Optimal ITB formation scenarios with reinforcement learning

R. Yoneda[1], T. Kojima[1], Y. Shirasawa[1], M. Takahashi[1], M. Honda[2] and A. Matsuyama[2]

1 Space Environment and Energy Laboratories, NTT, Inc., Tokyo, Japan
2 Kyoto University, Kyoto, Japan

E-mail: ryota.yoneda@ntt.com

1. Introduction

Internal transport barriers (ITBs) are attractive for advanced tokamak operation because they locally reduce heat transport and enable steep core temperature and pressure gradients [1,2]. Such improved confinement regimes are relevant for high-performance and steady-state operation scenarios [3–5]. However, ITB formation and sustainment are not simple scalar control problems. In ITB plasmas, pressure gradients, bootstrap current, q-profile, magnetic shear, and transport coefficients evolve as a coupled nonlinear system [2,6]. Therefore, a visible T_e-barrier is only one part of the ITB control problem, and a control objective based only on barrier height or barrier location may not be sufficient to evaluate whether the resulting state is useful for confinement.

A previous simulation study investigated reinforcement-learning (RL) based simultaneous control of the q-profile and normalized beta for JT-60SA ITB scenarios, highlighting the coupling between pressure profile, bootstrap current, q-profile, and beta as a central control issue [6]. RL has also recently been applied to several tokamak control problems, including magnetic control and instability avoidance [7,8]. The present study addresses a complementary problem: how an RL agent can access ITB-like T_e-barrier states in TASK/TR simulations, and how such states should be evaluated beyond scalar barrier metrics.

In this work, a sequential TASK/TR–RL environment was developed. Behavior cloning (BC) was used to initialize the policy near an ITB-forming trajectory, and proximal policy optimization (PPO) was then used for fine-tuning. The current ramp-up waveform was prescribed, while the RL agent controlled EC/ECCD power, EC deposition radius, and IC heating power. The purpose of this proceeding is to report the present formulation, PPO rollout results under normal and rapid current-ramp scenarios, and the resulting need for physics-informed reward design.

2. TASK/TR – RL formulation

A sequential control environment was constructed by coupling the TASK/TR transport simulation code to a Python-based RL framework. In the present study, density evolution and particle sources were fixed to focus on heat transport and current-profile evolution. TASK/TR solves heat transport coupled to current-profile evolution, and related transport simulation frameworks have been used to study improved confinement associated with current-profile modification [9].

The plasma current ramp-up waveform was prescribed and used as a background scenario. This choice reduced the action-space complexity and allowed the RL agent to focus on heating and localized EC/ECCD control. The PPO agent controlled three inputs,

$$a_t = [P_{\mathrm{EC}}\,,\ r_{\mathrm{EC}}\,,\ P_{\mathrm{IC}}],$$

where $P_{\mathrm{EC}}$is the EC power with ECCD enabled, r_{EC} is the EC deposition radius, and P_{IC} is the IC heating power. EC/ECCD provides localized heating and current drive, whereas IC

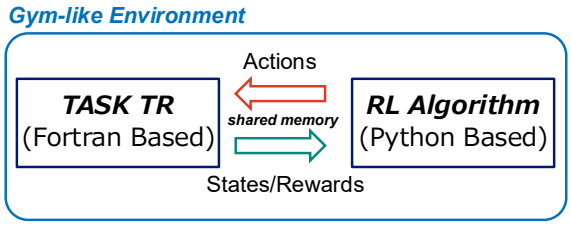
heating is treated as broad bulk heating. Separate policies were trained for normal and rapid prescribed current-ramp scenarios.

The observation vector included sampled q-profile and T_e-profile information, together with scalar quantities such as q_0, $q_{\min}$, q_{95}, and total auxiliary power. Although q_0, $q_{\min}$ and q_{95} are in principle contained in the sampled q-profile, they were included explicitly as scalar anchor quantities for control-relevant features. This representation helps the policy access key constraints and summary information, especially because $q_{\min}$ was also used in a penalty term to avoid trajectories with $q_{\min} < 1$. In this first formulation, barrier detection and the main reward were defined using the electron temperature T_e. Ion-channel quantities were not used directly as reward components in the present study, although they can be examined as TASK/TR diagnostics. This choice was made to focus on a minimal T_e-barrier formation problem, while leaving a more complete confinement-oriented formulation including ion-channel quantities and stored-energy-related metrics for future work. The main reward was based on a T_e-barrier metric defined as the maximum negative T_e gradient within a target radial window,

$$G_{T_e} = \max_{\rho \in \Delta\rho_{\text{target}}} \left(-\frac{dT_e}{d\rho} \right).$$

A location reward encouraged the detected barrier position ρ_{peak} to approach the target radius. Additional penalties were applied to avoid excessive power and abrupt actuator changes. Thus, the initial reward was designed to form a strong T_e-barrier near a target radial location while avoiding undesirable trajectories such as $q_{\min} < 1$.

