計測・通信の遅れに頑健なフィードフォワード方策を統合する深層強化学習技術

代表研究者　小林　泰介　奈良先端科学技術大学院大学・助教
共同研究者　芳澤　健太　奈良先端科学技術大学院大学・修士課程

1 はじめに

深層学習技術の発展に伴い，それを関数近似器として統合した強化学習の一種「深層強化学習」が驚くべき成果を上げ始めている．特に，ビデオ・ボードゲームでは人を打倒するレベルの方策を獲得できるようになってきており，社会的にも大きな話題を集めている．この深層強化学習技術はロボットの制御にも活用されており，ロボットの制御器に相当する方策を学習させて，例えばカメラ画像から物体の検出・認識を介することなく，物体の操作を直接実現している．

しかし，深層強化学習の方策が状態入力に基づいて行動を決定する，ある種のフィードバック制御器であることに着目すると，計測・通信などの遅れあるいは外れ値の影響を強く受けることが容易に想像できる．そのため，対象とする問題やロボットのハードウェアがよりダイナミックなものとなると，フィードバック制御が間に合わずに性能を最大限に引き出せない（図1）．制御工学においては，これらの計測・通信にまつわる問題はフィードフォワード制御器を併用することで補償する技術が確立している．特に，川人らが提案したフィードバック誤差学習では，フィードバック制御における誤差情報を基にフィードフォワード制御器を修正することで対象とする問題のモデル化誤差に対応することが可能となっている．ただし，この技術では参照軌道への追従問題のみ対象としているため，強化学習が扱えるより一般的な問題への適用は難しい．

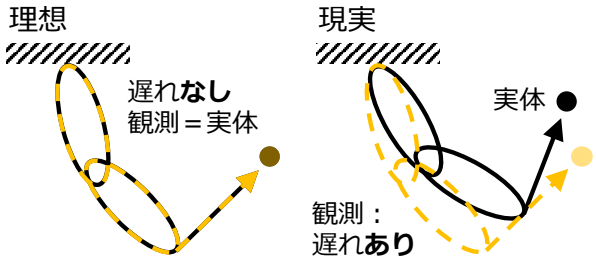

図 1 計測・通信遅れの悪影響

そこで本研究では，フィードバック制御器（方策）だけでなくフィードフォワード方策を内包した上で，それらを統一的に学習することが可能な新しい深層強化学習技術（図 2）の開発を目指す．この目的実現のために，本調査の範囲では，(1) リザーバコンピューティングによるフィードフォワード方策の実装，(2) 変分オートエンコーダを活用したフィードバック・フィードフォワード方策の同時学習，(3) 2 種類の方策から行動を決定するための方策合成則の考案，(4) システムの時間発展を扱えるオンライン深層強化学習の開発，の 4 項目について技術開発を図った．また，提案技術の検証に向けて，(1) 可変剛性アクチュエータを搭載したロボットアームの開発，(2) ロボットのシミュレーションモデルの構築，(3) シミュレーションにおけるボール遠投タスクの学習実験，の 3 項目について実施した．結果として，提案技術によりフィードバック方策の獲得と同時にフィードフォワード方策へと知識を転写することで，フィードバック方策と同等の性能を持つフィードフォワード方策を十分に学習可能であることを確認した．

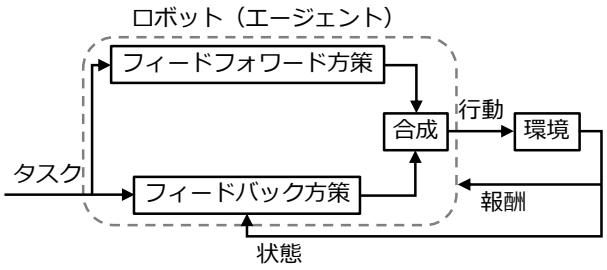

図 2 フィードフォワード方策を内包する深層強化学習

2 フィードフォワード方策を内包する深層強化学習

2-1 リザーバコンピューティングによるフィードフォワード方策の実装

フィードフォワード方策は状態に非依存の確率関数として定義される必要がある．最も単純なモデルとして，経過時間を入力して方策用の確率モデルを出力するニューラルネットワークが考えられるが，最大経過

