Combined learning and optimal power flow for storage dispatch in grids with renewables

Rodrigo Porteiro
School of Engineering
Universidad ORT Uruguay
Montevideo, Uruguay
rodrigo.porteiro@fi365.ort.edu.uy

Fernando Paganini
School of Engineering
Universidad ORT Uruguay
Montevideo, Uruguay
paganini@ort.edu.uy

Juan Andres Bazerque
Department of Electrical Engineering
University of Pittsburgh
Pittsburgh, PA, USA
juanbazerque@pitt.edu

***Abstract*—We propose an optimization and learning technique for controlling energy storage in power systems with renewables. A reinforcement learning (RL) approach is employed to bypass the need for an accurate stochastic dynamic model for wind and solar power; at the same time, the presence of the grid is explicitly accounted for through the "DC" approximation to the Optimal Power Flow (OPF) to impose line constraints. The key idea that allows the inclusion of such instantaneous constraints within the RL framework is to take as control actions the storage operational prices, which may be suitably discretized. A policy to select these actions as a function of the state is parameterized by a neural network model and trained based on traces of demand and renewables. We call this combined strategy RL-OPF. We test it on a trial network with real data records for demand and renewables, showing convergence to a control policy that induces arbitrage of energy across space and time.**

***Index Terms*—Energy storage, Power system optimization, Reinforcement learning.**

I. Introduction

Storage devices are being increasingly incorporated into power grids, giving them new capabilities. Storage enables the arbitrage of energy [1] over the time-scales of hours to days, reducing costs through the alignment of supply and demand. It also improves resilience since accidental islands may be powered by stored energy [2]. It supports microgrids [3], which may disconnect from the main network or reverse power flows with customers selling energy to utility companies.

Large storage systems could be strategically designed with a system-wide perspective and placed in a central bus [4]. Still, storage is naturally distributed: installed locally by consumers, incorporated to support solar and wind farms [5], or in the form of electric vehicles that connect to the grid occasionally and could move power in both directions [6].

In this context, we consider the problem of optimizing energy dispatch of a power system with storage, renewable sources, and grid constraints over a given discrete time horizon. Storage introduces arbitrage dynamics, where generation need not balance with demand instantly but can be accumulated and used at more convenient times in the future. This time, coupling makes the optimization fall in the category of *dynamic programming* [7]; there is also a significant amount of *uncertainty*, both in renewable (solar and wind) resources and in demand (e.g., from electric vehicles).

To bypass the need to model the complex solar, wind, and load dynamics, we consider *reinforcement learning* (RL) [8] as a tool for stochastic dynamic programming. Indeed, RL for control of energy storage is being actively explored [9], [10], [11], and has been proven effective for coping with the uncertainty of wind forecasts [12], [13]. However, standard RL is a black box approach that unduly abstracts other components for which we *do* have reliable models: in particular, energy exchanges in a grid obey the *power flow* equations commonly used in model-based optimization of energy storage management systems [14], [15]. The literature on incorporation of such instantaneous constraints into RL is quite limited, as the state of the art considers time-averaged constraints satisfied in expectation [16].

In this paper, we put forward a combined optimization strategy that is model-based regarding the power flows and learns from data how to respond optimally to the uncertainty in renewables and demand. The key idea lies in the choice of control action: rather than attempt to learn the power injections themselves, which are subject to the power flow constraints, we propose to learn *price* variables assigned to the storage units. Then, the injections are determined by these prices and the generation costs through a DC-OPF model and linear program [17]. The state variable in our RL setup includes storage levels and samples of renewables and demand, modeled as Markov processes with memory. In this way, learning is based on empirical traces of these variables.

The paper is organized as follows: Section II presents the model under consideration, and Section III describes our optimization and learning approach. An illustrative application example is given in Section IV, and conclusions in Section V.

II. Modeling Preliminaries

A. Notation

t, discrete time index with sampling interval ΔT.
$\mathbf{r}_t, \mathbf{d}_t \in \mathbb{R}^N$ distributed renewables and demands
$\mathbf{g}_t \in \mathbb{R}^{N_G}$ fuel-based generation
$\mathbf{p}_t \in \mathbb{R}^M$ line power flows
$\mathbf{b}_t, \mathbf{l}_t \in \mathbb{R}^N$ injections and levels of storage devices
$\mathbf{a}_t,\ \mathbf{k}_t \in \mathbb{R}^N$, $\mathbf{c}_t \in \mathbb{R}^{N_G}$, prices of $\mathbf{b}_t$, $\mathbf{r}_t$ and $\mathbf{g}_t$.

This work is supported by ANII-Uruguay, grants FSE_1_2019_1_159457 and FCE_1_2021_1_167301.

