Les Cahiers du GERAD ISSN: 0711–2440

Batch reinforcement learning for network-safe demand response in unknown electric grids

A. Lesage-Landry, D.S. Callaway

G–2021–55

October 2021
Revised: March 2022

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2021
– Bibliothèque et Archives Canada, 2021

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2021
– Library and Archives Canada, 2021

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Batch reinforcement learning for network-safe demand response in unknown electric grids

Antoine Lesage-Landry [a, b]

Duncan S. Callaway [c]

[a] *Department of Electrical Engineering, Polytechnique Montréal, Montréal (Qc), Canada, H3T 1J4*

[b] *GERAD, Montréal (Qc), Canada, H3T 1J4*

[c] *Energy & Resources Group, University of California, Berkeley, Berkeley, USA 94720*

antoine.lesage-landry@polymtl.ca
dcal@berkeley.edu

October 2021
Revised: March 2022
Les Cahiers du GERAD
G–2021–55

Abstract : We formulate a batch reinforcement learning-based demand response approach to prevent distribution network constraint violations in unknown grids. We use the fitted Q-iteration to compute a network-safe policy from historical measurements for thermostatically controlled load aggregations providing frequency regulation. We test our approach in a numerical case study based on real load profiles from Austin, TX. We compare our approach's performance to a greedy, grid-aware approach and a standard, grid-agnostic approach. The average tracking root mean square error is 0.0932 for our approach, and 0.0600 and 0.0614 for, respectively, the grid-aware and grid-agnostic implementations. Our numerical case study shows that our approach leads to a 95% reduction, on average, in the total number of rounds with at least a constraint violation when compared to the grid-agnostic approach. Working under limited information, our approach thus offers lower but acceptable setpoint tracking performance while ensuring safer distribution network operations.

Keywords: Batch reinforcement learning, demand response, frequency regulation, network-safe, thermostatically controlled loads

Acknowledgements: This work was funded in part by the Institute for Data Valorization (IVADO), in part by the Natural Sciences and Engineering Research Council of Canada, in part by the National Science Foundation, award 1351900, and in part by the Advanced Research Projects Agency-Energy, award DE–AR0001061.

1 Introduction

Demand response [16, 23, 24] is an effective way of increasing the flexibility of electric power systems. The flexibility can be used, for example, to increase the grid efficiency, e.g., peak-shaving via load-shifting [5], or to mitigate the intermittency of renewable generators, e.g., by providing frequency regulation services [4, 28]. The growing presence of renewable generation in power systems means that the latter application will become prevalent in modern grids as to ensure its stability. Combined with the fact that demand response requires limited infrastructure investment [28], the need for an increased grid flexibility motivates the development of large-scale demand response programs.

The increasing demand response capacity means that significant amounts of power adjustment can be made throughout distribution networks. This creates risks of distribution network constraint violations due to, e.g., a large number of air conditioner units being turned on simultaneously [20]. Network constraint violations represent, e.g., a voltage magnitude above (below) the maximum (minimum) operational limits at the different buses or line current values above its ampacity [21]. Ensuring grid constraint satisfaction is essential to guaranteeing the reliable and safe operation of distribution grids. Consequently, in our work, we prioritize grid safety at the expense of the demand response program performance.

Related work. We now review the relevant literature on network-safe demand response. Reference [18] studies the effect of demand response of TCL aggregations on distribution network constraints using numerical simulations. The authors of [20], building on their previous work [17, 18], propose a network-aware frequency regulation framework in which an aggregator first dispatches TCLs and then the system operator modifies the aggregator's controls to prevent network constraint violations. The operator uses either a blocking or a mode-count control method. In this setting, the aggregator does not require knowledge of the network, e.g., its parameters or topology. Such knowledge is, however, needed by the operator which intervenes in the demand response process and is in charge of meeting its network's constraints. In [2, 22, 30], demand response controls are computed via an optimal power flow formulation to model network constraints. Reference [10] and prior work [6] model TCL aggregations as an aggregated Markov process where each state corresponds to a power consumption level. The authors of [10] then use this model in a chance-constrained, relaxed power flow problem to optimize both the distribution network's and the aggregation's objectives while accounting for network constraints. In [3, 7], online primal-dual algorithms are proposed to control power injections of distributed energy resources, including TCLs, for setpoint tracking at a point of connection under linearized network constraints. In [3], aggregations are also considered. The online process of [3, 7] allows the decision algorithm to use measurements to rapidly correct controls that lead to constraint violations. In [15], a method based on convex relaxations of the optimal power flow is proposed to certify that a set of power injections at different nodes, e.g., changes in power consumption of flexible loads, will not lead to any network constraint violations. In [19], the authors establish sufficient conditions on the maximum change in power consumption to be computed by the operator of a network such that voltage constraints of the network are satisfied at all times. These conditions are then used to constrain the aggregator's controls. Reference [1] uses an actor-critic-based deep reinforcement learning approach for load-shifting. The approach ensures that power flow constraints are satisfied throughout the network by projecting the control onto a feasible convex set.

In this work, we consider network-safe demand response of thermostatically controlled load (TCL) aggregations in unknown networks [20]. We formulate a demand response approach for frequency regulation that accounts for distribution network constraints when the network's electrical parameters and/or its topology are unknown to the aggregator. We further assume that the aggregator acts independently of the system operator, i.e., no information exchange is possible between the two entities. The problem we considered is of particular importance as (i) private aggregators may not have access to the network information because of privacy reasons [20] and (ii) system operators may wish to fully outsource demand response. Because we do not assume knowledge of the network, its electrical

variables, e.g., the voltage magnitudes at the different nodes, the power flowing in or out of the network, and the aggregations' power adjustment, cannot be computed. We base our approach on batch reinforcement learning [8] to circumvent this problem. This allows us to learn the network and aggregation models instead of requiring detailed modelling and characterization before implementation.

