



Journal Website:
<http://sciencebring.com/index.php/ijasr>

Copyright: Original content from this work may be used under the terms of the creative commons attributes 4.0 licence.

Research Article

Real-Time Adaptive Virtual Inertia Control of Grid-Forming Battery Energy Storage Using Safe Reinforcement Learning

Submission Date: February 12, 2026, **Accepted Date:** March 10, 2026,

Published Date: March 31, 2026

Crossref doi: <https://doi.org/10.37547/ijasr-06-03-11>

 **Oybek Achilovich Kayumov**

Department of Computer Science and Programming, Jizzakh Branch of the National University of Uzbekistan named after Mirzo Ulugbek, Jizzakh, Uzbekistan

 **Nazokat Rashitovna Kayumova**

Department of Computer Science and Programming, Jizzakh Branch of the National University of Uzbekistan named after Mirzo Ulugbek, Jizzakh, Uzbekistan

ABSTRACT

Reinforcement learning (RL) can adapt the virtual inertia and damping of grid-forming battery energy storage systems (BESS) in real time, but an agent that optimizes performance alone may violate frequency-security limits, and a safety filter added to it may simply be exploited. This paper proposes a safe-RL controller that updates the virtual inertia H_v and damping D_v of a virtual synchronous machine every 20 ms. The constrained Markov decision process is solved by a Lagrangian evolution strategy, and a model-based shield enforces analytic lower bounds on H_v and D_v derived from an online estimate of the disturbance. A second Lagrange multiplier penalizes shield interventions, so that the agent learns to be safe by itself instead of relying on the shield. Domain randomization of the inertia estimate is used during training. On 500 test scenarios of a low-inertia system (inverter-based share 40–85 %, disturbances 3–12 %), the proposed controller produced no violation of the 0.5 Hz/s RoCoF and 49.5 Hz nadir limits for any of three training seeds. It used 32 % less BESS energy than a fixed VSM, which violated the limits in 3.2 % of scenarios, and 17 % less than a rule-based adaptive law, and the shield had to intervene in only 6.4 % of control steps. Unconstrained RL violated the limits in 80 % of scenarios, Lagrangian RL without a shield in 4.3 %, and a shielded agent without the intervention cost relied on the shield in 64 % of steps. Under out-of-distribution conditions with larger disturbances and ± 20 % inertia-estimation error, the proposed controller had the lowest rate of frequency-limit violation (2.1 % against 3.3–19.0 %). The policy requires about 3 μ s per decision.

KEYWORDS

Safe reinforcement learning; constrained Markov decision process; shielding; grid-forming inverter; virtual inertia; battery energy storage; frequency stability; evolution strategies.

INTRODUCTION

Grid-forming battery energy storage systems (BESS) controlled as virtual synchronous machines (VSMs) are a key means of supporting frequency in power systems with declining synchronous inertia [1–4]. Their virtual inertia H_v and damping D_v shape the rate of change of frequency (RoCoF), the frequency nadir and the energy drawn from the battery. Fixed parameters are a compromise: they must be large enough for the worst operating point and therefore use stored energy unnecessarily elsewhere. Adaptive laws that change the parameters with the frequency deviation and its derivative [5, 6] improve this trade-off, but they rely on hand-tuned rules.

Reinforcement learning (RL) can learn such state-dependent adaptation directly from interaction, and it has been applied to a range of power-system control problems [7, 8], including primary frequency control [9]. A learned controller that acts on a security-critical variable must, however, respect hard limits. Safe RL addresses this with two families of methods [10]. Constrained Markov decision processes (CMDPs) [11], solved by Lagrangian or trust-region methods [12, 13], keep the expected constraint cost below a threshold, but only on average and only after training. Shields and safety filters [14–16] correct

unsafe actions at run time, but an agent trained behind a shield can learn to push against it and leave the safety work to the shield. Such a controller is only as good as the model that the shield uses.

