

$$\boldsymbol{\theta}^* \leftarrow \arg\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \tag{5}$$

2.2 ガウス過程回帰

ガウス過程回帰とは, 入力から出力を予測する回帰関数の確率モデルである. 入力ベクトルを $\mathbf{x} \in \mathbb{R}^d$, 出力値を z, 観測値を y とする. 式 (6) に示すように, 観測値は出力値にガウス分布に従ったノイズ σ^2 が加わったものとする.

$$p(y|z) = \mathcal{N}(y|z, \sigma^2) \tag{6}$$

ノイズは各データに対して独立に決まるため, $\mathbf{z} = [z_1, \cdots, z_N]^{\mathrm{T}}$ が与えられた下での観測ベクトル $\mathbf{y} = [y_1, \cdots, y_N]^{\mathrm{T}}$ の同時分布は式 (7) に示すようなガウス分布に従う.

$$p(\mathbf{y}|\mathbf{z}) = \mathcal{N}(\mathbf{y}|\mathbf{z}, \sigma^2\mathbf{I}_N) \tag{7}$$

ここで $\mathbf{I}_N$ は $N \times N$ の単位行列であり, N は観測データ点数である. ガウス過程の定義より, 周辺分布 $p(\mathbf{z})$ は式 (8) に示すような, 平均が 0, 共分散がカーネルグラム行列 $\mathbf{K}$ のガウス分布となる.

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{K}) \tag{8}$$

なお, カーネルグラム行列の要素 K_{ij} は式 (9) に示すようなカーネル関数で与えられる.

$$k(\mathbf{x}_i, \mathbf{x}_j) = \theta_1 \exp\left(-\frac{\theta_2}{2}||\mathbf{x}_i - \mathbf{x}_j||^2\right) + \theta_3 \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j \tag{9}$$

ただし, $\theta_1, \theta_2, \theta_3$ はカーネル関数のハイパーパラメータである. 式 (7), 式 (8) より, 周辺分布 $p(\mathbf{y})$ は $\mathbf{z}$ について周辺化することで求まり, 式 (10) のようになる.

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{z})p(\mathbf{z})\mathbf{dz} = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K} + \sigma^{\mathbf{2}}\mathbf{I}) \tag{10}$$

回帰においては, 訓練データ集合 $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]^{\mathrm{T}}$ および, $\mathbf{y} = [y_1, \cdots, y_N]^{\mathrm{T}}$ が与えられたもとでの, テスト入力 $\mathbf{x}_*$ に対する観測値 y_* の予測分布 $p(y_*|\mathbf{x}_*)$ を考える. この予測分布はベイズの定理を用いると, 式 (11) に示すような, 平均 $m(\mathbf{x}_*)$ および, 分散 $\sigma^2(\mathbf{x}_*)$ を持つガウス分布として求まる.

$$\begin{aligned} p(y_*|\mathbf{x}_*) &= \mathcal{N}(y_*|m(\mathbf{x}_*), \sigma^2(\mathbf{x}_*)) \\ m(\mathbf{x}_*) &= \mathbf{k}_*^{\mathrm{T}}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y} \\ \sigma^2(\mathbf{x}_*) &= k_{**} - \mathbf{k}_*^{\mathrm{T}}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{k}_* \end{aligned} \tag{11}$$

ここで, $\mathbf{k}_* = k(\mathbf{X}, \mathbf{x}_*), k_{**} = k(\mathbf{x}_*, \mathbf{x}_*), K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ である.

2.3 EM 方策探索

期待報酬値 $J(\boldsymbol{\theta})$ の最大化において, 式 (4) で示した積分計算を解くためには, 全てのパラメータ $\boldsymbol{\theta}$ に対してエピソード d で周辺化する必要があり, 計算困難である. そこで, 閉形式で解析的に計算するために, EM 方策探索 [11] を用いる. EM 方策探索では期待報酬値の下限を最大化することで, パラメータ $\boldsymbol{\theta}$ を逐次的に更新し, $J(\boldsymbol{\theta})$ を最大化する. 式 (4) において, $R(d)$ は非負であると仮定し, イェンセンの不等式 [23] を用いると, $J(\boldsymbol{\theta})$ の対数の下限を求める E-step は式 (12) のように導ける. なお, 方策の更新回数を L としている.

$$\begin{aligned}\log J(\boldsymbol{\theta}) &\geq \int \frac{R(d)P(d;\boldsymbol{\theta}_L)}{J(\boldsymbol{\theta}_L)} \log \frac{P(d;\boldsymbol{\theta})}{P(d;\boldsymbol{\theta}_L)} dd + \log J(\boldsymbol{\theta}_L) \\ &\equiv \log J_L(\boldsymbol{\theta})\end{aligned} \tag{12}$$

次に式 (13) で示すように, M-step で式 (12) を最大化するパラメータ $\boldsymbol{\theta}$ を求める.

$$\begin{aligned}\boldsymbol{\theta}_{L+1} &\equiv \arg\max_{\boldsymbol{\theta}} \log J_L(\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \int R(d)P(d;\boldsymbol{\theta}_L) \log \pi(\mathbf{a}|\mathbf{S};\boldsymbol{\theta}) dd \\ &\approx \frac{1}{M} \arg\max_{\boldsymbol{\theta}} \sum_{m=1}^{M} R_m \log \pi(\mathbf{a}_m|\mathbf{S}_m;\boldsymbol{\theta})\end{aligned} \tag{13}$$

ここで, M は $P(d;\boldsymbol{\theta}_L)$ から生成されるエピソード数である. R_m は $P(d;\boldsymbol{\theta}_L)$ から生成されたエピソードのうち, m 番目のエピソードの累積報酬値とした. 式 (12) において, $\boldsymbol{\theta} = \boldsymbol{\theta}_L$ のとき $\log J(\boldsymbol{\theta}_L) = \log J_L(\boldsymbol{\theta}_L)$ となり, 式 (12) で求めた下限は, 目的関数である $J(\boldsymbol{\theta})$ の対数に接する. これら E-step と M-step を交互に繰り返し, 方策は期待報酬値の単調非減少性を保証して更新される.

$$J(\boldsymbol{\theta}_{L+1}) \geq J(\boldsymbol{\theta}_L) \tag{14}$$

3. 提案法

3.1 提案法 1：スパースガウス過程方策を用いた方策探索

本節では，方策の事前分布にスパースガウス過程 [20] を仮定し, 変分推論 [21] に基づいた近似 EM 方策探索法を提案する．3.1.1 節でスパースガウス過程の定義について述べる．式 (13) で示した報酬の重みを考慮した方策モデルについて, 3.1.2 節でスパースガウス過程方策を用いた方策探索法について述べる．3.1.3 節で提案法に対して未知入力を与えたときの予測分布の導出について述べる.

3.1.1 スパースガウス過程方策モデル

式（13）に示した M ステップについて，スパースガウス過程 (SPGP) 方策の方策改善法について考える．疑似入出力を $\overline{\mathbf{S}} = [\overline{\mathbf{s}}_1, \cdots, \overline{\mathbf{s}}_M]^{\mathrm{T}}$, $\overline{\mathbf{f}} = [\overline{f}_1, \cdots, \overline{f}_M]^{\mathrm{T}}$ とする. 状態 $\mathbf{S} = [\mathbf{s}_1, \cdots, \mathbf{s}_N]^{\mathrm{T}}$ と行動 $\mathbf{a} = [a_1, \cdots, a_N]^{\mathrm{T}}$ からなる観測データのセットを $\mathcal{D} = \{\mathbf{S}, \mathbf{a}\}$ とする. $\mathbf{S}$ に対する方策のノイズ無し関数を $\mathbf{f} = [f_1, \cdots, f_N]^{\mathrm{T}}$ とする. このとき, スパースガウス過程方策モデルは, 疑似入力を考慮した方策モデルであるため, 報酬重み付き対数尤度関数は解析的な計算が困難である. そこで, 式 (15) で示すような近似を導入する.

$$R \log \pi(\mathbf{a}|\mathbf{S}, \overline{\mathbf{S}}) \approx \log \int p(\mathbf{a}|\mathbf{f})^R p(\mathbf{f}|\overline{\mathbf{f}}, \mathbf{S}, \overline{\mathbf{S}}) p(\overline{\mathbf{f}}|\overline{\mathbf{S}}) d\mathbf{f} d\overline{\mathbf{f}} \tag{15}$$

ここで，$p(\overline{\mathbf{f}}|\overline{\mathbf{S}}) = \mathcal{N}(\overline{\mathbf{f}}|0, \mathbf{K}_M)$ は疑似出力に対するガウス過程事前分布，$\mathbf{K}_M$ は疑似入力 $\bar{\mathbf{S}}$ に関するカーネルグラム行列，残りの項は

$$\begin{aligned} p(\mathbf{a}|\mathbf{f})^R &= \prod_{n=1}^{N} p(a_n|f_n)^{R_n} \\ p(a_n|f_n) &= \mathcal{N}(a_n; f_n, \sigma^2 \mathbf{I}) \end{aligned} \tag{16}$$

$$\begin{aligned} p(\mathbf{f}|\overline{\mathbf{f}}, \mathbf{S}, \overline{\mathbf{S}}) &= \prod_{n=1}^{N} p(f_n|\mathbf{s}_n, \overline{\mathbf{f}}, \overline{\mathbf{S}}) \\ p(f_n|\mathbf{s}_n, \overline{\mathbf{f}}, \overline{\mathbf{S}}) &= \mathcal{N}(f_n|\mathbf{k}_n^{\mathrm{T}} \mathbf{K}_M^{-1} \overline{\mathbf{f}}, K_{nn} - \mathbf{k}_n^{\mathrm{T}} \mathbf{K}_M^{-1} \mathbf{k}_n) \end{aligned} \tag{17}$$

である．ここで $K_{nn} = \mathbf{k}(\mathbf{s}_n, \mathbf{s}_n)$, $\mathbf{k}_n = \mathbf{k}(\mathbf{s}_n, \bar{\mathbf{S}})$ とした．また σ^2 は加法ノイズである．このような方策モデルを用いる利点は，通常のガウス過程回帰と比べて，$M << N$ のときに，未知入力に対する出力の予測に要する計算量が削減されることにある．しかしながら，周辺尤度関数を疑似入力 $\overline{\mathbf{S}}$(およびハイパーパラメータ，ただし本稿では対象外) に関して直接最大化することは困難である．よって，次節において，変分推論 [21] を用いた近似最大化法を検討する．

3.1.2 変分推論による SPGP 方策改善

式 (15) において, 適当な汎関数 $q(\mathbf{f}, \overline{\mathbf{f}})$ を仮定すると，一般に以下のような分解が成り立つ．

$$\log \pi^R(\mathbf{a}|\mathbf{S}, \overline{\mathbf{S}}) = F(q, \overline{\mathbf{S}}) + KL(q, \overline{\mathbf{S}}) \tag{18}$$

ただし,

$$\begin{aligned} F(q, \overline{\mathbf{S}}) &= \int q(\mathbf{f}, \overline{\mathbf{f}}) \log \frac{p(\mathbf{f}, \overline{\mathbf{f}}, \mathbf{a})^R}{q(\mathbf{f}, \overline{\mathbf{f}})} d\mathbf{f} d\overline{\mathbf{f}} \\ KL(q, \overline{\mathbf{S}}) &= \int q(\mathbf{f}, \overline{\mathbf{f}}) \log \frac{q(\mathbf{f}, \overline{\mathbf{f}})}{p(\mathbf{f}, \overline{\mathbf{f}}|\mathbf{a})^R} d\mathbf{f} d\overline{\mathbf{f}} \end{aligned} \tag{19}$$

とした[1]．ここで, KL は非負であるので, KL の最小化は F の最大化と同義である．イェンセンの不等式と積分の解析計算 (付録 A) より，F を最大化する最適な q^* を代入した F_V は次式のように導出される．

$$F_V(\overline{\mathbf{S}}) = \log[\mathcal{N}(\tilde{\mathbf{a}}|0, \sigma^2\mathbf{I} + \tilde{\mathbf{Q}}_N)] - \frac{1}{2\sigma^2}\mathrm{Tr}(\mathbf{K}_N - \mathbf{Q}_N) \tag{20}$$

ただし, $\mathbf{Q}_N = \mathbf{K}_{NM}\mathbf{K}_M^{-1}\mathbf{K}_{MN}$, $\tilde{\mathbf{Q}}_N = \tilde{\mathbf{K}}_{NM}\mathbf{K}_M^{-1}\tilde{\mathbf{K}}_{MN}$, $\tilde{\mathbf{K}}_{NM} = \mathbf{K}_{NM}\mathbf{W}$, $\mathbf{K}_N = \mathbf{K}_{NN}$, $\tilde{\mathbf{K}}_N = \mathbf{W}\mathbf{K}_{NN}\mathbf{W}$, $\tilde{\mathbf{a}} = \mathbf{W}\mathbf{a}$ とした．$\mathbf{W}$ は $\sqrt{R_i}(i = 1, \cdots, N)$ を対角成分に持つ対角行列である．この後, 一般的な変分推論では勾配法などを適用し F_V を最大化する疑似入力 $\overline{\mathbf{S}}^*$ を求めることで，方策改善を行う．しかし, 本研究では, 入力が画像のため, 求める疑似入力 $\overline{\mathbf{S}}^*$ も画像である．そのため, 入力画像 $\mathbf{S}$ のうち

[1] 自明な変数は省略した

最も F_V を最大化する画像入力を疑似入力画像 $\overline{\mathbf{S}}^*$ として選択する手法をとった. なお，$\mathbf{W} = \mathbf{I}$ のとき，上記の計算は通常の変分スパースガウス過程回帰と等価となる [21]. F_V を最大化するときに必要な対数尤度の計算において $\sigma^2\mathbf{I} + \tilde{\mathbf{Q}}_N$ の逆行列の計算は行列の公式を利用して高速化を行った (付録 B).

