論文 / 著書情報
Article / Book Information

題目(和文)	多点探索に基づく強化学習とその応用
Title(English)	Reinforcement learning based on multipoint search and its application
著者(和文)	宮前惇
Author(English)	Atsushi Miyamae
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第8416号, 授与年月日:2011年3月26日, 学位の種別:課程博士, 審査員:小林 重信
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第8416号, Conferred date:2011/3/26, Degree Type:Course doctor, Examiner:
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

博士 (工学) 論文

多点探索に基づく強化学習とその応用

Reinforcement Learning based on Multipoint Search and its Application

指導教官　小林 重信　教授

学籍番号　08D35201

氏名　　宮前 惇

目 次

第1章　はじめに

1.1　背景と目的

センサーの入力からアクチュエータへの出力への写像を制御器 (controller) と呼ぶ．制御器の最適化はシステムの制御を伴う工学の様々な場面で現れる重要な問題クラスである．最適な制御器が解析的に求まる場合は限られており，近似解法によるアプローチが必要かつ重要である．

強化学習 (reinforcement learning) [36] によるアプローチでは制御対象の領域知識を必要とせず，報酬と呼ばれるスカラーの評価指標を設定し，試行錯誤によって得られる経験から期待獲得報酬を最大化する制御器を獲得できる．強化学習では制御器のことを政策 (policy) と呼ぶ．

従来の強化学習手法の多くは離散化された状態や行動を表形式で扱う価値関数法をベースにしている．TD-法や Q-learning，Sarsa などが代表例である．これらの方法は局所的表現能力に優れているが，高次元問題において組合せ的爆発の問題を内包していること，連続状態・連続行動を取り扱うとき不完全知覚に伴い発生する非マルコフ性への対処などの点で問題がある．

価値関数法に代わる新たなアプローチとして直接政策探索法 (Direct Policy Search; DPS) が近年注目されている．このアプローチでは政策を，例えば，ニューラルネットなどのパラメトリックモデルで表現し，そのパラメータ（以下，政策パラメータ）を直接最適化する．DPS の最適化手法として，例えば，進化計算のひとつである実数値 GA や勾配法に基づく政策勾配法 [40] がある．

DPS の実用化という観点では，多峰性の扱いが必要かつ重要である．多くの実問題にお

いて期待獲得報酬は政策パラメータに対して複数の局所解が存在するからである．また，実機の制御問題など環境とのインタラクションのコスト (時間，費用) が非常に高いといった特徴をもつことが多い．進化計算 (evolutionary computation) アプローチは，確率的多点探索によって初期値への影響を緩和し，多峰性の景観において比較的評価値の良い局所解を発見しやすいとして関数最適化の領域で近年注目されており，DPS のシナリオに適している [11]．しかし，多点を用いた探索がゆえに，環境とのインタラクション回数が多くなる傾向にある．一方，政策勾配法 (policy gradient methods) は政策の微分情報を利用することで比較的効率よく局所解を発見する DPS として多くのアプリケーションに適用されている近似解法であるが [18, 39, 29]，複数の局所解を有する多峰性の景観においては初期値の影響を強く受けることが知られている [11]．そのため，様々な初期値から探索を行う必要があるので試行数が増大し，結果的に環境とのインタラクション回数が多くなる．このため，多峰性の景観を対象とした効率的なアルゴリズム設計は重要な課題である．

本研究では，アルゴリズムの効率化を，

1. 大域的に最適な政策パラメータを獲得するための試行回数の削減，
2. 1 試行における局所最適な政策パラメータを探索するための環境とのインタラクション回数の削減，

と考え，多点探索によるアプローチをもとに，より効率的な DPS の設計を研究目的とする．

1.2 研究の方法

本研究は，実数値 GA および政策勾配法を基点として，二つの方向からのアプローチからなる．

実数値 GA は大域的探索能力に優れるため，1 試行におけるインタラクション回数の削減が目的となる．政策勾配法は局所探索の効率に優れるため，大域的に最適な政策の獲得に要する試行回数の削減が目的となる．以下では本研究の基本方針を述べる．

実数値 GA からのアプローチ　実数値 GA からのアプローチでは，大域的探索能力を維持しながら局所探索の効率改善を図る．従来の実数値 GA において交叉の際のペアリングは実数値ベクトル表現された個体の第 i 要素同士で行われる．しかし，政策パラメータ最適化では政策を表すモデルに，例えば，多層パーセプトロンの中間ユニットなど役割の入れ換えが可能である要素が存在する場合がある．そのため，役割の異なる要素 (中間ユニット) 同士がペアリングされて交叉が適用されることで，効率的な改善解の生成が妨げられると考えられる．それに対して，役割を考慮したペアリングを行う枠組みは探索を効率化する可能性を有する．そこで，役割を考慮したペアリングの枠組みのもとで効率的に探索を行う実数値 GA の枠組みを設計する．

政策勾配法からのアプローチ　政策勾配法からのアプローチでは，局所探索の効率を維持向上しながら大域的探索を行う枠組みへの拡張を図る．大域的探索の枠組みへの拡張では，進化計算が多点を用いた探索により多峰性の景観へ対処を行っていることから，政策勾配法を多点探索を行う枠組みに拡張する．拡張を行った政策勾配法の探索効率は各点の局所探索で生成したサンプルを再利用することで維持する．また，実数値 GA による DPS とは異なり，政策勾配法では制御器の微分情報が利用可能であることを想定しているが，そのような制約を排除しつつさらなる探索効率の向上を図る．

1.3　論文の構成

本研究は,「多点探索に基づく強化学習とその応用」と題し，全 6 章からなる．以下，章立ておよび，各研究の要旨を述べる．

第 2 章 問題設定　本研究における基本的概念である強化学習について概観し，制御器のパラメータ最適化問題を強化学習問題として定式化する．

第 3 章 実数値 GA によるインスタンスベース政策最適化と非ホロノミック系制御への適用
この研究では，政策のモデルとしてインスタンスベース政策 (IBP) を用いた実数値 GA に

基づく強化学習手法を提案する．インスタンスは状態と行動の対からなり，政策はインスタンス集合からなる．IBP を対象とした DPS がいくつか提案されているが，これらは状態行動空間が連続である場合に最適政策を獲得できない．ここでは状態行動空間が連続である場合の IBP を対象とする DPS のための実数値 GA を提案する．IBP の最適化には三つの困難さがあり，高次元性，変数間依存性，そして多峰性である．提案手法はこれらの困難さを克服するように設計されている．IBP の最適化は高次元問題となるが，評価値を改善することができるインスタンスは評価時に使われた活性インスタンスに限定される．そして，活性インスタンス数は少ないことが多いため，IBP の最適化を活性インスタンスのパラメータ探索に制限することで低次元の問題として扱うことができる．変数間依存性を持つ関数を効率的に最適化するためには統計量を保存する交叉を使うのが良いことが一般的に知られていることから，IBP を効率的に探索するためにインスタンスの周りに統計量を保存するよう新たなインスタンスを生成する交叉的突然変異（拡張 XLM）を提案する．多峰性を克服するために世代交代モデルには拡張 CCM を使う．拡張 CCM では親とその親から生成された子個体の中から生存できる個体数を制限することで，拡張 XLM を使用する場合に集団の多様性を維持する．拡張 XLM と拡張 CCM からなる提案手法を FLIP と名づけ，宇宙ロボット，車両ロボット，並列二重倒立振子での実験により有効性を示す．

第 4 章 集団を用いた政策勾配法と多峰性景観への適用　この研究では，従来の政策勾配法と比べて高い確率で最適政策を学習する新たな手法を提案する．提案手法は政策勾配法に基づいた多点探索手法であり，複数の政策パラメータの組を保持する．様々な政策パラメータから改善することで最適政策を学習する確率を高めることができる．そして，重点サンプリング手法により全ての政策パラメータの勾配方向を同一の履歴から推定することで学習を効率化する．さらに，複数の政策パラメータから最良のものを選択するための政策を導入する．この政策は政策勾配法によりエージェントの性能を最大化するように改善される．提案手法の有効性を 2 分木タスク，匍匐ロボットの前進タスクを用いて示す．

第 5 章 確率モデルで表した制御パラメータに対する自然政策勾配法　この研究では，確率モデルで表した制御器のパラメータに対する効率的な自然政策勾配法を提案する．提案手法は制御器の微分情報を必要としない Policy Gradients with Parameter-based Exploration の枠組みに基づいている．また，従来の自然政策勾配法とは異なり，提案手法では真のフィッシャー情報行列を使用しており，自然政策勾配の推定に要する計算量は通常の政策勾配のそれと等価である．提案手法の探索効率が従来の政策勾配法と比べて向上することを線形 2 次形式制御タスク，匍匐ロボットの前進タスクを用いて示す．

第 6 章 おわりに　本研究の成果を要約し，今後の研究課題をまとめる．

第2章　問題設定

2.1　強化学習による制御問題へのアプローチ

2.1.1　強化学習の枠組み

強化学習とは，学習主体であるエージェントが試行錯誤を通じて環境に適応する枠組みである．教師付き学習とは異なり，状態観測に対する正しい行動を明示的に示す教師が存在しない．かわりに報酬というスカラーの情報を手がかりに学習するが，報酬にはノイズや遅れがある．そのため，行動を実行した直後の報酬をみるだけでは，エージェントはその行動が正しかったかどうかを判断できないという困難を伴う．

エージェントは環境と以下のやり取りを行う．

step 1. エージェントは環境の状態観測に応じて意思決定を行い，行動を選択．

step 2. 選択した行動を環境へ実行することで状態遷移が起こり，エージェントは環境から次状態と状態遷移に伴う報酬を受け取る．

step 3. 時刻をひとつ進めて **step 1.** へ戻る．

エージェントは，報酬によって定まる評価値の最大化を目的として，状態観測から行動への写像である政策を獲得する．エージェントと環境には一般に下記の性質が想定される．

- エージェントは予め環境に関する知識を持たない．
- 環境の状態遷移は確率的．
- 報酬の与えられ方は確率的．

- 状態遷移を何度か繰り返した後で報酬が与えられるような，段取り的な行動を必要とする環境.

2.1.2 部分観測マルコフ決定過程と政策表現モデル

制御対象である環境はマルコフ決定過程 (Markov Decision Processes) でモデル化できると仮定し，制御問題を不十分なセンサーのため状態観測に不確実性や不完全性の存在する部分観測マルコフ決定過程 (Partially Observable MDPs; POMDPs) として扱う．強化学習における POMDPs は以下のように定義される．状態集合を $\mathcal{S}$，観測集合を $\mathcal{X}$，行動集合を $\mathcal{A}$ とする．エージェントは状態 $\mathbf{s}$ において確率 $p_X(\mathbf{x}|\mathbf{s})$ で観測 $\mathbf{x} \in \mathcal{X}$ を受け取り，$\mathbf{x}$ に基づいて行動 $\mathbf{a} \in \mathcal{A}$ を実行する．その結果として，次状態 $\mathbf{s}'$ へ確率 $p_T(\mathbf{s}'|\mathbf{s}, \mathbf{a})$ で遷移し，期待値が $\mathcal{R}(\mathbf{s}, \mathbf{a}) \in \Re$ である有界な報酬 r を得る．初期状態は確率 $p_I(\mathbf{s})$ で状態 $\mathbf{s}$ であるとする．エージェントは予め $p_X(\mathbf{x}|\mathbf{s})$ や $p_T(\mathbf{s}'|\mathbf{s}, \mathbf{a})$，$\mathcal{R}(\mathbf{s}, \mathbf{a})$，$p_I(\mathbf{s})$ に関する知識を持っていないものとする．

また，エージェントの政策 π をモデルで表し，そのパラメータ $\theta \in \Re^d$ を政策パラメータと呼ぶ．政策 π のもとで観測 $\mathbf{x}$ において行動 $\mathbf{a}$ を選択する確率を $\pi(\mathbf{a}|\mathbf{x}, \theta)$ で表す．本研究では制御器のパラメータを政策パラメータに含めることで，政策最適化により制御器のパラメータの最適化を行う．

2.2 政策の評価値

本研究では，政策の評価値としてエージェントが受け取る報酬の平均の期待値である平均報酬 (average reward)[35] を用いる．

制御問題はその性質から二種類のタスクに分類される．ひとつはエピソード的タスクで，もうひとつは連続タスクである．それぞれのタスクにおける平均報酬は以下のようである．エピソード的タスクは，エピソードが T ステップで終了するタスクをいう．このとき，平

均報酬は

$$\eta(\theta) = \frac{1}{T}\mathbb{E}[r_1 + r_2 + \cdots + r_T|\theta] \tag{2.1}$$

で定義される．

連続タスクは連続系のプロセス制御など限界なく継続されるタスクをいう．仮定としてすべての政策 π のもとで背景のマルコフ決定過程はエルゴート性を有するものとする．連続タスクにおける平均報酬は，

$$\eta(\theta) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}[r_1 + r_2 + \cdots + r_T|\theta] \tag{2.2}$$

$$= \int_{\mathcal{S}} \rho^{\pi}(\mathbf{s}|\theta) \int_{\mathcal{A}} \mathcal{R}(\mathbf{s},\mathbf{a}) p_{\pi}(\mathbf{a}|\mathbf{s},\theta) d\mathbf{a} d\mathbf{s} \tag{2.3}$$

で定義される．ここで，$\rho^{\pi}(\mathbf{s}|\theta)$ は θ のもとでの状態 $\mathbf{s}$ の定常分布，$p_{\pi}(\mathbf{a}|\mathbf{s},\theta)$ は，

$$p_{\pi}(\mathbf{a}|\mathbf{s},\theta) = \int_{\mathcal{X}} p_X(\mathbf{x}|\mathbf{s}) \pi(\mathbf{a}|\mathbf{x},\theta) d\mathbf{x} \tag{2.4}$$

である．

本研究では，3 章においてエピソード的タスクを扱い，4 章および 5 章では連続タスクを扱う．

第3章　実数値GAによるインスタンスベース政策最適化と非ホロノミック系制御への適用

3.1　はじめに

制御則の理論的な導出が困難な問題クラスとして非ホロノミック系の安定化制御がある．非ホロノミック系とは積分できない拘束条件を持つ系であり，例として車両ロボット，水中ロボット，ロボットの指，宇宙ロボット，体操ロボットなどがあり，応用分野が広い [59]．非ホロノミック系は Brockett の定理により，滑らかな時不変の状態フィードバックでは安定化制御できないという性質を持つ [7]．これは，標準的な制御則の設計方法をそのまま適用できないことを意味し，主な接近法としては制御タスクを二つ以上の部分問題に分割し，それぞれの制御則とその切り換えを設計することになるが，一般的に制御則の設計は困難である [45]．非ホロノミック系の安定化制御は，その応用分野の広さと，制御則導出の困難さから制御理論の枠を越えて多くの研究がなされている [28, 56, 13, 60, 43]．

本章では，非ホロノミック系の安定化制御を対象タスクとして，連続状態・連続行動を扱うインスタンスベース政策 (Instance-Based Policy; IBP）最適化を効率的に行う実数値 GA の設計を目的とする．IBP は状態と行動の対であるインスタンスの集合と行動選択器によって政策を表現する．行動選択器は状態に照合するインスタンスを選択し，その行動を出力する．行動選択器が所与のとき，各インスタンスを状態行動空間上に配置することを IBP 最適化と呼ぶ．IBP はインスタンスの部分集合が局所的な制御則を表し，部分集合同士の境界が制御則の切り換えを表す．

IBP に対する交叉は，インスタンス集合によって定まる隣接関係をできるだけ維持する範囲内で子個体を生成することが望ましいが，通常の実数値 GA の枠組みでは交叉のためのペアリングがインスタンスの隣接関係を考慮したものではない．そのため，GA を用いた IBP 最適化の先駆的な研究 [13, 60, 43] では隣接関係の維持を行うことで探索の効率化を行っている．しかし，これらの IBP 最適化手法では，各インスタンスの持つパラメータの精緻化が難しく，安定化制御が獲得しにくいことが指摘されている [43]．これに対して [44] は IBP と特殊な状態フィードバックを組み合わせることで安定化制御が適切に行えることを示している．しかし，この方法にはタスクに強く依存したユーザパラメータが存在すること，パラメータの精緻化は行えないので安定化に至る軌道の発見確率を改善できないことが問題として残る．これらのことから，従来手法は学習効率を高めるために大域的探索能力および局所探索の精度を犠牲にしていると考えられる．

上述の現状認識に基づき，本章では IBP 最適化のための実数値 GA の新しい枠組みを提案し，非ホロノミック系の安定化制御へ適用することによりその妥当性と有用性を示す．

3.2 対象タスクおよび接近法

非ホロノミック系の制御に関する研究は，制御理論的な接近と強化学習による接近があるが，従来の多くの研究は一つの非ホロノミック系を対象とし，その系に特化した制御則の獲得を目的とすることが多い．それに対して，本研究では異なる性質を持つ複数の非ホロノミック系に対して制御則を獲得する接近法を目指すため，三つの特徴的なタスクを対象とする．

3.2.1 対象タスク

宇宙ロボットの姿勢制御

宇宙ロボットは角運動量保存則より，腕関節角度を変化させることで本体姿勢も変わる非ホロノミック系である [45]．状態は 2 つの腕の曲げ角 ϕ_1, ϕ_2，本体姿勢 φ の 3 次元，行

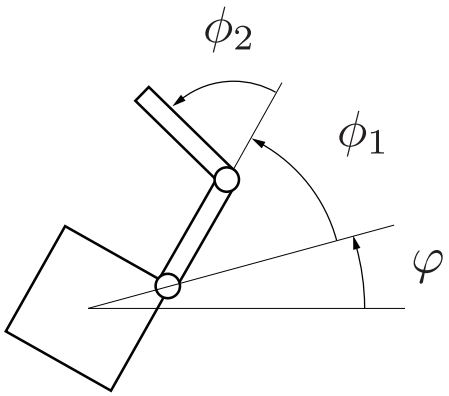

図 3.1: 宇宙ロボットのモデル

動は 2 つの腕の角速度 ω_1, ω_2 の 2 次元である．状態遷移は以下の式に基づく．

$$\begin{pmatrix} \dot{\phi_1} \\ \dot{\phi_2} \\ \dot{\varphi} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ f_1(\phi_1,\phi_2) & f_2(\phi_1,\phi_2) \end{pmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix} \tag{3.1}$$

ここで，f_1, f_2 は関節角の解析関数で，各リンクの幾何学パラメータと慣性パラメータから計算できる．初期状態は (ϕ_1,ϕ_2,φ)=(0,0,0)，目標状態を (0,0, $\pm\pi$) とする．意思決定および状態遷移は 0.1[sec] 毎に行われ，500 ステップを 1 エピソードとする．

報酬は即時報酬を以下のように与える．

$$r = -(\sum_i |\phi_i| + 2(\pi - |\varphi|) + 0.2 \sum_i |\omega_i|)/\pi \tag{3.2}$$

車両ロボットの車庫入れ

車両ロボットは局所的に真横に移動することができず，この拘束により非ホロノミック系となる [53]．状態は車両位置 x, y，車体角度 φ，操舵角度 ϕ の 4 次元，行動は車両速度

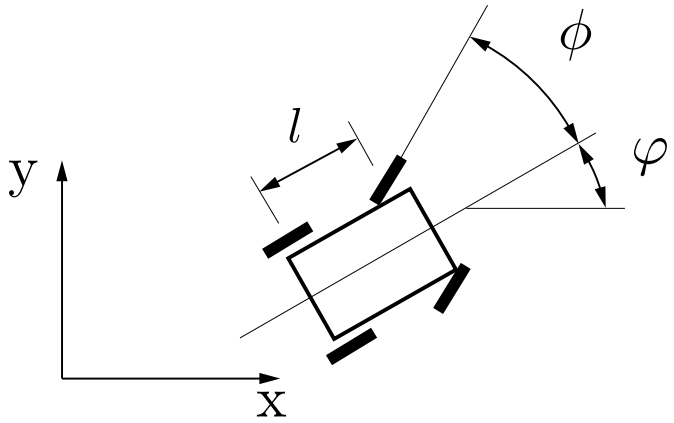

図 3.2: 車両ロボットのモデル

v，操舵角速度 ω の 2 次元である．状態遷移は以下の式に基づく．

$$\begin{pmatrix} \dot{\mathrm{x}} \\ \dot{\mathrm{y}} \\ \dot{\varphi} \\ \dot{\phi} \end{pmatrix} = \begin{pmatrix} \cos\varphi & 0 \\ \sin\varphi & 0 \\ \frac{\tan\phi}{l} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} v \\ \omega \end{pmatrix} \tag{3.3}$$

状態空間中の障害物が車庫の形状をしている．初期状態は (x, y, φ, ϕ)=(0.225, 0.05, π/2,0)，目標状態を (0.1, 0.2, 0,0) とする．意思決定および状態遷移は 0.1[sec] 毎に行われ，500 ステップを 1 エピソードとする．

報酬は即時報酬を以下のように与える．

$$\begin{aligned} r = -(&2\sqrt{(\mathrm{x}_{\mathrm{target}} - \mathrm{x})^2 + (\mathrm{y}_{\mathrm{target}} - \mathrm{y})^2}/0.3 \\ &+ (|\varphi| + 3|\phi|)/\pi + 0.1(|v| + |\omega|)) \end{aligned} \tag{3.4}$$

並列二重倒立振子の振上げ安定化

PDIP（並列二重倒立振子）の振上げ安定化とは，台車に取り付けられた 2 本の振子を，台車を横に振る力を制御することで倒立させる問題をいう [60]．振子の角度を保ったまま

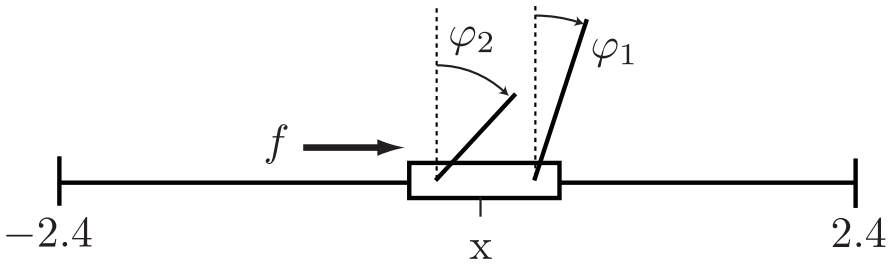

図 3.3: PDIP のモデル

台車を移動できないという幾何学的拘束を持つが可制御である [58] という非ホロノミック系である．状態は台車の位置 x と速度 $\dot{\mathrm{x}}$，2 つの振子の角度 φ_1, φ_2 と角速度 $\dot{\varphi}_1, \dot{\varphi}_2$ の 6 次元で，行動は台車に加える力 f の 1 次元である．状態遷移は以下の式に基づく．

$$\ddot{\mathrm{x}} = \frac{\sum_{i=1}^{2} m_i g \sin\varphi_i \cos\varphi_i - \frac{4}{3}\left(f + \sum_{i=1}^{2} l_i m_i \dot{\varphi}_i^2 \sin\varphi_i\right)}{\sum_{i=1}^{2} m_i \cos^2\varphi_i - \frac{4}{3}\left(m_0 + \sum_{i=1}^{2} m_i\right)} \tag{3.5}$$

$$\ddot{\varphi}_i = \frac{3(g\sin\varphi_i - \ddot{\mathrm{x}}\cos\varphi_i)}{4l_i} \tag{3.6}$$

初期状態は (x, $\dot{\mathrm{x}}$, φ_1, $\dot{\varphi}_1$, φ_2, $\dot{\varphi}_2$)=(0, 0, π, 0, π, 0)，目標状態を (#, 0, 0, 0, 0, 0) とする（#は任意の値）．意思決定および状態遷移は 0.01[s] 毎に行われ，500 ステップを 1 エピソードとする．

報酬は即時報酬を以下のように与える．

$$r_{\mathrm{clash}} = \begin{cases} 0 & if\ -2.4 < \mathrm{x} < 2.4 \\ -1000 & otherwise \end{cases} \tag{3.7}$$

$$r = \cos(\varphi_1) + \cos(\varphi_2) + r_{\mathrm{clash}} \tag{3.8}$$

[60] では r_{clash} の項が無いため，台車を定義域の端へ勢いよく衝突することで振り上げを行っているが，実システムでの運動を考えると衝突することは好ましくないため，衝突に対してペナルティを科している．これにより，[60] より難しいタスクになる．

3.2.2 非ホロノミック系の特徴

非ホロノミック系はBrockettの定理[7]より滑らかな時不変の状態フィードバックでは安定化できない．そのため二つ以上の制御則とそれらの切り換え制御が必要である．例えばPDIPの振上げ安定化には，振上げと安定化という相反する制御が要求されるサブタスクが含まれている．非ホロノミック系の安定化制御は，このような性質を共通して持っている．

一方で，対象とする三つのタスクは，非ホロノミック系となる拘束条件の性質が異なる．宇宙ロボットは運動学的拘束を，車両ロボットとPDIPは幾何学的拘束を受けることで非ホロノミック系となり，さらに宇宙ロボットと車両ロボットは位置と速度に拘束されている1階非ホロノミック系，PDIPは拘束条件に加速度まで含まれるので2階非ホロノミック系と呼ばれる．以上の運動力学的，幾何学的拘束の違いや，1階または2階の違いで系の性質が異なり，各性質ごとに制御則の設計方法が異なるのが一般的である[45]．

このように，非ホロノミック系の安定化制御は，二つ以上の制御が必要であるという共通の特徴に加え，非ホロノミック系となる拘束は複数あり，拘束により制御則の設計方法が異なるという特徴をもつ．以下では，制御理論的手法と強化学習手法に分けて，従来の接近法を概観する．

3.2.3 制御理論的手法による接近

非ホロノミック系ではBrockettの定理により滑らかな時不変の状態フィードバックで安定化制御できない．この条件を回避する制御則は，i) 滑らかな時変の状態フィードバック，ii) 滑らかでない状態フィードバックに分類できる[45]．これらを比較すると時変の制御は遅い収束に悩まされ[10]，滑らかでない制御を用いると，より早い収束が達成されることが知られている[6, 22]．このことから，滑らかでない制御による接近が好ましいと考えられる．

本研究で対象としているタスクへの接近として，車両ロボットはChained formに変換

して制御するのが一般的である [59, 53]．Chained form への変換による接近は，幾何学的拘束による 1 階非ホロノミック系である移動ロボット（トレーラなど）に対して汎用性がある．しかし，たとえば宇宙ロボットなどは必ずしも Chained form に変換できないため，[45]，この方法は性質の異なる非ホロノミック系へは適用できない．

宇宙ロボットに対する接近法は，[45] のように対象とする系に特化した二つの制御則とその切り換え制御則を設計するのが一般的である．そのため，他の宇宙ロボットや同じ性質に属する猫ひねり運動 [60] などにも適用できない．PDIP に関しては，目標状態付近での安定化を対象としているものがあるが，この接近法は対象を PDIP に限定している [58]．また，振り上げ安定化を対象とした例はない．

以上より，制御理論的な接近は系の性質に対する依存が強く，たとえモデル化しても制御則の一般的な設計方法がないため制御できないのが現状といえる．

3.2.4 強化学習手法による接近

価値関数法による非ホロノミック系の安定化制御への接近は [28, 56] がある．これらの手法では Acrobot の振上げ安定化に成功している．予めサブタスクごとに設計した低位の制御則を行動として，その低位の制御則の切り換えタイミングを学習により獲得している．低位の制御則はタスク依存であるため適応的に獲得することが好ましいが，そのような枠組みにはなっていない．

DPS による接近は政策表現モデルとして単一のニューラルネットがしばしば用いられてきたが [38, 21]，非ホロノミック系の安定化制御で使うにはタスクの性質に合わせた特別な工夫が必要である [42]．しかし，政策表現モデルをタスクごとに設計することは好ましくない．

近年，DPS の新しい接近法としてサブタスクでの制御則と，その切り換えを同時に獲得する IBP 最適化の枠組みが [13] により提案された．インスタンスは本来，過去の事例を表すが [33, 30]，IBP ではインスタンス集合を政策パラメータとみなして政策を表現する．[13] は IBP 最適化を Acrobot と PDIP の振上げ安定化に適用している．また，[60, 44] は

それぞれ異なる性質をもつ非ホロノミック系のタスクにIBP最適化を適用しており，非ホロノミック系の制御則を効率よく獲得する枠組みとしてIBP最適化が期待される．

3.3 インスタンスベース政策最適化

3.3.1 IBP最適化問題の定式化

状態観測と行動がそれぞれN次元，M次元の実数値ベクトルで表されるエピソード的タスクを対象とする．状態と行動の対でひとつのインスタンスを表す．サイズLのインスタンス集合$\mathcal{I} = \{I_1, ..., I_L\}$に対して状態観測$\mathbf{x}$が与えられたとき，1-nearest neighbor法により$\mathbf{x}$とのユークリッド距離が最小の状態$\mathbf{s}_i$をもつインスタンス$I_i = (\mathbf{s}_i, \mathbf{a}_i)$を選択し，その行動$\mathbf{a}_i$を出力する．

