Model-free Approaches for Real-time Distribution System Operation: A Comparison of Feedback Optimization and Reinforcement Learning

Sen Zhan*, Haoyang Zhang*, Koen Kok*, Nikolaos G. Paterakis*
*Department of Electrical Engineering, Eindhoven University of Technology, 5600MB Eindhoven, The Netherlands
Email: {s.zhan, h.zhang2, j.k.kok, n.paterakis}@tue.nl

***Abstract*—With the proliferation of distributed energy resources, real-time control becomes critical to ensure that voltage and loading limits are maintained in power distribution systems. The lack of an accurate grid model and load data, however, renders traditional model-based optimization inapplicable in this context. To overcome this limitation, this paper aims to present and compare two model-free and forecast-free approaches for real-time distribution system operation via Lyapunov optimization-based online feedback optimization (OFO) and deep reinforcement learning (DRL), respectively. Simulation studies performed on a 97-node low-voltage system suggest that OFO significantly outperforms DRL by 24% less PV energy curtailment over a test week relative to the total possible generation while enforcing distribution grid limits and requiring minimal effort for training.**

***Index Terms*—Distribution system operation, model-free control, online feedback optimization, deep reinforcement learning**

I. Introduction

The integration of distributed energy resources (DERs) such as photovoltaics (PVs) and battery storage has caused operational problems in power distribution systems. These include violations of voltage limits and overloading of cables and transformers. Fortunately, those inverter-interfaced DERs also provide opportunities for real-time control, which can potentially be leveraged to resolve network issues.

In the context of real-time distribution system operation, two significant challenges arise when considering traditional offline optimization, i.e. solving an optimization problem formulated based on a model of the system and estimated data. First, an accurate model of the distribution system is not always available. Creating a computationally tractable model of AC power flow poses further challenges. Second, load profiles in distribution systems are uncertain and volatile. Privacy concerns also prevent distribution system operators (DSOs) from obtaining these profiles in real time. These challenges motivated the recent development of online feedback optimization (OFO) and reinforcement learning (RL) approaches for the real-time operation of distribution systems.

OFO uses network measurements, for example voltage and current measurements as feedback and leverages optimization algorithms as feedback controllers to iteratively steer the distribution system towards its optimal operating points [1]. By incorporating measurements, OFO eliminates the need for a grid model and load data and is robust to uncertainties [1]. Indicatively, [2]–[4] present OFO implementations based on primal-dual algorithms that feature distributed calculations, while [5] and [6] conduct laboratory-scale and real-life demonstrations based on the gradient projection algorithm, respectively.

Although OFO does not need a detailed grid model, it takes as input network sensitivities, which reflect how power injections affect system states. These sensitivities can be derived through a grid model or by data-driven approaches, e.g. [7] to make OFO model-free. The zeroth-order optimization-based OFO bypasses the need for sensitivities using estimated gradients [8] at the cost of potential performance deterioration. Furthermore, as a real-time corrective algorithm, OFO by nature does not consider the intertemporal relation of some DERs such as battery storage, i.e. the action at one time step will impact the future. In this regard, a multiperiod extension was proposed in [9] using real-time forecasts, a two-stage approach was proposed in [10], while Lyapunov optimization was exploited in [11], [12].

Reinforcement learning (RL) algorithms have also been utilized to address the challenges of real-time distribution system operation, particularly in the presence of highly uncertain and dynamically changing DERs. These uncertainties encompass factors such as the generation and consumption patterns of prosumers and the dynamic state of DERs, including the state of charge (SoC) of battery storage. The operation of DERs can be effectively modeled as a Markov decision process (MDP), which can then be solved using a variety of RL algorithms. In [13], the authors proposed a real-time intelligent resilience controller that employs RL algorithms to dispatch DERs and expedite power restoration to customers following sudden outages. Based on this work, [14] introduced a safety layer to the actor network of the RL algorithm. This layer ensures recalibration of unsafe actions in safe domains through a quadratic programming formulation.

Although both OFO and RL have shown promise in addressing real-time operational challenges, a comparison of their effectiveness in distribution systems remains unexplored. In this regard, the main contributions of this paper are as follows:

- Two model-free and forecast-free approaches based on

This work is part of the NO-GIZMOS project (MOOI52109) which received funding from the Topsector Energie MOOI subsidy program of the Netherlands Ministry of Economic Affairs and Climate Policy, executed by the Netherlands Enterprise Agency (RVO).

OFO and DRL are developed, respectively, for the real-time operation of distribution systems by controlling PVs and battery storage.
- The approaches are simulated and compared on a 97-bus low-voltage system concerning PV curtailment, network constraint satisfaction, and computation time. Sensitivity analyses are performed on the key parameters of the Lyapunov optimization-based OFO.

The remainder of this paper is organized as follows: Section II formulates the real-time power flow optimization problem. Sections III and IV develop the OFO and DRL approaches, respectively. Section V presents the simulation results, while Section VI draws conclusions.

