# 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（延長）


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．強化学習は基本的に対象とするシステムのモデル情報が不要であるため，より複雑な問題への適用事例が報告され始めている．

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，対象とする問題によってはフィードバック制御の本質的な欠点により目標を達成できない．制御工学においては，これらの計測・通信にまつわるフィードバック制御器の欠点を補うにあたって，フィードフォワード制御器を併用する方法論が確立している．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．

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 1）の開発を目指す．この目的実現のために，前年度までの結果を踏まえて本調査の範囲では，(1)推論問題としての最適制御と変分リカレントニューラルネットワーク（VRNN）を応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理論の確立，(2)より汎用的な２つの方策の合成則の考案および解析，(3)シミュレーションおよび実機実験を通じた実証，の３項目について研究を進めた．結果として，導出した新しい強化学習則により，学習初期にはフィードバック方策を獲得し，時間経過とともにフィードフォワード方策へフィードバック方策の知識が転写されることをシミュレーション・実機実験で確認した．また，カメラによるロボットの位置計測に対してセンシング障害をエミュレートすることで，フィードバック方策のセンシング障害への脆弱性とフィードフォワード方策による補償の実現可能性を確認した．

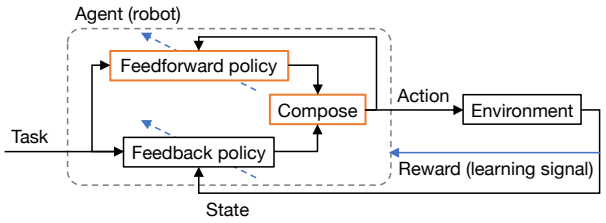


図 1 提案する強化学習の枠組み

## 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 軌道最適化問題の定式化

一般的な強化学習則はフィードバック方策を前提とした導出であるため，本研究には不適である．特に，フィードバック方策を内包する，前年度までの経験を踏まえると VRNN を用いた予測モデルを内包するような最適化問題を新たに導出する必要がある．このために，まず Sergey が近年提唱している概念である，制御器の最適化を推論問題として解く枠組みに則る．すなわち，現状あるいは将来全てが最適なのかどうかを示す確率二値変数を導入する．この変数は従来の強化学習における報酬あるいは価値関数により表現可能である．ただし，それらとは異なり確率変数であるため，ベイズ推定などに容易に組み込むことが可能となる．実際，本研究では適当な方策を事前分布としたときの最適あるいは非最適な事後分布を最適方策と非最適方策としてベイズ推定を用いて表現する．

この上で，最適方策と非最適方策が実際の環境（状態遷移確率）に作用した場合を仮定すると，最適軌道と非最適軌道の確率分布を得られる．ロボットはこの最適軌道に近づくようにしつつ，非最適軌道からは離れるような軌道を描くことが望ましい（図 2）．実際に描かれるであろう軌道を予測すべく，実際の状態遷移確率の予測モデルを導入する．この予測モデルにロボットの方策，すなわち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方策を作用させたときに得られるものが予測軌道の確率分布となる．

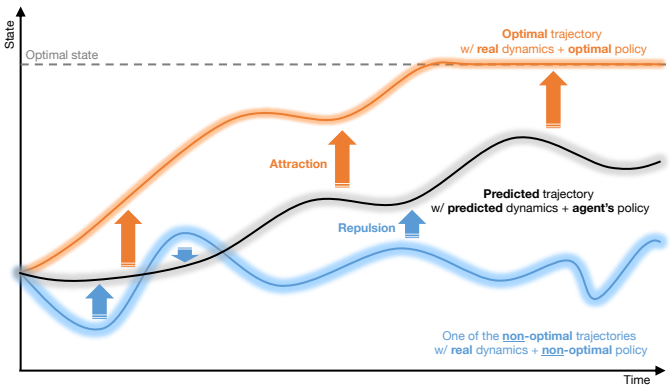


図 2 軌道間距離に関する最適化問題

確率分布間の乖離度を図るには Kullback-Leibler ダイバージェンスが広く使われるため，本研究でもそれに倣う．ただし，最適・非最適軌道の確率分布は実際にはブラックボックス関数となっているため，それらはモンテカルロ近似（その確率分布からのサンプリング値を用いた近似）しなければいけないことを念頭に置いて最適化問題を定式化した．その問題を方策勾配法に倣って解くと，従来の強化学習則，とりわけ Actor-Critic 法と似て非なるものが得られた．また，当初の目的通り，最適化用の勾配に予測モデルの対数尤度が含まれることを確認できた．