B. Dynamic programming formulation

Consider a transmission grid with N buses, M lines, and $N_s \leq N$ storage devices. For each bus $n = 1, \ldots, N$, we are given exogenous variables $d_t(n)$ and $r_t(n)$ representing demand and renewable generation. In turn, we must determine the variables $g_t(n)$ of fuel-based generations and $b_t(n)$ of injections from storage devices, to minimize the cost

$$E\left[\sum_{t=0}^{T-1} \mathbf{c}_t^T \mathbf{g}_t + \mathbf{k}_t^T \mathbf{r}_t\right], \tag{1}$$

where $\mathbf{c}_t$ and $\mathbf{k}_t$ represent the time-varying prices of fuel-based and renewable generation, respectively, and the expected value accounts for the randomness of $\mathbf{d}_t$ and $\mathbf{r}_t$.

A major source of coupling over time in the above optimization is the dynamics of charge/discharge of storage units:

$$\mathbf{l}_{t+1} = \mathbf{l}_t - \Delta T \mathbf{b}_t. \tag{2}$$

In addition, there may be time correlation in the demand and renewable processes; we will treat these as Markovian, i.e. where current values of $\mathbf{r}_t$, $\mathbf{d}_t$ define an appropriate state.

Variables $\mathbf{g}_t$, $\mathbf{r}_t$, $\mathbf{b}_t$, $\mathbf{d}_t$ are jointly constrained by power balance conditions. These take place over a grid with transmission lines $m = 1, \ldots, M$, that carry power flows $p_t(m)$, subject to capacity constraints

$$|p_t(m)| \leq \bar{p}(m). \tag{3}$$

To relate these constraints to the variables of our dynamic program requires a *power flow* model, which is now presented.

C. DC power flow model

We adopt the classical DC-power flow model for the grid [17]. Dropping the time index momentarily, the active power flowing on the lines satisfies $\mathbf{p} = \mathbf{X}^{-1}\mathbf{A}\boldsymbol{\theta} \in \mathbb{R}^M$ in terms of the voltage angles $\boldsymbol{\theta} \in \mathbb{R}^N$, and the node balances are collected in $\mathbf{A}^T\mathbf{p} = \mathbf{H}\mathbf{g} + \mathbf{r} + \mathbf{b} - \mathbf{d} \in \mathbb{R}^N$. The adjacency matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ specifies the connectivity of the grid, with all null entries except for $A(m,n) = 1$ and $A(m,n') = -1$, which define the sign of the flow on line m to be positive when moving from n to n'. Matrix $\mathbf{X} \in \mathbb{R}^{M \times M}$, is diagonal with $X(m,m) = X_m$ representing the reactance of line m, $m = 1, \ldots M$, and $\mathbf{H} \in \mathbb{R}^{N \times N_G}$ represents the matrix of zeros and ones that assigns generators to buses.

Although the equations for line flows and node balances described above are enough to impose capacity constraints, the following equivalent conditions remove the unnecessary explicit reference to the angles $\boldsymbol{\theta}_t$:

$$\mathbf{p} = \mathbf{F}(\mathbf{H}\mathbf{g} + \mathbf{r} + \mathbf{b} - \mathbf{d}), \tag{4}$$

$$\mathbf{1}^T\mathbf{d} = \mathbf{1}^T\mathbf{H}\mathbf{g} + \mathbf{1}^T\mathbf{b}, \tag{5}$$

where $\mathbf{F} = \mathbf{X}^{-1}\mathbf{A}\mathbf{L}^{\dagger} \in \mathbb{R}^{M \times N}$ is the matrix of *distribution factors*, defined in terms of the pseudo-inverse of the newtork Laplacian $\mathbf{L} = \mathbf{A}^T\mathbf{X}^{-1}\mathbf{A}$. For more details, we refer to [18].

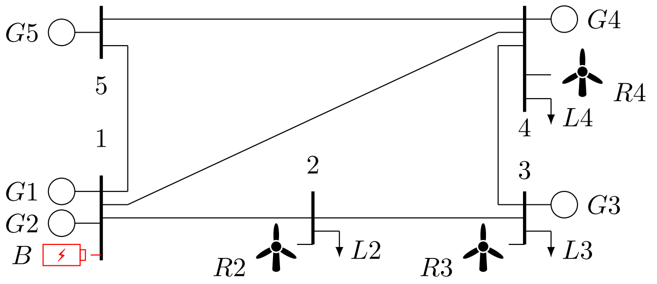

Fig. 1. Reduced PJM grid model [19] comprised of $N = 5$ buses and $M = 6$ lines connecting $N_G = 5$ fuel-based generators $G_1 - G_5$, one storage device and B, and three renewable sources $R2 - R4$ to three loads loads $L2 - L4$.