II. Real-time distribution system operation

A. Notation

Consider a distribution grid with $N+1$ buses collected in the set $\mathcal{N} := \{0, 1, \cdots, N\}$ and branches including cables and transformers collected in $\mathcal{E} := \{(i,j)\} \subseteq \mathcal{N} \times \mathcal{N}$. Define $\mathcal{N}^+ := \mathcal{N} \backslash \{0\}$, where bus 0 is the substation bus. Define $\mathcal{T}$ as the collection of time steps. For each bus $i \in \mathcal{N}$, let $v_{i,t}$ be its voltage magnitude at the time step t. Let $\underline{v}$ and $\overline{v}$ be the lower and upper voltage limits, respectively. Bus 0 is assumed to have a fixed voltage magnitude of v_0. For each branch $(i,j) \in \mathcal{E}$, let $P_{ij,t}$ and $Q_{ij,t}$ be the active and reactive power flow from buses i to j at time t, respectively, and let $\overline{S}_{ij}$ be its apparent power capacity.

Define $\mathcal{N}^{pv} \subseteq \mathcal{N}^+$ as the set of buses with PV installations, and $\mathcal{N}^b \subseteq \mathcal{N}^+$ the set with battery storage installations. For each PV $i \in \mathcal{N}^{pv}$, at time t, let $p^{pv}_{i,t}$ and $q^{pv}_{i,t}$ be its active and reactive power injections into the grid (negative values for offtake), respectively, let $\overline{P}^{pv}_{i,t}$ be its maximum power generation, and let $\overline{S}^{pv}_i$ be its inverter capacity. For each battery storage $i \in \mathcal{N}^b$, let $p^b_{i,t}$ be its charging power at time t (negative values for discharging), let $q^b_{i,t}$ be its reactive power consumption (negative values for injections), let $\overline{S}^b_i$ be its power rating, let $\overline{E}^b_i$ be its energy capacity, let $SoC_{i,t}$ be its state of charge at time t, let $\underline{SoC}_i$ and $\overline{SoC}_i$ be its lower and upper SoC limits, and let η^{ch}_i and η^{dis}_i be its charging and discharging efficiencies, respectively. Finally, bold letters will be used for column vectors with components defined earlier, for example, $\mathbf{v} := [v_i, i \in \mathcal{N}^+]^\top$.

B. PV and battery storage modeling

1) PV modeling: The PV model includes (1) and (2), where (1) enforces the maximum generation limit, while (2) enforces the inverter rating limit.

$$0 \le p^{pv}_{i,t} \le \overline{P}^{pv}_{i,t}, \forall i \in \mathcal{N}_{pv}, \forall t \in \mathcal{T} \tag{1}$$

$$(p^{pv}_{i,t})^2 + (q^{pv}_{i,t})^2 \le (\overline{S}^{pv})^2, \forall i \in \mathcal{N}_{pv}, \forall t \in \mathcal{T} \tag{2}$$

2) Battery storage modeling: The battery storage model includes (3)-(6). Specifically, (3) describes the SoC evolution, where $p^{ch}_{i,t}$ and $p^{dis}_{i,t}$ are defined in (4), and Δ_t is the length of each time step. The SoC limits are enforced in (5), while the battery inverter rating limit is enforced in (6). Note that the initial SoC $SoC_{i,0}$ is a known parameter.

$$SoC_{i,t} = SoC_{i,t-1} + (\eta^{ch}_i p^{ch}_{i,t} - p^{dis}_{i,t}/\eta^{dis}_i)\Delta_t/\overline{E}_i, \forall i \in \mathcal{N}^b, \forall t \in \mathcal{T} \tag{3}$$

$$p^{ch}_{i,t} = \max(0, p^b_{i,t}), p^{dis}_{i,t} = \max(0, -p^b_{i,t}), \quad \forall i \in \mathcal{N}^b, \forall t \in \mathcal{T} \tag{4}$$

$$\underline{SoC}_i \le SoC_{i,t} \le \overline{SoC}_i, \forall i \in \mathcal{N}^b, \forall t \in \mathcal{T} \tag{5}$$

$$(p^b_{i,t})^2 + (q^b_{i,t})^2 \le (\overline{S}^b_i)^2, \forall i \in \mathcal{N}^b, \forall t \in \mathcal{T} \tag{6}$$

For each battery i, at each time step t, if $SoC_{i,t-1}$ falls within the limit, the following limits for the battery charging power $p^b_{i,t}$ can be derived according to (3) and (5). Note that $SoC_{i,t-1}$ is a known parameter at time t.

$$\eta^{dis}_i \overline{E}_i \frac{\underline{SoC}_i - SoC_{i,t-1}}{\Delta t} \le p^b_{i,t} \le \overline{E}_i \frac{\overline{SoC}_i - SoC_{i,t-1}}{\eta^{ch}_i \Delta t} \tag{7}$$

C. Problem formulation

An optimization problem is formulated in (8)-(10) for the distribution system operation application. The objective function in (8) minimizes the active power curtailment from PVs plus reactive power use weighted by a positive term $\omega < 1$, where $\Omega := \{p^{pv}_{i,t}, q^{pv}_{i,t}, p^b_{j,t}, q^b_{j,t}, \forall i \in \mathcal{N}_{pv}, \forall j \in \mathcal{N}_b, \forall t \in \mathcal{T}\}$ is the set of decision variables. The problem admits a multi-period formulation because of (3). The system state variables $v_{i,t}$, $P_{ij,t}$, and $Q_{ij,t}$ are constrained in (9) and (10). Note that these state variables depend not only on the decision variables in Ω, but also on other non-controllable loads in the system. Explicitly modeling this dependency requires the full AC power flow equations, the topology of the underlying network, and data of the non-controllable loads.

