核融合科学学際連携センター　第1回 フュージョン・インフォマティクス勉強会

“Magnetic control of tokamak plasmas through deep reinforcement learning”を読む

草場 穫

核融合科学研究所　学際連携センター　特任助教

自己紹介

researchmap

↑統計数理研究所 (東京都　立川市)

経歴

2022年3月:　総合研究大学院大学　統計科学専攻
博士課程修了（統計科学）

~2024年8月:　統計科学研究所
ものづくりデータ科学研究センター　特任研究員

現職

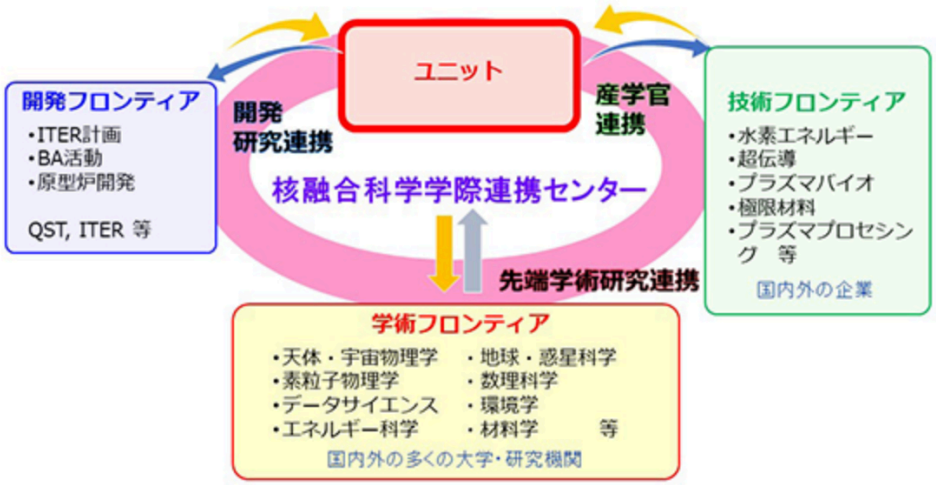
核融合科学研究所　学際連携センター　特任助教

専門

統計科学、機械学習、
マテリアルズインフォマティクス (MI)

当センターについて

核融合科学学際連携センターでは、核融合科学の学際化・学際連携、核融合開発研究との連携、産官学連携による核融合技術の社会実装に関する支援を行う。特に、先端的な学術研究領域との学際的な研究ネットワークの構築、およびオープンサイエンスの推進による核融合科学の学際化を進める。そのために、次の３部門を設置する。１）先端学術研究連携部門、２）開発研究連携部門、３）産学官連携部門



HP: https://www.nifs.ac.jp/about/fsicc.html

フュージョン・インフォマティクス勉強会について

目的: (1) データ科学を横串として、専門領域を横断した学際連携を深める。
(2) 勉強会を通した研究ネットワーク構築の支援。
(3) フュージョン・インフォマティクス (FI) の理解を深める。

テーマ: 核融合科学と情報科学の学際領域であるFIを会のテーマとする。

形式: 参加者が持ち回り（挙手制）でFIに関連する論文を紹介する輪読形式。聴講のみの参加も可。

対象論文の範囲: 核融合に関連する対象に何らかの形で情報科学の技術が応用されているもの。核融合に関連する対象としては、プラズマ物理学以外に材料科学、ロボット工学等も含む。学際的に高い汎用性が見込まれるものであれば、純粋に情報科学の論文でも良い (どう核融合科学に応用できるかの議論を加えるとなお良い)。複数本同時に紹介しても良い。

フュージョン・インフォマティクス勉強会について

発表形式: 現地とzoomのハイブリッド開催

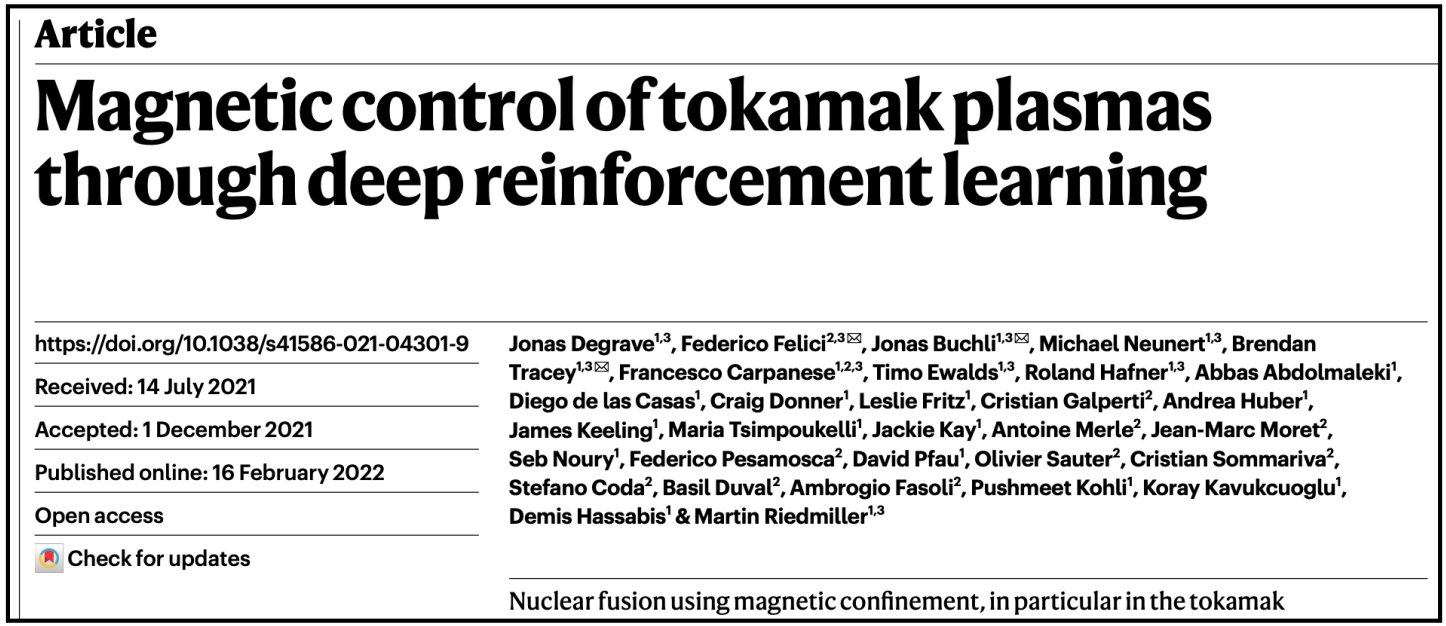
スケジュール: 発表時間は45~75分程度を想定。2週間に1回程度の開催を想定。開催日時は随時調整を取って柔軟に対応する。

特徴: 双方向性を重視し、議論の時間を長く取る。発表中に自由に発言しても良い。

webページ: https://projects.nifs.ac.jp/fi-workshop/ を公開した。ページ内には常設の参加申し込みフォームが設置されているので、所内外の方にこの勉強会を紹介する際には、是非このwebページを参照ください。

メーリングリストとTeam: 勉強会用のメーリングリスト（fi-workshop@nifs.ac.jp）とTeamを作成した。参加希望でまだ連絡していない方は私（kusaba.minoru@nifs.ac.jp）まで連絡ください。

今回紹介する論文



Article

Magnetic control of tokamak plasmas through deep reinforcement learning

https://doi.org/10.1038/s41586-021-04301-9

Received: 14 July 2021

Accepted: 1 December 2021

Published online: 16 February 2022

Open access

Check for updates

Jonas Degrave[1,3], Federico Felici[2,3]✉, Jonas Buchli[1,3]✉, Michael Neunert[1,3], Brendan Tracey[1,3]✉, Francesco Carpanese[1,2,3], Timo Ewalds[1,3], Roland Hafner[1,3], Abbas Abdolmaleki[1], Diego de las Casas[1], Craig Donner[1], Leslie Fritz[1], Cristian Galperti[2], Andrea Huber[1], James Keeling[1], Maria Tsimpoukelli[1], Jackie Kay[1], Antoine Merle[2], Jean-Marc Moret[2], Seb Noury[1], Federico Pesamosca[2], David Pfau[1], Olivier Sauter[2], Cristian Sommariva[2], Stefano Coda[2], Basil Duval[2], Ambrogio Fasoli[2], Pushmeet Kohli[1], Koray Kavukcuoglu[1], Demis Hassabis[1] & Martin Riedmiller[1,3]

Nuclear fusion using magnetic confinement, in particular in the tokamak

Degrave, Jonas, et al. "Magnetic control of tokamak plasmas through deep reinforcement learning." Nature 602.7897 (2022): 414-419. (https://doi.org/10.1038/s41586-021-04301-9)

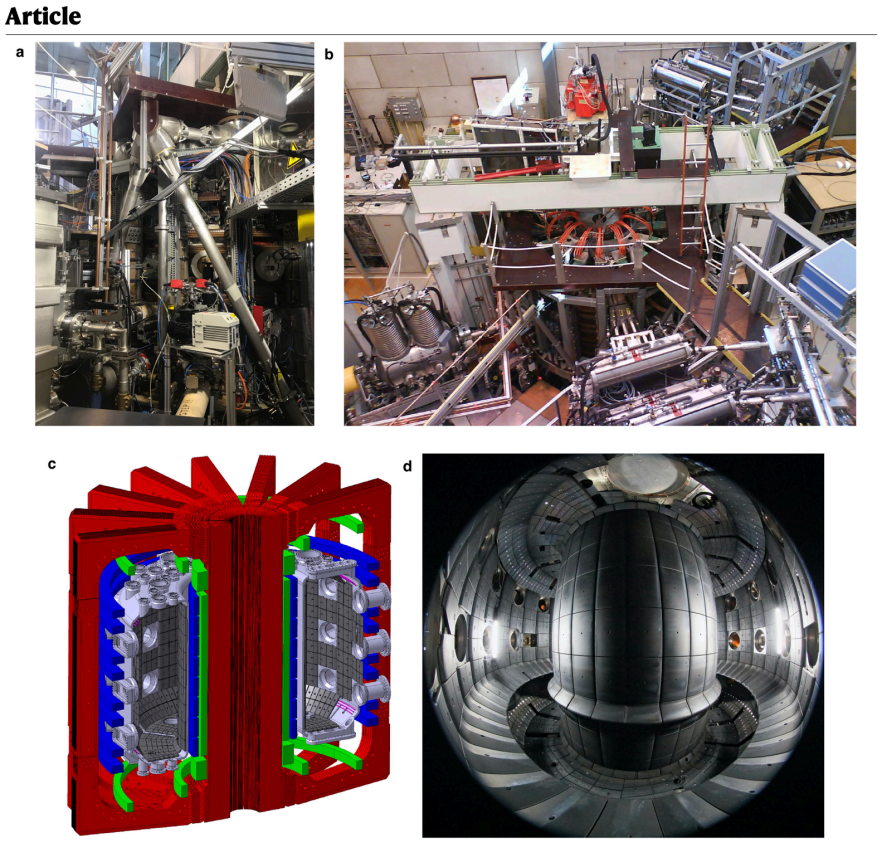
・DeepMindとSPCによる2022年の論文 (被引用数806/2024-10-28 Google Scholar)

・トカマク型核融合炉のプラズマ磁場制御において、従来の手法に代わり、深層強化学習（RL）を活用して非線形の制御器設計を行う新しいアプローチを提案している。

→「強化学習におけるSim2Real」を実際の核融合炉上で高度に実現した。

概要

Tokamak à configuration variable (TCV)



Extended Data Fig. 1 | Pictures and illustration of the TCV. a, b Photographs showing the part of the TCV inside the bioshield. **c** CAD drawing of the vessel and coils of the TCV. **d** View inside the TCV (Alain Herzog/EPFL), showing the limiter tiling, baffles and central column.