D. Reduced models

The number of optimization variables may be reduced if we assume that the grid operator has identified a subset of lines that typically operate at near capacity and the complementary set of lines that are safely over-dimensioned so their constraints do not activate. In this situation, we can partition the power flow vector as $\mathbf{p} = (\mathbf{p}_A, \mathbf{p}_B)$. The rows of (4) can be partitioned accordingly, and those corresponding to $\mathbf{p}_B$ can be discarded. The remaining model is $\mathbf{p}_A = \mathbf{F}_{AA}(\mathbf{H}_A\mathbf{g} + \mathbf{r}_A + \mathbf{b}_A + \mathbf{F}_{AB}(\mathbf{H}_B\mathbf{g} + \mathbf{r}_B) - \mathbf{d}'_A$, where node variables $\mathbf{r}$, $\mathbf{b}$, $\mathbf{d}$, and matrices $\mathbf{F}$ and $\mathbf{H}$ are partitioned accordingly, and with the loads across the network collapsed in the lower dimensional vector $\mathbf{d}'_A = \mathbf{F}_{AA}\mathbf{d}_A + \mathbf{F}_{AB}\mathbf{d}_B$.

All derivations henceforth will consider the entire network, but they admit lower dimensional counterparts in terms of the reduced model just described.

III. Distributed optimization and learning

The optimal control policy searches for the grid flows that minimize (1), with variables satisfying the instantaneous constraints (3),(4),(5), and subject to *dynamic*, inter-temporal constraints of two kinds: the storage balance (2), and the stochastic dependence of $\mathbf{d}_t$ and $\mathbf{r}_t$ on its past values. Since the latter are difficult to model, we apply RL to learn model-free based on data. It is non-trivial, however, to retain the partial model (2),(3),(4),(5) within the RL framework. Our main idea to allow this combination is to use as action variable the storage discharging *price* $\mathbf{a}_t$ per time t.

To fix ideas, imagine that the storage was not controllable but was managed by an external operator that sells and buys power to the grid operator at price $\mathbf{a}_t$. Given $\mathbf{a}_t = \mathbf{a}$, the grid operator would solve the following problem

$$\begin{aligned} \min_{(\mathbf{g},\mathbf{p},\mathbf{b}) \in \mathcal{C}} \quad & \mathbf{c}^T\mathbf{g} + \mathbf{k}^T\mathbf{r} + \mathbf{a}^T\mathbf{b} \\ \text{s. to: } & \mathbf{p} = \mathbf{F}(\mathbf{H}\mathbf{g} + \mathbf{r} + \mathbf{b} - \mathbf{d}) \\ & \mathbf{1}^T\mathbf{d} = \mathbf{1}^T(\mathbf{H}\mathbf{g} + \mathbf{r} + \mathbf{b}) \\ & \mathbf{0} \leq \mathbf{l} - \mathbf{b}\Delta T \leq \bar{\mathbf{l}} \end{aligned} \tag{6}$$

with $\mathcal{C} = \{\mathbf{0} \leq \mathbf{g} \leq \bar{\mathbf{g}},\ -\bar{\mathbf{p}} \leq \mathbf{p} \leq \bar{\mathbf{p}},\ -\bar{\mathbf{b}} \leq \mathbf{b} \leq \bar{\mathbf{b}}\}$.

The linear optimization problem (6) is not coupled across time, and solving it can be conceived as an operator that takes demand, renewables, storage levels, and prices as inputs and

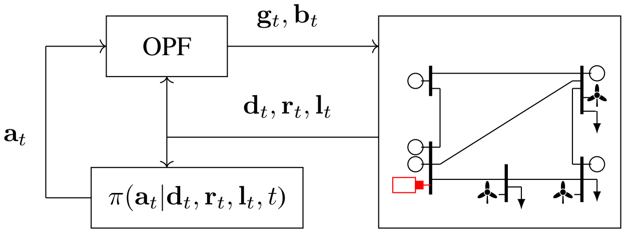
OPF
$\mathbf{g}_t, \mathbf{b}_t$
$\mathbf{d}_t, \mathbf{r}_t, \mathbf{l}_t$
$\mathbf{a}_t$
$\pi(\mathbf{a}_t|\mathbf{d}_t, \mathbf{r}_t, \mathbf{l}_t, t)$

Fig. 2. Feedback loop between the grid, OPF optimizer, and RL policy.

returns the optimized injections power injections from fuel-based generators and storage devices, i.e.,

$$(\mathbf{g}_t, \mathbf{b}_t) = OPF(\mathbf{d}_t, \mathbf{r}_t, \mathbf{l}_t, \mathbf{a}_t). \tag{7}$$

A. Learning the prices of storage

With this operator at hand, the remaining challenge is to learn the storage manager policy that adapts the prices $\mathbf{a}_t$ to the state variables $(\mathbf{d}_t, \mathbf{r}_t, \mathbf{l}_t)$ to optimize (1). This part fits nicely in the RL framework [8], since it is comprised of

- An expected cost (1) to optimize across time.
- Dynamics of loads, renewables, and storage $(\mathbf{d}_t, \mathbf{r}_t, \mathbf{b}_t)$
- Randomness of renewables and demand $(\mathbf{d}_t, \mathbf{r}_t)$
- Control actions $\mathbf{a}_t$ representing prices of stored power.