3.1.3 SPGP 方策の予測分布

スパースガウス過程では尤度が次式 (21) で与えられる.

$$p(a|\mathbf{s}, \overline{\mathbf{f}}, \overline{\mathbf{S}}) = \mathcal{N}(a|\mathbf{k}_n^{\mathrm{T}}\mathbf{K}_M^{-1}\overline{\mathbf{f}}, K_{nn} - \mathbf{k}_n^{\mathrm{T}}\mathbf{K}_M^{-1}\mathbf{k}_n) \tag{21}$$

よって, 状態 $\mathbf{S}$ に対する行動 $\mathbf{a}$ の確率は次のようになる.

$$\begin{aligned} p(\mathbf{a}|\mathbf{S}, \overline{\mathbf{f}}, \overline{\mathbf{S}}) &= \prod_{n=1}^{N} p(a_n|\mathbf{s}_n, \overline{\mathbf{f}}, \overline{\mathbf{S}}) \\ &= \mathcal{N}(\mathbf{a}|\mathbf{K}_{NM}\mathbf{K}_M^{-1}\overline{\mathbf{f}}, \mathbf{\Lambda} + \sigma^2\mathbf{I}) \end{aligned} \tag{22}$$

ただし, $\mathbf{\Lambda} = \mathrm{diag}\,(\boldsymbol{\lambda}), \lambda_n = K_{nn} - \mathbf{k}_n^T\mathbf{K}_M^{-1}\mathbf{k}_n$ とした. $\mathbf{K}_{NM}$ は入力の状態 $\mathbf{S}$ と疑似入力 $\bar{\mathbf{S}}$ に関するカーネルグラム行列である. 事後分布は式 (23) のようになる.

$$p(\overline{\mathbf{f}}|\mathcal{D}, \overline{\mathbf{S}}) = \mathcal{N}(\overline{\mathbf{f}}|\mathbf{K}_M\mathbf{Q}_M^{-1}\mathbf{K}_{MN}(\mathbf{\Lambda} + \sigma^2\mathbf{I})^{-1}\mathbf{a}, \mathbf{K}_M\mathbf{Q}_M^{-1}\mathbf{K}_M) \tag{23}$$

ただし

$$\mathbf{Q}_M = \mathbf{K}_M + \mathbf{K}_{MN}(\mathbf{\Lambda} + \sigma^2\mathbf{I})^{-1}\mathbf{K}_{NM} \tag{24}$$

とした. 式 (23) より, 疑似入力が報酬重みを考慮した $\overline{\mathbf{S}}^*$ であるならば, 未知入力 $\mathbf{s}_*$ に対応する出力 a_* の予測分布はガウス分布の積の公式 (付録 C) より, 式 (25) のようになる.

$$\begin{aligned} p(a_*|\mathbf{s}_*, \mathcal{D}, \overline{\mathbf{S}}) &= \int p(a_*|\mathbf{s}_*, \overline{\mathbf{f}}, \overline{\mathbf{S}}) p(\overline{\mathbf{f}}|\mathcal{D}, \overline{\mathbf{S}}) d\overline{\mathbf{f}} = \mathcal{N}(a_*|\mu_*, \sigma_*^2), \\ \mu_* &= \mathbf{k}_*^{\mathrm{T}}\tilde{\mathbf{Q}}_M^{-1}\tilde{\mathbf{K}}_{MN}(\mathbf{\Lambda} + \sigma^2\mathbf{I})^{-1}\tilde{\mathbf{a}} \\ \sigma_*^2 &= K_{**} - \mathbf{k}_*^{\mathrm{T}}(\mathbf{K}_M^{-1} - \tilde{\mathbf{Q}}_M^{-1})\mathbf{k}_* + \sigma^2 \end{aligned} \tag{25}$$

ここで $\mathbf{k}_* = \mathbf{k}(\mathbf{s}_*, \overline{\mathbf{S}}^*)$ である. また, $\tilde{\mathbf{Q}}_M = \mathbf{K}_M + \tilde{\mathbf{K}}_{MN}(\mathbf{\Lambda} + \sigma^2\mathbf{I})^{-1}\tilde{\mathbf{K}}_{NM}$ とする. なお, $\mathbf{\Lambda} + \sigma^2\mathbf{I}$ は対角行列のため, 式 (24) で示した $\mathbf{Q}_M$ の計算量は行列積に限定され $\mathcal{O}(M^2N)$ である. 平均と分散の計算量はそれぞれ $\mathcal{O}(M), \mathcal{O}(M^2)$ となる.

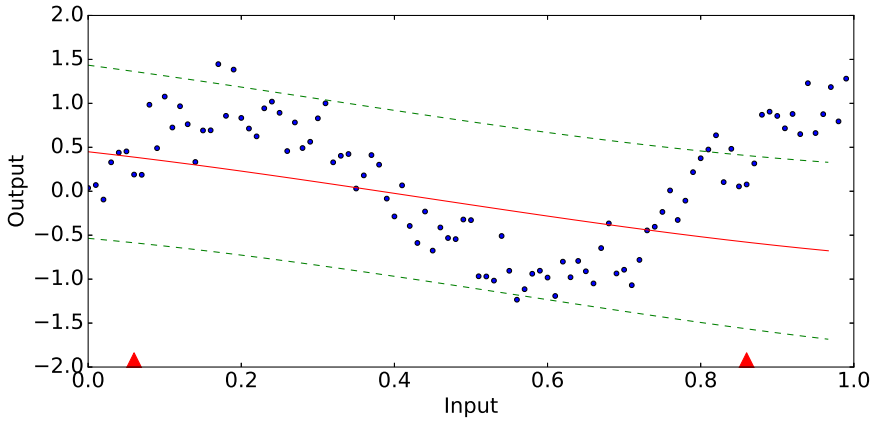


図 10　疑似入力を 2 点選んだときの提案法 1 による予測分布.

3.1.4　回帰問題による動作確認

二次元の観測データ 100 点に対して, 提案法 1 で方策モデルを学習した. 報酬重み付き観測データを考慮した疑似入力の選択が行えるか検証した. 観測データに対して以下のような報酬重みを付加し, 提案法 1 による回帰を行った.

$$R = \begin{cases} 1 & (\text{入力} < 0.7) \\ 10^{-2} & (\text{入力} \geq 0.7) \end{cases} \tag{26}$$

図 10 に疑似入力を 2 点選んだ時の提案法 2 の予測分布を示す. 赤の実線が予測されたガウス分布の平均, 緑の破線が標準偏差である. 図の下部にある赤の三角は疑似入力を表す. 分散が大きく, 平均も観測データの分布に一致していないことは明らかである. よって 2 点の疑似入力では適切な方策モデルを得られないことを確認した. 図 11 に疑似入力を 15 点選んだ時の提案法 2 の予測分布を示す. 選ばれた疑似入力の数は 0.7 以上が 1 点, 0.7 未満は 14 点であった. よって, 0.7 未満の報酬重みが高い観測データを重視した疑似入力の選択が行えていることを確認した. また, 予測分布に関しても入力値 0.7 以上では平均が観測データから大きく外れていることが分かる. これは入力が 0.7 以上の観測データの報酬重みが小さいからである. 一方, 入力が 0.7 未満の範囲では観測データの重みが大きいため, 予測分布の平均と分散によって, 観測データの分布が適切に表現できていることが分かる.

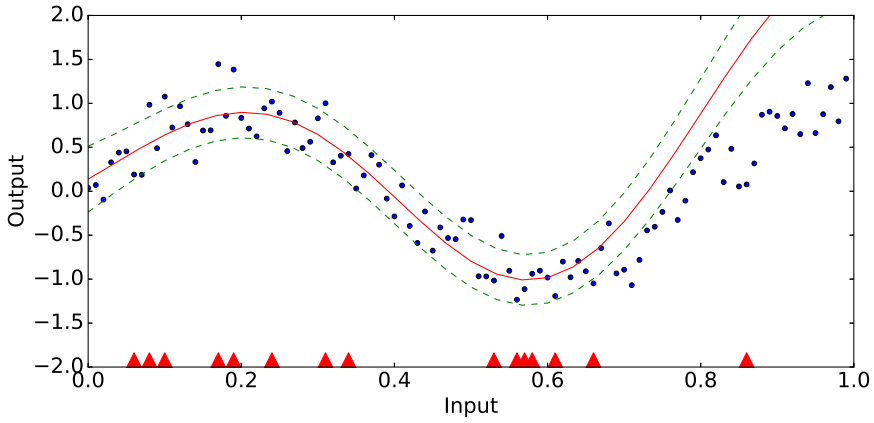


図 11　疑似入力を 15 点選んだときの提案法 1 による予測分布

3.2　提案法 2：重複混合ガウス過程方策を用いた多峰性方策探索

本節では，方策の事前分布に重複混合ガウス過程 [22] を仮定し, 変分推論 [21] に基づいた近似 EM 方策探索法を提案する．3.2.1 節で重複混合ガウス過程方策の定義について述べる．式 (13) で示した報酬の重みを考慮した方策モデルについて, 3.2.2 節で重複混合ガウス過程方策を用いた方策探索法について述べる．3.2.3 節で提案法に対して未知入力を与えたときの予測分布の導出について述べる．

3.2.1　重複混合ガウス過程方策モデル

重複混合ガウス過程 (OMGP) を用いた強化学習について考える．N 点の観測データ (行動 a) が, 状態 $\mathbf{s}$ を入力とする M 本の潜在関数 $\{f^{(m)}(\mathbf{s})\}_{m=1}^{M}$ にガウス性ノイズ ($\mathcal{N}(0,\sigma^2)$) を伴った結果, 独立に観測されたと仮定する．その仮定の下, 得られた観測データから潜在関数を推定する．観測データと潜在関数の関係性を表す指示行列として大きさ $N \times M$ の行列 $\mathbf{Z}$ を導入する．これは, バイナリー行列であり, 行列のそれぞれの要素は 1 か 0 の値を取る．例えば, n 番目の観測データが m 番目の潜在関数から得られたとみなすとき, 行列 $\mathbf{Z}$ の n 行 m 列目の要素が 1 となる．また, 行列 Z のそれぞれの要素を選択する確率値として大きさ $N \times M$ の行列 $\mathbf{\Pi}$ を導入する ($\sum_{m=1}^{M}[\mathbf{\Pi}]_{nm} = 1\ \forall n$)．この行列 $\mathbf{\Pi}$ によって, それぞれの観測

データの生成確率が求まる.

数式の簡略化のため, 本稿では N 点の行動データを $\mathbf{a} = [a_1 \cdots a_N]^{\mathrm{T}}$, 状態データを $\mathbf{S} = [\mathbf{s}_1 \cdots \mathbf{s}_N]^{\mathrm{T}}$, m 番目の潜在関数を $\mathbf{f}^{(m)} = [f_1^{(m)} \cdots f_N^{(m)}]^{\mathrm{T}}$ とする. そして, 潜在関数を全てまとめたものを $\{\mathbf{f}^{(m)}\}$ とする. 上記の定義を踏まえて, 事前分布を式 (27) で定義する. $\mathbf{K}$ はカーネルグラム行列であり, また, m 番目の潜在関数に対応するカーネルを $\mathbf{K}^{(m)}$, パラメータを $\boldsymbol{\theta}^{(m)}$ と表記する.

$$
\begin{aligned}
p(\mathbf{Z}) &= \prod_{n=1,m=1}^{N,M} [\mathbf{\Pi}]_{nm}^{[\mathbf{Z}]_{nm}} \\
p(\{\mathbf{f}^{(m)}\}|\mathbf{S};\boldsymbol{\theta}^{(m)}) &= \prod_{m=1}^{M} \mathcal{N}(\mathbf{f}^{(m)}|\mathbf{0},\mathbf{K}^{(m)})
\end{aligned}
\tag{27}
$$

また，重複混合ガウス過程の尤度関数は式 (28) で定義される.

$$
p(\mathbf{a}|\{\mathbf{f}^{(m)}\},\mathbf{Z}) = \prod_{n=1,m=1}^{N,M} \mathcal{N}([\mathbf{a}]_n|[\mathbf{f}^{(m)}]_n,\sigma^2)^{[\mathbf{Z}]_{nm}} \tag{28}
$$

3.2.2 OMGP 方策モデルの方策探索

式 (13) に基づいて，OMGP 方策モデルの方策探索における E-step として，報酬重み付き尤度関数を用いた周辺尤度関数を導出する．次に，M-step として，この周辺尤度関数を用いて，ガウス過程のハイパーパラメータを最適化する. まず，報酬重み付き尤度関数を次のように定義する.

$$
\begin{aligned}
\prod_{n=1}^{N} p(a_n|f_n^{(m)},\mathbf{Z})^{\mathbf{r}_n} &= \prod_{n=1}^{N} \mathcal{N}(a_n|f_n^{(m)},\sigma^2)^{[\mathbf{Z}]_{nm}\mathbf{r}_n} \\
&= \prod_{n=1}^{N} \mathcal{N}(w_n a_n|w_n f_n^{(m)},\sigma^2)^{[\mathbf{Z}]_{nm}} \\
&= \prod_{n=1}^{N} \mathcal{N}(\tilde{\mathbf{a}}|\tilde{\mathbf{f}}^{(m)},\sigma^2\mathbf{I})^{[\mathbf{Z}]_{nm}}
\end{aligned}
\tag{29}
$$

ここで，$\tilde{\mathbf{a}} = \mathbf{W}\mathbf{a}$, $\tilde{\mathbf{f}}^{(m)} = \mathbf{W}\mathbf{f}^{(m)}$, $\mathbf{W}$ は式 (30) に示すような大きさ $N \times N$ の対角行列である．また, 大きさ $T \times T$ の単位行列を $\mathbf{I}_{T\times T}$ とする．

$$\mathbf{W} = \begin{pmatrix} \sqrt{R(1)}\mathbf{I}_{T\times T} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \sqrt{R(D)}\mathbf{I}_{T\times T} \end{pmatrix} \tag{30}$$

式 (29) では $\mathbf{W}$ の対角要素を $w_n (n = 1, \cdots, N)$ としている．次に，この尤度関数を用いて行動系列の周辺尤度を考える．しかし, OMGP 方策モデルでは周辺尤度の解析的な計算が困難であるので, 式 (31) で示すような近似を導入する．

$$\begin{aligned} R\log\pi(\mathbf{a}|\mathbf{S};\theta^{(m)}) &= \log\left(\int p(\mathbf{a}|\mathbf{f}^{(m)},\mathbf{Z})p(\mathbf{Z})\prod_{m=1}^{M} p(\mathbf{f}^{(m)}|\mathbf{S};\boldsymbol{\theta}^{(m)})d\{\mathbf{f}^{(m)}\}d\mathbf{Z}\right)^{R} \\ &\approx \log\int p(\mathbf{a}|\mathbf{f}^{(m)},\mathbf{Z})^{R}p(\mathbf{Z})\prod_{m=1}^{M} p(\mathbf{f}^{(m)}|\mathbf{S};\boldsymbol{\theta}^{(m)})d\{\mathbf{f}^{(m)}\}d\mathbf{Z} \end{aligned} \tag{31}$$

ここで, 報酬の重みを付加した方策の確率モデル $\tilde{p}(\mathbf{a}|\mathbf{S};\boldsymbol{\theta}^{(m)})$ を式 (32) のように定義する．

$$\log\tilde{p}(\mathbf{a}|\mathbf{S};\boldsymbol{\theta}^{(m)}) = \log\int p(\tilde{\mathbf{a}}|\tilde{\mathbf{f}}^{(m)},\mathbf{Z})p(\mathbf{Z})\prod_{m=1}^{M} p(\tilde{\mathbf{f}}^{(m)}|\mathbf{S};\boldsymbol{\theta}^{(m)})d\{\tilde{\mathbf{f}}^{(m)}\}d\mathbf{Z} \tag{32}$$