In reinforcement learning [11, 26], a decision maker must choose, based on previous decision information, the best control to implement given the current state of a system. The state then changes and a reward is granted accordingly. The system is modelled as a Markov decision process where the dynamics are unknown. Batch reinforcement learning differs from traditional, online reinforcement learning because the control policy is computed from information taking the form of batch data available prior to implementation rather than being collected sequentially. The data represents a collection of historical transitions, i.e., initial state, control, reward and final state tuples. Online reinforcement learning requires the exploration of numerous state-control pairs. This conflicts with our objective of ensuring network-safe demand response because exploration will lead to several constraints violations because the unsafe state and control pairs to be identified. For this reason, we opt for batch reinforcement learning.

We train our algorithm with historical measurements of the power flow at the point of connection with the grid and of the voltage magnitudes at the buses of interest. We use the fitted Q-iteration (`FQI`) [8] to compute a network-safe control policy for several TCL aggregations located in a distribution grid. We model the aggregations as a stochastic battery using [9]'s model. Our approach provides power setpoint tracking at the point of connection with the grid while ensuring that the voltage magnitude is within the prescribed limits at all nodes equipped with metering infrastructure. In this work, we focus on voltage magnitude constraints but our setting can be extended to other types of network constraints, e.g., line ampacity.

To the authors' best knowledge, our work is the first to consider network-safe demand response when the network's topology and/or electric parameters are unknown to the aggregator and no interventions from the system operator is possible. Our main contribution is to formulate data-driven approach for network-safe demand response in unknown environment. We then numerically evaluate its performance using real data.

2 Fitted Q-iteration

We now introduce the fitted Q-iteration (`FQI`) algorithm. Consider the Markov decision process defined by the state space $\mathcal{X} \subseteq \mathbb{R}^{n_1}$, $n_1 \in \mathbb{N}$, the control space $\mathcal{U} \subseteq \mathbb{R}^{n_2}$, $n_2 \in \mathbb{N}$, the stochastic system's dynamics $f : \mathcal{X} \times \mathcal{U} \times \mathcal{W} \mapsto \mathcal{X}$ where $\mathcal{W} \in \mathbb{R}^{n_3}$, $n_3 \in \mathbb{N}$, is the set of random disturbances, and a reward function $r : \mathcal{X} \times \mathcal{U} \times \mathcal{X} \times \mathcal{W} \mapsto \mathbb{R}$. The disturbance set $\mathcal{W}$ models the environment/system's uncertainty, e.g., due measurement errors, erratic human behaviour, sudden weather changes, random failures in communication or TCLs, etc. The objective is to determine a policy $\pi : \mathcal{X} \mapsto \mathcal{U}$ to maximize the total expected reward given by:

$$\mathbb{E}\left[\sum_{t=1}^{+\infty} \beta^t r(\mathbf{x}_t, \pi(\mathbf{x}_t), f(\mathbf{x}_t, \pi(\mathbf{x}_t), \mathbf{w}_t), \mathbf{w}_t)\right], \tag{1}$$

where $\beta \in [0, 1)$ is the discount factor. The expectation in (1) is computed with respect to the joint conditional probability of $\mathbf{w}_t \in \mathcal{W}$ given the state $\mathbf{x}_t$ and the control $\mathbf{u}_t$ for all t. In a reinforcement learning problem, the system's model, f, is unknown.

In batch reinforcement learning, the decision maker has access to batch data prior to implementation. In this work, we use `FQI` to derive an approximate optimal policy for (1) based on the batch data. This process avoids the exploration process of typical reinforcement learning methods which can lead to unsafe distribution network operations. Let $\mathcal{D}$ denote the batch data defined as: $\mathcal{D} = \left\{\left(\mathbf{x}_t^l, \mathbf{u}^l, \mathbf{x}_{t+1}^l, r^l\right) \in \mathcal{X} \times \mathcal{U} \times \mathcal{X} \times \mathbb{R} \middle| \mathbf{x}_{t+1}^l = f(\mathbf{x}_t^l, \mathbf{u}^l, \mathbf{w}), \mathbf{w} \in \mathcal{W}, l = 1, 2, \ldots, L\right\}$, where $L \in \mathbb{N}$ is

the number of data points in the dataset. `FQI` is a dynamic programming-based batch reinforcement learning approach which iteratively computes an approximate Q-function that satisfies Bellman's equation [8]. `FQI` is presented in Algorithm 1. We provide next a summary of the algorithm's derivation. Interested readers are referred to [8] for a detailed coverage of `FQI`.

Let $Q : \mathcal{X} \times \mathcal{U} \mapsto \mathbb{R}$ be the Q-function or state-control value function. Let $Q_0(\mathbf{x}, \mathbf{u}) = 0$ for all $(\mathbf{x}, \mathbf{u}) \in \mathcal{X} \times \mathcal{U}$. For $N \geq 0$, let Q_N-function be defined recursively as:

$$Q_{N+1}(\mathbf{x}, \mathbf{u}) = \mathbb{E}\left[r(\mathbf{x}, \mathbf{u}, \mathbf{w}) + \beta \max_{\mathbf{u}' \in \mathcal{U}} Q_N(f(\mathbf{x}, \mathbf{u}, \mathbf{w}), \mathbf{u}')\right],$$

where $\mathbf{w} \in \mathcal{W}$. The above expectation is again taken with respect to the joint probability of $\mathbf{w}$ given the state $\mathbf{x}_t$ and the control $\mathbf{u}_t$. Then, by the contraction mapping theorem, $Q_N \to Q$ where Q satisfies the Bellman equation:

$$Q(\mathbf{x}, \mathbf{u}) = \mathbb{E}\left[r(\mathbf{x}, \mathbf{u}, \mathbf{w}) + \beta \max_{\mathbf{u}' \in \mathcal{U}} Q(f(\mathbf{x}, \mathbf{u}, \mathbf{w}), \mathbf{u}')\right]. \tag{2}$$