時間が未定義である以上，実用上は入力が発散する可能性があり不適であった．代替モデルとして，入力の時間発展を内部ダイナミクスを有することで表現可能なリカレントニューラルネットワークを用いたモデルである．前回の出力を入力へフィードバックすることで過去の出力系列を考慮可能となる．ただし，リカレントニューラルネットワークの欠点として学習の困難さが懸念された．具体的にはニューラルネットワークの学習手段として代表的な誤差逆伝播法を時間方向にも逆伝播する必要があるため，計算コストの高さやどの程度過去まで考慮すべきかのチューニングなどが障害として考えられた．

そこで，より学習が簡単な手法として，人の運動を制御している小脳のモデルとされるリザーバコンピューティング（図 3）を採用した．これは，大規模な中間層は事前にランダム生成された固定のダイナミクスに従って出力を生成し，この出力を線形回帰に則り目的の出力（今回はフィードフォワード方策）へと整形する手法である．学習は線形回帰となるため，非常に簡単化される特徴がある．どれだけ過去の情報を維持できるのか，すなわち時定数は事前のダイナミクス設計段階で決定可能であり，各ニューロンで異なる時定数を与えることも可能であるため，大量のニューロンの中から問題に適した時定数を持ったものが自動的に選出されることで適切にフィードフォワード方策を表現できる．

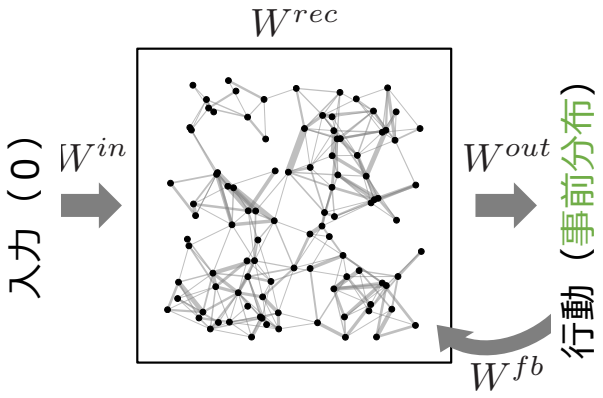

図 3 リザーバコンピューティングのモデル

具体的な実装として，深層学習用ライブラリである PyTorch をベースにして Python で記述した．入力項は常に 0（バイアス項の 1 のみ）とし，最終的に方策が出力して環境に渡した行動をフィードバック項として与えた．出力項は行動空間の次元数と一致する対角多変量ガウス分布とし，実際にはその平均と分散を出力させた．なお，リザーバコンピューティングの内部ダイナミクスが発散しないよう，活性化関数として飽和型の関数を採用した．

2-2 変分オートエンコーダを活用したフィードバック・フィードフォワード方策の同時学習

提案技術の核はフィードフォワード方策を如何にして獲得するかである．上記のリザーバコンピューティングにより方策のモデル化は為されたが，その学習則が不可欠である．特に，一般的な強化学習における方策改善は状態依存の方策を想定しているため，フィードバック方策の更新のための情報をフィードフォワード方策に逆伝播しても有効な更新になるとは限らない．

そこで，強化学習における直接的な方策改善に代わるフィードフォワード方策の学習則を提案する．ここで，方策が確率モデルで与えられていること，フィードバック方策とフィードフォワード方策は異なる情報で条件付けられた条件付き確率であることに注目する．すると，深層学習技術の一つである変分オートエンコーダを駆使して，それが抽出する低次元潜在空間の一部が行動空間で対応付けられると仮定することで，２つの方策間に制約を与えることができる．具体的には，変分オートエンコーダは状態入力の復元誤差と潜在空間を構築する潜在変数の状態入力によって定まる事後分布と状態非依存の事前分布との間のダイバージェンスを最小化する問題として定式化されている．この状態依存の事後分布と状態非依存の事前分布が乖離しないようにする制約の働きによって，フィードバック方策からフィードフォワード方策への知識の転写を可能とする．なお，変分オートエンコーダでは状態依存の事後分布からサンプルした値から状態入力を復元するのだが，提案技術では次節で述べるように，フィードバック方策とフィードフォワード方策を合成した新しい方策から行動をサンプルするため，サンプルする確率分布が変化することに注意する必要が生まれた．すなわち，フィードバック方策と合成後の方策での重点サンプリングの項が復元誤差に重みとして与えられることが数理的な導出よりわかった．ただし，重点サンプリングは数値的に不安定になり得るため，相対重点サンプリングに置き換えることで数値的安定な計算を可能とした．