$$\min_{\Omega} \sum_{t \in \mathcal{T}} \left[\sum_{i \in \mathcal{N}_{pv}} \left[(p^{pv}_{i,t} - \overline{P}^{pv}_{i,t})^2 + \omega (q^{pv}_{i,t})^2 \right] + \omega \sum_{j \in \mathcal{N}_b} (q^b_{j,t})^2 \right] \tag{8}$$

$$\text{s.t.} \quad (1) - (6)$$

$$\underline{v} \le v_{i,t} \le \overline{v}, \forall i \in \mathcal{N}^+, \forall t \in \mathcal{T} \tag{9}$$

$$\sqrt{P^2_{ij,t} + Q^2_{ij,t}} \le \overline{S}_{ij}, \forall (i,j) \in \mathcal{E}, \forall t \in \mathcal{T} \tag{10}$$

D. Real-time decision making

To summarize, modeling and solving (8)-(10) presents the following challenges:

- An accurate grid model is not always available;
- Estimated data of non-controllable loads are not available or highly uncertain;
- Power flow equations are non-convex;
- The battery model is non-convex because of (4).

Because of these challenges, the aim of this paper is to provide real-time oracles that make decisions over Ω only based on currently available information while trying to pursue

Algorithm 1: Real-time system operation process

Data: $\mathcal{O}_0$
for $t = 1, 2, \cdots$ **do**
  Implement $\mathcal{O}_{t-1}$ to the system;
  Acquire $\mathcal{I}_t$;
  Call an oracle to compute $\mathcal{O}_t$.
end

the solution to (8)-(10). Particularly, at each time step $t-1$, the input includes:

$$\begin{aligned}\mathcal{I}_{t-1} := \{&\overline{P}^{pv}_{i,t-1}, \overline{S}^{pv}_{i}, p^{pv}_{i,t-1}, q^{pv}_{i,t-1}, SoC_{j,t-1}, \eta^{ch}_{j}, \\ &\eta^{dis}_{j}, \Delta t, \overline{E}_j, \overline{S}^{b}_{j}, p^{b}_{j,t-1}, q^{b}_{j,t-1}, \omega, \tilde{v}_{k,t-1}, \underline{v}, \overline{v}, \\ &\tilde{P}_{mn,t-1}, \tilde{Q}_{mn,t-1}, \overline{S}_{mn}, \forall i \in \mathcal{N}^{pv}, \forall j \in \mathcal{N}^{b}, \\ &\forall k \in \mathcal{N}^{+}, \forall (m,n) \in \mathcal{E}\}\end{aligned} \tag{11}$$

Of particular importance are measurements of the state variables $\tilde{v}_{i,t-1}$, $\tilde{P}_{ij,t-1}$, and $\tilde{Q}_{ij,t-1}$, which enable a closed control loop and bring robustness [1]. The output includes:

$$\mathcal{O}_{t-1} := \{p^{pv}_{i,t}, q^{pv}_{i,t}, p^{b}_{j,t}, q^{b}_{j,t}, \forall i \in \mathcal{N}_{pv}, \forall j \in \mathcal{N}_b\} \tag{12}$$

The iterative process for the real-time distribution system operation is summarized in Algorithm 1. In the following sections, two oracles are presented based on OFO and DRL.

III. Online feedback optimization

A. Lyapunov optimization

This section describes an OFO oracle based on Lyapunov optimization, which makes OFO forecast-free and applicable to control battery storage. To handle the intertemporal relation, Lyapunov optimization starts with defining a virtual queue for each battery storage as in (13), where $\xi^{+}_{i,t}$ and $\xi^{-}_{i,t}$ represent the incoming and outgoing energy for the specific battery at the time step t, respectively. Physically, $V_{i,t}$ represents the energy stored in the battery.

$$V_{i,t} = V_{i,t-1} + \xi^{+}_{i,t} - \xi^{-}_{i,t}, \forall i \in \mathcal{N}_b, \forall t \in \mathcal{T} \tag{13}$$

Then, a quadratic Lyapunov function can be defined in (14), where e_i is a design parameter.

$$L_{i,t} = \frac{1}{2}(V_{i,t} - e_i)^2, \forall i \in \mathcal{N}_b, \forall t \in \mathcal{T} \tag{14}$$

The Lyapunov drift is defined in (15), where $B = \frac{1}{2}\overline{\xi}_i^2$ with $\overline{\xi}_i$ being an upper bound for $\xi^{+}_{i,t} - \xi^{-}_{i,t}$ potentially derived from the limit of the battery inverter capacity.

$$\begin{aligned}\Delta L_{i,t} &= L_{i,t} - L_{i,t-1} = \frac{1}{2}(V_{i,t} - e)^2 - \frac{1}{2}(V_{i,t-1} - e_i)^2 \\ &= (V_{i,t-1} - e_i)(\xi^{+}_{i,t} - \xi^{-}_{i,t}) + \frac{1}{2}(\xi^{+}_{i,t} - \xi^{-}_{i,t})^2 \\ &\le (V_{i,t-1} - e_i)(\xi^{+}_{i,t} - \xi^{-}_{i,t}) + B, \forall i \in \mathcal{N}_b, \forall t \in \mathcal{T}\end{aligned} \tag{15}$$