Pure PPO training was found to be unstable and sample inefficient. This is expected because ITB formation requires coordinated heating, current-profile evolution, and delayed transport response. Therefore, behavior cloning (BC) was introduced to initialize the policy from a physics-guided ITB-forming trajectory. PPO was then used to fine-tune the policy in the TASK/TR environment. PPO approximates trust-region policy optimization using a clipped surrogate objective [10].



Category	Quantity	Meaning
Actions	$P_{\text{EC}}(0\text{–}1.0\ \text{MW})$	EC/ECCD power
	$r_{\text{EC}}(0.20\text{–}0.75\ \text{m})$	EC deposition radius
	$P_{\text{IC}}(0\text{–}3.0\ \text{MW})$	IC heating power
States	$q(\rho), q_0, q_{\min}, q_{95}$	magnetic-profile information
	$T_e(\rho)$	thermal-profile and barrier information
	P_{aux}	total auxiliary-heating level
Rewards	T_e-barrier strength	encourages a steep internal T_e gradient
	target-location reward	encourages the barrier to form near the target radius
	power penalty	discourages excessive actuator use
	smoothness penalty	discourages abrupt control variations

Fig. 1. TASK/TR–RL environment and RL formulation in table. The current ramp-up waveform is prescribed, while PPO controls EC/ECCD power, EC deposition radius, and IC heating power. The main reward is based on T_e-barrier strength and target radial location

3. BC + PPO Training Rollout Results

Figure 2 compares the best-reward PPO rollouts selected from evaluation episodes for the normal and rapid prescribed current-ramp scenarios. The two ramp-up cases were introduced to examine whether the same TASK/TR–RL formulation could access ITB-like T_e-barrier states under different background current-profile evolutions, rather than only reproducing a single reference scenario. Because the present reward was based mainly on scalar T_e-barrier strength and target location, such behavior should not be interpreted as proof that the agent has learned the full confinement-relevant ITB physics.

Separate BC + PPO policies were trained for the two scenarios using the same observation design, action space, and reward structure. In each case, behavior cloning was first applied to initialize the policy from a physics-guided ITB-forming trajectory, providing PPO with an initial policy already near a reasonable T_e-barrier response. PPO was then used as an online fine-tuning method in the TASK/TR environment to adjust the timing and combination of EC/ECCD power, EC deposition radius, and IC heating power under the prescribed current-ramp condition. This two-stage procedure was important because pure PPO training was unstable and sample-inefficient in this problem, where successful barrier formation requires coordinated heating, current-profile evolution, and delayed transport response. Thus, BC was used not to replace reinforcement learning, but to provide a physically reasonable initial policy from which PPO could search for improved closed-loop trajectories.

Both BC-initialized PPO policies formed ITB-like T_e-barrier states near the target radius. The detected barrier location approached the target radius in both scenarios. However, the learned actuator trajectories differed between the two prescribed current ramps. The rapid-ramp policy used a stronger early-phase IC heating response and showed a different EC deposition trajectory, whereas the normal-ramp policy produced a more moderate but more sustained late-phase T_e-barrier response. The sharp actuator variation and corresponding reduction of the T_e-barrier metric around $t \simeq 4$ s may partly reflect the inherited structure of the BC expert trajectory, which included a heating transition around this time.

These results demonstrate that BC-initialized PPO can access reward-defined ITB-like T_e-barrier regimes in TASK/TR under different prescribed current-ramp scenarios. However, they should not be interpreted as proof of a globally optimal ITB controller or of a complete physical mechanism of ITB formation. Rather, they show that the proposed TASK/TR–RL formulation provides a practical framework for generating and comparing ITB-forming control trajectories, motivating future reward design beyond scalar T_e-barrier height and location.

