Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Physics-informed Actor-Critic for Coordination of Virtual Inertia from Power Distribution Systems

Simon Stock | **Davood Babazadeh** | **Sari Eid** | **Christian Becker**

[1]Institute of Power and Energy Technology, Hamburg University of Technology, Hamburg, Germany

Correspondence
Corresponding author Simon Stock
Email: simon.stock@tuhh.de

Author contributions
Simon Stock: Conceptualization, Data curation, Investigation, Methodology, Resources, Software, Visualization, Writing – original draft, editing
Davood Babazadeh: Conceptualization, Investigation, Project administration, Supervision, Writing – review & editing
Sari Eid: Data curation, Investigation, Software
Christian Becker: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing

Funding Information
Publishing fees supported by Funding Programme Open Access Publishing of Hamburg University of Technology (TUHH).

Abstract

The vanishing inertia of synchronous generators in transmission systems requires the utilization of renewables for inertial support. These are often connected to the distribution system and their support should be coordinated to avoid violation of grid limits. To this end, this paper presents the Physics-informed Actor-Critic (PI-AC) algorithm for coordination of Virtual Inertia (VI) from renewable Inverter-based Resources (IBRs) in power distribution systems. Acquiring a model of the distribution grid can be difficult, since certain parts are often unknown or the parameters are highly uncertain. To favor model-free coordination, Reinforcement Learning (RL) methods can be employed, necessitating a substantial level of training beforehand. The PI-AC is a RL algorithm that integrates the physical behavior of the power system into the Actor-Critic (AC) approach in order to achieve faster learning. To this end, we regularize the loss function with an aggregated power system dynamics model based on the swing equation. Throughout this paper, we explore the PI-AC functionality in a case study with the CIGRE 14-bus and IEEE 37-bus power distribution system in various grid settings. The PI-AC is able to achieve better rewards and faster learning than the exclusively data-driven AC algorithm and the metaheuristic Genetic Algorithm (GA).

KEYWORDS
Physics-informed Machine Learning, Reinforcement Learning, Virtual Inertia, Power Distribution Systems, Frequency Dynamics

1 | INTRODUCTION

In recent years, renewable energy resources have replaced conventional generation to reduce the carbon footprint of energy production. Their dynamic characteristics are substantially different from conventional generation, since many of them are connected via power electronics, i.e. inverters. These lack inherent inertia in case of frequency events, in contrast to synchronous generators. This leads to higher vulnerability of the power systems, which can be partially compensated for by the immediate release of power from renewables via the inverter. The corresponding technique is termed Virtual Inertia (VI).

Most renewable energy is connected to the distribution system, while conventional generation tends to be connected to the transmission system [1]. This can cause bidirectional power flows and violation of grid limits as a consequence. Thus, Distribution System Operators (DSOs) have been increasing the number of measurement devices in their grids to make full use of their operational capabilities and avoid structural grid extensions. These can be utilized in frameworks for the coordination of VI in distribution grids [2, 3].

However, it is difficult to apply commonly used coordination algorithms, since there is often no distribution system model available. For that reason, model-based optimization can rarely be applied and the optimality of the corresponding results is questionable when inaccurate models are used. This puts focus on model-free techniques. There has been considerable work on metaheuristic approaches, such as the Genetic Algorithm (GA) and Particle Swarm Optimizer (PSO) [4]. Although these approaches are able to optimize a given objective in a model-free manner, they do not obtain a strategy. Thus, they have to be rerun as soon as there are slight changes to the model or the objective. This brings up RL techniques, such as Actor-Critic (AC), which seek to obtain a coordination policy through offline

learning. However, offline training still requires substantial amounts of time [3].

In this paper, we strive to improve the training performance of an Actor-Critic (AC) approach that is model-free by design. To this end, we extend the AC training loss by a physics-regularized term driven by a generic formulation for power system dynamics, the swing equation. This approach, the PI-AC, is based on the idea that knowledge of the system can improve learning performance. Physics-based regularization has shown promising results in previous work applied to other Machine Learning (ML) approaches [5, 6]. Overall, loss regularization is widely used in ML, such as lasso and ridge regularization. We consider the PI-AC to be model-free despite being driven by a physics-regularized loss, because the utilized physics formulation is generic for all power systems and does not require specific knowledge of the individual grid structure or setting parameters.

In the literature, there exists considerable work on similar approaches, but most of them rely on an additional Physics-informed Neural Network (PINN) to operate as a transition model. In [5] a RL approach is proposed that follows a similar idea. However, the swing equation is utilized in a PINN transition model not as an additional regularization term. This approach increases the overall computational cost, since an additional PINN model has to be trained. A similar approach is proposed in [7] for non-power system-related problems. In contrast to [7, 5], the proposed PI-AC seeks to regularize the loss function of the critic to influence the quality function. This does not cause considerable additional computation cost, since the number of trained networks remains the same.

In conclusion, this paper proposes the PI-AC approach, which demonstrates superior performance with fewer training iterations compared to the AC and GA approach for coordination of VI from distribution systems.

This paper is structured as follows: section 2 presents the problem formulation and in the following the PI-AC methodology. It also introduces the fundamentals of the GA which is used for comparison in this paper. This section is followed by the specification of the problem formulation for a power system section 3, which includes the relaxation of the given problem through penalty functions. In section 4 we present the models that are used in the case study with the CIGRE 14-bus and IEEE 37-bus model and the corresponding equations. This section also presents the scenarios that will be used for evaluation in this paper. Section 5 presents the results for all scenarios and compares the approaches. This section closes with a discussion on specific aspects that we found throughout the investigation. Section 6 closes this paper and concludes the main outcomes.

2 | METHODOLOGY

In this paper, we strive to coordinate a dynamical system that can generally be described by a set of differential equations as follows

$$\dot{\boldsymbol{x}} = g(\boldsymbol{u}, \boldsymbol{x}; \boldsymbol{\lambda}). \tag{1}$$

The system states are described by $\boldsymbol{x}$ and its derivatives by $\dot{\boldsymbol{x}}$, $\boldsymbol{u}$ are the system inputs and $\boldsymbol{\lambda}$ the system parameters. Operator g maps the inputs and parameters to the system states. To establish a model-free optimization problem, we describe the system as a black-box environment that produces a set of measurable states $\boldsymbol{s} \in \mathcal{S}$ under an applied set of actions $\boldsymbol{a} \in \mathcal{A}$. These states can be a subset of system states $\boldsymbol{x}$ and the actions $\boldsymbol{a}$ can be a subset of inputs $\boldsymbol{u}$. Actions are applied in order to optimize a given objective

$$\min_{\boldsymbol{a}} f(\cdot; \boldsymbol{s}). \tag{2}$$

In this paper, we focus on the model-free coordination of VI in power distribution systems, which can be set up in a form similar to eq. (2). In the following, we will derive the PI-AC which seeks to optimize the given objective f through learning that is regularized by an approximation of eq. (1) for faster learning.