図3.4に状態と行動が2次元の場合のIBPの動作例を示す．図3.4の状態空間は1-nearest neighbor法の下でインスタンス集合によってボロノイ分割され，各領域で最近接のインスタンスの行動が選択される．政策実行時に参照されるインスタンスは**活性**である，参照されないものは**不活性**であるという．

IBP最適化問題は$\mathcal{I}$の各要素を状態・行動空間に最適配置することである．インスタンスI_iをベクトル$\vartheta_i \in \Re^{N+M}$で表せば，政策パラメータは$\theta = [\vartheta_1^{\mathrm{T}}, \vartheta_2^{\mathrm{T}}, ..., \vartheta_L^{\mathrm{T}}]^{\mathrm{T}} \in \Re^{L(N+M)}$である．教師信号である目標軌道が与えられないため，IBPの評価はエピソードを通しての報酬の総和によってなされる．そのため，IBP最適化問題は平均報酬$\eta(\theta)$を最大化する政策パラメータθを求める強化学習問題として定式化される．必要に応じてθに対応するインスタンス集合$\mathcal{I}$を用いて平均報酬を$\eta(\mathcal{I})$と表すことにする．

3.3.2 対象タスクのIBP最適化問題としての特徴

対象タスクにおけるIBP最適化問題には次のような特徴がある．

- **高次元性**　IBP最適化問題の次元数は$L(N+M)$である．インスタンス集合のサイ

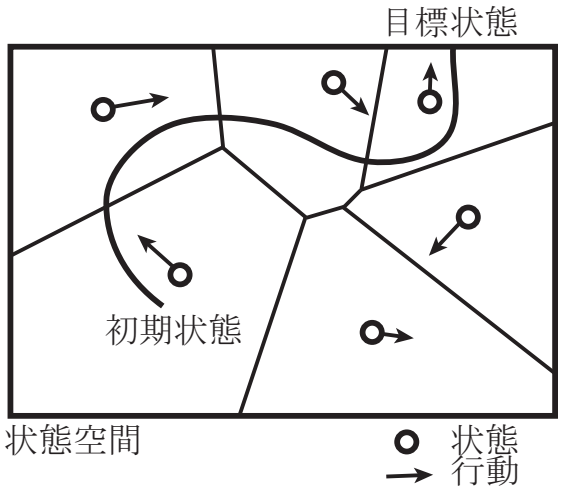

図 3.4: インスタンスベース政策の例：点の位置が状態を，矢印が行動を表し，この組がひとつのインスタンスに対応する

ズ L が 20 のとき，本研究で取り上げる宇宙ロボット，車両ロボット，倒立振子のタスクは，それぞれ 100 次元，120 次元，140 次元の高次元関数最適化問題となる．

- **多峰性**　対象タスクには目標状態へ至る軌道が数多くあり，適応度の景観は多峰性を有している．さらに，最適政策の実現に必要なインスタンス集合のサイズは未知であるため，十分なサイズのインスタンス集合を用意する必要がある．このような冗長性の導入は最適化問題の規模を大きくするだけでなく，局所最適解の数の増大，すなわち多峰性の程度を強めるため，最適化を難しいものにする．

- **インスタンスパラメータの強い依存関係**　状態と行動の対であるインスタンスのパラメータの次元数は $N + M$ である．上述の 3 つのタスクの次元数は，それぞれ 5 次元，6 次元，7 次元であるが，いずれも強い**変数間依存性**と**悪スケール性**を有する．

- **インスタンス間の局所的依存関係**　IBP における状態分割はインスタンスの相対的位置関係によって決まる．特に，**決定境界**に直接影響を与えるのは隣接するインスタンス同士であり，その意味でインスタンス間の依存関係は局所的であるといえる．

IBP 最適化の枠組みを設計するに当たっては，上述した特徴を十分考慮する必要がある．そのような観点から，3.3.3 では既存研究の妥当性を考察する．

3.3.3 既存研究

Ikeda[13] は，IBP の最適化問題をインスタンス集合の組み合わせ最適化問題として捉え，組合せ GA による最適化を行っている．インスタンスのパラメータ最適化は一部のインスタンスを再初期化するだけに留まり，ほとんど機能していない．

土谷ら [60] は，IBP の最適化問題をインスタンス集合の組み合わせ最適化と各インスタンスのパラメータ最適化の 2 つの部分問題に分け，各部分問題を担う GA を設計し，それらを統合したハイブリッド GA として SLIP を提案している．SLIP における組み合わせ GA はインスタンス集合で作られる決定境界の構造を最適化する機能を担い，実数値 GA は決定境界の位置と各領域での行動を最適化する機能を担う．SLIP はふたつの GA を一世代ごとに切り替えることで，両者の機能を相補的に働かせている．

SLIP における実数値 GA は交叉的突然変異 **INDX** によってインスタンスのパラメータ最適化を行っている．INDX は交叉的突然変異 XLM[51] における親個体数を 2 に限定したものにほぼ等価である．主親以外に 1 個体しか用いない INDX では，子個体生成分布は状態と行動の各次元に対して等方的であるため，INDX は変数間依存性や悪スケール性に対して適切に対処することができない．INDX の能力を補うために，SLIP では不活性インスタンスを再初期化するオペレータを導入している．これにより，SLIP の実数値 GA は辛うじてその機能を実現していると考えられる．

塩川ら [43] は，SLIP の INDX によるインスタンスパラメータの最適化の精度が低いために，安定化制御の性能が低い水準に留まっていることを指摘している．これは，上述したように，INDX が各次元のスケールの違いや変数間の依存関係に対応できないために，決定境界の位置や各領域での行動を微調整することが難しいためと考えられる．

塩川ら [44] は，SLIP に目標状態に対する状態フィードバック機能を付与した SLIP with Feedback を提案し，目標状態近傍での安定化制御の性能を改善している．しかし，この手

法は領域知識に基づいてタスクごとに調整する必要があり，汎用的な枠組みとはいえない．

3.4 実数値 GA の設計

前章での議論を踏まえて，IBP 最適化の設計方針を示し，それに基づく新しい枠組みを提案する．

3.4.1 設計方針

1. **実数値 GA 単独での IBP 最適化**　連続状態・連続行動を対象とする IBP 最適化問題は，本来，実数値 GA の枠組みだけで解くことが望ましい．そこで，本研究では性能の良い実数値 GA を構築できれば，SLIP における組み合わせ GA のような補完的な機能は不要との立場を取る．

2. **交叉的突然変異 XLM の拡張**　3.3.2 で述べたインスタンス間の依存関係の局所性を考慮して，交叉的突然変異の適用単位はインスタンスとする．インスタンスパラメータの強い依存関係に対処するためには，**統計量の遺伝**[46] という交叉の設計指針を考慮する必要があり，本研究ではこの指針に基づいて交叉的突然変異 XLM を拡張する．

3. **世代交代モデル CCM の拡張**　交叉に用いる親を選ぶ複製選択と次世代に残す個体を選ぶ生存選択を合わせて世代交代モデルという．複製選択については実数値 GA の標準的な世代交代モデル MGG[52] に準じて，集団から交叉的突然変異に必要な親個体をランダムに非復元抽出する．一方，生存選択については，交叉的突然変異が主親を中心としてその周辺に子個体を生成するので，主親と子個体は直系の関係にあることを考慮するものが望ましい．直系を重視した世代交代モデルとして CCM[23] がある．CCM の生存選択では主親またはその直系子個体を必ず次世代に残すことで多様性維持を図っている．しかし，3.3.2 で述べた冗長性が存在する IBP 最適化で

は，直系重視が探索性能の向上につながるとは限らない．そこで，本研究では直系保存の割合を調節できるように CCM の生存選択を拡張する．

3.4.2 探索オペレータの設計

交叉的突然変異 XLM の拡張

交叉的突然変異 XLM[51] は多親交叉 UNDX-m[47] をベースに提案されたオペレータである．個体が $\mathbf{w} \in \Re^n$ で表されるとき，m が変数の次元数 n に等しい UNDX-n による子個体生成は式 (3.9) で表される．

$$\mathbf{w}^c = \mathbf{w}^g + \sum_{i=1}^{n} \xi_i (\mathbf{w}^i - \mathbf{w}^g),\ \xi_i \sim \mathcal{N}(0, \sigma_\xi^2) \tag{3.9}$$

ここで，$\mathbf{w}^g$ は $n+1$ 個の個体 $\mathbf{w}^1, ..., \mathbf{w}^{n+1}$ の重心であり，$\sigma_\xi^2 = 1/n$ のとき，UNDX-n は統計量の遺伝を満たす．すなわち，UNDX-n は親個体群の平均値ベクトルと分散・共分散行列を保存するように子個体を生成する．

XLM は，m が変数の次元数 n に等しいとき，主親 $\mathbf{w}^p$ を中心に，式 (3.10) によって子個体を生成する．

$$\mathbf{w}^c = \mathbf{w}^p + \sum_{i=1}^{n} \xi_i (\mathbf{w}^i - \mathbf{w}^p),\ \xi_i \sim \mathcal{N}(0, \sigma_\xi^2) \tag{3.10}$$

式 (3.10) は式 (3.9) の $\mathbf{w}^g$ をすべて $\mathbf{w}^p$ に置き換えているために，統計量の遺伝を満たさない．

分散・共分散行列の保存は変数間依存性や悪スケール性を有する景観に対処する上で重要な役割を担っている．そこで，式 (3.10) を次のように修正する．

$$\mathbf{w}^c = \mathbf{w}^p + \sum_{i=1}^{n+k} \xi_i (\mathbf{w}^i - \mathbf{w}^g),\ \xi_i \sim \mathcal{N}(0, \sigma_\xi^2) \tag{3.11}$$

ここで，$\mathbf{w}^g$ は $n+k$ 個の個体 $\mathbf{w}^1, ..., \mathbf{w}^{n+k}$ の重心であり，$\sigma_\xi^2 = 1/(n+k)$ のとき，式

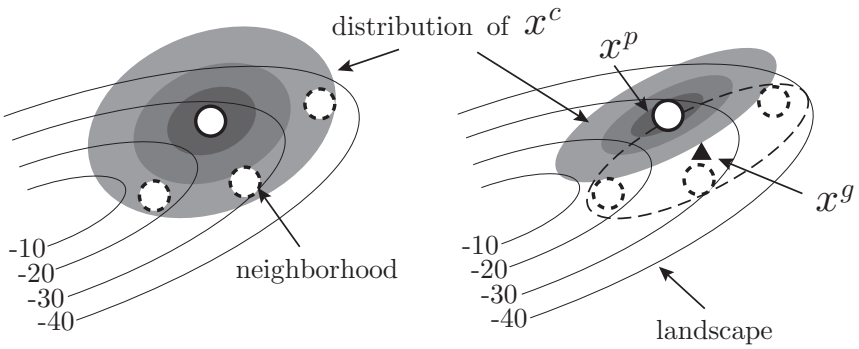

図 3.5: XLM による子個体生成分布（左）と，拡張 XLM による子個体生成分布（右）の概念図.

(3.11) は個体群 $\mathbf{w}^1, ..., \mathbf{w}^{n+k}$ の分散・共分散行列を保存する．式 (3.11) で表される交叉的突然変異を**拡張 XLM**[1] と呼ぶ．拡張 XLM において分布の中心となる個体 $\mathbf{w}^p$ を**主親**，それ以外の個体 $\mathbf{w}^1, ..., \mathbf{w}^{n+k}$ を**副親**という．

XLM と拡張 XLM による子個体生成分布の違いを図 3.5 に示す．XLM は親個体の分布に比べて子個体生成分布が広くなり冗長なサンプリングを行うことが予想されるのに対して，拡張 XLM は評価値の景観に沿った子個体生成分布を得ることが期待できる．

インスタンス間の局所的依存関係を考慮した子個体の生成

実数値 GA による IBP 最適化において個体はインスタンス集合 $\mathcal{I}$ であり，集団はその集合 $\mathcal{J} = \{\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_{n_{pop}}\}$ である．集団から選ばれた i 番目の親個体を $\mathcal{I}_i^p$ で表す．$\mathcal{I}_i^p$ は L 個のインスタンスの集合 $\{I_{i,1}^p, ..., I_{i,L}^p\}$ で特徴づけられる．このインスタンス集合の大部分を保ちながら局所的に変更を加えることで，有効な決定境界を残しつつ新たな決定境界を生成できる．IBP 最適化ではインスタンス間の局所的依存関係により，一部の変更が政策全体へ影響を与えないため局所的な更新が可能である．そこで，3.4.1 で述べた方針に従い，拡張 XLM の適用単位はインスタンスとする．状態が N 次元，行動が M 次元のとき，集団から $P \geq N + M + 2$ 個の個体を選べば，ある個体のインスタンスを主親に，残

[1] 拡張 XLM は多親交叉 REX($\mathcal{N}$,n+k)[54] において子個体生成の中心を $\mathbf{w}^g$ から $\mathbf{w}^p$ にシフトたものと等価である.

りの $P-1$ 個の個体から 1 つずつインスタンスを選んでそれらを副親とすれば，インスタンス単位で拡張 XLM を適用できる．

親 $\mathcal{I}_i^p$ の k 番目のインスタンス $I_{i,k}^p$ に対して，他の親 $\mathcal{I}_j^p$ の状態空間における最近接インスタンスを $nearest(I_{i,k}^p, \mathcal{I}_j^p)$ で表す．$P-1$ 個の各副親から最近接インスタンスを取り出して得られる集合を $\mathcal{Y}_{i,k}$ で表す．主親 $I_{i,k}^p$ を表すベクトルを ϑ^p，$\mathcal{Y}_{i,k}$ に含まれる各副親を表すベクトルを $\vartheta^1, ..., \vartheta^{P-1}$ として，式 (3.11) の拡張 XLM によりベクトル ϑ^c で表される新たなインスタンス $I_{i,k}^{c_j}$ を生成するものとする．

3.3.1 で述べたように，インスタンスは政策実行時に参照されるか否かで活性と不活性に類別される．IBP では活性インスタンス集合によって制御軌道が定まるので，これを洗練化するためには，活性インスタンスの状態と行動の両方に対し摂動を加える必要がある．しかし，すべての活性インスタンスを同時に変化させてしまうと，隣接関係を大きく変化させてしまう懸念がある．そこで，活性インスタンスについては適用確率 β で拡張 XLM による摂動を加えることとする．

不活性インスタンスは新しい軌道を発見するためのタネとしての役割が期待されているが，状態と行動の両方に摂動を加えた場合，活性インスタンス集合によって定まる隣接関係に対し破壊的に振舞う懸念がある．そこで，不活性インスタンスについては状態は不変として，行動に対してのみ摂動を加える．不活性インスタンスに対しては適用確率 1 で摂動を加えることとする．

以上の考えに基づく子個体生成の手続きを Algorithm 1 に示す．図 3.6 に拡張 XLM による子個体生成の概念図を示す．$\mathcal{I}_i^p$ を主親としてつくられる子 $\mathcal{I}_i^{c_j}$ を $\mathcal{I}_i^p$ の**直系子個体**，$\mathcal{I}_i^p$ と C 個の直系子個体を合わせたものを**直系家族**と呼ぶ．

3.4.3 世代交代モデルの設計

設計方針に従い，MGG[52] の複製選択と CCM[23] の生存選択をベースに，交叉的突然変異による IBP 最適化に適した世代交代モデルを設計する．IBP 最適化のための実数値 GA では直系を必ず保存することが適切かどうかは分らない．そこで，各直系から生存で

Algorithm 1 子個体生成の手続き

Input: $\mathcal{I}_i^p (i = 1, .., P)$：親個体
Output: $\mathcal{I}_i^{c_j} (i = 1, .., P, j = 1, ..., C)$：親 $\mathcal{I}_i^p$ の直系子個体
begin
for i=1,...,P **do**
  for j=1,...,P, $j \neq i$ **do**
    for k=1,...,L **do**
      $\mathcal{Y}_{i,k} \leftarrow \mathcal{Y}_{i,k} \cup \{nearest(I_{i,k}^p, \mathcal{I}_j^p)\}$
    end for
  end for
  for j=1,...,C **do**
    $\mathcal{I}_i^{c_j} \leftarrow \emptyset$
    for k=1,...,L **do**
      if $I_{i,k}^p$ is active **then**
        $I_{i,k}^{c_j} \leftarrow$ 拡張 XLM$(I_{i,k}^p, \mathcal{Y}_{i,k})$ (w.p. β)
      else
        $I_{i,k}^{c_j} \leftarrow$ 拡張 XLM$(I_{i,k}^p, \mathcal{Y}_{i,k})$
        $I_{i,k}^p$ の状態ベクトルを $I_{i,k}^{c_j}$ へ上書き
      end if
      $\mathcal{I}_i^{c_j} \leftarrow \mathcal{I}_i^{c_j} \cup \{I_{i,k}^{c_j}\}$
    end for
  end for
end for

きる最大の個体数を λ として直系保存の割合を調節できるように CCM を拡張することを考える．

まず直系家族内からベスト λ 個体を選択して $\lambda \times P$ 個体からなる家族を構成する．次に家族内からベスト P 個体を選択して生存個体とする．すなわち，直系家族内からの選択では直系を重視し，家族内からの選択では直系を考慮しない．

上述した生存選択と MGG の複製選択を合わせた世代交代モデルを**拡張 CCM** と呼ぶ．拡張 CCM では，優れた個体を含む直系家族からは多くの個体を選択することにより，悪平等を回避することができる．拡張 CCM の手続きを Algoritm 2 に示す．図 3.7 に拡張 CCM の概念図を示す．

Algorithm 2 拡張 CCM の手続き

複製選択

1: 集団から P 個体をランダムに非復元抽出する.

生存選択

1: **Input:** $\mathcal{I}_i^p (i = 1, .., P)$：親個体, $\mathcal{I}_i^{c_j} (j = 1, ..., C)$：親 $\mathcal{I}_i^p$ の直系子個体
2: **Output:** $\mathcal{I}_i^{\text{next}} (i = 1, ..., P)$：集団へ戻す個体
3: **begin**
4: 主親 $\mathcal{I}_i^p$ と直系子個体 $\mathcal{I}_i^{c_j} (j = 1, ..., C)$ を合わせて i 番目の直系家族とし，P 組の直系家族を用意する
5: 各直系家族からベスト λ 個体を選び $\lambda \times P$ 個体からなる家族とする
6: 家族内からベスト P 個体を選び $\mathcal{I}_i^{\text{next}} (i = 1, ..., P)$ とする

Algorithm 3 FLIP の手続き

1: **begin** 初期集団をランダムに生成する
2: **while** 終了条件を満たすまで **do**
3: 複製選択：〈**拡張 CCM の手続き**〉により集団から P 個体を取り出し，親とする
4: 子個体生成：〈**子個体生成の手続き**〉により各親の直系子個体を C 個ずつ生成する
5: 生存選択：〈**拡張 CCM の手続き**〉により P 個体を選び，集団へ戻す
6: **end while**

3.4.4 実数値 GA の構成

拡張 XLM を用いた子個体の生成及び拡張 CCM による世代交代に基づく IBP 最適化を行う実数値 GA の枠組みを **FLIP**[2] と名付ける．FLIP の手続きを Algorithm 3 に示す.

[2] Functionally sophisticated Learner for Instance-based Policy の略称で，機能的に洗練化された IBP 学習器という意味を表す.

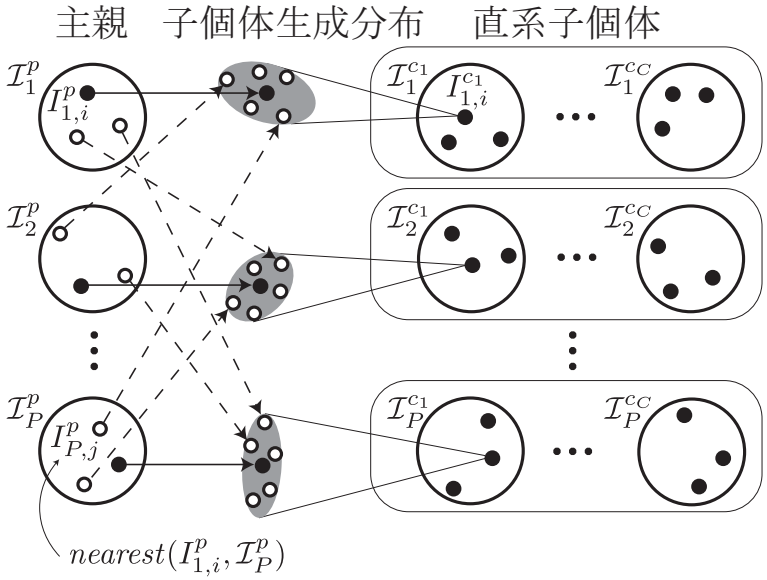

図 3.6: 拡張 XLM による子個体生成の概念図

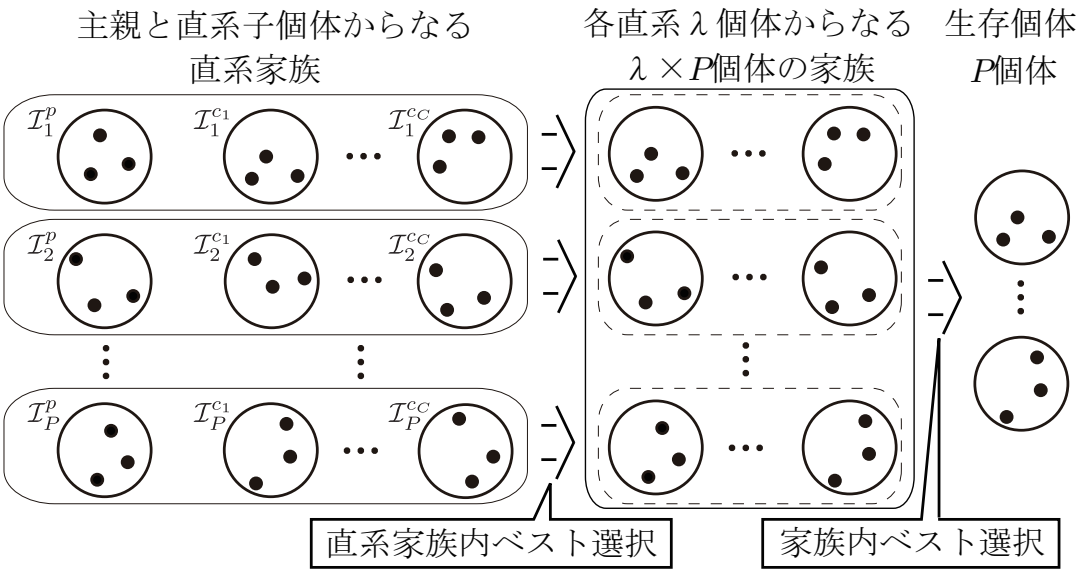

図 3.7: 拡張 CCM による生存選択の概念図

3.5 実験による評価

3.4 で提案した FLIP の性能を評価するために, 既存の IBP 最適化手法 SLIP，SLIP with Feedback との比較実験を 3.2.1 で示した 3 つのタスクで行う.

3.5.1 実験設定

各手法のパラメータ

各手法に共通の設定 共通のパタメータは SLIP で政策獲得が確認された値を選び，生成子個体数 C=100，インスタンス数 L=20 とした．集団サイズ n_{pop} は問題の複雑さに応じて設定し，宇宙ロボットの姿勢制御および車両ロボットの車庫入れに対しては n_{pop}=50 とし，PDIP の振上げ安定化に対しては n_{pop}=600 とした.

FLIP の設定 親個体数は 3 つのタスクで $P \geq N+M+2$ となる P=10 とした．拡張 XLM の適用確率は β=0.5，拡張 CCM の直系保存の係数を λ=2 とした.

SLIP および SLIP with Feedback の設定 SLIP のパラメータ設定は [60, 44] の推奨値を採用し，組合せ交叉 BDX の交叉率 ρ=0.7，交叉的突然変異 INDX の係数 α=0.1，適用確率 β=0.5，再初期化の適用確率 τ=0.1 とした.

SLIP with Feedback のゲインの範囲は [44] の推奨値を採用し，宇宙ロボットの姿勢制御では [-5, 5]，車両ロボットの車庫入れでは [-1, 1] とした．PDIP の振上げ安定化は SLIP with Freedback が適用できないので比較の対象外とした.

対象タスクのパラメータおよび定義域

各タスクの詳細設定は以下のように定めた.

宇宙ロボットの姿勢制御　本体の質量 $m_0 = 840$, 腕の質量 $m_1 = 39$,$m_2 = 139$，長さ $l_0 = 5860$,$l_1 = 75$,$l_2 = 262$, 重心周りの慣性モーメント $J_0 = 50$,$J_1 = 150$,$J_2 = 150$．状態は (ϕ_1,ϕ_2,φ) の 3 次元で定義域は ϕ_1=[-0.5π,0.5π], ϕ_2=[-0.6π,0.6π]，θ = [-π,π][rad]，行動は (ω_1,ω_2) の 2 次元で定義域は [-0.5π,0.5π] [rad/sec].

車両ロボットの車庫入れ　車輪間距離 l=0.04[m] とする．状態は (x,y,φ,ϕ) の 4 次元で定義域は x,y=[0, 0.3][m], φ=[-π, π][rad]，ϕ=[-π/6, π/6][rad]，行動は (v, ω) の 2 次元で定義域は v=[-0.1/3, 0.1],ω=[-1, 1][rad/sec]．障害物は (x1,y1,x2,y2) の矩形領域で (0.0, 0.0, 0.2, 0.19),(0.0, 0.21, 0.2, 0.3),(0.25, 0.0, 0.3, 0.3).

並列二重倒立振子の振上げ安定化　重力加速度 g=9.8 [m/s^2], 振子の質量 m_1=1.0, m_2=0.5 [kg], 台車の質量 m_0=1.0 [kg], 振子の長さ l_1=1.0, l_2=0.5 [m] とする．状態 (x, $\dot{\text{x}}$, φ_1, $\dot{\varphi}_1$, φ_2, $\dot{\varphi}_2$) の定義域はそれぞれ [-2.4, 2.4][m]，[-16, 16][m/s]，[-π, π][rad]，[-14, 14][rad/s]，[-π, π][rad]，[-14, 14][rad/s]，行動 f の定義域は [-40,40][N].