`FQI` aims to approximate the Q-function that meets (2) by sequentially building an approximation of the Q_N-function. This is done by first assuming $\hat{Q}_0(\mathbf{x}, \mathbf{u}) = 0$ for all $\mathbf{x}$ and $\mathbf{u}$. Then, at iteration N, (i^l, o^l)-pairs are computed using the previous approximation, $\hat{Q}_{N-1}$, as follows:

$$\begin{aligned} i^l &= \left(\mathbf{x}_t^l, \mathbf{u}^l\right) \\ o^l &= r^l + \beta \max_{\mathbf{u}' \in \mathcal{U}} \hat{Q}_{N-1}(\mathbf{x}_{t+1}^l, \mathbf{u}'), \end{aligned}$$

for $l = 1, 2, \ldots, L$. The $\hat{Q}_N$-function is finally obtained by fitting a function using a regression tree method over the pairs $\left\{(i^l, o^l), l = 1, 2, \ldots, L\right\}$. A regression tree method is used to approximate the information gathered in the batch data to other state-control pairs and the conditional expectation of the Bellman equation [8]. The latter comes from the fact that the regression tree averages the output o^l of all state-control pairs contained in a leaf node. This process is repeated until the maximum difference between consecutive $\hat{Q}_N$ is below a set tolerance ϵ. The convergence of the sequence $\hat{Q}_N$ is guaranteed as shown by [8] when specific regression tree methods are used, e.g., `KD-Tree` or `Totally Randomized Trees`.

Finally, the $\hat{Q}_N$-functions are used to determine the policy. The greedy policy $\pi : \mathcal{X} \mapsto \mathcal{U}$ for `FQI` is defined as: $\pi(\mathbf{x}) \in \arg\max_{\mathbf{u} \in \mathcal{U}} \hat{Q}_N(\mathbf{x}, \mathbf{u})$ for all $\mathbf{x} \in \mathcal{X}$.

Algorithm 1: Fitted Q Iteration (`FQI`)

Parameters: L, $\mathcal{D}$, $\beta \in [0, 1)$, $\epsilon > 0$
Initialization: $N = 0$, $\hat{Q}_0(\mathbf{x}, \mathbf{u}) = 0$ for all $\mathbf{x}, \mathbf{u}$.

while $\left\|\hat{Q}_N(\mathbf{x}, \mathbf{u}) - \hat{Q}_{N-1}(\mathbf{x}, \mathbf{u})\right\|_\infty \geq \epsilon$ **do**
  $N = N + 1$
  for $l = 1, 2, \ldots, L$ **do**
    Compute the pairs:

$$\begin{aligned} i^l &= \left(\mathbf{x}_t^l, \mathbf{u}^l\right) \\ o^l &= r^l + \beta \max_{\mathbf{u}' \in \mathcal{U}} \hat{Q}_{N-1}^j\left(\mathbf{x}_{t+1}^l, \mathbf{u}'\right). \end{aligned}$$

  end for
  Compute $\hat{Q}$-function for iteration N:

$$\hat{Q}_N(\mathbf{x}, \mathbf{u}) = \text{RegressionTree}\left(\left\{\left\{(i^l, o^l)\right\}_{l=1}^L\right\}; (\mathbf{x}, \mathbf{u})\right)$$

end while

3 Safe demand response in unknown networks

In this section, we use `FQI` for network-safe demand response of TCL aggregations under limited information. We consider a network consisting of multiple loads and TCL aggregations. We do not assume knowledge of the network's parameters nor of its topology. Moreover, we do not assume any collaboration between the system operator and the aggregator, i.e., the system operator does not intervene in the demand response process to block or limit the aggregator [19, 20]. We consider a demand response setting where the power consumption of TCL aggregations is modified such that the total power demand for the network at the point of connection tracks a setpoint, e.g., a frequency regulation signal [5].

We discretize the time horizon T into m-minute rounds. We denote the time indices by $t \in \{1, 2, \dots, T\}$. We consider a distribution network connected to the grid at node 0. Let $\mathcal{N}$ be the set of nodes of the network. Let $\mathcal{L} \subseteq \mathcal{N}$ be the set of loads and $\mathcal{K} \subseteq \mathcal{N}$ be the set of TCL aggregations in the network. We denote by $p_{0,t} \in \mathbb{R}$ the power flow in or out of the network at the point of connection and at time t. We let $d_{i,t} \in \mathbb{R}$ be the power consumption of load $i \in \mathcal{L}$. Our objective is to track the power setpoint s_t by adjusting the network demand $p_{0,t}$. This is done in turn by controlling the power consumption of TCL aggregations. Let $\mathbf{u}_t \in \mathcal{U}$ be the control signal sent the aggregations at time t. We now introduce the aggregation model and then state our `FQI`-based approach.

3.1 TCL aggregation model

We consider multiple aggregations each consisting of K TCLs. We model the aggregations as a stochastic battery [9]. Let $x^{\text{TCL}}(t)$ be the state of charge of the battery which models the TCL aggregation at time t and $\mu(t)$ be the power coming in or out the battery, i.e., the change in power consumption of the aggregation from its nominal level. From [9], we have for aggregation j, for all $t \geq 0$:

$$\dot{x}_j^{\text{TCL}}(t) = -\alpha_i x_j^{\text{TCL}}(t) - \mu_j(t) \tag{3}$$

$$-n_j^-(t) \leq \mu_j(t) \leq n_j^+(t), \tag{4}$$

with the boundary conditions: $x_j^{\text{TCL}}(0) = 0$ and $\left|x_j^{\text{TCL}}(t)\right| \leq C_j$ for all t. The variables $n_j^-(t)$ and $n_j^+(t)$ model the maximum charging and discharging rate of the equivalent battery, C_j its energy capacity, and α_j its dissipation rate. The authors of [9] define necessary and sufficient conditions for the battery model (C_j, n_j^-, n_j^+, α_j from (3)−(4)) to represent the flexibility of a TCL aggregation while ensuring that the temperature constraints of each TCL are met. The reader is referred to [9, Section IV] for the detailed derivation of these conditions. We use the sufficient battery model which maximizes its energy capacity [9]. Our approach does not assume knowledge of the equivalent battery model for the aggregations. For completeness, we now introduce the TCL thermal model. We then define the battery equivalent parameters for the TCL aggregations. Both will be used in Section 4 to simulate the aggregations' thermal and power variables.

