Two-Timescale Voltage Regulation in Distribution Grids Using Deep Reinforcement Learning

Qiuling Yang
School of Automation
Beijing Institute of Technology
Beijing 100081, China
yang6726@umn.edu

Gang Wang, Alireza Sadeghi, Georgios B. Giannakis
ECE Dept. and Digital Tech. Center
University of Minnesota
Minneapolis, MN 55455, USA
{gangwang, sadeghi, georgios}@umn.edu

Jian Sun
School of Automation
Beijing Institute of Technology
Beijing 100081, China
sunjian@bit.edu.cn

***Abstract*—Frequent and sizeable voltage fluctuations become more pronounced with the increasing penetration of distributed renewable generation, and they considerably challenge distribution grids. Voltage regulation schemes so far have relied on either utility-owned devices (e.g., voltage transformers, and shunt capacitors), or more recently, smart power inverters that come with contemporary distributed generation units (e.g., photovoltaic systems, and wind turbines). Nonetheless, due to the distinct response times of those devices, as well as the discrete on-off commitment of capacitor units, joint control of both types of assets is challenging. In this context, a novel two-timescale voltage regulation scheme is developed here by coupling optimization with reinforcement learning advances. Shunt capacitors are configured on a slow timescale (e.g., daily basis) leveraging a deep reinforcement learning algorithm, while optimal setpoints of the power inverters are computed using a linearized distribution flow model on a fast timescale (e.g., every few seconds or minutes). Numerical experiments using a real-world 47-bus distribution feeder showcase the remarkable performance of the novel scheme.**

***Index Terms*—two-timescale, voltage regulation, inverter, capacitor, deep reinforcement learning.**

I. Introduction

Contemporary distribution grids are undergoing a rapid evolution due to the growing deployment of electric vehicles and distributed renewable generators such as photovoltaic (PV) systems and wind turbines. Electricity utilities in the US are currently experiencing major challenges related to the unparallel levels of load peaks and large voltage fluctuations. For example, over-voltage happens during mid-day when PV generation peaks and load demand is relatively low; whereas, on the other hand, voltage sags considerably overnight due to low PV generation yet a high load demand [1].

To keep the voltages within an acceptable range, early approaches have relied on configuring on a daily or even slower basis the utility-owned devices, namely load-tap-changing transformers and capacitors [2]. Such configurations have been effective without (or with low) renewable generation, and with slowly varying aggregate load. Specifically, auto-transformers were controlled to address the over-voltage issues in [3]. To set up the tap positions, a semidefinite relaxation approach was proposed in [4].

With the ever growing renewable generation and electric vehicles, rapid voltage changes occur often.For instance, the PV generation can fluctuate up to 15% of their nameplate ratings within one-minute intervals [5], [6]. Voltage regulation in this case would entail more frequent switching actions as well as further installation of control devices. Smart power inverters, on the other hand, come with contemporary distributed generation units and electric vehicles. Equipped with two-way communication and computing capabilities, they can be commanded to adjust reactive power output within milliseconds and in a continuously-valued manner [7], [8]. Indeed, recent proposals have focused on engaging power inverters in energy management (including voltage regulation) of distribution networks; see, for example, [1], [6]. A second order cone program (SOCP) relaxation approach was adopted to tackle the nonconvex optimization of the inverter VAR control problem in [9]. A stochastic and online learning approach was devised to minimize the power loss through reactive power compensation in [8]. To minimize communications between power inverters and the controlling center, local or decentralized voltage regulation schemes were developed in [10]–[14].

Despite considerable success of those approaches, joint control of both the traditional utility-owned devices and contemporary power inverters has not been explored. In this work, we focus on shunt capacitors and PV inverters, and engage them in reactive power provision. In this context, a novel two-timescale voltage regulation scheme is developed. Discrete actions (corresponding to on-off commitments of capacitors) are found using a deep reinforcement learning algorithm [15] on a slow timescale (e.g., hourly basis), while the optimal setpoints of inverters are obtained based on a linearized distribution flow model every few seconds. Numerical experiments on a real-world distribution grid using real solar and load data corroborate the merits of our novel approach.

Notation. Lower- (upper-) case boldface letters denote column vectors (matrices), with the exception of power flow vectors $(\boldsymbol{P}, \boldsymbol{Q})$, and normal letters represent scalars. Calligraphic symbols are reserved for sets, while $\mathbf{1}$ denotes all-ones vector, $\mathbb{R}_+^{\mathrm{N}}$ denotes the set of all non-negative N-dimensional vectors. Symbol $^\top$ stands for transposition, and $\|\boldsymbol{x}\|$ is the l_2-norm of $\boldsymbol{x}$.