[論文より引用/Extended Data Fig.1]

1. シミュレーションを使って、強化学習モデルを学習した。

2. 学習済みモデルをそのまま、制御方策として実際の核融合炉に実装した (zero-shot)。

3. その結果提案手法は、様々な形状のプラズマを制御することに成功した。

4. この方法により、新しいプラズマ形状の試験と開発サイクルが短縮され、核融合研究の加速が期待される。

実験結果

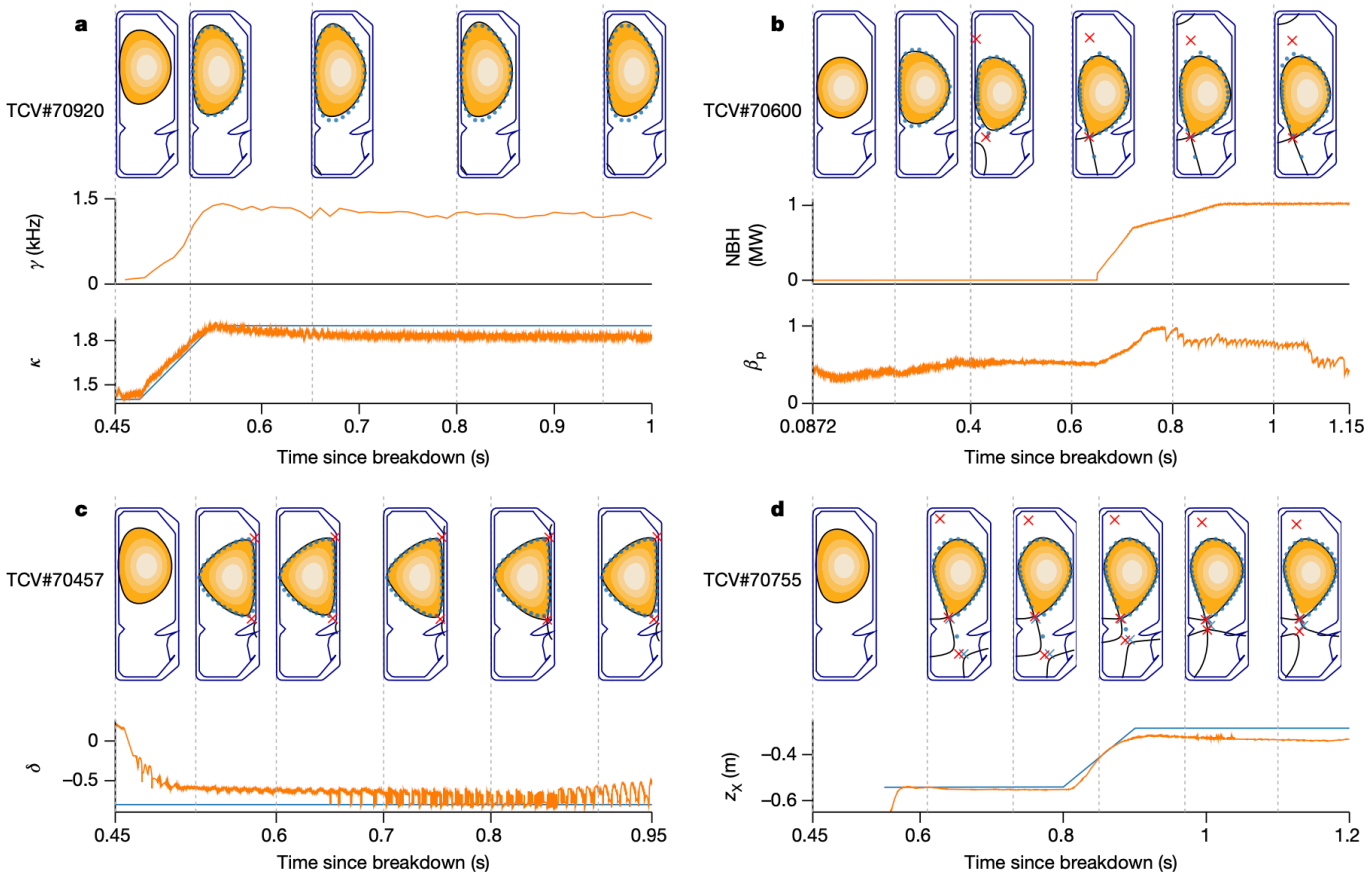


Fig. 3 | Control demonstrations. Control demonstrations obtained during TCV experiments. Target shape points with 2 cm radius (blue circles), compared with the equilibrium reconstruction plasma boundary (black continuous line). In all figures, the first time slice shows the handover condition. **a**, Elongation of 1.9 with vertical instability growth rate of 1.4 kHz. **b**, Approximate ITER-proposed shape with neutral beam heating (NBH) entering H-mode. **c**, Diverted negative triangularity of −0.8. **d**, Snowflake configuration with a time-varying control of the bottom X-point, where the target X-points are marked in blue. Extended traces for these shots can be found in Extended Data Fig. 2.

[論文より引用/Fig.3]

提案手法は、ITER（国際熱核融合実験炉）における標準的な形状や、先進的な「負の三角形形状」や「スノーフレーク」形状の制御を成功させており、複数のプラズマ「ドロップレット」を同時に安定化するデモも行われた。

実験結果

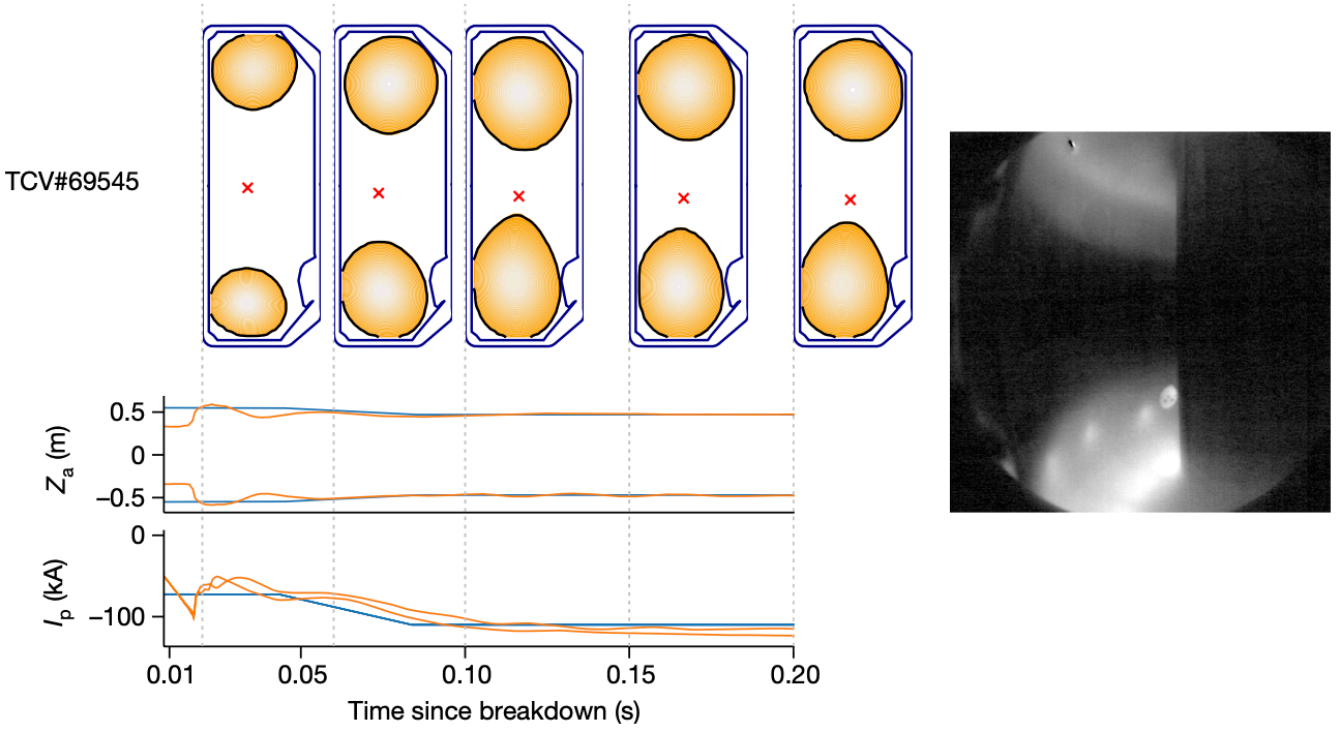
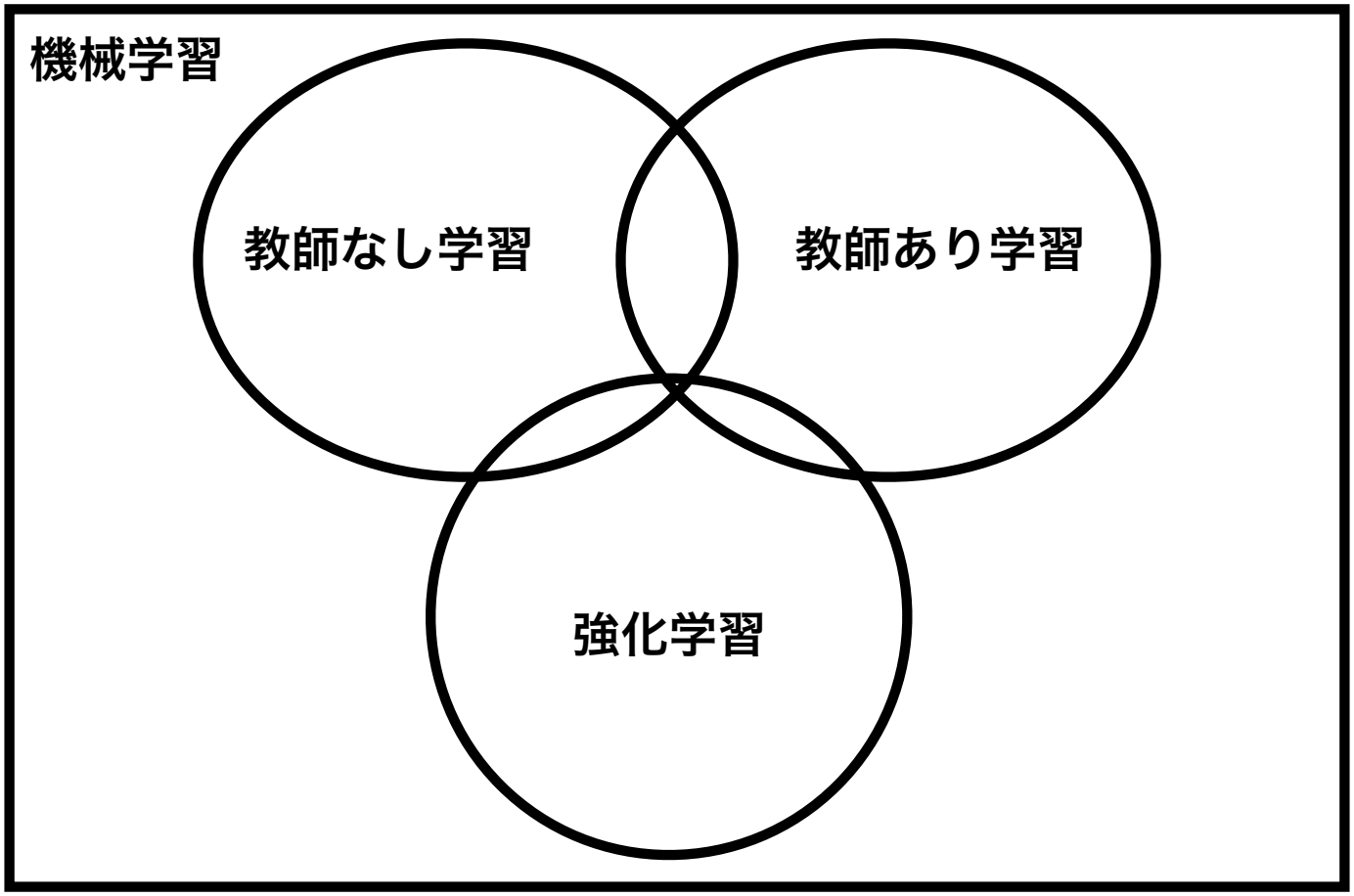


Fig. 4 | Droplets. Demonstration of sustained control of two independent droplets on TCV for the entire 200-ms control window. Left, control of I_p for each independent lobe up to the same target value. Right, a picture in which the two droplets are visible, taken from a camera looking into the vessel at $t = 0.55$.