ただし, $p(\tilde{\mathbf{f}}^{(m)}|\mathbf{S};\boldsymbol{\theta}^{(m)}) = \mathcal{N}(\tilde{\mathbf{f}}^{(m)}|0,\tilde{\mathbf{K}}^{(m)})$, $\tilde{\mathbf{K}}^{(m)} = \mathbf{W}\mathbf{K}^{(m)}\mathbf{W}^{\mathrm{T}}$ とする．式 (31) で示した, 近似後の周辺尤度関数は閉形式として解析的に得られないため，方策モデルに関する最大化は困難である．そこで，変分推論の枠組みを用いて下界を閉形式として導出する．OMGP の学習手法を参考に，まず E-step として，適当な汎関数 $q := q(\{\tilde{\mathbf{f}}^{(m)}\},\mathbf{Z})$ を用いて, 式 (31) を変分推論により最大化する．

$$\begin{aligned} \log\tilde{p}(\mathbf{a}|\mathbf{S};\boldsymbol{\theta}^{(m)}) &\geq \int q\log\frac{p(\tilde{\mathbf{a}}|\tilde{\mathbf{f}}^{(m)},\mathbf{Z})p(\mathbf{Z})\prod_{m=1}^{M} p(\tilde{\mathbf{f}}^{(m)}|\mathbf{S};\boldsymbol{\theta}^{(m)})}{q}d\{\tilde{\mathbf{f}}^{(m)}\}d\mathbf{Z} \\ &= \tilde{\mathcal{L}}(q;\boldsymbol{\theta}^{(m)}) \end{aligned} \tag{33}$$

等号成立条件は$q = p(\mathbf{Z}, \{\tilde{\mathbf{f}}^{(m)}\}|\mathbf{S}, \tilde{\mathbf{a}})$である. この$\tilde{\mathcal{L}}$の最大化によって, 式(29)で示した対数尤度関数を最大化する. ここで, $q = q(\{\tilde{\mathbf{f}}^{(m)}\})q(\mathbf{Z})$と分解し, それぞれを交互に最大化することで$\tilde{\mathcal{L}}$を最大にし, 対数尤度関数を最大化する. $q(\{\tilde{\mathbf{f}}^{(m)}\}), q(\mathbf{Z})$は式(34), 式(35)で求まる.

$$
\begin{aligned}
q(\mathbf{Z}) &= \prod_{n=1,m=1}^{N,M} [\hat{\mathbf{\Pi}}]_{nm}^{[\mathbf{Z}]_{nm}} \quad \text{with } [\hat{\mathbf{\Pi}}]_{nm}^{[\mathbf{Z}]_{nm}} \propto [\mathbf{\Pi}]_{nm} \exp(g_{nm}) \\
g_{nm} &= -\frac{1}{2}\left(\frac{([\tilde{\mathbf{a}}]_n - [\boldsymbol{\mu}^{(m)}]_n)^2 + [\mathbf{\Sigma}^{(m)}]_{nn}}{\sigma^2} + \log(2\pi\sigma^2)\right)
\end{aligned}
\tag{34}
$$

$$
\begin{aligned}
q(\tilde{\mathbf{f}}^{(m)}) &= \mathcal{N}(\tilde{\mathbf{f}}^{(m)}|\boldsymbol{\mu}^{(m)}, \mathbf{\Sigma}^{(m)}) \\
\mathbf{\Sigma}^{(m)} &= ((\tilde{\mathbf{K}}^{(m)})^{-1} + \mathbf{B}^{(m)})^{-1}, \boldsymbol{\mu}^{(m)} = \mathbf{\Sigma}^{(m)}\mathbf{B}^{(m)}\tilde{\mathbf{a}}
\end{aligned}
\tag{35}
$$

ただし, $\mathbf{B}^{(m)}$は要素$[\hat{\mathbf{\Pi}}]_{1m}/\sigma^2 \cdots [\hat{\mathbf{\Pi}}]_{Nm}/\sigma^2$を対角成分にもつ対角行列である.

次に，M-stepとして，上記の学習によって最大化された$\tilde{\mathcal{L}}$について，OMGP方策のハイパーパラメータを最適化する．本稿では文献[22]の最適化手法を実装し, ガウスカーネルのハイパーパラメータのみを最適化する.

式（29）において，$\mathbf{W} = \mathbf{I}$ ($\mathbf{I}$は単位行列)の場合には，上記のアルゴリズムは，入力を状態，出力を行動とする，OMGPの変分学習アルゴリズムと等価になる．報酬の導入方法を示したことが，本アルゴリズムの貢献である．OMGP方策探索をまとめると以下の2ステップを収束するまで繰り返す.

E-step: $\hat{q} \leftarrow \arg\max_q \tilde{\mathcal{L}}(q, \hat{\boldsymbol{\theta}}^{(m)})$ (式(34)および式(35))

M-step: $\hat{\boldsymbol{\theta}}^{(m)} \leftarrow \arg\max_{\boldsymbol{\theta}^{(m)}} \tilde{\mathcal{L}}(\hat{q}, \boldsymbol{\theta}^{(m)})$ (付録D)

3.2.3 OMGP 方策の予測分布

提案法によって学習された $p(\mathbf{a}|\mathbf{S};\boldsymbol{\theta})$ について, 未知入力 $\mathbf{s}_*$ に対する出力 a_* の予測分布 $\pi(a^*|\mathbf{s}^*,\tilde{\mathbf{a}},\mathbf{S})$ を考える. OMGP では m 個の予測分布が得られるが, 本手法では式 (36) の基準で, そのうちの 1 つを選択する.

$$\hat{m} = \arg\min_m \sigma_*^{(m)} \tag{36}$$

選択した $\hat{m}$ について, 予測分布は式 (37), 式 (38) で求まる.

$$\pi(a_*|\mathbf{s}_*,\tilde{\mathbf{a}},\mathbf{S}) = \mathcal{N}(a_*|\mu_*^{(\hat{m})},\sigma_*^{(\hat{m})}) \tag{37}$$

$$\begin{aligned}\mu_*^{(\hat{m})} &= (\tilde{\mathbf{k}}_*^{(\hat{m})})^{\mathrm{T}}(\tilde{\mathbf{K}}^{(\hat{m})} + (\mathbf{B}^{(\hat{m})})^{-1})^{-1}\tilde{\mathbf{a}} \\ \sigma_*^{(\hat{m})} &= \sigma^2 + k_{**}^{(\hat{m})} - (\tilde{\mathbf{k}}_*^{(\hat{m})})^{\mathrm{T}}(\tilde{\mathbf{K}}^{(\hat{m})} + (\mathbf{B}^{(\hat{m})})^{-1})^{-1}\tilde{\mathbf{k}}_*^{(\hat{m})}\end{aligned} \tag{38}$$

ただし, $\tilde{\mathbf{k}}_*^{(\hat{m})} = \mathbf{W}\mathbf{k}_*^{(\hat{m})}$ としている. また, 加法ノイズを σ^2 としている. しかし, この行動選択では行動選択による未知の方策の探索を十分に行えていないと判断し, 提案法 2 の 3 段階スリット問題によるシミュレーション実験では, 次式 (39) のソフトマックス関数を導入した行動選択を行った.

$$p(m) = \frac{\exp(-\beta_m/\tau)}{\sum_{m=1}^{M}\exp(-\beta_m/\tau)} \tag{39}$$

ただし, $\beta_m = \sigma_*^{(m)}$ であり, τ は温度変数である. $p(m) \in \{0,1\}$ は m の選択確率であり, 選択された m を $\hat{m}$ と表記すると, 方策は次式 (40) で表せる.

$$\pi(a_*|\mathbf{s}_*) = \mathcal{N}(a_*|\mu_*^{(\hat{m})},\sigma_*^{(\hat{m})}) \tag{40}$$

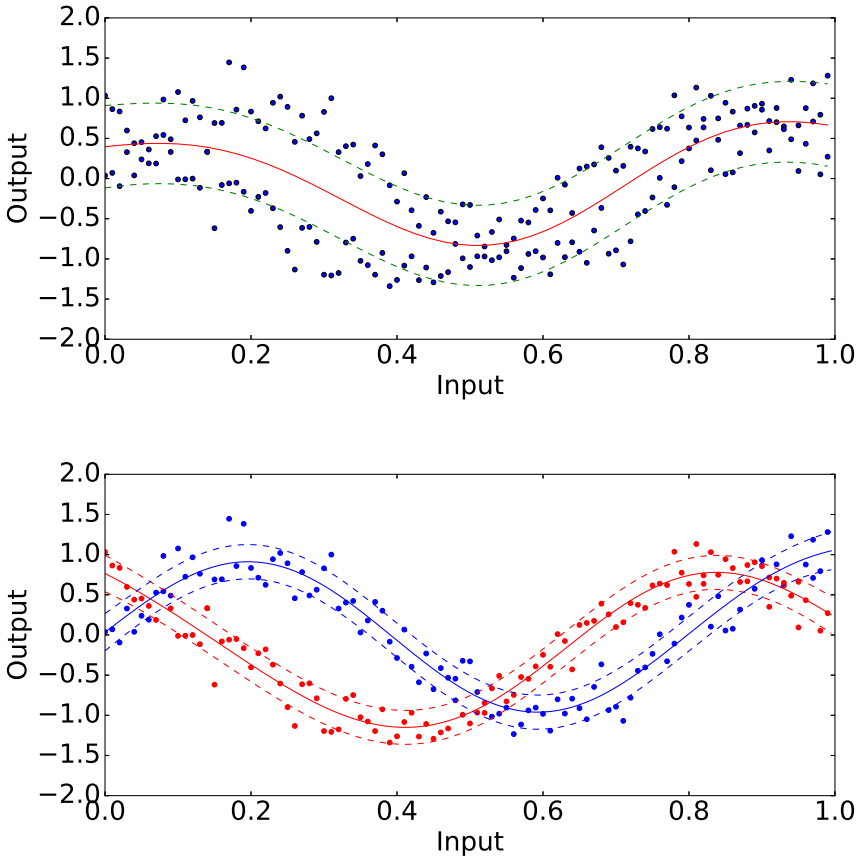


図 12　従来法 (上段) と提案法 2(下段) から得た方策の予測分布

3.2.4　回帰問題による動作確認 1

二次元の観測データ 100×2 点に対して, 従来法と提案法 2 で方策モデルを学習した. 報酬重み付き観測データを考慮した方策が得られるか検証した. 観測データに付加する報酬重みは全て 1 とし, 混合数は $M = 2$ とした.

図 12 に従来法による回帰結果を示す. 図 12 の上段において, 青点は観測データである. 赤線は得られた予測分布の平均であり, 緑線は標準偏差である. 図 12 より, 従来法では観測データの分布を表現するために, 大きな分散が発生しており, 観測データの分布と予測分布が合致していないことは明らかである.

提案法 2 による回帰結果を図 12 の下段に示す. 混合数 M を 2 としているため,

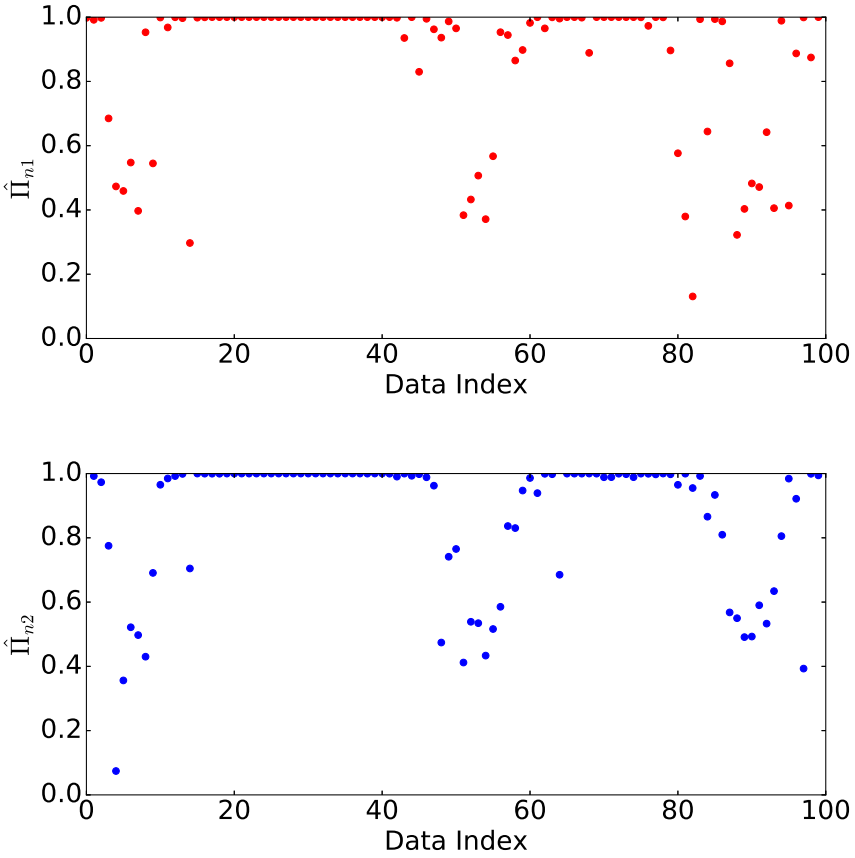


図 13　提案法 2 から得られた確率値 $\hat{\mathbf{\Pi}}$

予測分布の平均と分散もそれぞれ 2 種類ずつ得られた. 得られた 2 つの予測分布を赤線, 青線で示した. 図 12 より, 提案法 2 では, 観測データの分布を従来法より正しく表現しているといえる. また, 図 13 に指示多項分布 $\hat{\mathbf{\Pi}}$ に基づいて, 観測データの確率値を示した. 図 13 より, 図 12 における予測分布の平均が交差する 2 地点以外では $\hat{\mathbf{\Pi}}$ の値が 1 となり, 確実性が高い分類ができていることが分かる.

3.2.5 回帰問題による動作確認 2

二次元の観測データ 100×4 点に報酬重みを付加したときの提案法 2 の有効性を確認した. 観測データには異なる報酬重みを付加した. 図 14 に示すように, 高い報酬重みを付加した観測データを赤色, 青色で示す. 低い報酬重みを付加した観測データを水色, 紫色で示す. 実際に用いた報酬値は以下の値である.

$$R = \begin{cases} 1 & (\text{赤色と青色の観測データ}) \\ 10^{-3} & (\text{水色と紫色の観測データ}) \end{cases} \tag{41}$$

図 14 に示すような観測データに対して, 提案法 2 による予測を行った結果を図 15 に示す. 図 15 より, 報酬重みの高い赤色と青色の観測データを重視した, 予測分布が得られていることが分かる. 報酬重みの低い水色と紫色の観測データは予測分布に対する影響が低いことから, 提案法 2 によって報酬重みを考慮した多峰な方策を学習できているといえる.

