

DER Allocation without Load Prediction via Reinforcement Learning

Abed AlRahman Al Makdah, Aravind Ramana, Shaofeng Zou, Oliver Kosut, and Lalitha Sankar

Abstract—The growing variability of renewable generation increases the need for fast and flexible grid-balancing mechanisms. Existing frameworks for distributed energy resource aggregations (DERAs) rely on short-term forecasts of net demand, making their performance highly sensitive to prediction errors. In this paper we present a forecast-free reinforcement learning (RL) framework for DERA allocation that learns optimal policies directly from operational data. We model the DERA dynamics as a deterministic linear system and the exogenous net load as a feature-based linear Markov process, capturing short-range temporal dependencies without explicit forecasting. We derive a closed-form expression for the optimal policy, which is learned through a least-squares value iteration (LSVI) algorithm using data collected across episodes. The proposed framework preserves the interpretability and constraint satisfaction of DER model while adapting to stochastic demand variations through data-driven updates. Numerical experiments on real California Independent System Operator (CAISO) net-demand data demonstrate that the learned controller achieves high tracking accuracy and stable regulation across heterogeneous DER aggregators without requiring any demand prediction.

Index Terms—Reinforcement learning, DER allocation.

I. INTRODUCTION

The growth of renewable generation introduces uncertainty in the power grid leading to increased variability and ramping requirements in net demand. Traditionally, such variability is addressed using fossil fuel-based generators such as coal and natural gas, which are both costly and environmentally unsustainable. A promising alternative is to leverage distributed energy resources (DERs) — including buildings, thermostatically controlled loads, storage devices, and electric vehicles (EV)— via DER aggregators (DERAs). An aggregator can provide fast, flexible responses when it coordinates across different DER types effectively. Current approaches to coordinating DERs are predominantly forecast-driven. Forecast errors not only reduce tracking performance but also complicate aggregator operation, as errors propagate through ramping and state-of-charge dynamics.

In this paper, we develop a forecast-free alternative. In particular, we develop a reinforcement learning (RL) framework that learns an allocation policy directly from operational data, rather than predicting future demand. We model the net-load as a feature-based linear Markov process that captures its short-range dependence. At a DERA, each DER is represented by a generalized battery model with state-of-charge leakage and ramp-rate dynamics, yielding a tractable deterministic linear system. Combining the two yields a hybrid structure

involving the classical deterministic linear dynamics of the device combined with a stochastic Markovian demand. The resulting model allows us to combine a quadratic tracking objective that preserves the analytical simplicity of the linear quadratic regulator (LQR) while exploiting policy updates using reinforcement learning.

Related work. A substantial portion of the literature relies on forecast-based, finite-horizon optimal control. In [1], a nonlinear AC optimal power flow including thermostatically controlled loads is cast as a Markov decision process and reformulated as a convex finite-horizon problem. Similar setups appear in [2], [3]. In [4], a centralized model predictive control (MPC) framework is used for DERA allocation. Distributed MPC formulations have been proposed for economic dispatch [5] and frequency regulation [6]–[8], but these works either omit DERs or focus on device-level tracking without aggregate battery models or grid-level allocation. In [9], a packetized energy management (PEM) Macro model approach is introduced. Unlike forecast-based or finite-horizon optimization schemes, the PEM framework matches demand using myopic, real-time modulation of packet acceptance probabilities rather than solving predictive control problems. Recent work has increasingly explored reinforcement learning and data-driven methods for grid flexibility and distributed energy resource coordination. In [10], a multi-agent deep RL strategy for coordinating inverter-based DERs to maintain voltage stability under high renewable penetration is proposed, which demonstrates the ability of RL to handle nonlinearity and uncertainty in distribution networks. In [11], a deep RL-based residential demand response controller that learns consumption strategies under stochastic load and price signals is developed. In [12], a data-driven scheduling model for a virtual power plant that participates in day-ahead and real-time markets is proposed. These formulations generally rely on explicit forecasts or stochastic models of prices, loads, or renewable outputs and often optimize economic or network-regulation objectives over long horizons. In contrast, our work develops a forecast-free RL controller that directly allocates DERA in real time to track an exogenous load demand signal, without relying on load forecasting. This shifts the focus to model-free demand matching using online data.

Contributions. This paper features three contributions. First, we eliminate the need for demand forecasts by coupling DERA dynamics with a feature-based linear Markov model for the net-load, allowing adaptive allocation policies. Second, we derive a closed-form expression for the optimal policy, which is learned and updated using online data, without requiring explicit identification of the demand model. Finally, we demonstrate the effectiveness of our framework through numerical experiments on real net-load data.