[論文より引用/Fig.4]

提案手法は、ITER（国際熱核融合実験炉）における標準的な形状や、先進的な「負の三角形形状」や「スノーフレーク」形状の制御を成功させており、複数のプラズマ「ドロップレット」を同時に安定化するデモも行われた。

強化学習とは?



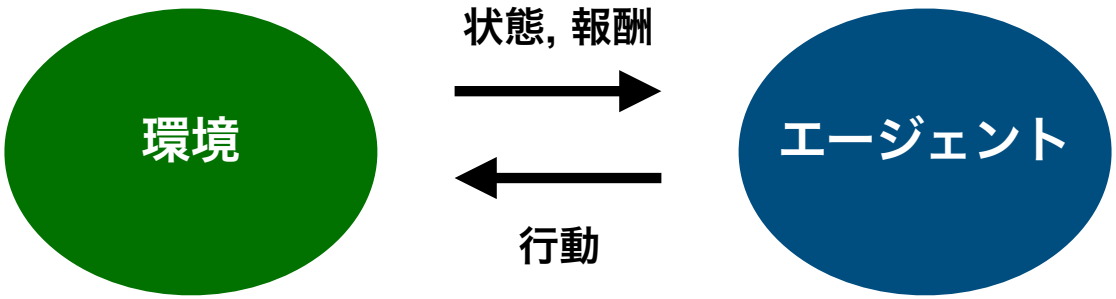
- **強化学習は機械学習における大きな学習の枠組みの内の一つ**

教師なし学習: 入力データのみを使い、データの構造や関連性を学習する

教師あり学習: 入力データと出力データの関係性を学習する

強化学習: エージェントが環境との相互作用を通じて、報酬を最大化する行動方針を学習する

強化学習とは?



・エージェントは行動方針に基づいて行動する。

・一般的に行動方針は、将来得られる報酬の合計（累積報酬）を最大化するように学習される。

核融合炉制御問題はなぜ強化学習の問題と見做せるか?

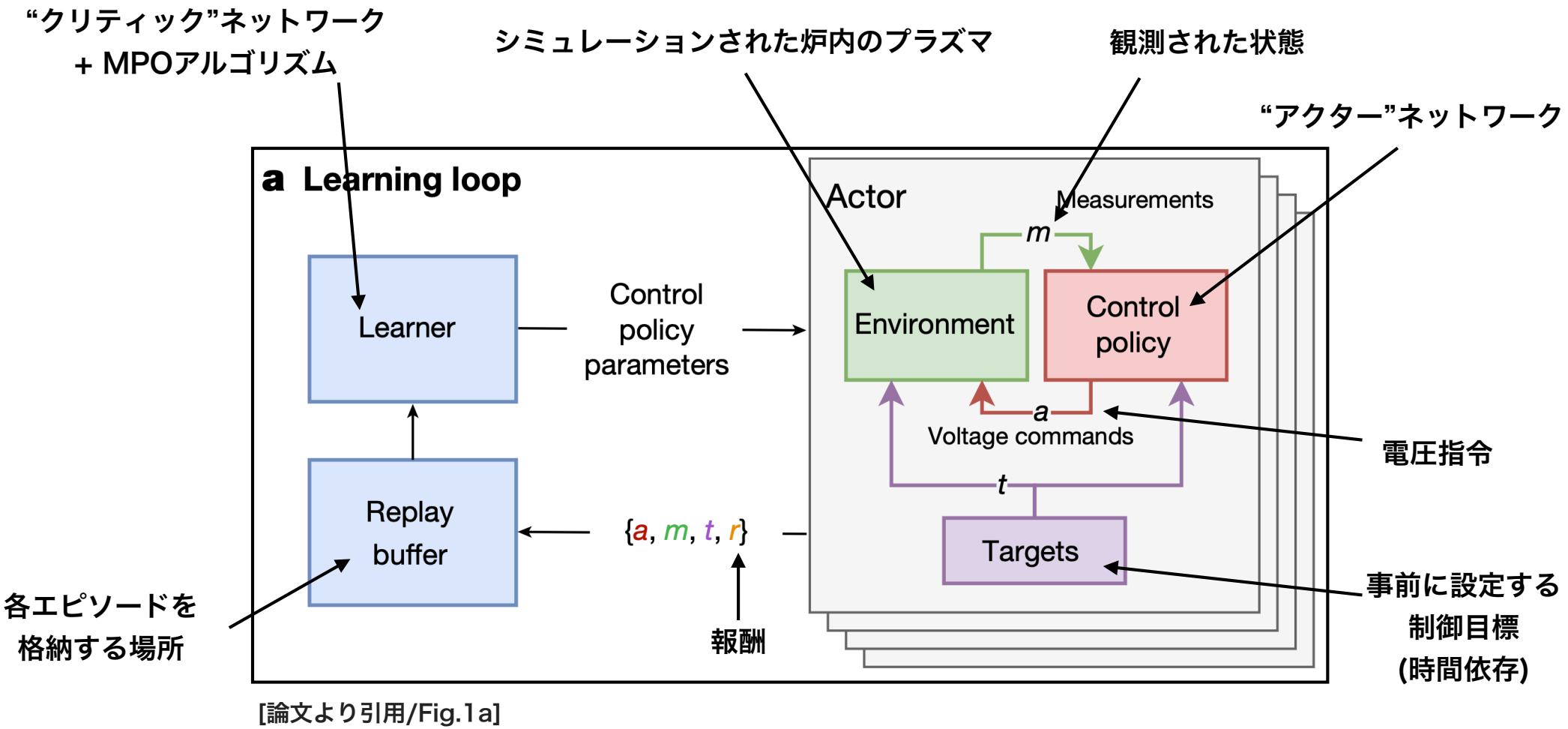
・環境は核融合炉内のプラズマの状態であり、これはセンサーを通して観測される (m)

・制御の目標 (t) を事前に定義しておけば、報酬 (r) を算出できる

・エージェントは制御器に相当し、各制御用コイルに送る電圧指令 (a) を通して炉内のプラズマを制御する

・プラズマ制御は最適な制御器を設計することを目的とし、これは強化学習の目的と一致する

本論文の強化学習手法概要



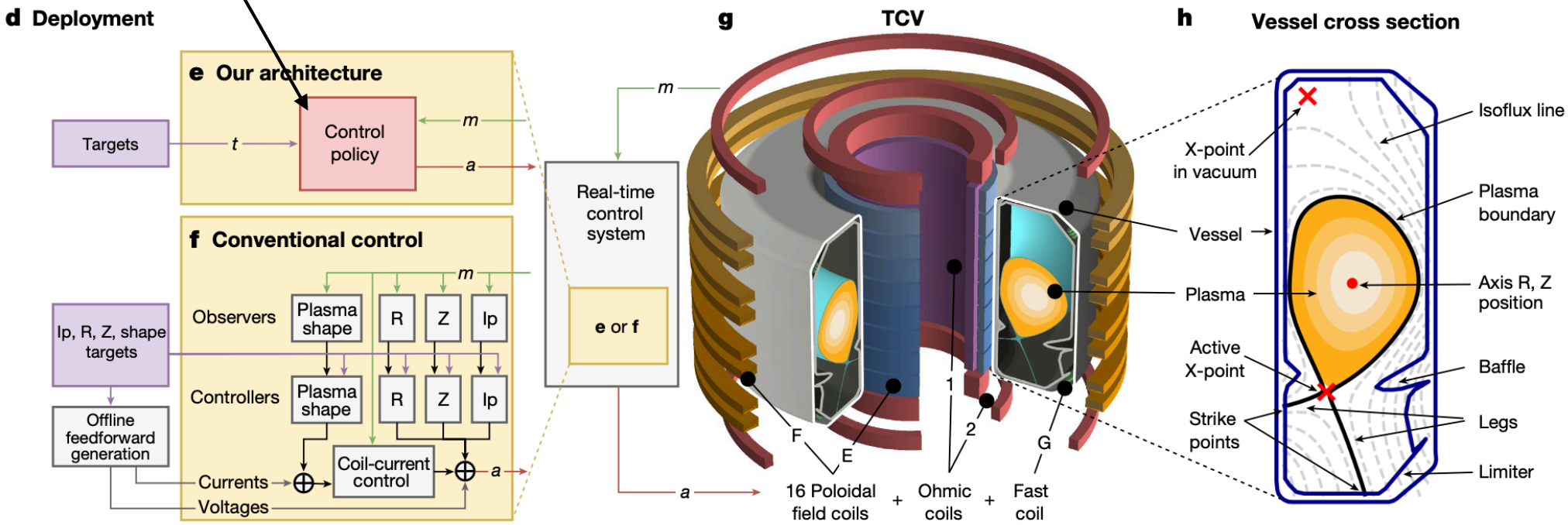
[論文より引用/Fig.1a]

・上記をループさせることにより、強化学習モデルは学習される。

・強化学習法として「アクター・クリティック法」が採用されている。

本論文の強化学習手法概要

“アクター”ネットワーク (学習済み)

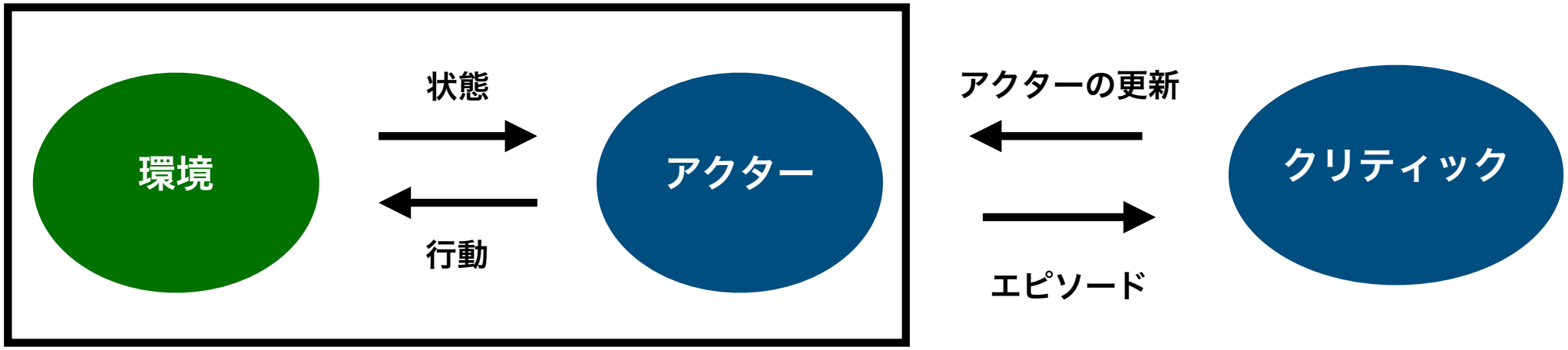


[論文より引用/Fig.1d]

- 学習済み強化学習モデルは上記のようにTCVに搭載される。
- 実用する時点では、アクターネットワークのみ使う

アクター・クリティック法とは?