また，本来の変分オートエンコーダでは状態入力を復元するモデルであるが，次状態を予測するモデルへと変換（復元誤差に代わり予測誤差を最小化）することで，潜在空間および行動空間の方策にダイナミクスの情報が逆伝播可能となる．これは特にフィードフォワード方策の学習の補助的な役割を果たすことが期待できる．ただし，全ての状態遷移を一律に扱ってしまうと，場合によっては強化学習における価値（将来受

け取る報酬の総和の期待値）が低下してしまう状態遷移を促すような方策に学習されかねない．価値を高める状態遷移を優先して学習してフィードフォワード方策へと反映するために，価値の改善量を示す TD 誤差を予測誤差の重みとして用いる手法を加えた．これにより，価値が改善する状態遷移を予測可能な変分オートエンコーダとして学習が進み，同時に価値の改善を促すようなフィードフォワード方策の獲得を補助できる．

2-3 フィードバック・フィードフォワード方策の合成則の考案

次に，提案技術が有する２種類の方策を適切に合成して行動をサンプルすることを考える（図 4）．当初は２つを混合分布のコンポーネントと見立てて混合係数を学習する手法を考えていたが，この場合の混合係数を学習する簡単かつ効果的な手法がなかった．代わりに，フィードバック・フィードフォワード方策ともどもガウス分布でモデル化しており，ガウス分布同士の積もまたガウス分布になる特性に注目し，合成後の方策を２つの方策の積により定義することとした．合成後の平均は各ガウス分布の精度（分散の逆数）の比に基づいた値となるため，基本的にはフィードバック方策が優先されながら，学習が進みフィードバック方策の知識が上記の手法でフィードフォワード方策へ転写されると，２つを対等に扱うことが期待できる．また，計測・通信の遅れや外れ値によってフィードバック方策の精度が低下した場合には，フィードフォワード方策を優先利用することでそれらの問題を補償することが可能となる．

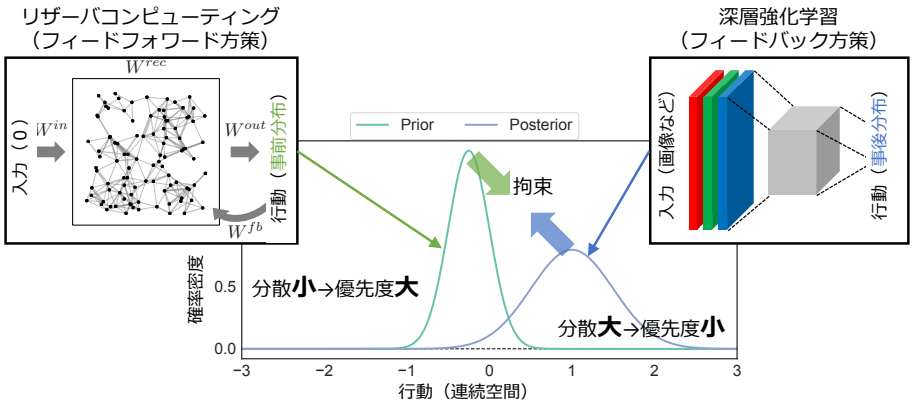

図 4 フィードバック・フィードフォワード方策の合成則

2-4 オンライン深層強化学習の開発

提案技術の問題点として，フィードフォワード方策としてリザーバコンピューティングというリカレントニューラルネットワークの一種を用いていることで状態と行動の時間発展が各時刻の出力に影響を与えてしまうことに付随する，経験データの再利用が困難になる点が挙げられる．近年の深層強化学習ではこの経験データの再利用を前提とした技術体系が形成されているため，ストリーミング形式に経験データを１回のみ用いるオンライン深層強化学習を実現するには別の技術体系を利用する必要がある．特に，経験データの再利用ができないことで低下する学習効率を再度向上させる手法が新たに必要となる．また，ミニバッチデータの平均的振る舞いに従って更新できないことで生じる更新ノイズの影響を外れ値に頑健な最適化技術で除外することも重要である．