2.1 | Fundamentals of Physics-informed Actor-Critic

Most RL techniques follow a model-free approach that considers the system as a black-box environment. To this end, we extend the previous formulation to a Markov decision process $(\mathcal{S}, \mathcal{A}, \mathcal{R}, p)$ with reward signal $r \in \mathcal{R}$ based on the objective f. The probability to transition from a given state and action to another state is given by p. The goal of RL is to find the optimal policy $\pi(\boldsymbol{s})$ to control the environment to obtain the best objective.[‡] The policy learning itself is primarily guided by a quality function. This is learned separately using the reward r based on the objective f. Thus, the policy learning is driven by the objective f.

This quality function can be represented by a Neural Network (NN) parameterized by a set of weights and biases $\boldsymbol{\Theta}^Q$. The NN is called the critic, since it criticizes the actions taken by the current policy. The objective is incorporated into the learning of the quality function through the reward with $r = -f$

$$\mathcal{L}_{\text{critic}} = \mathbb{E}[(Q(\boldsymbol{s}_k, \boldsymbol{a}_k; \boldsymbol{\Theta}^Q) - (r + \gamma Q(\boldsymbol{s}_{k+1}, \boldsymbol{a}_{k+1}; \boldsymbol{\Theta}^Q)))^2]. \tag{3}$$

[‡] This is different for metaheuristic methods, which obtain a given set of parameters for the current environment and settings. They must retrain when significant changes are made in the environment or settings.

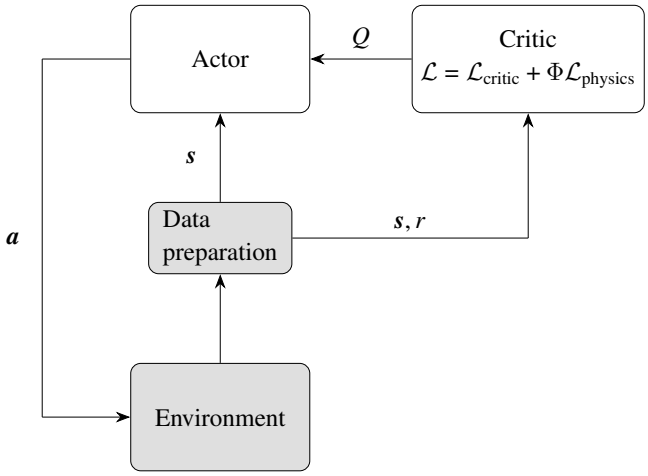

FIGURE 1 Schematic representation of the proposed PI-AC structure

Parameter γ represents a discount factor that weights future expected rewards. The parameters $\boldsymbol{\Theta}^Q$ are adapted based on $\mathcal{L}_{\text{critic}}$. The described approach is called the AC. In the following, we extend it to the PI-AC.

Physics regularized learning: In this paper, we aim to regularize the learning using the physical knowledge of the system. The goal is to achieve better performance, i.e. better objectives and faster learning. To this end, we augment the above equation eq. (3) with a system model that approximates eq. (1)

$$\mathcal{L}_{\text{physics}} = (\dot{\boldsymbol{x}} - d(\boldsymbol{u}, \boldsymbol{x}; \hat{\boldsymbol{\lambda}}))^2. \tag{4}$$

This equation will approach zero if the estimates of the system parameters $\hat{\boldsymbol{\lambda}}$ are accurate. It is based on the measured system states $\boldsymbol{x}$ and a surrogate system model d that utilizes the inputs $\boldsymbol{u}$ and the estimated system parameters $\hat{\boldsymbol{\lambda}}$. The estimates $\hat{\boldsymbol{\lambda}}$ are given by the critic. Note that d approximates g, but does not necessarily has the same structure or rank. Instead, it can be a simplified representation of the system.

Ultimately, this leads to the following critic loss formulation

$$\mathcal{L} = \mathcal{L}_{\text{critic}} + \Phi\mathcal{L}_{\text{physics}} \tag{5}$$

with weighting factor Φ. The critic steers the learning of the policy, which itself can be represented by a NN that is termed the actor.

We call the proposed approach the Physics-informed Actor-Critic (PI-AC) since it augments the Actor-Critic (AC) algorithm with a physics-informed loss term.[§] This structure is schematically shown in fig. 1. The parameterized actor function $\mu(\boldsymbol{s}, \boldsymbol{\Theta}^\mu)$ represents the current policy. This NN is updated by applying the chain rule to the expected return from the initial distribution, with respect to the actor parameters. The resulting policy score function is calculated as follows

$$\nabla_{\boldsymbol{\Theta}^\mu} J \approx \mathbb{E}[\nabla_{\boldsymbol{a}} Q(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{\Theta}^Q)|\nabla_{\boldsymbol{\Theta}^\mu}\mu(\boldsymbol{s}, \boldsymbol{\Theta}^\mu)]. \tag{6}$$

Note that this formulation relies on the quality function, i.e. the critic network, thus, policy learning is driven by the quality function and ultimately by the reward in eq. (3).

Throughout this paper, we apply the Deep Deterministic Policy Gradient (DDPG) algorithm for learning. NNs are used to approximate both the actor and critic functions, which allows for learning in large state and action spaces. For efficient computation, learning can be performed in minibatches rather than online. This issue is addressed through a replay buffer. This buffer is a finite sized cache $\mathcal{R}$, which stores the tuple $(\boldsymbol{s}_k, \boldsymbol{a}_k, r_k, \boldsymbol{s}_{k+1})$ after environment exploration. Once the buffer has reached its maximum capacity, the oldest sample is discarded to make room for the new one. The DDPG introduces duplicates of the actor and critic networks to ensure stable learning, namely $Q'(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{\Theta}^{Q'})$ and $\mu'(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{\Theta}^{\mu'})$, which are used to compute target values and are thus referred to as target networks. The network parameters $\boldsymbol{\Theta}^{Q'}$ and $\boldsymbol{\Theta}^{\mu'}$ are updated utilizing a soft update parameter τ. This parameter slowly tracks them back to the learned networks $\boldsymbol{\Theta}' \leftarrow \tau\boldsymbol{\Theta} + (1-\tau)\boldsymbol{\Theta}'$. It is recommended to set $\tau << 1$ for stable convergence. The target values are limited to gradual modifications, which facilitates steady learning and, consequently, produces stable targets [9]. Exploring continuous action spaces can be difficult. Off-policy algorithms, such as DDPG, are beneficial, as they decouple exploration and learning. The exploration policy can thus be build by adding random samples from a noise process following $\mathcal{N}$

$$\mu_{ex}(\boldsymbol{s}_k) = \mu(\boldsymbol{s}_k, \boldsymbol{\Theta}_k^\mu) + \mathcal{N}. \tag{7}$$

For $\mathcal{N}$ an Ornstein-Uhlenbeck process is recommended for generation of temporally correlated exploration centered around 0.