・強化学習法の一種であり、方策と価値関数を別々に学習するオンポリシーで、特に連続行動空間や大規模空間で有効とされる。→強化学習として他にはQ学習等がある



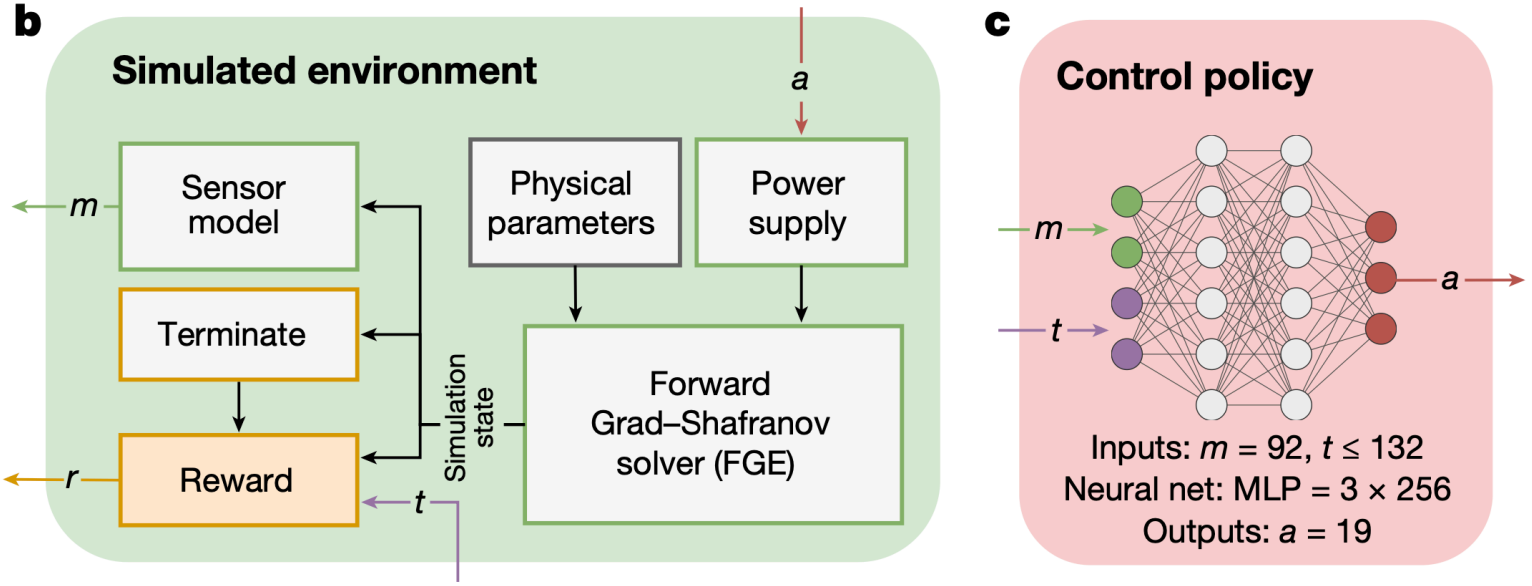
・”アクター”ネットワークと”クリティック”ネットワークから構成される。

・”アクター”は状態を入力とし、行動を出力する方策であり、制御器に相当する。

・”クリティック”は状態と行動のペアに対して、累積報酬の期待値 (Q値) を出力する。訓練時のみ使われる、”制御器の教官”に相当する。

→クリティックとして大型のネットワーク、アクターとして小型のネットワークを使えば、「高度な非線形性を捉える」ことと「実践における高速な制御」を両立できることも、この手法が選ばれた大きな理由である (制御器は超高速に応答する必要があるため)。

具体的な学習手順



[論文より引用/Fig.1b, c]

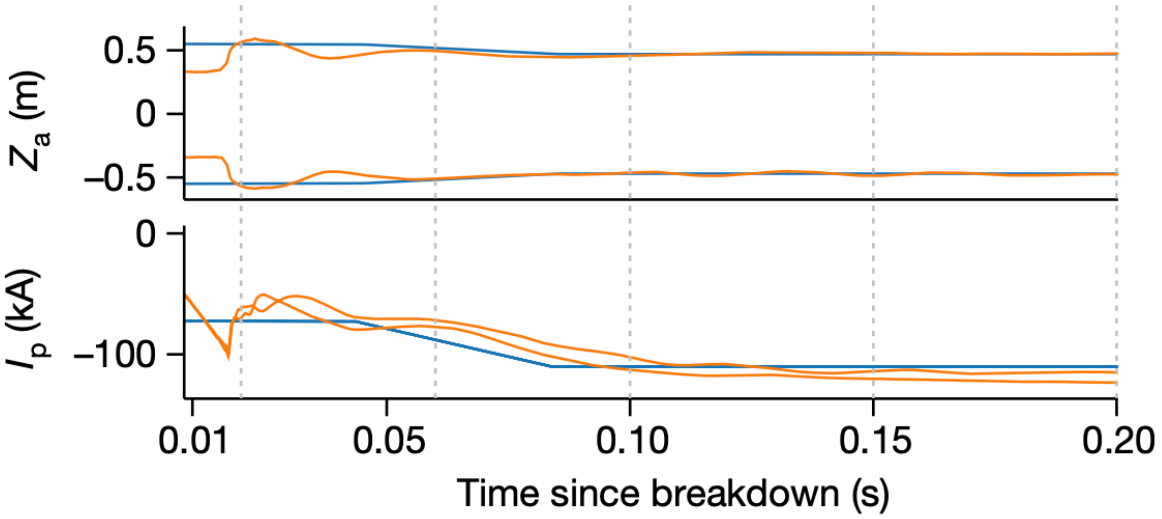
*行動aはトカマクの19個の制御コイルに対する電圧支持に対応する

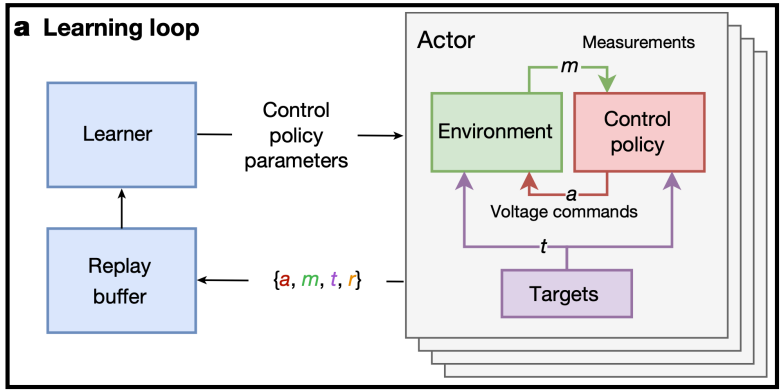
・アクターネットワークとしてc図、クリティックとしてLSTM (後述) が使われている。

・シミュレーションは上記のように、プラズマの形状と電流の進化を物理的にリアルに再現するように計算された。

・シミュレーションの1ショットが0.2秒 (droplet), 0.5秒 (ED fig.2 a.c), または1秒 (その他) と設定され、状態mの観測と行動aの出力は0.1ミリ秒ごとに行われた。この制御データー本分をエピソードという。

具体的な学習手順





[論文より引用/Fig.1aとFig.4の一部]

・状態mはセンサーによって観測された、タイムステップごとの、プラズマの電流 (Ip)、プラズマの位置 (R, Z)、X点（磁気的に中立な点）の位置等からなる。

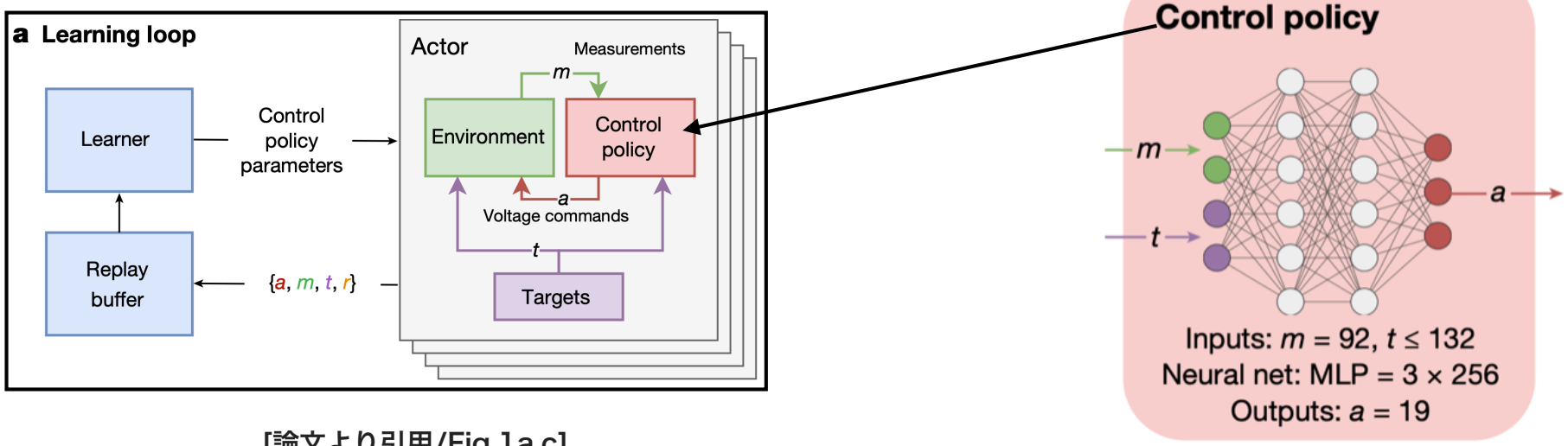
・制御目標 t は上記物理量の理想的な時間発展として、事前に与えられる。

・報酬rはタイムステップごとの、上記物理量の観測値と制御目標の誤差に基づいて計算される。物理量ごとに算出された報酬は重み付き線型結合により、総報酬としてスカラー値に変換される。

・１エピソードは、0.1msごとに測定された定区間の状態m, 行動a, 報酬r, 制御目標tの時系列データである。

具体的な学習手順

アクターネットワークと同じもの



[論文より引用/Fig.1a,c]