3.5.2　実験結果

基本性能

即時報酬の総和である評価値は，目標状態との差で与えられる．そのため，より多くの step で，より目標状態の近くにいることができれば評価値は高くなる．しかし，評価値だけではタスクの成否を判断できない．そこで，全ての状態変数の目標状態との差が，定義域の幅に対して 1% 以内である時，目標状態へ到達したとみなし，さらに到達してから 100step 以上到達を維持できた場合に安定化できていると考え，タスクに成功したとする．

表 3.1 にタスクごとに各手法が獲得した政策の評価値の 10 試行平均と標準偏差および成功回数を示す．表からすべてのタスクにおいて FLIP が他の手法に比べて最も評価値の高い政策を獲得していることが分かる．成功回数を比較しても，FLIP は安定化に優れる SLIP with Feedback より多く成功している．これは，宇宙ロボットの姿勢制御で顕著で

あり，目標状態へ到達することがSLIPに基づく手法では困難であることがわかる．それに対して，FLIPは安定化だけでなく軌道生成にも優れている．

図3.8，図3.10，図3.12は各手法の10試行平均の収束曲線をタスクごとに示したものである．これらの図からFLIP $\succ$ SLIP with Feedback $\succ$ SLIPの順に収束が早いことが分かる．例えば，宇宙ロボットではFLIPは凡そ2×10^5エピソードで目標状態に至る軌道を発見し，それ以降は軌道の洗練化を行っていると見られるのに対し，SLIPは目標状態に至る軌道の発見までに約2倍の4×10^5エピソード程度を費やしていると見られる．数値実験において環境とのインタラクションが計算の大部分を占めるため，最終性能に優れているだけでなく，探索の序盤から性能の改善が早いことが好ましい．探索の進みが早いこともFLIPがSLIPなどと比較して有利な点である．

以上の結果より，実数値GA単独のFLIPが，ハイブリッドGAのSLIPや領域知識を用いたSLIP with Feedbackに比べて，より優れた政策をより早く獲得することに成功しているといえる．

FLIPが獲得した政策の挙動

FLIPが各タスクで獲得した典型的な政策の挙動を図3.9，図3.11，図3.13に示す．

宇宙ロボットでは，図3.9から腕を回すように繰り返し動かしながら本体の姿勢を制御して，目標状態に到達していることが分かる．車両ロボットでは，図3.11から一度車庫の入り口を通り過ぎてその後車両の向きを大きく変化させて，その後バックで車庫に入ることで目標状態に到達していることが分かる．

PDIPでは，図3.13からレールの可動域の端で動作を切り返して短い振子を一回転させる間に長い振子を振り上げ，その後二本の振子を安定化している様子が分かる．PDIPではこれ以外の振り上げ方法が多数存在するが，FLIPが獲得した軌道での振り上げが最も早いことを確認している．一方，SLIPでは同じような振り上げ方法はいずれの試行でも獲得されておらず，さらに集団サイズを大きくとっても獲得されなかった．

表 3.1: 3 つの手法の性能比較：タスクごとに評価値の 10 試行平均と標準偏差，成功回数を表す．タスク名の後の数字は評価回数を表す

宇宙ロボットの姿勢制御 (8×10^5)			
	平均	標準偏差	成功回数
FLIP	**-258.54**	**5.58**	**9**
SLIP	-286.31	24.00	4
SLIP with Feedback	-266.93	19.69	6
車両ロボットの車庫入れ (6×10^5)			
	平均	標準偏差	成功回数
FLIP	**-91.72**	**11.25**	**9**
SLIP	-98.48	13.96	7
SLIP with Feedback	-95.28	16.60	8
PDIP(3×10^6)			
	平均	標準偏差	成功回数
FLIP	**718.29**	40.78	**10**
SLIP	644.77	**40.68**	9

以上より，FLIP は目標状態での安定化や適切なタイミングでの切り返しを担うインスタンスのパラメータを精度高く最適化していると考えられる．

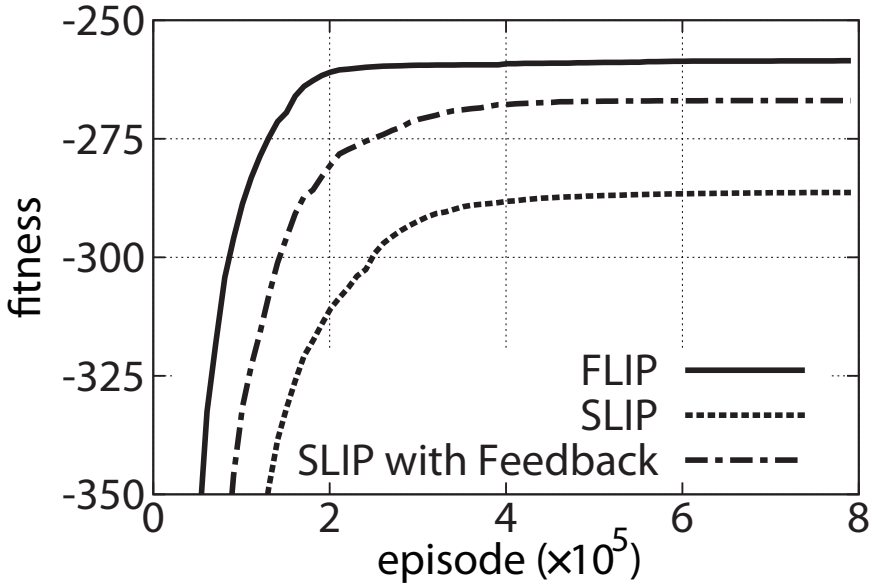

図 3.8: 宇宙ロボットの姿勢制御タスクにおける評価値の推移

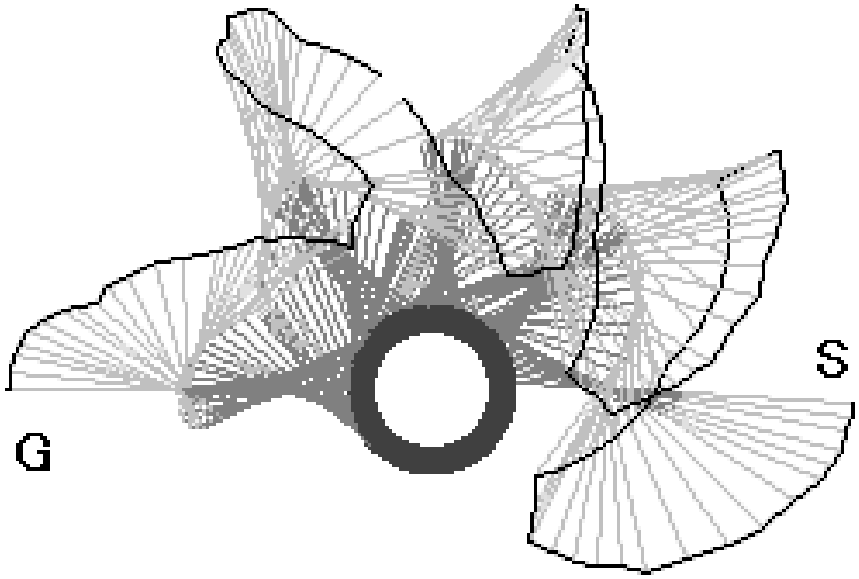

図 3.9: FLIP が獲得した政策による宇宙ロボットの挙動

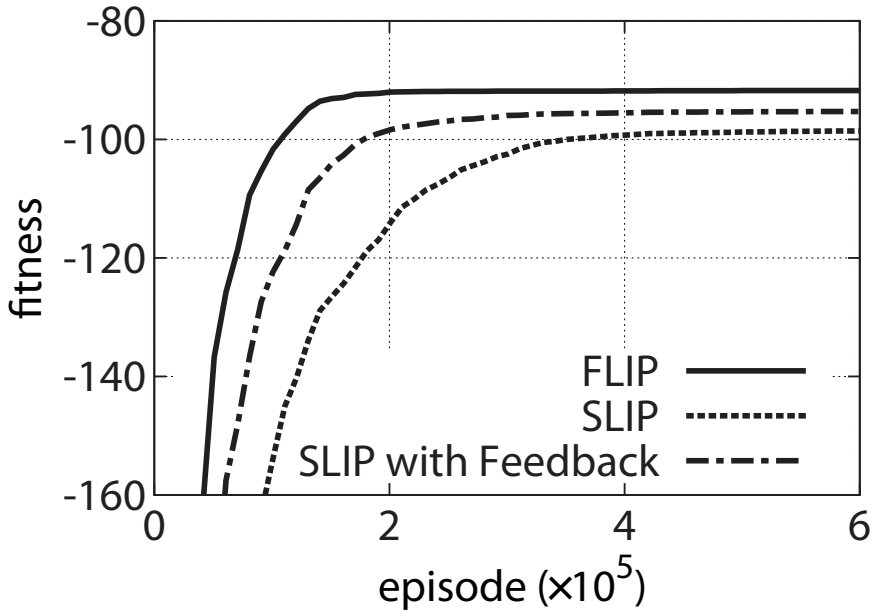

図 3.10: 車両ロボットの車庫入れタスクにおける評価値の推移

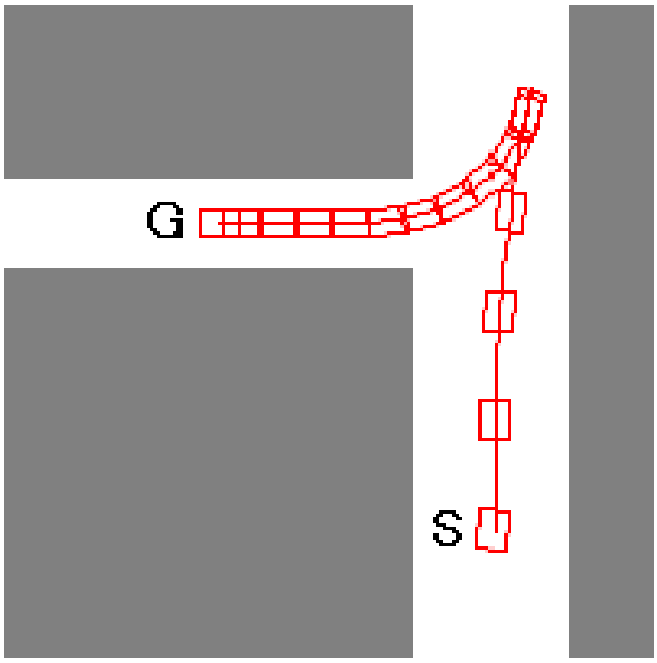

図 3.11: FLIP が獲得した政策による車両ロボットの挙動

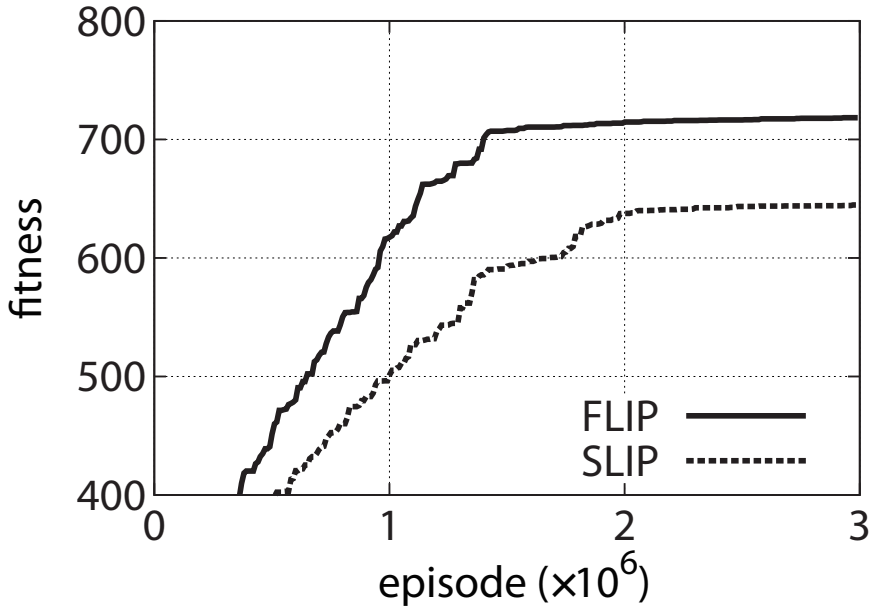

図 3.12: PDIP の振り上げ安定化タスクにおける評価値の推移

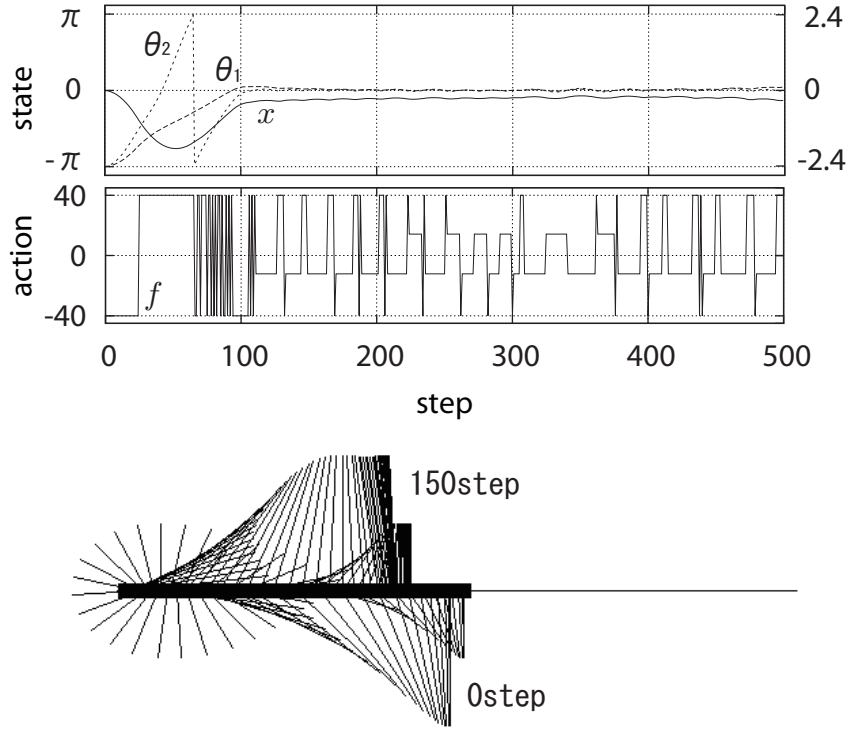

図 3.13: FLIP が獲得した政策よる PDIP の挙動

3.5.3 考察

FLIP と SLIP の探索性能の違いについて

FLIP と SLIP の探索性能の違いについて，政策改善率の推移および活性インスタンス数の推移を調べることにより考察する．

改善率の推移 探索が進むにつれて政策の改善率がどのように推移していくかを調べるために，**SR 値**(Success Rate)[60] を導入する．

$\mathcal{P}$ を親個体集合 $\{\mathcal{I}_1^p, \mathcal{I}_2^p, ..., \mathcal{I}_P^p\}$，$\mathcal{F}$ を親個体と子個体の集合 $\{\mathcal{I}_1^p, ..., \mathcal{I}_P^p, \mathcal{I}_1^{c_1}, ..., \mathcal{I}_P^{c_C}\}$ として，SR 値は探索オペレータ適用 1 世代当たりの評価値の改善確率として式 (3.12) で定義される．

$$SR := \Pr\left(\max_{\mathcal{I}\in\mathcal{F}} \eta(\mathcal{I}) - \max_{\mathcal{I}\in\mathcal{P}} \eta(\mathcal{I}) > 0\right) \tag{3.12}$$

図 3.14 に宇宙ロボットの姿勢制御タスクにおける SR 値の推移を示す．この図から，SLIP では探索が進むにつれて SR 値が急激に減少し，中盤から終盤にかけては 0.3 前後の低い水準に留まっているのに対し，FLIP では探索過程を通じて SR 値が 0.7 ∼ 0.8 の高い水準を維持している．すなわち，FLIP は SLIP に比べて探索中盤以降も獲得した軌道を着実に洗練化し続けており，それが評価値の高い政策の獲得につながっていると考えられる．

活性インスタンス数の推移 図 3.15 に宇宙ロボットの姿勢制御タスクにおける FLIP と SLIP の活性インスタンス数の推移を示す．

FLIP では活性インスタンス数が序盤で 10 個程度に増加した後，8 個前後に収束しているのに対し，SLIP では活性インスタンス数が単調増加し続けて終盤では FLIP の約 2 倍の 16 個前後にまで増えている．

FLIP では軌道獲得後の洗練化過程では活性インスタンス数を増やすことなく，活性インスタンスのパラメータを精緻化することによって性能向上が達成されているのに対し，SLIP ではパラメータの精緻化が難しいために活性インスタンス数を増やすことで性能を改善しているものと考えられる．

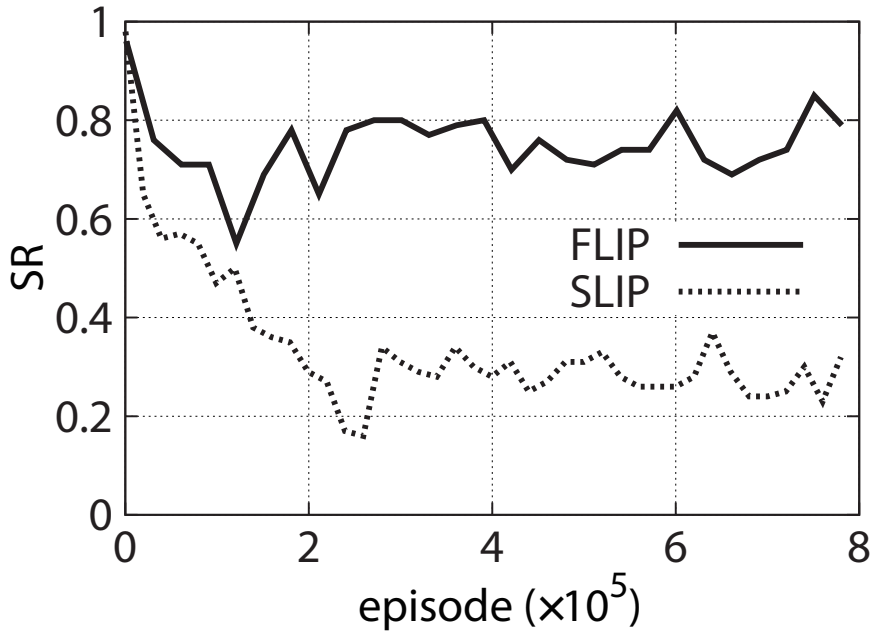

図 3.14: 宇宙ロボットの姿勢制御タスクにおける SR 値の推移

ちなみに，最初に与えるインスタンス数を半分の 10 程度に設定した場合，FLIP では 20 の場合とほぼ同程度の性能の政策をより早く獲得できるのに対し，SLIP では目標状態に到達する軌道を獲得することができない．

同様の傾向は他のタスクでも観察される．SLIP で活性インスタンスの数が増えてしまうのは，探索中盤以降でも組み合わせ GA が働いて必要以上に活性インスタンス数を増加させてしまうためと考えられる．

拡張 XLM の性能について

FLIP では交叉的突然変異として 3.4.2 で提案した拡張 XLM を採用している．拡張 XLM の探索オペレータとしての妥当性を検証するために，他の探索オペレータを用いた場合との比較実験を行う．

比較対象として，交叉的突然変異 XLM，INDX および交叉 REX($\mathcal{N}$,n+k) を選び性能を比較する．拡張 XLM，XLM，REX($\mathcal{N}$,n+k) については親個体数 $P = 10$，INDX については $P = 2$ とする．

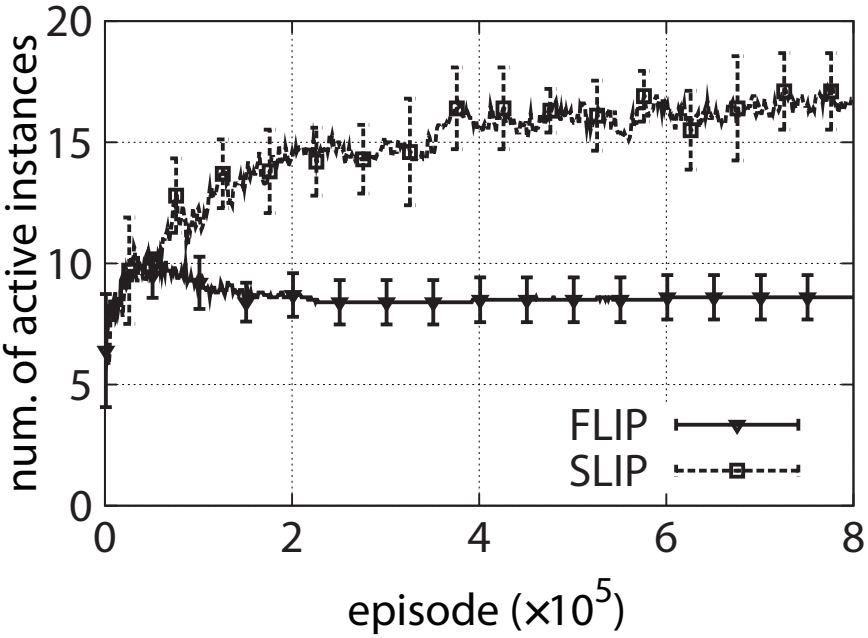

図 3.15: 宇宙ロボットの姿勢制御タスクにおける活性インスタンス数の推移

実験結果を表3.2に示す．この表から，すべてのタスクで拡張XLM $\succ$ XLM $\succ$ REX($\mathcal{N}$,n+k) $\succ$ INDX の順に評価値が高いことが分かる．タスクの成功回数に関しても同様の順序となっている．XLM は 2 番目に高い性能を示しているが，拡張 XLM に比べて標準偏差が大きい．これは，XLM の子個体生成分布が親個体群の分散・共分散行列を保存しないことに起因して，探索性能の安定性に影響が出たためと考えられる．2 親の INDX では変数間依存性や悪スケール性に対応できないため，最も性能が劣っていると考えられる．拡張 XLM と REX($\mathcal{N}$,n+k) の比較から，IBP 最適化では主親を中心に子個体を生成する交叉的突然変異の方が交叉より適していることが分かる．

拡張 CCM のパラメータ λ について

3.5.1 では，CCM の生存選択にパラメータ λ を導入した拡張 CCM において $\lambda = 2$ に設定している．ここでは，CCM との比較として λ=1 の場合，SLIP で採用しているモデルとの比較として λ=C+1 の場合，および λ=2,3,4 の場合について性能を比較する．

実験結果を表 3.3 に示す．CCM と等価である $\lambda = 1$ のときの性能が最も低い．$\lambda = 1$

の場合，評価値を無視してすべての直系を保存するために，集団の多様性は維持される．しかし，IBP 最適化では直系保存の悪平等という面が影響して，探索が停滞すると考えられる．

直系をまったく考慮しない $\lambda = C{+}1$ の場合は，性能が $\lambda = 1$ の場合に次いで低く，特に標準偏差が大きい．これは探索初期に多くの直系が淘汰されることで集団の多様性が失われ，探索が停滞してしまうためと考えられる．

$\lambda = 2 \sim 4$ の場合はいずれも性能が高い．これは，できるだけ直系を保存しつつ，しかし評価値の劣る直系は淘汰するという拡張 CCM の考え方の妥当性を示すものといえる．$\lambda = 2 \sim 4$ の中では，$\lambda = 2$ のとき最も性能が高く，標準偏差も小さいので，本研究では拡張 CCM のパラメータ値として $\lambda = 2$ を推奨する．

QL との比較

価値関数法との比較として，QL で実験を行う．QL は表形式の強化学習法であり，状態行動空間を予め分割しておく必要がある．

分割は予備実験により決定し，宇宙ロボットの姿勢制御に対しては，状態空間の各次元を等間隔に 20 分割し，行動は ω_1,ω_2 ともに {-0.5π,-0.25π,0,0.25π,0.5π} の 5 種類ずつとした．車両ロボットの車庫入れに対しては，状態空間の各次元を等間隔に 20 分割し，行動は v を {-0.1/3,0,0.1/3,0.2/3, 0.1}，ω を {-1,-0.5,0,0.5,1} の 5 種類ずつとした．PDIP の振り上げ安定化制御に対しては，状態空間の各次元を等間隔に 10 分割し，行動は {-40,-20,0,20,40} の 5 種類とした．全てのタスクで，学習率 α=0.01，割引率 γ=0.95，行動選択は ε-greedy(ε=0.01) として，実験をそれぞれ 10 試行行った．

表 3.4 に各タスクで最終的に得られた Q 値で greedy な行動選択による 1 エピソードあたりの総報酬和 (IBP 最適化における評価値) の平均と標準偏差，およびタスクの成功回数を示す．表 3.2 との比較から，QL は IBP 最適化手法と比べて評価値が低く，すべてのタスクで一度も成功していないことがわかる．

QL により獲得した振る舞いを確認すると，宇宙ロボットの姿勢制御や車両ロボットの

表 3.2: FLIP に組み込む探索オペレータの違いによる性能差（10 試行平均）.

宇宙ロボットの姿勢制御 (8×10^5)			
	平均	標準偏差	成功回数
拡張 XLM	**-258.54**	**5.58**	**9**
XLM	-276.41	42.90	8
INDX	-550.34	68.57	2
REX($\mathcal{N}$,$n+k$)	-303.83	47.06	8
車両ロボットの車庫入れ (6×10^5)			
	平均	標準偏差	成功回数
拡張 XLM	**-91.72**	**11.25**	**9**
XLM	-98.28	19.21	7
INDX	-391.60	149.20	0
REX($\mathcal{N}$,$n+k$)	-166.61	109.08	4
PDIP(3×10^6)			
	平均	標準偏差	成功回数
拡張 XLM	**718.29**	**40.78**	**10**
XLM	627.13	89.17	9
INDX	428.35	81.30	6
REX($\mathcal{N}$,$n+k$)	465.90	66.92	8

車庫入れでは初期状態付近で動作を終えてしまう．これらのタスクでは目標状態へ移動するために何度か即時報酬が低くなる行動をとらなくてはならないが，目標状態までの系列が長く，将来的な報酬を反映した価値の推定に失敗していると思われる．PDIP の振上げ安定化制御の振る舞いは二本の振子を振り回し続けるものであった．このタスクではサブタスクとして安定化をする行動を獲得できていない．

QL を非ホロノミック系のタスクに適用した結果，安定化制御は獲得されなかった．[28, 56] のように局所的に有効な制御則を行動として与えられないタスクでは，価値関数法による強化学習は困難であると思われ，IBP 最適化が有効である．

表 3.3: 拡張 CCM において λ の違いによる性能差（10 試行平均）.

