Adaptive Reward Shaping via Meta Reinforcement Learning for NPP Aggressive Cooldown Control

YooJoon Seoung, 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 : Meta reinforcement learning, Reward shaping, SAC, Nuclear power plant, Aggressive cooldown

1. Introduction

Deep reinforcement learning (DRL) has emerged as a promising approach for autonomous control of nuclear power plant (NPP) operations, particularly in emergency scenarios that demand rapid and precise control under strict safety constraints [1,2]. Among DRL algorithms, Soft Actor-Critic (SAC) has demonstrated effectiveness in continuous NPP control tasks due to its off-policy nature and entropy-regularized objective [3].

However, reward engineering remains a critical bottleneck in applying DRL to NPP operations. The reward function serves as the de facto problem definition, yet designing effective rewards for safety-critical systems is fraught with challenges: (1) reward sparsity, where meaningful signals appear only at episode end after hours of simulation; (2) scale mismatch between multiple constraint violation terms; and (3) exploration collapse, where strong penalties prevent learning near safety boundaries [4]. Our prior work demonstrated empirically that different reward function designs produce qualitatively different control behaviors for the same aggressive cooldown task, with performance varying by orders of magnitude depending on the reward shape.

When operational objectives change, the entire reward function must be manually redesigned. This paper proposes a meta reinforcement learning (Meta-RL) framework that automatically determines optimal reward shaping parameters conditioned on user-specified operational weights, eliminating the need for manual reward engineering.

2. Modeling and Methodology

2.1 Problem Formulation

The aggressive cooldown task during a small-break loss-of-coolant accident (SBLOCA) is formulated as a Markov Decision Process. The state vector comprises the current cooling rate and target cooling rate (55.6°C/hr), and the action controls the atmospheric dump valve (ADV) position via a continuous signal. The cooling rate is calculated as a sliding-window average over 1800 seconds. The base-level control policy is learned using Soft Actor-Critic (SAC), an off-policy DRL algorithm well-suited for continuous control. Ten parallel simulator instances share a common replay buffer to accelerate training.

2.2 Adaptive Reward Shaping via Meta-RL

A central challenge in applying DRL to NPP operations is that the reward function must be manually redesigned whenever operational objectives change. For instance, shifting from a tracking-priority to a safety-priority strategy requires fundamentally different reward landscapes, and hand-tuning these for each scenario is labor-intensive and non-transferable. This motivates the use of meta reinforcement learning to automate the reward shaping process itself.

The proposed framework employs a bi-level optimization architecture. The inner loop (base-level) consists of the SAC agent learning a control policy under the shaped reward parameterized by θ. The outer loop (meta-level) consists of a meta-policy neural network (128-64 units, sigmoid output) that observes recent base-level performance and outputs updated reward parameters.

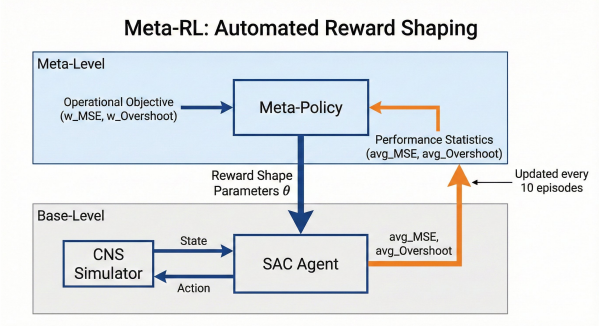

Fig. 1. Reward Shaping via Meta-RL Framework

The shaped reward combines a task-specific component with a potential-based shaping term [5]:

$$r_{shaped} = r_{task} + \gamma\Phi(s';\theta) - \Phi(s;\theta) \quad (1)$$

$$r_{task} = -\text{MSE} \times w_{\text{MSE}} - \text{Overshoot} \times w_{\text{Overshoot}} \quad (2)$$

where MSE = (cooling_rate − target)² quantifies tracking accuracy, and Overshoot = max(0, cooling_rate − target)² penalizes safety violations. The weights w_MSE and w_Overshoot are user-specified

operational priorities summing to 1.0, encoding the desired trade-off between tracking accuracy and safety.