The work of Q. Yang and J. Sun was supported in part by the National Natural Science Foundation of China under Grants 61621063, 61522303, 61720106011, 61621063, and in part by the Program for Changjiang Scholars and Innovative Research Team in University (IRT1208). Q. Yang was also supported by the China Scholarship Council. The work of G. Wang, A. Sadeghi, and G. B. Giannakis was supported in part by the National Science Foundation under Grants 1508993, 1509040, and 1711471.

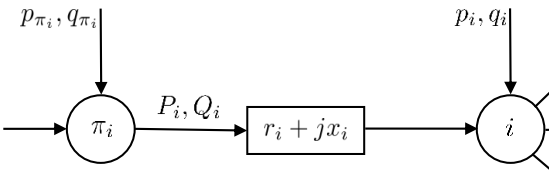

Fig. 1. Bus i is connected to its unique parent π_i via line

II. Problem Formulation

Consider a distribution grid comprising $N+1$ buses as a graph $\mathcal{G} := (\mathcal{N}_0, \mathcal{L})$, where $\mathcal{N}_0 := \{0, 1, \ldots, N\}$ all buses, and $\mathcal{L}$ collects all lines. The grid is typically radially as a tree, with its root at the substation bus by $i = 0$, and thus $|\mathcal{L}| = N$. The substation is connected to a transmission grid, at which the squared voltage magnitude is regulated to a constant. All buses except the substation comprise the set $\mathcal{N} := \{1, 2, \ldots, N\}$. For $\forall i \in \mathcal{N}$, let v_i denote the squared voltage magnitude, and $p_i + jq_i$ be its complex power injection into the grid, where $p_i := p_i^g - p_i^c$ and $q_i := q_i^g - q_i^c$ are found as the surplus between corresponding generation and consumption. For notational brevity, column vectors $\boldsymbol{v}$, $\boldsymbol{p}$, $\boldsymbol{q}$, $\boldsymbol{p}^g$, $\boldsymbol{q}^g$, $\boldsymbol{p}^c$, and $\boldsymbol{q}^c$ collect the corresponding quantities of all buses. In a radial grid, every leaf bus $i \in \mathcal{N}$ has a unique parent bus π_i, and the two are connected via the i-th transmission line denoted by $(\pi_i, i) \in \mathcal{L}$. Let $P_i + jQ_i$ represent the complex power flow from π_i to i, and $r_i + jx_i$ represent the impedance of line $i \in \mathcal{L}$, depicted in Fig. 1. Throughout, $\boldsymbol{p}$ and $\boldsymbol{q}^c$ are supposed to be given quantities.

In this paper, we consider shunt capacitors and power inverters for voltage regulation of a distribution network. Yet, our novel two-timescale approach based on deep reinforcement learning can also account for other types of utility-owned devices that have discrete actions and slow responses. On a fast timescale, smart inverters are controlled on a minute or 30-second basis, while shunt capacitors are configured on a slow timescale, say e.g., every hour or day. Suppose there is a total of N_a capacitors in the grid, whose indices are collected in $\mathcal{N}_a$ and are in one-to-one correspondence with entries of $\mathcal{K} := \{1, 2, \ldots, N_a\}$. Assume that every bus is installed with either a capacitor or a PV system (thus, an inverter), but not both. The remaining buses, after excluding $\mathcal{N}_a$ from $\mathcal{N}$, which are collectively denoted by $\mathcal{N}_r$, are equipped with inverters. This assumption is made without loss of generality as one can simply set the upper and lower bounds on the inverters' output to zero at buses having no PVs installed. In our model, we divide a day into $N_{\bar{T}}$ intervals indexed by $\tau = 1, \ldots, N_{\bar{T}}$. Each of these $N_{\bar{T}}$ intervals is further partitioned into N_T time slots which are indexed by $t = 1, \ldots, N_T$, as depicted in Fig. 2. To match the slow load variations, the on-off decisions of capacitors are made (at the end of) every interval τ, which can be chosen to be e.g., an hour; yet, to accommodate the rapidly changing renewable generation, the inverter output is adjusted (at the beginning of) every slot t, taken to be e.g., a minute. We assume that quantities $\boldsymbol{p}^g(\tau, t)$, $\boldsymbol{p}^c(\tau, t)$, and $\boldsymbol{q}^c(\tau, t)$ remain the same within each t-slot, but may change from slot t to $t+1$.