Let $c^{j,k}$, $r^{j,k}$ be, respectively, the thermal capacitance and the thermal resistance of TCL $k \in \{1, 2, \dots, K\}$ from aggregation $j \in \mathcal{K}$. Let $P_{\text{r}}^{j,k}$ and $\eta^{j,k}$ be the rated electrical power and the coefficient of performance of TCL k's cooling unit, respectively. Let $\theta_{\text{d}}^{j,k}$ be the desired temperature of TCL k and $\Delta^{j,k}$ be its temperature deadband. Considering a continuous thermal model [9] for the temperature dynamic of TCLs, we let $P_{0,t}^{j,k}$ be the nominal power consumption of TCL k needed to keep it at its desired temperature: $P_{0,t}^{j,k} = \frac{\theta_t^{\text{ambient}} - \theta_{\text{d}}^{j,k}}{\eta^{j,k} r^{j,k}}$. Any deviation of load k from $P_{0,t}^{j,k}$ at time t then represents its consumption flexibility. The flexibility of each load is bounded by the maximum power consumption of the unit and zero, and must not push the TCL's temperature out of its deadband. We remark that even if the model uses a continuous model, the controls sent to TCLs are all binary [9]. Finally. at

time t, the battery model parameters of aggregation j are:

$$C_j = \sum_{k=1}^{K} f^{j,k} \qquad \alpha_j = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{v^{j,k}}$$
$$n_{j,t}^{-} = \left(\sum_{k\in=1}^{K} f^{j,k}\right) \min_{k\in\{1,2,\dots,K\}} \frac{P_{0,t}^{j,k}}{f^{j,k}}$$
$$n_{j,t}^{+} = \left(\sum_{k\in=1}^{K} f^{j,k}\right) \min_{k\in\{1,2,\dots,K\}} \frac{P_{\mathrm{r}}^{j,k} - P_{0,t}^{j,k}}{f^{j,k}},$$

where $v^{j,k} = r^{j,k}c^{j,k}$ and $f^{j,k} = \frac{\Delta^{j,k}c^{j,k}}{\eta^{j,k}(1+|\alpha_i v^{j,k}-1|)}$. An aggregation dispatches ON and OFF signals to their individual TCLs to track the power adjustment $a_{j,t} \in \mathbb{R}$ [9]. Let $q_t^{j,k} = 1$ if the cooling unit of aggregation j's TCL k at time t is ON or 0 otherwise. Let $P_{j,t}^{\mathrm{agg}} = \sum_{k=1}^{K} q_t^{j,k} P_{r,t}^{j,k}$ be the power consumption of the aggregation at time t. Let $P_{j,t}^{\mathrm{baseline}} = \sum_{k=1}^{K} P_{0,t}^{j,k}$ be the baseline power consumption at time t. At each time step, the aggregation either turns ON or OFF TCLs such that the power deviation $\delta_{j,t} = P_{j,t}^{\mathrm{agg}} - P_{j,t}^{\mathrm{baseline}}$ matches the adjustment $a_{j,t}$. If $a_{j,t} < \delta_{i,t}$, then the aggregation is interpreted as a battery discharging and injecting power to the grid. This corresponds to turning OFF cooling units that were ON at $t-1$, thus reducing the power consumption of the aggregation with respect to its baseline. Conversely, if $a_{j,t} > \delta_{j,t}$, the aggregation is perceived as a battery being charged by the grid and TCLs are turned ON. Lastly, for $\delta_{j,t} = 0$, the battery is idle and the aggregation must keep its power consumption constant. The cooling units to be turned ON or OFF are determined using the priority-stacks described in [9, Section V]. The power adjustment $a_{j,t}$ is set for all aggregations by the control $u_{j,t}$ which is in turn computed by the network-safe demand response model presented next.

3.2 `FQI`-based safe demand response

At time step t, the control $u_{j,t} \in \mathcal{U}$ is sent to the TCL aggregation j. The aggregation then converts the control to a power adjustment $a_{j,t}$ depending on their maximum charging and discharging rate at t, $n_{j,t}^{+}$ and $n_{j,t}^{-}$, respectively. Finally, the power adjustment is used locally at the aggregation to turn ON and OFF TCLs so that the network demand $p_{0,t}$ matches the power setpoint s_t. To ensure safe distribution grid operations, the control dispatched by the decision-maker must not lead to any network constraint violations.

We define our loss function $\ell_t : \mathcal{X} \times \mathcal{U} \times \mathcal{W} \mapsto \mathbb{R}$ as the sum between the square setpoint tracking error and the penalty function $\rho(|v_{i,t}|) = [|v_{i,t}| - \overline{v}]^{+} + [\underline{v} - |v_{i,t}|]^{+}$ where $|v_{i,t}|$ is the voltage magnitude at node i and $\overline{v}$ and $\underline{v}$ are, respectively, the upper and lower voltage magnitude limits. We use the $[\cdot]^{+} = \max\{0, \cdot\}$-operator as the penalty because it does not impact the loss function if the voltage constraints are met but highly increases its value otherwise. The loss function at time t is: $\ell_t(\mathbf{x}_t, \mathbf{u}_t, \mathbf{w}_t) = (s_t - p_{0,t})^2 + \lambda \sum_{i\in\mathcal{N}^m} \rho(|v_{i,t}|)$, where λ controls the magnitude of the constraint violation penalty and the set $\mathcal{N}^m \subseteq \mathcal{N}$ represents the nodes equipped with voltage metering capacity. The variables $\mathbf{x}_t, \mathbf{u}_t$ and $\mathbf{w}_t$ implicitly define $p_{0,t}$ and $|v_{i,t}|$ as it will be presented shortly.

Let $f_i^{\mathrm{network}} : \mathcal{U} \times \mathcal{X} \times \mathcal{W} \mapsto \mathbb{R}$ be a function that maps the state of the network, the controls sent to the TCL aggregations and some disturbances to the voltage magnitude at a node $i \in \mathcal{N}$. Similarly, let $g^{\mathrm{network}} : \mathcal{U} \times \mathcal{X} \times \mathcal{W} \mapsto \mathbb{R}$ be a function that returns the power injection at the point of connection of the network with the grid given the state of the network, the aggregation controls, and a disturbance vector. These mappings represent solving the power flow equations for the network at a given state to determine the voltage magnitude at node i and the power injection at the point of connection. Let $b : \mathcal{U} \times \mathcal{X} \times \mathcal{W} \mapsto \mathbb{R}^{\mathrm{card}\,\mathcal{K}}$ be a function that returns the TCL aggregation power consumptions as modeled by the battery equivalent representation where b_j refers to the aggregation $j \in \mathcal{K}$. Lastly, we let $h : \mathcal{U} \times \mathcal{X} \times \mathcal{W} \mapsto \mathcal{X}$ be the system dynamics or transition function of the problem. For all states

except for the aggregation's state of charge, the transition function is exogenous to the problem, e.g., ambient temperature evolves independently of the system. For the state of charge $x_{j,t}^{\mathrm{TCL}}$, h is implicitly a function of b_j.

Because we do not assume any prior knowledge of the network parameters nor about its topology, all these mappings are unknown. We assume that the infrastructure present in the network allows for voltage magnitude metering, $|v_{i,t}|$, at nodes $i \in \mathcal{N}^m$, and for power readings at the point of connection, $p_{0,t}$. Using these measurements and the corresponding controls $\mathbf{u}_t$, the loss function $\ell\left(\mathbf{x}_t, \mathbf{u}_t, \mathbf{w}_t\right)$ can be evaluated.

In our implementation, the state variable $\mathbf{x}_t \in \mathcal{X}$, i.e., the information available when the controls $u_{j,t}$, $j \in \mathcal{K}$, are computed, are: (i) the ambient temperature $\theta_t^{\mathrm{ambient}} \in \mathbb{R}$, (ii), the setpoint $s_t \in \mathbb{R}$, (iii) the total power demand of the network $\sum_{i \in \mathcal{L}} d_{i,t}$, and (iv) the state of charge of each TCL aggregation $x_{j,t}^{\mathrm{TCL}}$, $j \in \mathcal{K}$. Other states could be considered to improve performance, e.g., the time or the calendar day. Finally, the disturbance $\mathbf{w}_t$ represents the uncertainty due to, e.g., measurement errors, erratic human behavior, sudden weather changes, random failures in communication or TCLs, etc [27].

Finally, we formulate the network-safe demand response problem as:

$$
\begin{aligned}
\min_{\{\mathbf{u}_t\}_{t=1}^{+\infty}} \quad & \mathbb{E}\left[\sum_{t=1}^{+\infty} \beta^t \left(\left(s_t - p_{0,t}\right)^2 + \lambda \sum_{i \in \mathcal{N}^m} \rho(|v_{i,t}|) \right)\right] \\
\text{subject to} \quad & u_{j,t} \in \mathcal{U}, \text{ for all } j \in \mathcal{K} \\
& \mathbf{u}_t = \left(u_{1,t}, u_{2,t}, \ldots, u_{\operatorname{card}\mathcal{K},t}\right) \\
& \mathbf{x}_{t+1} = h\left(\mathbf{x}_t, \mathbf{u}_t, \mathbf{w}_t\right) \\
& \underline{\mathbf{x}}_t \leq \mathbf{x}_t \leq \overline{\mathbf{x}}_t \\
& p_{i,t} = d_{i,t} \text{ for all } i \in \mathcal{L} \setminus \mathcal{K} \\
& p_{j,t} = d_{j,t} + b_j(u_{j,t}, \mathbf{x}_t, \mathbf{w}_t) \text{ for all } j \in \mathcal{K} \\
& |v_{i,t}| = f_i^{\mathrm{network}}\left(\mathbf{u}_t, \mathbf{x}_t, \mathbf{w}_t\right) \text{ for all } i \in \mathcal{N}^m \\
& p_{0,t} = g^{\mathrm{network}}\left(\mathbf{u}_t, \mathbf{x}_t, \mathbf{w}_t\right),
\end{aligned}
$$

where $\underline{\mathbf{x}}_t$ and $\overline{\mathbf{x}}_t$ are constraints on the state variables, e.g., maximum and minimum state of charge of the battery equivalent model, $n_{j,t}^+$ and $n_{j,t}^-$ for all j. We use `FQI` to compute an approximation of the optimal control at each round.

We use historical voltage magnitude measurements at nodes $i \in \mathcal{N}^m$ and power readings $p_{0,t}$ to compute the loss function for different states and controls. The loss function value at a given time t is then collected with its associated states $\mathbf{x}_t$, control $\mathbf{u}_t$ and resulting state $\mathbf{x}_{t+1}$ to form the batch data set. The data can be gathered, for example, from the normal operation of the TCL aggregations or during their participation to a grid-agnostic demand response program. Given a large enough data set, the risk of network constraint violation is unavoidably non-zero as stated in the introduction.

4 Case study

In this section, we present a numerical case study in which two aggregations are controlled to provide safe demand response in a distribution network.

4.1 Setup

We consider the 18-node distribution network presented in Figure 1. The network is based on the residential part of the Cigré LV benchmark network given in [12, 25]. The distribution network is connected at node 0 to the rest of the grid. It includes 6 loads and 2 TCL aggregations. The loads are located at nodes 1, 11, 15, 16, 17, and 18. The aggregations are located at nodes 5 and 12. We set the

voltage magnitude limits to $\underline{v} = 0.95$ and $\overline{v} = 1.05$ per unit. We assume that all nodes can measure the voltage magnitude, i.e., $\mathcal{N}^m = \mathcal{N}$.

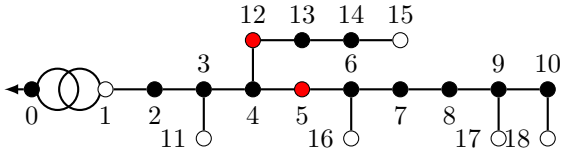


Figure 1: 18-node distribution network (white: load nodes, red: aggregation nodes, black: regular nodes)

We use the demand profiles for six residential loads located in Austin, TX, and provided by Pecan Street, Inc. We scale up the load consumption by a factor of eight to emphasize the need for safe demand response. The power flow computations use the full load profile whereas `FQI` only uses the total power consumption to reduce to the state space size. We further reduce its size and consider 60 discrete total consumption states. Historical ambient temperatures for Austin, TX which correspond to the same period as the demand profiles are used. The temperature data is also provided by Pecan Street, Inc. The temperature is discretized into 25 states. The demand regulation setpoints are randomly sampled according to: $s_t = 2\text{Uniform}(-\overline{D}, \overline{D}) + \sum_{i\in\mathcal{N}} d_{i,t}$, for $t = 1, 2, \ldots, T$, and where $\overline{D}$ is the average total demand for the considered period. The signal is similarly discretized in 60 values. We use `pandapower` [29] to generate the power component of the data set given the control $\mathbf{u}_t$, i.e., to compute the power flow at the point of connection and the voltage magnitudes at each node given the state of the network. For deployment, the batch data would strictly be from network measurements.

We set $K = 50$ TCLs per aggregation and sample their thermal parameters similarly to [9]. A TCL's cooling unit is initially ON with a probability of a half. The aggregation's state of charge is computed with respect to 25 discrete states. The initial state of charge of aggregation $j \in \mathcal{K}$ is set to 0 kWh for each day. Once the control is received, the aggregation determines locally which TCL to turn ON or OFF according to the process described in Section 3.1.

The performance of `FQI` is compared to two approaches: (i) a greedy, grid-agnostic (GGA) and (ii) a greedy, full information (GFI), i.e., grid-aware, approach. The former corresponds to a standard online setpoint tracking algorithm where the power consumption of flexible loads is adjusted to match the setpoint without incorporating network aspects like line losses or voltage constraints, e.g., [13, 14]. This approach uses controls defined as:

$$\mathbf{u}_t^{\text{GGA}} \in \underset{\substack{\mathbf{u}_t\in\mathcal{U} \\ \mathbf{u}_t=(u_{1,t},u_{2,t},\\ \ldots,u_{\text{card}\,\mathcal{K},t})}}{\arg\min} \left(s_t - \sum_{j\in\mathcal{K}} b_j(u_{j,t}, \mathbf{x}_t, \mathbf{w}_t) - \sum_{i\in\mathcal{N}} d_{i,t} \right)^2 .$$

The GFI approach solves the tracking problem given the aforementioned f_i^{network}, g^{network} and b_i functions, and the disturbance $\mathbf{w}_t$. Specifically, the GFI's control is:

$$\begin{aligned}
\mathbf{u}_t^{\text{GFI}} \in \underset{\mathbf{u}_t\in\mathcal{U}}{\arg\min} \quad & (s_t - p_{0,t})^2 \\
\text{subject to} \quad & \mathbf{u}_t = (u_{1,t}, u_{2,t}, \ldots, u_{\text{card}\,\mathcal{K},t}) \\
& \underline{\mathbf{x}}_t \le \mathbf{x}_t \le \overline{\mathbf{x}}_t \\
& p_{i,t} = d_{i,t} \;\forall\; i \in \mathcal{L} \setminus \mathcal{K} \\
& p_{j,t} = d_{j,t} + b_j(u_{j,t}, \mathbf{x}_t, \mathbf{w}_t) \;\forall\; j \in \mathcal{K} \\
& |v_{i,t}| = f_i^{\text{network}}(\mathbf{u}_t, \mathbf{x}_t, \mathbf{w}_t) \;\forall\; i \in \mathcal{N}^m \\
& p_{0,t} = g^{\text{network}}(\mathbf{u}_t, \mathbf{x}_t, \mathbf{w}_t) \\
& \underline{v} \le |v_{i,t}| \le \overline{v} \;\forall\; i \in \mathcal{N}^m.
\end{aligned}$$

The GFI approach does not necessarily provide the best control at time t. While GFI has access to full information, it is greedy, e.g., it does not consider the state of charge of the aggregation throughout time. Finally, as no prior work can be implemented under our assumptions, we limit our comparison to GGA and GFI only.

We consider batch data representing 108 days from 11 am to 9 pm with $m = 1$ minute-rounds ($L = 64{,}800$) spanning from June 1$^{\text{st}}$, 2018 to September 16$^{\text{th}}$, 2018. For each data point, a random control is sampled and the loss ℓ_t is computed accordingly by solving the power flow [29] to replicate grid measurements. We use $\mathcal{U} = \{-1, 0, 1\}^2$ and let the power adjustment of aggregation j be defined as:

$$a_{j,t} = \begin{cases} \min\left\{50, n^+_{j,t}\right\}, & \text{if } u_{j,t} = 1 \\ \max\left\{-50, -n^-_{j,t}\right\}, & \text{if } u_{j,t} = -1 \\ 0, & \text{otherwise.} \end{cases}$$

For the algorithm, we set $\epsilon = 0.5$ and $\beta = 0.1$. The penalty scaling factor is set to $\lambda = 3.5 \times 10^6$. We use `Totally Randomized Trees` with 10 trees in the ensembles and a minimum of 10 data points in a leaf node. The regression tree kernels are computed at the first iteration and kept constant to guarantee the `FQI`'s convergence [8]. We evaluate our approach performance over a three day-period between 11 am to 9 pm each day.

4.2 Results

We use our approach on 20 different batch datasets and run the simulation with new setpoints each time. The average setpoint tracking loss improvement relative to the benchmark case where no DR approaches are used and the tracking root mean square error (RMSE) are compared, respectively, in the first and second column of Table 1 for `FQI`, GFI and GGA. Given the network information, GFI necessarily meets the constraints at all time. Consequently, its setpoint tracking performance is slightly lower than GGA which ignores the network. Because `FQI` must learn the network constraint thresholds based on the available measurements and then satisfy them during the simulations, its tracking performance is reduced in comparison to both GFI and GGA.

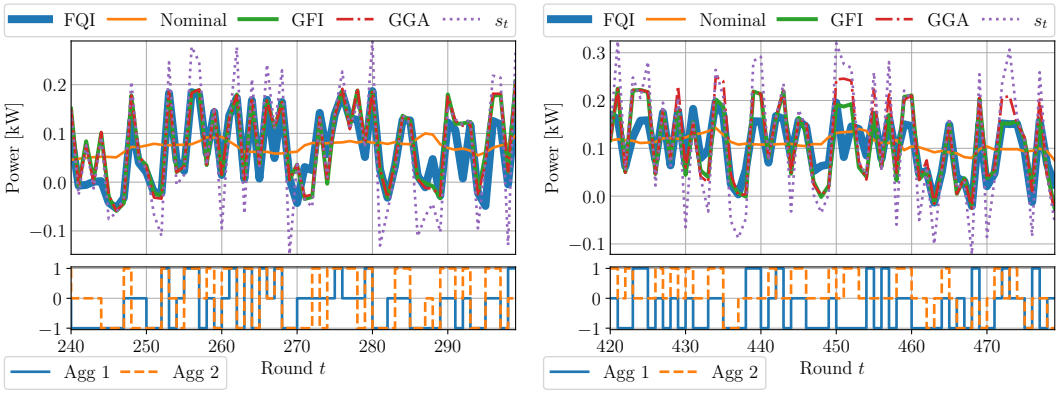
Table 1: Performance comparison (averaged over 20 simulations)

Approach	**Setpoint tracking** [%]	**RMSE** [kW]	**Nb. of rounds**
Nominal	—	0.1290	0.00
`FQI`	−47.55	0.0932	2.75
GFI	−77.29	0.0614	0.00
GGA	−78.10	0.0600	60.10

The average number of rounds with a least a constraint violation is presented in Table 1, third column, for the three approaches and the benchmark case. For `FQI`, the average number of rounds with a constraint violation is low at 2.75 in comparison to 60.1 rounds when the network is not accounted for by GGA. We observe that although the difference in setpoint tracking performance between `FQI` and GGA represents about a third of the latter's, on average, `FQI` leads to a 95% decrease in the number of rounds with at least a constraint violation. This is achieved without requiring the grid parameters, topology nor coordination with the system operator, and is in line with the priority we give to safe distribution grid operations.

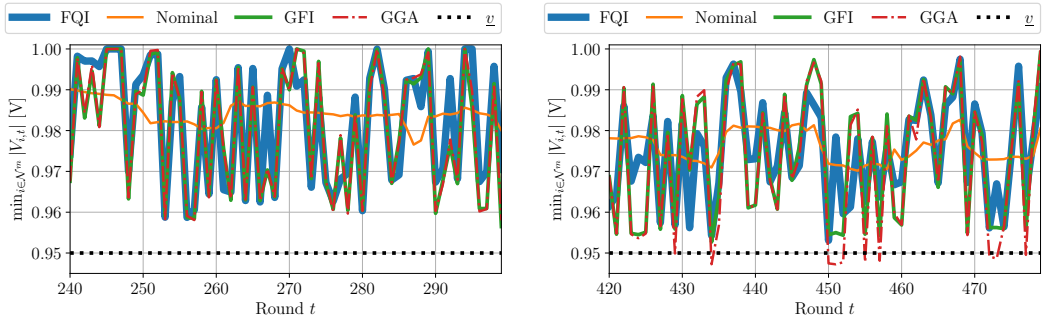
We conclude by presenting setpoint tracking curves for a specific data set simulation. Throughout the selected 30-hour simulation, `FQI` caused only a single constraint violation whereas 74 rounds with a least a constraint violation were registered when GGA was used. The minimum voltage magnitude in the distribution feeder when `FQI`, GFI and GGA are implemented is shown in Figure 3. Figures 2a and 3a shows that when the voltage magnitude is safely within the limit, `FQI`'s tracking is similar to GGA and GGI. The main difference in tracking occurs when the risks of constraint violation increases.

Figure 3b shows that between 12 pm and 1 pm, multiple violations occurs when GGA is implemented. All these violations are avoided using `FQI` which operates under similar assumptions. However, `FQI` tends to be conservative when setpoints are large. For example, this is observed on Figure 2b: (i) between rounds 450 and 453 where violations are avoided as shown by the overlap between `FQI` and GFI, (ii) and between rounds 458 and 460 where the power adjustment is decreased unnecessarily resulting in poor tracking as shown by the mismatch between `FQI` and GFI. We note that increasing the batch data size and number of bins should improve the `FQI` ability to identify violation-prone settings.Finally, the computation burden of the approach is manageable when the number of aggregation is limited like in the current case. For problems with a larger number of aggregations where the risk for network constraint violations is even higher, a scalable multi-agent version of `FQI` would be required. This is a topic for future work.



(a) 9 – 10 am (b) 12 – 1 pm

Figure 2: Setpoint tracking for September 19th, 2018 data (top: setpoint tracking performance, bottom: dispatched controls $u_{1,t}$ and $u_{2,t}$)



(a) 9 – 10 am (b) 12 – 1 pm

Figure 3: Minimum voltage magnitude across the feeder for September 19th, 2018 data

5 Conclusion

In this work, we design an approach for network-safe demand response in unknown environments. We do not assume knowledge of the network parameters nor of its topology. We further suppose no communication nor coordination between the aggregator and the system operator. We use `FQI` to control TCL aggregations with unknown dynamics and track a regulation signal at the point of connection with the grid. Our batch reinforcement learning approach leverages historical network

measurements to minimize the number of voltage magnitude constraint violations while tracking the power setpoint. Finally, we present a case study based on real data for a 18-node distribution network with two 50-TCL aggregations located in Austin, Texas, USA. The number of rounds where at least a constraint violation occurred drops to 2.75, on average, with our approach instead of 60.10 when a grid-agnostic approach is implemented, corresponding to a 95% reduction. The `FQI`-based approach leads to an RMSE of 0.0932 kW whereas a greedy, grid-agnostic, approach has an RMSE of 0.0600 kW. Working under limited information, the numerical simulations show that our approach avoids most under-voltage incidents while providing acceptable setpoint tracking. In future work, we will investigate multi-agent reinforcement learning to scale the approach to a large number of aggregations in the distribution network and ensure the its safe operation.