Specifically, we want to learn a parametric random policy $\pi_{\boldsymbol{\lambda}}$ [8, p.312], with parameter $\boldsymbol{\lambda} \in \mathbb{R}^P$ such that

$$\mathbf{a}_t \sim \pi_{\boldsymbol{\lambda}}(\mathbf{d}_t, \mathbf{r}_t, \mathbf{l}_t, t). \tag{8}$$

The control outputs produced by this policy $\pi_{\boldsymbol{\lambda}}$ are regarded as actions in the literature of RL, and these actions are driven by the state of the system $(\mathbf{d}_t, \mathbf{r}_t, \mathbf{l}_t)$. The main reason to randomize the policy in RL, instead of searching for a deterministic control law, is to be able to take exploratory random actions $\mathbf{a}_t$ at the beginning of the learning process when the best policy is still unknown and progressively adapt the parameter $\boldsymbol{\lambda}$ taking into account the costs $\mathbf{c}_t^T\mathbf{g}_t + \mathbf{k}_t^T\mathbf{r}_t$ resulting from these actions. Deciding the optimal actions amounts to finding the policy parameters that solve

$$\min_{\lambda \in \mathbb{R}^P} E_{\pi_\lambda}\left[\sum_{t=1}^{T} \mathbf{c}_t^T\mathbf{g}_t + \mathbf{k}_t^T\mathbf{r}_t\right] \tag{9}$$

$$\text{s. to: } \ (\mathbf{g}_t, \mathbf{b}_t) = OPF(\mathbf{d}_t, \mathbf{r}_t, \mathbf{l}_t, \mathbf{a}_t) \tag{10}$$

where the expectation is taken with respect to the policy and the distribution of the stochastic processes of $\mathbf{d}_t$ and $\mathbf{r}_t$.

The feedback loop between the power system, the OPF optimizer, and the price policy $\pi_{\boldsymbol{\lambda}}$ is illustrated in Fig. 2. The controller is divided into two blocks. The control policy $\pi_{\boldsymbol{\lambda}}(\cdot)$ measures the state ($\mathbf{d}_t$, $\mathbf{r}_t$, $\mathbf{l}_t$) from the grid and declares a control action $\mathbf{a}_t$. This $\mathbf{a}_t$ does not control the system directly, but it is converted by the second block (OPF), with additional inputs ($\mathbf{d}_t$, $\mathbf{r}_t$, $\mathbf{l}_t$), into the pair ($\mathbf{g}_t$,$\mathbf{b}_t$) that is used to operate the grid. After an interval ΔT, the system transits into a new state ($\mathbf{d}_{t+1}$, $\mathbf{r}_{t+1}$, $\mathbf{l}_{t+1}$) following the unknown dynamics of the demand and renewables, and the storage charge/discharge (2). Then, a new control cycle begins.

Algorithm 1 RL-OPF

```
repeat
    for $t = 0, \dots, T-1$ do
        Draw $a_t \sim \pi_\lambda(a|\mathbf{d}_t, \mathbf{r}_t, \mathbf{l}_t, t)$
        Solve (6) and obtain $G_t = \mathbf{c}_t^T\mathbf{g}_t + \mathbf{k}_t^T\mathbf{r}_t$
        Update $\lambda$ −= $G_t\nabla_\lambda \log\pi(a_t|\mathbf{d}_t, \mathbf{r}_t, \mathbf{l}_t, t)$ (13)
    end for
until convergence
```

Substituting (10) into (9) the problem becomes unconstrained, with cost

$$\mathbf{V}(\boldsymbol{\lambda}) = E_{\pi_\lambda}\left[\sum_{t=1}^{T} \mathbf{c}_t^T\mathbf{g}_t\left(\mathbf{d}_t, \mathbf{r}_t, \mathbf{l}_t, \mathbf{a}_t\right) + \mathbf{k}_t^T\mathbf{r}_t\left(\mathbf{d}_t, \mathbf{r}_t, \mathbf{l}_t, \mathbf{a}_t\right)\right]$$