### 2-2 予測モデルとしての VRNN

前述の通りに予測モデルに対して VRNN を導入すれば，その内部で事前分布と事後分布間の乖離度を最小化しようとする正則化機能が得られる．ただし，既存の VRNN はダイナミクスを陽に含んでおらず，予測モデルとして機能しない．そこで，既存の VRNN を改良して，ダイナミクスを含み，かつフィードバック・フィードフォワード方策両方を持つ構造となるようにした（図 3）．この改良後でも従来同様に観測データの対数尤度を最大化する際の変分下界を解析的に得られることを確認した．これを 2-1 で得られた最適化問題（の勾配）に代入すると，方策がより良い方向へと改善されるときには２つの方策間の正則化が強化される挙動を示すようになった．なお，逆に悪い方向への更新時には正則化の符号が反転してしまい学習が不安定化する懸念があったため，予測モデル単体での最適化を同時にすることで常に正の正則化となるよう調整した．

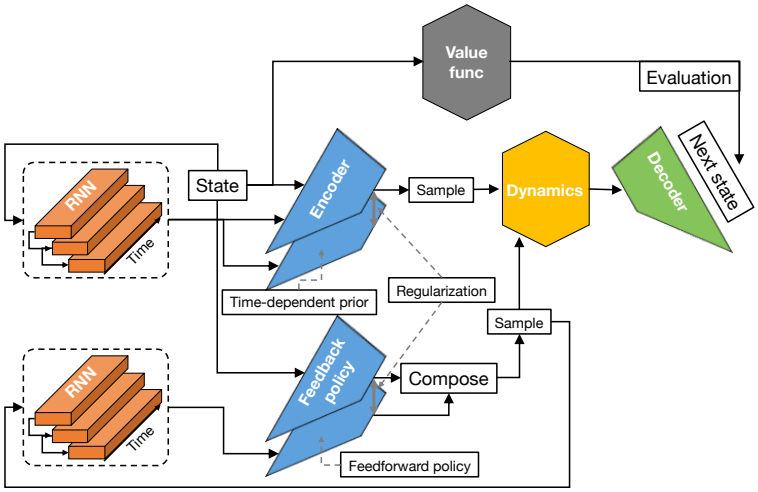


図 3 ネットワークの全体構成

### 2-3 混合分布による方策の合成

上記の導出はフィードバック・フィードフォワード方策を合成して１つの方策にする方法については言及していない．そこで，前年度の反省を踏まえて新しい合成則を検討した．今回は一般性を重視するものとし，その中で最も代表的な方法である混合分布を採用するものとした．こうすることで，２つの方策は任意の確率分布モデルで定義可能となり，かつ異なるモデルであっても支障がない利点がある．ただし，混合分布は複数の確率分布の中から１つを選ぶ上位のカテゴリカル分布が存在するモデルといえるため，このカテゴリカル分布のパラメータ，すなわち混合比の設計が必要になる．本研究ではこれをヒューリスティックな設計で与えることにした．考慮すべきな点は，i）確信度の高い方策を優先すべきである，ii）２つの方策が近いのであればどちらを選んでも結果は変わらず，離れているのであれば選択の重要性が高まる，という２点である．これらを満たす設計として，各方策の負のエントロピーを入力としたソフトマックス関数をベースに，その逆温度パラメータ（ランダム性の強さ）を２つの方策の平均の距離で設計するようにした．