This paper combines both mechanisms for real-time virtual-inertia control and addresses their weaknesses. The contributions are: **(i)** a CMDP formulation in which an agent updates H_v and D_v of a grid-forming BESS every 20 ms to minimize battery energy and frequency deviation, subject to RoCoF, nadir and converter-saturation constraints; **(ii)** a model-based shield with closed-form lower bounds on H_v and D_v computed from an online estimate of the power imbalance; **(iii)** a second Lagrangian constraint on the magnitude of shield interventions, which, together with domain randomization of the inertia estimate, makes the learned policy itself safe and leaves the shield as a rarely used backstop; and **(iv)** a systematic comparison with fixed, rule-based, shield-only, unconstrained and Lagrangian controllers over three training seeds, both in and out of the training distribution.

METHODS

The proposed framework is shown in Fig. 1.

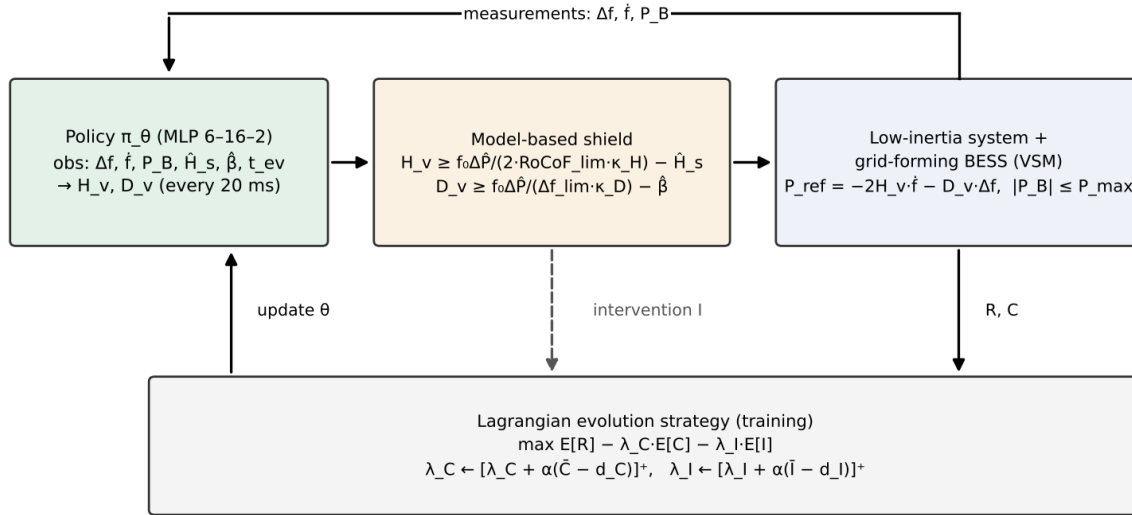


Fig. 1. Proposed safe-RL framework: policy, model-based shield, plant and Lagrangian evolution-strategy training.

1. Plant

The power system is an aggregated single-area model with a grid-forming BESS [17, 18]:

$$2H_s \cdot d\Delta f/dt = \Delta P_m + \Delta P_B - \Delta P_L - D \cdot \Delta f \quad (1)$$

$$T_g \cdot dx_g/dt = -K \cdot \Delta f - x_g, \quad (2)$$

$$T_t \cdot d\Delta P_m/dt = x_g - \Delta P_m$$

$$T_c \cdot d\Delta P_B/dt = -(2H_v \cdot d\Delta f/dt + D_v \cdot \Delta f) - \Delta P_B \quad (3)$$