Notation. A Gaussian random variable x with mean μ and covariance Σ is denoted as $x \sim \mathcal{N}(\mu, \Sigma)$. The $n \times n$ identity

This material is based upon work supported by the National Science Foundation under Grant No. EPCN-2246658 and a PSERC grant S-114. A. A. Al Makdah, S. Zou, O. Kosut, and L. Sankar are with the School of Electrical, Computer and Energy Engineering at the Arizona State University, {aalmakda,zou,okosut,lsankar}@asu.edu. A. Ramana is with the Department of Physics at the Indian Institute of Technology Madras, ep23b003@smail.iitm.ac.in.

matrix is denoted by I_n . The expectation operator is denoted by $\mathbb{E}[\cdot]$. A positive definite (semidefinite) matrix A is denoted as $A \succ (A \succeq 0)$. The Kronecker product is denoted by $\otimes$.

II. PROBLEM FORMULATION

We consider a generalized battery model for each aggregator. The i -th DER aggregator obeys:

$$\begin{aligned} z_{i,t+1} &= \alpha_i z_{i,t} - \beta_i p_{i,t}, \quad t \in \{0, \dots, T-1\}, \\ p_{i,t+1} &= p_{i,t} + u_{i,t} \end{aligned} \quad (1)$$

where $z_{i,t} \in \mathbb{R}$ is the state of charge (SoC), $\beta_i \in \mathbb{R}$ is a sampling time constant, $\alpha_i \in [0, 1]$ is a leakage parameter, $p_{i,t} \in \mathbb{R}$ is the power supplied, and $u_{i,t} \in \mathbb{R}$ is the ramp rate at time $t \geq 0$. We re-write (1) in linear state-space form as

$$\underbrace{\begin{bmatrix} z_{i,t+1} \\ p_{i,t+1} \end{bmatrix}}_{x_{i,t+1}} = \underbrace{\begin{bmatrix} \alpha_i & -\beta_i \\ 0 & 1 \end{bmatrix}}_{A_i} \underbrace{\begin{bmatrix} z_{i,t} \\ p_{i,t} \end{bmatrix}}_{x_{i,t}} + \underbrace{\begin{bmatrix} 0 \\ 1 \end{bmatrix}}_{B_i} u_{i,t}.$$

For M aggregators, the complete linear state space model is

$$\begin{aligned} \underbrace{\begin{bmatrix} x_{1,t+1} \\ \vdots \\ x_{M,t+1} \end{bmatrix}}_{x_{t+1}} &= \underbrace{\begin{bmatrix} A_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & A_M \end{bmatrix}}_A \underbrace{\begin{bmatrix} x_{1,t} \\ \vdots \\ x_{M,t} \end{bmatrix}}_{x_t} \\ &+ \underbrace{\begin{bmatrix} B_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & B_M \end{bmatrix}}_B \underbrace{\begin{bmatrix} u_{1,t} \\ \vdots \\ u_{M,t} \end{bmatrix}}_{u_t}, \end{aligned} \quad (2)$$

where $x_t \in \mathcal{X} = \mathbb{R}^{2M}$ and $u_t \in \mathcal{U} = \mathbb{R}^M$. Let

$$y_t = \underbrace{\left(\mathbb{1}_M^\top \otimes \begin{bmatrix} 0 & 1 \end{bmatrix} \right)}_C x_t, \quad (3)$$

where $y_t \in \mathbb{R}$ is the total power supplied by the aggregators. We consider net load demand evolving according to the discrete-time Markov process

$$s_{t+1} | s_t \sim \mathbb{P}_t(s_{t+1} | s_t), \quad t \in \{0, \dots, T-1\}, \quad (4)$$

where $s \in \mathcal{S} \subseteq \mathbb{R}$ denotes net load demand and $\mathbb{P}_t(s' | s)$ denotes the transition probability from state s to s' . We introduce the following assumption on the Markov Process in (4).

Assumption II.1. (Linear Markov Process) Let $w \in \mathcal{S}^r$, $\phi : \mathcal{S}^r \rightarrow \mathbb{R}^d$ be a known feature vector, and $\mu_t \in \mathbb{R}^d$ a vector of d unknown signed measures over $\mathcal{S}$. For $s \in \mathcal{S}$, we have

$$\underbrace{\begin{bmatrix} s_{t-r+2} \\ \vdots \\ s_t \\ s_{t+1} \end{bmatrix}}_{\tilde{A}} = \underbrace{\begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \end{bmatrix}}_{\tilde{A}} \underbrace{\begin{bmatrix} s_{t-r+1} \\ \vdots \\ s_{t-1} \\ s_t \end{bmatrix}}_{w_t} + \underbrace{\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}}_{\tilde{B}} s_{t+1}, \quad (5)$$

$$\mathbb{P}_t(s_{t+1} | w_t) = \phi(w_t)^\top \mu_t(s_{t+1}), \quad (6)$$

We assume $\|\phi(w)\| \leq 1/\sqrt{d}$ and $\|s\| \leq \delta_s$ for all $s \in \mathcal{S}$, $\mathbb{E}[\phi(w_t)\phi(w_t)^\top] \succ 0$, and $\|\mu_t\| \leq 1$ for all t .