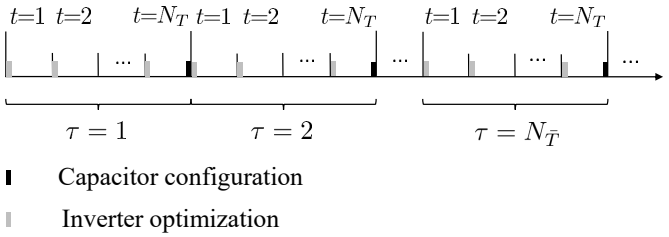

Fig. 2. Two-timescale partitioning of a day for joint capacitor and inverter control.

Since the shunt capacitor configuration is updated on a slow timescale (i.e., every τ), the reactive compensation $q_i^g(\tau, t)$ provided by capacitor $k_i \in \mathcal{K}$ (or, the capacitor installed at bus i) is represented by

$$q_i^g(\tau, t) = \hat{y}_{k_i}(\tau) q_{a,k_i}^g, \ \forall i \in \mathcal{N}_a, \tau, t \tag{1}$$

where $\hat{y}_{k_i}(\tau) \in \{0, 1\}$ is the on-off commitment of capacitor k_i for the entire interval τ. Clearly, if $\hat{y}_{k_i}(\tau) = 1$, a constant amount (nameplate value) of reactive power q_{a,k_i}^g is injected into the grid during this interval, and 0 otherwise. For convenience, the on-off decisions of capacitor units at interval τ are collected in a column vector $\hat{\boldsymbol{y}}(\tau)$. On the other hand, the reactive power $q_{r,i}^g(\tau, t)$ generated by inverter i is adjusted on a faster scale (every t), which is constrained as

$$|q_{r,i}^g(\tau, t)| \leq \bar{q}_i^g := \sqrt{(\bar{s}_i^g)^2 - (\bar{p}_i^g)^2}, \ \forall i \in \mathcal{N}_r, t \tag{2}$$

where $\bar{s}_i^g$ and $\bar{p}_i^g$ are the nameplate values of apparent power and active power for inverter i, respectively; see e.g., [8].

Given real-time load consumption and generation that are modeled as Markovian processes [16], the objective of voltage regulation is minimizing the *long-term* average voltage deviation by finding the optimal reactive power support per slot through configuring capacitors every interval and adjusting inverters' outputs every slot. As voltage magnitudes $\boldsymbol{v}(\tau, t)$ depend solely on the control variables $\boldsymbol{q}^g(\tau, t)$, they are expressed as implicit functions of $\boldsymbol{q}^g(\tau, t)$, yielding $\boldsymbol{v}_{\tau,t}(\boldsymbol{q}^g(\tau, t))$, whose actual function forms for postulated grid models are postponed to Section III. The novel two-timescale voltage regulation scheme entails solving the following stochastic optimization problem

$$\underset{\substack{\{\boldsymbol{q}_r^g(\tau,t)\} \\ \{\boldsymbol{y}(\tau) \in \{0,1\}^{N_a}\}}}{\text{minimize}} \quad \mathbb{E}\left[\sum_{\tau=1}^{\infty} \sum_{t=1}^{N_T} \gamma^\tau \left\| \boldsymbol{v}_{\tau,t}(\boldsymbol{q}^g(\tau, t)) - v_0 \mathbf{1} \right\|^2\right] \tag{3a}$$

$$\text{subject to} \quad q_i^g(\tau, t) = \hat{y}_{k_i}(\tau) q_{a,k_i}^g, \quad \forall i \in \mathcal{N}_a, \tau, t \tag{3b}$$

$$q_i^g(\tau, t) = q_{r,i}^g(\tau, t), \quad \forall i \in \mathcal{N}_r, \tau, t \tag{3c}$$

$$|q_{r,i}^g(\tau, t)| \leq \bar{q}_i^g, \quad \forall i \in \mathcal{N}_r, \tau, t \tag{3d}$$

for some discount factor $\gamma \in (0, 1]$, where the expectation is taken over the joint distribution of $(\boldsymbol{p}^c(\tau, t), \boldsymbol{q}^c(\tau, t), \boldsymbol{p}^g(\tau, t))$ across all intervals and slots. Clearly, discrete variables $\hat{\boldsymbol{y}}(\tau) \in \{0, 1\}^{N_a}$ render problem (3) nonconvex and *NP-hard* in general. Moreover, it is a multi-scale optimization, whose decisions

are not all made at the same timescale and must also account for the power variability during real-time operation. In words, tackling (3) exactly is challenging.