また，混合分布として合成するにあたって，上記で得られていた最適化問題（の勾配）に代入・式展開と不要な項の除外を進めて解析していくと，最終的にはフィードバック方策とフィードフォワード方策間の Kullback-Leibler ダイバージェンスにフィードバック方策を重視する混合比の２乗が係数として残ることが判明した．これはすなわち，フィードバック方策の最適化が進み確信度が高くなる（エントロピーが小さくなる）と，正則化が強化されてフィードバック方策の知識がフィードフォワード方策へ転写されることを示唆している．また経験的にではあるが，学習初期のようにランダムな行動を繰り返していく内はフィードフォワード方策の最適化は進まずにフィードバック方策が安定して最適化できることがわかっている．すなわち，２つの方策を同時学習する枠組みであるにも関わらず，自然とフィードバックを先に獲得した後に徐々にフィードフォワード方策を獲得するプロセスが期待できる．これはフィードバック制御器の誤差信号を基にフィードフォワード制御器が最適化されていくフィードバック誤差学習と類似したプロセスといえ，それ自体が我々生物の学習構造を模倣して設計されたものであることから，得られた新しい強化学習則は生物と似た挙動をもたらすことを期待させる．

## 3 実証実験

### 3-1 倒立振子シミュレーション

まずはじめに開発した強化学習則の統計的な学習性能や挙動を確認することを狙い，動力学シミュレーションエンジンである Pybullet の OpenAI Gym ラッパーが提供する InvertedPendulum 環境での学習を試みた．この環境は観測状態が 5 次元で，ロボットの行動は土台の左右への加速度に相当する 1 次元であり，報酬はポールを直立状態に維持した分貰えるよう設定されている．前年度開発したオンライン強化学習手法を併用して学習した結果が図 4 である．ランダムシードを変えて計 30 試行実施したが，その内の 25 試行で学習に成功したことが窺える．一方で残りの 5 試行では，強化学習で誤った解に陥ってしまったときの挙動と合致している．とはいえ，80%以上の確率で学習に成功していることから，新しい強化学習則は十分に有用であるといえる．

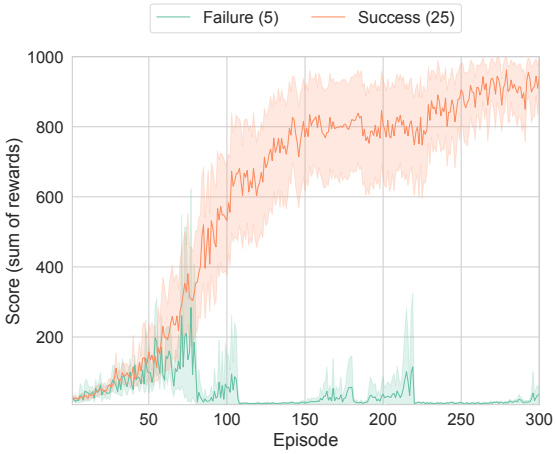


図 4 シミュレーションの学習結果

また，それぞれの学習過程における混合比（1 に近いほどフィードバック方策が優勢）は図 5 の通りである．見てわかるように，学習に成功した事例では，学習初期にフィードバック方策が優先的に使用されるようになったものの，学習の経過とともに徐々にフィードフォワード方策が活性化している．これは期待した通りの挙動といえる．一方で失敗した事例では，ほぼフィードバック方策のみが用いられているといえる．これは，フィードフォワード方策が持つリカレントニューラルネットワークの挙動の不安定さから非常に極端な値を出力するようになってしまい，フィードバック方策もそれに正則化項の影響で引き寄せられ続けた結果，フィードバック方策が終始優勢であるものの，目的を達成できずに終わってしまったと考えられる．


図 5 シミュレーション時の混合比

学習後のフィードバック・フィードフォワード方策それぞれの性能を検証するために，合成時の混合比を正弦波に従って与えたときの一例を図 6 に示す．フィードフォワード方策が主体となっている期間が 1 秒周期で訪れるにも関わらず，ロボットの行動は一貫して周期的に生成されている．すなわち，フィードバック方策の知識が正則化項を通じて正しく転写されたことで，フィードフォワード方策は過去の行動の履歴から最適な行動を生成できたといえる．