We consider the cost following cost at time t ,

$$c(x_t, s_t, u_t) = x_t^\top W_x x_t + u_t^\top R u_t + \kappa (s_t - y_t)^2 \quad (7)$$

where $W_x \succeq 0$, $R \succ 0$, and $\kappa > 0$. The cost in (7) can be written in the following form

$$c(x_t, s_t, u_t) = \underbrace{\begin{bmatrix} x_t \\ s_t \\ u_t \end{bmatrix}^\top \begin{bmatrix} W_x + \kappa C^\top C & -\kappa C^\top & 0 \\ -\kappa C & \kappa & 0 \\ 0 & 0 & R \end{bmatrix} \begin{bmatrix} x_t \\ s_t \\ u_t \end{bmatrix}}_P,$$

The aggregators follow a control policy $\pi_t : \mathcal{X} \times \mathcal{S}^r \rightarrow \mathcal{U}$, where $u_t = \pi_t(x_t, w_t)$ is the action that the agent takes at state x_t and s_t at time t , for $t \geq 0$. We seek to find an optimal control policy, $\pi = (\pi_0, \dots, \pi_T)$ for the following task:

$$\begin{aligned} \underset{\pi}{\text{minimize}} \quad & \mathbb{E} \left[\sum_{t=0}^T c(x_t, s_t, u_t) \right], \\ \text{subject to} \quad & x_{t+1} = Ax_t + Bu_t, \\ & s_{t+1} \sim \mathbb{P}_t(s_{t+1} | w_t), \\ & w_{t+1} = \hat{A}w_t + \hat{B}s_{t+1} \\ & u_t = \pi_t(x_t, w_t), \end{aligned} \quad (8)$$

with $u_T = \pi_T(x_T, w_T) = 0$. We define the state-action value function $Q_t^\pi : \mathcal{X} \times \mathcal{S}^r \times \mathcal{U} \rightarrow \mathbb{R}$ as the expected cumulative cost under policy π starting from state x_t , s_t , w_t , and action u_t at time t , given by

$$\begin{aligned} Q_t^\pi(x, s, w, u) &\triangleq c(x, s, u) \\ &+ \mathbb{E} \left[\sum_{i=t+1}^T c(x_i, s_i, \pi_t(x_i, w_i)) \middle| x_t = x, s_t = s, w_t = w, u_t = u \right]. \end{aligned}$$

III. REINFORCEMENT LEARNING ALGORITHM

In this section, we leverage the hybrid structure, deterministic linear DERA dynamics coupled with a feature-based linear Markov model of the net load and the quadratic structure of the cost, to derive a parametric expression for the optimal policy. Then, we introduce a least-squares value iteration (LSVI) algorithm to learn the parameters from online data. In [13], we derived a parametric expression of the state-action value function that is linear in the feature map, ϕ , along with a parametric expression for the corresponding optimal policy. In the next result, we present the parametrized expression of optimal policy, which is adopted from [13].

Theorem III.1. (Optimal policy representation [13]) For any $x \in \mathcal{X}$, $s \in \mathcal{S}$, $w \in \mathcal{S}^r$, and $t \in \{0, \dots, T-1\}$

$$\begin{aligned} u_t^*(x, w) &= \pi_t^*(x, w) \\ &= K_{t,x}x + K_{t,s}s + K_{t,h} \left(\phi(w)^\top \otimes Z \right) \theta_{t+1}, \end{aligned} \quad (9)$$

where $Z = [I_n, 0_{n \times 1}]$,

$$\begin{aligned} K_{t,x} &= -(R + B^\top G_{t+1}B)^{-1} (B^\top G_{t+1}A + D^\top), \\ K_{t,s} &= -(R + B^\top G_{t+1}B)^{-1} H^\top, \\ K_{t,h} &= -(R + B^\top G_{t+1}B)^{-1} B^\top, \end{aligned} \quad (10)$$

and G_{t+1} satisfies,

$$\begin{aligned} G_t &= A^\top G_{t+1} A + W \\ &-(A^\top G_{t+1} B + D)(R + B^\top G_{t+1} B)^{-1}(B^\top G_{t+1} A + D^\top), \end{aligned} \quad (11)$$

and $\theta_{t+1} \in \mathbb{R}^{d(n+1)}$.