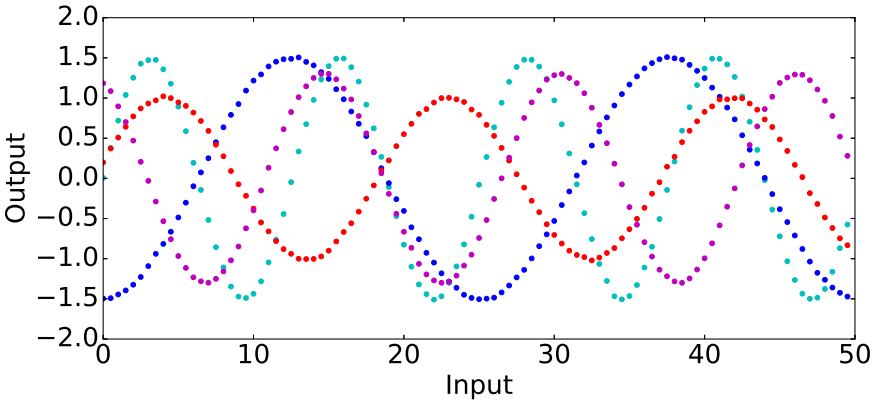


図 14　一次元の入力に対して, 4 種類の異なる出力を伴う観測データ.

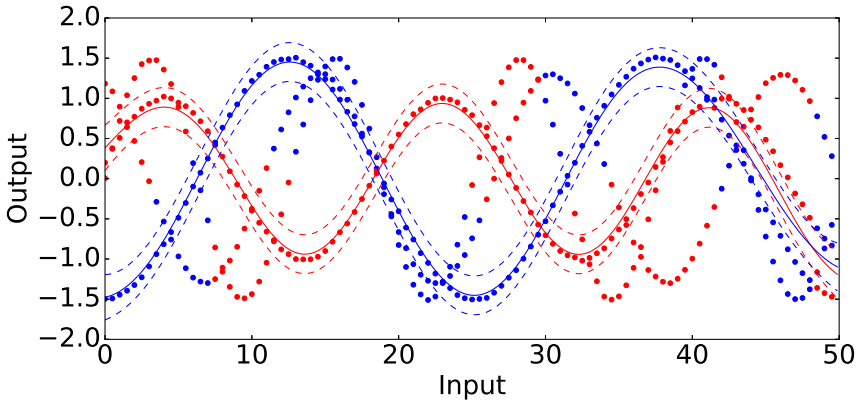


図 15　提案法 2 の方策から得られた予測分布

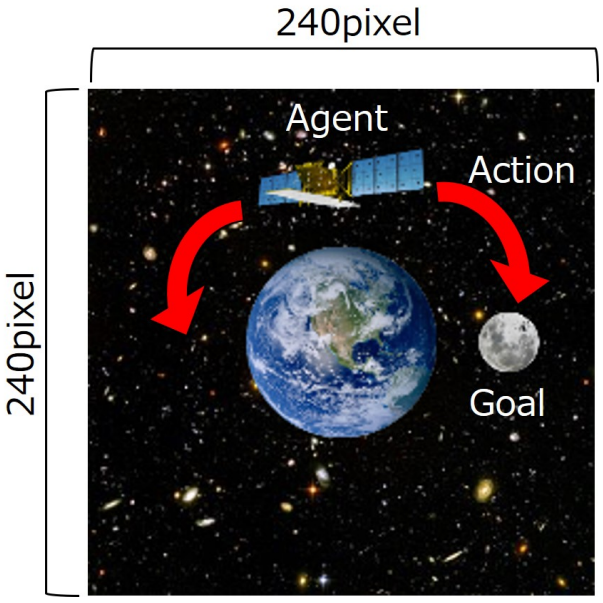


図 16　衛星問題の概念図

4. 提案法 1 によるシミュレーション実験

衛星画像を入力とするリーチングタスクを用いて学習の有効性を検証した. 衛星問題では, 従来法と提案法の学習の遷移を比較し, 計算量と精度の比較を行った. 従来法にはガウス過程方策 (GP 方策) を用いた.

4.1 衛星問題に対する方策探索

4.1.1 設定

画像を入力とするリーチングタスクにより提案法 1 の有効性を確認した. 提案法 1 および, 従来法を用いた学習を行い, 結果を比較することで有効性を示す. 図 16 に示すような人工衛星 (80×36) を操作する問題を考える. 状況に応じて適切な行動を選択することで月の上に到達することを目標とした. 状態空間は 240×240 の画像を入力とする高次元空間とした. 画像入力を状態入力として扱うために, 空間的ピラミッドカーネル (SPM カーネル) を用いた (付録 E). 各時刻において行動

は 1 次元かつ連続であるが, 式 (42) に示すように閾値を設定して行動に制限を与えた.

$$a_t = \min\left\{\max\left(a_t, -\frac{\pi}{3}\right), \frac{\pi}{3}\right\} \tag{42}$$

衛星の位置を表す θ は式 (43) で与えた. ただし, 12 時の方向に衛星があるときに $\theta = 0$ とした.

$$\theta_{t+1} = \theta_t + a_t \tag{43}$$

衛星のピクセル座標は式 (44) で与えた.

$$\begin{aligned} s_x(t) &= x_{center} + \alpha \sin\theta_t \\ s_y(t) &= y_{center} - \alpha \cos\theta_t \end{aligned} \tag{44}$$

ただし, (x_{center}, y_{center}) は宇宙画像の中心ピクセル座標とし, 本実験では $(x_{center}, y_{center}) = (120, 120)$ とした. また係数 $\alpha = 80$ とした. 報酬は式 (45) で示すように, 時刻 t における報酬 $r(t)$ は衛星のピクセル座標と, 目標位置 $s_x^{target}, s_y^{target}$ とのユークリッド距離を考えて与えた.

$$r(t) = 10\exp\left\{-\frac{(s_x(t) - s_x^{end})^2 + (s_y(t) - s_y^{end})^2}{2}\right\} \tag{45}$$

ここで, 目標とする画像上での月のピクセル座標は $s_x^{target}, s_y^{target} = (200, 120)$ とした. 初期の状態は $\theta_0 = 0$ で与えた. 強化学習のパラメータは, ステップ数を 10, エピソード数を 30, 方策の更新回数を 10 とした. 累積報酬の計算に必要な割引率は $\gamma = 0.99$ とした. なお, SPGP の疑似入力点数は $M = 100$ とした. 加法ノイズは $\sigma^2 = 0.005$ とした. また, SPM カーネル (付録 E) に用いたレベルは $L = 2$, ヒストグラムの重みは $(\alpha_0, \alpha_1, \alpha_2) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{2})$, ヒストグラムの横軸は $Z = 300$ としている.

4.1.2 結果

図 17 に従来法と提案法 1 それぞれの手法で, 同じ初期値から 5 回学習を行った結果を報酬値の平均と標準偏差を用いて示す. 緑色の破線が従来法であり, 青色の実線が提案法 1 である. 図 17 より, 提案法 1 の方が従来法よりも期待報酬値の平均値は高いが, 分散は従来法の方が小さいことが分かる. これは提案法 1 を用いることで過学習を抑えれているため, 従来法の方策の更新回数が増えた終盤でも, 適切な分散によって, より良い行動の探索が行えていたからである. 一方, 従来法では報酬重みの高い観測データを重視し過ぎたため, 更新回数が増えるにつれ, 行動の探索が局所的になっている. その結果, 期待報酬値の分散は提案法 1 よりも小さい.

図 18 は初期方策を用いたときの遷移図である. 初期方策では衛星を月の上に到達させれないことが確認できる. 図 19 は方策を 14 回更新した時の遷移図である. 衛星を月の上に到達できることを確認した.

表 1 は従来法と提案法 1, それぞれの手法を用いて, 行動の平均と分散の予測に要した計算時間を示している. 測定値は千回の計算の平均値で示している. 2 手法の計算時間の差はあるものの, 差は大きくはない. これはシミュレーション時間の都合上, 従来法では画像数が 300 枚に対して, 提案法 1 で用いた疑似入力画像が 100 枚としたからである. このデータ数の差がより大きくなると 一層顕著に計算時間の差が現れることが期待できる.

表 1 行動決定に要した計算時間

予測対象	提案法 1	従来法
平均	$0.24 \times 10^{-5}[s]$	$0.33 \times 10^{-5}[s]$
分散	$0.26 \times 10^{-3}[s]$	$1.30 \times 10^{-3}[s]$

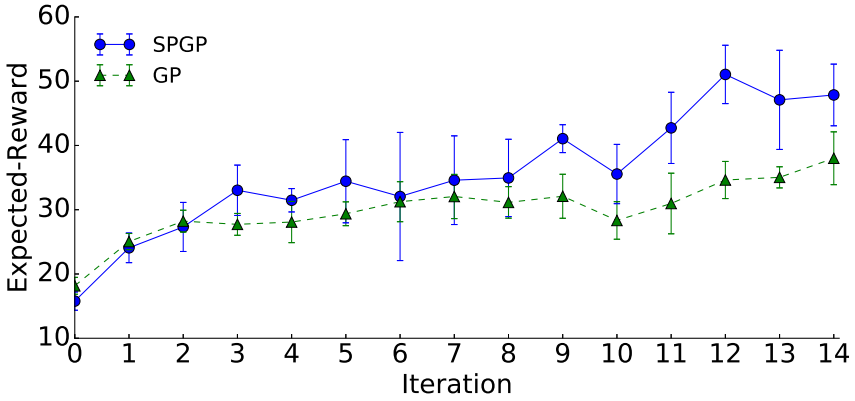


図 17　衛星問題での従来法 (緑) と提案法 1(青) の学習曲線

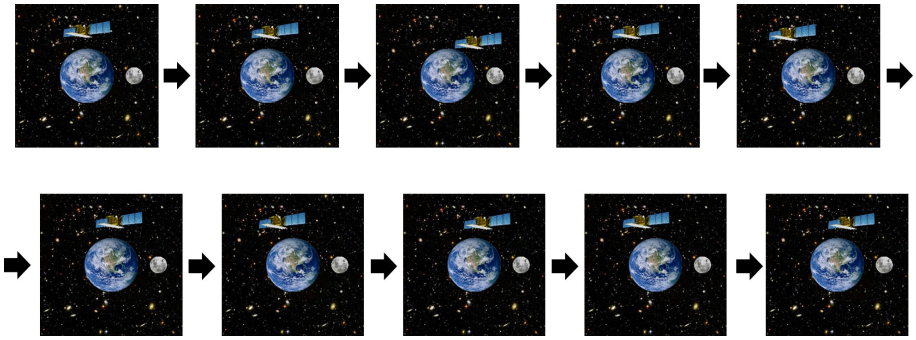


図 18　初期方策を用いたときの遷移図

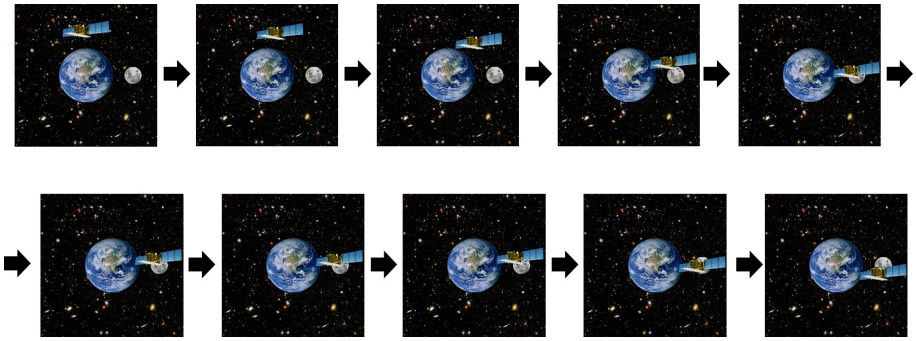


図 19　10 回更新後の方策を用いたときの遷移図

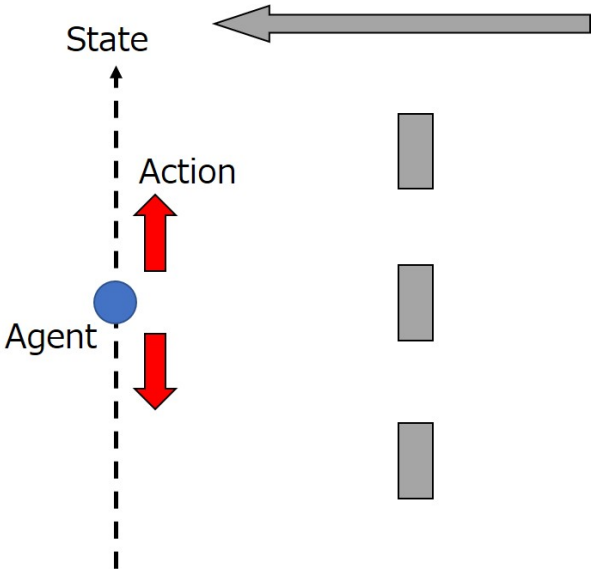


図 20　1 段階スリット問題の概念図

5. 提案法 2 によるシミュレーション実験

スリット通り抜け問題 [24] を用いて学習の有効性を検証した. 従来法と提案法 2 の学習の遷移を比較し, 解が複数存在する問題における提案法の有効性について確認した. 従来法には単峰な方策であるガウス過程方策 (GP 方策) を用いた.

5.1　1 段階スリット通り抜け問題

5.1.1　設定

同一の状態に対して, 2 種類の行動候補が考えられる問題に対して, 提案法 2 を適用し有効性を確認した. 図 20 に示すような複数の壁が 1 回迫ってくる問題を考えた. 状況に応じて適切な行動を選択することで壁を避けることを目標とした. 状態空間は 1 次元かつ連続で $s \in \{-3, 3\}$ とした. 行動も同じく 1 次元かつ連続であるが, 式 (46) に示すように閾値を設定して行動に制限を与えた.

$$a_t = \min\left\{\max\left(a_t, -1.5\right), 1.5\right\} \tag{46}$$

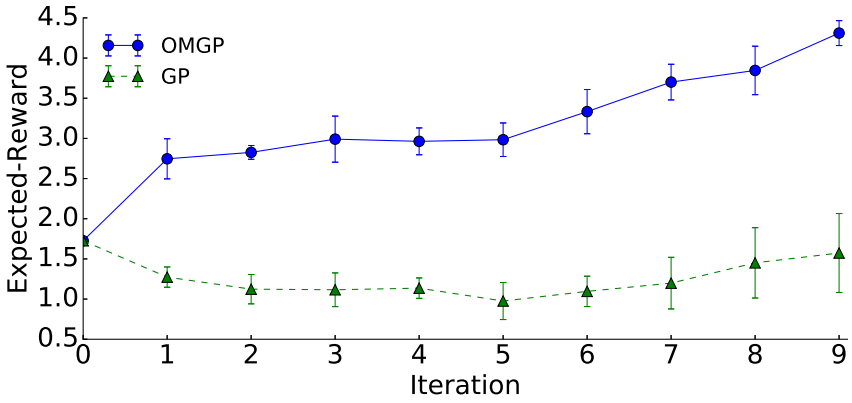


図 21　1 段階スリット問題での従来法 (緑) と提案法 2(青) の学習曲線

状態遷移モデルは式 (47) で与えた.

$$s_{t+1} = s_t + a_t \tag{47}$$

初期の状態は式 (48) で与えた.

$$s_0 \sim \mathrm{rand}(-0.5, 0.5) \tag{48}$$

報酬はスリットを通過したときに報酬値 5 を得るものした. 強化学習のパラメータは, ステップ数を 1, エピソード数を 200, 方策の更新回数を 9 とした. なお, 行動選択に用いた加法ノイズは $\sigma^2 = 0.1 - 0.001l$ $(l = 1, \cdots 9)$ とした. 各更新ごとに 200 点の観測データを用いて方策のモデルを学習した. 混合数は $M = 2$ とした.