References

[1] Shahab Bahrami, Yu Christine Chen, and Vincent WS Wong. Deep reinforcement learning for demand response in distribution networks. IEEE Transactions on Smart Grid, 2020.

[2] Mostafa Bakhtvar, Carlos Andrade Cabrera, Giuseppina Buttitta, Olivier Neu, and Andrew Keane. A study of operation strategy of small scale heat storage devices in residential distribution feeders. In 2017 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), pages 1–6, 2017.

[3] Andrey Bernstein and Emiliano Dall'Anese. Real-time feedback-based optimization of distribution grids: A unified approach. IEEE Transactions on Control of Network Systems, 6(3):1197–1209, 2019.

[4] Duncan S Callaway. Tapping the energy storage potential in electric loads to deliver load following and regulation, with application to wind energy. Energy Conversion and Management, 50(5):1389–1400, 2009.

[5] Duncan S Callaway and Ian A Hiskens. Achieving controllability of electric loads. Proceedings of the IEEE, 99(1):184–199, 2010.

[6] Michael Chertkov, Deepjyoti Deka, and Yury Dvorkin. Optimal ensemble control of loads in distribution grids with network constraints. In 2018 Power Systems Computation Conference (PSCC), pages 1–7. IEEE, 2018.

[7] Emiliano Dall'Anese, Swaroop S Guggilam, Andrea Simonetto, Yu Christine Chen, and Sairaj V Dhople. Optimal regulation of virtual power plants. IEEE Transactions on Power Systems, 33(2):1868–1881, 2017.

[8] Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. Journal of Machine Learning Research, 6(Apr):503–556, 2005.

[9] He Hao, Borhan M Sanandaji, Kameshwar Poolla, and Tyrone L Vincent. Aggregate flexibility of thermostatically controlled loads. IEEE Transactions on Power Systems, 30(1):189–198, 2014.

[10] Ali Hassan, Robert Mieth, Michael Chertkov, Deepjyoti Deka, and Yury Dvorkin. Optimal load ensemble control in chance-constrained optimal power flow. IEEE Transactions on Smart Grid, 10(5):5186–5195, 2018.

[11] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. Journal of artificial intelligence research, 4:237–285, 1996.

[12] Stavros Karagiannopoulos, Petros Aristidou, and Gabriela Hug. Data-driven local control design for active distribution grids using off-line optimal power flow and machine learning techniques. IEEE Transactions on Smart Grid, 10(6):6461–6471, 2019.

[13] Antoine Lesage-Landry and Joshua A Taylor. Setpoint tracking with partially observed loads. IEEE Transactions on Power Systems, 33(5):5615–5627, 2018.

[14] Johanna L Mathieu, Stephan Koch, and Duncan S Callaway. State estimation and control of electric loads to manage real-time energy imbalance. IEEE Transactions on Power Systems, 28(1):430–440, 2012.

[15] Daniel Molzahn and Line A Roald. Grid-aware versus grid-agnostic distribution system control: A method for certifying engineering constraint satisfaction. In Proceedings of the 52nd Hawaii International Conference on System Sciences, 2019.

[16] Peter Palensky and Dietmar Dietrich. Demand side management: Demand response, intelligent energy systems, and smart loads. IEEE transactions on industrial informatics, 7(3):381–388, 2011.

[17] Stephanie C Ross, Petter Nilsson, Necmiye Ozay, and Johanna L Mathieu. Managing voltage excursions on the distribution network by limiting the aggregate variability of thermostatic loads. In 2019 American Control Conference (ACC), pages 4260–4267. IEEE, 2019.

[18] Stephanie C Ross, Necmiye Ozay, and Johanna L Mathieu. Coordination between an aggregator and distribution operator to achieve network-aware load control. In 2019 IEEE Milan PowerTech, pages 1–6. IEEE, 2019.

[19] Stephanie Crocker Ross and Johanna L Mathieu. A method for ensuring a load aggregator's power deviations are safe for distribution networks. Electric Power Systems Research, 189:106781, 2020.

[20] Stephanie Crocker Ross and Johanna L Mathieu. Strategies for network-safe load control with a third-party aggregator and a distribution operator. IEEE Transactions on Power Systems, 2021.

[21] Stephanie Crocker Ross, Gabrielle Vuylsteke, and Johanna L Mathieu. Effects of load-based frequency regulation on distribution network operation. IEEE Transactions on Power Systems, 34(2):1569–1578, 2018.

[22] Wenbo Shi, Na Li, Xiaorong Xie, Chi-Cheng Chu, and Rajit Gadh. Optimal residential demand response in distribution networks. IEEE journal on selected areas in communications, 32(7):1441–1450, 2014.

[23] Pierluigi Siano. Demand response and smart grids—a survey. Renewable and sustainable energy reviews, 30:461–478, 2014.

[24] Goran Strbac. Demand side management: Benefits and challenges. Energy policy, 36(12):4419–4426, 2008.

[25] Kai Strunz, Ehsan Abbasi, Chad Abbey, Christophe Andrieu, F Gao, T Gaunt, A Gole, N Hatziargyriou, and R Iravani. Benchmark systems for network integration of renewable and distributed energy resources. Cigre Task Force C, 6(04-02):78, 2009.

[26] Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction. MIT press, 2018.

[27] J.A. Taylor and J.L. Mathieu. Uncertainty in Demand Response – Identification, Estimation, and Learning, chapter 5, pages 56–70. Tutorials in Operations Research. INFORMS, 2015.

[28] Josh A Taylor, Sairaj V Dhople, and Duncan S Callaway. Power systems without fuel. Renewable and Sustainable Energy Reviews, 57:1322–1336, 2016.

[29] L. Thurner et al. Pandapower—an open-source python tool for convenient modeling, analysis, and optimization of electric power systems. IEEE Transactions on Power Systems, 33(6):6510–6521, 2018.

[30] Evangelos Vrettos and Göran Andersson. Combined load frequency control and active distribution network management with thermostatically controlled loads. In 2013 IEEE International Conference on Smart Grid Communications (SmartGridComm), pages 247–252, 2013.