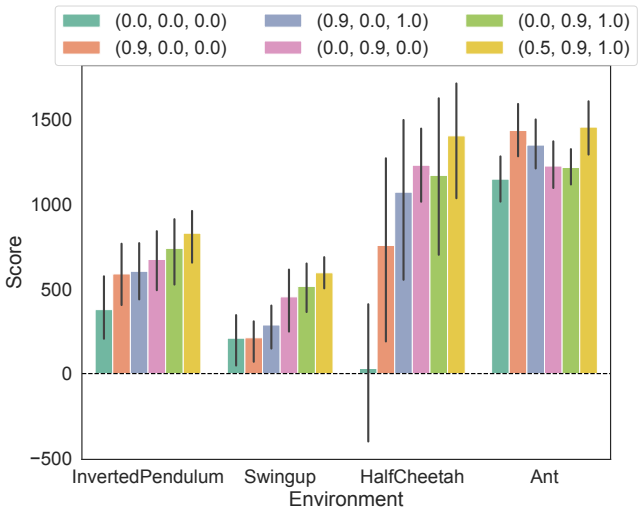

図 5 オンライン深層強化学習の検証
黄色（0.5，0.9，1.0）が提案手法

そこで本研究調査では，新たにオンライン深層強化学習に関する開発を進め，学習効率の良い手法を提案した．具体的には，オンライン強化学習で利用されていた適正度履歴を深層強化学習においても利用可能となるように，その障害となっていた過去の情報の一律な減衰率を適応的な減衰へと転換する手法を開発した．特に，過去の情報が現在も利用可能であるかを出力の乖離度合いから確認し，それに応じて減衰を促すことに加え，履歴を多層化することで減衰率を段階的に切り替えるようにした．これらの改善により適正度履歴

は学習効率の改善性能を保持しながら深層強化学習においても利用可能となった．また，外れ値に頑健な最適化技術として，申請者が開発を進めている確率的勾配降下法の一種 TAdam を利用することとした．結果として，一般的な強化学習用ベンチマークでその学習効率の改善を確認できた（図 5）．

3 実証実験

2-1 可変剛性アクチュエータを搭載したロボットアームの開発

提案技術の実証実験に向けたシステムとして，可変剛性アクチュエータを搭載したロボットアームを製作した．まず，要となる可変剛性アクチュエータとして，予定していた qb robotics 社製の qbmove advanced を 4 個含む qbmove base kit を購入した．この qbmove advanced はモジュール型のアクチュエータとなっており，付属のアタッチメントを用いてアクチュエータ同士を適切な軸構成で接続することで多自由度ロボットアームを構築した．また，実証実験のタスクとしてボールの遠投を想定していたため，ボールを投げやすいようにロボットアームを中空に吊るためのフレームを別途設計・製作した（図 6 左）．単一アクチュエータとの通信には，qb robotics 社より提供されている Robot Operating System (ROS)で開発されたパッケージをベースに実装して，目標関節角度と目標剛性を指令値として送信するとともに，関節角度・角速度・トルク・剛性などを一定周期で計測・受信可能なシステムを構築した．また，同 ROS パッケージを適切に組み合わせることで複数のアクチュエータとの通信も一定周期で実行可能となり，ロボットアームとしての制御に向けた枠組みを完成させた．