5.1.2　結果

図 21 より, 提案法 2 を用いた学習は従来法よりも期待報酬値が高いことが確認できる. 9 回目の更新時に得られていた観測データの例を図 22, 図 23 に示す. 図 22 より, 従来法では単峰性を仮定して学習しているので, 得られた結果も単峰となっていることが確認できる. 一方, 図 23 より, 提案法 2 では多峰性を考慮した学習を行うため, 得られた結果も多峰となっていることが確認できる.

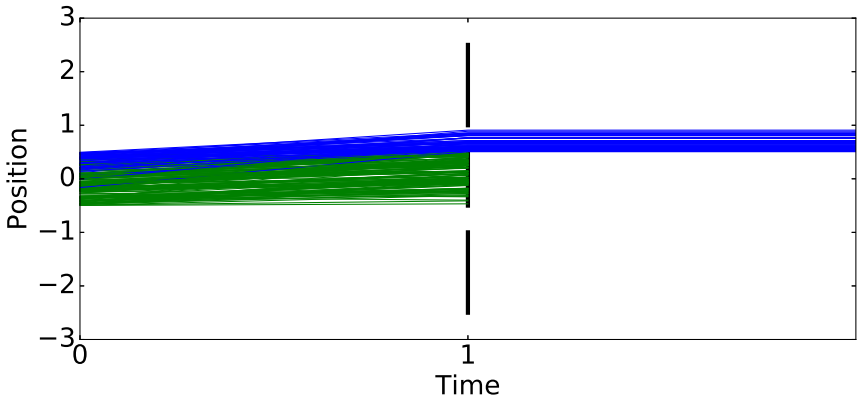


図 22　従来法による 9 回更新後の方策から得られた観測データ

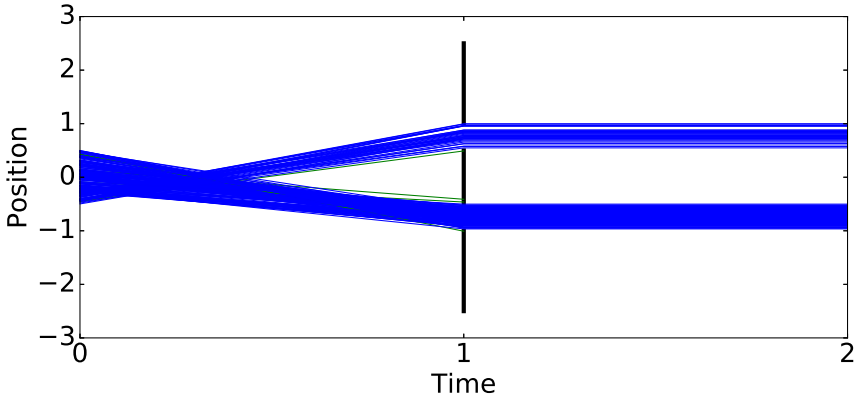


図 23　提案法 2 による 9 回更新後の方策から得られた観測データ

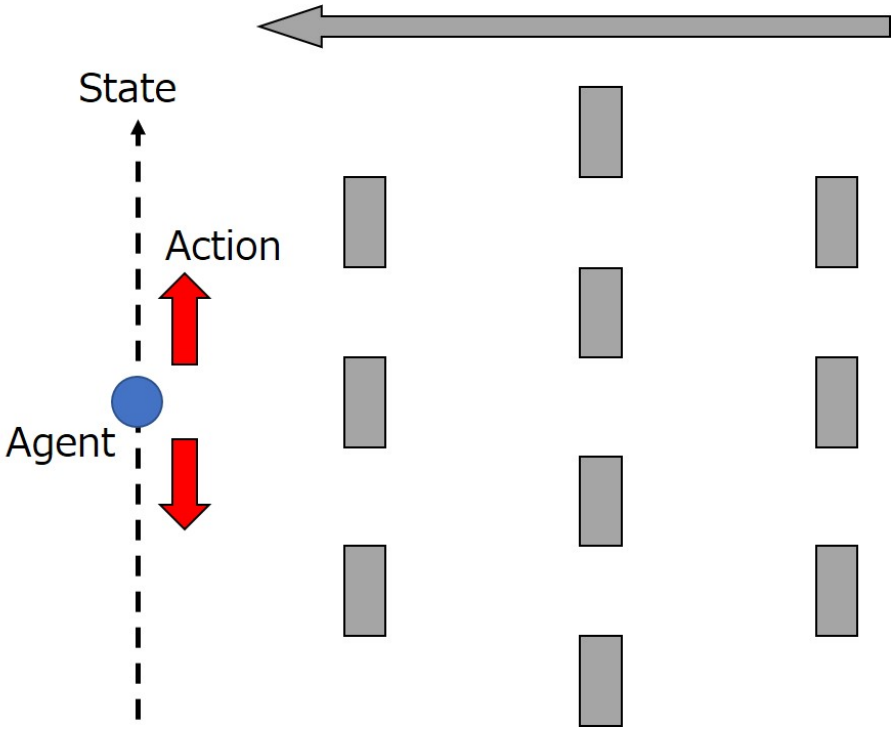


図 24　3 段階スリット問題の概念図

5.2　3 段階スリット通り抜け問題

5.2.1　設定

1 段階スリット問題を階層化した 3 段階スリット問題に対しても, 提案法 2 を適用して有効性を確認した. 図 24 に示すような複数の壁が 3 回迫ってくる問題を考えた. 状況に応じて適切な行動を選択して壁を避けることを目標とした. 状態空間は 1 次元かつ連続で $s \in \{-3.5, 3.5\}$ とした. 行動も同じく 1 次元かつ連続であるが, 式 (49) に示すように閾値を設定して行動に制限を与えた.

$$a_t = \min\left\{\max\left(a_t, -1.5\right), 1.5\right\} \tag{49}$$

状態遷移モデルは式 (50) で与えた.

$$s_{t+1} = s_t + a_t \tag{50}$$

各時刻ごとに与えられる報酬は式 (51) で与えた.

$$r(t) = \begin{cases} (t+1)^2 & (\text{壁を回避}) \\ 0 & (\text{壁に衝突}) \end{cases} \tag{51}$$

初期の状態は式 (52) で与えた.

$$s_0 \sim \mathrm{rand}(-0.5, 0.5) \tag{52}$$

強化学習のパラメータは, ステップ数を 3, エピソード数を 200, 方策の更新回数を 9 とした. なお, 行動選択に用いた加法ノイズは $\sigma^2 = 0.1 - 0.001l \quad (l = 1, \cdots 9)$, 温度パラメータは $\tau = 0.01$ とした. 各更新ごとに 600 点の観測データを用いて方策のモデルを学習した. 混合数は $M = 2$ とした.

5.2.2 結果

図 25 より, 提案法 2 を用いた学習は従来法よりも期待報酬値が高いことが確認できる. 特に更新回数が 1〜4 回のときに, 提案法 2 の期待報酬値が大きく上がっていることが分かる. 学習中に得られていた従来法と提案法 2 の観測データを図 26, 図 27, に示す. 青線が 3 段階のスリットを全て通過したときである. 黄線は 2 段階のスリットを通過したときである. 赤線は 1 段階のスリットを通過したときである. 緑線はスリットを 1 つも通過しないときである. 図 27 より, 1 段階のスリット問題と同様, 提案法 2 では結果が多峰となっている. 図 26 より, 従来法による学習においても, 9 回目の更新時では期待報酬値が高くなっていることが確認できた. これは有限サンプルデータからの従来法による推定量の偏りが原因である. 観測データが少ないため, 結果的に 1 つの解に収束している. しかし, 図 25 に示したように, 提案法 2 と比べて従来法では学習に時間がかかっている. よって, 複数の高い報酬経路が存在する問題においては, 提案法 2 による学習を行う必要があるといえる.

図 28 に方策の更新の遷移図を示す. 方策の更新をしていない状態では多峰な方策となりつつも, 状態に応じた適切な行動を得られていないことが分かる. これは図 27 に示したように, 観測データに青線以外の線が多いことが原因である. 図

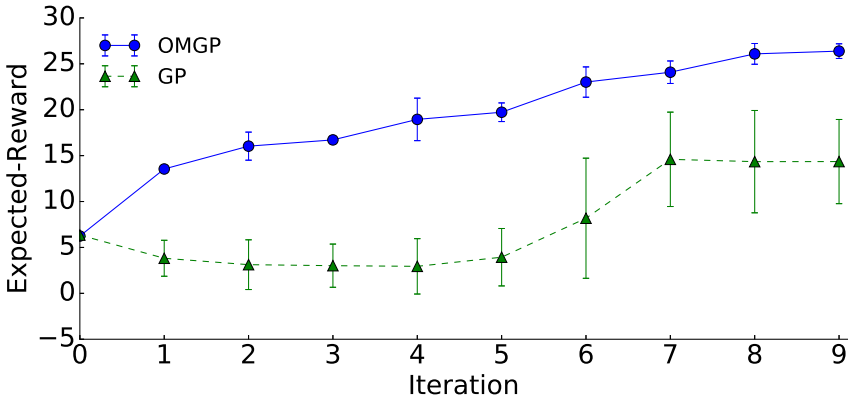
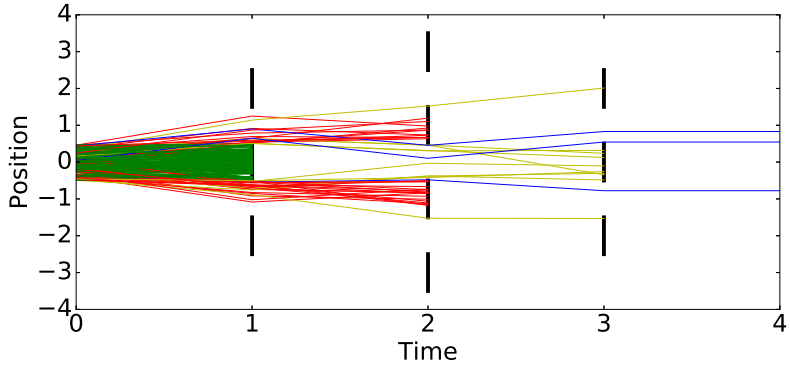
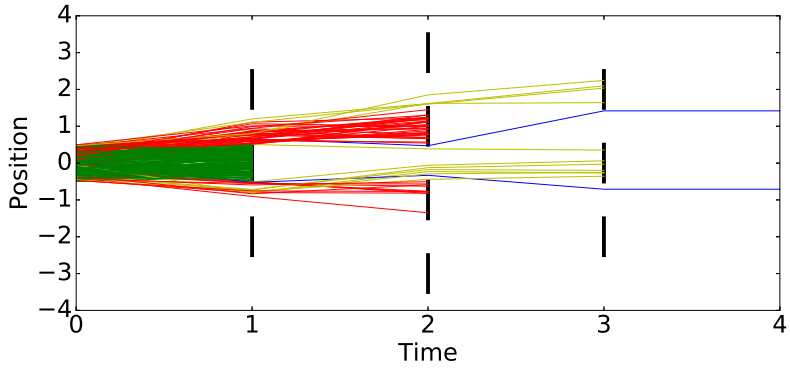


図 25　3 段階スリット問題での従来法 (緑) と提案法 2(青) の学習曲線

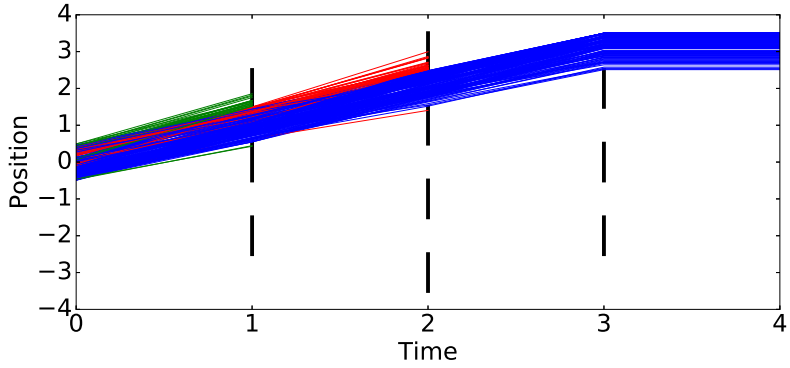
28 より, 方策の更新を増やすと, 予測分散が小さくなりつつ, 状態に応じた行動が得られている. 特に, 5 回目の方策モデルではデータが多く存在する -2 から 2 の状態範囲において, 青と赤で明らかに異なる行動が予測できていることが分かる. 図 29 に従来法の結果と提案法 2 の方策モデルを重ねた結果を示す. 従来法では状態に依存せず, 1 以上の行動が予測されていることが分かる. この結果として図 26 のような上側のスリットを通り抜けてる結果となっていた.