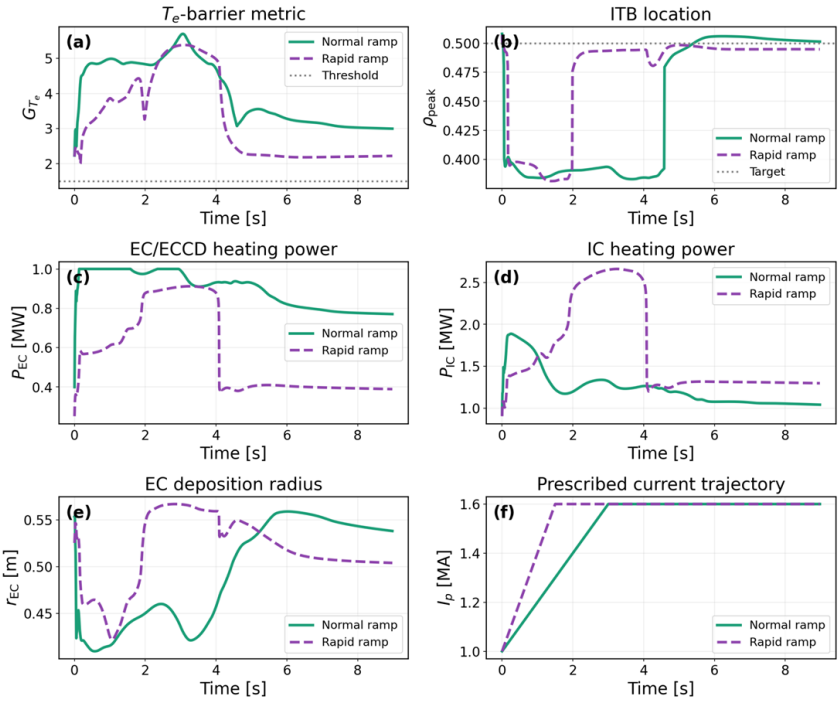


Fig. 2. Best-reward PPO rollout comparison for normal and rapid prescribed current-ramp scenarios. The plotted trajectories were selected as the highest-reward episodes during evaluation. (a) Time evolution of the T_e-barrier metric G_{T_e}, where the dotted line indicates the threshold used to identify barrier formation. (b) Detected ITB location ρ_{peak}, defined by the peak position of G_{T_e} with the target radius of 0.5. (c) EC/ECCD heating power P_{EC}. (d) IC heating power P_{IC}. (e) EC deposition radius r_{EC}. (f) Prescribed plasma current trajectory I_p.

4. Conclusions

A sequential TASK/TR–RL environment was developed to investigate reward-defined ITB-like T_e-barrier formation using reinforcement learning. The plasma current ramp-up waveform was prescribed, while the BC-initialized PPO agent controlled EC/ECCD power, EC deposition radius, and IC heating power. Behavior cloning provided a physics-guided initial policy near an ITB-forming trajectory, enabling PPO fine-tuning in the nonlinear TASK/TR environment where pure PPO was unstable and sample-inefficient.

Separate BC + PPO policies were trained for normal and rapid prescribed current-ramp scenarios using the same observation design, action space, and reward structure. These two cases tested whether the same RL formulation could access T_e-barrier states under different background current-profile evolutions. Both policies formed ITB-like T_e-barrier states near the target radius, but their learned EC/IC heating and EC deposition trajectories differed, indicating that the prescribed current ramp affected the reward-seeking control trajectory.

The rollout comparison also clarifies the present limitation of the formulation. The BC-initialized PPO policies satisfied the scalar reward based on T_e-barrier strength and target location, but this does not prove that the agent has learned the full confinement-relevant physics of ITB formation. Part of the rollout timing may also reflect the inherited structure of the BC expert trajectory. Thus, scalar T_e-barrier height and location are useful first proxies for accessing ITB-like regimes, but they are not sufficient as final control objectives.

The present result should therefore be regarded not as a completed optimal ITB controller, but as a step toward a practical RL framework for ITB control studies. Future work will test physics-informed rewards using confinement-relevant quantities such as stored energy, H_{98y}, input-power efficiency, ion-channel quantities, bootstrap current, magnetic shear, and transport reduction. Reward design will be extended from initial T_e-barrier formation to barrier sustainment, where ECCD-driven current-profile control may help maintain the ITB-like state after the early formation phase. The RL methodology and its generalization capability will also be validated through baseline comparisons, seed and hyperparameter studies, and robustness tests across ramp scenarios, transport parameters, actuator constraints, and initial plasma conditions.

References

[1] F. M. Levinton et al., “Improved confinement with reversed magnetic shear in TFTR,” *Phys. Rev. Lett.* **75**, 4417–4420 (1995).

[2] X. Litaudon, “Internal transport barriers: critical physics issues?” *Plasma Phys. Control. Fusion* **48**, A1–A34 (2006).

[3] C. Gormezano et al., “Chapter 6: Steady state operation,” *Nucl. Fusion* **47**, S285–S336 (2007).

[4] N. Hayashi et al., “Transport modelling of JT-60U and JET plasmas with internal transport barriers towards prediction of JT-60SA high-beta steady-state scenario,” *Nucl. Fusion* **57**, 126037 (2017).

[5] L. Garzotti et al., “Analysis of JT-60SA operational scenarios,” *Nucl. Fusion* **58**, 026029 (2018).

[6] T. Wakatsuki et al., “Simultaneous control of safety factor profile and normalized beta for JT-60SA using reinforcement learning,” *Nucl. Fusion* **63**, 076017 (2023).

[7] J. Degrave et al., “Magnetic control of tokamak plasmas through deep reinforcement learning,” *Nature* **602**, 414–419 (2022).

[8] J. Seo et al., “Avoiding fusion plasma tearing instability with deep reinforcement learning,” *Nature* **626**, 746–751 (2024).

[9] A. Fukuyama et al., “Transport simulation on L-mode and improved confinement associated with current profile modification,” *Plasma Phys. Control. Fusion* **37**, 611 (1995).

[10] J. Schulman et al., “Proximal policy optimization algorithms,” arXiv:1707.06347 (2017).