with $|\Delta P_B| \leq P_{\max} = 0.12$ p.u., $T_g = 0.2$ s, $T_t = 0.3$ s, $T_c = 20$ ms, droop $R = 0.05$ and $K = (1 - \eta)/R$. Each episode lasts 10 s, and a step loss of generation ΔP_L occurs at $t = 0.2$ s. The operating condition is drawn at random: $H_0 \in [4, 6]$ s, inverter-based share $\eta \in [0.40, 0.85]$ (so $H_s = H_0(1 - \eta)$), $D \in [0.5, 1.5]$ p.u. and $\Delta P_L \in [0.03, 0.12]$ p.u. The controller receives estimates $\hat{H}_s$ and $\hat{\beta} = \hat{D} + \hat{K}$ with relative errors $N(0, 8\%)$ and $N(0, 2\%)$, as could be obtained from a physics-

informed estimator. The model is integrated with a 2 ms explicit Euler step, and the controller acts every 20 ms. Frequency and its derivative are measured with Gaussian noise (2 mHz and 0.01 Hz/s).

2. Constrained MDP

The observation is $o = [\Delta f, \dot{f}, \Delta P_B/P_{\max}, \hat{H}_s, \hat{\beta}, t_{ev}]$, where t_{ev} is the time since the event was detected. The action $a \in [-1.5, 1] \times [-3, 1]$ sets $H_v = 2 \cdot e^{a_1}$ s and $D_v = 8 \cdot e^{a_2}$ p.u. The episode return and the constraint cost are

$$R = -(\int |\Delta P_B| dt + 0.1 \cdot \int |\Delta f| dt) \quad (4)$$

$$C = [RoCoF_{500} - 0.5]^+/0.5 + [49.5 - f_{\min}]^+/0.5 + t_{sat} \quad (5)$$

where the first term of R is the BESS energy throughput (p.u.·s), the second is the integrated absolute frequency deviation (Hz·s), $RoCoF_{500}$ is measured over 500 ms after the event, $f_{\min}$ is the nadir and t_{sat} is the time the converter

spends at its limit. The CMDP is $\max_{\theta} E[R]$ subject to $E[C] \leq d_C$, with $d_C = 0.002$.

3. Model-based shield

When the event is detected, the controller estimates the imbalance from the swing equation using its own measured power, $\Delta \hat{P}(t) = \max_{\tau \leq t} [\Delta P_B(\tau) - 2\hat{H}_s \cdot \dot{f}(\tau) - \Delta f(\tau)]$, and enforces the closed-form SFR bounds

$$H_v \geq f_0 \Delta \hat{P} / (2 \cdot \text{RoCoF}_{\text{lim}} \cdot \kappa_H) - \hat{H}_s \quad (6)$$

$$(t_{\text{ev}} < 1.5 \text{ s}), \quad D_v \geq f_0 \Delta \hat{P} / (\Delta f_{\text{lim}} \cdot \kappa_D) - \hat{\beta}$$

with $\kappa_H = 0.85$, $\kappa_D = 0.8$, $\text{RoCoF}_{\text{lim}} = 0.5 \text{ Hz/s}$ and $\Delta f_{\text{lim}} = 0.5 \text{ Hz}$. The bounds are capped at the actuator limits. The first bound follows from the initial RoCoF of the equivalent system with inertia $H_s + H_v$. The second is the steady-state requirement $f_0 \Delta P / (\beta + D_v) \leq \Delta f_{\text{lim}}$. Their accuracy depends on $\hat{H}_s$. The intervention magnitude per step is $I = (H_{v,\text{shield}} - H_v)/2 + (D_{v,\text{shield}} - D_v)/8$, averaged over the episode.

4. Lagrangian evolution strategy with intervention cost

The policy is a multilayer perceptron (6-16-2, tanh; 146 parameters). It is trained with an evolution strategy (ES) [19], which estimates the gradient from antithetic parameter perturbations and is well suited to vectorized simulation and episodic constraints. Each generation evaluates 48 perturbed policies ($\sigma = 0.1$) on the same eight random scenarios (common random numbers). The fitness of each policy is

$$F = \bar{R} - \lambda_C \cdot \bar{C} - \lambda_I \cdot \bar{I}, \quad \lambda_C \leftarrow [\lambda_C + \alpha(\bar{C} - d_C)]^+, \quad \lambda_I \leftarrow [\lambda_I + \alpha(\bar{I} - d_I)]^+ \quad (7)$$

with rank-based fitness shaping, Adam updates (learning rate 0.03), $\alpha = 0.5$ and $d_I = 0.02$. Training runs for 400 generations (153 600 episodes), and the error of $\hat{H}_s$ is randomized with a standard deviation of 15 % instead of 8 %, so that the shield is sometimes wrong during training (domain randomization [20]). All environments of a generation are simulated in parallel in NumPy, so one training run takes about 145 s on one CPU core.