1 回更新後の方策から得られた観測データ

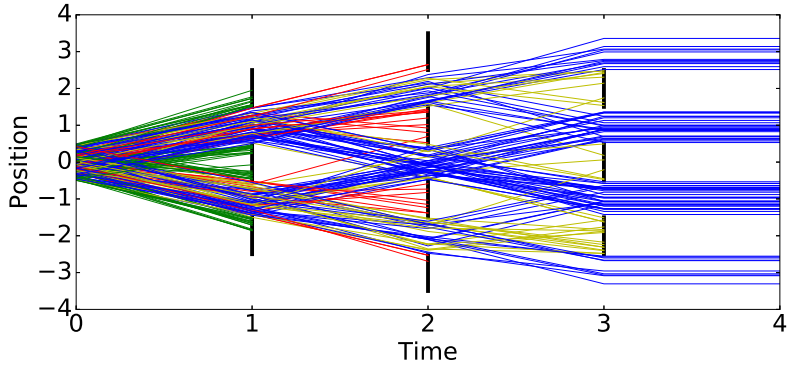


4 回更新後の方策から得られた観測データ

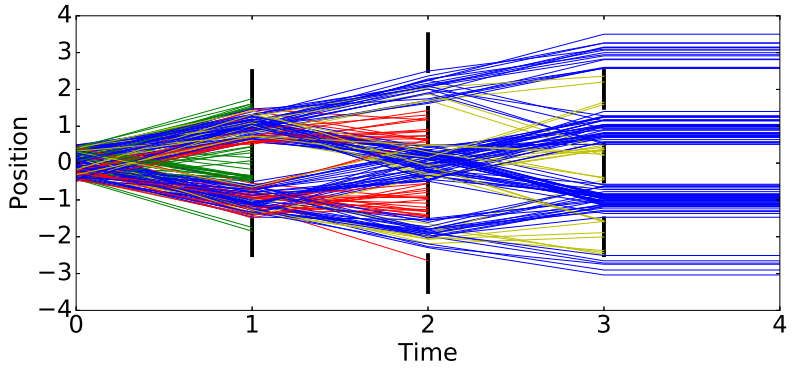


9 回更新後の方策から得られた観測データ

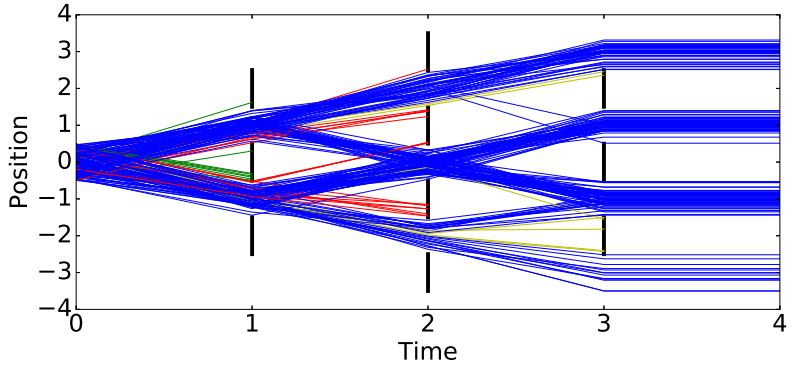
図 26　3 段階スリット問題での従来法の方策から得られた観測データ



1 回更新後の方策から得られた観測データ

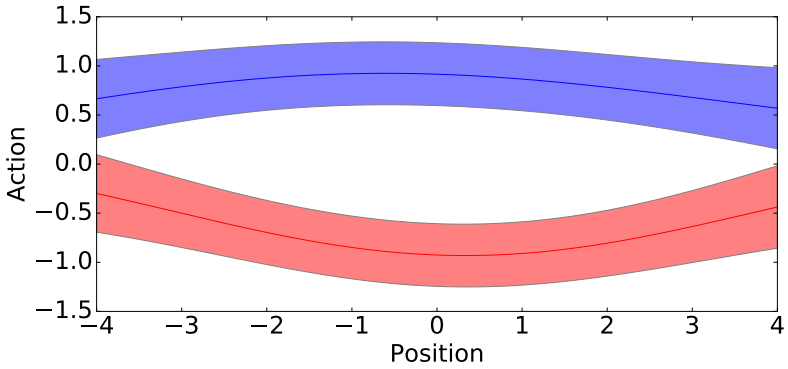


4 回更新後の方策から得られた観測データ

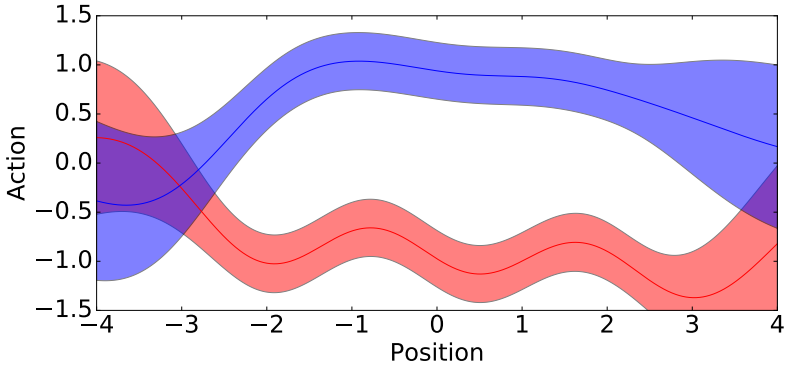


9 回更新後の方策から得られた観測データ

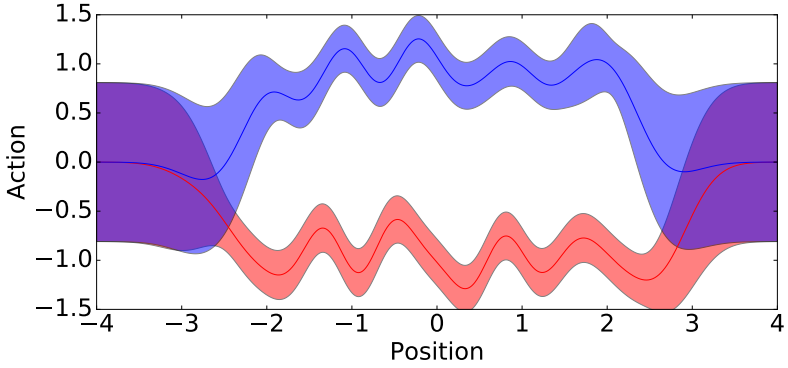
図 27　3 段階スリット問題での提案法 2 の方策から得られた観測データ



初期方策



2 回更新後の方策



4 回更新後の方策

図 28　3 段階スリット問題での提案法 2 の方策

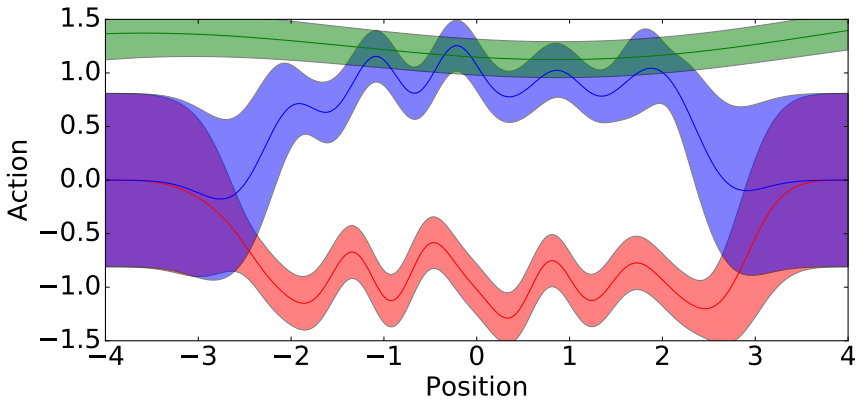


図 29　従来法の方策 (緑) と提案法 2 の方策 (赤と青)

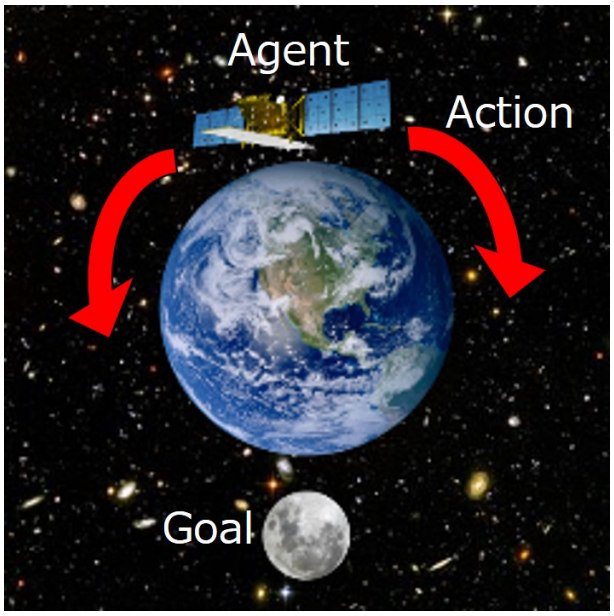


図 30　衛星問題の概念図

5.3 衛星問題に対する方策探索

5.3.1 設定

画像を入力とし, 多峰性のある問題として衛星問題を考え, 提案法 2 による学習の有効性を確認した. 図 30 に示すような人工衛星 (80×36) を操作する問題を考えた. 状況に応じて適切な行動を選択することで月の上に到達することを目標とした. 状態空間は 240×240 の画像を入力とする高次元空間とした. 画像入力を状態入力として扱うために, 空間的ピラミッドカーネルを用いた (付録 E). 各時刻において行動は 1 次元かつ連続であるが, 式 (53) に示すように閾値を設定して行動に制限を与えた.

$$a_t = \min\left\{\max\left(a_t, -\frac{\pi}{3}\right), \frac{\pi}{3}\right\} \tag{53}$$

衛星の位置を表す θ は式 (54) で与えた.

$$\theta_{t+1} = \theta_t + a_t \tag{54}$$

ただし, 12 時の方向に衛星があるときに $\theta = 0$ とした. 衛星のピクセル座標は式 (55) で与えた.

$$\begin{aligned} s_x(t) &= x_{center} + \alpha \sin\theta_t \\ s_y(t) &= y_{center} - \alpha \cos\theta_t \end{aligned} \tag{55}$$

ただし, (x_{center}, y_{center}) は宇宙画像の中心ピクセル座標とし, 本実験では $(x_{center}, y_{center}) = (120, 120)$ とした. また係数 $\alpha = 80$ とした. 報酬は式 (56) で示すように, ステップ数だけ衛星を動かした時の衛星のピクセル座標 s_x^{end}, s_y^{end} と, 目標位置 $s_x^{target}, s_y^{target}$ とのユークリッド距離を考えて与えた.

$$R = 10 \exp\left\{-\frac{(s_x^{target} - s_x^{end})^2 + (s_y^{target} - s_y^{end})^2}{2}\right\} \tag{56}$$

ここで, 目標とする画像上での月のピクセル座標は $s_x^{target}, s_y^{target} = (120, 200)$ とした. 初期の状態は式 (57) で与えた.

$$\theta_0 \sim \frac{\pi}{9}\mathrm{rand}(-1, 1) \tag{57}$$

強化学習のパラメータは, ステップ数を 5, エピソード数を 50, 方策の更新回数を 14 とした. なお, 行動選択に用いた加法ノイズは $\sigma^2 = 0.005$ とした. また, 画像を入力とすることで, 十分な観測データから方策モデルを求める必要があったため, サンプルリユースを用いた. 具体的には, 方策の更新毎に得られる 250 枚の画像に, 過去の経験データから報酬値の高い 150 枚の画像を追加したものを学習データとして追加して方策モデルを求めた. SPM カーネル (付録 E) に用いたレベルは $L = 2$, ヒストグラムの重みは $(\alpha_0, \alpha_1, \alpha_2) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{2})$, ヒストグラムの横軸は $Z = 300$ としている.

5.3.2 結果

図 31 に得られた学習曲線を示す. 青色の実線が提案法 2 の結果であり, 緑色の破線が従来法の結果である. 図 31 より, 従来法では, 方策の更新を行っても, 報酬値は上がらず, 学習が行えていないことが確認できる. これは従来法は単峰であるため, 月に到達する 2 つの経路の平均を取った結果, ほとんど動かない結果となったからである. これは, スリット問題とは異なり, 衛星問題ではサンプルリユースを導入したからである. サンプルリユースを利用しない場合, スリット問題のように, 従来法において観測データの片寄りによって方策の更新によって報酬値が上がる可能性がある. しかし, サンプルリユースにより, 過去の報酬値が高いデータが学習データに残り続けるため, 従来法において, 常に報酬値が高い 2 つの経路を含む学習となり, 学習が行えなかった. 一方, 提案法 2 では方策を更新するにつれ, 報酬値が増加しており学習できていることが確認できる. これはサンプルリユースによって, 報酬値が高い 2 つの経路を含む学習データから, 正しく状態を捉えて多峰な方策モデルを作ることができたからである. 図 32 に提案法 2 によって 14 回方策を更新した後の衛星の動きの一例を示す. この結果より, 右回り, 左回りの 2 つの経路を学習していることが分かる.

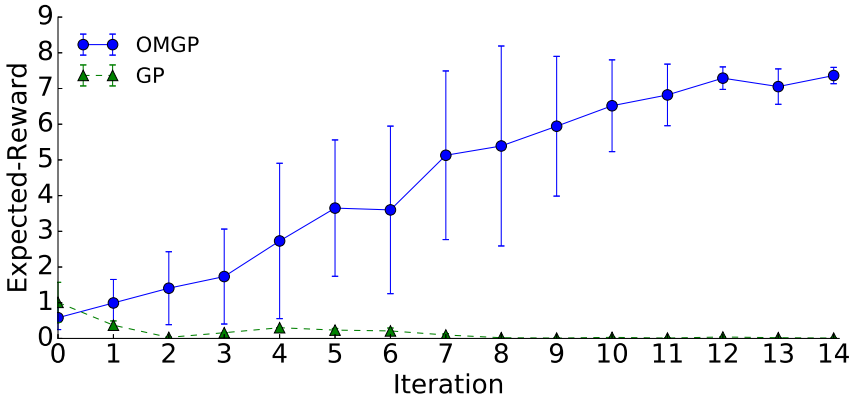


図 31　衛星問題での従来法 (緑) と提案法 2(青) の学習曲線

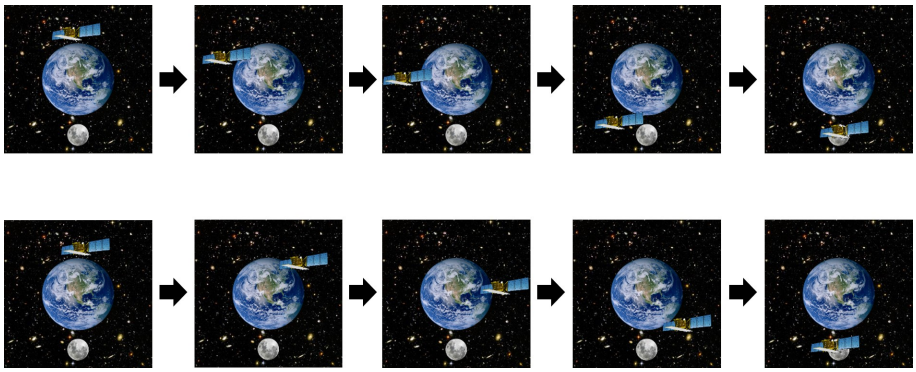


図 32　14 回方策を更新した後の衛星の遷移図

図 33　布展開タスクの目標とする遷移図

6. 実機実験

6.1 設定

実機による実用的な柔軟物操作の実現にあたって, 提案法 1 による学習の有効性を検証した. 図 33 に示すように, 4 つ折りに折りたたまれた布を複数回操作して展開することを目標とした. 初期の 4 つ折りに折りたたまれた状態では青色の面が多く, 布を展開するにつれて, 赤色, 黄色という順に観測される布の色が変わる. 図 34 に示すような布展開タスクを行う. 実機には図 7 に示すような双腕ロボット BAXTER を用いた. また, 画像を取得するために kinectv2 を 1 台, システムの上部に取り付けた. 画像入力を状態入力として扱うために, 空間的ピラミッドカーネルを用いた (付録 E). BAXTER には棒状の物体を掴ませている. 棒は布と緑のゴムによって結びつけられており, 棒を操作することで布を間接的に操作することができる. 状態空間は 240×240 の画像を入力とする高次元空間とした. 時刻 t での行動 a_t は 1 次元かつ連続であるが, 式 (58) に示すように閾値を設定した.

$$a_t = \min\{\max(a_t, -0.1), 0.1\} \tag{58}$$

また, 実機の可能範囲の制限より, 実際に実機を動かす前に, 時刻 $t+1$ での実機の状態 B_{t+1} を式 (59) で計算した.

$$B_{t+1} = B_t + a_t \tag{59}$$

なお, 時刻 0 における実機の状態は $B_0 = 0$ とした. 式 (59) において, $B_{t+1} \in \{0, 0.3\}$ とし, 制限を超える行動 a_t が求まった時は, $a_t = 0$ として実機を動かした.