宇宙ロボットの姿勢制御 (8×10^5)			
	平均	標準偏差	成功回数
1	-327.77	7.75	0
2	**-258.54**	**5.58**	**9**
3	-267.39	20.32	7
4	-278.13	28.16	7
C+1	-302.01	47.45	6
車両ロボットの車庫入れ (6×10^5)			
	平均	標準偏差	成功回数
1	-121.54	10.53	0
2	**-91.72**	**11.25**	**9**
3	-113.43	16.48	5
4	-116.08	24.35	3
C+1	-116.76	39.83	3
PDIP(3×10^6)			
	平均	標準偏差	成功回数
1	581.97	51.13	2
2	**718.29**	**40.78**	**10**
3	673.32	80.21	9
4	669.74	93.46	9
C+1	621.55	110.40	8

表 3.4: QL による性能:タスクごとに評価値の 10 試行平均と標準偏差，成功回数を表す. タスク名の後の数字は評価回数を表す

宇宙ロボットの姿勢制御 (8×10^5)		
平均	標準偏差	成功回数
-967.11	5.85	0
車両ロボットの車庫入れ (6×10^5)		
平均	標準偏差	成功回数
-681.55	12.28	0
PDIP(3×10^6)		
平均	標準偏差	成功回数
13.66	145.18	0

3.6 おわりに

本章では，連続状態・連続行動を扱うインスタンスベース政策最適化問題を対象に，その特徴は，高次元性，多峰性，インスタンスパラメータの強い依存関係，インスタンス間の局所的依存関係にあることを指摘し，これらの特徴を踏まえて実数値 GA に基づく新しい枠組み FLIP を提案した．FLIP は統計量の遺伝を満たす交叉的突然変異の拡張 XLM および直系保存の割合を調整できる世代交代モデルの拡張 CCM から構成される．

FLIP を宇宙ロボットの姿勢制御，車両ロボットの車庫入れ，並列二重倒立振子振り上げ安定化制御のタスクに適用した結果，従来手法の SLIP や SLIP with Feedback に比べて極めて高い性能を示すことを確認した．

今後は，FLIP をインスタンスベース政策以外の政策表現モデルへ適用できる枠組みへ拡張することが考えられる．現在の FLIP では交叉の際にインスタンスの役割を考慮したペアリングをアドホックな方法で行い探索効率を高めている．より汎用性の高いペアリングの方法を構築することで，多層パーセプトロンなど他の政策表現モデルに対しても探索効率の良い実数値 GA を実現することができると考えている．

第4章　集団を用いた政策勾配法と多峰性景観への適用

4.1　はじめに

直接政策探索法の一種である政策勾配法 (policy gradient methods)[40, 48, 5] は，政策パラメータに関する平均報酬の勾配を山登りする手法であり，政策表現モデルとしてニューラルネットなどを用いることにより連続な状態や行動を扱うことができる．さらに，政策勾配法は Q-learning に代表される価値関数法 [36] では扱いが困難な部分観測マルコフ決定過程 (partially observable MDPs; POMDPs) にも適用できる利点を持つ [29, 39, 15]．しかし，政策勾配法は局所探索に基づいているため，匍匐ロボットの前進動作獲得 [57]，並列二重倒立振子安定化制御 [11] など平均報酬の景観が多峰性を有する問題に適用した場合，最適政策の獲得に失敗する可能性が高い．

本章では，平均報酬が多峰性の景観を持つタスクを扱えるように，進化計算の考え方を導入した政策勾配法として Population-based Policy Gradient 法を提案する．提案手法では，エージェントが初期値の異なる複数の政策を持ち，それらの局所探索により得られた局所最適政策のうち，最良のものを選択することで大域的探索を行う．提案手法において，エージェントの持つ政策の数を多く取ることで大域的探索能力が向上すると考えられるが，全ての政策に対して局所探索を適用するため，環境とのインタラクション数が増加する．この問題を解決するために，重点サンプリングを用いた政策勾配法を使用する．これにより，ある政策の局所探索のために生成した履歴を利用して，それ以外の政策も局所探索を行うことができるため，環境とのインタラクション数の増加を抑えられる．

4.2 準備

4.2.1 政策勾配法

政策勾配法は，政策パラメータに関する平均報酬の勾配を山登りする局所探索法である．状態行動価値を，

$$Q_\theta(\mathbf{s},\mathbf{a}) = \mathbb{E}[\sum_{t=1}^{\infty} r_t - \eta(\theta)|\mathbf{s}_1 = \mathbf{s}, \mathbf{a}_1 = \mathbf{a}, \theta]$$

とし，すべての θ に関して $\pi(\mathbf{a}|\mathbf{x},\theta)$ が微分可能と仮定する．このとき，平均報酬の勾配は政策勾配定理 [35] により以下で与えられる．

$$\nabla_\theta \eta(\theta) = \int_{\mathcal{S}} \rho^\pi(\mathbf{s}|\theta) \int_{\mathcal{A}} Q_\theta(\mathbf{s},\mathbf{a}) \nabla_\theta p_\pi(\mathbf{a}|\mathbf{s},\theta) d\mathbf{a} d\mathbf{s}, \tag{4.1}$$

ここで，

$$\nabla_\theta p_\pi(\mathbf{a}|\mathbf{s},\theta) = \int_{\mathcal{X}} p_X(\mathbf{x}|\mathbf{s}) \nabla_\theta \pi(\mathbf{a}|\mathbf{x},\theta)$$

であり，完全知覚問題であれば $\nabla_\theta p_\pi(\mathbf{a}|\mathbf{s},\theta) = \nabla_\theta \pi(\mathbf{a}|\mathbf{s},\theta)$ である．

式 (4.1) の計算には $\rho^\pi(\mathbf{s}|\theta)$ や $p_X(\mathbf{x}|\mathbf{s})$，$Q_\theta(\mathbf{s},\mathbf{a})$ に関する知識が必要であるが，エージェントと環境のインタラクションから得られた履歴から式 (4.1) をモンテカルロ近似する GPOMDP アルゴリズム [5] が提案されている．GPOMDP アルゴリズムによる政策勾配の近似は以下のように計算される．

$$\nabla_\theta \eta(\theta) \approx \frac{1}{T} \sum_{t=1}^{T} r_t \mathbf{z}_t, \tag{4.2}$$

ここで，

$$\mathbf{z}_t = \beta \mathbf{z}_{t-1} + \mathbf{e}_t \tag{4.3}$$

は適正度の履歴 (eligibility trace) [48],

$$\mathbf{e}_t = \nabla_\theta \ln \pi(\mathbf{a}_t|\mathbf{x}_t, \theta) \tag{4.4}$$

は *characteristic eligibility* [40] と呼ばれ，$\beta(0 \leq \beta < 1)$ は割引率と呼ばれる係数である．GPOMDP アルゴリズムに基づく政策勾配法は，環境に関する知識を必要とせず，インタラクションにより得られた観測と行動，報酬から政策の改善方向を推定することができる．

4.2.2 重点サンプリングを用いた政策勾配の推定

重点サンプリングは，乱数を用いて積分値を近似するモンテカルロ積分の分散を減少させる統計手法として発展したものであり，異なる分布の期待値を推定する手法として強化学習での利用が注目されている [27]．ここでは，強化学習において重点サンプリングを用いた既存の研究について紹介する．

エージェントが行動選択で用いる政策を π_b で表し，挙動政策と呼ぶ．π_b を実行して得られた遷移の履歴 $h_t = \{(\mathbf{x}_1, \mathbf{a}_1), (\mathbf{x}_2, \mathbf{a}_2), ..., (\mathbf{x}_t, \mathbf{a}_t)\}$ から，他の政策 π の政策勾配 $\nabla_\theta \eta(\theta)$ を推定するために，重点サンプリングを利用することを考える．政策 π および政策 π_b に従ったとき，h_t が発生する確率を $p^\pi(h_t)$ と $p^{\pi_b}(h_t)$ で表す．このとき，$p^\pi(h_t)$ は，$\tilde{p}_\pi(\mathbf{a}_t|\mathbf{s}_t, \theta) = p_X(\mathbf{x}_t|\mathbf{s}_t)\pi(\mathbf{a}_t|\mathbf{x}_t, \theta)$ を用いて，

$$p^\pi(h_t) = p_I(\mathbf{s}_1) \prod_{k=1}^{t} \tilde{p}_\pi(\mathbf{a}_k|\mathbf{s}_k, \theta) p_T(\mathbf{s}_{k+1}|\mathbf{s}_k, \mathbf{a}_k) \tag{4.5}$$

となる [32]．$p^\pi(h_t)$ と $p^{\pi_b}(h_t)$ の違いを補正する係数 W_t は重要度 (importance weight) と呼ばれ，

$$W_t = \frac{p^\pi(h_t)}{p^{\pi_b}(h_t)} = \prod_{k=1}^{t} \frac{\pi(\mathbf{a}_k|\mathbf{x}_k, \theta)}{\pi_b(\mathbf{a}_k|\mathbf{x}_k)} = \prod_{k=1}^{t} w_k \tag{4.6}$$

で表される [27]．ここで，W_t を計算するためには環境に関する知識を必要とせず，観測

$\mathbf{x}_t$ および行動 $\mathbf{a}_t$ と，行動選択確率の比率だけを必要とすることに注意されたい．式 (2.3) の平均報酬は重要度を用いると，

$$\eta(\theta) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}[r_1 W_1 + r_2 W_2 + \cdots + r_T W_T | \pi_b]$$

と記述できる．[17] より，

$$\nabla_\theta W_t = W_t \sum_{k=1}^{t} \mathbf{e}_k$$

であることから，π_b の下で式 (4.1) のモンテカルロ近似は，式 (4.2) の代わりに次のように計算される．

$$\nabla_\theta \eta(\theta) \approx \frac{1}{T} \sum_{t=1}^{T} r_t \tilde{\mathbf{z}}_t, \tag{4.7}$$

ここで，

$$\tilde{\mathbf{z}}_t = \beta \tilde{\mathbf{z}}_{t-1} w_t + \mathbf{e}_t W_t$$

である．

重点サンプリングを導入することで，挙動政策とは異なる政策の改善が行える．しかし，重点サンプリングを用いた政策探索は，動作が不安定であることが知られている [32, 24]．そのため，実用上は重点サンプリングとヒューリスティクスを組合せて用いることが一般的である．

4.3 Population-based Policy Gradient 法

本章では，多点探索を行う政策勾配法の設計方針を示し，それに基づく新しい枠組みおよびアルゴリズムを提案する．

4.3.1 設計方針

集団を用いた大域的探索

既存の進化計算手法の探索効率を向上するために政策勾配法を補助的に組み込むのではなく，政策勾配法そのものを多峰性の景観を持つ政策に対処できる枠組みにすることが望ましいとの立場を取る．そこで，本論文では政策勾配法を，政策を個体と見なして集団で探索を行う枠組みに拡張する．

重点サンプリングを用いた経験の共有

集団を用いた枠組みにおいて，同じ局所解を重複して探索する個体が複数存在する場合に，類似した履歴を繰り返し生成することになり，インタラクションの回数を増加させる要因になる．本論文では，ロボット制御など実機への適用を想定し，政策更新の計算コストは環境とのインタラクションのコストに比べて十分に小さいものとして，他個体によって得られた経験を有効利用するために，重点サンプリングに基づいてすべての個体の政策パラメータを各時刻更新する．

4.3.2 Population-based Policy Gradient 法の枠組み

提案手法の枠組みを図 4.1 に示す．エージェントは政策パラメータ $\theta \in \Re^d$ と重みパラメータ $\vartheta \in \Re$ によって特徴づけられた n 個の個体 ξ をメンバーとする集団 Ξ を統括する．各個体の政策表現モデルと重み関数は共通であるとし，それぞれ $\pi(a|x,\theta)$，$m(\vartheta)$ で表す．i 番目の個体を $\xi_i \in \Xi$ で表し，その政策パラメータを $\theta^{(i)}$，重みパラメータを $\vartheta^{(i)}$ で表す．エージェントは，各時刻において，重み $m(\vartheta^{(i)})$ に従うルーレット選択によって個体をひとつ選択する．選択された個体を ξ_b で表し，そのパラメータ $\theta^{(b)}$ で表される政策に従って行動を選択し，環境とのインタラクションを行う．

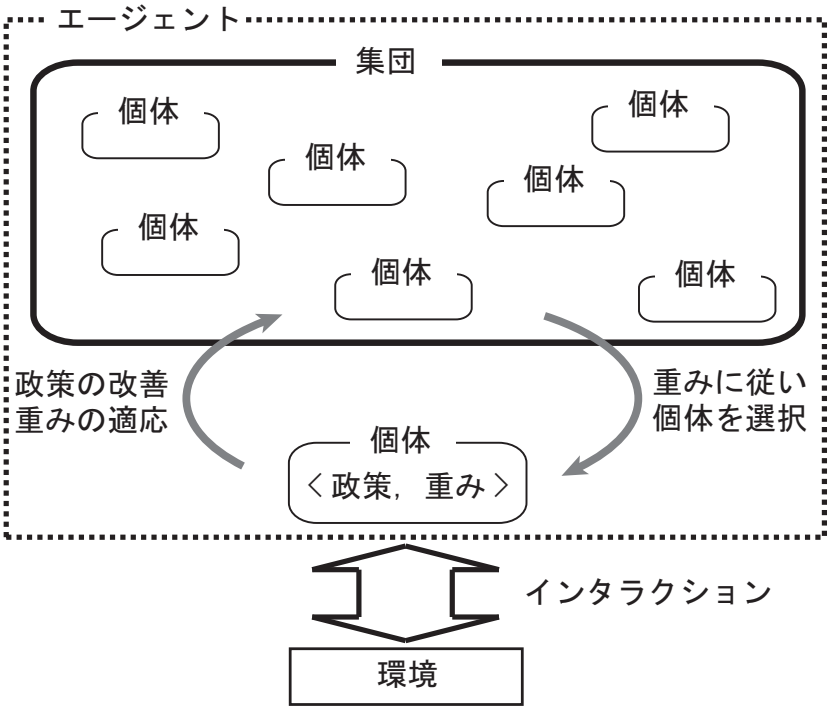

図 4.1: 提案手法の枠組み

提案手法では，多峰性の景観を持つ政策に対して，図 4.2 に示すように上位と下位の 2 階層での学習により大域的探索を行う．下位層では各個体の局所探索を行い，初期値に対応した局所最適政策を獲得する．上位層では，下位層で獲得した局所最適政策の中から最良の政策を獲得することにより大域的探索を行う．

本論文では，個体 ξ_i のパラメータ $\theta^{(i)}$ および $\vartheta^{(i)}$ の更新に GPOMDP アルゴリズム [5] に基づく手法を用いる．個体 ξ_b の政策パラメータおよび重みパラメータの *characteristic eligibility* は，

$$\mathbf{e}_{b,t} = \nabla_{\theta^{(b)}} \ln \pi(\mathbf{a}_t|\mathbf{x}_t, \theta^{(b)}), \epsilon_{b,t} = \nabla_{\vartheta^{(b)}} \ln M(\vartheta^{(b)}), \tag{4.8}$$

ここで，$M(\vartheta^{(b)}) = m(\vartheta^{(b)}) / \sum_i m(\vartheta^{(i)})$ である．また，個体 $\xi_i (i \neq b)$ では挙動政策 $\pi(\mathbf{a}|\mathbf{x}, \theta^{(b)})$ と政策パラメータ $\theta^{(i)}$ は独立であるため，

$$\bar{\mathbf{e}}_{i,t} = \nabla_{\theta^{(i)}} \ln \pi(\mathbf{a}_t|\mathbf{x}_t, \theta^{(b)}) = \mathbf{0}, \epsilon_{i,t} = \nabla_{\vartheta^{(i)}} \ln M(\vartheta^{(b)}) \tag{4.9}$$

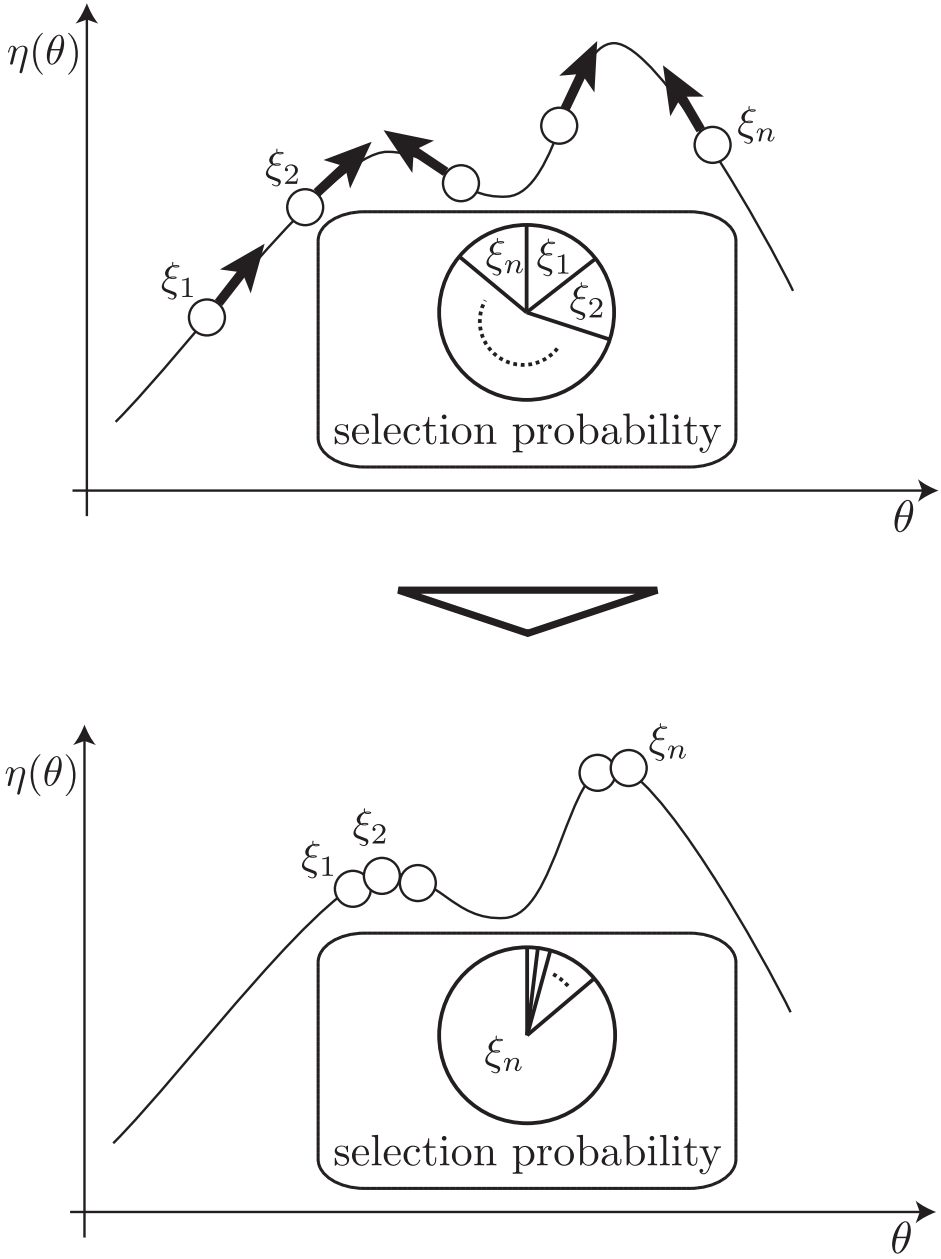

図 4.2: 提案手法による政策探索の概念図．上段：個体の政策はランダムに，選択確率は等確率に初期化する．下段：各個体は初期値に対応した局所最適政策を獲得し，集団中の最良個体の選択確率が最大となるように重みを適応する．

である．$\bar{\mathbf{e}}$ は行動選択確率の計算に $\theta^{(b)}$ を用いることを表す．

従来の政策勾配法は，提案手法において集団サイズが 1 の場合に相当する．提案手法は集団サイズを 1 より大きくすることで，多峰性の景観を持つ政策においても最適政策を獲得することが期待できる．

4.3.3 環境とのインタラクションの抑制

未知の環境において，予め適切な集団サイズを設計することは困難であるため，集団サイズは大きめに設計することが妥当である．そのような場合に，環境とのインタラクションの回数を抑制するために，重点サンプリングを用いた政策勾配法を導入する．各個体の行動選択確率は容易に計算できるため，時刻 t での個体 ξ_i と挙動政策の行動選択確率の比率を $w_{i,t} = \quad \pi(a_t|x_t, \theta^{(i)})/\pi_b(a_t|x_t)$ とし，重要度を $W_{i,t}$ とすれば，政策パラメータの適正度の履歴は，

$$\mathbf{e}_{i,t} = \nabla_{\theta^{(i)}} \ln \pi(\mathbf{a}_t|\mathbf{x}_t, \theta^{(i)}) \tag{4.10}$$

を用いて，

$$\tilde{\mathbf{z}}_{i,t} = \beta \tilde{\mathbf{z}}_{i,t} w_{i,t} + \mathbf{e}_{i,t} W_{i,t} \tag{4.11}$$

となる．重点サンプリングを用いた経験の共有により，$i \neq b$ でも $w_{i,t}$ や $W_{i,t}$ がゼロでない限り，政策改善を行うことができる．

ただし，4.2.2 で述べたとおり，政策勾配法で重点サンプリングを使用する際には，適切なヒューリスティクスを組合せる必要がある．ヒューリスティクスが必要な場合は 2 つある．ひとつは，W_t がゼロになる場合で，このとき学習が進まなくなる．もうひとつは，$W_t \gg 1$ となる場合で，政策勾配の推定値が大きくなり，政策パラメータが発散する．

これらの問題点へ対処するために，本論文では 2 つのヒューリスティクスを用いる．W_t がゼロになる場合への対処として，[26] は，各時刻において重要度に正の定数を加えることで，$W_t > 0$ としている．しかし，この方法では，勾配推定にバイアスが加わり，挙動政策と大きく異なる政策では，改善方向が推定されるとは限らない．そこで，提案手法は集団を用いた手法であることに着目して，時刻 t において挙動政策である個体 ξ_b の重要度を $W_t = 1$ にする**リセット**と呼ぶ操作を導入する．これは，履歴の生成に使用している挙動政策の学習は重要度が 1，つまり重点サンプリングを行わないことが自然であるのと，

他の個体は重要度がゼロのままであれば少なくとも改悪はしないためである．

$W_t \gg 1$ となる場合への対処として，[17] は勾配ベクトルを正規化することにより，政策パラメータの発散を抑制している．しかし，ここで使用されている方法では，政策表現モデルの設計に依存して政策改善が上手く行えないことが報告されている [17]．正規化が必要な理由は，政策更新 $\theta \leftarrow \theta + \alpha \nabla_\theta \eta(\theta)$ における学習率 α が，重点サンプリングを用いない場合でチューニングされていることが挙げられる．重点サンプリングを用いると，$W_t > 1$ である場合に式 (4.11) の第二項が想定よりも大きくなり，α を大きく設定したことになってしまう．そこで，適正度の履歴 $\tilde{\mathbf{z}}_{i,t}$ を $\min(1, 1/W_t)$ 倍する**リデュース**と呼ぶ操作を導入する．これにより，式 (4.11) の第二項の増大が抑えられ，学習率を減少させることに相当する効果が得られる．また，学習率が単調減少するため，挙動政策として選択された個体の学習率は初期値へ戻す．

提案手法における重点サンプリングを用いた政策探索の概念図を図 4.3 に示す．重点サンプリングを用いた勾配推定は挙動政策と同じ時系列を生成しやすい政策で有効である．また，そうでない政策は重要度がほぼゼロになり推定した勾配ベクトルが $\mathbf{0}$ に近くなるため効果が期待できない．集団を用いた枠組みでは，挙動政策と同じ局所解を探索している個体は同じ時系列を生成しやすいと予想され，重点サンプリングが有効に機能することが期待できる．そして，異なる局所解を探索している個体は，[17] で指摘されているように挙動政策との行動選択確率の比 $w_{i,t}$ がゼロに近い値になりやすいと予想され，推定した勾配を用いて政策を更新しても政策はほとんど変化しないと考えらる．そのため，改善は期待できないが少なくとも探索が妨げられることは無いと思われる．これらのことから，集団サイズが大きく同じ局所解を探索する個体数が多い場合に，経験共有による探索の効率化が期待できる．

4.3.4 アルゴリズムの提案

前節までに述べた集団を用いた政策勾配法の枠組み，および重点サンプリングを用いた手法を Population-based Policy Gradient 法 (以下，PPG) として提案する．PPG の手続

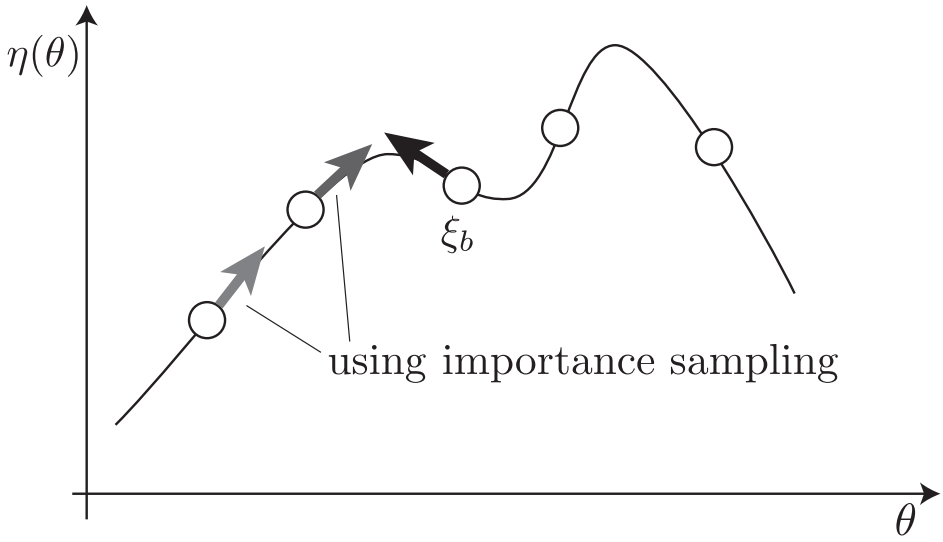

図 4.3: 提案手法における経験共有の概念図．環境とのインタラクションを行っている個体 ξ_b と同じ局所最適政策を探索している個体は経験共有により局所探索を進めることが期待できる．他の局所最適政策を探索している個体の改善は期待できないが，改悪もしないと考えられる．

きを Algorithm 1 に示す．PPG は集団サイズ $n = 1$ のとき，GPOMDP アルゴリズムに基づいて政策の逐次更新を行う OLPOMDP アルゴリズム [5] と等価である．Algorithm 1 の 10 行目から 18 行目までが，重点サンプリングの操作になる．12 行目から 16 行目までがヒューリスティクスの手続きであり，リデュースは w_t および W_t を減少することで実現している．17 行目で式 (4.11) を使用した式 (4.7) の計算により，重点サンプリングを用いた政策更新を行っている．19 行目は式 (4.2) の計算により個体の重みの更新をしている．

4.3.5 提案手法の位置付けと特徴

本論文では提案手法 PPG を進化計算手法の一種として位置付けている．PPG の手続きは図 4.4 に示すように，GA と対応付けて解釈することができる．まず，集団中からルーレット選択で個体を選択する操作は複製選択に相当する．次に，政策勾配法による政策改善は GA における突然変異による局所探索と対応している．そして，個体の重み適応では探索領域の絞込みが行われるため生存選択に相当すると考えられる．以上のように，集団を用いて GA と類似した手続きにより探索を行う PPG は，政策勾配法から派生した進化

Algorithm 4 Population-based Policy Gradient (PPG)

Notation:
$\pi(\mathbf{a}|\mathbf{x},\theta)$: 政策; $\mathbf{a}$: 行動, $\mathbf{x}$: 観測, θ: 政策パラメータ
$m(\vartheta)$: 個体の重み関数; ϑ: 重みパラメータ
$\tilde{\mathbf{z}}$: θ の適正度の履歴; $\mathbf{e} = \nabla_\theta \ln \pi(\mathbf{a}|\mathbf{x},\theta)$
ζ: ϑ の適正度の履歴; $\epsilon_i = \nabla_{\vartheta^{(i)}} \ln M(\vartheta^{(b)})$
Input: $\xi_i = \langle \pi(\mathbf{a}|\mathbf{x},\theta^{(i)}), m(\vartheta^{(i)}) \rangle$, $i = 1, ..., n$: 個体, n: 集団サイズ, γ: 割引率, α: 学習係数
Output: $\xi_i = \langle \pi(\mathbf{a}|\mathbf{x},\theta^{(i)}), m(\vartheta^{(i)}) \rangle$, $i = 1, ..., n$
begin x_1 を観測する.
for $t = 1, ...$ **do**
  個体選択: $\xi_b \sim m(\vartheta^{(b)}) / \sum_i m(\vartheta^{(i)})$
  行動選択: $\mathbf{a}_t \sim \pi(\mathbf{a}_t|\mathbf{x}_t,\theta^{(b)})$
  インタラクション: $\mathbf{a}_t$ を実行, $\mathbf{x}_{t+1}$ と r_t を得る.
  for $i = 1, ..., n$ **do**
    /* begin 重点サンプリングの操作 */
    重要度の計算: $w_{i,t} = \frac{\pi(\mathbf{a}_t|\mathbf{x}_t,\theta^{(i)})}{\pi(\mathbf{a}_t|\mathbf{x}_t,\theta^{(b)})}$, $W_{i,t} = W_{i,t-1} \times w_{i,t}$
    if $i == b$ **then**
      リセット: $W_{i,t} \leftarrow 1$
    else if $W_{i,t} > 1$ **then**
      リデュース: $w_{i,t} \leftarrow w_{i,t}/W_{i,t}$, $W_{i,t} \leftarrow 1$
    end if
    個体の政策パラメータの更新:
    $\tilde{\mathbf{z}}_{i,t} = \gamma \tilde{\mathbf{z}}_{i,t-1} w_{i,t} + \mathbf{e}_{i,t} W_{i,t}$,
    $\theta^{(i)} \leftarrow \theta^{(i)} + \alpha r_t \tilde{\mathbf{z}}_{i,t}$
    /* end 重点サンプリングの操作 */
    個体の重み更新:
    $\zeta_{i,t} = \gamma \zeta_{i,t-1} + \epsilon_{i,t}$,
    $\vartheta^{(i)} \leftarrow \vartheta^{(i)} + \alpha r_t \zeta_{i,t}$
  end for
end for

計算手法であると考えられる.

PPG の特徴を以下にまとめる.

- 多点探索を行うため, 多峰性の景観を持つ政策に適用した場合に, 集団サイズを大きくとることで最適政策を獲得する可能性が高くなる.