5. Benchmarks and test sets

Seven controllers were compared: (1) a fixed VSM with $H_v = 3 \text{ s}$ and $D_v = 10 \text{ p.u.}$; (2) a rule-based adaptive law, $H_v = H^*(1 + 0.6 \cdot \tanh(\Delta f \cdot \dot{f}/\epsilon))$ and $D_v = D^*(1 + 0.5 \cdot \tanh(|\Delta f|/\delta))$, with H^* , D^* computed from the SFR bounds for a 0.1 p.u. design disturbance [5, 6]; (3) shield only, with minimum gains corrected by the shield; (4) unconstrained RL ($\lambda_C = \lambda_I = 0$); (5) Lagrangian RL without a shield; (6) Lagrangian RL with a shield but without the intervention cost; and (7) the proposed controller. RL controllers 4–7 were trained with three random seeds and are reported as mean $\pm$ standard deviation. Two test sets of 500 scenarios each were used: in-distribution (ID), drawn from the training ranges with an 8 % inertia-estimation error; and out-of-distribution (OOD), with η up to 0.92, ΔP_L up to 0.15 p.u. (above the converter rating) and a 20 % estimation error. Table 1 summarizes the settings.

Table 1. Simulation and training settings

Item	Setting
Plant parameters	$H_0 \in [4, 6] \text{ s}$, $\eta \in [0.40, 0.85]$, $D \in [0.5, 1.5]$, $R = 0.05$, $T_g = 0.2 \text{ s}$, $T_t = 0.3 \text{ s}$

BESS	$P_{\max} = 0.12$ p.u., $T_c = 20$ ms, $H_v \in [0.45, 5.4]$ s, $D_v \in [0.4, 21.7]$ p.u.
Disturbance	$\Delta P_L \in [0.03, 0.12]$ p.u. (OOD: up to 0.15) at $t = 0.2$ s
Limits	$\text{RoCoF}_{500} \leq 0.5$ Hz/s, $f_{\min} \geq 49.5$ Hz, no saturation
Control / integration step	20 ms / 2 ms, episode 10 s
Policy	MLP 6-16-2, tanh, 146 parameters
ES	48 members (antithetic), $\sigma = 0.1$, Adam lr 0.03, 8 scenarios/generation, 400 generations
Dual updates	$\alpha = 0.5$, $d_C = 0.002$, $d_I = 0.02$
Estimation error $\hat{H}_s$	training 15 %, ID test 8 %, OOD test 20 % (standard deviation)

RESULTS

1. Training

Fig. 2 shows the training curves. Unconstrained RL quickly converged to the highest return by setting the gains near their minimum, and violated the limits in about 75 % of validation

episodes. Lagrangian RL reduced the violation rate to 2–8 % but did not eliminate it, because the CMDP constrains only the expected cost. Both shielded agents had zero violations throughout training. Without the intervention cost, however, the shield had to correct the agent with an average magnitude $I \approx 0.40$. With the intervention cost, I fell below 0.01 within 50 generations, and λ_I settled at 0.6–0.8.