Physics regularized learning for power system dynamics: In the context of power systems dynamics, the Single Machine Infinite Bus (SMIB) model should be chosen as the surrogate model d that regularizes learning in eq. (4). This model is a generic aggregated description of the dynamics of every power system, thus, the learning is still model-free, since no previous knowledge of the actual system is required. It can be described by [10]

$$\dot{\delta} = \Delta\omega \tag{8}$$

$$\Delta\dot{\omega} = \frac{1}{2\hat{H}}(P_m - \hat{D}\Delta\omega - \frac{\hat{EV}}{X_d}\sin(\delta - \varphi)) \tag{9}$$

with system parameter estimates $\hat{\boldsymbol{\lambda}} = \{\hat{H}, \hat{D}, \frac{\hat{EV}}{X_d}\}$. These represent the overall system inertia H, the damping D, the synchronous voltage E, the terminal voltage V and the synchronous

[§] In contrast to the PINN [8], we are not using the critic network to function as a surrogate model itself estimating the system states, but rather rely on the measurements of $\boldsymbol{x}$.

reactance X_d. The states of the system are the frequency deviation $\Delta\omega$ and the rotor angle δ. The terminal voltage angle is described by φ. The parameter estimates $\hat{\boldsymbol{\lambda}}$ are determined by the critic, which therefore provides four outputs instead of one, namely the Q and the system parameter estimates $\hat{\boldsymbol{\lambda}}$. This requires tuning the critic parameters $\boldsymbol{\Theta}^Q$, so the obtained Q and $\hat{\boldsymbol{\lambda}}$ lead to a small overall loss eq. (5).

2.2 | Fundamentals of Genetic Algorithm

The GA is a metaheuristic optimization technique that seeks to minimize a given objective function. It is based on the fundamental concept that a group of system realizations, called the parents, are randomly placed throughout the search space. These are iteratively updated throughout the generations following genetic rules until convergence is reached. The set of parents can be expressed as

$$A = \{\boldsymbol{a}_1, \boldsymbol{a}_2, ...\boldsymbol{a}_N\} \tag{10}$$

with N parameters to influence the fitness, i.e. objective function $f(\cdot; \boldsymbol{a}_i)$. After evaluating the corresponding fitness functions of the initial parameters, they are sorted after the fitness achieved. A high probability $p(\boldsymbol{a}_i)$ is assigned to those who achieve a high fitness. Mutation, replication, and crossover are employed to update the realizations and find the next generation [11].

3 | COORDINATION FRAMEWORK AND OBJECTIVE

In this paper, we strive to provide VI from the distribution system for frequency support in a coordinated way. The superordinate framework was proposed in previous work [3]. It is based on the assumption that the Transmission System Operator (TSO) requests a certain amount of inertia that must be provided by the distribution system. The DSO is obligated to coordinate the provision considering economic and system operational aspects. To this end, the VI plant setpoints can be optimized using a certain objective. The framework can be concluded by the following steps:

1. TSO evaluates the system dynamics and determines the required amount of additional inertia and damping.
2. TSO optimizes the inertia provision structure of the overall grid and assigns set points to specific regions.
3. DSO receives the set points, i.e. inertia and damping budget for its region.
4. DSO optimizes the provision considering certain preferences in its grid and assigns the corresponding set points to the VI plants.

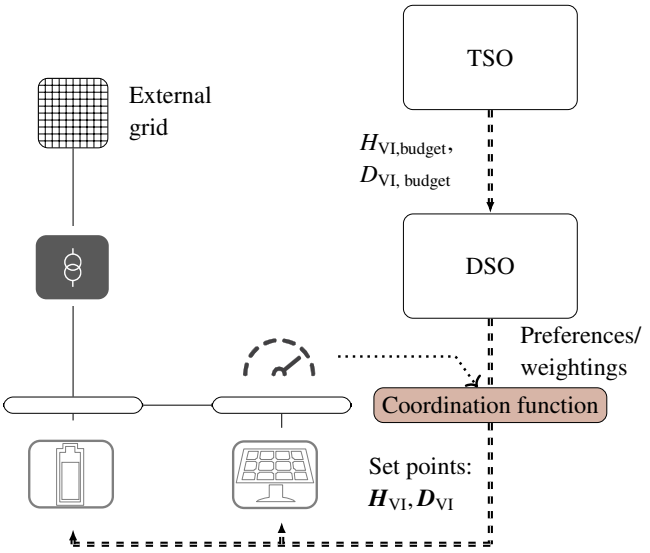

FIGURE 2 The inertia support framework surrounding the coordination function

This structure is also visualized in fig. 2.

Often times, there is no accurate system model available of the distribution system. To this end, we apply the model-free PI-AC approach. The optimization task can be written as follows, taking into account various VI resources in the distribution system that can be controlled through their inertia and damping constants $H_{\text{VI},i}$ and $D_{\text{VI},i}$:

$$\min_{H_{\text{VI},i}, D_{\text{VI},i}} f(C, \xi) \tag{11}$$

In this formulation, C represents the economic costs and ξ the system operational costs. The latter can be defined by the operator, for example, voltage deviation or transformer loading. In this paper, we set the economic costs to [3]

$$C = \sum c_i P_{\text{max},i} \tag{12}$$

and $P_{max,i}$ represents the maximum power provided by an individual plant considering the maximum expected Rate of Change of Frequency (ROCOF) and frequency deviation. These are weighted in the sum by an individual cost factor c_i, which is randomly chosen in the range of 0.5 to 1.5. The system operational costs consist of the voltage deviation at and individual bus V_i from the nominal value $V_{i,set}$, as follows

$$\xi = \sum c_{V,i}(|V_i| - |V_{i,set}|). \tag{13}$$

We constrain eq. (11) by the given inertia and damping budget, the available maximum inertia and damping per plant and the voltage limits. This extends the previous formulation eq. (11),

so

$$\begin{aligned} &\min_{H_{\mathrm{VI},i}, D_{\mathrm{VI},i}} f(C, \xi) \\ \text{s.t.} \quad & H_{\mathrm{VI,sys}} \geq H_{\mathrm{VI,budget}} \\ & D_{\mathrm{VI,sys}} \geq D_{\mathrm{VI,budget}} \\ & 0 \leq H_{\mathrm{VI},i} < \frac{P_{\max,i}}{2\dot{\Delta}\omega_{\max}} \\ & 0 \leq D_{\mathrm{VI},i} < \frac{P_{\max,i}}{\Delta\omega_{\max}} \\ & \Delta V_i \leq \Delta V_{\max} \end{aligned} \tag{14}$$

The VI provided from the distribution system $H_{\mathrm{VI,sys}}$ is calculated based on the sum of the individual contributions $H_{\mathrm{VI},i}$ weighted by the corresponding nominal power and the system power. The damping is calculated in a similar way. For this formulation, we assume that the maximum ROCOF and maximum frequency deviation will not appear at the same time.