The potential function $\Phi(s; \theta)$ is a novel 5-point piecewise linear function parameterized by seven learnable parameters $\theta = [x_1, x_2, x\_target, x_3, x_4, y_2, y_3]$. Five control points $(x_1, 0)$, (x_2, y_2), $(x\_target, 1000)$, (x_3, y_3), $(x_4, 0)$ define a continuous piecewise-linear potential landscape with a fixed peak of 1000 at x_target. This structure was chosen because it provides sufficient flexibility to represent symmetric, asymmetric, narrow, or wide reward shapes while keeping the parameter space compact (7 dimensions) and thus tractable for meta-learning. The position parameters (x_1 through x_4) control where the reward gradient is steep or gentle, while the height parameters (y_2, y_3) govern asymmetry. Notably, x_target is itself learnable, allowing the meta-policy to shift the effective target away from the nominal value when doing so improves performance under a given operational objective.

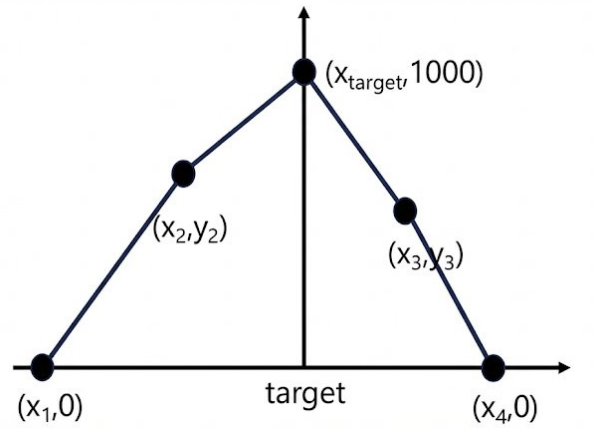

Fig. 2. 7-Parameter Piecewise Potential Function

3. Result

3.1 Experimental Setup

The aggressive cooldown task during a small-break loss-of-coolant accident (SBLOCA) is formulated as a Markov Experiments were conducted on the Compact Nuclear Simulator (CNS), a low-fidelity, high-speed simulator modeling a Westinghouse 1000 MWe 3-loop PWR [6]. The SBLOCA scenario has a break size of 5.20 cm² injected at 60 s, with agent control at 300 s. Each episode simulates 4 hours with a 12 s control interval. Six weight configurations spanning the tracking–safety spectrum were evaluated (Table I).

Table I: Weight for experimental evaluation

w_MSE	w_Ovs	Description
1.0	0.0	Tracking Only
0.8	0.2	Tracking Priority
0.6	0.4	Tracking Preferred
0.4	0.6	Safety Preferred
0.2	0.8	Safety Priority
0.0	1.0	Safety Only

3.2 MSE and Overshoot Performance

Table II presents the integrated results across all weight configurations, evaluated over the final 10 episodes.

Table II: MSE and Overshoot across weight configurations

(w_MSE, w_Ovs)	**MSE**	**Overshoot**
(1.0, 0.0)	78.18±0.39	0.355±0.063
(0.8, 0.2)	79.17±1.63	0.362±0.029
(0.6, 0.4)	80.96±2.80	0.356±0.049
(0.4, 0.6)	81.05±7.54	0.342±0.060
(0.2, 0.8)	81.67±9.45	0.312±0.062
(0.0, 1.0)	**1532.65±68.60**	**0.000**

Result for MSE weight dominant cases, as w_MSE increases, MSE monotonically decreases. Ranks 1–5 form a stable cluster (78–82) with marginal differences. The (0.0, 1.0) configuration is a pathological outlier (MSE ≈ 1533, ~19× higher), as the agent receives no tracking signal. Variance also increases systematically as w_MSE decreases (±0.39 to ±9.45), indicating reduced consistency.