where functions $\mathbf{g}_t(\cdot)$ and $\mathbf{r}_t(\cdot)$ are the solutions to (6). Hence, the optimal $\boldsymbol{\lambda}$ is obtained by minimizing the unconstrained value function $V(\boldsymbol{\lambda})$ via stochastic gradient descent. One major breakthrough in the literature of RL was to obtain the gradient of $V(\boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$, that is $\nabla_{\boldsymbol{\lambda}}V(\lambda) = E\left[G_t\nabla_\lambda \log\pi_\lambda(\mathbf{a}_t|\mathbf{d}_t, \mathbf{r}_t, \mathbf{l}_t)\right]$, with $G_t = \sum_{t=1}^{T}\mathbf{c}_t^T\mathbf{g}_t + \mathbf{k}_t^T\mathbf{r}_t$ [8, p.327]. Still, with this form of $\nabla_{\boldsymbol{\lambda}}V(\boldsymbol{\lambda})$, the gradient descent algorithm is not implementable since the expected value in $\nabla_{\boldsymbol{\lambda}}V(\lambda)$ also depends of the state transition probabilities of $\mathbf{d}_t$ and $\mathbf{r}_t$, and these transitions are unknown under our RL working assumptions. To bypass this hindrance, the RL policy gradient method resorts to a stochastic version of gradient descent such that the expectation in $\nabla_{\boldsymbol{\lambda}}V(\lambda)$ is dropped, and the parameter $\boldsymbol{\lambda}$ is updated in the direction of [8, p.328]

$$\hat{\nabla}_{\boldsymbol{\lambda}}V(\lambda) = G_t\nabla_\lambda \log\pi_\lambda(\mathbf{a}_t|\mathbf{d}_t, \mathbf{r}_t, \mathbf{l}_t) \tag{11}$$

All the information to compute (11) is available. Specifically, given samples $\mathbf{d}_t$ and $\mathbf{r}_t$, we measure $\mathbf{l}_t$ and compute the cost G_t solving (6) sequentially following the loop in Fig. 2. In addition, π_λ and $\mathbf{a}_t$ are part of our design, so they are also at hand. This results in the stochastic Algorithm 1.

It remains to design π_λ. A sensible choice should result in a simple closed form for $\nabla_\lambda \log\pi_\lambda$ in (11). The standard choices in RL are Gaussian when the actions are modeled as continuous random variables and *soft-max* (defined below) when the actions live in a finite set. To guide this decision, we introduce a viable simplification when all storage is injected in a single bus, and the costs $\mathbf{c}_t$ take values on a finite set. In this case, it is sufficient to consider a_t as a variable taking values on a finite set $\mathcal{A} = \{\bar{a}_1, \dots, \bar{a}_P\} \subset R$. The rationale for this is that the nodal price α at the storage bus, defined as a variable of the dual problem to (6) only takes a finite number of values $\alpha_1, \alpha_2, \dots, \alpha_P$ related to the costs of those generators that are active. Hence, if two different storage prices a and a' are such that $a, a' \in (\alpha_p, \alpha_{p+1})$, then the fuel-based generation induced by these prices coincide, that is, $\mathbf{g} = \mathbf{g}'$, when $(\mathbf{g}, \mathbf{b}) = OPF(\mathbf{d}, \mathbf{r}, a)$ and $OPF(\mathbf{d}, \mathbf{r}, a') = (\mathbf{g}', \mathbf{b}')$. As long as a is in this interval, the set of active generators is unchanged, and thus, the amount of power injected from fuel-based generators and storage is determined by the demand but

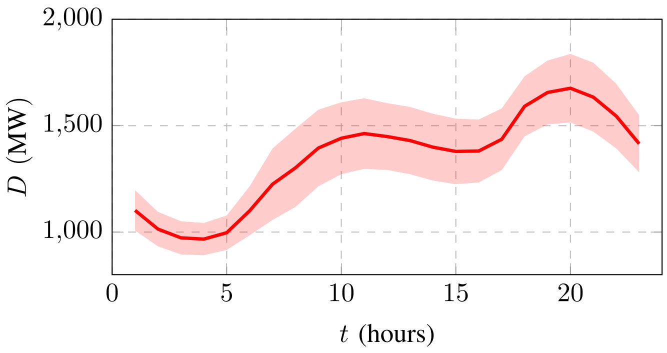

Fig. 3. Average and dispersion of the aggregate demand as a function of the time of the day during the winters of 2017-2019 [20]. $D_t = \sum_{n=1}^{N} d_t - \sum_{n=1}^{N} r_t$ represents the net demand, subtracting the renewable generation.

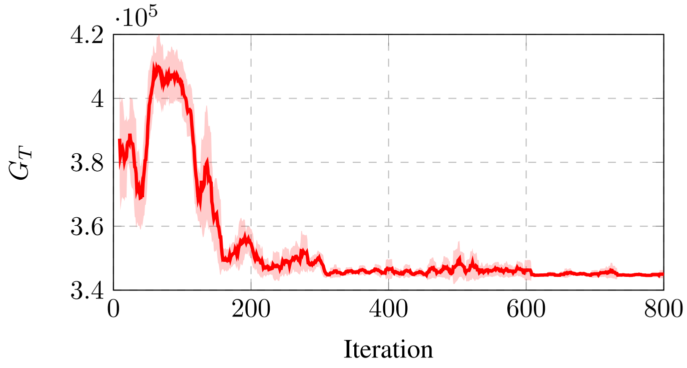

Fig. 4. Mean and deviation of the moving average of G_T averaged across 10 episodic iterations. The parameters $\boldsymbol{\lambda}$ evolve to a stabilized minimum cost after approximately 300 iterations.

not by the price a. We will adopt this simplification, which facilitates the convergence of RL. But if the storage is not injected in one bus, then the Gaussian distribution can be used instead.