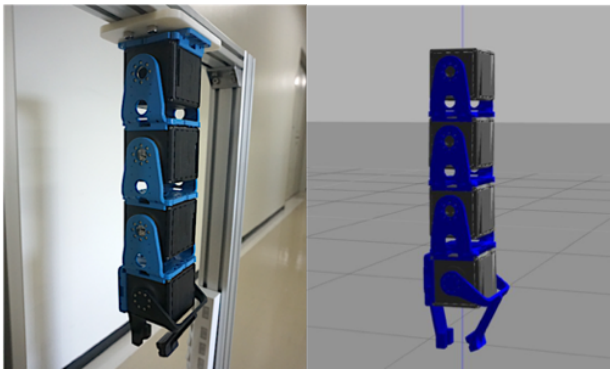

図 6 可変剛性アクチュエータを搭載したロボットアーム

構築した実証実験用のロボットシステムを，強化学習における環境として一般的なベンチマーク環境と同じように定義・利用するために， 強化学習のベンチマーク用フレームワークである OpenAI Gym と ROS を連携可能なオープンソースパッケージをベースにして，ソフトウェアレベルでのシステム構築および動作確認を進めた．特に，強化学習のための空間定義として，可変剛性アクチュエータは目標関節角度と目標剛性を指令値として受け取るものの，確率的な方策関数による位置レベルでのランダムな行動指令はハードウェアレベルでの負担が大きいことが懸念された．そこで行動空間には，目標関節角加速度と目標剛性変化速度を定義し，環境内部でそれらを数値積分することでアクチュエータへの指令値を生成するようにした．それに伴い，強化学習を成立させるために，状態空間には計測可能な関節角度・角速度・トルク・剛性だけでなく，数値積分の結果得られた指令値も加えることとした．

最終的に，グリッパの 1 自由度とアームの 3 自由度で構成される 4 自由度ロボットアームを構築し，行動空間は前述の通り，各アクチュエータの目標関節角加速度と目標剛性変化で計 8 次元，状態空間は，各アクチュエータの角度・角速度・トルク・剛性・目標角度・目標剛性とボールの水平・鉛直方向の位置・速度で計 28 次元と定めた．

2-2 ロボットのシミュレーションモデルの構築

提案技術を実証するにあたり，高速かつ安定した学習結果を得るためにシミュレーションモデルを構築した．シミュレーション環境には ROS との親和性の高い Gazebo を採用した．可変剛性アクチュエータ単体との通信用 ROS パッケージをシミュレーションでも利用可能なよう修正し，同等の情報の相互通信を実現した．アクチュエータレベルでの URDF モデルは github にて公開されていたため，それを基に実機の仕様に合わせてロボットアームの URDF モデルを記述した（図 6 右）．ただし，公開されていたモデルの動作範囲などが実機と一致しておらず，動作確認時に不自然な挙動となってしまったため，公開されていたアクチュエータ単体での URDF モデルを適宜修正することとした．

また，シミュレーションにおいてボールの遠投タスクを実行可能とすべく，強化学習における試行開始時の初期化処理として，ボールをロボットアームが把持するように初期化処理を追加した．特に，フィードフ

ォワード方策は初期状態のゆらぎに脆弱であることが懸念されたため，可能な限り一定の初期状態となるように微調整をした．結果として，初期状態をほぼ一定にでき，フィードフォワード方策でも安定してタスクを実現可能な条件を整えられた．

2-3 シミュレーションにおけるボール遠投タスクの学習実験

深層強化学習では方策の学習のために多くの試行を必要とする．特に，本提案では学習効率を重視せずにフィードフォワード方策を包括的に学習する技術に注目をしているため，この試行数を軽減することは期待できない．そこで，提案技術のコンセプトレベルでの検証として動力学シミュレーションである Gazebo でのボール遠投タスクの学習を実施し，その傾向から提案技術の妥当性を検証した．

タスクを達成する上で不可欠な報酬関数の設計として，ボールが地面にぶつかる直前の飛距離（単位は cm）に加えて，ボールの位置・速度を初期条件として自由落下した場合に想定される仮想的な飛距離（単位は m）を毎時刻与えることにした．この仮想的な飛距離の項は，１試行終了時の報酬だけでは学習が難化することが知られているため，学習を少しでも加速するための対策である．単位を変えることで実際のボールの飛距離に関する報酬のスケールを 100 倍しており，ボールを投げたほうがより高い価値となるように調整した．