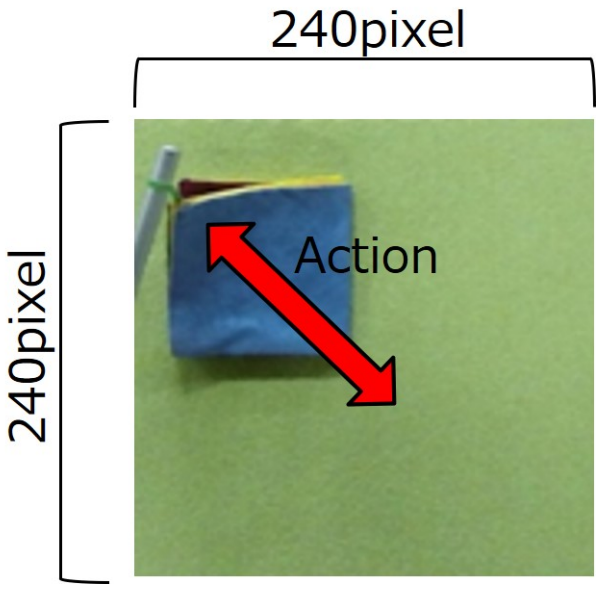


図 34　布展開タスクの概念図

ただし, 報酬計算には, 求まった行動を用いて計算した. 時刻 t での報酬 $r(t)$ は式 (60) で示すように, 行動量と色の付いた布の面積に応じて与えた.

$$\begin{aligned} r(t) &= \gamma^t C^{area} \exp(-|5a_t|) \\ C^{area} &= \frac{p_{red} + 3p_{yellow}}{100} \end{aligned} \tag{60}$$

ここで, $\gamma = 0.99$ は割引率であり, p_{red}, p_{yellow} はそれぞれ, 赤色の布のピクセル数と, 黄色の布のピクセル数を表す. 強化学習のパラメータは, ステップ数を 10, エピソード数を 30, 方策の更新回数を 10 とした. なお, 行動選択に用いた加法ノイズは $\sigma^2 = 0.01$ とした. 各更新ごとに, 疑似入力に使用した画像は 100 枚である. SPM カーネル (付録 E) に用いたレベルは $L = 2$, ヒストグラムの重みは $(\alpha_0, \alpha_1, \alpha_2) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{2})$, ヒストグラムの横軸は $Z = 300$ としている.

6.2　結果

方策を 10 回更新した結果を図 35 に示す. 図 35 のそれぞれの折れ線は 3 回提案法 1 で学習したことを示している. 全ての学習において報酬値が増加しているこ

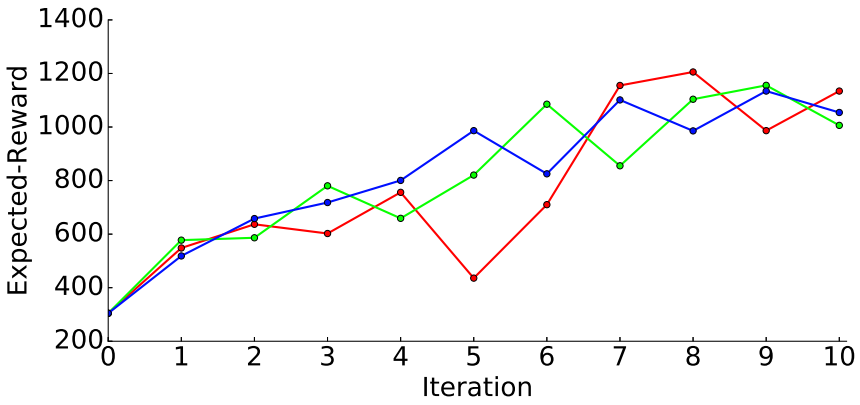


図 35　布展開タスクでの 3 回の実験結果

とから, 実機による実験においても, 提案法 1 によって学習できることを確認した. また, 図 36〜図 38 に, 方策の更新に伴って得られた布の遷移結果を示す. 図 36 に示すように, 初期方策では布を展開できないことを確認した. 図 37 は 4 回方策を更新したときの布の遷移図である. 棒を斜め方向に移動させると布を展開できることを学習し始めていた. 布を最大に展開する動きが, 学習データにまだ十分に含まれていなかったので, 棒を布の中央あたりに移動させた後の動きは不確実性が高く安定していなかった. 図 38 は 10 回方策を更新したときの布の遷移図である. 棒を画像右下にしっかりと移動させ, 布を展開している動きが観測できた. また, 布を広げ切った後は, 棒を静止させている動きも確認できた.

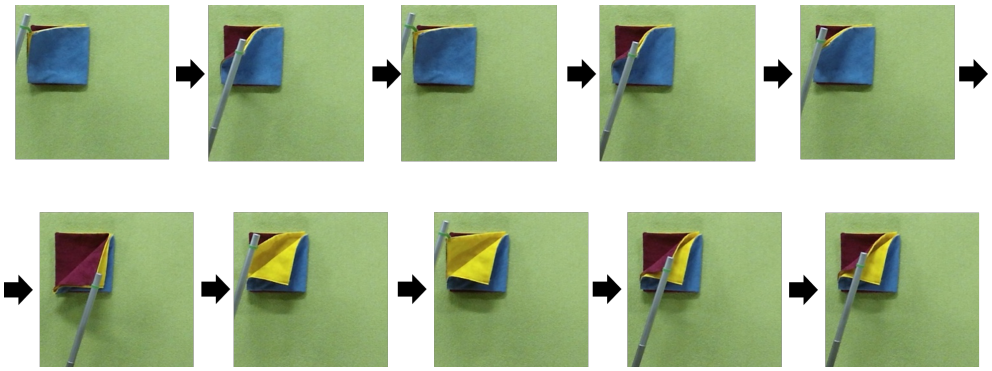


図 36　初期方策を用いたときの遷移図

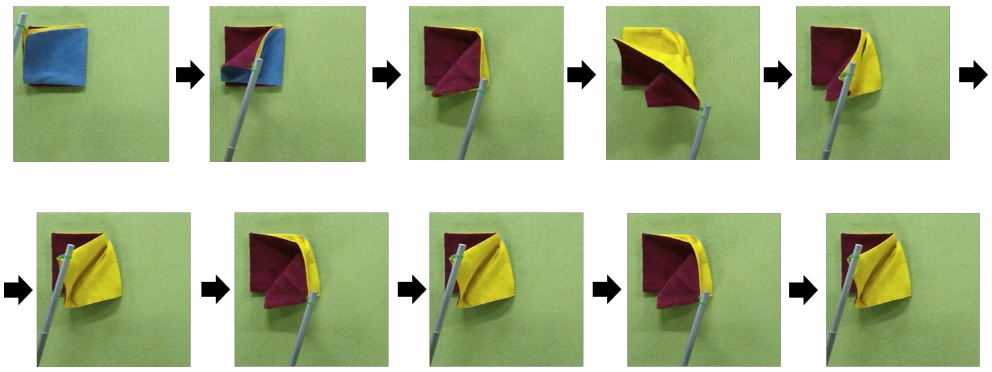


図 37　4 回更新後の方策を用いたときの遷移図

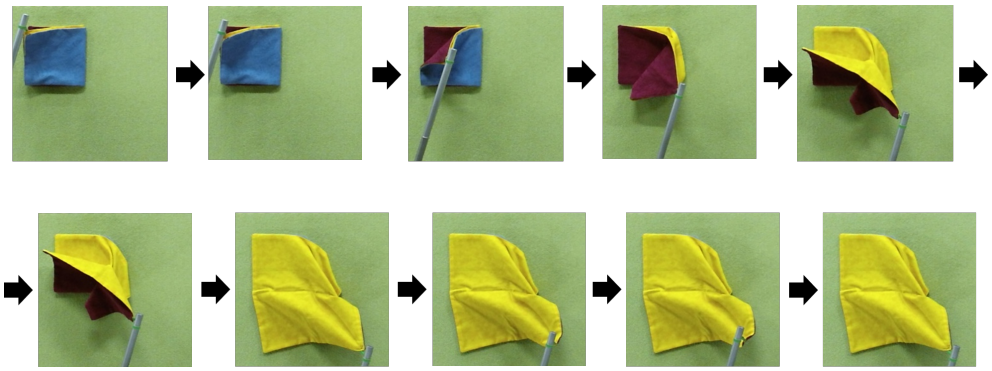


図 38　10 回更新後の方策を用いたときの遷移図

7. おわりに

7.1 まとめ

本研究では, 高次元入力に頑健なガウス過程方策を用いた方策探索を提案した. 具体的には, 2 つの手法を提案した. 1 つ目に, 学習や予測に要する計算量の削減と過学習の回避を目標に, 事前分布にスパースガウス過程を仮定した方策に対して, 変分推論に基づく近似 EM 方策探索を行う手法を提案した. 2 つ目に, ガウス過程方策の単峰性という課題を解決するために, 事前分布に重複混合ガウス過程を仮定した方策に対して, 変分推論に基づく近似 EM 方策探索を行う手法を提案した.

前者のスパースガウス過程方策の方策探索については, シミュレーション実験として衛星問題を考え, 従来法との比較を行った. この結果, 計算速度と学習精度において提案法 1 が優位であることを示した. 後者の重複混合ガウス過程方策の方策探索ついては, スリット問題と衛星問題の実験を行い, 多峰性のある問題に対して提案法 2 が有効であることを確認した.

最後に実機実験として, 提案法 1 を用いて, 4 つ折りに折りたたまれた布の展開タスクを行い, ロボットを用いた実験においても有効であることを示した.

7.2 今後の課題

1 つ目の課題は, 2 つの提案法の報酬重み付き対数尤度関数の計算における, 近似計算の理論的証明である. 本研究では, この近似計算の有効性を, 報酬値の増加という実験的検証によって証明している. しかし, 解析的な計算が困難な積分計算を近似することで発生する, 近似誤差の影響を学習によって補っている可能性もある. したがって, 今後は有効性を理論的に証明する必要がある.

2 つ目の課題は提案法 2 における実機実験である. 様々なシミュレーション実験によって多峰性のある問題に対して提案法 2 が有効であることは示せた. しかし, 提案法 2 を実機に応用した実験は行えていない. そのため, 今後の課題として提案法 2 を用いた実機による学習問題に取り組みたい.

謝辞

本研究を進めるにあたり，知能システム制御研究室の杉本謙二教授には研究のご指導のみならず，研究者，技術者としての姿勢や考え方など，様々な面でご指導い ただきましたこと，心より御礼申し上げます.

また，お忙しい中，副指導教員を引き受けていただきましたロボティクス研究室の小笠原司教授に深く御礼申し上げます.

松原崇充准教授には，知識，研究の方向性，論文の書き方や助言のみならず，生活面でも，様々なアドバイスや熱心なご指導を賜りました. 改めて厚く御礼申し上 げます.

研究室での研究環境の整備や，生活面でのサポートしていただいた小林泰介助教に深く感謝いたします.

研究に関するご指摘や，色々教えていただきました小蔵正輝助教に深く感謝します.

学会への参加に関する事務処理など，その他学生生活を送るうえでの様々なサポートをして下さった秘書の林英子さんに深く感謝いたします.

最後に，研究にのみならず，学生生活において多大なサポートをして頂いた両親と，同じ学習班であり生活面と研究面両方において様々な助言やご支援をしてくれた知能システム制御研究室の諸氏に深く心より感謝致します.

参考文献

[1] M. Saha and P. Isto. Motion planning for robotic manipulation of deformable linear objects. In *International Conference on Robotics and Automation (ICRA)*, pp. 2478–2484, 2006.

[2] 鷲見和彦. 柔軟物も取り扱える生産用ロボットシステムの開発. 日本ロボット学会誌, Vol. 10, pp. 1082–1085, 2009.

[3] 鈴木彼方ら. 深層学習を用いた多自由度ロボットによる柔軟物の折り畳み動作生成. 情報処理学会, Vol. 78, pp. 97–111, 2016.

[4] 山崎公俊, 稲葉雅幸. 布の折れ重なりやしわの状態, 布地に着目した画像特徴量による無造作に置かれた布製品の個体識別. 計測自動制御学会論文集, Vol. 49, No. 7, pp. 661–669, 2013.

[5] L. Sun, G. Aragon-Camarasa, A. Khan, S. Rogers, and P. Siebert. A precise method for cloth configuration parsing applied to single-arm flattening. *International Journal of Advanced Robotic Systems*, Vol. 13, No. 2, p. 70, 2016.

[6] J. Kober and J. Peters. Learning motor primitives for robotics. In *International Conference on Robotics and Automation (ICRA)*, pp. 2112–2118, 2009.

[7] P. Kormushev, S. Calinon, and D. G. Caldwell. Robot motor skill coordination with EM-based reinforcement learning. In *International Conference on Intelligent Robots and Systems (IROS)*, pp. 3232–3237, 2010.

[8] H. Hachiya, J. Peters, and M. Sugiyama. Reward-weighted regression with sample reuse for direct policy search in reinforcement learning. *Neural Computation*, Vol. 23, No. 11, pp. 2798–2832, 2011.

[9] S. Calinon, P. Kormushev, and D. G. Caldwell. Compliant skills acquisition and multi-optima policy search with em-based reinforcement learning. *Robotics and Autonomous Systems*, Vol. 61, No. 4, pp. 369–379, 2013.