The PI-AC and RL techniques in general are mostly guided by the reward function, thus, it is difficult to integrate constraints into the learning process. In this paper, we apply the penalty method, which relaxes the constraints by augmenting the objective with a corresponding penalty function. This additionally prevents the PI-AC from being trapped in local minima that are close to the boundary of the feasible space. Allowing the algorithm to briefly step into the infeasible space tends to support broader exploration after all. Given that, the given constraints can be ensured in two ways. First, the penalty can be introduced earlier than the actual physical limit or constraint. Second, the constraints can be manually ensured, thus, a certain action is adapted after training to guarantee compliance with the given limits. In the following, we rely on the second approach, since the first does not guarantee compliance with the constraints. The third and fourth constraints are incorporated by limiting the action space in a corresponding way. All other constraints are used in the objective as penalties

$$\min_{H_{\mathrm{VI},i}, D_{\mathrm{VI},i}} f(C, \xi, p_{H_{\mathrm{budget}}}, p_{D_{\mathrm{budget}}}, p_{\Delta V_{\mathrm{lim}}})$$

s.t.

$$p_{H_{\mathrm{VI,budget}}} = \begin{cases} 0 & \text{if } (H_{\mathrm{VI,budget}} - H_{\mathrm{VI,sys}}) \leq 0 \\ c_H \cdot (H_{\mathrm{VI,budget}} - H_{\mathrm{VI,sys}}) & \text{if } (H_{\mathrm{VI,budget}} - H_{\mathrm{VI,sys}}) > 0 \end{cases}$$

$$p_{D_{\mathrm{VI,budget}}} = \begin{cases} 0 & \text{if } (D_{\mathrm{VI,budget}} - D_{\mathrm{VI,sys}}) \leq 0 \\ c_D \cdot (D_{\mathrm{VI,budget}} - D_{\mathrm{VI,sys}}) & \text{if } (D_{\mathrm{VI,budget}} - D_{\mathrm{VI,sys}}) > 0 \end{cases}$$

$$p_{\Delta V_i} = \begin{cases} 0 & \text{if } (|\Delta V_i - \Delta V_{\max}|) \leq 0 \\ c_{\Delta V} \cdot (|\Delta V_i - \Delta V_{i,\max}|) & \text{if } (|\Delta V_i - \Delta V_{\max}|) > 0 \end{cases}$$

The penalties are described by the corresponding penalty function p. Ultimately, $f(\cdot)$ can be used as a reward in the reinforcement learning task with $r = -f(\cdot)$, so the reward can be maximized.

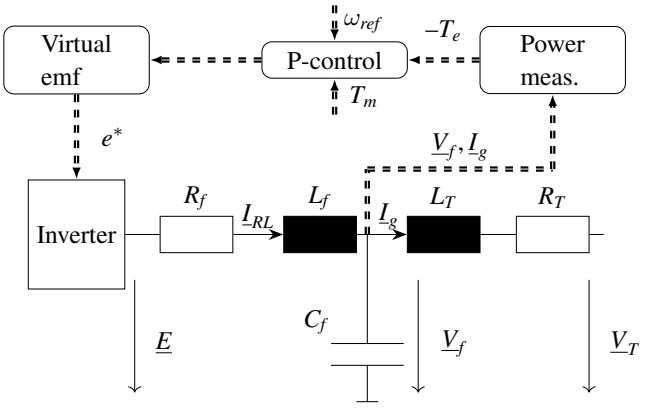

FIGURE 3 Simplified representation of the synchronverter [12]

4 | CASE STUDY

In this paper, we perform a case study based on the CIGRE 14-bus and IEEE 37-bus grid to explore the PI-AC for coordination of multiple VI sources. The superordinate transmission system is represented by an aggregated synchronous generator that is modeled with [10]

$$\dot{\delta} = \Delta\omega \tag{15}$$

$$\Delta\dot{\omega} = \frac{1}{2H_{gen}}(P_m - D_{gen}\Delta\omega - \frac{EV}{X_d}\sin(\delta - \varphi) + P_{gov}) \tag{16}$$

$$\dot{P}_{gov} = -\frac{1}{T_s}(\Delta\omega + P_{gov}). \tag{17}$$

The inertia and damping of the generator are represented by H_{gen} and D_{gen}. The synchronous longitudinal reactance X_d and synchronous voltage E are coefficients to the electrical torque, while the terminal voltage V is assumed to be 1. The corresponding terminal voltage angle is described by φ. Overall, the generator is driven by the mechanical power P_m. In addition, it provides a governor control through P_{gov} with time constant T_s.

In this paper, we draw the energy for VI provision from Battery Energy Storage Systems (BESSs), each BESS has its own inverter, which is operated as a synchronverter [12], as shown in fig. 3.

The RLC-filter and transformer (R_T, L_T) of the model follow the equations

$$\dot{\underline{I}}_{RL} = \frac{1}{L_f}(\underline{E} - \underline{V}_f - R_f\underline{I}_{RL}) \tag{18}$$

$$\dot{\underline{V}}_f = \frac{1}{C_f}(\underline{I}_{RL} - \underline{I}_g) \tag{19}$$

$$\dot{\underline{I}}_g = \frac{1}{L_t}(\underline{V}_f - \underline{V}_T - R_T\underline{I}_g). \tag{20}$$

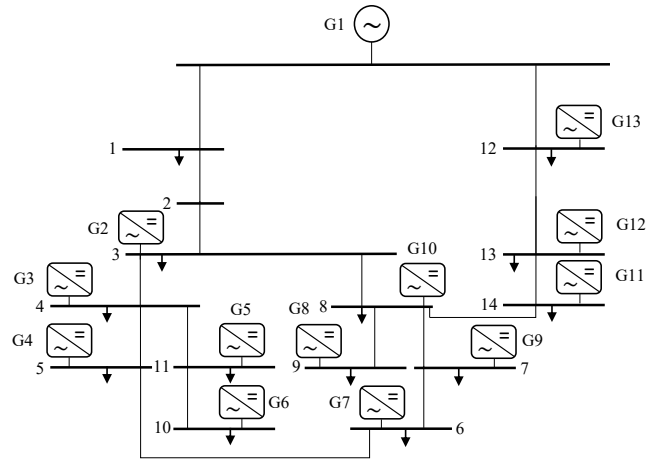

FIGURE 4 CIGRE 14-bus MV system structure

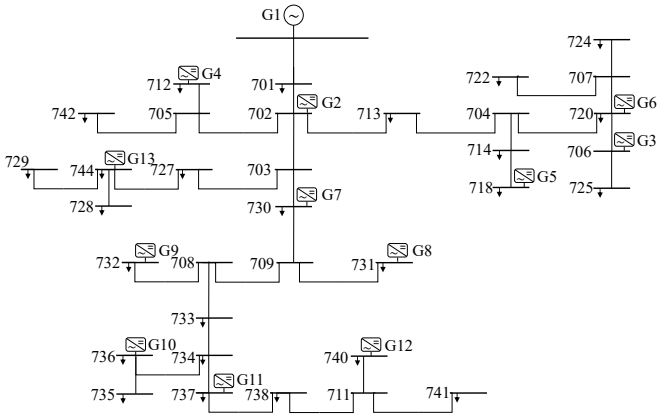

FIGURE 5 IEEE 37-bus MV system structure

The active power control follows the swing equation by calculating a virtual angular frequency ω_c and a virtual angle δ_c [12]