フィードバック・フィードフォワード方策の関数近似に用いたネットワーク設計は，フィードバック方策に関しては 1 層当たり 100 個のニューロンを含む 5 層の中間層（＋入力層と出力層）で設計した．フィードフォワード方策に関してはリザーバコンピューティングのリザーバ層に 500 個のニューロンを持たせ，各ニューロンがランダムな時定数で定義されるダイナミクスを持つよう設計した．

ロボットは 30fps で制御し，その都度計測した状態からフィードバック方策を生成，フィードフォワード方策と合成して行動をサンプルさせた．なお，実際に 30fps でロボットの状態を計測すると，特にトルク・剛性にノイズが多く含まれてしまったため，指数移動平均によるローパスフィルタでノイズ除去を試みた．ボールを投擲しなかった場合でも試行を終了させるために最大 300 ステップ経った場合には強制的に試行を終了させて初期化，次の試行に移行させた．学習のために合計 500 試行実施した．

まず，予備的な検証として，設計したタスクがどの程度難しいものなのかを確認した．これには従来のフィードバック方策のみを学習する深層強化学習を利用した．検証の中で，経験的にだが，全状態入力から全行動へ写像する方策を直接的に学習しようとするには，設計したボール遠投タスクの難易度が高く安定した結果が得られないことが判明した．タスクを簡単化する手段として，特に行動空間（方策）を位置制御用と剛性制御用に分割して，それらを順番に学習するカリキュラムを構築すると，ある程度安定した学習が実現できた．ただし，この学習カリキュラムによる解決策はフィードフォワード方策を同時に学習する場合には学習プロセスが複雑化してしまい，挙動の定量的・定性的な分析が困難になることを懸念して，次からの検証では除外した．

学習時のボールの飛距離（単位は cm）をスコアとして図 7 にプロットした．学習初期はボールを投げられないか反対方向に投げてしまっていたところ，試行を繰り返す内に前方への投げ方を習得して，最終的には 400 試行辺りから安定して 50cm 程度の投擲を実現した．途中でボールの飛距離が 0 cm となっているのは，ボールを投げずに前方に突き出す局所解（仮想的な飛距離により発生）に陥りかけたためだと考えられる．

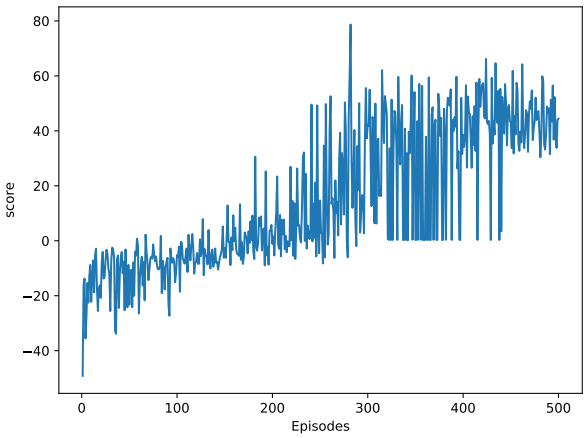

図 7 ボールの飛距離（cm）の学習過程

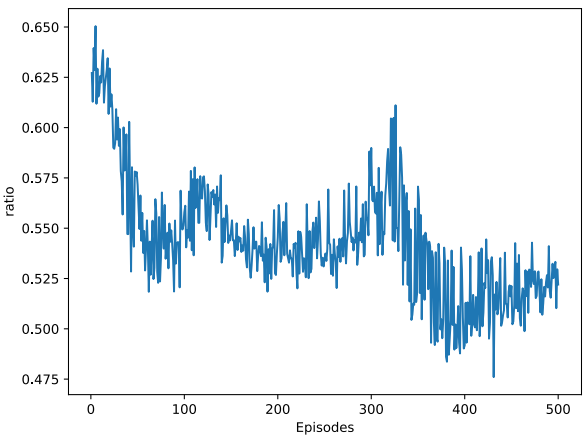

図 8 方策の合成比の学習過程

学習時のフィードバック・フィードフォワード方策の合成比の 1 試行当たりの平均を図 8 にプロットした．なお，1 に近いとフィードバック方策が，0 に近いとフィードフォワード方策が支配的となっている．学習途中，特に学習初期やボールの飛距離が伸びた 300 試行辺りで合成比が 0.5 以上となっており，フィードバック方策が先に学習を進めて高い精度を獲得したことがわかる．また，その直後には合成比が低下することから，フィードバック方策の知識が適切にフィードフォワード方策に転写され，フィードフォワード方策も高い精度を獲得したことが読み取れる．最終的には 400 試行辺りから合成比が 0.5 程度に収束しており，安定したボール投擲が可能な方策をフィードバック・フィードフォワード方策ともども獲得したといえる．