B. Soft-max RL policy

If $\mathcal{A}$ is the finite set of mid-points $\bar{a}_p = (\alpha_{p-1} + \alpha_p)/2, p = 2, \ldots, P$, we adopt a discrete soft-max policy [8, p.322]

$$a_t \sim \pi_{\boldsymbol{\lambda}}(\bar{a}_p \mid \mathbf{d}_t, \mathbf{r}_t, \mathbf{l}_t, t) = \frac{e^{\mu_p(\mathbf{d}_t, \mathbf{r}_t, \mathbf{l}_t, t)}}{\sum_{p'=1}^{P-1} e^{\mu_{p'}(\mathbf{d}_t, \mathbf{r}_t, \mathbf{l}_t, t)}}. \quad (12)$$

with $\mu_p(\mathbf{d}_t, \mathbf{r}_t, \mathbf{l}_t, , t),\ p = 1, \ldots, P-1$ being the outputs of a parametric vector-valued function $\boldsymbol{\mu}_{\boldsymbol{\lambda}}(\mathbf{d}, \mathbf{r}, \mathbf{l}, t) \in \mathbb{R}^{P-1}$. The log-derivative for the soft-max policy is given by [8, p.329]

$$\nabla_\lambda \log \pi(a_t|\mathbf{d}_t, \mathbf{r}_t, \mathbf{l}_t, t) = \sum_p 1[a_t = \bar{a}_p] \nabla_{\boldsymbol{\lambda}} \mu_p(\mathbf{d}_t, \mathbf{r}_t, \mathbf{l}_t, t)$$
$$- \sum_p \pi(\bar{a}_p|\mathbf{d}_t, \mathbf{r}_t, \mathbf{l}_t, t) \nabla_{\boldsymbol{\lambda}} \mu_p(\mathbf{d}_t, \mathbf{r}_t, \mathbf{l}_t, , t). \quad (13)$$

Although different parameterizations of $\boldsymbol{\mu}_{\boldsymbol{\lambda}}$ have been proposed, including linear basis expansions and reproducing kernel Hilbert spaces, the state of the art has moved towards neural networks [8, p. 224]. In particular, we used a fully connected network with one hidden layer. In this case, $\nabla_{\boldsymbol{\lambda}} \mu_p$ in (13) is obtained via back-propagation [8, p. 238].

Remark: The RL-OPF strategy described above has the following attractive properties. It decouples the learning process from the OPF optimization. By these means, it retains the well-accepted DC-OPF model for the grid. In turn, RL handles the unknown statistics for demand and renewables with the mild Markov assumption that the distribution of $\mathbf{d}_t$ and $\mathbf{r}_t$ is conditionally independent of the history given $\mathbf{d}_{t-1}$ and $\mathbf{r}_{t-1}$.

IV. Numerical Experiments

We tested our RL-OPF algorithm on the grid of Fig. 1.

Lines $m = 1$,2,3,4,5, and 6, are oriented from bus 5 to bus 4, 5 to 1, 1 to 4, 4 to 3, 1 to 2, and 3 to 2, respectively. With this line numbering, reactances are given by $X_1 = 0.0297$, $X_2 = 0.0064$, $X_3 = 0.0304$, $X_4 = 0.0297$, $X_5 = 0.0281$, $X_6 = 0.0108$, in per unit on a basis of 100MW and 150KV, and line capacities by $\mathbf{p} = (240, 400, 400, 400, 400, 400)$ MW.

Variables $\mathbf{d}_t$ and $\mathbf{r}_t$ are not simulated but formed from real data in [20]. These records are actual measurements of demand and wind power in the Uruguayan power grid during the winter months from June to August 2017-2019. Although these measurements do not correspond to the PJM grid in Fig. 1, using them for these experiments serves the purpose of corroborating that RL can adapt to real samples of renewables and demand and the Markov assumption is not critical. The aggregated net demand in [20] is presented as a time-varying random variable in Fig. 3 along the day. It shows a typical winter pattern where the wind power is abundant in the early morning hours and the demand peaks during the evening. Each record of aggregate demand in [20], for a particular time and day, is divided into 38%, 29%, and 33%, which are served from buses $n = 2,\ 3$, and 4, respectively in our test PJM grid. Similarly, the total renewable power in [20] is divided in 42%, 33%, and 25%, injected to the same buses.

The maximum $\bar{\mathbf{g}} = (40, 170, 520, 200, 600)$MW and prices $\mathbf{c}_t = (14, 100, 30, 40, 10)$ in \$/MWh describe the generators.

The storage system, located at bus $n = 1$, is purposely large, with capacity $\bar{l} = 1500$MWh and a maximum $\bar{b} = 300$MW. The finite set $\mathcal{A} = \{5, 15, 25, 28, 35, 45, 150\}$ of activation prices was determined experimentally by running OPF in (6) for multiple values of $l_t \in (0, \bar{l})$, $a_t \in (0, 150)$, and with the records of $\mathbf{d}_t$ and $\mathbf{r}_t$ described above. These experiments indicated that the nodal price α for bus 1 only takes values in the finite set $\{10, 18, 27, 30, 40, 100\}$ as l_t $\mathbf{d}_t$ and $\mathbf{r}_t$ move, and the set of active constraints does not change when a_t moves in between these α's, so it suffices to take prices $a_t \in \mathcal{A}$.

The exponents $\boldsymbol{\mu}_\lambda(\mathbf{d}_t, \mathbf{r}_t, \mathbf{l}_t, t)$ in (12) were designed as a fully-connected neural network with 1 hidden layer of 256 neurons, dropout 0.7 (see [8, p.226]), 8 inputs $(d_2, r_2, d_3, r_3, d_4, r_4, l_1, t)$, and $|\mathcal{A}| = 7$ outputs.

With these parameters and data, we run 800 iterations of the loop in Algorithm 1, collecting in each iteration an episode of $T = 24$ samples of states and actions in intervals of $\Delta T = 1$ hour. The evolution of Algorithm 1 is shown in Fig. 4, where the convergence of the parameters $\boldsymbol{\lambda}$ are manifested by the

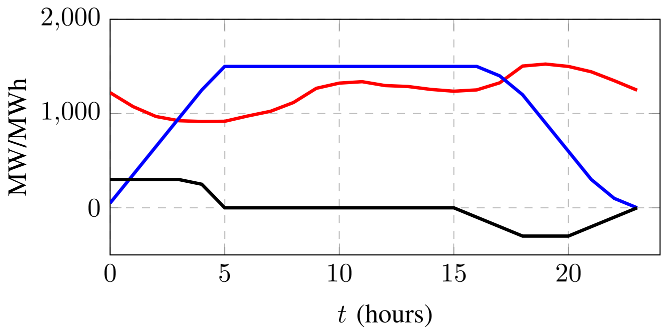

Fig. 5. Storage operation, as driven by the prices, effected from the learned policy; (red) total net demand in MW for a representative day; (black)) flow b_t in MW from and into the storage; (blue) resulting storage level in MWh.

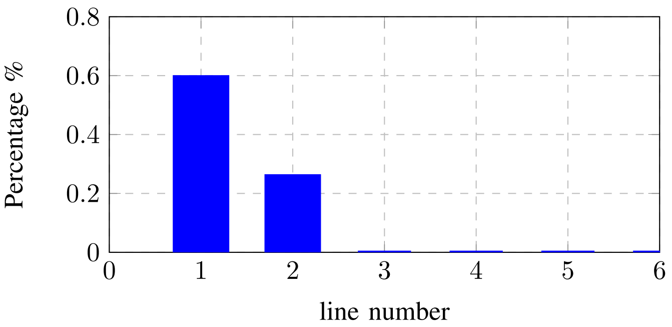

Fig. 6. Percentage of time instants in which the line constraints (3) are activated in the solution to OPF (6).

stabilization to the average cost in (1) around \$344500. Upon convergence of the parameters $\boldsymbol{\lambda}$, the resulting policy $\pi_{\boldsymbol{\lambda}}$ induces the energy arbitrage shown in Fig. 5. Specifically, in Fig. 5, we show how the storage is operated for a representative day of demand and renewables in the records of [20]. With the storage started at null-level, the prices given by $\pi_{\boldsymbol{\lambda}}$ result in the storage charging on the first hours of the morning when the net demand is low because of high wind power, and then discharging during the hours of high demand. Finally, Fig. 6 shows that the constraints in (6) activate for lines $m = 1$ and $m = 2$. The OPF optimizer ensured that the grid constraints were accounted for and satisfied. This observation highlights the pertinence of the combined RL-OPF strategy proposed here since OPF alone could not accommodate the dynamics, and unconstrained RL alone would produce power flows that could not be implemented without overloading the system. In addition to the grid, the battery model (capacity and power rate) is also factored in. In general, the OPF module can handle more general modeling assumptions, e.g., battery efficiencies, transmission losses, and AC-OPF models, to name a few.