In this work, a stochastic optimization approach combining physics principles as well as data-driven advances is pursued. Specifically, on a slow timescale, say at the end of each interval $\tau-1$, the optimal on-off capacitor decisions $\boldsymbol{y}(\tau)$ are approached through a deep reinforcement learning algorithm that can learn from the predictions collected within the current interval $\tau-1$; while, on a fast timescale, say at the beginning of each slot t within interval τ, our two-stage control paradigm computes the optimal reactive power setpoints for inverters, by minimizing instantaneous bus voltage deviations while respecting physical constraints, given the current on-off commitment of capacitor units $\hat{\boldsymbol{y}}(\tau)$ found at the very end of interval $(\tau-1)$. These two timescales are elaborated in Sections III and IV, respectively.

III. Fast-timescale Control of Inverters

As alluded earlier, the actual forms of $\boldsymbol{v}_{\tau,t}(\boldsymbol{q}^g(\tau,t))$ will be specified in this section, relying on a linearized approximation model. Throughout this section, the interval index τ will be omitted when clear from context. Leveraging the linearized distribution flow model in [17], the power flow equations for all buses $i\in\mathcal{N}$, and for all t within every interval τ is dictated as follows

$$p_i(t)=\sum_{j\in\chi_i}P_j(t)-P_i(t) \tag{4a}$$

$$q_i(t)=\sum_{j\in\chi_i}Q_j(t)-Q_i(t) \tag{4b}$$

$$v_i(t)=v_{\pi_i}(t)-2(r_iP_i(t)+x_iQ_i(t)). \tag{4c}$$

Adopting the approximate grid model in (4), the optimal setpoints of smart inverters can be found by solving the following optimization problem at every slot t within interval τ, given again $\hat{\boldsymbol{y}}(\tau)$ available from the last interval on the slow timescale

$$\underset{\boldsymbol{v}(t),\boldsymbol{q}^g_\tau(t),\boldsymbol{P}(t),\boldsymbol{Q}(t)}{\text{minimize}} \quad \|\boldsymbol{v}(t)-v_0\mathbf{1}\|^2 \tag{5a}$$

$$\text{subject to} \quad (4a)-(4c)$$

$$q^g_i(t)=\hat{y}_{k_i}(\tau)q^g_{a,k_i},\quad \forall i\in\mathcal{N}_a \tag{5b}$$

$$q^g_i(t)=q^g_{r,i}(t),\quad \forall i\in\mathcal{N}_r \tag{5c}$$

$$|q^g_{r,i}(t)|\le\bar{q}^g_i,\quad \forall i\in\mathcal{N}_r. \tag{5d}$$

Observing that all constraints are linear and the cost function is quadratic, problem (5) constitutes a standard convex quadratic program. As such, it can be solved efficiently by e.g., primal-dual type algorithms, or off-the-shelf convex programming toolboxes, whose implementation details are skipped due to space limitations.

IV. Slow-timescale Reconfiguration of Capacitors

Due to the discrete nature of configuring shunt capacitors (determining their on-off state), traditional approaches were mainly based on heuristics or relied on convex relaxation. They often yield sub-optimal performance while incurring high computational and storage complexities. To address these challenges, we draw from recent advances in artificial intelligence, and advocate a deep reinforcement learning (DRL) approach for optimally and efficiently configuring shunt capacitors using minimal information available from the last interval. Toward this goal, the optimal capacitor configuration is cast as an MDP, which is solved using the DRL algorithm [15]. An MDP is defined as a 5-tuple $(\mathcal{S},\mathcal{A},\mathcal{P},c,\gamma)$, where $\mathcal{S}$ is a set of states; $\mathcal{A}$ is a set of actions; $\mathcal{P}$ is a set of transition matrices; $c:\mathcal{S}\times\mathcal{A}\mapsto\mathbb{R}$ is a cost function such that, for $\boldsymbol{s}\in\mathcal{S}$ and $\boldsymbol{a}\in\mathcal{A}$, $c=(c(\boldsymbol{s},\boldsymbol{a}))_{\boldsymbol{s}\in\mathcal{S},\boldsymbol{a}\in\mathcal{A}}$ are the real-valued immediate costs after the system operator takes an action $\boldsymbol{a}$ at state $\boldsymbol{s}$; and $\gamma\in[0,1)$ is the discount factor. These components are defined next before introducing our voltage regulation scheme.