$$\dot{\omega}_c = \frac{1}{2H_c}(T_m - D_p(\omega_{\text{ref}} - \omega_c) - T_e) \tag{21}$$

$$\dot{\delta}_c = \omega_c \tag{22}$$

$$e^* = \dot{\delta}_c M_f i_f \sin(\delta_c). \tag{23}$$

The inertia and damping constants of the synchronverter are described by H_c and D_p respectively, the mechanical and electrical torque by T_m and T_e, the mutual inductance by M_f and the rotor excitation current i_f, which is assumed to be constant [12]. We place twelve VIs, i.e. BESS with synchronverter, in both the CIGRE 14-bus and the IEEE 37-bus grid, as shown in fig. 4 and fig. 5.

For investigation purposes, we simulate different grid situations, with varying shares of VI and different load levels. The individual settings for each scenario are shown in table 1, the

TABLE 1 Evaluation scenarios for test grids

Scenario	H_{TS} in p.u.	D_{TS} in p.u.	$H_{\text{VI,budget}}$ in p.u.	$D_{\text{VI,budget}}$ in p.u.	P_{load} in p.u.
Reference	1.1	0.8	13	13	1.0
$H_{\text{DS, budget}}$ variation	1.1	0.8	8	8	1.0
	1.1	0.8	18	18	1.0
H_{TS} variation	0.9	0.6	13	13	1.0
	1.8	1.5	13	13	1.0
Load variation	1.1	0.8	13	13	0.7
	1.1	0.8	13	13	1.2

transmission system inertia H_{TS} is equal to the generator inertia in eq. (17). The budgets $H_{\text{VI,budget}}$ and $D_{\text{VI,budget}}$ are assigned to the individual VIs by the coordination function, which results in two vectors $\boldsymbol{H}_{\text{VI}}, \boldsymbol{D}_{\text{VI}}$ of 12 individual $H_{\text{VI},i}$, $D_{\text{VI},i}$. We measure the system states $\Delta\omega$ and δ at the generator and provide these as states for the PI-AC training, so $\boldsymbol{s}_k = \{\Delta\omega, \delta\}$. The actions applied are the individual VI parameters, so $\boldsymbol{a}_k = \{\boldsymbol{H}_{\text{VI}}, \boldsymbol{D}_{\text{VI}}\}$ with $H_{\text{VI},i}, D_{\text{VI},i}$ being the parameters for the ith plant. The power system simulations are run for $T = 15$ s and step size $\Delta t = 0.05$ s after reaching a stable operating point, so $\dot{\boldsymbol{x}} = 0$. At that point a step change in the mechanical power of the generator is applied with $P_m = 0.1$ p.u.. This provokes a change in the frequency, thus the VIs that are placed in the distribution system react. The parameters of PI-AC and DDPG respectively are presented in table 2. The Parameters for the benchmark systems, namely the 14-bus and 37-bus system, were taken from [13] and [14].

TABLE 2 List of Deep Deterministic Policy Gradient (DDPG) Hyperparameters

Hyperparameter	Value
Optimizer	ADAM
Actor learning rate	$1 \cdot 6^{-5}$
Critic learning rate	$2 \cdot 6^{-5}$
Discount factor γ	0.995
Experience replay buffer $\mathcal{D}$ size	300
Target step size	50
Repeat times	10
Minibatch size	50
Soft update coefficient τ	$2 \cdot 10^{-4}$
Noise std σ_{Noise}	a_{max}
Actor NN size (neurons)	$\dim(\boldsymbol{s}_k)$; 100; $\dim(\boldsymbol{a}_k)$
Critic NN size (neurons)	$\dim(\boldsymbol{s}_k + \boldsymbol{a}_k)$; 100; 3

5 | RESULTS

We start this evaluation by comparing the overall reward achieved from the three approaches PI-AC, AC and GA. Table 3

shows, that the PI-AC achieves larger rewards in nearly all scenarios compared to AC and GA. This can be found for both grids, the 14-bus and 37-bus. It should be mentioned here, that the achieved rewards are not necessarily comparable between the systems, 14-bus and 37-bus, since the dynamics and therefore the achievable rewards are different. In the results of the 14-bus system, we see that the PI-AC always achieves the best rewards, except for when $H_{\text{DS, budget}}$ is high, meaning that high inertia and damping budgets are assigned. The largest difference can be found in the reference scenario and in high IBR penetration, i.e. low H_{TS}. This is different for the 37-bus system, where the largest difference can be found in low $H_{\text{DS, budget}}$. Additionally, the results of the 37-bus system reveal that PI-AC achieves the final reward faster than the AC and GA in nearly all cases. This is strongly pronounced in the 37-bus reference scenario and low H_{TS}. In the 14-bus system tests, it can also be seen that the PI-AC achieves the final reward slightly faster in low H_{TS} low compared to the reference scenario. This indicates that the higher share of IBRs in these scenarios, i.e. H_{TS} low, results in faster learning of the PI-AC. Overall, the difference in rewards between the PI-AC, AC and GA is smaller in the 37-bus system. This is caused by the larger size of the 37-bus system and thus slower dynamics due to a smaller overall share of IBRs.