### 3-2 ヘビ型ロボットの開発

次に，実用性の検証に向けて，実機での実証実験を実施する．そのためのロボットとして，qb robotics 社製の可変剛性アクチュエータ qbmove を直列繋ぎにしたヘビ型ロボット（図 7）を構築した．全 8 軸をヨー軸回転となるように配置することで，揺動運動による移動を可能にする．移動をする上で，各アクチュエータ下部にキャスターを取り付けることで摩擦抵抗を減少させた．また，ロボットの先端に AR マーカを設置して，上部の定点カメラ（Stereo labs 社の ZED2）からロボットの路面上の位置座標を計測するようにした．これらのシステムは Robot Operating System (ROS)で連携して，50fps で状態観測値の受信と指令値（目標関節角度および目標剛性）の送信を可能にした．

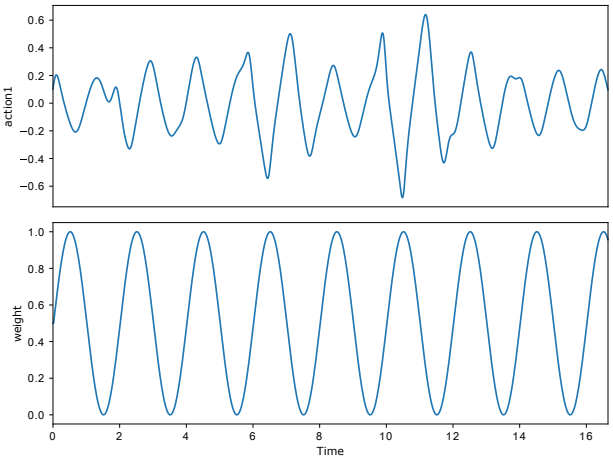


図 6 周期的な混合比に対する行動系列

制御器として，ヘビ型ロボットの移動は多くの場合が Central Pattern Generators (CPGs)が採用されることに注目した．具体的には，Cohen らの正弦波ベースの CPG を各アクチュエータの目標角度生成に用いた．残りの目標剛性に関しては強化学習により最適化するものとした．こうすることで，基本的な移動は容易に生成可能な一方で，詳細な挙動については剛性によって調整しなければならない．本実験では調整の目的としてヘビ型ロボットの直進移動を目指すものとした．

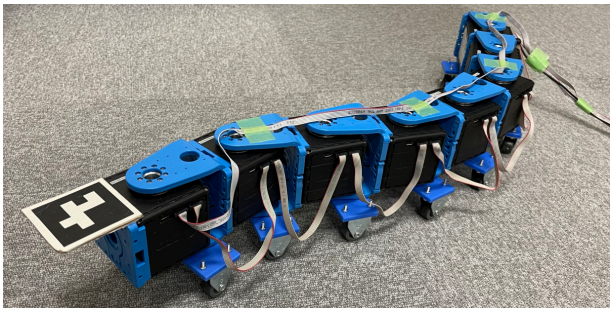

図 7 開発したヘビ型ロボット

まとめると，ロボットは各アクチュエータの関節情報と自身の位置座標で合計 34 次元の状態を観測し，各アクチュエータの目標剛性を個別に最適化，すなわち 8 次元の行動空間を有する．報酬は移動自体は CPGs によって自然と行われるため，横方向へのドリフト量に対するペナルティのみと定めた．

### 3-3 実証実験

図 8 に各エピソードにおける報酬の総和に関する学習曲線を，図 9 に学習前後でのロボットの挙動の一例を示す．学習前のランダムな目標剛性では直進移動できずに画面手前に曲がっていってしまったものの，学習が進むにつれて安定して概ね直進移動できるように改善されたことがわかる．なお，ロボットは揺動運動をしているため，報酬の総和はゼロになり得ないことに注意したい．また，学習時間の都合上，試行回数は 1 回のみである．混合比の学習曲線を図 10 にまとめると，シミュレーション結果と同様に，フィードバック方策が先に獲得された後にフィードフォワード方策へ知識が転写されていることが窺える．