Following the *minimizing drift-plus-penalty* approach [11], [12], [15], the original multiperiod problem is transformed to the following real-time optimization problem. The upper bound of the Lyapunov drift is minimized in this formulation, where γ is a weighting parameter. Furthermore, the inclusion of (7) ensures that the SoC limit is always kept. Note that the subscript t in the constraints implies that only the constraints at time t are considered for this real-time problem.

$$\begin{aligned}\min_{\Omega_t} f_t := &\sum_{i \in \mathcal{N}_{pv}} \left[(p^{pv}_{i,t} - \overline{P}^{pv}_{i,t})^2 + \omega (q^{pv}_{i,t})^2\right] \\ &+ \omega \sum_{j \in \mathcal{N}_b} (q^{b}_{j,t})^2 + \gamma (V_{i,t-1} - e_i)(\xi^{+}_{i,t} - \xi^{-}_{i,t}) \\ \text{s.t.} \quad &(1)_t, (2)_t, (6)_t, (7)_t, (9)_t, (10)_t\end{aligned} \tag{16}$$

B. Primal-dual projected gradient

Characterized by the incorporation of measurements, distributed calculations, and a gather-and-broadcast communication architecture, a primal-dual projected gradient (PDPG) algorithm is developed to solve the real-time optimization problem. Define the partial Lagrangian function as in (17), where $\boldsymbol{\mu}_t$, $\boldsymbol{\lambda}_t$, and $\boldsymbol{\rho}_t$ are the nonnegative Lagrangian dual variables associated with the respective inequality constraint.

$$\begin{aligned}\mathcal{L}_t := f_t &+ \boldsymbol{\mu}_t^{\top}(\underline{\mathbf{v}} - \mathbf{v}_t) + \boldsymbol{\lambda}_t^{\top}(\mathbf{v}_t - \underline{\mathbf{v}}) \\ &+ \sum_{(i,j) \in \mathcal{E}} \rho_{ij,t}\left(\sqrt{P^2_{ij,t} + Q^2_{ij,t}} - \bar{S}_{ij}\right)\end{aligned} \tag{17}$$

At its core, the PDPG algorithm sequentially updates the dual and primal variables. At each time step $t-1$, the following steps are taken:

S1. Collect voltage measurements $\tilde{\mathbf{v}}_{t-1}$ for each bus and update $\boldsymbol{\lambda}_t$ and $\boldsymbol{\mu}_t$ using (18)-(19), where α (with or without superscript) is the step size, and $\text{proj}_{\mathcal{X}}[\cdot]$ denotes the projection into $\mathcal{X}$. Here, the projection into $\mathbb{R}^N_+$ ensures that the dual variables are nonnegative.

$$\boldsymbol{\lambda}_t = \text{proj}_{\mathbb{R}^N_+}[\boldsymbol{\lambda}_{t-1} + \alpha^{\lambda}(\tilde{\mathbf{v}}_{t-1} - \overline{\mathbf{v}})] \tag{18}$$

$$\boldsymbol{\mu}_t = \text{proj}_{\mathbb{R}^N_+}[\boldsymbol{\mu}_{t-1} + \alpha^{\mu}(\underline{\mathbf{v}} - \tilde{\mathbf{v}}_{t-1})] \tag{19}$$

Collect power measurements $\tilde{P}_{ij,t-1}$ and $\tilde{Q}_{ij,t-1}$ for each branch (i,j) and update $\rho_{ij,t}$ using (20).

$$\rho_{ij,t} = \text{proj}_{\mathbb{R}_+}[\rho_{ij,t-1} + \alpha^{\rho}(\sqrt{\tilde{P}^2_{ij,t-1} + \tilde{Q}^2_{ij,t-1}} - \bar{S}_{ij})] \tag{20}$$

S2. The DSO broadcasts the updated dual variables and power flow measurements, i.e. $\boldsymbol{\lambda}_t$, $\boldsymbol{\mu}_t$, $\boldsymbol{\rho}_t$, $\tilde{\boldsymbol{P}}_{t-1}$, and $\tilde{\boldsymbol{Q}}_{t-1}$.

S3. For each PV $i \in \mathcal{N}_{pv}$, update its active and reactive power setpoints based on (21), where $\mathcal{X}_{i,t}$ is the feasible operation region of PV i defined by $(1)_t$ and $(2)_t$. Note that $\overline{P}^{pv}_{i,t}$ can be replaced by $\overline{P}^{pv}_{i,t-1}$ to avoid forecasts.

$$\begin{aligned}\begin{bmatrix} p^{pv}_{i,t} \\ q^{pv}_{i,t} \end{bmatrix} \leftarrow \text{proj}_{\mathcal{X}_{i,t}} \Bigg\{ &\begin{bmatrix} p^{pv}_{i,t-1} \\ p^{pv}_{i,t-1} \end{bmatrix} \\ &- \alpha \nabla_{[p^{pv}_i, q^{pv}_i]} \mathcal{L}_t \Big|_{p^{pv}_{i,t-1}, q^{pv}_{i,t-1}, \boldsymbol{\lambda}_t, \boldsymbol{\mu}_t, \boldsymbol{\rho}_t} \Bigg\}\end{aligned} \tag{21}$$