V. Conclusions

We proposed a combined strategy for controlling storage systems that enforces grid constraints via DC-OPF and learns from data of demand and renewables via reinforcement learning. These two optimization techniques interact through the price of storage, which is adapted to the data. By these means, the proposed RL-OPF algorithm bypasses the need for an accurate stochastic dynamic model for wind and solar power. We tested this strategy in the PJM five-bus system with records of real data from the Uruguayan power grid, converging to a policy that effects an intuitive arbitrage of energy along the day. The chosen OPF formulation with distribution factors is amenable for a model reduction in case of large networks focusing on parts of the grid that are known to overload.

References

[1] S. Vazquez, S. M. Lukic, E. Galvan, L. G. Franquelo, and J. M. Carrasco, “Energy storage systems for transport and grid applications,” *IEEE Trans. on Industrial Electronics*, vol. 57, no. 12, pp. 3881–3895, 2010.

[2] D. Kottick, M. Blau, and D. Edelstein, “Battery energy storage for frequency regulation in an island power system,” *IEEE Transactions on Energy Conversion*, vol. 8, no. 3, pp. 455–459, 1993.

[3] N. Löhndorf and S. Minner, “Optimal day-ahead trading and storage of renewable energies—an approximate dynamic programming approach,” *Energy Systems*, vol. 1, pp. 61–77, 2010.

[4] F. Keck, M. Lenzen, A. Vassallo, and M. Li, “The impact of battery energy storage for renewable energy power grids in australia,” *Energy*, vol. 173, pp. 647–657, 2019.

[5] A. V. Savkin, M. Khalid, and V. G. Agelidis, “A constrained monotonic charging/discharging strategy for optimal capacity of battery energy storage supporting wind farms,” *IEEE Transactions on Sustainable Energy*, vol. 7, no. 3, pp. 1224–1231, 2016.

[6] G. Pulazza, N. Zhang, C. Kang, and C. A. Nucci, “Transmission planning with battery-based energy storage transportation for power systems with high penetration of renewable energy,” *IEEE Transactions on Power Systems*, vol. 36, no. 6, pp. 4928–4940, 2021.

[7] D. Gayme and U. Topcu, “Optimal power flow with large-scale storage integration,” *IEEE Tran. Power Sys.*, vol. 28, no. 2, pp. 709–717, 2012.

[8] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[9] F. S. Gorostiza and F. M. Gonzalez-Longatt, “Deep reinforcement learning-based controller for soc management of multi-electrical energy storage system,” *IEEE Trans. on Smart Grid*, vol. 11, no. 6, pp. 5039–5050, 2020.

[10] R. Xiong, J. Cao, and Q. Yu, “Reinforcement learning-based real-time power management for hybrid energy storage system in the plug-in hybrid electric vehicle,” *Applied Energy*, vol. 211, pp. 538–548, 2018.

[11] H. Wang and B. Zhang, “Energy storage arbitrage in real-time markets via reinforcement learning,” in *2018 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, 2018, pp. 1–5.

[12] J. Yang, M. Yang, M. Wang, P. Du, and Y. Yu, “A deep reinforcement learning method for managing wind farm uncertainties through energy storage system control and external reserve purchasing,” *Int. Journal of Electrical Power & Energy Systems*, vol. 119, p. 105928, 2020.

[13] E. Oh and H. Wang, “Reinforcement-learning-based energy storage system operation strategies to manage wind power forecast uncertainty,” *IEEE Access*, vol. 8, pp. 20 965–20 976, 2020.

[14] O. Mégel, G. Andersson, and J. L. Mathieu, “Reducing the computational effort of stochastic multi-period dc optimal power flow with storage,” in *IEEE Power Sys. Comp. Conf. (PSCC)*, 2016, pp. 1–7.

[15] E. Dall’Anese, K. Baker, and T. Summers, “Chance-constrained ac optimal power flow for distribution systems with renewables,” *IEEE Transactions on Power Systems*, vol. 32, no. 5, pp. 3427–3438, 2017.

[16] S. Paternain, M. Calvo-Fullana, L. F. Chamon, and A. Ribeiro, “Safe policies for reinforcement learning via primal-dual methods,” *IEEE Transactions on Automatic Control*, vol. 68, no. 3, pp. 1321–1336, 2022.

[17] J. D. Glover, M. S. Sarma, and T. Overbye, *Power system analysis & design, SI version*. Cengage Learning, 2012.

[18] S. H. Low, “Lecture notes for power system analysis,” http://netlab.caltech.edu/book/book.html.

[19] F. Li and R. Bo, “Small test systems for power system economic studies,” in *IEEE PES General Meeting*, 2010, pp. 1–4.

[20] A. Castellano and J. A. Bazerque, “Learning the operation of energy storage systems from real trajectories of demand and renewables,” in *IEEE Innovative Smart Grid Technologies Conf. (ISGT)*, 2020, pp. 1–5.