Theorem III.1 shows that the optimal control at time t admits a feedback form in the physical DERA state, x , the current net load, s , and a feature summary of recent net-load history, w . The feedback gains, $K_{t,x}$ and $K_{t,s}$ are LQR-type expressions computed from a Riccati recursion for that incorporates the tracking cross-terms, while the feature feedforward term, $K_{t,h}$ captures how the linear Markov process shapes future costs through parameters θ_{t+1} . Next, we leverage this structure and propose a least-squares value iteration algorithm to learn θ directly from data, avoiding explicit demand forecasting. We adapt the least-squares value iteration algorithm (Algorithm 1) from our prior work [13]. Our algorithm consists of an outer loop over L episodes, where each episode consists of two loops: 1) backward-in-time weight update loop (lines 5-11) and 2) forward roll-out and data collection loop (lines 12-16). During the first pass of episode ℓ (lines 5-11), we treat the data collected in the previous $\ell-1$ episodes as a fixed dataset

$$\mathcal{D}_{\ell-1} := \{(x_t^i, s_t^i, u_t^i, x_{t+1}^i, s_{t+1}^i) : i < \ell, 0 \leq t < T\}. \quad (12)$$

At each time step t , θ minimizes a regularized least-squares loss. Solving this problem on past trajectory data yields an accurate value-function approximation and enables closed-form greedy policy updates without estimating the transition probabilities. The regularized least-squares regression is given by the following optimization solved in Algorithm 1.

$$\begin{aligned} \theta_{t+1}^\ell = \arg \min_{\theta \in \mathbb{R}^{d(n+1)}} & \sum_{i=1}^{\ell-1} \left(\phi(w_t^i)^\top Y(x_t^i, u_t^i) \theta \right. \\ & \left. - \epsilon_{t+1}^\ell(x_{t+1}^i, s_{t+1}^i, w_{t+1}^i) \right)^2 + \lambda \|\theta\|^2, \end{aligned} \quad (13)$$

Algorithm 1 Least-Squares Value Iteration

```

1: Given:  $L, R_\theta, \lambda$ 
2: for episode  $\ell = 1, \dots, L$  do
3:    $x_0^\ell = x_T^{\ell-1}$ 
4:    $w_0^\ell = w_T^{\ell-1}$ 
5:   for step  $t = T-1, \dots, 0$  do
6:      $\Lambda_t^\ell \leftarrow \sum_{i=1}^{\ell-1} Y(x_t^i, u_t^i)^\top \phi(w_t^i) \phi(w_t^i)^\top Y(x_t^i, u_t^i) + \lambda I_{dn+d}$ 
7:      $\theta_{t+1}^\ell \leftarrow (\Lambda_t^\ell)^{-1} \sum_{i=1}^{\ell-1} Y(x_t^i, u_t^i)^\top \phi(w_t^i) \epsilon_{t+1}^\ell(x_{t+1}^i, s_{t+1}^i, w_{t+1}^i)$ 
8:     if  $\|\theta_{t+1}^\ell\| > R_\theta$  then
9:        $\theta_{t+1}^\ell \leftarrow \frac{R_\theta}{\|\theta_{t+1}^\ell\|} \theta_{t+1}^\ell$ 
10:    end if
11:  end for
12:  for step  $t = 0, \dots, T-1$  do
13:     $u_t^\ell \leftarrow K_{t,x} x_t^\ell + K_{t,s} s_t^\ell + K_{t,h} (\phi(w_t^\ell)^\top \otimes Z) \theta_{t+1}^\ell$ 
14:    Take action  $u_t^\ell$ 
15:    Observe  $x_{t+1}^\ell, s_{t+1}^\ell$ , and  $w_{t+1}^\ell$ 
16:  end for
17: end for
```

where

$$\epsilon_{t+1}^\ell(x, s, w) = 2(x)^\top h_{t+1}^\ell(s, w) + q_{t+1}^\ell(s, w), \quad (14)$$

$$\begin{aligned} h_{t+1}^\ell(s, w) &= (A^\top + K_{t,x}^\top B^\top) (\phi(w)^\top \otimes Z) \theta_{t+2}^\ell \\ &+ (F + K_{t,x}^\top H^\top) s, \end{aligned} \quad (15)$$

$$\begin{aligned} q_{t+1}^\ell(s, w) &= \left(\phi(w)^\top \otimes \bar{Z} \right) \theta_{t+2}^\ell + s^\top (M + H K_{t,s}) s \\ &+ \theta_{t+2}^{\ell^\top} (\phi(w) \otimes Z^\top) B K_{t,h} (\phi(w)^\top \otimes Z) \theta_{t+2}^\ell \\ &+ 2s^\top H K_{t,h} \left(\phi(w)^\top \otimes Z \right) \theta_{t+2}^\ell, \end{aligned} \quad (16)$$