1. 適当なアクターと複数の制御目標・初期状態から、シミュレーションを回し多数のエピソードをReplay bufferに記録する。

2. Replay bufferからエピソードをランダムに選択し、クリティックネットワークを学習する

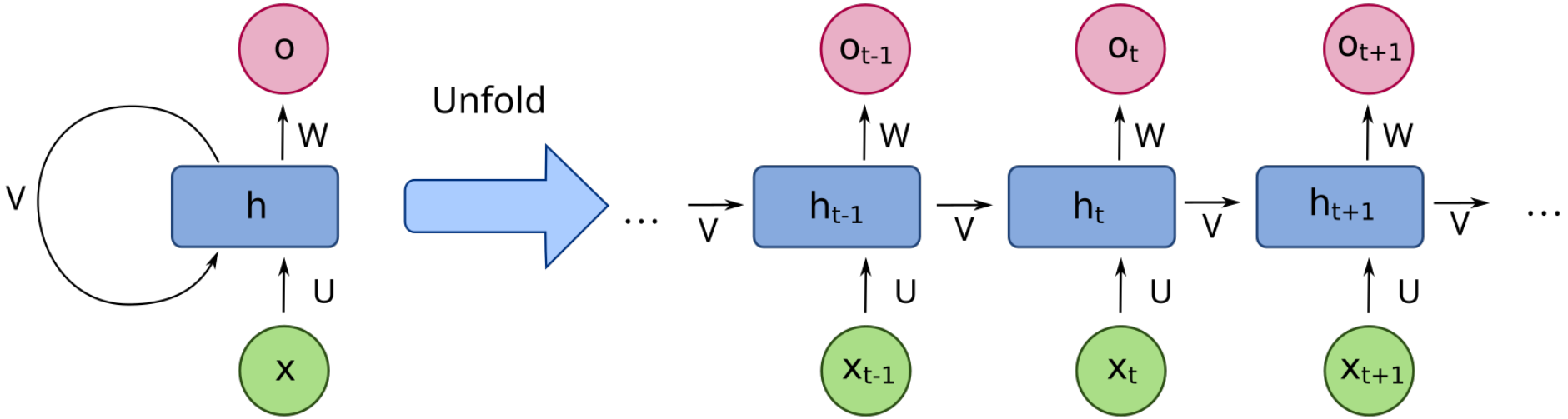
3. 学習されたクリティックを使ってMPOアルゴリズム（後述）によりアクターを更新する

4. 1~3を繰り返す。

クリティックネットワークの詳細

・クリティックネットワークは再帰型ニューラルネットワークモデル (RNN) の一種である、Long short-term memory モデル (LSTM) によってモデル化された。

RNNの基本構造

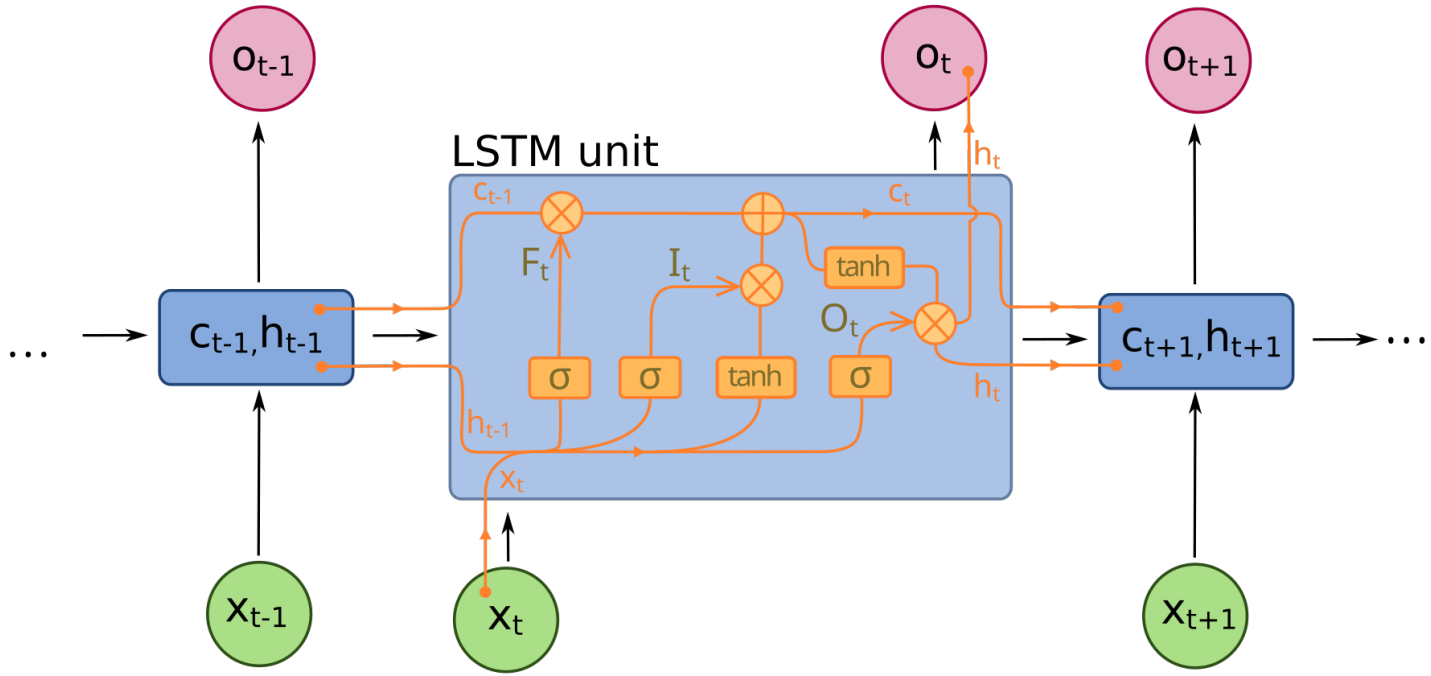


[Wikipediaより引用/fdeloche CC 4.0]

クリティックネットワークの詳細

・クリティックネットワークは再帰型ニューラルネットワークモデル (RNN) の一種である、Long short-term memory モデル (LSTM) によってモデル化された。

LSTMとは?　→基本的な構造はRNNと同じだが、中間層を以下のLSTM unitに置き換えたもの



[Wikipediaより引用/fdeloche CC 4.0]

忘却ゲート(Ft)、入力ゲート(It)、出力ゲート(Ot)、記憶セル(Ct)からなり、従来のRNNと比べて、長期依存関係や複雑な依存関係を持つ系列データの学習に向いている。

補足

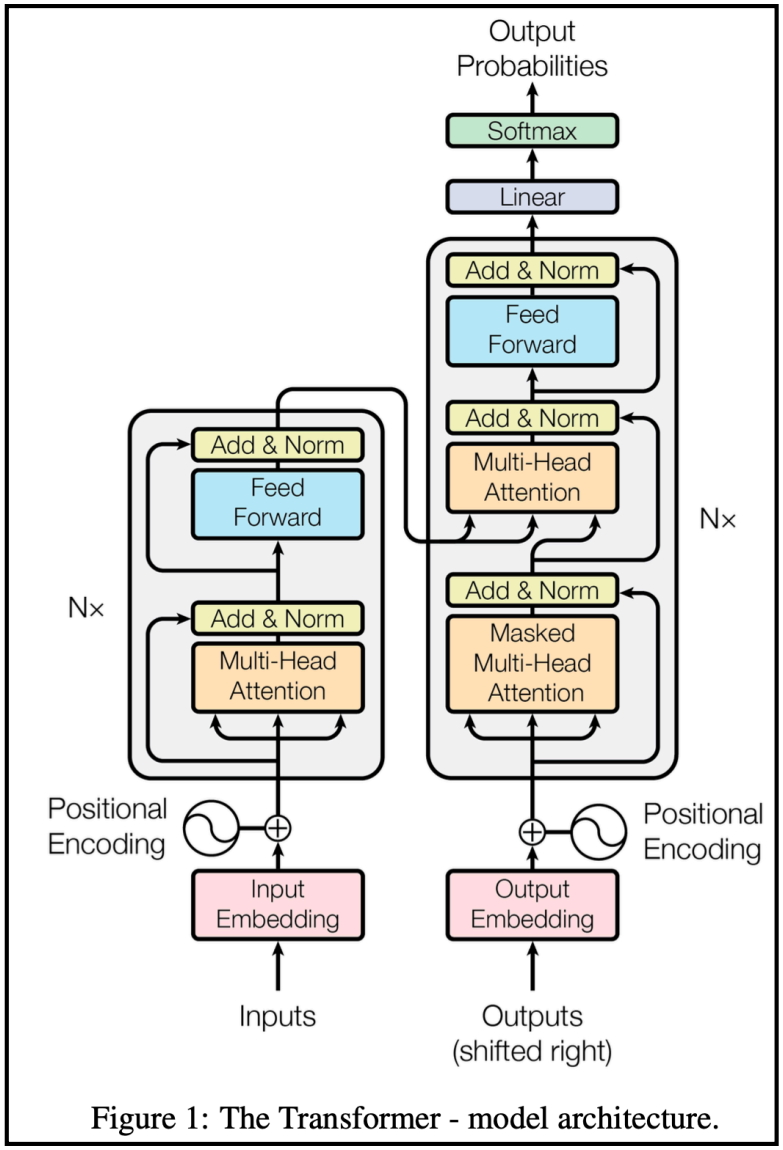


Figure 1: The Transformer - model architecture.

・LSTMは自然言語処理において長らく支配的な地位を保っていたが、近年は左図のTransformerというモデルに取って変わられ、あまり使われなくなった。

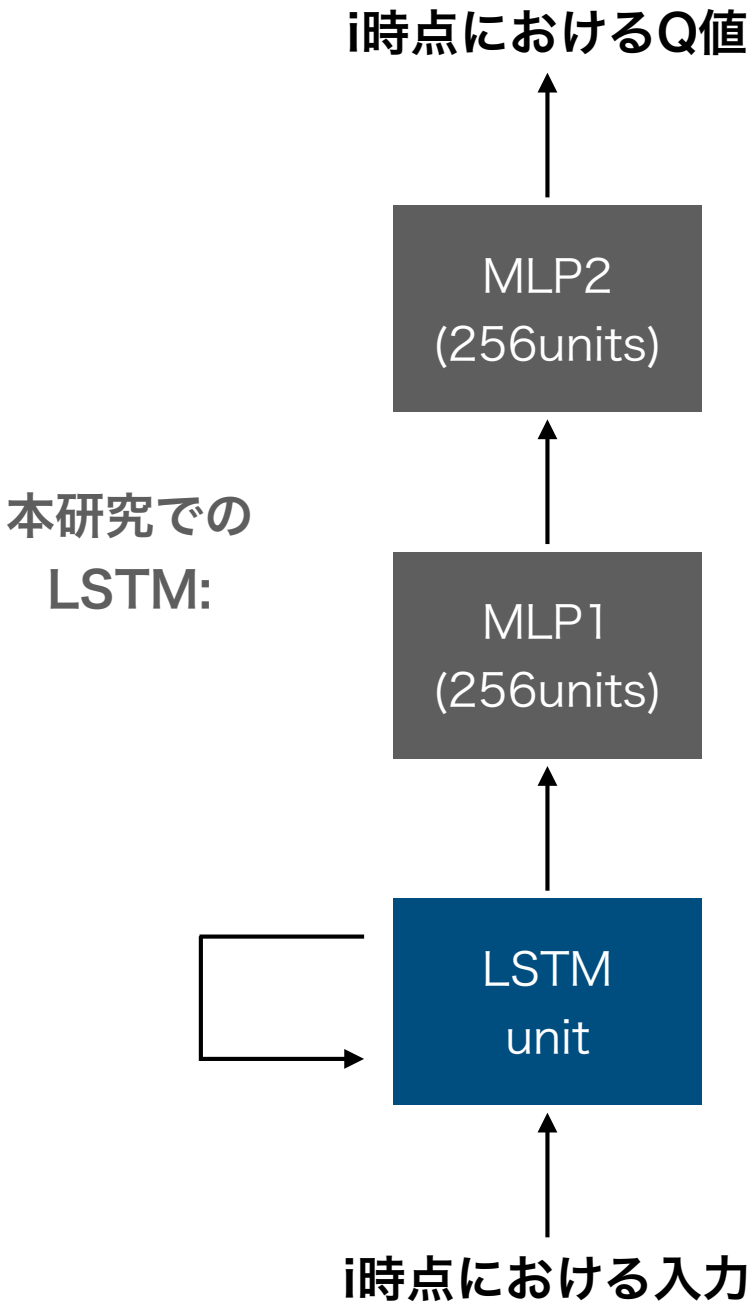
・TransformerはAttentionのみを用いたアーキテクチャであり、長いシーケンスに対する依存関係を効率的に処理でき、高い並列化と計算効率を持つ。