[10] J. Kober, A. Wilhelm, E. Öztop, and J. Peters. Reinforcement learning to adjust parametrized motor primitives to new situations. *Autonomous Robots*, Vol. 33, No. 4, pp. 361–379, 2012.

[11] P. Dayan and G. E. Hinton. Using expectation-maximization for reinforcement learning. *Neural Computation*, Vol. 9, pp. 271–278, 1997.

[12] G. Neumann. Variational inference for policy search in changing situations. In *International Conference on Machine Learning (ICML)*, pp. 817–824, 2011.

[13] O. Kroemer C. Daniel, G. Neumann and J. Peters. Hierarchical relative entropy policy search. *Journal of Machine Learning Research(JMLR)*, Vol. 17, No. 93, pp. 1–50, 2016.

[14] C. E. Rasmussen. The infinite gaussian mixture model. In *Advances in Neural Information Processing Systems 12*, pp. 554–560, 2000.

[15] D. Bruno, S. Calinon, and D. G. Caldwell. Bayesian nonparametric multi-optima policy search in reinforcement learning. In *AAAI Conference on Artificial Intelligence*, pp. 1374–1380, 2013.

[16] H. van Hoof, J. Peters, and G. Neumann. Learning of non-parametric control policies with high-dimensional state features. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 995–1003, 2015.

[17] H. van Hoof, J. Peters, and G. Neumann. Non-parametric policy search with limited information loss. *Journal of Machine Learning Research (JMLR)*, No. 73, pp. 1–46, 2017.

[18] V. Mnih, et al. Human-level control through deep reinforcement learning. *Nature*, Vol. 518, No. 7540, pp. 529–533, 2015.

[19] S. Levine, P. Pastor, A. Krizhevsky, and D. Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *arXiv preprint arXiv:1603.02199*, 2016.

[20] E. Snelson and Z. Ghahramani. Sparse gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1257–1264, 2006.

[21] M. K. Titsias. Variational model selection for sparse gaussian process regression. Technical report, School of Computer Science, University of Manchester, UK, 2009.

[22] M. Lázaro-Gredilla, S. Van Vaerenbergh, and N. Lawrence. Overlapping mixtures of gaussian processes for the data association problem. *Pattern Recognition*, Vol. 45, No. 4, pp. 1386–1395, 2012.

[23] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006.

[24] H. J. Kappen. Path integrals and symmetry breaking for optimal control theory. *Journal of Statistical Mechanics: Theory and Experiment*, No. 11, p. P11011, 2005.

[25] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 2169–2178, 2006.

[26] L. Bo, X. Ren, and D. Fox. Depth kernel descriptors for object recognition. In *International Conference on Intelligent Robots and Systems (IROS)*, pp. 821–826, 2011.

付録

A. 提案法 1 における F_V の導出

汎関数 $q(\mathbf{f}, \bar{\mathbf{f}}) = p(\mathbf{a}|\mathbf{f})\phi(\bar{\mathbf{f}})$ となるような $(\bar{\mathbf{f}})$ に関する確率分布 $\phi(\bar{\mathbf{f}})$ を仮定する. イェンセンの不等式より式 (61) のような下界を得る.

$$\begin{aligned} F_V(\overline{\mathbf{S}}, \phi) &= \int p(\mathbf{f}|\bar{\mathbf{f}})\phi(\bar{\mathbf{f}}) \log \frac{p(\mathbf{a}|\mathbf{f})p(\mathbf{f}|\bar{\mathbf{f}})p(\bar{\mathbf{f}})}{p(\mathbf{f}|\bar{\mathbf{f}})\phi(\bar{\mathbf{f}})} d\mathbf{f} d\bar{\mathbf{f}} \\ &= \int p(\mathbf{f}|\bar{\mathbf{f}})\phi(\bar{\mathbf{f}}) \log \frac{p(\mathbf{a}|\mathbf{f})p(\bar{\mathbf{f}})}{\phi(\bar{\mathbf{f}})} d\mathbf{f} d\bar{\mathbf{f}} \end{aligned} \tag{61}$$

これは次式 (62) のように書ける.

$$F_V(\overline{\mathbf{S}}, \phi) = \int \phi(\bar{\mathbf{f}}) \left\{ \int p(\mathbf{f}|\bar{\mathbf{f}}) \log p(\mathbf{a}|\mathbf{f}) d\mathbf{f} + \log \frac{p(\bar{\mathbf{f}})}{\phi(\bar{\mathbf{f}})} \right\} d\bar{\mathbf{f}} \tag{62}$$

$\mathbf{f}$ に関する積分計算は次式 (63) のようになる.

$$\begin{aligned} \log G(\bar{\mathbf{f}}, \mathbf{a}) &= \int p(\mathbf{f}|\bar{\mathbf{f}}) \log p(\mathbf{a}|\mathbf{f}) d\mathbf{f} \\ &= \int p(\mathbf{f}|\bar{\mathbf{f}}) \left\{ -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \mathrm{Tr}[\mathbf{a}\mathbf{a}^{\mathrm{T}} - 2\mathbf{a}\mathbf{f}^{\mathrm{T}} + \mathbf{f}\mathbf{f}^{\mathrm{T}}] \right\} d\mathbf{f} \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \mathrm{Tr}[\mathbf{a}\mathbf{a}^{\mathrm{T}} - 2\mathbf{a}\boldsymbol{\alpha}^{\mathrm{T}} + \boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathrm{T}} + K_{nn} - Q_{nn}] \end{aligned} \tag{63}$$

ただし, $\boldsymbol{\alpha} = E[\mathbf{f}|\bar{\mathbf{f}}] = K_{nm}K_{mm}^{-1}\bar{\mathbf{f}}, Q_{nn} = K_{nn}K_{mm}^{-1}K_{mn}$ とした. 式 (63) についてガウス分布を考慮すると, 次式 (64) のように表せる.

$$\log G(\bar{\mathbf{f}}, \mathbf{a}) = \log[\mathcal{N}(\mathbf{a}|\boldsymbol{\alpha}, \sigma^2 I)] - \frac{1}{2\sigma^2} \mathrm{Tr}[K_{nn} - Q_{nn}] \tag{64}$$

よって, $F_V(\overline{\mathbf{S}}, \phi)$ は次式 (65) のように表せる.

$$F_V(\overline{\mathbf{S}}, \phi) = \int \phi(\bar{\mathbf{f}}) \log \frac{\mathcal{N}(\mathbf{a}|\boldsymbol{\alpha}, \sigma^2 I)p(\bar{\mathbf{f}})}{\phi(\bar{\mathbf{f}})} d\bar{\mathbf{f}} - \frac{1}{2\sigma^2} \mathrm{Tr}[K_{nn} - Q_{nn}] \tag{65}$$

式 (65) の第一項において, イェンセンの不等式を用いると, 次式 (66) のようになる.

$$\begin{aligned} F_V(\overline{\mathbf{S}}) &= \log \int \mathcal{N}(\mathbf{a}|0, \sigma^2\mathbf{I}) p(\overline{\mathbf{f}}) d\overline{\mathbf{f}} - \frac{1}{2\sigma^2}\mathrm{Tr}(K_{nn} - Q_{nn}) \\ &= \log[\mathcal{N}(\mathbf{a}|0, \sigma^2\mathbf{I} + \mathbf{Q}_N)] - \frac{1}{2\sigma^2}\mathrm{Tr}(\mathbf{K}_N - \mathbf{Q}_N) \end{aligned} \tag{66}$$

B. 逆行列の計算に利用した行列の公式

行列 A, B, C を考える. このとき式 (67) で逆行列は計算できる.

$$(\mathbf{A} + \mathbf{CBC}^{\mathrm{T}})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{C}(\mathbf{B}^{-1} + \mathbf{C}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{C})^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{A}^{-1} \tag{67}$$

C. ガウス分布の積の公式

平均を $\boldsymbol{\mu}$, 分散を $\boldsymbol{\Sigma}$ とする 2 つのガウス分布 $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ を考える. このとき, これらのガウス分布の積は式 (68) となる.

$$\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) = c\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \tag{68}$$

ただし, $c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c$ は式 (69) である.

$$\begin{aligned} c &= \frac{1}{\sqrt{\det(2\pi(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2))}} \exp\left(-\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right) \\ \boldsymbol{\mu}_c &= (\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})^{-1}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2) \\ \boldsymbol{\Sigma}_c &= (\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})^{-1} \end{aligned} \tag{69}$$

D. 重複混合ガウス過程のハイパーパラメータの最適化

式 (70) を最大化することで提案法 2 のハイパーパラメータの最適化を行う. D は出力次元数であるが, 本研究では $D = 1$ としている.

$$\begin{aligned} L = \sum_{m=1}^{M} \left(-\frac{1}{2} \sum_{d=1}^{D} ||\mathbf{R}^{(m)\mathrm{T}} \backslash (\mathbf{B}^{(m)1/2} \mathbf{a}_d^{(m)}||^2 - D \sum_{n=1}^{N} \log [\mathbf{A}^{(m)}]_{nn} \right) \\ - \mathrm{KL}(q(\mathbf{Z})||p(\mathbf{Z})) - \frac{D}{2} \sum_{n=1,m=1}^{N,M} [\hat{\mathbf{\Pi}}]_{nm} \log(2\pi\sigma^2) \end{aligned} \tag{70}$$

ただし, $\mathbf{A}$ は式 (71) とした. また $\backslash$ は線形方程式の解としている[2].

$$\mathbf{A}^{(m)} = \mathrm{chol}(\mathbf{I} + \mathbf{B}^{(m)1/2} \mathbf{K}_R^{(m)} \mathbf{B}^{(m)1/2}) \tag{71}$$

E. 空間的ピラミッドカーネル

2 つのベクトル (画像) の類似度を求めるカーネル関数として, SIFT 特徴量をベースにした空間的ピラミッドカーネル (SPM カーネル) [25] を用いる. SPM カーネルでは入力画像を $0, \ldots, L$ レベルの異なる解像度に分解し, SIFT 特徴量を求める. SIFT 特徴量とは回転変化や明暗の変化に頑健な画像特徴量である. 例えば, レベル 0 の解像度の画像に対して SIFT 特徴を求めると図 39 のようになり, 色のついた中抜きの円が SIFT 特徴量を検出した位置となっている. また, 入力画像を異なるレベルの解像度に分解する際, レベル l の画像は 2^{2l} の領域に分割される. 本研究では $L = 2$ とし, そのときの画像を図 40 に示した. 図 40 の上段より, レベル 0 では入力画像をそのまま扱う. レベル 1 では入力画像を 4 分割し, レベル 2 では入力画像を 16 分割する. 異なる解像度に分解後, レベル l における画像領域 2^{2l} に対して, それぞれの領域から SIFT 特徴量を求める. そして求めた SIFT 特徴量を横軸を Z としたヒストグラムで表す. $Z = 300$ のときのヒストグラムを求めた結果が図 40 の下段である. SPM カーネルでは画像 $\mathbf{s}$ の特徴ベクトル $\phi(\mathbf{s})$ は各解像度のヒストグラムの重み付き和で求まる. 解像度 l のヒストグラムを H_l, 重み

[2] $\mathbf{C}\backslash\mathbf{c}$ は線形方程式 $\mathbf{Cx} = \mathbf{c}$ の解 $\mathbf{x}$ を示す.

を α_l とすると, $\phi(\mathbf{s})$ は式 (72) で表せる.

$$\phi(\mathbf{s}) = \sum_{l=0}^{L} \alpha_l H_l \tag{72}$$

また, $\phi(\mathbf{s})$ の次元 $\dim(\phi(\mathbf{s}))$ は解像度のレベル L と抽出するヒストグラムの横軸 Z に依存し, 式 (73) で表せる.

$$\dim(\phi(\mathbf{s})) = Z \sum_{l=0}^{L} 4^l = \frac{Z}{3}(4^{L+1} - 1) \tag{73}$$

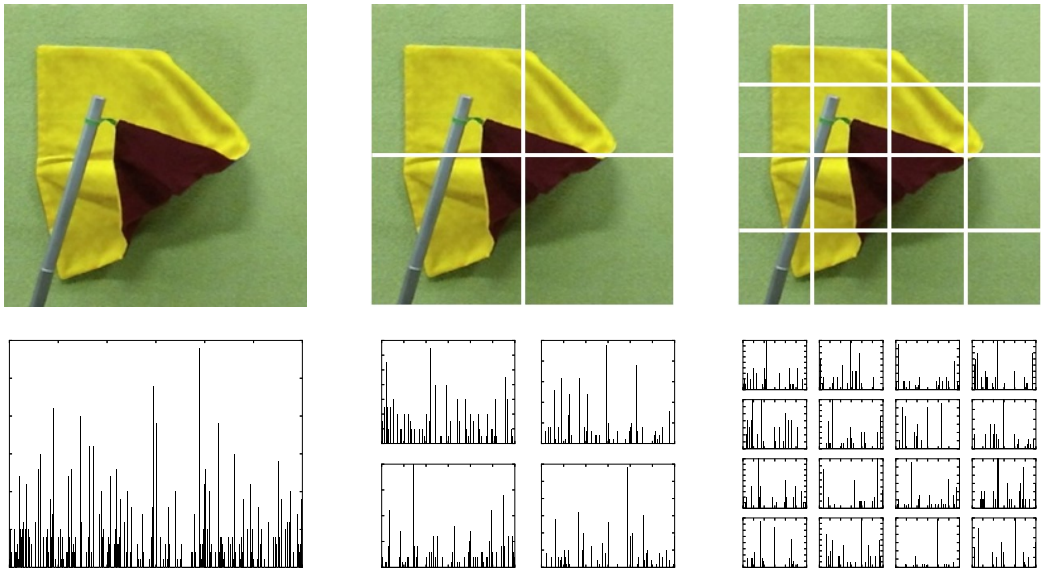
特徴ベクトル $\phi(\mathbf{s})$ を用いると, 文献 [26] より, 2 画像 $\mathbf{s}_1, \mathbf{s}_2$ の類似度を求めるカーネル $\mathbf{k}(\mathbf{s}_1, \mathbf{s}_2)$ は式 (74) で表せる.

$$\mathbf{k}(\mathbf{s}_1, \mathbf{s}_2) = \phi(\mathbf{s}_1)^{\mathrm{T}} \phi(\mathbf{s}_2) \tag{74}$$

本研究では特徴ベクトル $\phi(\mathbf{s})$ に対して主成分分析を適用して 100 次元ベクトル $\omega\phi(\mathbf{s})$ を求め, 式 (75) に示すようなカーネルを利用した.

$$\mathbf{k}(\mathbf{s}_1, \mathbf{s}_2) = (\omega\phi(\mathbf{s}_1))^{\mathrm{T}} (\omega\phi(\mathbf{s}_2)) \tag{75}$$

図 39　レベル 0 のときに検出された SIFT 特徴量



Level 0　　Level 1　　Level 2

図 40　レベル毎の区切り (上段) とヒストグラム (下段)