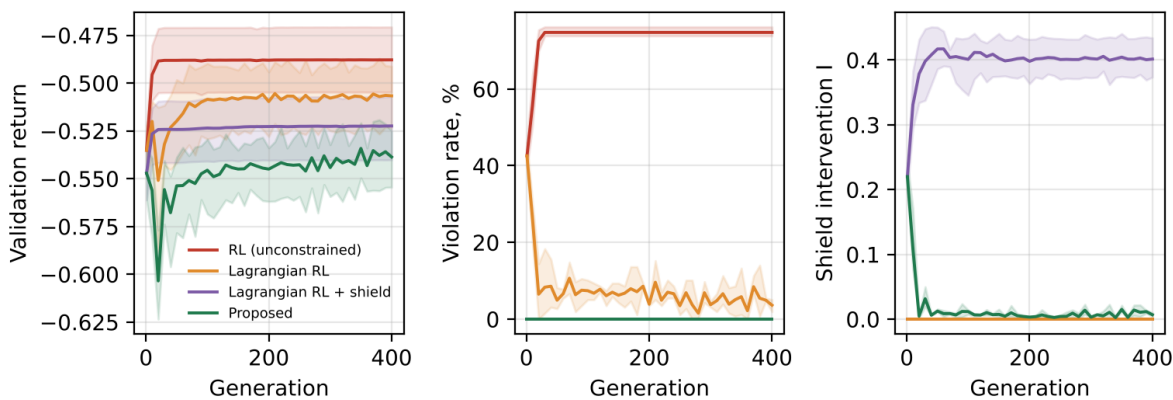


Fig. 2. Validation return, violation rate and shield-intervention magnitude during training (mean and range over three seeds).

2. In-distribution performance

Table 2 lists the results. The proposed controller produced no frequency-limit violation for any seed. Its maximum RoCoF_{500} was 0.368

Hz/s, a 26 % margin, and its lowest nadir was 49.57 Hz. Compared with the fixed VSM, which violated the RoCoF limit in 3.2 % of scenarios, it used 32 % less BESS energy (0.297 against 0.436

p.u.s). Compared with the rule-based law it used 17 % less energy and kept a larger RoCoF margin (0.368 against 0.451 Hz/s), but its integrated frequency deviation was 18 % larger. The shield-only controller used even less energy (0.237), but it allowed a 16 % larger frequency deviation and depended on the shield in 64 % of control steps, against 6.4 ± 1.3 % for the proposed controller. Lagrangian RL without a shield used the least

energy apart from unconstrained RL, but violated the nadir limit in 4.3 ± 1.6 % of scenarios. The shielded agent without the intervention cost converged to the shield-only behaviour (64 % of steps corrected). This confirms the shield-exploitation effect that the intervention constraint is designed to remove. Fig. 4 (left) places the controllers in the energy–frequency-deviation plane.

Table 2. Test results (RL controllers: mean $\pm$ s.d. over three seeds; s.d. omitted when below 0.05 %)

Controller	Viol. ID, %	Energy ID, p.u.s	$\int \Delta f dt$ ID, Hz.s	Max RoCoF ID, Hz/s	Shield active, % steps	Freq. viol. OOD, %	Saturation OOD, %
Fixed VSM	3.2	0.436	2.06	0.547	–	19.0	1.4
Rule-based adaptive	0	0.356	2.37	0.451	–	6.6	4.4
Shield only	0	0.237	3.26	0.447	64.0	4.2	5.0
RL, unconstrained	80.4	0.046	4.81	1.987	–	85.0	0
Lagrangian RL	4.3 ± 1.6	0.196	3.50	0.492	–	18.6 ± 2.6	4.6
Lagrangian RL + shield	0	0.234	3.28	0.423	63.6	3.3 ± 0.3	4.0
Proposed	0	0.297	2.80	0.368	6.4 ± 1.3	2.1 ± 0.2	6.9

3. Learned behaviour

Fig. 3 shows a representative event ($\eta = 82$ %, $\Delta P_L = 0.11$ p.u.). Before the event, the proposed policy kept the gains low. Within the first 20–40 ms after detection it raised H_v to about 5.4 s and D_v to about 21 p.u., which is close to the actuator limits. It then reduced D_v to about 12 p.u. as the governors took over and the frequency

approached its quasi-steady value. This learned emergency-boost-and-release pattern resembles the alternating-inertia principle [5], but it is applied to damping as well and is timed by the policy. Lagrangian RL without a shield reacted more slowly and reached a lower nadir. The fixed VSM kept its gains constant and used the most energy.