5.1 | Comparison of AC and PI-AC

We continue this investigation by comparing the PI-AC and the AC in more detail. These are only different by the augmented loss function, as shown in eq. (5). fig. 6 reveals that the loss of PI-AC is substantially greater. This stems from the additional physics regularization term. It can also be seen that the physics loss is much higher in scenarios with high shares of IBRs. This is caused by the deviation of the swing-equation-based regularization term from behavior of the real system. A large number of IBRs cause the system response to strongly deviate from the swing equation behavior, which is the basis for the physics-informed loss function in PI-AC. This can also be seen in fig. 7, which shows the physics regularization over the critic loss part. Furthermore, increasing IBR penetration causes a combined growth of $\mathcal{L}_{\text{physics}}$ and $\mathcal{L}_{\text{critic}}$. This raises the general question of whether a more influential physics regularization in the loss function improves or hampers the learning of the algorithm. This question can be answered by comparing the reward of both PI-AC and AC for low and high H_{TS} in table 3. It can be seen that the reward in low H_{TS}, that is, a high share of IBRs, increases slightly compared to the reference scenario, while a higher H_{TS}, i.e., a low share of IBRs, leads to a significantly lower reward for PI-AC. The rewards achieved by the AC are comparable for both low and high IBR penetration in the 14-bus system, thus, the differences can be assumed to result from

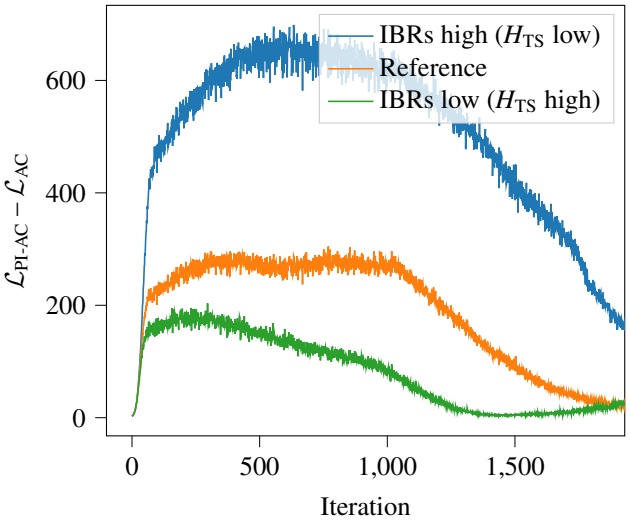

FIGURE 6 Critic loss difference of PI-AC and AC for different IBR penetrations (14-bus system)

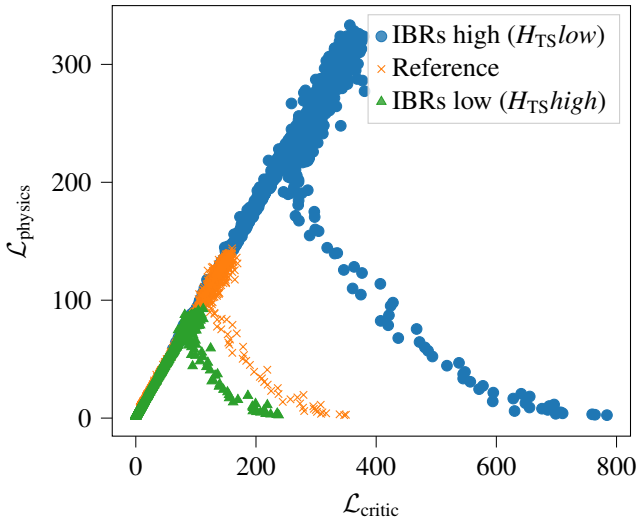

FIGURE 7 Influence of IBR penetration on $\mathcal{L}_{\text{physics}}$ and $\mathcal{L}_{\text{critic}}$ (14-bus system)

the physics regularization term. However, the results from the 37-bus system show another influential factor. It is generally more difficult to obtain a better reward in the slower dynamics, i.e. H_{TS} is high, since the influence of the actions, i.e. VI parameters, cannot be clearly obtained. Thus, the AC reward also decreases with slower dynamics, i.e. smaller share of IBRs, in the 37-bus case, this cannot be found for the 14-bus system, which has generally faster dynamics compared to the 37-bus system.

fig. 8 a, b compare the learning behavior of the PI-AC and AC in the C-ξ plane. They indicate that the PI-AC reaches the

TABLE 3 Final reward r_{final} and required iterations (it.) for coordination of CIGRE 14-bus grid and IEEE 37-bus grid

Scenario	Algorithm	Reference		$H_{\text{DS, budget}}$ low		$H_{\text{DS, budget}}$ high		H_{TS} low		H_{TS} high		P_{load} low		P_{load} high	
		r_{final}	it.	r_{final}	it.	r_{final}	it.	r_{final}	it.	r_{final}	it.	r_{final}	it.	r_{final}	it.
14-bus	PI-AC	-6.93	704	-5.41	775	-15.15	2277	-6.39	631	-8.13	676	- 6.93	704	-6.93	704
	AC	-10.69	662	-6.48	991	- 13.15	1773	-10.89	775	-10.45	369	-10.69	662	-10.69	662
	GA	-17.56	910	-11.21	672	-21.71	398	- 17.48	782	-16.75	910	-17.09	985	-16.93	833
37-bus	PI-AC	-8.86	365	-4.38	352	-12.51	262	-7.62	340	-10.12	760	-8.86	365	-8.86	365
	AC	-9.42	686	-6.66	583	-12.49	275	-7.21	640	-9.23	459	-9.42	686	-9.42	686
	GA	-15.68	928	-14.51	981	-20.99	919	-15.01	868	-17.10	760	-16.45	779	-17.07	527

convergence region at an earlier point in time, i.e. at earlier iteration while the AC moves through the plane at later iterations. A similar behavior has been found throughout this paper for most scenarios. It indicates that the physics regularization helps to drive the reward to the convergence region faster.

In the following, we investigate how different shares of IBRs and the corresponding losses influence the behavior of the PI-AC in the optimization space, presented in the C-ξ plane in fig. 9. This figure shows that the increased IBR penetration, low H_{TS}, leads to a stronger movement in the optimization landscape and ultimately results in lower costs. This is different for a lower share of IBRs, i.e. high H_{TS}. In this situation, the PI-AC mostly moves around one spot and converges early with higher costs.

5.2 | Comparison of PI-AC and GA

In this section, we evaluate the differences in the search strategies between the PI-AC and the GA algorithm. To this end, we compare the achieved costs for both algorithms in fig. 8 a,c in the C-ξ plane. It can be seen that the PI-AC follows a straight path to the convergence region and spends the majority of iterations in that region to find the best setting. Focusing on the convergence region, it becomes visible that it is very shallow, due to the closely interwoven infeasible and feasible region. The GA constantly moves in the direction of lower costs, however, the behavior is substantially different from the PI-AC. The search is spread throughout the space, thus, the GA settles in a much higher reward region than the PI-AC. It should be highlighted that the PI-AC learns a policy, while the GA optimizes the problem at hand. To this end, the PI-AC can quickly obtain results after training. In contrast, the GA needs to be retrained for new results.

5.3 | Influence of the weighting factor Φ

The influence of the physics-based loss has been shown in previous results. We saw that higher physics loss often results in faster learning compared to the AC. The higher physics loss was caused by different power system situations, where the increased share of IBRs led to an increased physics loss. These findings lead to a consecutive questions: does the influence of the physics loss saturate, so a higher loss reduces the achieved reward. This question will be evaluated in the following.

Figure 10 demonstrates that an increased weighting Φ strongly influences the reward and also changes the reward trajectory. It can be seen that the best rewards are achieved consecutively in the area around $\Phi = 5000$. A higher loss leads to lower rewards. This is highly pronounced when Φ is larger than 100000. The rewards finally achieved with the different weightings are compared in fig. 11. This supports the results found in fig. 10. Based on these findings, we set $\Phi = 5000$ for all investigations.