・ChatGPTもこのアーキテクチャに従うモデルである

・この研究でも、LSTMの代わりにTransformerを使った方が精度や計算効率が上がるかもしれない。

[Vaswani, A. "Attention is all you need." Advances in Neural Information Processing Systems (2017). のFig1より引用]

クリティックネットワークの詳細



・入力はi時点における、観測された状態m, 電圧指令a, 制御目標tからなる。

・i時点におけるQ値はスカラー値であり、
i, i+1, i+2, …, 終了までの累積報酬の期待値を示す。

・各時点における**Q値とターゲットQ値の誤差の総和が最小**になるようにモデルは学習される。

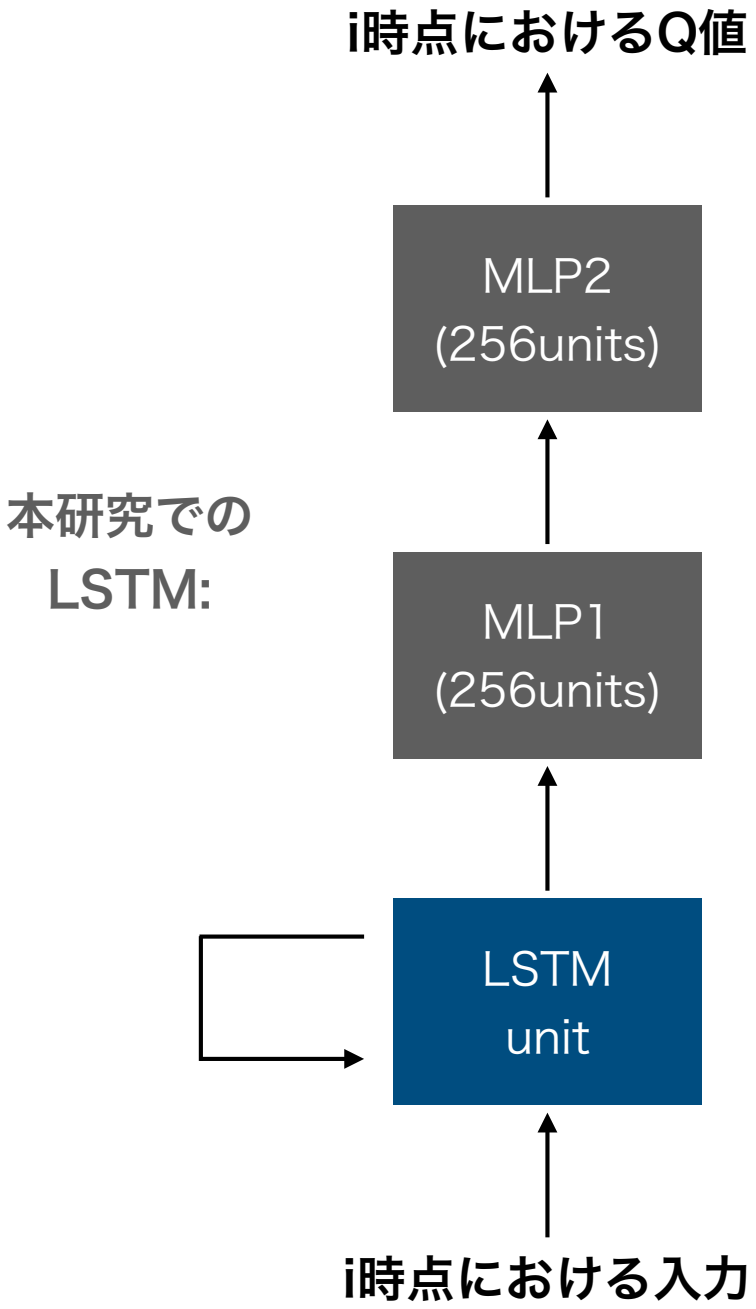
ターゲットQ値は以下のように計算される

$$Q_{target}(i) = r(i) + \gamma r(i+1) + \gamma^2 r(i+2) \cdots$$

i時点での総報酬

割引係数 ($\gamma = 0.99$と設定された)

クリティックネットワークの詳細



・Replay bufferからランダムに選択されたエピソードからクリティックを学習する。

・ターゲットQ値と誤差の小さいQ値が出力されるまで訓練する。

・その結果、状態mと行動aのペアに対して、精度の高い累積報酬の期待値が出力される関数が手に入る。

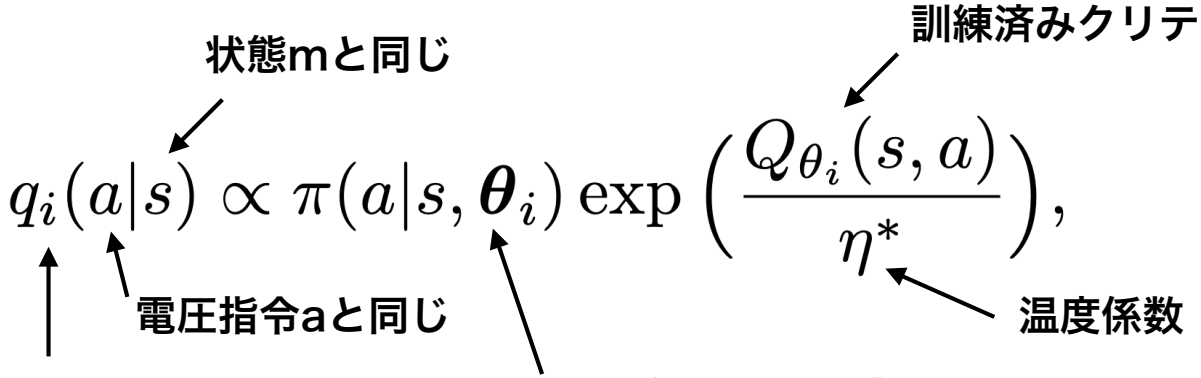
アクターネットワークの更新 (MPOアルゴリズム)

・次に、訓練済みクリティックネットワークを使って、アクターネットワークのパラメータを更新する。

・本論文では、この更新にMaximum a Posteriori Policy Optimisation (MPO)というアルゴリズムを採用している。

MPOアルゴリズムでは以下のEステップとMステップを繰り返し、アクターを更新する

・Eステップ



方針: 訓練済みクリティックのQ値出力を元に、理想の方策を作る

*方策は状態が与えられた時の行動の分布として表される

(続く)

[Abdolmaleki, Abbas, et al. "Maximum a posteriori policy optimisation." arXiv preprint arXiv:1806.06920 (2018).より上記の数式を引用]

アクターネットワークの更新 (MPOアルゴリズム)

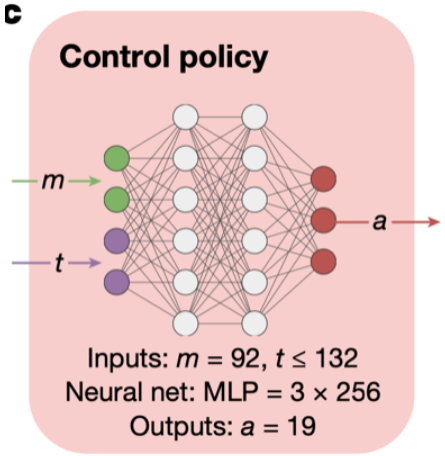
・Mステップ

状態の頻度　理想の方策で期待値　更新する方策 (アクター)

$$\max_{\pi} \mathbb{E}_{\mu_q(s)} \left[\mathbb{E}_{q(a|s)} \left[\log \pi(a|s, \boldsymbol{\theta}) \right] \right]$$

$$s.t.\ \mathbb{E}_{\mu_q(s)} \left[\mathrm{KL}(\pi(a|s, \boldsymbol{\theta}_i), \pi(a|s, \boldsymbol{\theta})) \right] < \epsilon.$$

確率分布の距離　更新前の方策　更新する方策 (アクター)



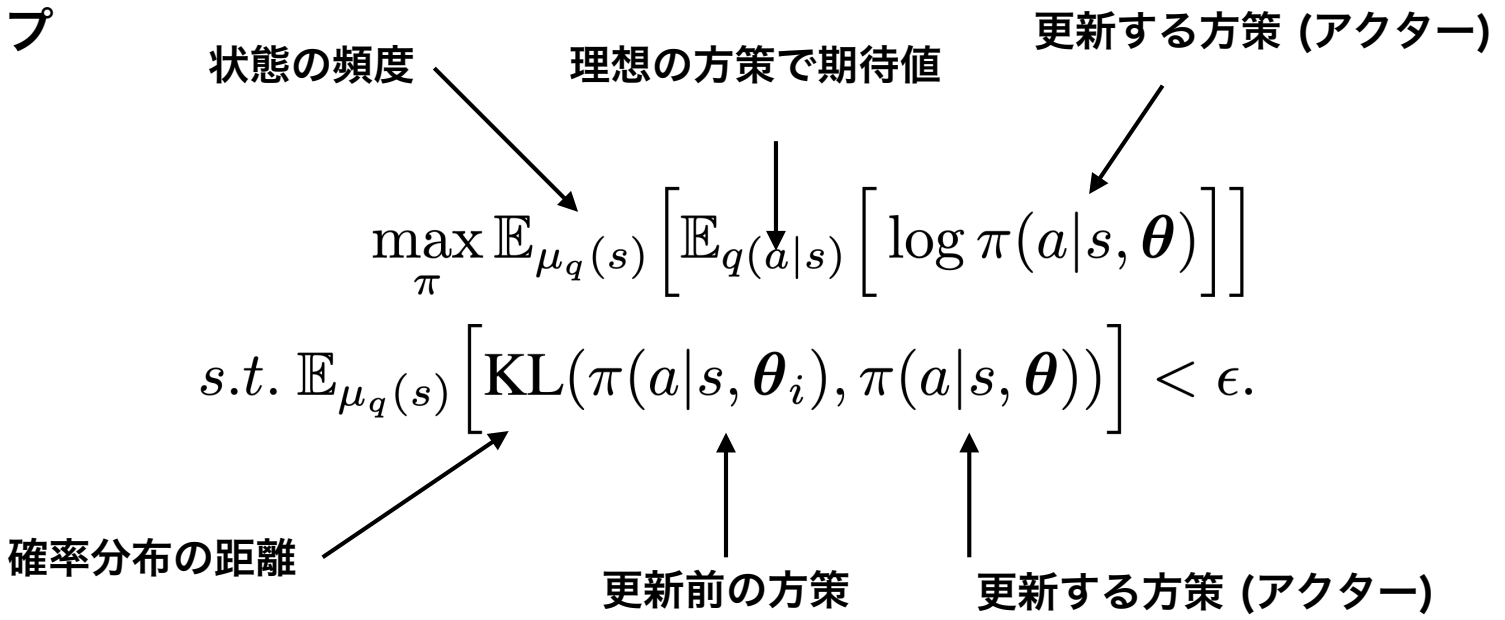
アクターはaの各要素の平均と標準偏差
を出力するネットワークであり、これにより
19次元のガウス分布が定義される。
これが $\pi(a|s)$ に相当する。