実際に，学習後の方策によって生成したロボットの挙動を図 9 に示す．合成した方策を用いた場合，ロボットはボールを前に突き出しながら手放すことで飛距離を稼いでいることがわかる．大域的最適解は一旦後ろに振ってから反動を利用してボールを投げる動作であり，その場合は 80cm 程度投げられるものの，安定した解としてはこのような動作となった．一方で，合成せずにフィードフォワード方策のみを用いて運動を生成した場合においても，合成した方策を用いた場合と軌道・タイミングともどもほぼ同じ挙動を示すことがわかった．ボールの飛距離自体もどちらも 50cm 程度であった．このことから，提案技術によって本来はフィードバック方策しか学習できない深層強化学習が同時にフィードフォワード方策を最適化して，状態入力を計測できない場合であっても運動を生成可能であることが確認できた．

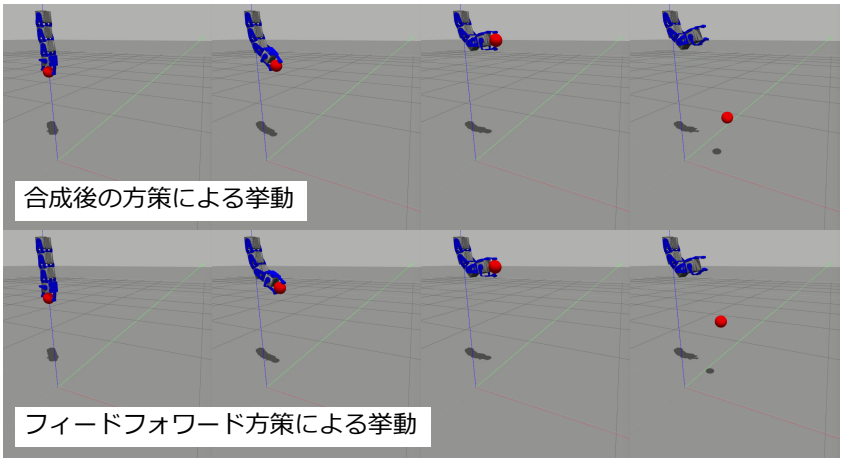

図 9 学習後の方策で生成した運動の様子

4 おわりに

本研究調査では，ロボットの動特性を最大限に発揮することでダイナミックな運動を実現するための新しい深層強化学習技術を開発・検証することを目的とした．その中で，一般的な強化学習で獲得される方策は状態依存のフィードバック制御器である点と，制御工学の文脈では状態入力の計測・通信遅れで生じてしまう運動性能の低下を補償するためにフィードフォワード制御器を併用する概念に注目した．すなわち，ダイナミックな運動生成に必要不可欠なフィードフォワード方策を同時に学習・獲得可能な深層強化学習技術を開発した．

まずはフィードフォワード方策を出力するニューラルネットワークとして，内部ダイナミクスを持つことでシステムの時間発展を扱えるリザーバコンピューティングを採用した．また，フィードバック方策と同時に学習する上で，変分オートエンコーダを援用した２つの方策間の制約により，フィードバック方策の知識をフィードフォワード方策へ転写する手法を提案した．その際に，変分オートエンコーダを予測モデルへと転換させつつ，より価値の高い状態へ遷移した時のみ予測精度向上を促すことで，方策がより価値の高い状態への遷移を促すような補助的な学習信号を生成した．そして，フィードバック・フィードフォワード方策の２つを適切に合成した上で行動を生成するために，それぞれをガウス分布でモデル化した後に，その積もまたガウス分布となる特性を利用して，合成後の方策をそれらの積で定義した．