Action space $\mathcal{A}$. Each action corresponds to one possible on-off commitment of capacitors 1 to N_a. Hence, one action is described as $\boldsymbol{a}=\boldsymbol{y}$. Specially, at interval τ, the action is $\boldsymbol{a}(\tau)=\boldsymbol{y}(\tau)$. Note that the action is discrete in the capacitor configuration problem. The set of all actions constitute the action space $\mathcal{A}$, whose cardinality grows exponentially with the number of capacitors; indeed, we have that $|\mathcal{A}|=2^{N_a}$.

State space $\mathcal{S}$. Define the average active power at all buses except for the substation over the current interval τ, along with the current capacitor configuration as the state at the interval τ; that is, $\boldsymbol{s}(\tau):=[\bar{\boldsymbol{p}}^\top(\tau),\hat{\boldsymbol{y}}^\top(\tau)]^\top$, which contains both continuous and discrete variables. Clearly, it holds for the state space that $\mathcal{S}\subseteq\mathbb{R}^N\times 2^{N_a}$.

The action is determined by the configuration policy π which is a function of the most recent state $\boldsymbol{s}(\tau-1)$, given as

$$\boldsymbol{a}(\tau)=\pi(\boldsymbol{s}(\tau-1)). \tag{6}$$

Cost function c. The cost on a slow timescale is defined as

$$c(\boldsymbol{s}(\tau-1),\boldsymbol{a}(\tau))=\sum_{t=1}^{N_T}\|\boldsymbol{v}_{\tau,t}(\boldsymbol{q}^g(\tau,t))-v_0\mathbf{1}\|^2. \tag{7}$$

Set of transition probability matrices $\mathcal{P}$ and discount factor γ. While being in a state $\boldsymbol{s}\in\mathcal{S}$ upon taking an action $\boldsymbol{a}$, the system moves into a new state $\boldsymbol{s}'\in\mathcal{S}$ probabilistically. Let us define, the transition probability matrix from any state $\boldsymbol{s}$ to any next state $\boldsymbol{s}'$ under a given action $\boldsymbol{a}$ as $P^{\boldsymbol{a}}_{\boldsymbol{s}\boldsymbol{s}'}$; evidently, it holds that $\mathcal{P}:=\{P^{\boldsymbol{a}}_{\boldsymbol{s}\boldsymbol{s}'}|\forall\boldsymbol{a}\in\mathcal{A}\}$. The discount factor $\gamma\in[0,1)$, trades off the current versus future costs. The smaller γ is, the more weight the current cost has in the overall cost.

Given the current state and action, the action-value function under the control policy π is defined as

$$Q_\pi(\boldsymbol{s}(\tau-1),\boldsymbol{a}(\tau)):=\mathbb{E}\left[\sum_{\tau'=\tau}^{\infty}\gamma^{\tau'-\tau}c(\boldsymbol{s}(\tau'-1),\boldsymbol{a}(\tau'))\Big|\pi,\boldsymbol{s}(\tau-1),\boldsymbol{a}(\tau)\right] \tag{8}$$

where the expectation $\mathbb{E}$ is taken with respect to all sources of randomness.

To find the optimal capacitor configuration policy π^*, that minimizes the average voltage deviation in the long run, we resort to the Bellman optimality equations; see e.g., [18].