Similarly, for each battery storage $i \in \mathcal{N}_b$, update its active and reactive power setpoints based on (22), where

$\mathcal{Y}_{i,t}$ is the feasible operation region of battery i defined by $(6)_t$ and $(7)_t$.

$$\begin{bmatrix} p^b_{i,t} \\ q^b_{i,t} \end{bmatrix} \leftarrow \text{proj}_{\mathcal{Y}_{i,t}} \left\{ \begin{bmatrix} p^b_{i,t-1} \\ p^b_{i,t-1} \end{bmatrix} - \alpha \nabla_{[p^b_i, q^b_i]} \mathcal{L}_t \Big|_{p^b_{i,t-1}, q^b_{i,t-1}, \boldsymbol{\lambda}_t, \boldsymbol{\mu}_t, \boldsymbol{\rho}_t} \right\} \tag{22}$$

Remark 1. The current presentation assumes that every bus or branch is monitored. In practice, the DSO can also choose to monitor only critical buses located at the end of distribution feeders and critical branches such as transformers and cables that are at the beginning of feeders to save costs. This however reduces visibility across the grid and could lead to issues in unmonitored buses/branches.

Remark 2. The partial derivatives in (21) and (22) are explained in detail in [10]. An important ingredient is the network sensitivities, e.g. $\nabla_{\mathbf{p}^{pv}} \mathbf{v}$. In this work, to turn OFO model-free, Ridge regression is used based on yearly data of nodal net loads and measurements of voltages and power flow. The data set is divided into training and test sets to find the optimal regularization parameter.

Remark 3. The works in [2], [16] leveraged primal and dual regularization to establish theoretical convergence results for the PDPG algorithm. The regularization parameters must be small to reduce the discrepancy between the output of the PDPG algorithm and the underlying optimizer of the real-time optimization problem. In practice, the algorithm already works well without regularization.

IV. Reinforcement learning

The real-time operation of distribution systems that integrate DERs presents a variety of challenges due to their dynamic and uncertain nature. These challenges, including uncertainty in DER profiles, the temporal correlation of battery storage, and the need to make real-time decisions without full system knowledge, make the problem of distribution system operation well-suited for formulation as a MDP. For example, when the operator decides to charge or discharge a battery, this decision directly affects the battery's SoC at the next time step, which in turn influences subsequent decisions about when and how much to charge or discharge. By framing the problem within the MDP framework, the operator takes action $a_t = \mathcal{O}_t$ by controlling the DERs at the time step t based on the private partial observation $o_t = \mathcal{I}_t$ of the current state $s_t \in \mathcal{S}$ of the distribution system under policy $\pi(o_t)$. Subsequently, the operator receives a reward denoted as r_t from the system state, and the state is updated to the next state s_{t+1} with a probability governed by $p\big(s_{t+1}|s_t, a_t\big)$. The operator's reward r_t is comprised of components reflecting voltage violations at buses $r^v_{i,t}$, congestion $r^c_{ij,t}$, and the curtailment of PV generation $r^{pv}_{i,t}$ as follows:

$$r^c_{ij,t} = -\left(\frac{\sqrt{P^2_{ij,t} + Q^2_{ij,t}}}{\overline{S}_{ij,t}} - 1.0\right), \forall (i,j) \in \mathcal{E}, \forall t \in \mathcal{T} \tag{23}$$

$$r^v_{i,t} = \begin{cases} \alpha^v & \text{if } \underline{v}' \le v_{i,t} \le \overline{v}', \\ 0 & \text{if } \underline{v} \le v_{i,t} < \underline{v}', \\ 0 & \text{if } \overline{v}' < v_{i,t} \le \overline{v}, \\ -(\underline{v} - v_{i,t})^2 & \text{if } v_{i,t} < \underline{v}, \\ -(v_{i,t} - \overline{v})^2 & \text{if } \overline{v} < v_{i,t}, \end{cases} \tag{24}$$

$$r^{pv}_{i,t} = (p^{pv}_{i,t} - \overline{P}^{pv}_{i,t})^2, \forall i \in \mathcal{N}_{pv}, \forall t \in \mathcal{T} \tag{25}$$

$$r_t = \sum_{i \in \mathcal{N}^+} \omega^v r^v_{i,t} + \sum_{(i,j) \in \mathcal{E}} \omega^c r^c_{ij,t} + \sum_{i \in \mathcal{N}_{pv}} \omega^{pv} r^{pv}_t, \forall t \in \mathcal{T} \tag{26}$$

where ω^v, ω^c, and ω^{pv} represent the weight factors for the rewards associated with voltage violations, congestion constraints, and PV curtailment, respectively. The voltage regulation reward $r^v_{i,t}$ is defined by a piecewise function, incorporating a constant reward α^v and voltage thresholds: $\underline{v} < \underline{v}' < \overline{v}' < \overline{v}$. The congestion reward $r^c_{ij,t}$ penalizes branch overloading. Specifically, this study focuses on congestion at the transformer connected to the upper-level grid. Given the discount factor $\gamma_d \in [0,1]$, that weighs present and future rewards, the operator's objective is to maximize its cumulative discounted return, defined as $G_t = \sum_{\delta t=1}^{T} \gamma_d^{\delta t-1} r_{t+\delta t}$. To achieve this, the Q-value function $Q(s_t, a_t) = \mathbf{E}_\pi(G_t|s_t, a_t)$, is introduced to estimate the expected discounted return of the operator given the current state s_t and action a_t under policy π, respectively. The Bellman equation facilitates a recursive formulation of the Q-value function, enabling its update through the temporal difference (TD) method, which operates in a bootstrapping manner as follows:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha^{TD} \left[r_t + \gamma_d \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right] \tag{27}$$

where $\alpha^{TD} \in [0,1]$ is the step size. To address the aforementioned MDP, this work employs the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm [17] which is an actor-critic DRL algorithm that extends the DDPG algorithm [18]. It consists of two critic networks, $Q_1(s_t, a_t|\theta^{Q_1})$ and $Q_2(s_t, a_t|\theta^{Q_2})$, which estimate the Q-value by using the smaller of the two Q-values as the target. This approach helps mitigate overestimation in the Q-function. Additionally, the algorithm includes an actor network $\mu(s_t|\theta^\mu) = \text{argmax}_{a_t} Q(s_t, a_t)$ that determines the action to take. In addition to the regular critic and actor networks, there are target critic networks $Q'_1(s_t, a_t|\theta^{Q'_1})$ and $Q'_2(s_t, a_t|\theta^{Q'_2})$, and target actor networks $\mu'(s_t|\theta^{\mu'})$, which mirror the structure of the regular networks and are periodically updated via soft updates to stabilize the learning process. The regular critic networks are then updated towards the target by minimizing the loss functions, $L(\theta^{Q_1})$ and $L(\theta^{Q_2})$, as follows:

$$y = r_t + \gamma_d \min_{i=1,2} Q'_i(s_{t+1}, a_{t+1}|\theta^{Q'_i}) \tag{28}$$

$$L(\theta^{Q_1}) = \mathbf{E}_{(s_t, a_t, r_t, s_{t+1}) \sim D} \left[Q_1(s_t, a_t|\theta^{Q_1}) - y \right] \tag{29}$$

$$L(\theta^{Q_2}) = \mathbf{E}_{(s_t, a_t, r_t, s_{t+1}) \sim D} \left[Q_2(s_t, a_t|\theta^{Q_2}) - y \right] \tag{30}$$

where y represents the target Q-value, and D denotes the replay buffer, which stores training data in the form of (s_t, a_t, r_t, s_{t+1}). The objective of the critic networks' loss function is to minimize the TD error, defined as the difference between the estimated Q-values of the target networks and the regular networks. The objective of the actor network is maximizing the expected profit return as follows:

$$J(\theta^{\mu}) = \mathbf{E}_{(s_t,a_t,r_t,s_{t+1})\sim D}\left[\min_{i=1,2} Q_i\left[s_t, a_t(\theta^{\mu})\right]\right] \tag{31}$$

TD3 trains a deterministic policy in an off-policy way, and Gaussian noise $\mathcal{N}_t^{\epsilon}$ is added to the action for exploring the environment and finding the global optimum as follows:

$$a_t = clip(\mu(s_t|\theta^{\mu}) + \mathcal{N}^{\epsilon}(0, \sigma^{\epsilon}), \underline{a}, \overline{a}) \tag{32}$$

where σ^{ϵ} is the standard deviation of the Gaussian noise, and $\underline{a}$ and $\overline{a}$ represent the lower and upper bounds of the action space, respectively. To diminish exploration over time and facilitate convergence, σ^{ϵ} is linearly decayed across episodes by multiplying it with a decay factor, κ, in each episode.

V. Simulation Study

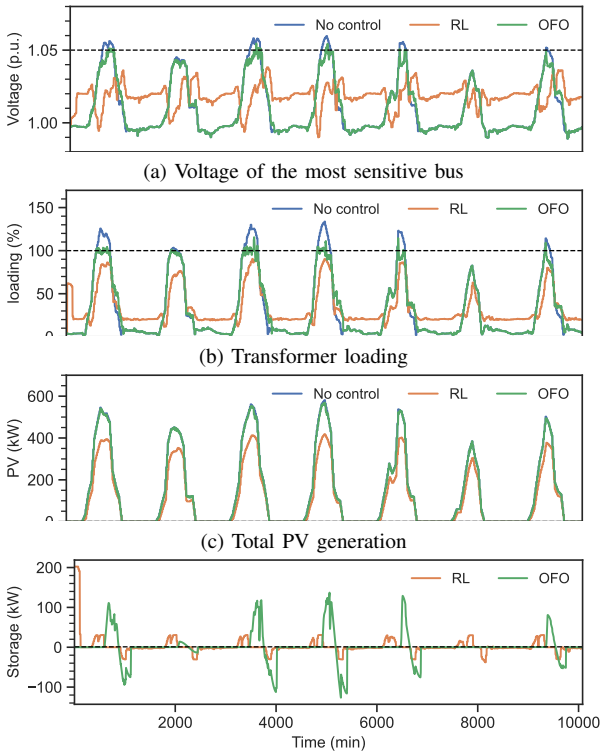
A. Case description

To demonstrate and compare the two real-time control approaches, a case study is conducted on a 97-bus LV system, which is adopted from Simbench (*1-LV-rural2–2-no_sw*) [19]. To simulate the congested case, 54 PV units and 36 battery storage units are randomly distributed throughout the system. The transformer capacity is increased from 250 to 400 kVA. The tap position is set as neutral. The slack bus voltage is 1.0 pu. Time-series maximum PV generation and load profiles from a summer week (days 245-251) with a 15-minute resolution are used for simulation. The lower and upper voltage limits are set to 0.95 and 1.05 pu, respectively. For the batteries, the charging and discharging efficiencies are assumed to be 95%. The initial, minimum, and maximum SoC are 0.2, 0.2, and 0.8, respectively. Finally, a highly-efficient AC power flow solver *Power-Grid-Model* [20] is run to simulate system response to the optimized power setpoints.

For the OFO simulation, it is assumed that the controller runs every minute. For Lyapunov optimization, e_i is taken as the initial energy stored in the battery i. Using a trial-and-error strategy, the drift weighting factor is set to 0.5, and the step sizes are chosen as: $\alpha = 0.5$, $\alpha^{\lambda} = \alpha^{\mu} = 10^5$, and $\alpha^{\rho} = 10^2$. PVs are initialized to operate at their maximum generation, while active power of storage and all reactive power are initialized to 0. For the RL implementation, simulation is performed using PyTorch [21], with hyperparameter optimization of the DRL algorithm performed through the open-source framework Optuna [22].

B. Results

The simulation results are summarized in Fig. 1, using the case without control as a benchmark. Figures 1a and 1b show the voltage profiles for the most sensitive bus and the transformer loading, respectively. Although both OFO and DRL enforce the grid limits, OFO makes more efficient use of the grid capacity. Temporary violations of the grid limits are, however, inevitable given the corrective nature of the algorithm and the primal-dual implementation. In particular, it takes some iterations for the dual variables to accumulate grid constraint violations, which consequently drive the DERs to steer the distribution system to secure operating regions. Figure 1c shows that OFO leads to negligible PV curtailment, which is, however, significant for DRL. As seen in Fig. 1d, OFO drives the batteries to charge during PV generation peaks, while in DRL, batteries tend to charge earlier than the PV peaks and DRL has to resort to PV curtailment to manage network congestion and voltage constraints. As a result, DRL leads to 24% more PV generation curtailment than OFO with respect to the maximum. A detailed comparison of DRL and OFO is provided in Table I, where the average voltage/loading violations (AVV/ALV) are defined as in [8]. Regarding computation, both approaches are suitable for online operation, averaging a few milliseconds to derive the updated setpoints at each time step. Aside from training the network sensitivities, OFO does not require additional training, whereas DRL demands significant training effort.



(a) Voltage of the most sensitive bus

(b) Transformer loading

(c) Total PV generation

(d) Total storage operation (positive for charging)

Fig. 1: DER and grid profiles using OFO and DRL.

TABLE I: Summary of comparison between DRL and OFO.

	Curtail. (kWh)	AVV (pu)	ALV (%)	Training (s)	Inference (ms)
DRL	5574 (25.2%)	0.0	0.0	93100	2.78
OFO	229 (1.0%)	6.1×10^{-6}	0.12	5	1.23

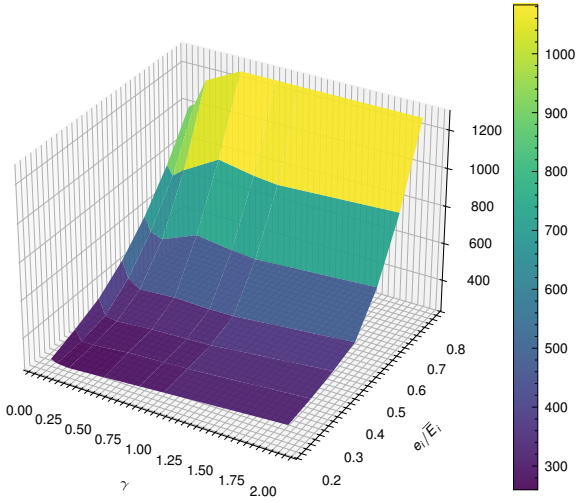


Fig. 2: Sensitivity analysis results for Lyapunov optimization showing PV curtailment (kWh).

C. Sensitivity analysis

This section describes the impact of the design parameters of Lyapunov optimization. In Fig. 2, the impact of e_i from (14) and the weighting parameter γ is shown. To ensure that the OFO controller works over a wide range of parameters, a regularization term of the active power of battery storage is included in the Lagrangian function (17), i.e. $\omega_b(p^b_{i,t})^2$ where ω_b is chosen as 0.02. Figure 2 shows that e_i significantly affects the controller performance, substantialized by the power curtailment. In this specific case, a small e_i leads to more freed-up energy capacities of the batteries to accommodate surplus PV generation as the batteries stabilize at a lower SoC. The impact of γ is less significant. A higher γ leads to a higher priority of stabilizing battery SoC over reducing PV curtailment, therefore, increasing PV curtailment. A too low γ results in the batteries discharging too slowly after reaching its maximum SoC from absorbing PV generation, leaving less room to absorb PV surplus the next day.