with $Y(x, u) = I_d \otimes [2(Ax + Bu)^\top, 1]$, $Z = [I_n, 0_{n \times 1}]$, and $\bar{Z} = [0_{1 \times n}, 1]$. And $\lambda > 0$ is the regularization parameter. The closed-form parameter update are given by lines 6 and 7 of Algorithm 1. At episode $\ell = 1$, we assume that we have access to the net-load values from time $t = -r + 1$ to $t = 0$ so that the initial vector w_0^ℓ is well defined for any $r \geq 1$. Further, at $\ell = 1$, we set $x_0^\ell \sim \mathcal{N}(0, I_n)$, $\theta_{t+1}^\ell \sim \mathcal{N}(0, I_{dn+d})$ and $\Lambda_t^\ell = \lambda I_{d(n+1)}$ for $t \in \{0, \dots, T-1\}$. The regularizer term $\lambda I_{d(n+1)}$ ensures numerical stability, the projection step in lines 8-10 makes sure that the norm of the learned parameters is uniformly bounded for $t \in \{0, \dots, T-1\}$ and $\ell \in \{1, \dots, L\}$. In the second pass (lines 12-16) the newly computed parameters θ_{t+1}^ℓ are plugged into the greedy closed-form policy (9) as shown in step 13 of Algorithm 1 to generate a new trajectory $(\{x_t^\ell, s_t^\ell, u_t^\ell\}_{t=0}^T)$. These samples are appended to the collected data (12), and will be used in the next episode's backward update.

The control formulation in this paper builds on our framework in [13], which establishes closed-form optimal policies for linear dynamics driven by feature-based Markov processes. In this work, we specialize our framework in [13] to real-time grid balancing with heterogeneous DER aggregations. In particular, we introduce a generalized battery and ramping model for DERAs, a forecast-free net-load tracking formulation using real CAISO data, and a regulation-oriented objective aligned with system operator dispatch. This enables real-time DERA allocation without short-term demand forecasting while preserving the structure and interpretability of classical LQR controllers. Furthermore, unlike [13], in this paper we model the evolution of the exogenous state s_{t+1} to be conditioned on the window w_t , which captures the short-range temporal dependencies required by the linear Markov process.

IV. NUMERICAL EXPERIMENTS

We adopt a similar experimental setup to [4]. For reference, we compare against a forecast-based MPC benchmark, similar to [4], which computes DERA allocations using short-term net-load predictions. Implementation details are omitted since MPC is used only as a baseline. We evaluate our forecast-free reinforcement learning framework using real net-demand data from the California Independent System Operator (CAISO) for the period from May 15, 2023 to September 15, 2023 [14]. The dataset contains net-load measurements (total demand minus renewable generation) sampled at each 5 minutes, represented by the matrix $S \in \mathbb{R}^{1 \times N_d}$, where N_d denotes the total number of samples, which is written as

$$S = [s_{-r+1} \quad \dots \quad s_0 \quad \dots \quad s_{N_d}]. \quad (17)$$

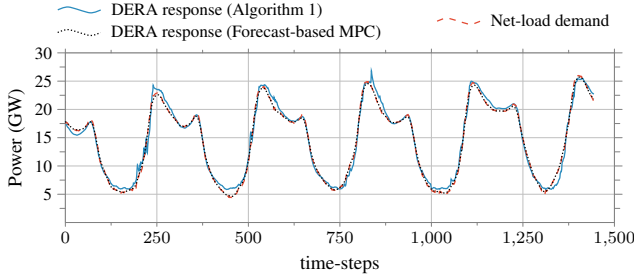


Fig. 1. This figure shows the net-load demand (dashed red line), the cumulative DERA response under the proposed forecast-free RL controller (solid blue line), and the cumulative DERA response under a forecast-based MPC benchmark (dotted black line) over a five day period, where each time-step corresponds to 5-minute interval. The cumulative DERA response corresponds to the output y in (3) when the system (2) is driven by the policy learned via Algorithm 1 over $L = 407$ episodes (blue line) and by the forecast-based MPC benchmark (dotted black line). We observe that the cumulative DERA responses track the net-load demand.

We partition the data into L episodes, each with a fixed time horizon $T = 289$ time-steps corresponding to one day of measurements, yielding $N_d = LT$. We consider $M = 5$ heterogeneous distributed energy resource aggregators — air conditioners (ACs), electric water heaters (E-WHs), building HVACs (bldgs), refrigerators (RFGs), and electric vehicles (EVs) — obeying the linear dynamics in (2), with the corresponding parameters summarized in Table I. We assume the net load from the CAISO data evolves according to the linear Markov process in (6) with $d=3$ and $r=2$, where the measures μ_t are unknown. We choose $\phi(\cdot)$ as follows

$$\phi(w_t) = \begin{bmatrix} f_1(w_t)/(f_1(w_t) + f_2(w_t) + f_3(w_t)) \\ f_2(w_t)/(f_1(w_t) + f_2(w_t) + f_3(w_t)) \\ f_3(w_t)/(f_1(w_t) + f_2(w_t) + f_3(w_t)) \end{bmatrix},$$

where

$$f_1(w_t) = 1, \quad f_2(w_t) = \exp\left(\frac{(w_t - \nu_2)^T \Sigma_2^{-1} (w_t - \nu_2)}{2r}\right),$$

$$f_3(w_t) = \exp\left(\frac{(w_t - \nu_3)^T \Sigma_3^{-1} (w_t - \nu_3)}{2r}\right).$$

In order to choose the means $\nu_2 \in \mathbb{R}^r$ and $\nu_3 \in \mathbb{R}^r$ and the covariance matrices $\Sigma_2 \in \mathbb{R}^{r \times r}$ and $\Sigma_3 \in \mathbb{R}^{r \times r}$, we first re-write S in (17) as

$$S_r = \begin{bmatrix} s_{-r+1} & s_{-r+2} & \cdots & s_{N_d-r+1} \\ s_{-r+2} & s_{-r+3} & \cdots & s_{N_d-r+2} \\ \vdots & \vdots & \ddots & \vdots \\ s_0 & s_1 & \cdots & s_{N_d} \end{bmatrix} = [w_0 \quad \cdots \quad w_{N_d}]. \quad (18)$$

Then, we use the K-means algorithm over the data w_t for $t \in \{0, \dots, N_d\}$ using $d-1 = 2$ clusters to obtain the feature centers ν_2 and ν_3 . For each cluster, we compute the sample covariance of the associated w_t vectors, which yields the covariance matrices Σ_2 and Σ_3 , respectively. We choose the weights of cost function in (7) as

$$W_x = \text{blkdiag}(W_x^1, W_x^2, W_x^3, W_x^4, W_x^5) \quad W_x^i = \begin{bmatrix} \kappa_z^i & 0 \\ 0 & \kappa_p^i \end{bmatrix},$$

where κ_z^i and κ_p^i are in Table I for $i \in \{1, \dots, 5\}$, $R = 10^4 I_5$ and $\kappa = 10^2$. We use Algorithm 1 to learn the weight parameters, θ , that parametrizes the optimal policy in (9). We set $L = 407$, $R_\theta = 10^4$, and $\lambda = 0.1$. For $\ell = 1$, we set $x_0^1 \sim \mathcal{N}(0, I_{10})$, $\Lambda_t^1 = \lambda I_{33}$, and $\theta_{t+1}^1 \sim \mathcal{N}(0, I_{33})$, for $t \in \{0, \dots, T-1\}$. During deployment, at each dispatch interval of 5 minutes, the system operator broadcasts the net-load signal s_t , DER aggregators report their aggregate states, $x_{i,t}$, and the controller transmits ramp-rate commands, $u_{i,t}$. Since the controller in (9) is in closed-form and requires only matrix-vector multiplications, the required latency is in the order of seconds. The framework avoids solving online optimization problems as in forecast-based MPC.

Fig. 1 shows the net-load demand (depicted by the dashed red line), the cumulative DERA response under the proposed forecast-free RL controller in Algorithm 1 (depicted by the solid blue line), and the cumulative DERA response under the forecast-based MPC benchmark (depicted by the dotted black line) over a five day period, where each time-step corresponds to 5-minute interval. The cumulative DERA response corresponds to the output y in (3) when the system (2) is driven by the policy learned via Algorithm 1 over $L = 407$ episodes (blue line) and by the forecast-based MPC benchmark (dotted black line). We observe that both cumulative DERA responses track the net-load demand. While the forecast-based MPC benchmark can achieve lower tracking error due to access to short-term predictions, the proposed forecast-free RL controller achieves comparable tracking performance without requiring demand forecasts or online optimization. Moreover, the RL controller continuously adapts to changes in net-load statistics through online learning (Remark 2), whereas forecast-based approaches rely on fixed prediction models that must be retrained when data distributions shift.

Fig. 2 and Fig. 3 show the power and SoC trajectories of each DERA, respectively, over a five day period, where each time-step corresponds to a 5-minute interval. These trajectories correspond to the state trajectories of the system (2) when driven by the policy learned via Algorithm 1 over $L = 407$ episodes. Fig. 4 shows the cumulative cost incurred by the policy learned from Algorithm 1 starting from the initial state x_0^ℓ , s_0^ℓ , and w_0^ℓ for episodes $1 \leq \ell \leq 407$. For $1 \leq \ell \leq 387$, we add a random exploration term sampled from $\mathcal{N}(0, 0.25I_5)$ to the policy in Algorithm 1 line 13, and for $387 < \ell \leq 401$, we add a random exploration term sampled from $\mathcal{N}(0, 0.0025I_5)$. We observe that the cumulative cost decreased as the number of episodes increases after the exploration term is removed.

Remark 1. (Need for exploration) In contrast to our earlier formulation in [13], in this work, the initial states are not i.i.d., but instead inherit temporal dependence from the underlying Markov process carried over from the previous episode. This lack of episodic reset weakens the natural excitation present in the data. To compensate, we incorporate a mild exploration term to ensure convergence of the learned weights.

Remark 2. (Dataset diversity and robustness) The CAISO net-load dataset that we use spans May-September 2023, which captures seasonal transitions (e.g., solar ramps, heat waves, and peak demand events). Our control policy is based on a structured linear policy rather than a black-box func-

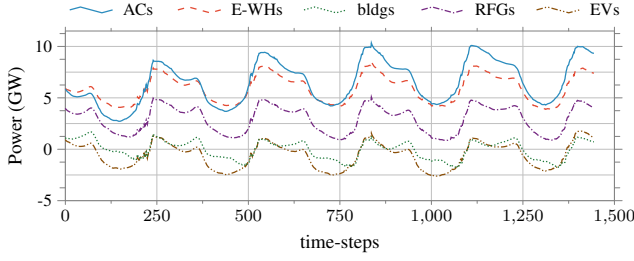


Fig. 2. This figure shows the power trajectories of each DERA: ACs (depicted by the solid blue line), E-WHs (depicted by the dashed red line), bldgs (depicted by the dotted green line), RFGs (depicted by the dash-dotted purple line), and EVs (depicted by the dash-double-dotted brown line) over a five day period, where each time-step corresponds to 5-minute interval. Each trajectory corresponds to the states p_1, p_2, p_3, p_4 , and p_5 of the system (2), respectively, when driven by the policy learned via Algorithm 1 over $L=407$ episodes.

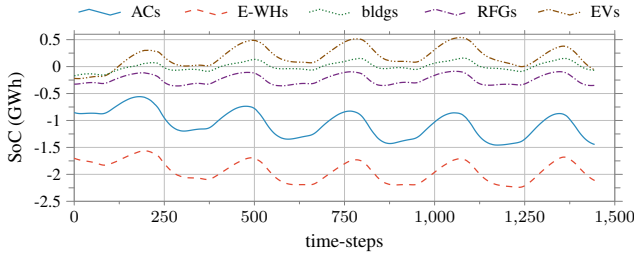


Fig. 3. This figure shows the SoC trajectories of each DERA: ACs (depicted by the solid blue line), E-WHs (depicted by the dashed red line), bldgs (depicted by the dotted green line), RFGs (depicted by the dash-dotted purple line), and EVs (depicted by the dash-double-dotted brown line) over a five day period, where each time-step corresponds to 5-minute interval. Each trajectory corresponds to the states z_1, z_2, z_3, z_4 , and z_5 of the system (2), respectively, when driven by the policy learned via Algorithm 1 over $L=407$ episodes.

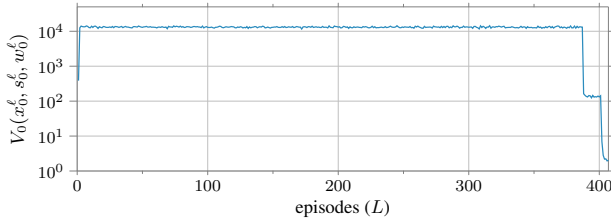


Fig. 4. This figure shows the cumulative cost incurred by the policy learned from Algorithm 1 starting from the initial state x_0^L, s_0^L , and w_0^L for episodes $1 \leq \ell \leq 407$. For episodes $1 \leq \ell \leq 387$, we add a random exploration term sampled from $\mathcal{N}(0, 0.25I_5)$ to the policy (given by line 13 of Algorithm 1), and for episodes $387 < \ell \leq 401$, we add a random exploration term sampled from $\mathcal{N}(0, 0.0025I_5)$. We observe that the cumulative cost decreases as the number of episodes increases after the exploration term is removed or reduced.

tion approximator (e.g., neural network), reducing risk of overfitting. Furthermore, explicit exploration (Remark 1) and online episodic policy update ensure sufficient excitation and robustness to temporal correlations.

V. CONCLUSION

In this paper, we develop a forecast-free RL framework for enabling heterogeneous distributed energy resource aggregators, DERAs, to track net-load demand without relying on short-term predictions. By coupling deterministic linear DERA dynamics with a feature-based linear Markov model for the net load, we derived closed-form expression for the optimal policy.

Par.	Unit	DER1	DER2	DER3	DER4	DER5
Type	—	ACs	E-WHs	bldgs	RFGs	EVs
N	million	10	10	1	10	1
α_i	—	0.98	0.99	0.97	0.96	0.99
$\kappa_{z_i}^i$	—	1e-3	2e-3	5e-3	5e-3	2e-3
$\kappa_{p_i}^i$	—	1e-1	2e-1	5e-1	5e-1	2e-1
β_i	hr	1/300	1/300	1/300	1/300	1/300

TABLE I
PARAMETERS FOR EACH CLASS OF DER, CORRESPONDING TO THE DISCRETE-TIME LINEAR DYNAMICAL MODEL IN (2).

This expression enabled a least-squares value iteration scheme that learns directly from operational data while preserving the interpretability and structure of classical LQR-type controllers. Numerical experiments using real CAISO net-demand data demonstrated that the learned controller achieves high tracking accuracy, smooth actuation, and stable regulation across DER classes. Our results show that reliable real-time DERA allocation can be achieved without explicit load forecasting, highlighting the value of hybrid model-based and data-driven control frameworks for modern grid-balancing applications.

REFERENCES

- [1] E. Benenati, M. Colombino, and E. Dall’Anese. A tractable formulation for multi-period linearized optimal power flow in presence of thermostatically controlled loads. In *IEEE Conf. on Decision and Control*, pages 4189–4194, Nice, France, 2019.
- [2] N. Cammardella, J. Mathias, M. Kiener, A. Bušić, and S. Meyn. Balancing california’s grid without batteries. In *IEEE Conf. on Decision and Control*, pages 7314–7321, Miami, FL, USA, 2018.
- [3] J. Mathias, S. Meyn, H. Ballouz, and M. Ansari. A distributed control architecture for optimal allocation of grid-responsive load aggregations. In *IEEE Power & Energy Society Innovative Smart Grid Technologies Conference*, pages 1–5, New Orleans, LA, USA, 2022.
- [4] J. Mathias, R. Angluri, O. Kosut, and L. Sankar. Model predictive control for joint ramping and regulation-type service from distributed energy resource aggregations. In *IEEE Power & Energy Society General Meeting*, pages 1–5, Seattle, WA, USA, 2024.
- [5] J. Köhler, M. A. Müller, N. Li, and F. Allgöwer. Real time economic dispatch for power networks: A distributed economic model predictive control approach. In *IEEE Conf. on Decision and Control*, pages 6340–6345, Melbourne, VIC, Australia, 2017.
- [6] M. Liu, Y. Shi, and X. Liu. Distributed mpc of aggregated heterogeneous thermostatically controlled loads in smart grid. *IEEE Transactions on Industrial Electronics*, 63(2):1120–1129, 2016.
- [7] M. Ma, C. Zhang, X. Liu, and H. Chen. Distributed model predictive load frequency control of the multi-area power system after deregulation. *IEEE Transactions on Industrial Electronics*, 64(6):5129–5139, 2017.
- [8] A. Parisio, C. Wieczorek, T. Kyntäjä, J. Elo, K. Strunz, and K. H. Johansson. Cooperative mpc-based energy management for networked microgrids. *IEEE Transactions on Smart Grid*, 8(6):3066–3074, 2017.
- [9] L. A. Duffaut Espinosa and M. Almassalkhi. A packetized energy management macromodel with quality of service guarantees for demand-side resources. *IEEE Transactions on Power Systems*, 35(5):3660–3670, 2020.
- [10] D. Cao, J. Zhao, W. Hu, F. Ding, Q. Huang, Z. Chen, and F. Blaabjerg. Data-driven multi-agent deep reinforcement learning for distribution system decentralized voltage control with high penetration of pvs. *IEEE Transactions on Smart Grid*, 12(5):4137–4150, 2021.
- [11] S. Bahrami, Y. C. Chen, and V. W. S. Wong. Deep reinforcement learning for demand response in distribution networks. *IEEE Transactions on Smart Grid*, 12(2):1496–1506, 2021.
- [12] H. Liu, J. Qiu, and J. Zhao. A data-driven scheduling model of virtual power plant using wasserstein distributionally robust optimization. *International Journal of Electrical Power & Energy Systems*, 137:107801, 2022.
- [13] A. A. Al Makdah, O. Kosut, L. Sankar, and S. Zou. Linear dynamics meets linear mdps: Closed-form optimal policies via reinforcement learning. *arXiv preprint arXiv: 2508.17185*, 2026.
- [14] California ISO – Folsom, CA 95763-9014. ISO Today. Online www.aiso.com/Pages/TodaysOutlook.aspx.