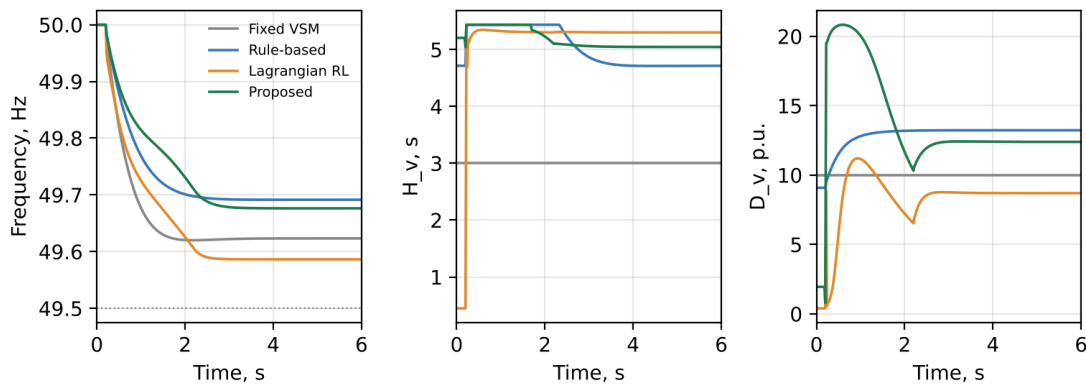


Fig. 3. Frequency, virtual inertia and damping during a 0.11 p.u. loss of generation at $\eta = 82\%$.

4. Out-of-distribution robustness

Fig. 4 (right) and Table 2 report the OOD results. Because some OOD disturbances exceed the converter rating, saturation is unavoidable there and is reported separately from frequency-limit violations. The proposed controller had the lowest frequency-limit violation rate ($2.1 \pm 0.2\%$). The other rates were 3.3 % for the shielded agent without intervention cost, 4.2 % for shield only, 6.6 % for the rule-based law, 18.6 % for

Lagrangian RL and 19.0 % for the fixed VSM. Its saturation rate was the highest (6.9 %), because it is the most willing to use the full converter capability in severe events. With the shield removed, the proposed agent performed identically (0 % ID and 2.1 % OOD violations). Its safety had therefore been internalized during training, and the shield acts only as a backstop. One policy decision took $3.3 \mu\text{s}$ in NumPy on one CPU core, four orders of magnitude below the 20 ms control period.