VI. Conclusion

This paper presented and compared Lyapunov optimization-based OFO and DRL as two viable model-free and forecast-free approaches for real-time distribution system operation using battery storage and PVs. Simulation results showed that OFO significantly outperformed DRL by reducing 24% PV energy curtailment because of its more efficient use of the available grid capacity. The sensitivity analyses showed the significant impact of the design parameter e_i over the performance of Lyapunov optimization. In future work, a systematic approach to defining the Lyapunov function for optimal battery storage control warrants further investigation. Additionally, studying the impact of battery degradation on control strategies would provide valuable insights. Extending the single-agent DRL framework to a multi-agent DRL setup could also facilitate decentralized control of DERs.

References

[1] A. Hauswirth, Z. He, S. Bolognani, G. Hug, and F. Dörfler, "Optimization algorithms as robust feedback controllers," *Annu. Rev. Control*, 2024.

[2] E. Dall'Anese and A. Simonetto, "Optimal power flow pursuit," *IEEE Trans. Smart Grid*, vol. 9, 2018.

[3] S. Bolognani, R. Carli, G. Cavraro, and S. Zampieri, "On the need for communication for voltage regulation of power distribution grids," *IEEE Trans. Control Netw. Syst.*, vol. 6, pp. 1111–1123, 2019.

[4] S. Zhan, J. Morren, W. van den Akker, A. van der Molen, N. G. Paterakis, and J. G. Slootweg, "Fairness-incorporated online feedback optimization for real-time distribution grid management," *IEEE Trans. Smart Grid*, vol. 15, pp. 1792–1806, 2024.

[5] L. Ortmann, A. Hauswirth, I. Cadu, F. Dörfler, and S. Bolognani, "Experimental validation of feedback optimization in power distribution grids," *Electr. Power Syst. Res.*, vol. 189, 2020.

[6] L. Ortmann, C. Rubin, A. Scozzafava, J. Lehmann, S. Bolognani, and F. Dörfler, "Deployment of an online feedback optimization controller for reactive power flow optimization in a distribution grid," in *IEEE PES ISGT Europe*, 2023.

[7] M. Picallo, L. Ortmann, S. Bolognani, and F. Dörfler, "Adaptive real-time grid operation via online feedback optimization with sensitivity estimation," *Electr. Power Syst. Res.*, vol. 212, p. 108405, 2022.

[8] Y. Chen, A. Bernstein, A. Devraj, and S. Meyn, "Model-free primal-dual methods for network optimization with application to real-time optimal power flow," in *Proc. Am. Control Conf.*, no. 5, 2020, pp. 3140–3147.

[9] X. Zhou, E. Dall'anese, and L. Chen, "Online stochastic optimization of networked distributed energy resources," *IEEE Trans. Automat. Contr.*, vol. 65, no. 6, pp. 2387–2401, 2020.

[10] S. Zhan, J. Morren, W. van den Akker, A. van der Molen, N. G. Paterakis, and J. G. Slootweg, "Multi-timescale coordinated distributed energy resource control combining local and online feedback optimization," *Electr. Power Syst. Res.*, vol. 234, 2024.

[11] E. Stai, C. Wang, and J. Y. Le Boudec, "Online battery storage management via Lyapunov optimization in active distribution grids," *IEEE Trans. Control Syst. Technol.*, vol. 29, no. 2, pp. 672–690, 2021.

[12] S. Fan, G. He, X. Zhou, and M. Cui, "Online optimization for networked distributed energy resources with time-coupling constraints," *IEEE Trans. Smart Grid*, vol. 12, pp. 251–267, 2021.

[13] M. M. Hosseini and M. Parvania, "Resilient operation of distribution grids using deep reinforcement learning," *IEEE Trans. Ind. Inform.*, vol. 18, no. 3, pp. 2100–2109, 2021.

[14] S. Hou, A. Fu, E. M. S. Duque, P. Palensky, Q. Chen, and P. P. Vergara, "Distflow safe reinforcement learning algorithm for voltage magnitude regulation in distribution networks," *J. Mod. Power Syst. Clean Energy*, pp. 1–12, 2023.

[15] M. J. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Springer Nature, 2022.

[16] A. Bernstein and E. Dall'Anese, "Real-time feedback-based optimization of distribution grids: A unified approach," *IEEE Trans. Control Netw. Syst.*, vol. 6, pp. 1197–1209, 2019.

[17] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Int. Conf. Mach. Learn.* PMLR, 2018, pp. 1587–1596.

[18] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Int. Conf. Mach. Learn.* PMLR, 2014, pp. 387–395.

[19] S. Meinecke, D. Sarajlić, S. R. Drauz, A. Klettke, L. P. Lauven, C. Rehtanz *et al.*, "Simbench-a benchmark dataset of electric power systems to compare innovative solutions based on power flow analysis," *Energies*, vol. 13, 2020.

[20] Y. Xiang, P. Salemink, B. Stoeller, N. Bharambe, and W. van Westering, "Power grid model: A high-performance distribution grid calculation library," in *CIRED 2023*, vol. 2023, 2023, pp. 1–5.

[21] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[22] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *25th ACM SIGKDD*, 2019, pp. 2623–2631.