- 提案手法では使用する記憶領域のサイズが集団サイズで決まるため, 集団サイズは

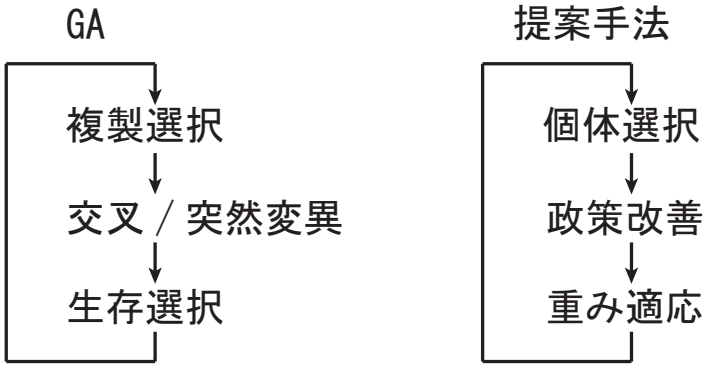

図 4.4: 進化計算手法としての提案手法の位置づけ

使用可能な計算資源に従い上限を定めることができる．

- 下位層の学習では，重点サンプリングにより，同じ局所解を探索している個体同士では経験を共有することができると考えられるため，集団サイズを大きく設定しても環境とのインタラクションを少なく抑えることが期待できる．

- しかし，上位層の学習は，各個体の政策が非定常である期間は非定常環境での学習となり最良個体の発見に失敗する可能性がある．また，対象タスクが UV 構造 [55] を持つ場合には，さらに最良個体の発見確率は低下することが予想される．これらの問題は集団サイズの増加では対処できないと考えられるため，上位層での学習方法を改良することは今後の課題である．

4.4 実験

4.4.1 実験計画

提案した PPG の有効性を確認するために，[37] の 2 分木タスクと，[15] の匍匐ロボットの前進動作獲得タスクに適用した．

2 分木タスクは解析的に局所最適政策が求まるタスクである．このタスクでは，集団サイズの増加により最適政策獲得の割合が増加することを確認する．また，PPG の収束速

度を，経験共有をしない設定の PPG と，局所最適政策の知識を利用してリスタートの条件を設計したマルチスタートローカルサーチと比較する．

匍匐ロボットは実機でも実験が行われている実問題である．このタスクでは，政策パラメータが高次元となる場合にも PPG が有効に機能することを確認する．

4.4.2 2 分木タスク

実験設定

2 分木タスクは実験 2 で用いる匍匐ロボットを単純化したベンチマークタスクである [37]．初期状態は 2 分木の根ノードにあたる状態 $s^{(0)}$ であり，行動 a_L で左，a_R で右の状態へ移動し，葉ノードにあたる状態に遷移したときに報酬を得て根ノードへ戻る．根ノードへ戻るまでを 1 エピソードとする．本論文では深さ 2 の 2 分木を扱う．環境は図 4.5 に示す 3 状態 2 行動 2 観測の POMDP 環境であり，報酬は [37] の設定に修正を加え，状態 $s^{(1)}$ で a_L を選択すると 1/2，状態 $s^{(2)}$ で a_R を選択すると 1，それ以外では 0 である．エージェントは現在の階層 $x^{(i)}$ を観測する．観測ベクトル $\mathbf{x}$ は i 番目の要素が 1 の単位ベクトルで与えられる．

政策表現モデルは以下のように設計した． i 番目の個体 ξ_i の重みを，

$$m(\vartheta^{(i)}) = \frac{1}{1 + \exp(\vartheta^{(i)})} \tag{4.12}$$

で表し，政策は，

$$\pi(a_L|\mathbf{x}, \theta^{(i)}) = \frac{1}{1 + \exp(-\sum_j \theta_j^{(i)} \mathbf{x}_j)}$$
$$\pi(a_R|\mathbf{x}, \theta^{(i)}) = 1 - \pi(a_L|\mathbf{x}, \theta^{(i)})$$

とする．この政策表現により平均報酬の景観は図 4.6 に示すように θ に対して多峰性関数になる．

初期集団は，政策パラメータの各次元の $[-1, 2]$ の領域を 4×4 に区切った格子点から一様ランダムに選択して生成した．ただし，集団中の個体には全て異なる初期値を与えた．学習率は $\alpha = 0.01$ とした．

結果

表 4.1 に集団サイズと最適政策の獲得確率を示す．ここでは，最終的な平均報酬が 0.49 以上になった試行では最適政策を獲得したと判断した．集団サイズの増加と共に最適政策の獲得確率も高くなっていることから，多点政策勾配法の上位層での学習による大域的探索が有効に機能していると考えられる．

図 4.7 は経験を共有する PPG を $n = 16$(share) とし，経験を共有しない PPG を $n = 16$(unshare)，およびマルチスタートローカルサーチを行う OLPOMDP を $n = 1$(restart) として，最適政策を獲得するまでの環境とのインタラクションの回数を比較している．$n = 1$(restart) のリスタートの条件は次のように設定した．2 分木タスクでは局所最適解が既知であり，$\theta = (\infty, \infty)$ で低いほうの報酬を獲得する局所解になる．$\theta = (5, 5)$ で 98.7%の確率で低いほうの報酬を獲得する局所解の経路を選択するので，$\theta_1 > 5$ かつ $\theta_2 > 5$ の場合には，平均報酬の低い局所最適政策に収束したと判断し，θ をランダムに再初期化する．

図から，経験を共有することで学習効率が向上していることが確認できる．図 4.8 は，$n = 16$(share) において，上位層の学習による最適政策を探索している (解析的に求めた勾配が最適解方向である) 個体の選択確率の推移を示している．この図からも，上位層での学習が有効であることが分かる．また，図 4.8 で最適政策を探索している個体の選択確率が約 60%まで高くなってから，図 4.7 において $n = 16$(share) の学習が加速している．同じ解を探索する個体から履歴が生成される頻度が高くなることで，経験共有の効果がより顕著になったと考えられる．

図 4.7 の $n = 16$(share) と $n = 1$(restart) の比較から，マルチスタートローカルサーチの方が分散は大きいが学習の立ち上がりが早いことが確認できる．しかし，最適政策へ収束するまでのインタラクションの回数は両者に大きな差はない．マルチスタートローカル

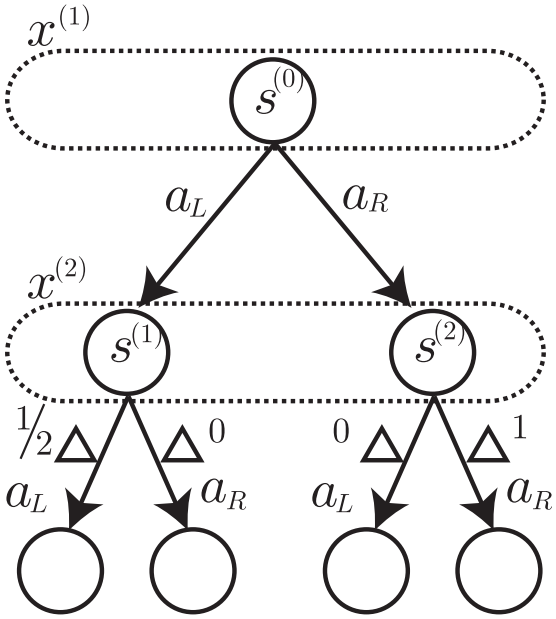

図 4.5: 深さ 2 の 2 分木タスク．根ノード $s(0)$ を初期状態として，葉ノードへ到達すると根ノードへ遷移する．エージェントは現在の階層 $x(i)$ を観測する．

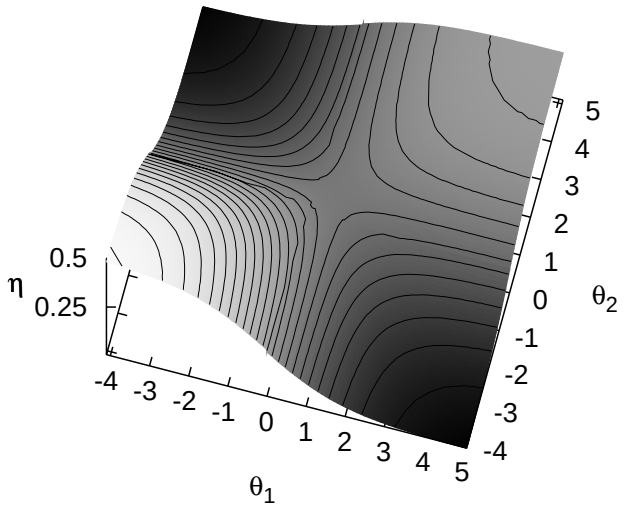

図 4.6: 深さ 2 の 2 分木タスクにおける政策パラメータに対する平均報酬の景観.

表 4.1: 2 分木タスクにおいて，各集団サイズにおける PPG が最適政策を獲得した回数 (全 100 試行).

population size (n)	1	2	4	8	16
optimum(%)	54	59	72	82	100

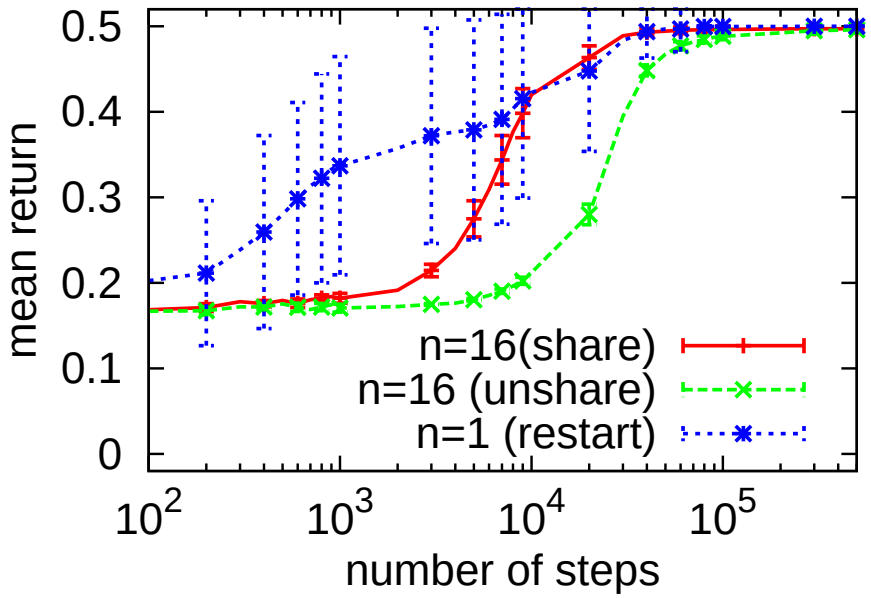

図 4.7: 2 分木タスクにおいて，集団サイズ $n = 16$ の PPG (経験共有あり/なし) と，従来手法 OLPOMDP をマルチスタートローカルサーチの枠組みで使用した場合の学習曲線 (100 試行平均).

サーチの適切なリスタートの条件の設計は，タスクに関する領域知識を必要とする．そのため，領域知識を使うことのできない実問題においては，PPG のほうが実用的であると考えられる．

考察

重要度の推移　ヒューリスティクスの設計において想定している状況がどのような場合に発生するかを検証する．3 つのシナリオにおいてヒューリスティクスを使用しないで重要度 W_t を計算し，その推移を確認する．

case1 π_b が π より局所最適政策に近い

case2 π_b が π より局所最適政策から遠い

case3 π_b が π と異なる局所最適政策を探索している

図 4.9 に，それぞれのシナリオでの政策 π の重要度の推移を示している．

case1 から，自身より局所最適政策に近い個体により生成された履歴で W_t を計算すると，W_t が緩やかに減少する傾向が確認できる．case2 の，自身より遠い個体によるものでは，W_t は最終的に減少するが，一時的に増加する試行があり，そのような試行では $W_t = 20$ 程度まで増加した．case3 より，異なる政策を探索している個体同士では，W_t が 1step 目から非常に小さくなることが確認できる．

以上より，case3 の異なる政策を探索している個体同士では経験の共有は困難であること，case1 や case2 でも，履歴が長くなると $W_t \simeq 0$ になることや，case2 では一時的に $W_t \gg 1$ になることが確認された．

ヒューリスティクスの妥当性　PPG において最適政策が獲得できる $n = 16$ で，重要度のリセットを行わない場合について図 4.10 に 100 試行分の学習曲線を重ねて示している．この結果から，リセットを行わない場合には，局所最適政策ですらない政策で学習が停滞していることが確認できる．これは，履歴が長くなると $W_t \simeq 0$ となることが起因していると予想される．図 4.7 でリセットを行う場合には最適政策が獲得できていることから，PPG においてリセットを導入することが妥当であることがわかる．

リデュースに関しては，このタスクでの影響が有意ではなかったため，その妥当性の検証は次の匍匐ロボットを使ったタスクで行う．

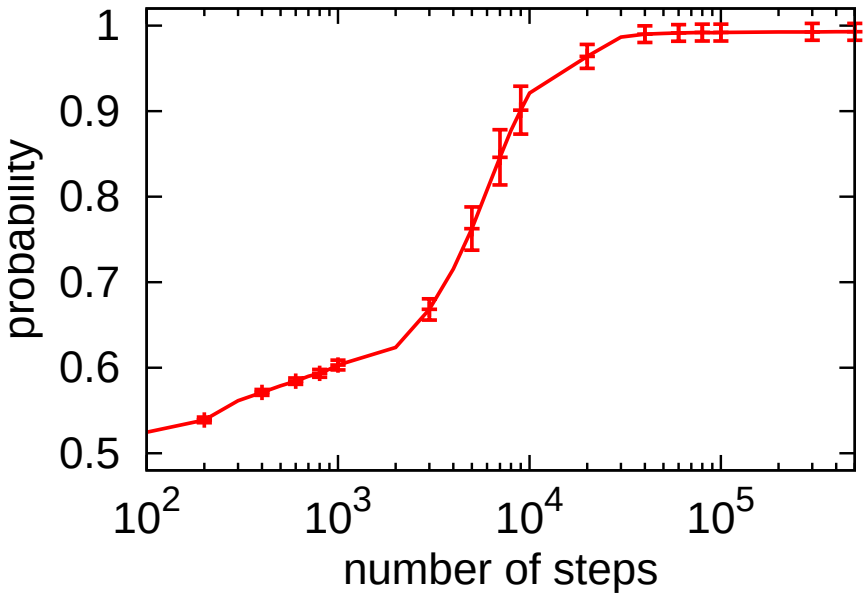

図 4.8: 2 分木タスクにおいて，集団サイズ $n = 16$ の PPG (経験共有あり) の，最適解を探索している個体の選択確率の推移 (100 試行平均).

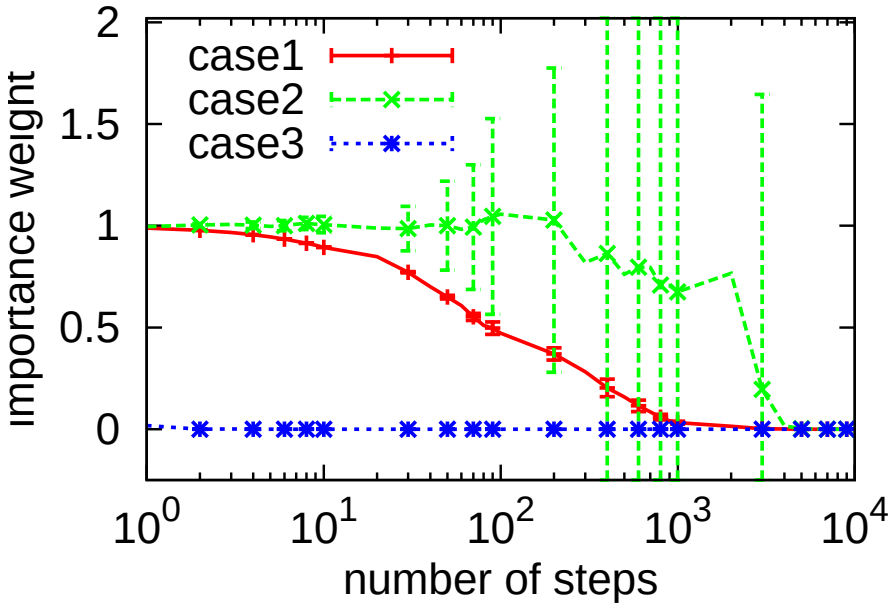

図 4.9: 2 分木タスクにおいて，様々な挙動政策の下での重要度の推移 (100 試行平均)．重要度を計算する政策は $\theta = (-4, -4)^{\mathrm{T}}$．挙動政策のパラメータは，case1: $\theta^{(b)} = (-5, -5)^{\mathrm{T}}$, case2: $\theta^{(b)} = (-3, -3)^{\mathrm{T}}$，case3: $\theta^{(b)} = (5, 5)^{\mathrm{T}}$ とした．それぞれ政策パラメータは固定とし，学習は行わない．

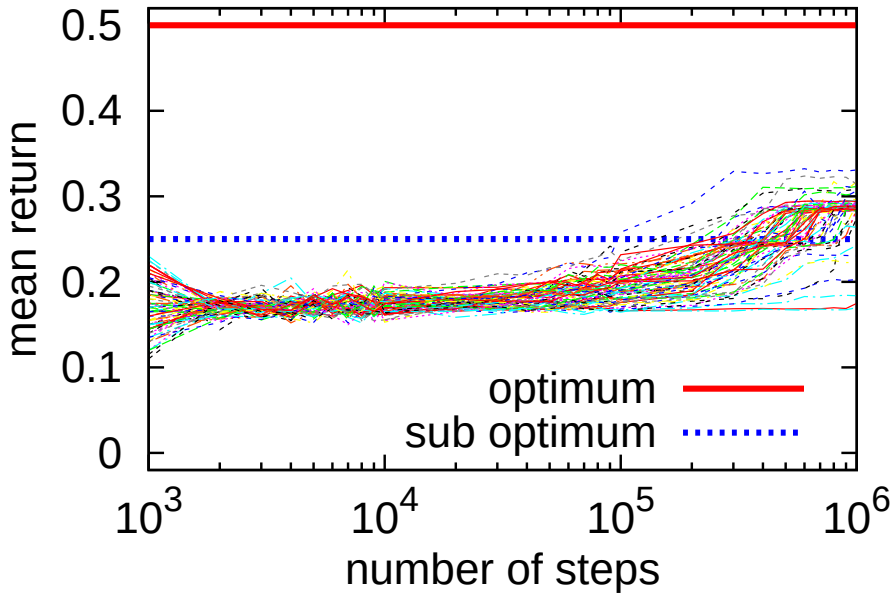

図 4.10: 2 分木タスクにおいて，重要度のリセットを行うヒューリスティクスを使わない場合の PPG の学習曲線 (全 100 試行)．集団サイズ $n = 16$，経験共有あり．

4.4.3 匍匐ロボットの前進動作獲得タスク

実験設定

[50] により実機での学習も行われている匍匐ロボットのシミュレータ [15] により，複雑なダイナミクスを有する連続タスクにおいて，提案手法の有用性を検証する．図 4.11 に示す 2 本のアームを持つ匍匐ロボットの前進動作を獲得するタスクに PPG を適用する．このタスクでは，適切なリスタートの条件設計が困難であるため，マルチスタートローカルサーチとの比較は行わない．

各アーム毎に行動次元数は 2 次元で，それぞれ $[0, 1]$ の区間を持つ．k 番目の足に関する観測は各関節の角度 (各要素 $[0, 1]$) の 2 次元ベクトル $(x_{k,1}, x_{k,2})^{\mathrm{T}}$ で，エージェントの観測する特徴ベクトルは $\mathbf{x} = (x_{1,1}, x_{1,2}, x_{2,1}, x_{2,2}, 1)^{\mathrm{T}}$ である．最後の要素はバイアス項である．j 番目の関節に関するパラメータを $\theta_{j,\mu} \in \Re^5$，$\theta_{j,\sigma} \in \Re$ で表すことにする．

政策表現モデルは以下のように設計した．個体の重みは 2 分木タスクと同じ式 (4.12) を用いて，政策は次のような単層パーセプトロンを用いる．j 番目の関節の行動 a_j は $\theta_{j,\mu}$，$\theta_{j,\sigma}$ を用いて，

$$
\begin{aligned}
\mu &= \frac{1}{1 + \exp(-\theta_{j,\mu}^{\mathrm{T}} \mathbf{x})}, \\
\sigma &= \frac{1}{1 + \exp(-\theta_{j,\sigma})} + 0.001, \\
a_j &= \mathcal{N}(\mu, \sigma)
\end{aligned}
$$

により生成する．ここで，$\mathcal{N}(\mu, \sigma)$ は中心 μ，標準偏差 σ の正規分布である．エージェントが $[0, 1]$ の範囲外の行動を実行しても，環境では制限の範囲でしか実行されないとした．政策パラメータの次元数は $d = 24$ である．

この政策表現モデルは分散 σ が 0 へ近づくと $\mathbf{e}_t$ が発散することに注意しなければならない．この問題に対処するためのひとつの方法として学習係数を σ^2 に比例させる方法 [40] が一般に用いられるため，本研究でもその設定を採用する．

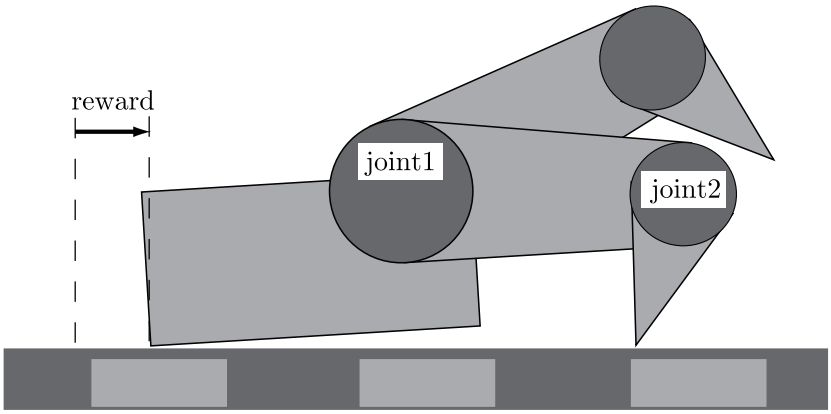

図 4.11: アームが 2 本の匍匐ロボット．胴体と 2 つの関節から成るアームを持つ．右方向へ移動することで報酬が得られる．

初期集団は，$\theta_{j,\mu}$ の各要素を $[-0.1, 0.1]$ の範囲で一様ランダムに与えて生成した．$\theta_{j,\sigma}$ の初期値は 0 として，$\alpha = 0.01$ とした．報酬は，各時刻に前進した距離が正の報酬，後退の場合は負の報酬が与えられる．

結果

図 4.12，図 4.13，図 4.14 に PPG の集団サイズを $n = 1, 25, 100$ とした場合の学習曲線と最終性能のヒストグラムを示している．学習曲線は 100 試行の結果を重ねて示している．ヒストグラムは学習を打ち切る 8×10^5step までの最後の 10^4step での即時報酬の平均値を，区間の幅を 1 として区切り，100 試行分の結果の頻度を示している．

まず，$n = 1$ では，いくつかの試行で性能の低い政策に収束していることが確認できる．ロボットの振舞いを確認したところ，性能の高い政策はアームを 2 本使った動作であり，性能の低い政策はアームを 1 本しか使わない動作であった．このことから，このタスクには大きく分けて 2 つの局所最適政策が存在すると予想される．

次に，$n = 25$ まで集団サイズを増やしたところ，性能の低い政策に収束する試行は無くなった．このことから，PPG の上位層での学習が有効に機能していることが分かる．また，学習効率は $n = 1$ の場合とほとんど変わらないことから，下位層での重点サンプリングの効果も確認できる．

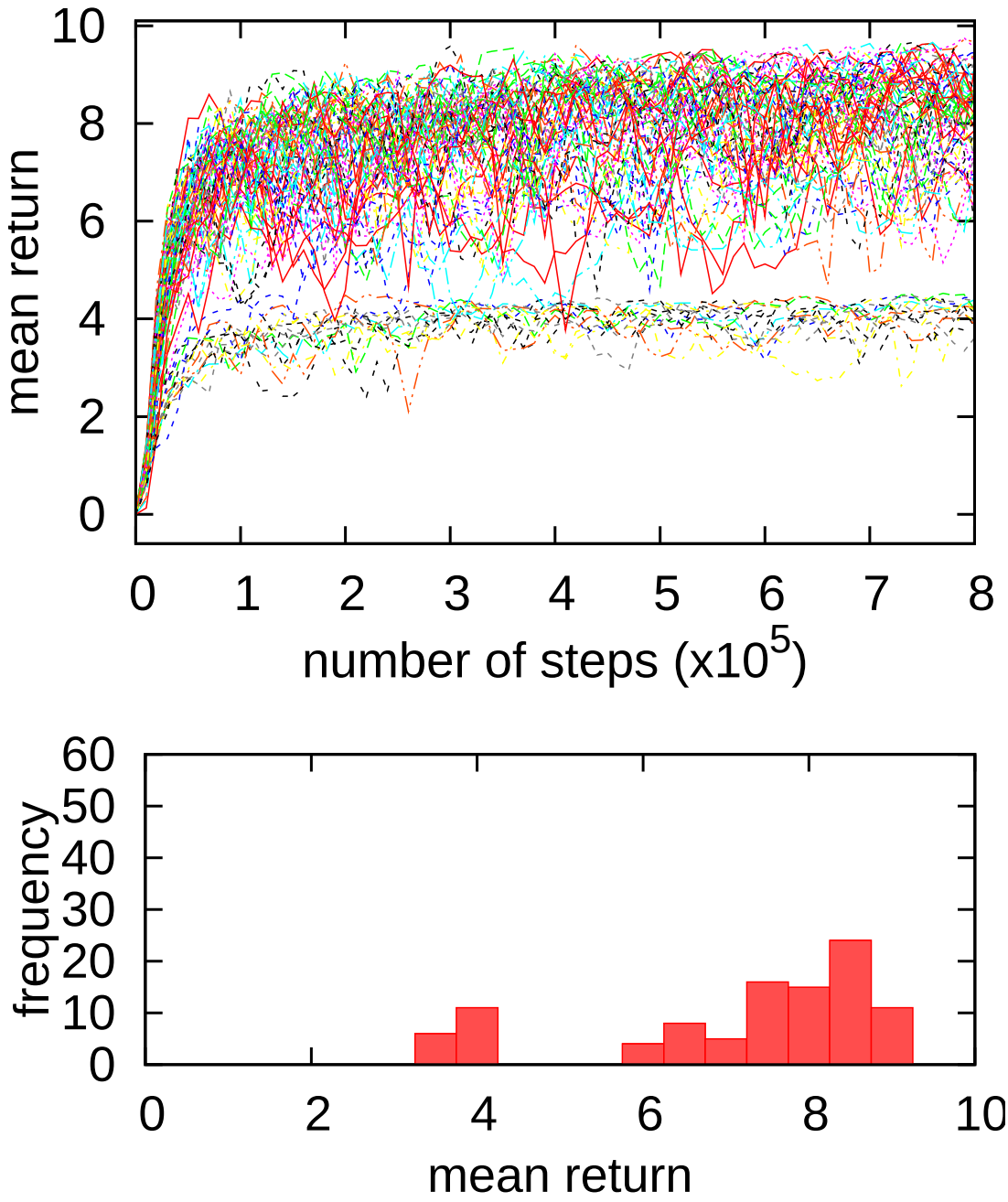

図 4.12: 集団サイズ $n = 1$ における匍匐ロボットでの学習曲線と学習結果のヒストグラム (全 100 試行)．上段：各試行での学習曲線．下段：各試行の最終性能のヒストグラム．

最後に，$n = 100$ では，学習曲線を見ると $n = 1$ や $n = 25$ と比べて，学習効率は低下したものの，学習曲線のばらつきが小さくなっている．ヒストグラムからも，獲得した政策の性能のばらつきが小さくなり，$n = 25$ と比べて性能が向上していることが確認できる．

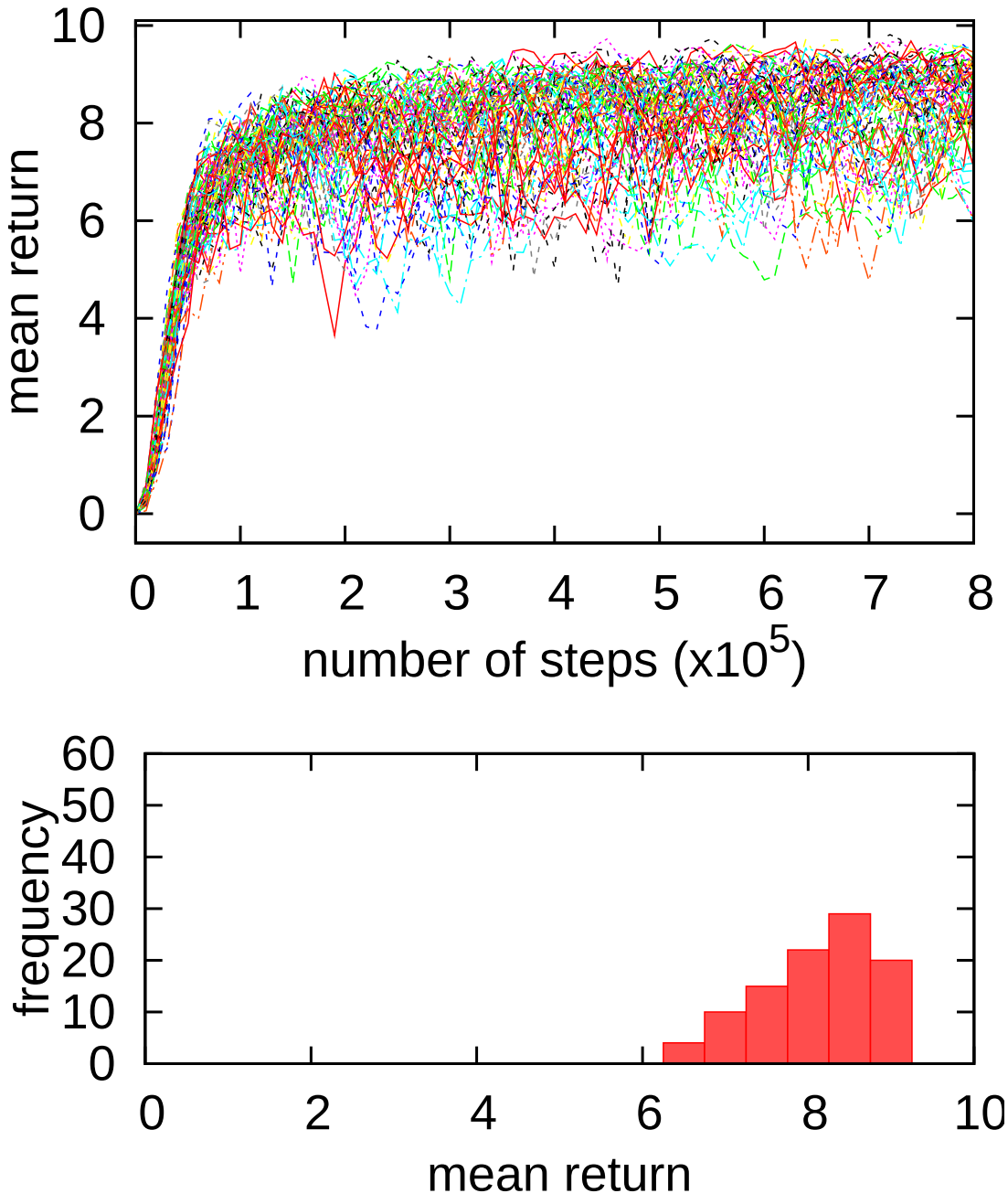

図 4.13: 集団サイズ $n = 25$ における匍匐ロボットでの学習曲線と学習結果のヒストグラム (全 100 試行)．上段：各試行での学習曲線．下段：各試行の最終性能のヒストグラム．