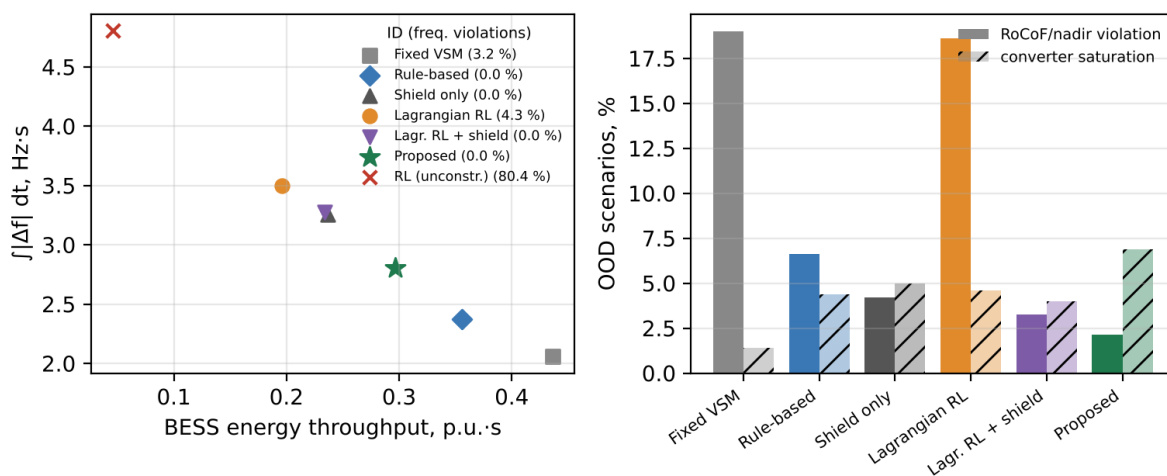


Fig. 4. Left: BESS energy versus integrated frequency deviation on the ID test set (legend: frequency-limit violation rate). Right: frequency-limit violations and converter saturation on the OOD test set.

DISCUSSION

The experiments separate three failure modes of learning-based frequency control. Unconstrained RL exploits the absence of a safety signal and minimizes energy at the expense of security. Lagrangian RL satisfies the constraint on average but still violates it in individual scenarios, which is unacceptable for grid codes that apply to every event. A shield alone removes violations within its model's validity, but an agent trained behind it becomes dependent on it. Because the shield relies on the inertia estimate, it shares that estimate's errors. The intervention cost turns the shield from a crutch into a teacher: the agent is rewarded for staying away from the shield's boundary. Combined with domain randomization, this taught the agent to keep a margin where the shield is likely to be wrong, which explains the lower OOD violation rate.

The proposed controller is not uniformly better. It used more energy than shield-only or Lagrangian RL control, and it allowed a larger frequency deviation than the rule-based law. Its operating point reflects the chosen return weights and the conservative training distribution. A different weighting would trade energy for frequency quality, and the Lagrangian framework allows such priorities to be set explicitly. The evolution strategy was chosen for its simplicity, robustness to episodic constraints and ease of vectorization. Gradient-based safe-RL algorithms such as constrained policy optimization [12] or PPO-Lagrangian [13] could be used instead, and might be more sample-efficient.

The study has limitations. (i) The plant is a single-area aggregated model without network, voltage or phase-locked-loop dynamics. Interactions among several learning controllers (cf. multi-unit coordination) were not considered. (ii) The shield bounds are derived from the SFR model and are only as reliable as $\hat{H}_s$ and $\hat{\beta}$. (iii) The state of charge was not modelled. (iv) No formal guarantee is given for scenarios outside the model class. (v) Validation was simulation-based, and hardware-in-the-loop testing with a real-time controller is the next step. The policy's small size (146 parameters) and microsecond inference make embedded deployment realistic.

CONCLUSIONS

A safe-RL controller that updates the virtual inertia and damping of a grid-forming BESS every 20 ms was developed. It combines a Lagrangian evolution strategy, a model-based shield and an explicit constraint on shield interventions. The main findings are as follows:

1. The proposed controller satisfied the 0.5 Hz/s RoCoF and 49.5 Hz nadir limits in all 500 in-distribution scenarios for all training seeds, with a 26 % RoCoF margin. It used 32 % less BESS energy than a fixed VSM and 17 % less than a rule-based adaptive law.
2. Unconstrained RL (80 % violations) and Lagrangian RL without a shield (4.3 %) were not safe enough. A shielded agent without an intervention cost relied on the shield in 64 % of control steps.

3. The intervention constraint reduced shield activity to 6.4 % of steps, and the agent remained safe even with the shield removed.

4. Out of distribution, the proposed controller had the lowest frequency-limit violation rate (2.1 % against 3.3–19.0 %), and its inference time ($\approx 3 \mu\text{s}$) is compatible with real-time embedded control.

DECLARATIONS

Funding: this research received no external funding. Conflict of interest: the authors declare no conflict of interest. Data and code availability: the Python code used to produce all results is available from the corresponding author on reasonable request. Author contributions: O. A. Kayumov — conceptualization, methodology, software, formal analysis, writing – original draft; N. R. Kayumova — validation, visualization, data curation, writing – review and editing.