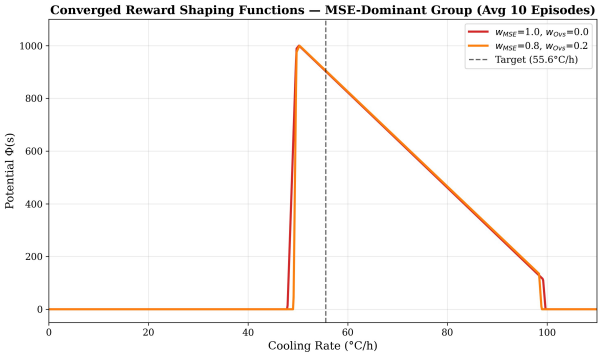

Fig. 3. Reward Shape for weight (1.0, 0.0), (0.8, 0.2)

For Overshoot weight dominant case, a threshold effect is observed. For w_Overshoot ≤ 0.4, overshoot converges to ~0.35 regardless of weight ratio. Meaningful reduction appears only when w_Overshoot ≥ 0.6. The (0.0, 1.0) case achieves zero overshoot at catastrophic MSE cost.

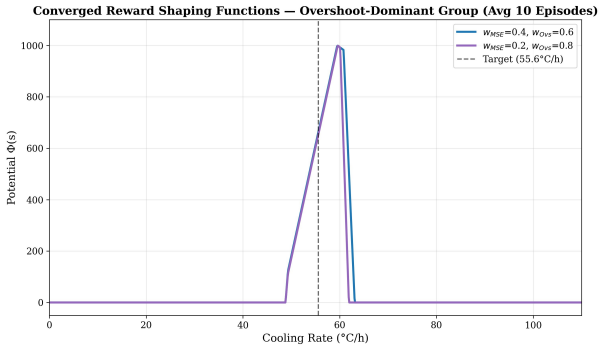

Fig. 4. Reward Shape for weight (0.4, 0.6), (0.2, 0.8)

3.3 Meta-Learned Reward Strategies

For safety-dominant configurations (0.4, 0.6) and (0.2, 0.8), the meta-policy discovers two qualitatively distinct strategies. Strategy A positions the potential peak below the actual target (x_target ≈ 50.3) with an asymmetric high left wall ($y_2 >> y_3$), creating a conservative approach that discourages overshoot by steering the agent below the limit. Strategy B positions the peak above the target (x_target ≈ 59.7) with a high right wall ($y_3 >> y_2$), encouraging the agent to achieve higher cooling rates for better tracking. The emergence of these dual strategies for identical weight configurations demonstrates the meta-policy's ability to discover non-obvious, multi-modal reward structures, offering operators a choice between qualitatively different control philosophies.

4. Conclusion

This paper proposed a Meta-RL framework for adaptive reward shaping in RL-based NPP aggressive cooldown control during SBLOCA. Reward engineering was identified as the primary bottleneck. The weight-conditioned Meta-RL framework successfully automates optimal reward shape determination across six configurations: increasing w_MSE monotonically reduces tracking error (H1 confirmed), while overshoot reduction requires a dominant overshoot weight ≥ 0.6 (H2 partially confirmed with threshold effect). The meta-policy autonomously discovers distinct reward strategies depending on the operational emphasis. Future work will extend to multi-objective constraints, dynamic weight adaptation during episodes, and validation on high-fidelity simulators.

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] D. Lee, J. Son, J. Kim, and S. J. Lee, "Comparison of deep reinforcement learning and PID controllers for automatic cold shutdown operation," Energies, vol. 15, no. 8, p. 2834, 2022.
[2] J. Bae, J. M. Kim, and S. J. Lee, "Deep reinforcement learning for a multi-objective operation in a nuclear power plant," Nucl. Eng. Technol., vol. 55, no. 9, pp. 3277–3290, 2023.
[3] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in Proc. ICML, PMLR, pp. 1861–1870, 2018.
[4] K. H. N. Nguyen et al., "Reinforcement learning-based control sequence optimization for advanced reactors," J. Nucl. Eng., vol. 5, no. 3, pp. 209–225, 2024.
[5] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in Proc. ICML, pp. 278–287, 1999.
[6] M. C. Kim and D. W. Jerng, "Feasibility analysis of aggressive cooldown in OPR-1000 nuclear power plants," Ann. Nucl. Energy, vol. 68, pp. 89–95, 2014.
[7] M. Andrychowicz et al., "Hindsight experience replay," in Advances in Neural Information Processing Systems, vol. 30, 2017.