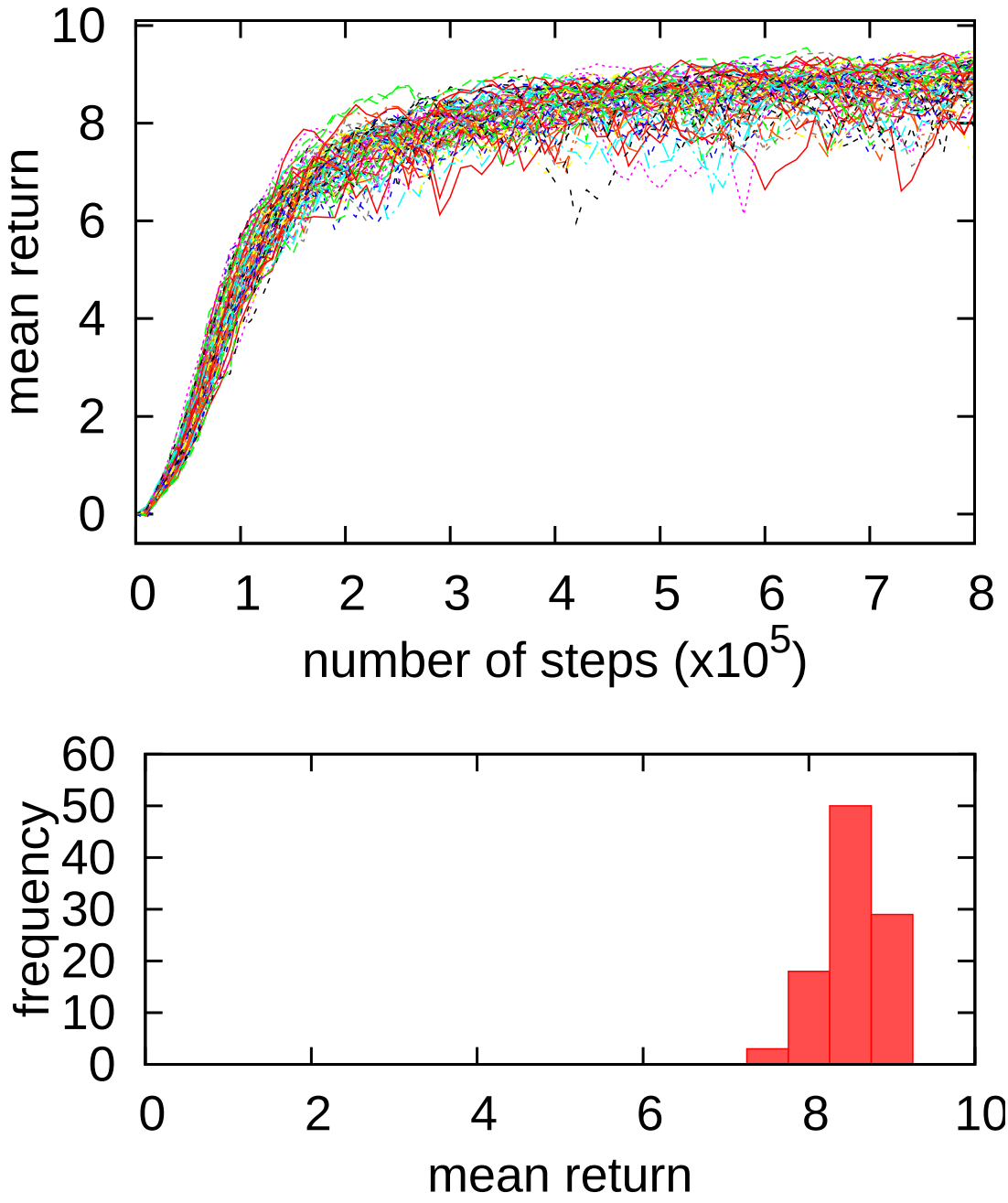

図 4.14: 集団サイズ $n = 100$ における匍匐ロボットでの学習曲線と学習結果のヒストグラム (全 100 試行)．上段：各試行での学習曲線．下段：各試行の最終性能のヒストグラム．

考察

集団サイズと学習曲線のばらつきの関係　図 4.13，図 4.14 において，$n = 25$ と $n = 100$ では試行ごとの性能のばらつきが異なることについて考察を行う．まず，原因を考えるために $n = 1$ での各試行の学習曲線を観察したところ，アームを 2 本使った動作を獲得した試行は，図 4.15 に示す 3 つの傾向に分かれた．well は上手く学習できている試行である．slow は学習が遅い試行であり，初期値が局所最適政策から遠くに生成されたと考えられる．wavering は性能が安定せずに振動している．

学習曲線が wavering となる原因としては，アームを 2 本使う政策にはさらに 2 つ以上の局所最適政策が存在することが考えられる．すなわち， well となる試行では政策パラメータの変化が性能に鈍感な政策を獲得しており， wavering となる試行では敏感な政策を獲得していると予想される．

政策パラメータの変化が大きすぎて性能が振動するならば，学習率を小さくすることで振動を抑えることができるはずである．そこで，$n = 1$ の場合について，学習率を下げて実験を行った．しかし，初めから小さな学習率を使うと時間がかかるため，初めの 10^5step は通常の $\alpha = 0.01$ で学習を行い，それ以降は 1/10 の $\alpha = 0.001$ とした．結果を図 4.16 に示している．図 4.16 の学習曲線の横軸は図 4.12 とスケールが異なることに注意されたい．図 4.12 の $n = 1$ と比べると，性能の振動が抑えられ， well と slow のみになった．このことから，アームを 2 本使う政策には，政策パラメータの変化が性能に与える影響が異なる複数の局所最適政策が存在すると考えられる．

$n = 100$ では，$n = 25$ と比べて個体数が多いため，学習曲線が well となる個体を含みやすい．そのため，図 4.14 において，$n = 100$ の性能が良いのは，上位層の学習により well の個体を選択することで slow や wavering による影響を抑えることができたためと考えられる．

ヒューリスティクスの妥当性　PPG においてリデュースを行わなかった場合について，$n = 25$ での結果を図 4.17 に示している．多くの試行で学習が正常に行われず，性能の高

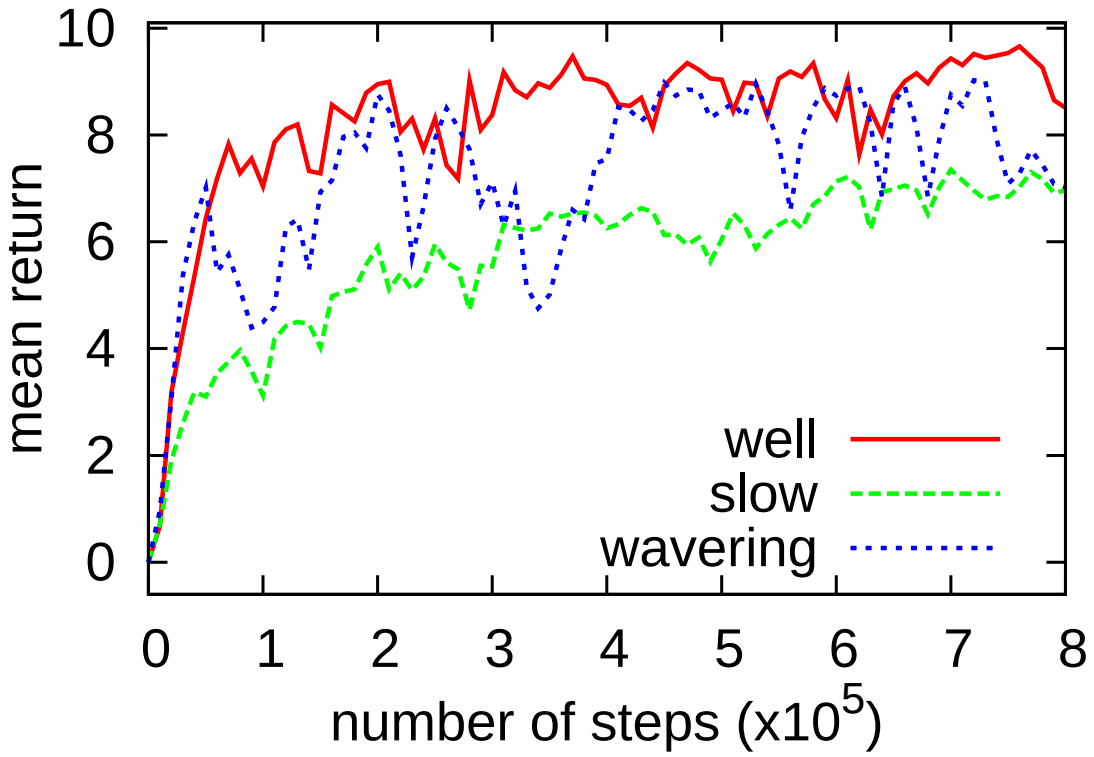

図 4.15: 匍匐ロボットでの集団サイズ $n = 1$ の典型的な試行の学習曲線.

い政策を獲得している試行も，学習が遅いことが確認できる．このタスクは，2 分木タスクよりも政策パラメータの探索に精度が要求される．そのため，$W_t > 1$ では式 (4.11) が増大し学習率が大きくなるため，政策改善に失敗するものと考えられる．

このことから，政策パラメータの精度が性能に与える影響が大きいタスクでは，リデュースを使用することが妥当であると考えられる．

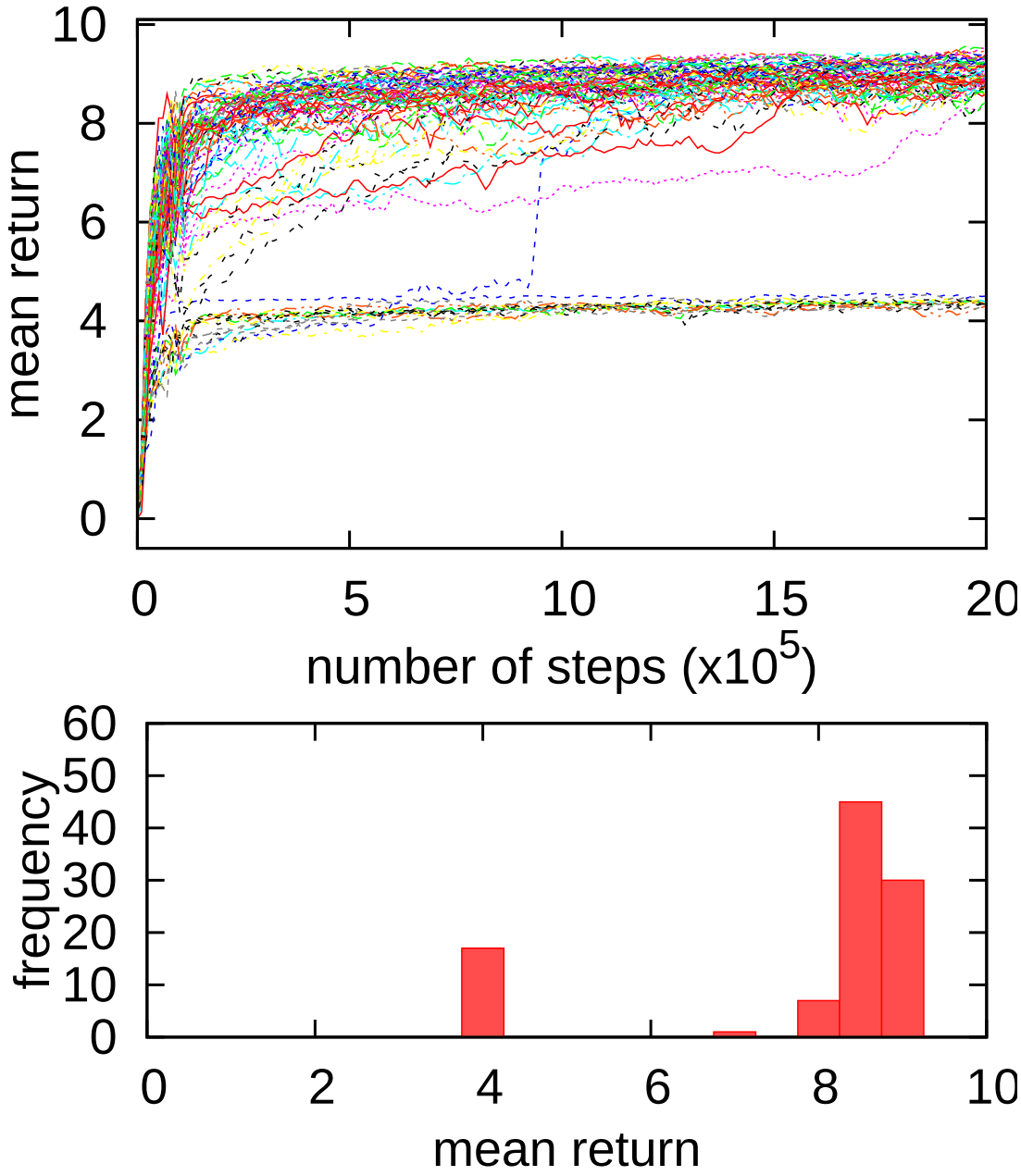

図 4.16: 匍匐ロボットでの学習曲線と学習結果のヒストグラム (全 100 試行)．集団サイズ $n = 1$，10^5step まで学習率 $\alpha = 0.01$，それ以降 $\alpha = 0.001$．上段：各試行での学習曲線．下段：各試行の最終性能のヒストグラム．

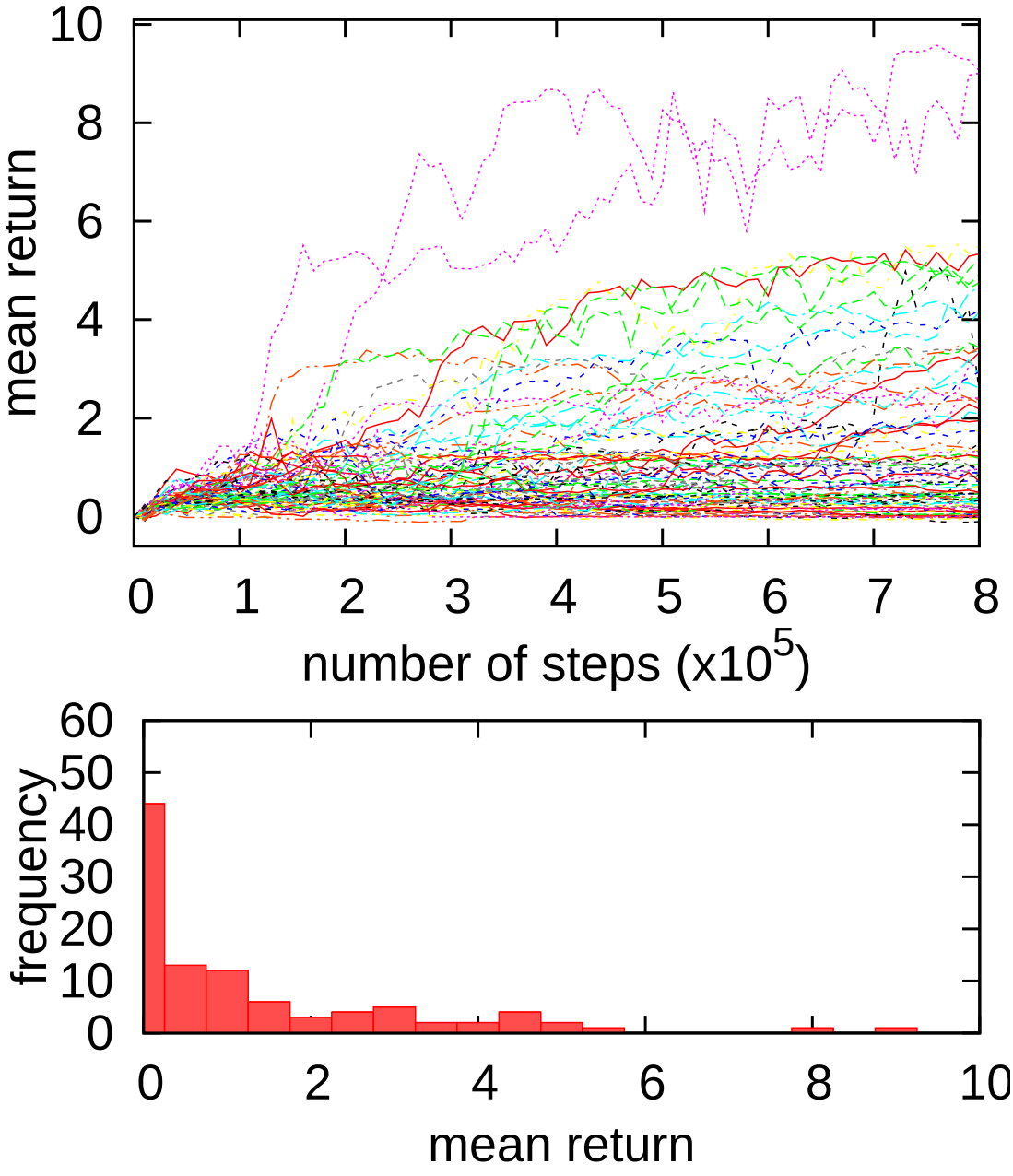

図 4.17: 匍匐ロボットでの学習曲線と学習結果のヒストグラム (全 100 試行)．集団サイズ $n = 25$ でステップサイズの調節を行うヒューリスティクスを使わない場合．上段：各試行での学習曲線．下段：各試行の最終性能のヒストグラム．

4.5 おわりに

本章では集団の概念を導入した新たな政策勾配法を提案した．これは多峰性の景観を持つ政策への対処を目的とした手法であり，単なる局所探索手法であった従来手法と比較すると，複数の局所最適政策が存在するタスクで最適政策を獲得する頻度の向上が認められた．さらに，重点サンプリングを導入し，集団サイズを大きく取ることに伴う環境とのインタラクションの増加へ対処した．

この新しい政策勾配法は，従来手法と同様に Actor-Critic 法の枠組み [49] や自然政策勾配法 [14, 25] を容易に適用することができると考えられる．さらに，上位層での学習は，個体の重みを表現する関数に確率的 2 分木 [50] などを用いれば，効率の良い学習が期待できる．提案手法にこれらの手法を適用することは今後の課題である．

また，上述の政策勾配法に関する既存技術の導入以外に，PPG を進化計算に基づく強化学習手法としてさらに発展させることが考えられる．現在の PPG は，初期集団が最適政策に対応した個体を含んでいない場合に，最適政策の獲得に失敗してしまう．対象とする政策がより高次元になると，初期集団が政策パラメータ空間においてスパースに配置される．そのため，多峰性の景観の下で最適政策を獲得するには非常に大きな集団サイズが必要となる．そのような場合に，政策更新の計算量がもはや無視できなくなると予想される．より少ない集団サイズで多峰性の景観へ対処する方法として，実数値 GA の交叉に相当する個体間の情報交換を行う操作を導入し，集団から推定される有望領域へ新たな個体を生成することが有効であると考えられるため，3 章で提案した FLIP との融合が求められる．

第5章　確率モデルで表した制御パラメータに対する自然政策勾配法

5.1　はじめに

4 章では多峰性の景観へ対処できる政策勾配法として PPG を設計した．PPG の性能は使用する政策勾配法の性能に依存するため，本章では政策勾配法の性能向上を目的とする．従来の政策勾配法は制御器の微分情報が利用できることが想定されているため，3 章で用いた IBP などの不連続な制御器を扱うことはできない．また，制御器の設計により，例えば，3.3.2 で述べた悪スケール性が発生することで性能に鈍感な政策パラメータの勾配が小さくなり，局所最適政策への収束が極端に遅くなる場合がある．そのため，政策表現モデルの設計は適切に行われている必要があり，これは設計者にとって負担となる．

これらの問題へ対処する方法として下記の 2 つの方法がある．まず，制御器の微分情報を用いない政策勾配法の枠組みとして *policy gradients with parameter-based exploration* (PGPE) [31] がある．PGPE は制御問題において効率的に政策改善を行う政策勾配法の枠組みとして提案されているが，結果的に制御器の微分情報を必要としない枠組みになっている．次に，制御器の設計におけるモデルのパラメータの取り方が政策更新に影響を与える問題に対して，自然勾配 [2] を用いた自然政策勾配法 [14, 4, 25] が提案されている．自然勾配を用いることで悪スケール性が存在するモデルなどにおいて通常の政策勾配法よりも高い収束率を実現している．そこで，PGPE の枠組みで自然政策勾配法を使用することにより上述の問題点が解決されると考えられる．

しかし，自然政策勾配法を PGPE の枠組みで使用することには計算量的な困難を伴う．自然政策勾配法は政策に対するフィッシャー情報行列 (Fisher information matrix; FIM)

の逆行列を求める必要があるが，政策パラメータが高次元になる場合にその計算コストは無視できなくなる．PGPE での政策表現モデルは通常の枠組みでのモデルに対してパラメータが高次元になる傾向がある．そのため，従来研究では自然政策勾配法と PGPE の組合せは 1 次元のパラメータを持つ制御器へのみ適用されるに留まり [41]，実用的な制御問題への適用はなされていない．

本章では，自然政策勾配法と PGPE を効果的に組合せた政策勾配法を提案する．従来手法ではサンプルから推定した FIM を使用するのに対して提案手法は真の FIM を使用し，自然政策勾配を推定するための計算量は通常の政策勾配におけるそれと等価である．また，提案手法は従来の政策勾配法と比べて学習性能が向上することを実験により示す．

5.2 準備

ここでは，4.2.1 の政策勾配法をベースラインアルゴリズムとして自然政策勾配法および PGPE の説明を行う．本研究では，説明の簡単のため不完全知覚の存在しない MDP 環境を想定するが，POMDP 環境への拡張は容易に行える．

5.2.1 自然政策勾配法

自然政策勾配法 [14] は政策更新が政策パラメータの取り方に依存しないことを目的とした政策勾配法である．自然勾配 [2] はパラメータ空間がリーマン構造であっても最急勾配方向を表す．確率分布に対する自然な計量は FIM であり [3]，リーマン多様体上での最急勾配方向は通常の勾配の左から計量行列の逆行列を掛けることで与えられる [2]．自然勾配は通常の勾配と FIM を計算することで求まり，これを用いた勾配法は通常の勾配を用いた場合よりも収束が速くなる傾向がある．

Kakade[14] は自然勾配を政策探索に導入した自然政策勾配法を提案している．もし FIM の逆行列が存在するのであれば，通常の勾配の左から FIM の逆行列を掛けることで自然政策勾配 $\tilde{\nabla}_\theta \eta(\theta) \equiv \mathbf{F}_\theta^{-1} \nabla_\theta \eta(\theta)$ が得られる．本研究では政策に対する FIM として [14] と

同様に以下で定義される FIM を用いる．

$$\mathbf{F}_\theta = \int_{\mathcal{S}} \rho^\pi(\mathbf{s}|\theta) \int_{\mathcal{A}} \pi(\mathbf{a}|\mathbf{s},\theta) \nabla_\theta \log \pi(\mathbf{a}|\mathbf{s},\theta) \nabla_\theta \log \pi(\mathbf{a}|\mathbf{s},\theta)^{\mathrm{T}} d\mathbf{a} d\mathbf{s}. \tag{5.1}$$

通常の政策勾配と同様に，自然政策勾配はサンプルから推定することができる．自然政策勾配に対する *characteristic eligibility* を $\tilde{\nabla}_\theta \log \pi(\mathbf{a}_t|\mathbf{s}_t,\theta) \equiv \mathbf{F}_\theta^{-1} \nabla_\theta \log \pi(\mathbf{a}_t|\mathbf{s}_t,\theta)$ と表記する．これを用いて自然政策勾配の近似は以下のように計算される．

$$\tilde{\nabla}_\theta \eta(\theta) \approx \frac{1}{T} \sum_{t=1}^{T} \mathbf{F}_\theta^{-1} r_t \mathbf{z}_t = \frac{1}{T} \sum_{t=1}^{T} r_t \tilde{\mathbf{z}}_t, \tag{5.2}$$

ここで，$\tilde{\mathbf{z}}_t = \beta \tilde{\mathbf{z}}_{t-1} + \tilde{\nabla}_\theta \log \pi(\mathbf{a}_t|\mathbf{s}_t,\theta)$ である．FIM をオンラインで推定するためのヒューリスティクスとして Kakade [14] は以下の方法を用いている．

$$\mathbf{F}_{\theta,t} = (1 - \frac{1}{t}) \mathbf{F}_{\theta,t-1} + \frac{1}{t} (\nabla_\theta \log \pi(\mathbf{a}_t|\mathbf{s}_t,\theta) \nabla_\theta \log \pi(\mathbf{a}_t|\mathbf{s}_t,\theta)^{\mathrm{T}} + \lambda \mathbf{I}), \tag{5.3}$$

ここで λ は小さな正の定数である．

5.2.2 確率モデルで表した制御パラメータに対する政策勾配法

政策勾配法を制御問題へ適用する場合には，政策を制御器 $\psi(\mathbf{u}|\mathbf{s},\mathbf{w})$ と探査戦略で構成する．ここで，$\mathbf{u} \in \mathcal{U} \subseteq \Re^m$ は制御信号，$\mathbf{w} \in \mathcal{W} \subseteq \Re^n$ は制御器のパラメータ（以下，制御パラメータ）である．学習の目的は制御パラメータ $\mathbf{w}$ の最適化であり，探査戦略には現在のパラメータの近傍を確率的にサンプリングすることが求められる．一般的に用いられる探査戦略のモデルは図 5.1 (上段) に示すように制御信号に対して正規分布に従う摂動を加えるものである．これを本論文では *control-based exploration*(CE) と呼ぶことにする．このモデルでは，制御信号をエージェントの行動として扱い，政策は以下のようである．

$$\pi_U(\mathbf{u}|\mathbf{s},\theta) = \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{u} - \psi(\mathbf{s},\mathbf{w}))^{\mathrm{T}} \Sigma^{-1} (\mathbf{u} - \psi(\mathbf{s},\mathbf{w}))\right) : \mathcal{S} \to \mathcal{U},$$

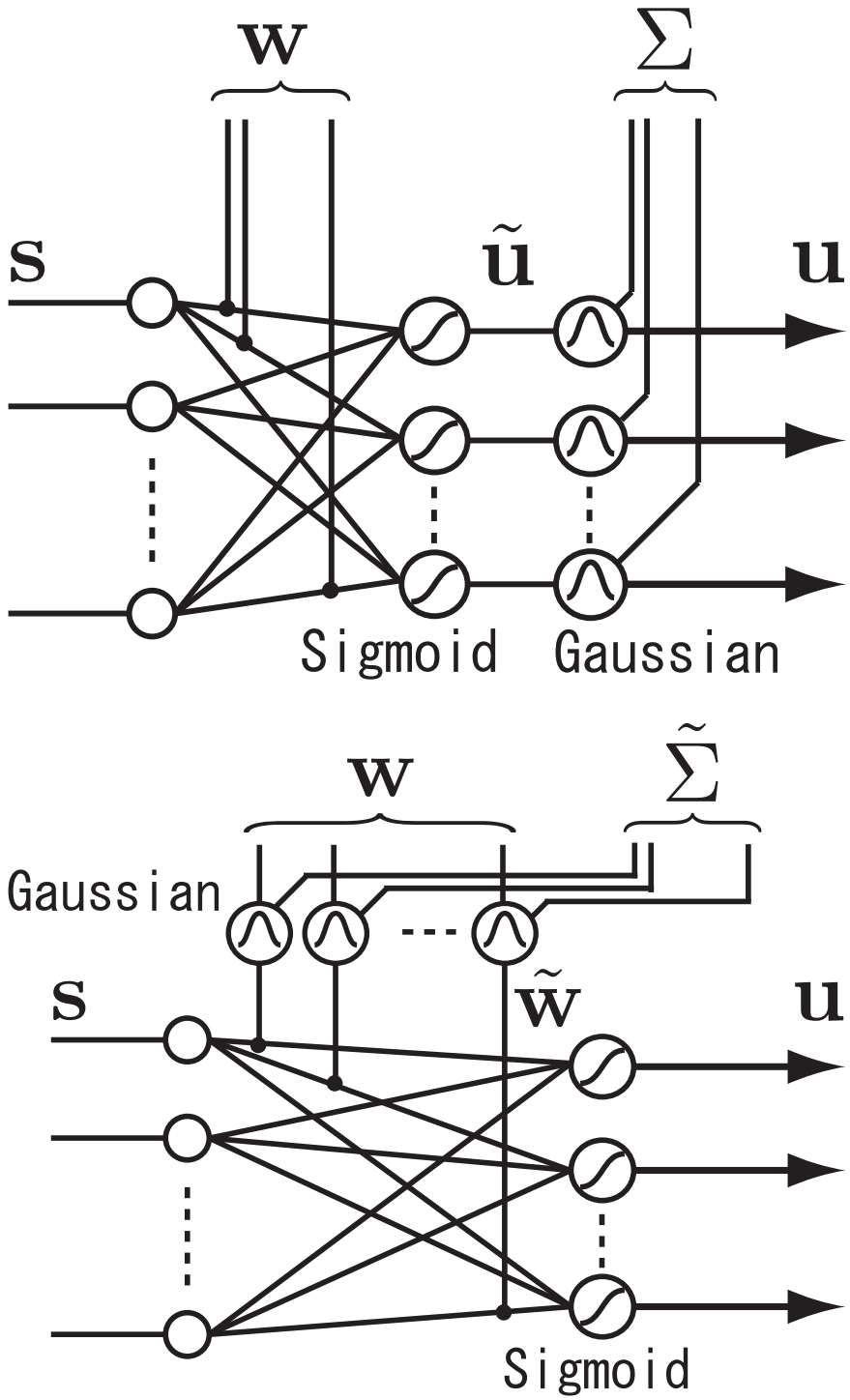

図 5.1: 制御器 $\psi(\mathbf{u}|\mathbf{s},\mathbf{w})$ が単層パーセプトロンである場合の Control-based Exploration (CE) と Parameter-based Exploration (PE) の模式図. CE (上段) が制御信号に対して摂動を加えるのに対して，PE (下段) は制御パラメータに対して摂動を加える.