In addition to its effect on the total loss $\mathcal{L}$, the physics loss also influences the critic loss $\mathcal{L}_{\text{critic}}$, shown in fig. 7. This stems from the fact that the optimizer minimizes the total loss $\mathcal{L}$. Higher weighting shifts the focus to the physics loss, which allows a larger critic loss during training. In this case, a generally higher loss is not critical, since the parameter updates for $\boldsymbol{\Theta}$ of both NNs are performed based on the gradients $\nabla\mathcal{L}$. This has already been visualized in fig. 7, but can also be seen in fig. 12 for different Φ. It can be seen here, that $\Phi = 5000$ results in a loss trajectory that is close to the bisecting line, while a smaller weighting causes a higher physics loss and a much higher weighting a smaller physics loss. In the latter case, the optimizer partly neglects the critic loss to minimize the physics loss.

5.4 | Discussion

This chapter discusses additional thoughts on the proposed PI-AC approach.

5.4.1 | Model-free operation

In the beginning of this paper, we aimed to formulate a model-free coordination function. However, we utilize a generic system model, the swing equation, in the physics loss term. This can still be considered model-free since the swing equation formulation works equally as an aggregated model for all power

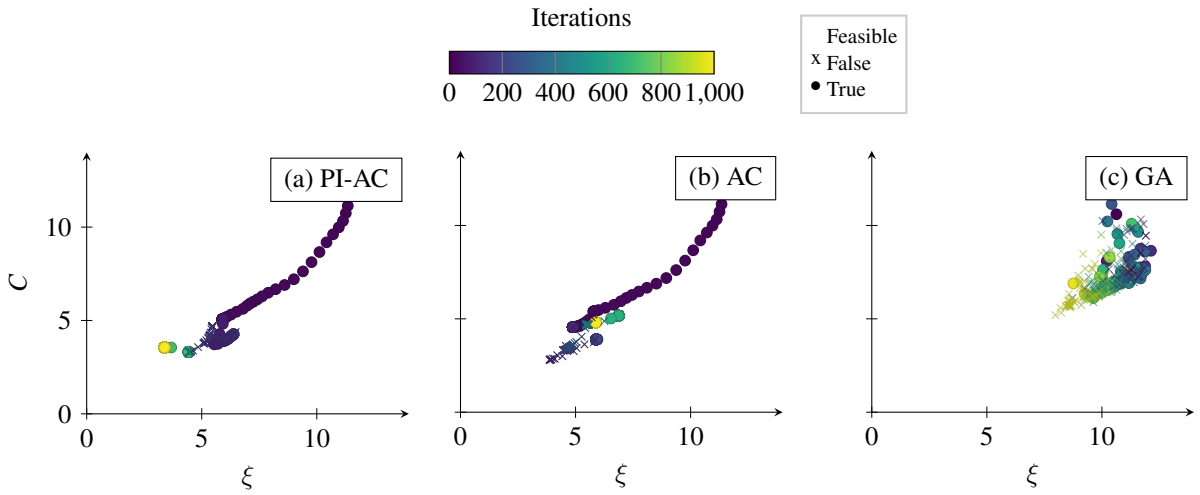

FIGURE 8 Comparison of PI-AC, AC and GA algorithms' movement through the optimization landscape (14-bus Reference scenario)

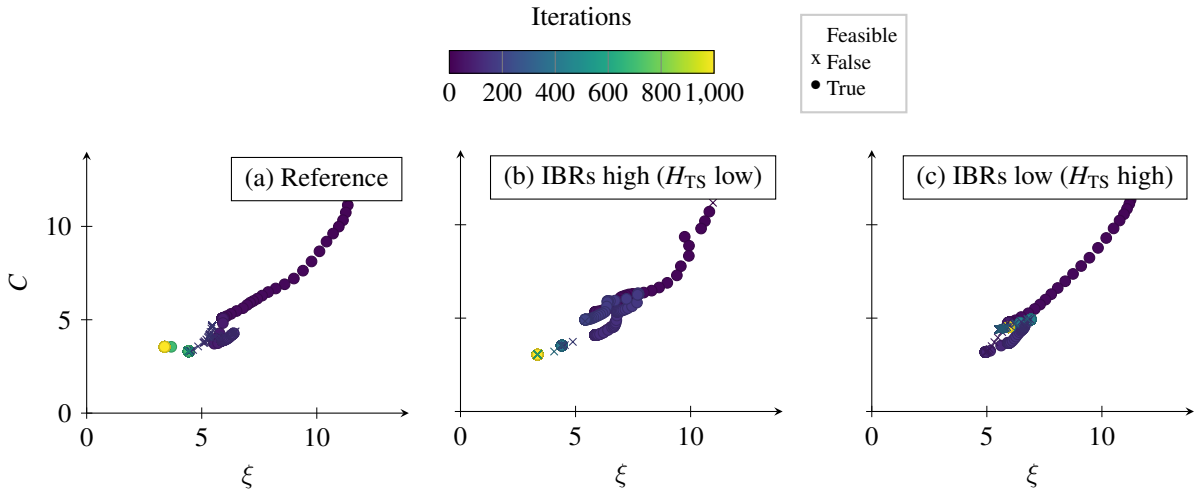

FIGURE 9 Objective analysis considering the evolution in a ξ and C space over time for PI-AC algorithm in a reference, high and low IBR penetration scenario (14-bus system)

systems, only the parameters change. These parameters are obtained by the critic network, thus, no prior knowledge about the system is required.

5.4.2 | PI-AC vs. GA

It has been found that the PI-AC and the GA apply different search strategies. The PI-AC produces a learning trajectory that points in the direction of convergence and arrives in this area after a few iterations. On the contrary, GA explores the space through its particles, causing a much more dispersed behavior. This ultimately causes a slower convergence.