→アクターは訓練時のみ確率分布として扱われる。
実用時点では、ガウス分布の平均のみ使われる。

[Abdolmaleki, Abbas, et al. "Maximum a posteriori policy optimisation." arXiv preprint arXiv:1806.06920 (2018).より上記の数式を引用]

アクターネットワークの更新 (MPOアルゴリズム)

・Mステップ



方針: 理想の方策にできるだけ近づくように方策を更新する。

$\mathbb{E}_{q(a|s)} \left[\log \pi(a|s, \boldsymbol{\theta}) \right]$ は$\pi(a \mid s)$が$q(a \mid s)$に一致するときに最大となる

KLダイバージェンスの制約により、方策が1回の反復で急激に変化しないようにする。

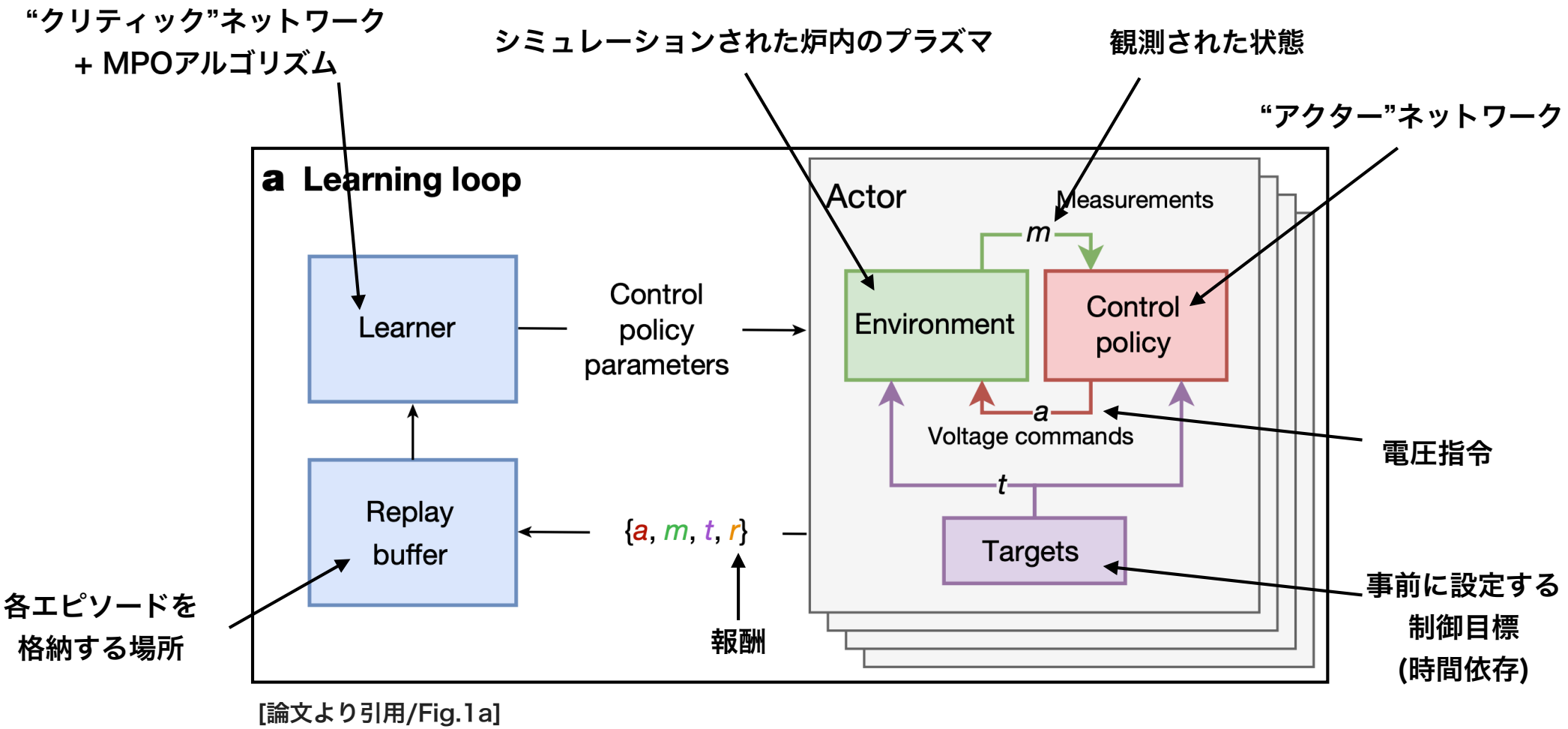
[Abdolmaleki, Abbas, et al. "Maximum a posteriori policy optimisation." arXiv preprint arXiv:1806.06920 (2018).より上記の数式を引用]

補足・MPOアルゴリズムを採用した理由

The RL algorithm uses the collected simulator data to find a near-optimal policy with respect to the specified reward function. The data rate of our simulator is markedly slower than that of a typical RL environment due to the computational requirements of evolving the plasma state. We overcome the paucity of data by optimizing the policy using maximum a posteriori policy optimization (MPO)[23], an actor-critic algorithm. MPO supports data collection across distributed parallel streams and learns in a data-efficient way. We additionally

[論文より引用/416page]

強化学習手法全体を再訪

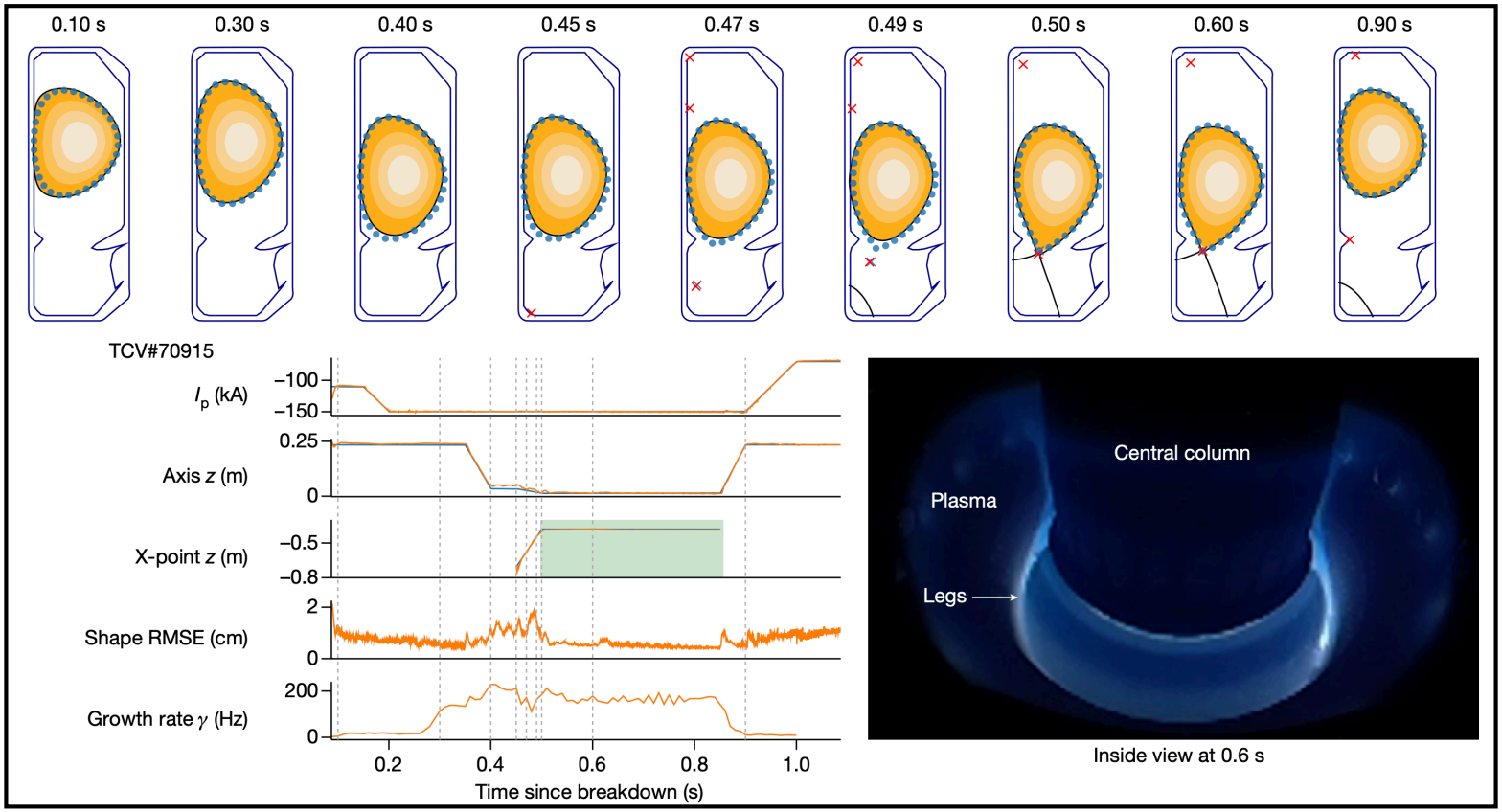


[論文より引用/Fig.1a]