ここで，Σ は $m \times m$ 次元の分散共分散行列（以下，共分散）であり，政策パラメータは $\theta = \langle \mathbf{w}, \Sigma \rangle$ となる．このように政策パラメータに探査戦略である正規分布の共分散のパラメータまで含める *Gaussian unit* [40] の利点は，探査の割合を自己調節することにある．この政策のもとで時刻 t での制御信号は以下で生成される．

$$\tilde{\mathbf{u}}_t = \psi(\mathbf{s}_t, \mathbf{w}),$$
$$\mathbf{u}_t \sim \mathcal{N}(\tilde{\mathbf{u}}_t, \Sigma).$$

CE では制御信号の近傍からサンプリングを行っている．しかし，制御パラメータでの近傍と制御信号の近傍は多くの場合一致しない．そのため，CE は分散を小さくしたとしても現在の制御パラメータの近傍を探査するとは限らない．この性質が政策改善を停滞させる原因のひとつになっていると考えられる．

この問題に対処するために，Sehnke ら [31] は CE とは異なる探査戦略を導入した新たな政策勾配法の枠組み PGPE を提案している．この枠組みでは，制御パラメータをエージェントの行動として扱い，政策は以下のようである．

$$\pi_W(\tilde{\mathbf{w}}|\mathbf{s}, \theta) = \frac{1}{(2\pi)^{n/2}|\tilde{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\tilde{\mathbf{w}} - \mathbf{w})^{\mathrm{T}}\tilde{\Sigma}^{-1}(\tilde{\mathbf{w}} - \mathbf{w})\right) : \mathcal{S} \to \mathcal{W},$$

ここで，$\tilde{\Sigma}$ は $n \times n$ 次元の共分散であり，政策パラメータは $\theta = \langle \mathbf{w}, \tilde{\Sigma} \rangle$ となる．また，制御器は環境の一部として扱われ，時刻 t での制御信号は以下で生成される．

$$\tilde{\mathbf{w}}_t \sim \mathcal{N}(\mathbf{w}, \tilde{\Sigma}),$$
$$\mathbf{u}_t = \psi(\mathbf{s}_t, \tilde{\mathbf{w}}_t).$$

PGPE の探査戦略 parameter-based exploration（PE）[31] は図 5.1 (下段) に示すように制御パラメータに対して摂動を加える．そのため，PE では分散は現在の制御パラメータのどれほど近傍を探査するかを定めることになる．

Sehnke ら [31] は PGPE が CE を用いた政策勾配法に対して収束速度が速くなるという実験結果を示している．さらに，PGPE の枠組みでは制御パラメータを表す確率モデルの微分情報は用いるが，制御器は環境の一部として扱いるため制御器の微分情報は必要ない．そのため，3 章で用いている IBP など不連続な制御器に対して適用可能である．

5.3 確率モデルで表した制御パラメータに対する自然政策勾配法

本節では，自然政策勾配を用いる PGPE として *natural policy gradients with parameter-based exploration* (NPGPE) を提案する．NPGPE は NPG と PGPE の利点に加え，i) 従来の自然政策勾配法とは異なり，真の FIM を使用する．ii) 自然政策勾配を推定するための計算量が通常の政策勾配のそれと同等であるという特徴を持つ．

5.3.1 探査戦略の設計

制御パラメータを多変量正規分布で表す探査戦略を採用する．政策は制御器のモデルを含まず探査戦略のみで構成され，政策パラメータ $\theta = \langle \mathbf{w}, \mathbf{C} \rangle$ で表される多変量正規分布 $\mu(\tilde{\mathbf{w}}|\theta)$ である．ここで，$\mathbf{w} \in \Re^n$ は中心ベクトル，$\mathbf{C} \in \Re^{n \times n}$ は共分散 $\tilde{\Sigma}$ の上三角行列となるコレスキー分解，つまり，$\tilde{\Sigma} = \mathbf{C}^{\mathrm{T}}\mathbf{C}$ である．Sun ら [34] はこの多変量正規分布のパラメータの取り方について二つの利点を挙げている，i) $\mathbf{C}$ は n^2 の要素を持つ共分散 $\tilde{\Sigma}$ を表すために $n(n+1)/2$ 個のパラメータしか使わない．ii) $\mathbf{C}$ の対角要素の絶対値は $\tilde{\Sigma}$ の固有値であり，$\mathbf{C}^{\mathrm{T}}\mathbf{C}$ は常に半正定値対称行列になる．政策パラメータ θ は $\mathbf{w}$ と $\mathbf{C}$ の上三角要素からなる以下の $[n(n+3)/2]$-次元の列ベクトルで表される．

$$\theta = [\mathbf{w}^{\mathrm{T}}, (\mathbf{C}_{1,1:n}), (\mathbf{C}_{2,2:n}), ..., (\mathbf{C}_{n,n:n})]^{\mathrm{T}}. \tag{5.4}$$

ここで，$\mathbf{C}_{k,k:n}$ は $\mathbf{C}$ の k 行の k から n 列要素からなる部分行列である．

5.3.2 フィッシャー情報行列とその逆行列

従来の自然政策勾配法 [14] では，サンプルの時系列から推定した FIM を使用している．ここでは，PGPE の枠組みにおいて真の FIM が利用できることを述べ，5.3.1 の政策に対する FIM およびその逆行列を示す．

政策 π を $\mu(\tilde{\mathbf{w}}|\theta)$ として (2.3) 式に代入することで，政策勾配は以下のようになる．

$$\nabla_\theta \eta(\theta) = \int_{\mathcal{S}} \rho^\pi(\mathbf{s}|\theta) \int_{\mathcal{W}} \mu(\tilde{\mathbf{w}}|\theta) \nabla_\theta \log \mu(\tilde{\mathbf{w}}|\theta) Q_\theta(\mathbf{s}, \tilde{\mathbf{w}}) d\tilde{\mathbf{w}} d\mathbf{s}. \tag{5.5}$$

また，(5.3) 式に代入することで，FIM は以下のようになる．

$$\mathbf{F}_\theta = \int_{\mathcal{S}} \rho^\pi(\mathbf{s}|\theta) \int_{\mathcal{W}} \mu(\tilde{\mathbf{w}}|\theta) \nabla_\theta \log \mu(\tilde{\mathbf{w}}|\theta) \nabla_\theta \log \mu(\tilde{\mathbf{w}}|\theta)^{\mathrm{T}} d\tilde{\mathbf{w}} d\mathbf{s} \tag{5.6}$$

$$= \int_{\mathcal{W}} \mu(\tilde{\mathbf{w}}|\theta) \nabla_\theta \log \mu(\tilde{\mathbf{w}}|\theta) \nabla_\theta \log \mu(\tilde{\mathbf{w}}|\theta)^{\mathrm{T}} d\tilde{\mathbf{w}}. \tag{5.7}$$

ここで，$\int_{\mathcal{W}} \mu(\tilde{\mathbf{w}}|\theta) \nabla_\theta \log \mu(\tilde{\mathbf{w}}|\theta) \nabla_\theta \log \mu(\tilde{\mathbf{w}}|\theta)^{\mathrm{T}} d\tilde{\mathbf{w}}$ が状態観測 $\mathbf{s}$ に対して独立であることと，$\int_{\mathcal{S}} \rho^\pi(\mathbf{s}|\theta) d\mathbf{s} = 1$ であることを用いた．以上より，政策勾配は依然として状態の定常分布 $\rho^\pi(\mathbf{s}|\theta)$ の項を含んでいるが，FIM からはその項が消えるため解析的に FIM を求めることができることがわかる．

Sun ら [34] は多変量正規分布 $\mathcal{N}(\mathbf{w}, \mathbf{C}^{\mathrm{T}}\mathbf{C})$ の FIM がブロック対角行列 $\mathrm{diag}(\mathbf{F}_0, ..., \mathbf{F}_n)$ になることを示している．それによると，最初の部分行列 $\mathbf{F}_0$ は，

$$\mathbf{F}_0 = \tilde{\Sigma}^{-1} \tag{5.8}$$

であり，k 番目 $(1 \leq k \leq n)$ の部分行列 $\mathbf{F}_k$ は，

$$\mathbf{F}_k = \begin{bmatrix} c_{k,k}^{-2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \tilde{\Sigma}_{k:n,k:n}^{-1} \tag{5.9}$$

$$= \begin{bmatrix} \mathbf{0} & \mathbf{I}_{\bar{k}} \end{bmatrix} \mathbf{C}^{-1} \left(\mathbf{v}_k \mathbf{v}_k^{\mathrm{T}} + \mathbf{I} \right) \mathbf{C}^{-\mathrm{T}} \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{\bar{k}} \end{bmatrix}, \tag{5.10}$$

で与えられる．ここで，$\mathbf{v}_k$ は第 k 要素のみが 1 でありその他の要素が全て 0 である n 次元列ベクトル，$\mathbf{I}_{\bar{k}}$ は $[n-k+1]$-次元の単位行列である．

Akimoto ら [1] はこの FIM の逆行列を導出しており，k 番目 $(1 \leq k \leq n)$ の部分行列 $\mathbf{F}_k$ の逆行列は以下のようである．

$$\mathbf{F}_k^{-1} = \begin{bmatrix} \mathbf{0} & \mathbf{I}_{\bar{k}} \end{bmatrix} \mathbf{C}^{\mathrm{T}} \left(-\frac{1}{2} \mathbf{v}_k \mathbf{v}_k^{\mathrm{T}} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{\bar{k}} \end{bmatrix} \right) \mathbf{C} \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{\bar{k}} \end{bmatrix}. \tag{5.11}$$

これは，$\mathbf{F}_\theta$ がブロック対角行列であることと，$\mathbf{C}$ が上三角行列であること，および以下の性質を用いることで容易に確かめられる．

$$\mathbf{v}_k^{\mathrm{T}} \mathbf{C} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{\bar{k}} \end{bmatrix} \mathbf{C}^{-1} = \mathbf{v}_k^{\mathrm{T}} \quad \text{and} \quad \begin{bmatrix} \mathbf{0} & \mathbf{I}_{\bar{k}} \end{bmatrix} \mathbf{C} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{\bar{k}} \end{bmatrix} \mathbf{C}^{-1} = \begin{bmatrix} \mathbf{0} & \mathbf{I}_{\bar{k}} \end{bmatrix}. \tag{5.12}$$

5.3.3 自然政策勾配

従来の自然政策勾配法 [14] では，通常の政策勾配に対して自然政策勾配の推定には計算量の増大が伴う．ここでは，PGPE の枠組みにおいて自然政策勾配が通常の政策勾配の推定と同等の計算量で行えることを述べる．

Akimoto ら [1] と同様に $\tilde{\nabla}_\theta \log \mu(\tilde{\mathbf{w}}_t|\theta) = \mathbf{F}_\theta^{-1} \nabla_\theta \log \mu(\tilde{\mathbf{w}}_t|\theta)$ を導出する．$\nabla_{\mathbf{w}} \log \mu(\tilde{\mathbf{w}}_t|\theta)$

は以下で与えられる．

$$\nabla_{\mathbf{w}} \log \mu(\tilde{\mathbf{w}}_t|\theta) = \tilde{\Sigma}^{-1}(\tilde{\mathbf{w}}_t - \mathbf{w}). \tag{5.13}$$

$\mathbf{F}_0^{-1}$ は明らかに $\tilde{\Sigma}$ であるため，$\tilde{\nabla}_{\mathbf{w}} \log \mu(\tilde{\mathbf{w}}_t|\theta)$ は以下のようになる．

$$\tilde{\nabla}_{\mathbf{w}} \log \mu(\tilde{\mathbf{w}}_t|\theta) = \mathbf{F}_0^{-1} \nabla_{\mathbf{w}} \log \mu(\tilde{\mathbf{w}}_t|\theta) = \tilde{\mathbf{w}}_t - \mathbf{w}. \tag{5.14}$$

$\frac{\partial}{\partial c_{i,j}} \log \mu(\tilde{\mathbf{w}}_t|\theta)$ は以下で与えられる．

$$\frac{\partial}{\partial c_{i,j}} \log \mu(\tilde{\mathbf{w}}_t|\theta) = \mathbf{v}_i^{\mathrm{T}} \left(\mathrm{triu}(\mathbf{Y}_t \mathbf{C}^{-\mathrm{T}}) - \mathrm{diag}(\mathbf{C}^{-1}) \right) \mathbf{v}_j, \tag{5.15}$$

ここで，$\mathrm{triu}(\mathbf{M})$ は上三角行列を表し，$i \leq j$ に対して (i,j) 成分が行列 $\mathbf{M}$ の (i,j) 成分と一致し，それ以外の要素は 0 である．また，$\mathbf{Y}_t = \mathbf{C}^{-\mathrm{T}}(\tilde{\mathbf{w}}_t - \mathbf{w})(\tilde{\mathbf{w}}_t - \mathbf{w})^{\mathrm{T}} \mathbf{C}^{-1}$ は対称行列である．

行列 $\mathbf{C}$ の k 行目の上三角要素からなる部分行列を $\mathbf{c}_k = (c_{k,k}, ..., c_{k,n})^{\mathrm{T}} \in \Re^{n+1-k}$ とする．このとき，$\frac{\partial}{\partial \mathbf{c}_k} \log \mu(\tilde{\mathbf{w}}_t|\theta)$ は以下で与えられる．

$$\nabla_{\mathbf{c}_k} \log \mu(\tilde{\mathbf{w}}_t|\theta) = \begin{bmatrix} \mathbf{0} & \mathbf{I}_{\bar{k}} \end{bmatrix} \left(\mathbf{C}^{-1} \mathbf{Y}_t - \mathrm{diag}(\mathbf{C}^{-1}) \right) \mathbf{v}_k. \tag{5.16}$$

行列の性質として，(5.12) 式および，

$$\mathrm{diag}(\mathbf{C}^{-1}) \mathbf{v}_k = c_{k,k}^{-1} \mathbf{v}_k \tag{5.17}$$

$$\mathbf{v}_k^{\mathrm{T}} \mathbf{C} \mathbf{v}_k = c_{k,k} \tag{5.18}$$

$$\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{\bar{k}} \end{bmatrix} \mathbf{C}\mathbf{v}_k = c_{k,k}\mathbf{v}_k, \tag{5.19}$$

を用いると，k 番目の部分行列 $\mathbf{F}_\theta^{-1}$ に対応する $\mathbf{F}_\theta^{-1}\nabla_\theta \log \mu(\tilde{\mathbf{w}}_t|\theta)$ は以下のようになる．

$$\begin{aligned}
\tilde{\nabla}_{\mathbf{c}_k} \log \mu(\tilde{\mathbf{w}}_t|\theta) &= \mathbf{F}_k^{-1}\nabla_{\mathbf{c}_k} \log \mu(\tilde{\mathbf{w}}_t|\theta) \\
&= \begin{bmatrix} \mathbf{0} & \mathbf{I}_{\bar{k}} \end{bmatrix} \mathbf{C}^{\mathrm{T}} \left(-\frac{1}{2}\mathbf{v}_k\mathbf{v}_k^{\mathrm{T}} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{\bar{k}} \end{bmatrix} \right) \mathbf{C} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{\bar{k}} \end{bmatrix} \left(\mathbf{C}^{-1}\mathbf{Y}_t - \mathrm{diag}(\mathbf{C}^{-1}) \right) \mathbf{v}_k \\
&= \begin{bmatrix} \mathbf{0} & \mathbf{I}_{\bar{k}} \end{bmatrix} \mathbf{C}^{\mathrm{T}} \left(-\frac{1}{2}\mathbf{v}_k\mathbf{v}_k^{\mathrm{T}} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{\bar{k}} \end{bmatrix} \right) (\mathbf{Y}_t - \mathbf{I})\mathbf{v}_k. \qquad (5.20)
\end{aligned}$$

そして，$\tilde{\nabla}_{\mathbf{c}_k} \log \mu(\tilde{\mathbf{w}}_t|\theta)^{\mathrm{T}} = \left(\tilde{\nabla}_{\mathbf{C}} \log \mu(\tilde{\mathbf{w}}_t|\theta) \right)_{k,k:n}$ であることから以下を得る．

$$\tilde{\nabla}_{\mathbf{C}} \log \mu(\tilde{\mathbf{w}}_t|\theta) = \left(\mathrm{triu}(\mathbf{Y}_t) - \frac{1}{2}\mathrm{diag}(\mathbf{Y}_t) - \frac{1}{2}\mathbf{I} \right) \mathbf{C}. \tag{5.21}$$

自然政策勾配を推定するための各ステップでの

$$\tilde{\nabla}_\theta \log \mu(\tilde{\mathbf{w}}_t|\theta) = [\tilde{\nabla}_{\mathbf{w}} \log \mu(\tilde{\mathbf{w}}_t|\theta)^{\mathrm{T}}, \tilde{\nabla}_{\mathbf{c}_1} \log \mu(\tilde{\mathbf{w}}_t|\theta)^{\mathrm{T}}, ..., \tilde{\nabla}_{\mathbf{c}_n} \log \mu(\tilde{\mathbf{w}}_t|\theta)^{\mathrm{T}}]^{\mathrm{T}}$$

の計算量は $O(n^3)$ となり，$\nabla_\theta \log \mu(\tilde{\mathbf{w}}_t|\theta)$ の計算量と等しくなる．PGPE の枠組みで Kakade[14] の方法を用いた自然勾配の推定では数値的に $\mathbf{F}_\theta^{-1}$ と $\nabla_\theta \log \mu(\tilde{\mathbf{w}}_t|\theta)$ を求めてから掛け合わせるため $O(n^6)$ であることと比べれば大幅に計算量が削減される．また，行列 $\mathbf{C}$ を対角行列に制限すれば計算量は $O(n)$ になる．そのようなモデルは制御パラメータ間の相関を考慮したサンプリングを行うことはできないが，$O(n^3)$ の計算が困難となる高次元の制御パラメータをもつ制御器に対して有効である．

Algorithm 5 Natural Policy Gradient Method with Parameter-based Exploration

Require: $\theta = \langle \mathbf{w}, \mathbf{C} \rangle$: policy parameters, $\psi(\mathbf{u}|\mathbf{s}, \mathbf{w})$: controller, α: step size, β: discount rate, b: baseline.

1: Initialize $\tilde{\mathbf{z}}_0 = \mathbf{0}$, observe $\mathbf{s}_1$.
2: **for** $t = 1, ...$ **do**
3: Draw $\xi_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, compute action $\tilde{\mathbf{w}}_t = \mathbf{C}^{\mathrm{T}}\xi_t + \mathbf{w}$.
4: Execute $\mathbf{u}_t \sim \psi(\mathbf{u}_t|\mathbf{s}_t, \tilde{\mathbf{w}}_t)$, obtain observation $\mathbf{s}_{t+1}$ and reward r_t.
5: $\tilde{\nabla}_{\mathbf{w}} \log \mu(\tilde{\mathbf{w}}_t|\theta) = \tilde{\mathbf{w}}_t - \mathbf{w}$, $\tilde{\nabla}_{\mathbf{C}} \log \mu(\tilde{\mathbf{w}}_t|\theta) = \{\mathrm{triu}(\xi_t \xi_t^{\mathrm{T}}) - \frac{1}{2}\mathrm{diag}(\xi_t \xi_t^{\mathrm{T}}) - \frac{1}{2}\mathbf{I}\}\mathbf{C}$
6: $\tilde{\mathbf{z}}_t = \beta \tilde{\mathbf{z}}_{t-1} + \tilde{\nabla}_{\theta} \log \mu(\tilde{\mathbf{w}}_t|\theta)$
7: $\theta \leftarrow \theta + \alpha(r_t - b)\tilde{\mathbf{z}}_t$
8: **end for**

5.3.4 アルゴリズム

制御器 $\psi(\mathbf{u}|\mathbf{s}, \mathbf{w})$ の制御パラメータ $\mathbf{w}$ を探査戦略 $\mu(\tilde{\mathbf{w}}|\theta)$ を用いて学習する NPGPE アルゴリズムを Algorithm 5 に示す．各時刻 t における制御パラメータ $\tilde{\mathbf{w}}_t$ は標準正規分布から生成される $\xi_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ を用いて $\tilde{\mathbf{w}}_t = \mathbf{C}^{\mathrm{T}}\xi_t + \mathbf{w}$ で表されることから，$\mathbf{Y}_t = \mathbf{C}^{-\mathrm{T}}(\tilde{\mathbf{w}}_t - \mathbf{w})(\tilde{\mathbf{w}}_t - \mathbf{w})^{\mathrm{T}}\mathbf{C}^{-1} = \xi_t \xi_t^{\mathrm{T}}$ を用いている．また，勾配推定の分散を抑えるために [9] の最適ベースライン b を用いる．

5.4 実験

ここでは，提案した NPGPE の性能評価を行う．PGPE の有効性はエピソード的タスクで報告されているが [31]，本研究では連続タスクにおいてその有効性を検証する．まず，単純な線型 2 次形式制御タスクにおいて探査戦略として CE と PE，勾配法として通常の政策勾配 (”Vanilla” Policy Gradient; VPG) および自然政策勾配 (Natural Policy Gradient; NPG) をそれぞれ組合せた場合の性能比較を行う．次に，匍匐ロボットの前進動作獲得タスクにおいて NPGPE の有効性を検証する．

5.4.1 実験で用いる探査戦略

探査戦略のモデルとして以下の 2 つを用いる．PE では 5.3.1 で導入した $\mu(\tilde{\mathbf{w}}|\theta)$ を用いる．CE では制御信号を正規分布で表す $\epsilon(\mathbf{u}|\tilde{\mathbf{u}}, \mathbf{D})$ を用いる．ここで，$\tilde{\mathbf{u}}$ は制御器 ψ により生成された制御信号の中心ベクトル，$\mathbf{D}$ は共分散 Σ の上三角行列となるコレスキー分解，つまり，$\Sigma = \mathbf{D}^{\mathrm{T}}\mathbf{D}$ である．政策 $\pi_U(\mathbf{u}|\mathbf{s}, \theta)$ のパラメータは $\theta = \langle \mathbf{w}, \mathbf{D} \rangle$ であり，θ は $\mathbf{w}$ と $\mathbf{D}$ の上三角要素からなる $[n + m(m + 1)/2]$-次元ベクトルである．

5.4.2 線型 2 次形式制御タスク

線型 2 次形式制御タスクは報酬の遅れが存在するベンチマーク問題である [16]．環境は以下に示す単純なダイナミクスからなる．

$$\mathbf{s}_{t+1} = \mathbf{s}_t + \mathbf{u}_t + \delta,$$

ここで，$\mathbf{s} \in \Re^1$，$\mathbf{u} \in \Re^1$ そして $\delta \sim \mathcal{N}(0, 0.5^2)$ である．各時刻の即時報酬は $r_t = -\mathbf{s}_t^2 - \mathbf{u}_t^2$ で与えられる．本実験では，遷移可能な状態空間は $[-4, 4]$ に制限する．$\mathbf{s}_t$ が上記の範囲を超える場合は，制限範囲までしか移動できないものとする．制御信号についても同様に $[-4, 4]$ の範囲外の値が提示されても，制限の範囲でしか実行されないとした．制御器は $\psi(\mathbf{u}|\mathbf{s}, \mathbf{w}) = \mathbf{s} \cdot \mathbf{w}$ で与えられ，制御パラメータは $\mathbf{w} \in \Re^1$ である．最適な制御パラメータは解析的に求めることができ，Riccati 方程式より，最適パラメータは $\mathrm{w}^* = 2/(1 + 2\beta + \sqrt{4\beta^2 + 1}) - 1$ で与えられる．

可読性を高めるために，自然政策勾配を用いるものを NPG，通常の政策勾配を用いるものを VPG で表し，NPG(w) および VPG(w) を制御パラメータ空間で探査を行う PE を用いたもの，NPG(u) および VPG(u) を制御信号空間で探査を行う CE を用いたものを表す．提案手法 NPGPE は NPG(w) である．

図 5.2 に各手法の学習曲線を示している．同じ勾配法を用いる場合に PE は CE と比べて学習曲線の立ち上がりが早くなることが確認できる．同じ探査戦略を用いる場合には自

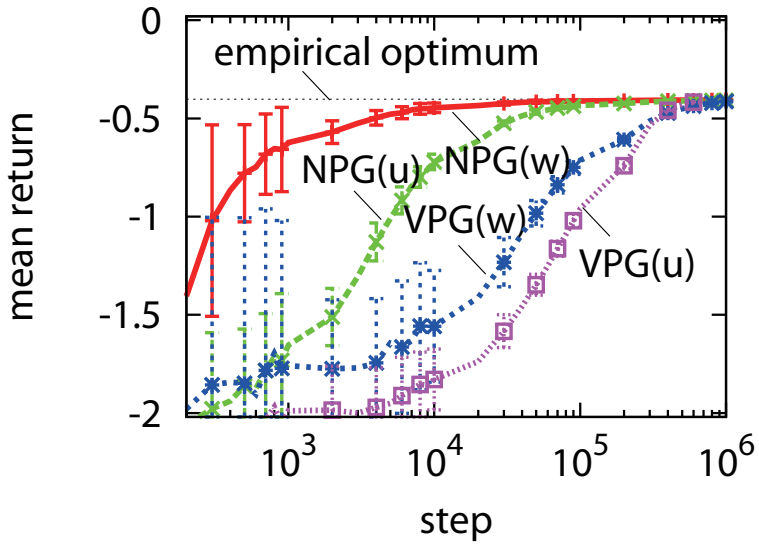

図 5.2: 線型 2 次形式制御タスクにおける NPG(w) と NPG(u)，VPG(w) および VPG(u) の学習曲線の比較 (100 試行平均)． empirical optimum は最適パラメータのもとでの性能を表している．