5.4.3 | Runtime

For runtime comparison, some calculations are run on an Intel i7 11700 CPU. The PI-AC algorithm achieves an average runtime of 1425.86 s for 100 training iterations. In comparison, the AC algorithm yields similar runtimes, namely 1425.94 s for 100 training iterations. It is evident that both algorithms have a comparable runtime. Hence, the extended loss function does not require a noticeable extra computational effort. This stems from the fact that no additional NN has to be trained. The GA achieves an average runtime of 4190 s for 100 iterations. It has to rerun the simulation for every particle in each iteration, which makes training heavily dependent on the simulation time

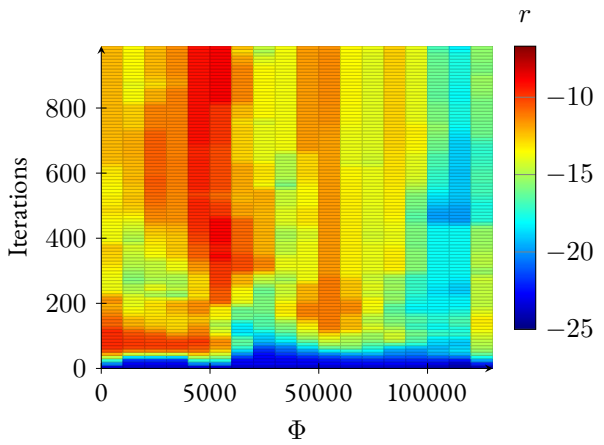

FIGURE 10 Reward trajectories during training for different Φ (14-bus reference scenario)

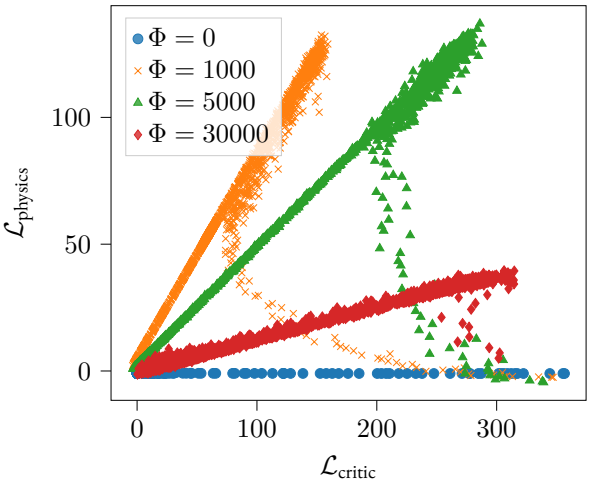

FIGURE 12 Comparison of different Φ weightings influence on $\mathcal{L}_{\text{physics}}$ and $\mathcal{L}_{\text{critic}}$(14-bus reference scenario)

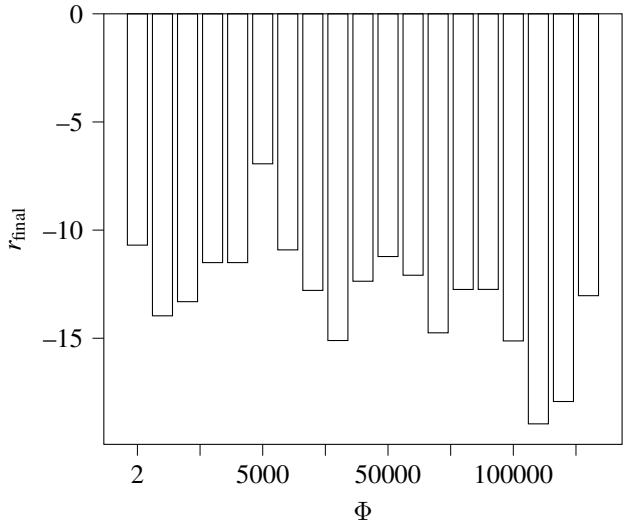

FIGURE 11 Comparison of achieved rewards for different Φ (14-bus reference scenario)

of the power system. The GA does not acquire a policy that can be used after training for online execution. Instead, it must be rerun every time a parameter is altered. On the contrary, the PI-AC performs a complete training only once to discover the underlying policy of the problem. Subsequently, the policy can be applied as long as there are no substantial changes to the problem. When a retraining is necessary, the online optimization process can still be performed using the policy that was previously acquired.

6 | CONCLUSION

In this paper, we presented the Physics-informed Actor-Critic (PI-AC) algorithm for model-free coordination of Virtual Inertia (VI) provision from a power distribution system. The PI-AC utilizes a physics regularization term in the RL loss formulation driven by a generic representation of the power system dynamics, the swing equation. We show that the physics regularization improves the learning, the PI-AC achieves better rewards in fewer iterations than the purely data-driven Actor-Critic (AC) in a case study based on the CIGRE 14-bus and IEEE 37-bus system. The results indicate that the physics-driven regularization is more pronounced in higher IBR penetration and therefore leads to better results and superiority of the PI-AC over the AC in these scenarios. This is important to note for future inverter-dominated power systems. Throughout the case study, we also compare the PI-AC to the metaheuristic Genetic Algorithm (GA) approach. The results show that the GA takes significantly longer to converge and achieves worse rewards than the PI-AC.

From the authors' perspective, the PI-AC is not limited to the presented problem. Physics-based regularization in RL can be a valuable tool for a variety of problems in power systems.

References

[1] BDEW, MITNETZ STROM, E-Bridge Consulting GmbH. Die zukünftige Rolle der Verteilnetzbetreiber in der Energiewende: Eine Studie [The future role of the distribution system operators in the energy transition: A study]. BDEW. 2017;.

[2] Del Nozal ÁR, Kontis EO, Mauricio JM, Demoulias

CS. Provision of inertial response as ancillary service from active distribution networks to the transmission system. IET Generation, Transmission and Distribution. 2020;14(22):5123–5134.

[3] Stock S, Hund P, Babazadeh D, Becker C. Reinforcement Learning based Coordination of Virtual Inertia Provision from Inverter-dominated Distribution Grids. In: 2023 32nd International Symposium on Industrial Electronics. 2023 32nd International Symposium on Industrial Electronics; 2023. .

[4] Magdy G, Shabib G, Elbaset AA, Mitani Y. A novel coordination scheme of virtual inertia control and digital protection for microgrid dynamic security considering high renewable energy penetration. IET Renewable Power Generation. 2019;13(3):462–474.

[5] Mahapatra K, Fan X, Li X, Huang Y, Huang Q. Physics Informed Reinforcement Learning for Power Grid Control using Augmented Random Search. HICSS. 2022;.

[6] Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. Nature Reviews Physics. 2021;3(6):2522–5820.

[7] Faria RR, Capron BDO, Secchi AR, De Souza MB. A data-driven tracking control framework using physics-informed neural networks and deep reinforcement learning for dynamical systems. Engineering Applications of Artificial Intelligence. 2024;127:107256.

[8] Raissi M, Perdikaris P, Karniadakis GE. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics. 2019;378:686–707.

[9] Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, et al.. Continuous control with deep reinforcement learning. arXiv. arXiv 2015.

[10] Sauer PW, Pai MA. Power System Dynamics and Stability. University of Illinois; 2006.

[11] Brunton SL, Kutz JN. Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control. Cambridge University Press; 2019.

[12] Zhong QC, Weiss G. Synchronverters: Inverters that mimic synchronours generators. IEEE Transactions on Industrial Electronics. 2011;58(4).

[13] Strunz K. Benchmark Systems for Network Integration of Renewable and Distributed Energy Resources. Electra. 2014;273.

[14] Rudion K, Orths A, Styczynski ZA, Strunz K. Design of benchmark of medium voltage distribution network for investigation of DG integration. In: 2006 IEEE Power Engineering Society General Meeting; 2006. .