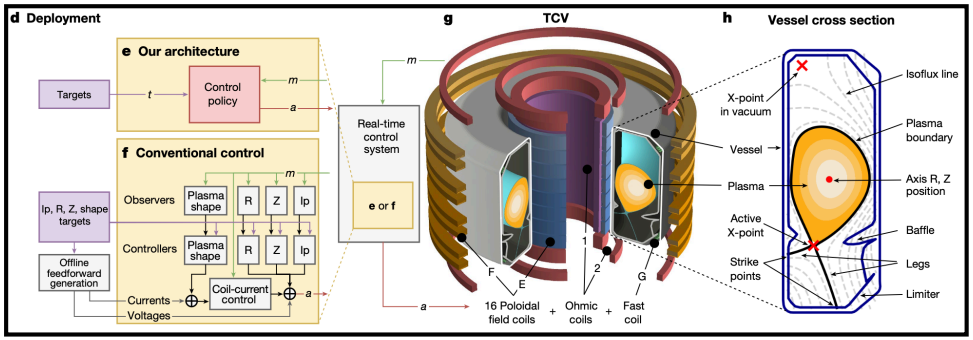
・以上のように、強化学習モデルはシミュレーション上で学習される。

・アクターは5,000個並列に学習され、実用時点ではそのアンサンブルが使われた。
（クリティックネットワークは1つ）

実験結果再訪



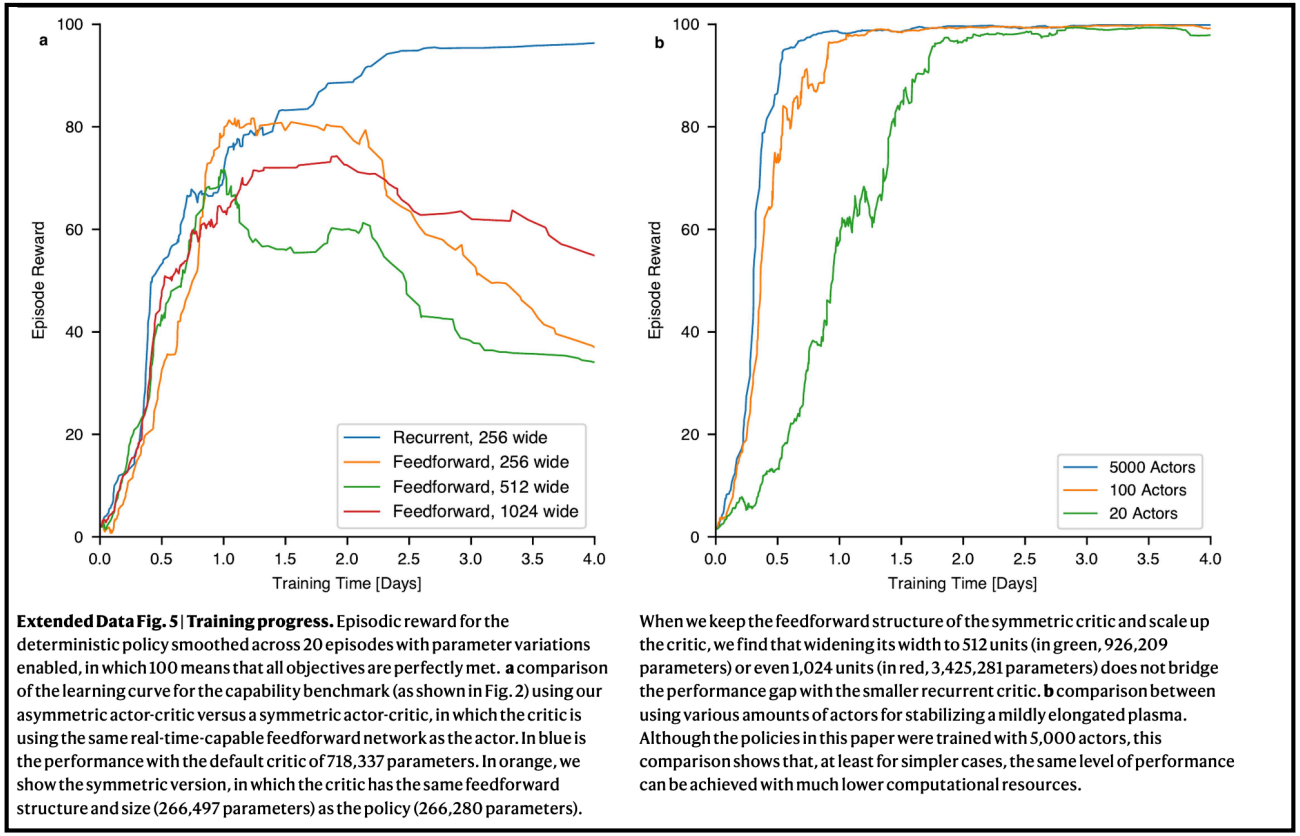
[論文より引用/Fig.2]



・訓練されたアクターをそのまま制御器として使った (zero-shot転移)。

・上図は1秒間の制御に成功していることを示している。

細かい点



Extended Data Fig. 5 | Training progress. Episodic reward for the deterministic policy smoothed across 20 episodes with parameter variations enabled, in which 100 means that all objectives are perfectly met. **a** comparison of the learning curve for the capability benchmark (as shown in Fig. 2) using our asymmetric actor-critic versus a symmetric actor-critic, in which the critic is using the same real-time-capable feedforward network as the actor. In blue is the performance with the default critic of 718,337 parameters. In orange, we show the symmetric version, in which the critic has the same feedforward structure and size (266,497 parameters) as the policy (266,280 parameters). When we keep the feedforward structure of the symmetric critic and scale up the critic, we find that widening its width to 512 units (in green, 926,209 parameters) or even 1,024 units (in red, 3,425,281 parameters) does not bridge the performance gap with the smaller recurrent critic. **b** comparison between using various amounts of actors for stabilizing a mildly elongated plasma. Although the policies in this paper were trained with 5,000 actors, this comparison shows that, at least for simpler cases, the same level of performance can be achieved with much lower computational resources.

[論文より引用/Extended Data Fig.5]

・左図はクリティックネットワークとして、RNNではなくMLPを使った場合を図示し、RNNの優位性を示している。

・右図はアクターネットワークのアンサンブル数を5,000個よりも少なくした場合のエピソード報酬の推移を示しており、アンサンブル数がずっと少なくても同じようなパフォーマンスが出ることを示している。

細かい点

data, to account for varying, uncontrolled experimental conditions. This provides robustness while ensuring performance. Although the simulator is generally accurate, there are known regions where the dynamics are known to be poorly represented. We built 'learned-region avoidance' into the training loop to avoid these regimes through the use of rewards and termination conditions (Extended Data Table 5), which halt the simulation when specified conditions are encountered.

[論文より引用/416page]

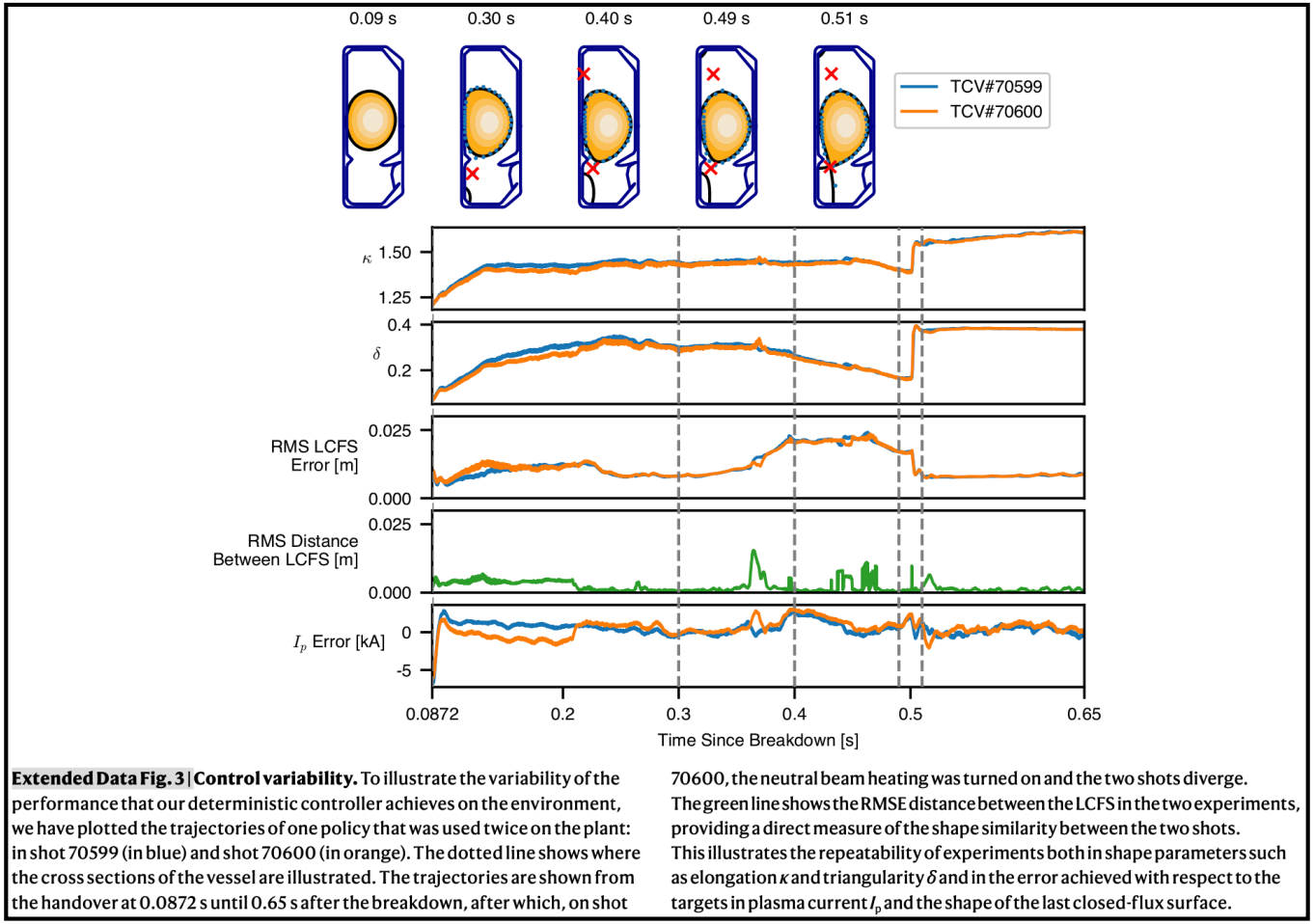
・シミュレーションが信頼できない領域で打ち切りを行なっている。

The control loop of TCV runs at 10 kHz, although only half of the cycle time, that is, 50 µs, is available for the control algorithm due to other signal processing and logging. Therefore we created a deployment system that compiles our neural network into real-time-capable code that is guaranteed to run within this time window. To achieve this, we remove superfluous weights and computations (such as the exploration variance) and then use tfcompile[43] to compile it into binary code, carefully avoiding unnecessary dependencies. We tailored the neural network structure to optimize the use of the processor's cache and enable vectorized instructions for optimal performance. The table of time-varying control targets is also compiled into the binary for ease of deployment. In future work, targets could easily be supplied at runtime to dynamically adjust the behaviour of the control policy. We then test all compiled policies in an automated, extensive benchmark before deployment to ensure that timings are met consistently.

[論文より引用/Deploymentセクション]

・制御器（アクター）が超高速で応答できるように工夫している。

細かい点



Extended Data Fig. 3 | Control variability. To illustrate the variability of the performance that our deterministic controller achieves on the environment, we have plotted the trajectories of one policy that was used twice on the plant: in shot 70599 (in blue) and shot 70600 (in orange). The dotted line shows where the cross sections of the vessel are illustrated. The trajectories are shown from the handover at 0.0872 s until 0.65 s after the breakdown, after which, on shot 70600, the neutral beam heating was turned on and the two shots diverge. The green line shows the RMSE distance between the LCFS in the two experiments, providing a direct measure of the shape similarity between the two shots. This illustrates the repeatability of experiments both in shape parameters such as elongation κ and triangularity δ and in the error achieved with respect to the targets in plasma current I_p and the shape of the last closed-flux surface.

[論文より引用/Extended Data Fig.3]

・この図は同じ制御器、制御目標を使うと、実験条件が異なっても (片方の実験（ショット70600）では、実験中に中性粒子ビーム加熱が追加されている)、同じような目標が達成できることを示している（制御器の安定性・再現性の確認）。

まとめと議論

・強化学習によるzero-shot Sim2Realにより、TCV上でプラズマ制御に成功した。

・筆者らは基本的な修正を加えることで、本手法がいくつかの仮定と技術的要件を満たす他のトカマクにもそのまま適用できると確信していると述べている。

・筆者らは、本手法が今日配備されている複雑な制御システムの設計や委託、建設前の設計案の評価を行う必要なく、新しいトカマクに迅速に配備することができる*と述べている。さらに、本手法はプラズマ形状、センシング、作動、壁設計、熱負荷、磁気制御を総合的に最適化することで、全体的なパフォーマンスを最大限に高める新しい原子炉設計の発見が可能になるかもしれないとも述べている。

*例えば「ドロップレット」の例において、既存のアプローチでも実現可能かもしれないが、その場合フィードフォワード・コイル電流プログラミングの開発、リアルタイム推定器の実装、コントローラ利得の調整、プラズマ生成後の制御の成功等に多大な投資が必要である。その一方本手法では、このケースに最適な報酬関数を定義するだけで済むと述べている。