REFERENCES

1. F. Milano, F. Dörfler, G. Hug, D. J. Hill, G. Verbič, "Foundations and challenges of low-inertia systems," in Proc. Power Systems Computation Conference (PSCC), Dublin, 2018, pp. 1–25.
2. F. Dörfler, D. Groß, "Control of low-inertia power systems," Annual Review of Control, Robotics, and Autonomous Systems, vol. 6, pp. 415–445, 2023.
3. Q.-C. Zhong, G. Weiss, "Synchronverters: Inverters that mimic synchronous generators," IEEE Trans. Ind. Electron., vol. 58, no. 4, pp. 1259–1267, 2011.
4. U. Tamrakar, D. Shrestha, M. Maharjan, B. P. Bhattarai, T. M. Hansen, R. Tonkoski, "Virtual inertia: Current trends and future directions," Applied Sciences, vol. 7, no. 7, art. 654, 2017.
5. J. Alipoor, Y. Miura, T. Ise, "Power system stabilization using virtual synchronous generator with alternating moment of inertia," IEEE J. Emerg. Sel. Topics Power Electron., vol. 3, no. 2, pp. 451–458, 2015.
6. X. Hou, Y. Sun, X. Zhang, J. Lu, P. Wang, J. M. Guerrero, "Improvement of frequency regulation in VSG-based AC microgrid via adaptive virtual inertia," IEEE Trans. Power Electron., vol. 35, no. 2, pp. 1589–1602, 2020.
7. M. Glavic, "(Deep) reinforcement learning for electric power system control and related problems: A short review and perspectives," Annual Reviews in Control, vol. 48, pp. 22–35, 2019.
8. X. Chen, G. Qu, Y. Tang, S. Low, N. Li, "Reinforcement learning for selective key applications in power systems: Recent advances and future challenges," IEEE Trans. Smart Grid, vol. 13, no. 4, pp. 2935–2958, 2022.
9. W. Cui, Y. Jiang, B. Zhang, "Reinforcement learning for optimal primary frequency control: A Lyapunov approach," IEEE Trans. Power Syst., vol. 38, no. 2, pp. 1676–1688, 2023.
10. J. García, F. Fernández, "A comprehensive survey on safe reinforcement learning," J. Mach. Learn. Res., vol. 16, pp. 1437–1480, 2015.
11. E. Altman, Constrained Markov Decision Processes. Boca Raton: Chapman & Hall/CRC, 1999.
12. J. Achiam, D. Held, A. Tamar, P. Abbeel, "Constrained policy optimization," in Proc. 34th Int. Conf. Machine Learning (ICML), 2017, pp. 22–31.
13. A. Ray, J. Achiam, D. Amodei, "Benchmarking safe exploration in deep reinforcement learning," OpenAI Technical Report, 2019.

14. M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, U. Topcu, "Safe reinforcement learning via shielding," in Proc. AAAI Conf. Artificial Intelligence, 2018, pp. 2669–2678.
15. K. P. Wabersich, M. N. Zeilinger, "A predictive safety filter for learning-based control of constrained nonlinear dynamical systems," *Automatica*, vol. 129, art. 109597, 2021.
16. A. D. Ames, S. Coogan, M. Egerstedt, G. Notomista, K. Sreenath, P. Tabuada, "Control barrier functions: Theory and applications," in Proc. European Control Conference (ECC), 2019, pp. 3420–3431.
17. P. M. Anderson, M. Mirheydar, "A low-order system frequency response model," *IEEE Trans. Power Syst.*, vol. 5, no. 3, pp. 720–729, 1990.
18. P. Kundur, *Power System Stability and Control*. New York: McGraw-Hill, 1994.
19. T. Salimans, J. Ho, X. Chen, S. Sidor, I. Sutskever, "Evolution strategies as a scalable alternative to reinforcement learning," arXiv:1703.03864, 2017.
20. J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), 2017, pp. 23–30.