然政策勾配法は通常の政策勾配法と比べて収束が速いことが確認できる．そして，自然政策勾配法と PE の組合せの性能が最も良いことが分かる．図 5.3 は状態-制御信号空間および状態-制御パラメータ空間における PE と CE におけるサンプリング領域を模式的に表したものである．PE を用いる場合に学習効率が向上する理由は明確ではないが以下のように考えられる．探査戦略の選択によらず，学習目的は制御パラメータの調節である．そのため，サンプリング領域は制御パラメータ空間での距離が重要な意味を持つ．PE における共分散は制御パラメータ空間におけるサンプリング領域を表しており，大きく取ることで制御パラメータは大きく移動し小さく取ることで精度を高めることができる．それに対して，CE は制御信号空間において現在の制御器の出力の近傍からサンプリングするため，制御パラメータ空間では状態に依存したサンプリングになっており，特に $\mathbf{s} = 0$ 付近では共分散を小さくしても一様分布に近いサンプリングを行うことになる．そのため，CE では共分散が制御パラメータの調節へ適切に反映されず，学習効率が低下しているものと思われる．このことから，PE は CE と比べて有効なサンプリングを実現していると考えられる．

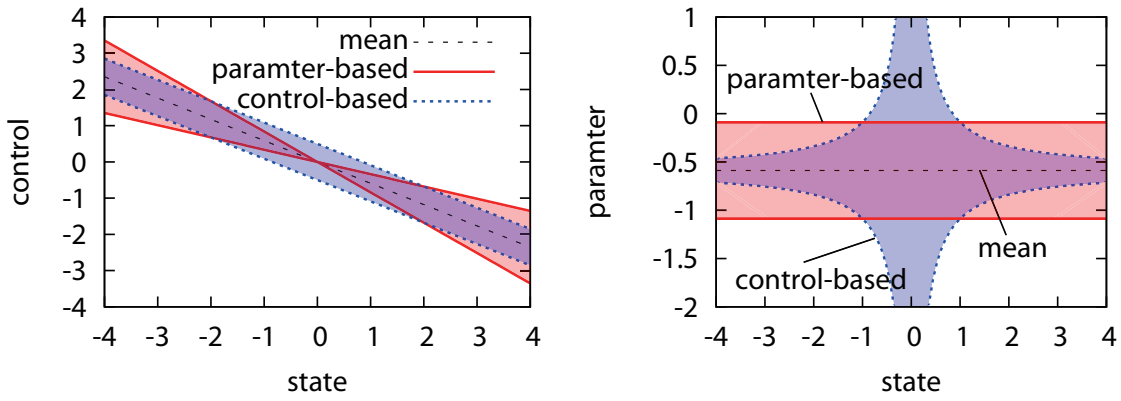

図 5.3: 線型 2 次形式制御タスクにおける CE と PE の違いを表した模式図．適当な σ に対して 1σ の領域を状態制御信号空間 (左) と状態-制御パラメータ空間 (右) に関して示している．

5.4.3 匍匐ロボットの前進動作獲得タスク

4.4.3 で用いた匍匐ロボットに提案手法を適用する．但し，4.4.3 では複数の局所解が存在するアームが 2 本のロボットを対象としたが，ここでは多峰性への対処は考えていないため図 5.4 に示すようにアームが 1 本のロボットを扱う．また，学習係数に関する特別なヒューリスティクスも用いない．政策パラメータの次元数は PE を用いる場合には $d_W = n(n+3)/2 = 27$，CE では $d_U = n + m(m+1)/2 = 9$ である．

比較は NPG(w)，つまり NPGPE，と NPG(u) および NAC(u) で行う．NAC[25] は政策勾配法に基づく手法として最も性能が良いとされている手法であり，自然政策勾配法と価値関数法である LSTD-Q[20] を Actor-Critic 法 [19] の枠組みで効果的に組合せたものである．NAC の計算量は NPG と等価であり，NAC(w) は NPGPE のような効率的な計算方法が無いため比較対象から除外した．

図 5.5 に学習曲線を示す．学習初期において NPG(w) は NAC(u) よりも性能が悪いが，目的の政策への収束は速い．NAC(u) は学習の終盤において NPG(u) の学習曲線と一致することも確認できる．NAC(u) は価値関数法と組合せることにより学習初期の政策改善が加速されているが，探査戦略が CE であるため性能が充分に引き出せていないものと思われる．図 5.6 は NPG(w) と NPG(u) による制御パラメータの推移を示している．初めか

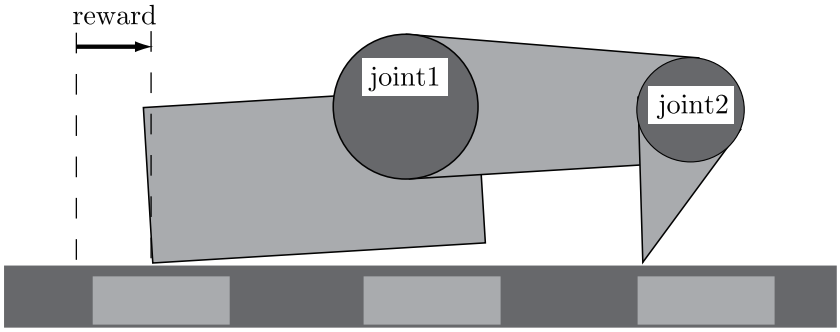

図 5.4: アームが 1 本の匍匐ロボット.

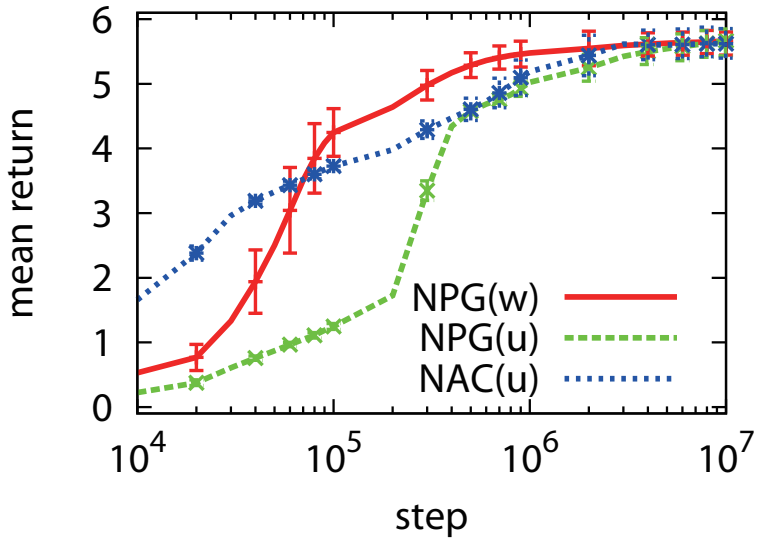

図 5.5: 匍匐ロボットでの NPG(w) と NPG(u) および NAC(u) の学習曲線の比較 (100 試行平均).

ら目標のパラメータへ向かわないのは，ロボットの獲得動作が徐々に変化していることを表している．NPG(w) と NPG(u) はパラメータの軌跡が異なることは，学習過程において現れるロボットの動作が異なることを表している．また，図中の矢印で示したパラメータ同士の関係の入れ替わりにより目的の動作の獲得フェーズへ移行するが，NPG(w) は NPG(u) よりフェーズの移行が早いことも確認できる．この差が学習曲線に現れていると考えられる．

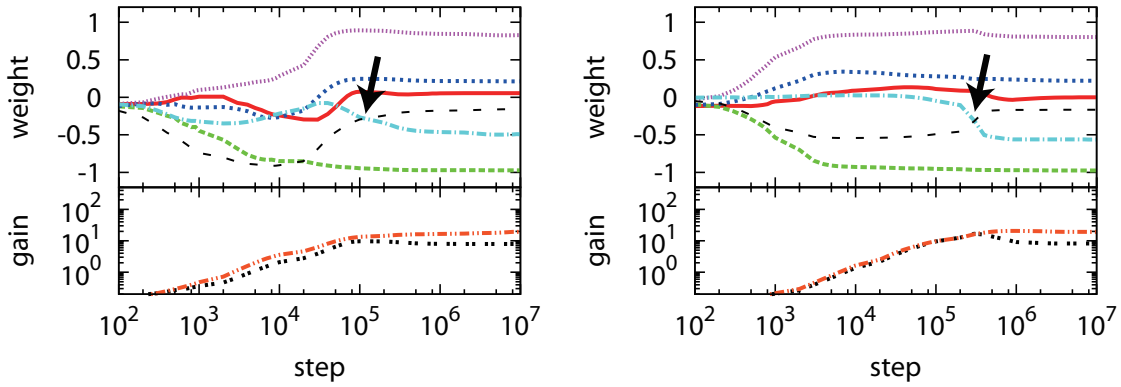

図 5.6: 匍匐ロボットでの NPG(w) と NPG(u) による制御パラメータの推移 (100 試行平均). NPG(w) での推移を左に，NPG(u) を右に示している．$gain_i = \sqrt{\sum_j \mathbf{w}_{i,j}}$，$weight_{i,j} = \mathbf{w}_{i,j}/gain_i$ として正規化したパラメータをプロットしている．ここで，$\mathbf{w}_{i,j}$ は i 番目の関節の j 番目のパラメータを表す.

5.5 おわりに

本章では確率モデルで表した制御パラメータに対する自然政策勾配法 NPGPE を提案した．提案手法は真の FIM を用いて計算量の面で効率的に自然政策勾配を推定することができる．実験により従来の政策勾配法と比べて学習性能が向上することを確認した.

今後の展望としては，局所最適政策への収束をさらに加速するために NPGPE を NAC の枠組みへ拡張することや，自然共役勾配法 [12] を導入することが考えられる．共分散のパラメータの取り方として，現在は共分散のコレスキー分解を用いることで共分散を半正定値にしているが，[8] の方法では正定値になるため，それを導入することもアルゴリズムの挙動を安定させるために有効であると考えられる.

第6章　おわりに

6.1　研究成果のまとめ

本研究では，多峰性の景観を有する制御パラメータの最適化を行う強化学習手法を実現する上で，多点探索に基づく DPS 法の構築が重要であるとの考えから，実数値 GA および政策勾配法に基づく手法を提案した．その成果を以下にまとめる．

成果 1　実数値 GA に基づく DPS の大域的探索能力を維持しながら効率化を行うために，インスタンスベース政策最適化においてインスタンスの役割に基づいたペアリングを行う枠組みに注目した．そして，インスタンスの役割に基づいたペアリングを利用して効率的に大域的探索を行うために，統計量の遺伝を満たす交叉的突然変異の拡張 XLM および直系保存の割合を調整できる世代交代モデルの拡張 CCM を導入した新しい実数値 GA の枠組み FLIP を提案した．FLIP を宇宙ロボットの姿勢制御，車両ロボットの車庫入れ，並列二重倒立振子振り上げ安定化制御のタスクに適用した結果，従来手法に比べて高い性能を示すことを確認した．

成果 2　政策勾配法が局所探索に基づく手法であるため政策の評価値の景観が多峰性を有する問題に適用した場合，最適政策の獲得に失敗する可能性が高いことを問題とし，集団で探索を行う政策勾配法として Population-based Policy Gradient 法を提案した．提案手法では各個体が局所探索により局所解を獲得し，それらの中から最良個体を選択することで大域的探索を行う．また，各個体の局所探索では環境とのインタラクション数を削減するために重点サンプリングを用いた経験の共有を行っている．提案手法を，複数の局所最適政策が存在する 2 分木タスクおよびアームが 2 本の匍匐ロボットで最適政策を獲得する

頻度の向上が認められた．

成果 3　政策勾配法は制御器に対して微分情報が利用できることや探索が停滞しにくいように設計されていることが想定されているため設計者の負担が大きいことを指摘した．そして，制御器のパラメータを確率モデルで表し，そのモデルのパラメータを自然勾配を用いて探索する NPGPE を提案した．提案手法は計算量の面で効率的に自然政策勾配を推定することができ，制御器の微分情報を必要としないという特徴を持つ．線型 2 次形式制御タスクおよびアームが 1 本の匍匐ロボットへ適用することで従来の政策勾配法と比べて学習性能が向上することを確認した．

6.2　今後の展望

本研究では，これまで余り考慮されてこなかった DPS における多峰性の景観を取り上げた意味では，DPS の応用範囲を広げたと考えられる．今後の課題として，実数値 GA と政策勾配法の融合・統合および政策勾配法の成果である PPG と NPGPE の融合・統合が挙げられる．これらについて以下で述べる．

FLIP と PPG の融合・統合　FLIP は多くの局所解が存在する多峰性の景観のもとでの大域的探索能力に優れている．PPG は数個程度の局所解を想定しており FLIP と比べて大域的探索能力は低いが勾配法に基づくため局所探索は高速である．そこで，実数値 GA による大域的探索能力を有しながら政策勾配法の局所探索の効率を実現する枠組みの構築が今後の課題となる．

PPG と NPGPE の融合・統合　PPG の局所探索性能は使用する政策勾配法に依存する．NPGPE は従来の政策勾配法と比べて優れた探索性能を示し，IBP のような微分情報が利用できない制御器に対しても適用可能であるという性質を有する．そのため，PPG の枠組みで NPGPE を利用することは PPG の探索効率の向上および FLIP との融合・統合を行ううえで実数値 GA と政策勾配法の差を埋める役割を担うと考えられる．

謝辞

本研究を遂行するにあたり，御指導御協力を頂いた，東京工業大学大学院の小林重信 教授，様々な御指摘，御助言をいただいた，同大学院 小野功 准教授，永田裕一 助教，大学評価・学位授与機構 宮崎和光 准教授，筑波大学大学院 佐久間淳 准教授に深く感謝する．

研究生活をともにした秋本洋平氏をはじめ，小林研究室，小野研究室の学生の方々には様々な面で支えていただいた．また，学生生活を通して多くの友人達に助けられ，支えていただいた．そのすべての方々に心より感謝する．

公表論文

学術論文

1 宮前 惇, 佐久間 淳, 小野 功, 小林 重信. インスタンスベース政策最適化のための実数値 GA と非ホロノミック系制御への適用. **人工知能学会論文誌**. 24(1):104 – 115 (2009)

2 大嶋 彈, 宮前 惇, 永田 裕一, 小林 重信, 小野 功, 佐久間 淳. UV 構造を考慮した適応的複製選択による実数値 GA の提案. **人工知能学会論文誌**. 25(2):290 – 298 (2010)

3 宮前 惇, 永田 裕一, 小野 功, 小林 重信. 多峰性景観下での直接政策探索：Population-based Policy Gradient 法の提案. **進化計算学会論文誌**, 2(1): 1 – 11 (2011)

国際会議

1 A. Miyamae, J. Sakuma, I. Ono, and S. Kobayashi. Optimization of Instance-based Policy Based on Real-coded Genetic Algorithms. In *Proceedings of the 2008 IEEE Conference on Soft Computing in Industrial Applications, SMCia/08*, pages 338 – 343 (2008)

2 D. Oshima, A. Miyamae, J. Sakuma, S. Kobayashi, and I. Ono. A New Real-coded Genetic Algorithm Using the Adaptive Selection Network for Detecting Multiple Optima. In *Proceedings of the 2009 IEEE Congress on Evolutionary Computation*, pages 1912 – 1919 (2009)

3 A. Miyamae, Y. Nagata, I. Ono, and S. Kobayashi. Natural Policy Gradient Methods with Parameter-based Exploration for Control Tasks. In *Advances in Neural Information Processing Systems 23*, pages 1660 – 1668 (2010)

国内学会等発表論文

1 宮前 惇, 佐久間 淳, 小野 功, 小林 重信. 実数値 GA によるインスタンスベース政策の最適化 〜FLIP の提案と非ホロノミック系制御への適用〜. **進化計算シンポジウム** *2007* **講演論文集** pages 127 – 234 (2007)

2 宮前 惇, 佐久間 淳, 小野 功, 小林 重信. 実数値 GA によるインスタンスベース政策の最適化. **第** *20* **回自律分散システムシンポジウム資料集** (2008)

3 大嶋 彈, 宮前 惇, 佐久間 淳, 小林 重信, 小野 功. ネットワークを利用した実数値 GA の新しい枠組みの提案. **計測自動制御学会 システム・情報部門学術講演会 講演論文集**, pages 589 – 594 (2008)

4 大嶋 彈, 宮前 惇, 佐久間 淳, 小林 重信, 小野 功. UV 構造を考慮した適応的複製選択に基づく実数値 GA の提案. **進化計算シンポジウム** *2008* **講演論文集**, pages 137 – 142 (2008)

5 宮前 惇, 佐久間 淳, 小野 功, 小林 重信. 確率的傾斜法に基づくオフポリシー型強化学習. **第** *21* **回自律分散システムシンポジウム資料集** (2009)

6 宮前 惇, 佐久間 淳, 小野 功, 小林 重信. 政策勾配法における挙動政策の適応的選択. **第** *36* **回知能システムシンポジウム資料集** (2009)

7 大嶋 彈, 宮前 惇, 佐久間 淳, 小林 重信, 小野 功. ネットワークを利用した適応的複製選択に基づく実数値 GA の提案. *21* **世紀** *COE* **プログラム 「エージェントベース社会システム科学の創出」 第** *4* **回若手フォーラム予稿集**, pages 56 – 61 (2009)

8 宮前 惇, 佐久間 淳, 小野 功, 小林 重信. 経験共有型多点政策勾配法. **第 *2* 回進化計算フロンティア研究会資料集**, pages 1–10 (2009)

9 宮前 惇, 佐久間 淳, 小野 功, 小林 重信. 多点政策勾配法による政策最適化. **進化計算シンポジウム *2009* 講演論文集**, pages 1–9 (2009)

参考文献

[1] Youhei Akimoto, Yuichi Nagata, Isao Ono, and Shigenobu Kobayashi. Bidirectional Relation between CMA Evolution Strategies and Natural Evolution Strategies. *Parallel Problem Solving from Nature XI*, pp. 154–163, 2010.

[2] S. Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, Vol. 10, No. 2, pp. 251–276, 1998.

[3] S. Amari and H. Nagaoka. *Methods of Information Geometry*. American Mathematical Society, 2007.

[4] J. Andrew Bagnell and Jeff Schneider. Covariant policy search. In *IJCAI'03: Proceedings of the 18th international joint conference on Artificial intelligence*, pp. 1019–1024, 2003.

[5] Jonathan Baxter and Peter L. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, Vol. 15, pp. 319–350, 2001.

[6] A. M. Bloch, M. Reyhanoglu, and N. H. McClamroch. Control and stabilizability properties of a nonholonomic controlsystem. In *Proc. 29th IEEE Conf. on Decision and Control*, pp. 1312–1314, 1990.

[7] R.W. Brockett. Asymptotic stability and feedback stabilization. *Differential Geometric Control Theory, Progress in Mathematics*, Vol. 27, pp. 180–191, 1983.

[8] Tobias Glasmachers, Tom Schaul, S. Yi, D. Wierstra, and J. Schmidhuber. Exponential natural evolution strategies. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, pp. 393–400, 2010.

[9] Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *The Journal of Machine Learning Research*, Vol. 5, pp. 1471–1530, 2004.

[10] L. Gurvits and Z. X. Li. Smooth time-periodic feedback solutions for nonholonomic motion planning. 1992.

[11] V. Heidrich-Meisner and C. Igel. Evolution strategies for direct policy search. In G. Rudolph, editor, *Parallel Problem Solving from Nature (PPSN X)*, Vol. 5199 of *LNCS*, pp. 428–437. Springer-Verlag, 2008.

[12] Antti Honkela, Matti Tornio, Tapani Raiko, and Juha Karhunen. Natural conjugate gradient in variational inference. In *ICONIP 2007*, pp. 305–314, 2008.

[13] Kokoro Ikeda. Exemplar-Based Direct Policy Search with Evolutionary Optimization. In *Proceedings of 2005 Congress on Evolutionary Computation CEC2005*, pp. 2357–2364, 2005.

[14] S. A. Kakade. A natural policy gradient. In *In Advances in Neural Information Processing Systems*, pp. 1531–1538, 2001.

[15] H. Kimura. Natural gradient actor-critic algorithms using random rectangular coarse coding. In *SICE Annual Conference*, pp. 2027 –2034, 2008.

[16] H. Kimura and S. Kobayashi. Reinforcement learning for continuous action using stochastic gradient ascent. In *Intelligent Autonomous Systems (IAS-5)*, pp. 288–295, 1998.

[17] H. Kimura and S. Kobayashi. Reinforcement learning by policy improvement making use of experiences of the other tasks. In *The 8th Conference on Intelligent Autonomous Systems (IAS-8)*, pp. 413–421, 2004.

[18] Hajime Kimura, Kazuteru Miyazaki, and Shigenobu Kobayashi. Reinforcement learning in pomdps with function approximation. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 152–160, 1997.

[19] V.R. Konda and J.N. Tsitsiklis. In *SIAM Journal on Control and Optimization*, pp. 1008–1014, 2000.

[20] M.G. Lagoudakis and Ronald Parr. Model-free least-squares policy iteration. *Advances in neural information processing systems*, Vol. 2, pp. 1547–1554, 2002.

[21] D. E. Moriarity, A. C. Schultz, and J. J. Grefenstette. Evolutionary algorithms for reinforcement learning. *Journal of Artificial Intelligence Research*, Vol. 11, pp. 199–229, 1999.

[22] R.M. Murray and S. S. Sastry. Nonholonomic motion planning: Steering using sinusoids. *IEEE Trans. on Automatic Control*, Vol. 38, pp. 700–713, 1991.

[23] Isao Ono, Yuichi Nagata, and Shigenobu Kobayashi. A Genetic Algorithm Taking Account of Characteristics Preservation for Job Shop Scheduling Problems. In *Proceedings of the International Conference on Intelligent Autonomous Systems 5*, pp. 711–718, 1998.

[24] Leonid Peshkin and Christian R. Shelton. Learning from scarce experience. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 498–505, 2002.

[25] Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomputing*, Vol. 71, No. 7–9, pp. 1180–1190, 2008.

[26] Doina Precup, Richard S. Sutton, and Sanjoy Dasgupta. Off-policy temporal-dierence learning with function approximation. In *In Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

[27] Doina Precup, Richard S. Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, pp. 759–766. Morgan Kaufmann, 2000.

[28] J. Randløv, A. G. Barto, and M. T. Rosenstein. Combining reinforcement learning with a local control algorithm. In *Proc. Seventeenth International Conference on Machine Learning*, pp. 775–782, 2000.

[29] Silvia Richter, Douglas Aberdeen, and Jin Yu. Natural actor-critic for road traffic optimisation. In *Advances in Neural Information Processing Systems 19*, pp. 1169–1176. MIT Press, Cambridge, MA, 2007.

[30] Juan Carlos Santamar'ia, Richard S. Sutton, and Ashwin Ram. Experiments with reinforcement learning in problems with continuous state and action spaces. *Adaptive Behavior*, Vol. 6, pp. 163–218, 1998.

[31] Frank Sehnke, C Osendorfer, T Rueckstiess, A. Graves, J. Peters, and J. Schmidhuber. Policy gradients with parameter-based exploration for control. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, pp. 387–396, 2008.

[32] Christian R. Shelton. Policy improvement for pomdps using normalized importance sampling. In *In Proceedings of the Seventeenth International Conference on Uncertainty in Artificial Intelligence*, pp. 496–503, 2001.

[33] J.W. Sheppard and S.L. Salzberg. A teaching strategy for memory-based control. *Artificial Intelligence Review*, Vol. 11, pp. 343–370, 1997.

[34] Yi Sun, Daan Wierstra, Tom Schaul, and Juergen Schmidhuber. Efficient natural evolution strategies. In *GECCO '09: Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, pp. 539–546, 2009.

[35] R. S. Sutton. Policy gradient method for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, Vol. 12, pp. 1057–1063, 2000.

[36] R. S. Sutton and A. Barto. *Reinforcement Learning : An Introduction.* The MIT Press, 1998.

[37] C. Tsuchiya, K. Ikeda, J. Sakuma, I. Ono, and S. Kobayashi. Instance-Based Policy Search Using Binomial Distribution Crossover and Iterated Refreshent. In *Proceedings of 2006 Congress on Evolutionary Computation CEC2006*, pp. 1102–1109, 2006.

[38] D. Whitley, S. Dominic, R. Das, and C. W. Anderson. Genetic reinforcement learning for neurocontrol problems. *Machine Learning*, Vol. 13, pp. 259–283, 1993.

[39] Daan Wierstra, Er Foerster, Jan Peters, and Juergen Schmidhuber. Solving deep memory pomdps with recurrent policy gradients. In *In International Conference on Artificial Neural Networks*, 2007.

[40] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Machine Learning*, pp. 229–256, 1992.

[41] NAKAMURA Yutaka. Policy gradient method for a policy function with probabilistic parameters. *IEICE technical report. Neurocomputing*, Vol. 107, No. 542, pp. 343–348, 2008-03-05.

[42] 井口 圭一, 木村 元, 小林 重信. GA による並列二重倒立振子の振り上げ安定化制御. 第 13 回自律分散システムシンポジウム, pp. 277–282, 2001.

[43] 塩川 祐介, 土谷 千加夫, 佐久間 淳, 小野 功, 小林 重信. インスタンスベース政策学習による非ホロノミック系の制御. 第 49 回自動制御連合講演会, 2006.

[44] 塩川 祐介, 土谷 千加夫, 佐久間 淳, 小野 功, 小林 重信. インスタンスベース政策学習による非ホロノミック系制御の実験的考察. 第 19 回自律分散システムシンポジウム, pp. 73–78, 2007.

[45] 外本 伸治, 松野 崇. マニフォールドを利用した宇宙ロボットの制御に関する検討. 日本ロボット学会誌, Vol. 23, No. 5, pp. 594–601, 2005.

[46] 喜多 一, 小野 功, 小林 重信. 実数値 GA のための正規分布交叉に関する理論的考察. 計測自動制御学会論文集, Vol. 35, No. 11, pp. 1333–1339, 1999.

[47] 喜多 一, 小野 功, 小林 重信. 実数値 GA のための正規分布交叉の多数の親を用いた拡張法の提案. 計測自動制御学会論文集, Vol. 36, No. 10, pp. 875–883, 2000.

[48] 木村元, 山村雅幸, 小林重信. 部分観測マルコフ決定過程下での強化学習： 確率的傾斜法による接近. 人工知能学会誌, Vol. 11, No. 5, pp. 761–768, 1996.

[49] 木村元, 小林重信. Actor に適正度の履歴を用いた actor-critic アルゴリズム：不完全な value-function のもとでの強化学習. 人工知能学会誌, Vol. 15, No. 2, pp. 267–275, 2000.

[50] 木村元, 小林重信. 確率的 2 分木の行動選択を用いた antor-critic アルゴリズム - 多数の行動を扱う強化学習 -. 計測自動制御学会論文集, Vol. 37, No. 12, pp. 1147–1155, 2001.

[51] 高橋 治, 木村 周平, 小林 重信. 交叉的突然変異による適応的近傍探索. 人工知能学会論文誌, Vol. 16, No. 2, pp. 175–184, 2001.

[52] 佐藤 浩, 小野 功, 小林 重信. 遺伝的アルゴリズムにおける世代交代モデルの提案と評価. 人工知能学会誌, Vol. 12, No. 5, pp. 734–744, 1997.

[53] 小林 啓吾, 潮 俊光. 大域的な座標変換による非ホロノミック車両システムのハイブリッド制御. システム制御情報学会論文誌, Vol. 17, No. 1, pp. 1–9, 2004.

[54] 小林 重信. 実数値 GA のブレークスルーへ向けて. 進化計算シンポジウム 2007, pp. 1–10, 2007.

[55] 池田心, 小林重信. Ga の探索における uv 現象と uv 構造仮説. 人工知能学会論文誌, Vol. 17, No. 3, pp. 239–246, 2002.

[56] 西村 政哉, 吉本 潤一郎, 時田 陽一, 中村 泰, 石井 信. 複数制御器の切換学習法による実アクロボットの制御. 電子情報通信学会論文誌, Vol. J88-A, pp. 646–657, 2005.

[57] 土谷千加夫, 木村元, 小林重信. 重点サンプリングを用いた政策勾配の推定による子個体生成. 計測自動制御学会 第 31 回知能システムシンポジウム, pp. 145–150, 2004.

[58] 川谷 亮治, 山口 尊志. 並列型 2 重倒立振子系の解析とその安定化. 計測自動制御学会論文誌, Vol. 29, No. 5, pp. 572–580, 1993.

[59] 中村 仁彦. 非ホロノミックロボットシステム. 日本ロボット学会誌, Vol. 11, No. 4, pp. 521–528, 1993.

[60] 土谷 千加夫, 塩川 祐介, 池田 心, 佐久間 淳, 小野 功, 小林 重信. ハイブリッド GA によるインスタンスベース政策学習　〜SLIP の提案と評価〜. 計測自動制御学会論文誌, Vol. 42, No. 12, pp. 1344–1352, 2006.