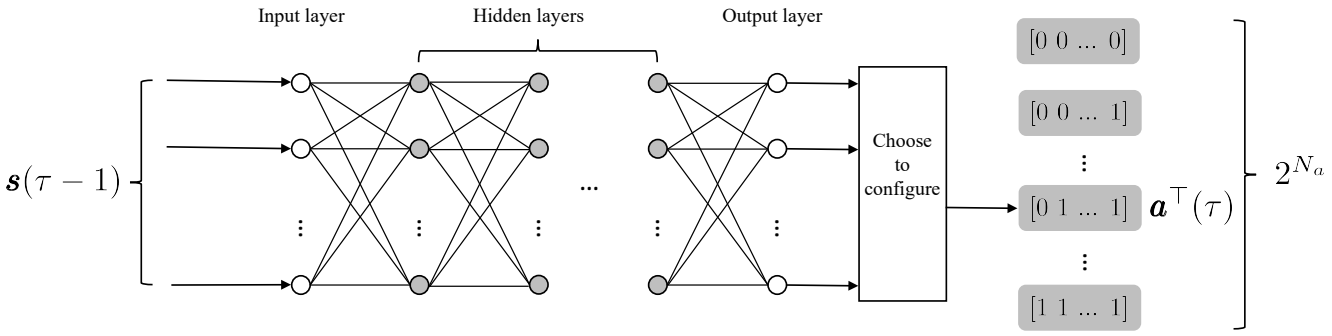

Fig. 3. Deep Q-network

Solving those yields the action-value function under the optimal policy π^* on the fly, given by

$$Q_{\pi^*}(\boldsymbol{s},\boldsymbol{a}) = \mathbb{E}[c(\boldsymbol{s},\boldsymbol{a})] + \gamma \sum_{\boldsymbol{s}'\in\mathcal{S}} P^{\boldsymbol{a}}_{\boldsymbol{s}\boldsymbol{s}'} \min_{\boldsymbol{a}\in\mathcal{A}} Q_{\pi^*}(\boldsymbol{s}',\boldsymbol{a}'). \quad (9)$$

With $Q_{\pi^*}(\boldsymbol{s},\boldsymbol{a})$ obtained, the optimal capacitor configuration policy can be obtained as

$$\pi^*(\boldsymbol{s}) = \arg\min_{\boldsymbol{a}} Q_{\pi^*}(\boldsymbol{s},\boldsymbol{a}). \quad (10)$$

It is clear from (9) that, if all transition probabilities $\{P^{\boldsymbol{a}}_{\boldsymbol{s}\boldsymbol{s}'}\}$ were available, we can derive $Q_{\pi^*}(\boldsymbol{s},\boldsymbol{a})$, and subsequently the optimal policy π^* from (10). Unfortunately, it is impossible to obtain those transition probabilities in a real-world cyber-physical distribution network. Targeting directly the optimal policy π^* without having the knowledge of $P^{\boldsymbol{a}}_{\boldsymbol{s}\boldsymbol{s}'}$, Q-learning approaches find π^* by learning $Q_{\pi^*}(\boldsymbol{s},\boldsymbol{a})$ 'on-the-fly' [18]. However, they are infeasible for the problem at hand, due to the large-size continuous state space $\mathcal{S}$. This motivates function approximation based reinforcement learning (RL) schemes that can reliably work and generalize on continuous state domains [18]. Using a neural network function approximator to estimate the action-value function, DRL approaches have gained popularity because they perform remarkably well in dealing with high-dimensional and/or continuous state spaces [15], [19], [20].

Specifically, this work considers a feed-forward neural network, shown in Fig. 3. It takes as input the state vector $\boldsymbol{s}(\tau-1)$, followed by L fully connected hidden layers, and has a separate output unit for each possible action (clearly, a total of 2^{N_a} outputs in our context). These outputs correspond to the predicted Q-values (i.e., estimated costs) of the individual actions (on-off configurations of all capacitors) for the input state vector. By stacking all the weight parameters of a deep neural network into a vector $\boldsymbol{\theta}$, we have a function approximation to estimate the action-value function $Q_\pi(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}) \approx Q_{\pi^*}(\boldsymbol{s},\boldsymbol{a})$. Finally, the grid operator can take an action according to (10) to configure the network capacitors.

To enable the DQN to estimate the costs for all possible capacitor configurations, one needs to train the DQN by iteratively updating $\boldsymbol{\theta}_\tau$ (the weights of the Q-network at iteration τ). To this end, we draw from recent advances in deep reinforcement learning [15]. In particular, the new experience is denoted by $e(\tau') := (\boldsymbol{s}(\tau'-1),\boldsymbol{a}(\tau')), c(\boldsymbol{s}(\tau'-1),\boldsymbol{a}(\tau')),\boldsymbol{s}(\tau'))$. Consider having a replay buffer $\mathcal{R}(\tau)$, which stores the most recent R experiences. That is, at a given interval τ, the replay buffer is $\mathcal{R}(\tau) := \{e(\tau-R+1), e(\tau-R+2),\ldots,e(\tau)\}$. Moreover, to stabilize the DQN updates, we create and maintain a second deep Q-network, commonly referred to as the *target* network, with weight parameters denoted by $\boldsymbol{\theta}^{\mathrm{Tar}}$. