Evaluation of Noise Robustness of DRL-Based Autonomous Aggressive Cooldown Control under Different Reward

YooJoon Seoung, Junyong Bae, Seung Jun Lee*
Department of Nuclear Engineering
Ulsan National Institute of Science and Technology (UNIST)
50 UNIST-gil, Ulju-gun, Ulsan, 44949, Republic of Korea
**Corresponding author: sjlee420@unist.ac.kr*

***Keywords : deep reinforcement learning, autonomous operation, aggressive cooldown**

1. Introduction

Small modular reactors (SMRs) are moving from concept to deployment as a pragmatic option for decarbonized, dispatchable power. Their core premises compact and modular plant layout, factory fabrication, and standardized designs promise lower capital risk and shorter construction schedules. Technically, SMRs emphasize passive safety (e.g., natural circulation and gravity-driven injection), simplified balance-of-plant, and operational flexibility that allows following variable demand. A distinctive operational feature is the multi-module configuration, in which a single control room crew or even a single licensed operator may supervise several reactor modules concurrently. While this architecture improves staffing efficiency, it also redistributes cognitive workload: operators must monitor more channels, diagnose concurrent transients, and coordinate multiple actuation pathways in real time. This human-factors pressure strengthens the case for autonomous operation that can shoulder time-critical control while keeping humans "on the loop" for supervision and authorization.

In parallel, the nuclear community has begun testing artificial intelligence (AI) in particular deep reinforcement learning (DRL) as a control paradigm for complex, nonlinear, and continuous action problems. Two strands of prior work are especially relevant. First, studies comparing DRL with traditional PID control for automatic cold shutdown reported that learned policies can coordinate multi-actuator actions and reduce operator burden in routine yet extended operations. Second, Bae et al. (2023) demonstrated SAC (Soft Actor–Critic) with HER (Hindsight Experience Replay) for multi-objective start-up automation, simultaneously regulating reactor coolant temperature, pressurizer pressure, and inventory within a compact simulator. Collectively, these findings establish the feasibility of AI assisted or even AI driven autonomous operation in nuclear settings, at least for nominal or planned procedures.

However, two important gaps remain. First, prior DRL applications have rarely targeted emergency operations, where the safety margin is thin, time pressure is high, and control objectives/constraints can conflict (e.g., depressurization versus structural limits). Second, most demonstrations were validated only in the training environment; they did not systematically examine generalization when targets change (e.g., a different cooling rate set-point) or when relevant but unobserved plant variables deviate from nominal (e.g., auxiliary feedwater status). Addressing these gaps is crucial if autonomous control is to contribute to emergency operating procedures without eroding safety.

Present study extends our prior investigation by considering the gap between simulator-based training and real plant environments. In actual nuclear power plants, sensor and process signals inevitably contain measurement noise and disturbances. Therefore, we performed additional experiments to evaluate whether the trained DRL agent can remain effective under noisy conditions, and how the reward shaping strategy influences the agent's adaptability. This extension aims to bridge the gap toward real-world applicability, testing not only nominal performance but also robustness under signal uncertainty.

2. Modeling and Methodology

This study develops the DRL framework that enables autonomous control of aggressive cooldown operations. The task is formulated as a Markov Decision Process (MDP), with states representing thermal-hydraulic variables, actions corresponding to continuous valve operations, and rewards designed to reflect safety and performance criteria. The agent is trained using Soft Actor–Critic (SAC) in combination with Hindsight Experience Replay (HER).

2.1 Soft Actor-Critic (SAC)

SAC is a model-free, off-policy DRL algorithm designed for continuous control tasks. Its main advantage lies in balancing reward maximization with entropy regularization, which encourages broad exploration and reduces the risk of converging to poor local optima. Another key feature is the use of two critic networks; by updating policies against the minimum of the two Q-value estimates, SAC alleviates overestimation bias and improves training stability. The algorithm's ability to generate continuous control signals (e.g., a partial valve opening) is particularly suitable for nuclear power plants, where precise and smooth actuation is essential. Moreover, the off-policy design allows efficient reuse of

collected experiences, which is critical in high-cost simulation environments

2.2 Hindsight Experience Replay (HER)

To address the sparse-reward problem, we incorporate HER, which enhances sample efficiency by relearning failed episodes with goals that were achieved. Instead of discarding trajectories where the original goal was missed, HER retrospectively substitutes alternative goals and recalculates rewards. This process transforms otherwise unsuccessful experiences into useful training data, substantially improving convergence speed and robustness. By combining HER with SAC, our framework can learn effective emergency control strategies even when informative feedback signals are scarce

3. Experiments

3.1 Aggressive cooldown

Aggressive cooldown is one of the most critical emergency operating procedures, typically initiated in accident scenarios such as a small break loss-of-coolant accident (SBLOCA) combined with a failure of the Safety Injection System (SIS). The procedure aims to rapidly lower both the temperature and pressure of the reactor coolant system (RCS). Fast depressurization is essential for activating low-pressure safety injection (LPSI) systems, including the Shutdown Cooling System (SCS), while maintaining adequate core cooling to keep fuel cladding temperatures within safety margins.

The process is usually carried out by opening atmospheric dump valves (ADVs) on the secondary side, which releases steam and accelerates depressurization. At the same time, the auxiliary feedwater (AFW) system delivers water to the steam generators to sustain heat removal. This coordinated operation enables a gradual reduction of RCS pressure and temperature until conditions are favorable for LPSI initiation.

Strict operational limits add complexity to aggressive cooldown. The cooling rate must remain below 55.6 °C/hr to prevent excessive thermal stresses that could damage major components such as the reactor pressure vessel, steam generators, or piping systems. Furthermore, successful SCS injection requires reducing RCS temperature to 177 °C or lower, and pressure to approximately 285.1 psia. Achieving these targets while avoiding structural risks is critical for the success of the procedure.

Because aggressive cooldown involves simultaneous, tightly coupled objectives rapid depressurization, rate limitation, and reaching precise target conditions it represents a highly challenging control problem. Reinforcement learning (RL) is well suited to this context, as an agent can learn to balance these trade-offs by continuously adjusting control variables such as ADV opening fractions and AFW flow rates. The inherently continuous nature of these actions highlights the appropriateness of RL algorithms designed for continuous control domains.

3.1 Reward function

In reinforcement learning, the reward function directly shapes agent behavior and thus strongly influences both training efficiency and final performance. For aggressive cooldown control, different operational priorities may call for different reward structures for example, rapid convergence to the target cooling rate, strict suppression of overshoot, or maximization of steady-state precision. Furthermore, robustness against measurement noise is not intrinsic to the learning algorithm but depends critically on how the reward is formulated. Therefore, this study systematically compares multiple reward shapes to examine how reward design affects tracking accuracy and adaptability under noisy conditions.

To investigate the effect of reward shaping on learning performance and robustness, we conducted comparative experiments using four distinct reward functions (Shape #1–#4), as illustrated in Fig. 1.

$$r_1(\mathrm{d}) = \begin{cases} 1000(1\text{-}\mathrm{d}/5), d < 5 \\ 0, \mathrm{d} \geq 5 \end{cases}$$

$$r_2(\mathrm{d}) = \begin{cases} 1000(1\text{-}\mathrm{d}/5), \mathrm{d} < 5, \mathrm{cooling\_rate} \leq \mathrm{target} \\ 0, \mathrm{otherwise} \end{cases}$$

$$r_3(\mathrm{d}) = \begin{cases} 1000(1\text{-}\mathrm{d}/5)^3, \mathrm{d} < 5, \mathrm{cooling\_rate} \leq \mathrm{target} \\ 0, \mathrm{otherwise} \end{cases}$$

$$r_4(\mathrm{d}) = \begin{cases} 1000, \mathrm{d} < 0.05 \\ 0, \mathrm{d} \geq 0.05 \end{cases}$$

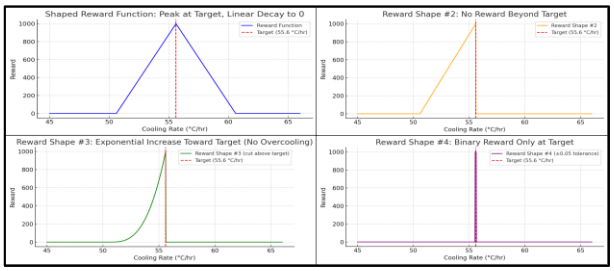

Fig. 1. Reward function shapes for comparative experiments: (a) Shape #1 – Linear decay around the target, (b) Shape #2 – No reward beyond the target, (c) Shape #3 – Exponential increase toward the target with overshoot suppression, (d) Shape #4 – Binary reward only at the target.

4. Result

The effect of reward shaping on the performance and robustness of the trained DRL agents was systematically evaluated. For each reward function (Shape #1–#4), the trained model was tested in ten independent runs under nominal conditions. The tracking error with respect to the reference cooling rate of 55.6 °C/hr was quantified using the mean squared error (MSE).

In addition, the ratio of overshoot occurrences beyond the operational limit of 55.6 °C/hr was measured to assess safety. To further evaluate robustness, noise was

injected into the AFW flow signal during testing, and the resulting MSE values were compared across reward functions.