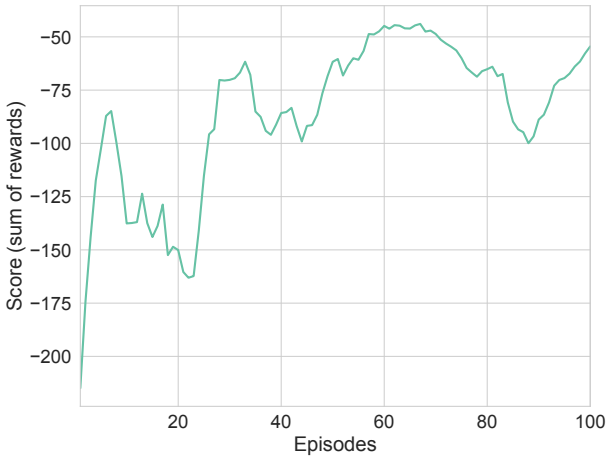


図 8 実証実験の学習結果

獲得したフィードバック・フィードフォワード方策それぞれの性能を確認するために，合成前に各方策をロボットに適用した場合の挙動の一例を図 11 上に示す．見てわかるように，ロボットはどちらの方策を用いてもほぼ同じ挙動を示していることがわかる．このことから，開発した

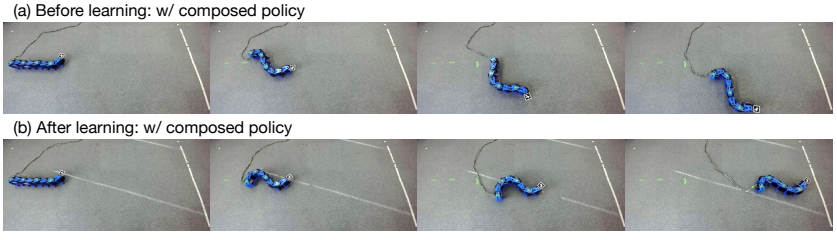


図 9 学習前後のロボットの挙動

強化学習則の正則化項の働きによって正しく知識の転写が実現できたといえる．また，ロボットの位置座標をカメラで取得していたことから，ここにセンシング障害が起きた場合をエミュレートした結果が図 11 下になる．画面左においてロボットは適切な位置座標が取得できないため，フィードバック方策は誤った状態に応じた誤った行動をして画面手前側にずれてしまい，障害から復旧した後にもフィードバックが間に合わず，あるいは制御可能な領域から外れてしまったためか，ずれたままゴールに向かっていることが読み取れる．このようにフィードバック方策はセンシング障害に脆弱であることが確認できた一方で，フィードフォワード方策のみを用いた場合にはセンシング障害の影響が方策に一切反映されないことから，図 11 上と同様に直進移動に成功している．このようにフィードフォワード方策は一般的な強化学習が最適化するフィードバック方策の欠点を補う能力を十分に秘めていることが確認でき，また，開発した新しい強化学習則であれば，これらを同時に学習可能であることを示せた．