提案技術の検証用環境として，可変剛性アクチュエータを搭載したロボットアームを製作し，そのシミュレーションモデルを用いて，ボールの遠投タスクの学習を試みた．強化学習のベンチマーク問題として標準化するために，状態・行動空間を定義し，ボールの（仮想）飛距離に基づく報酬関数を設計した．予備検証として，定義したタスクが可解であるかを従来の深層強化学習を用いて確認した．結果として，比較的難度が高いためか学習が不安定であるものの，例えば位置制御・剛性制御を分割して学習カリキュラムを構築することで十分に学習可能な問題であることを確認した．

最後に，提案技術を構築したボール遠投タスクに適用した．その結果，ボールの飛距離は試行を繰り返す過程で増加していき，最終的には安定して 50cm 程度投擲できるようになった．このとき，２つの方策の剛性比を確認すると，フィードバック方策が支配的な学習初期から，学習の進展に合わせてフィードフォワード方策の影響度が高まり，適切にフィードフォワード方策への知識の転写ができることがわかった．実際，学習後のフィードフォワード方策で動作を生成してみると，合成後の方策と遜色ない結果を獲得できた．

今回の調査を通して,新たに状態遷移の予測モデルを適切に利用することが重要であることが見えてきた．そこで，さらなる調査として，変分オートエンコーダを改良した変分リカレントニューラルネットワークを援用することで状態遷移の予測･フィードフォワード方策の獲得を同時に進行する技術開発を目指す.また，フィードバック・フィードフォワード方策で制御周期が異なる場合に２つをどのように合成するか新たに考える必要がある．最後に，実機での検証が不足してしまったため，今後は実機で検証しやすいタスクを設定しつつ提案技術の検証と改善を進めたい.

【参考文献】

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. nature, Vol. 518, No. 7540, pp. 529–533, 2015.

Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction. MIT press, 2018.

Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In IEEE international conference on robotics and automation, pp. 3389–3396. IEEE, 2017.

Mitsuo Kawato and Hiroaki Gomi. A computational model of four regions of the cerebellum based on feedback-error learning. Biological cybernetics, Vol. 68, No. 2, pp. 95–103, 1992.

Jun Nakanishi and Stefan Schaal. Feedback error learning and nonlinear adaptive control. Neural Networks, Vol. 17, No. 10, pp. 1453–1465, 2004.

Mantas Lukoˇsevi ˇcius and Herbert Jaeger. Reservoir computing approaches to recurrent neural network training. Computer Science Review, Vol. 3, No. 3, pp. 127–149, 2009.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

Manuel G Catalano, Giorgio Grioli, Manolo Garabini, Fabio Bonomo, Michele Mancini, Nikolaos Tsagarakis, and Antonio Bicchi. Vsa-cubebot: A modular variable stiffness platform for multiple degrees of freedom robots. In IEEE international conference on robotics and automation, pp. 5090–5095. IEEE, 2011.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In Advances in neural information processing systems, pp. 8026–8037, 2019.

Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. Neural computation, Vol. 25, No. 5, pp. 1324–1370, 2013.

Harm Seijen and Rich Sutton. True online td (lambda). In International Conference on Machine Learning, pp. 692–700, 2014.

Wendyam Eric Lionel Ilboudo, Taisuke Kobayashi, and Kenji Sugimoto. Tadam: A robust stochastic gradient optimizer. arXiv preprint arXiv:2003.00179, 2020.

Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. Ros: an open-source robot operating system. In ICRA

workshop on open source software, Vol. 3, p. 5. Kobe, Japan, 2009.
Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. arXiv preprint arXiv:1606.01540, 2016.
Nathan Koenig and Andrew Howard. Design and use paradigms for gazebo, an open-source multi-robot simulator. In IEEE/RSJ International Conference on Intelligent Robots and Systems, Vol. 3, pp. 2149–2154. IEEE, 2004.
Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In Advances in neural information processing systems, pp. 2980–2988, 2015.

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
オンライン深層強化学習に向けた適応型適正度履歴	自律分散システム・シンポジウム	2020 年 1 月
深層強化学習を用いた柔軟ロボットアームの投擲動作の獲得	ロボティクス・メカトロニクス講演会	2020 年 5 月