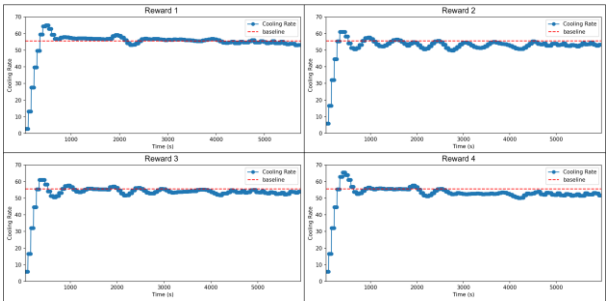

Fig. 2. Cooling rate trajectories under different reward

The comparative results are summarized in Tables 1 and 2. Table 1 reports the overshoot ratios obtained under different reward structures. Reward Shape #1 yielded the highest overshoot ratio (61.54%), indicating frequent violation of the operational limit. In contrast, Reward Shapes #2–#4, which assign zero reward once the target cooling rate is exceeded, effectively suppressed overshoot, achieving ratios of 10.91%, 17.49%, and 19.19%, respectively. These results confirm that explicitly penalizing overshoot in the reward function substantially improves safety by reducing the likelihood of excessive cooling.

Table 2 presents the tracking performance in terms of mean squared error (MSE) under nominal and noisy conditions. Among the tested reward structures, Reward Shape #3 achieved the lowest MSE in both nominal (4.25) and noisy conditions (4.31), with only a minor degradation (ΔMSE = 1.43%). Reward Shape #2 also exhibited relatively stable performance, while Reward Shape #1 showed moderate accuracy and slightly higher sensitivity to noise (ΔMSE = 1.51%). Reward Shape #4, although comparable under nominal conditions, demonstrated the largest degradation under noise (ΔMSE = 2.26%), indicating limited robustness.

Taken together, these findings highlight that reward design critically influences both control accuracy and robustness. Reward Shape #3 provides the most balanced performance, combining low tracking error with resilience to noise, while Reward Shape #1 and #4 reveal clear limitations in safety and robustness, respectively.

Table. 1. Overshoot ratios under different rewards

Reward	Overshoot ratio (%)
1	61.54
2	10.91
3	17.49
4	19.19

Table. 2. MSE comparison under different rewards

Reward	MSE	MSE (with noise)	ΔMSE
1	9.5103	9.6536	1.51
2	8.7559	8.9194	1.87
3	4.2521	4.3129	1.43
4	10.164	10.394	2.26

5. Conclusion

This study investigated the effect of reward function design on deep reinforcement learning (DRL) for autonomous aggressive cooldown control. Several reward structures were designed and systematically compared to assess their impact on learning efficiency, control accuracy, and robustness.

The results demonstrated that the shape of the reward function substantially influences both convergence speed and final precision. Reward structures with relatively flat gradients around the target promoted stable learning but limited ultimate accuracy, whereas those with steeper reward gradients increased learning difficulty but enabled higher precision once converged.

Robustness tests under noisy AFW flow signals further confirmed that reward design is critical for resilience: some reward structures preserved stable tracking, while others exhibited significant degradation.

These findings highlight that careful reward shaping tailored to operational objectives such as accuracy, stability, and robustness is essential for practical deployment of DRL-based controllers in nuclear power plant operations. Future research will explore more complex observation spaces, multi-objective reward formulations, and validation against real plant scenarios.

ACKNOWLEDGMENTS

This work was supported by Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the Korea government (MOTIE) (RS-2024-00403194, Next-Generation Nuclear Technology Creation IP-R&D Talent (Human Resources) Development Project)

This research was supported by the National Research Council of Science & Technology(NST) grant by the Korea government (MSIT) (No. GTL24031-400)

REFERENCES

[1] Kim, Man Cheol, and Sang Hoon Han. "Variability of plant risk due to variable operator allowable time for aggressive cooldown initiation." Nuclear Engineering and Technology 51.5 (2019): 1307-1313.

[2] Bae, Junyong, Jae Min Kim, and Seung Jun Lee. "Deep reinforcement learning for a multi-objective operation in a nuclear power plant." Nuclear Engineering and Technology 55.9 (2023): 3277-3290

[3] Lee, Daeil, et al. "Comparison of deep reinforcement learning and PID controllers for automatic cold shutdown operation." Energies 15.8 (2022): 2834.

[4] Haarnoja, Tuomas, et al. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor." International conference on machine learning. Pmlr, 2018.

[5] Andrychowicz, Marcin, et al. "Hindsight experience replay." Advances in neural information processing systems 30 (2017).