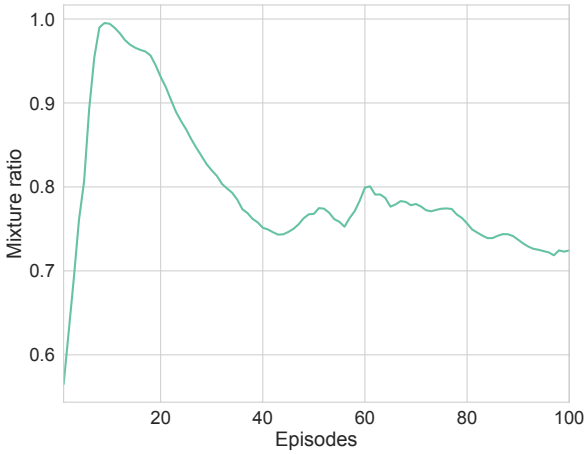


図 10 実証実験時の混合比

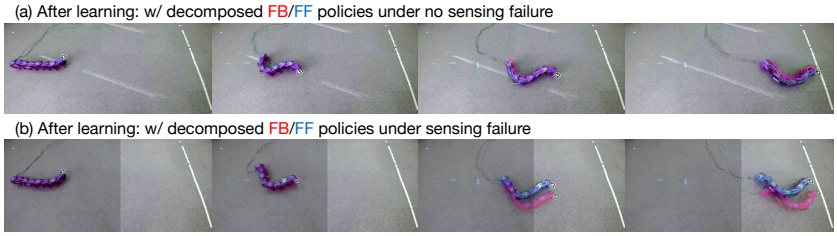


図 11 各方策を個別に適用した際のロボットの挙動

## 4 おわりに

本研究調査では，一般的な強化学習が最適化する方策がフィードバック制御器に相当することに着目し，その欠点となる通信・センシング障害への脆弱性を補うためのフィードフォワード方策を同時に学習できる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．この目的実現に向けて，まず Sergey が近年提唱している，推論問題として制御器の最適化を取り扱う枠組みに従って，最適性変数を陽に導入した．これを用いて最適方策および非最適方策をベイズ推定により定義した．これらが実際の環境に作用することで生成されるであろう最適軌道・非最適軌道に対して，強化学習エージェントが持つフィードバック・フィードフォワード方策を合成した方策が実際の環境をシミュレート可能な予測モデルに作用することで生成される予測軌道を漸近させる，あるいは乖離させることを目的とした最適化問題を導出した．これを解くと，従来の強化学習とは似て非なる学習則が得られるとともに，その内部に予測モデルの項が陽に含まれることを示した．そこで，この予測モデルとして VRNN を相互作用を陽に考慮するよう改良したモデルを採用することで，フィードバック・フィードフォワード方策間の正則化，すなわち知識の転写が自然と行われるようにした．また，２つの方策の合成則として，より汎用的な混合分布を採用することとし，このときの学習則を更に展開すると，フィードバック方策が優勢となっているときほど強くフィードフォワード方策への正則化が働くことを解析的に示した．一般的にフィードバック方策のほうが学習が容易であるため，提案する学習則ではフィードバック方策が学習初期に獲得され，その知識がフィードフォワード方策へと転写される，という我々生物と類似したプロセスが期待できることがわかった．

提案する強化学習則の実現可能性および挙動を確認するために，OpenAI Gym での倒立振子シミュレーションとヘビ型ロボットの直進移動スキル獲得を目的とした実証実験を実施した．結果として，どちらの環境においても，期待通りにフィードバック方策が学習初期に獲得されていき，その後でフィードフォワード方策が正則化の効力でフィードバック方策と類似した行動を出力するようになることを確認した．実際に，２つを合成せずに，あるいは指定した比率で強制的に合成してエージェントへ適用してみると，フィードフォワード方策はフィードバック方策とほぼ同じ挙動をもたらせることを確認した．この特性を活用することで，実機実験において，意図的にセンシング障害を発生させた場合には，フィードバック方策が破綻する中でも

フィードフォワード方策は正しく機能することを実験的に示し，その有用性を確認できた．

今回開発した技術は理論的に自然な形で２つの方策を同時に学習できる枠組みとなっているものの，その学習の安定性はまだ乏しいことも検証よりわかった．そのため，今後は学習の安定化（特にリカレントニューラルネットワークの安定化について）を推進するとともに，適当な応用事例を通じてフィードフォワード方策による補完性能の有用性を実証していきたい．その一例として，通信・センシング周期が遅いシステムにおいてもダイナミックな制御を実現するためのフィードフォワード方策の活用方法に関して，制御周期を擬似的に高めるような方法論について調査・研究を進めていく．

## 【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 Lukoˇseviˇcius 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.

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.

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.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. In arXiv preprint arXiv:1805.00909, 2018.

Taisuke Kobayashi. Adaptive and multiple time-scale eligibility traces for online deep

reinforcement learning. In arXiv preprint arXiv:2008.10040, 2020.
Avis H. Cohen, Philip J. Holmes, and Richard H. Rand. The nature of the coupling between segmental oscillators of the lamprey spinal generator for locomotion: A mathematical model. In Journal of mathematical biology, Vol. 13, No. 3, pp. 345-369, 1982.

## 〈発　表　資　料〉

| 題　　名 | 掲載誌・学会名等 | 発表年月 |
|---|---|---|
| フィードバック・フィードフォワード方策を内包する強化学習アルゴリズム | ロボティクスシンポジア | 2021 年 3 月 |
| Optimization Algorithm for Feedback and Feedforward Policies towards Robot Control Robust to Sensing Failures | arXiv (submitted for publication) | 2021 年 4 月 |
| | | |
| | | |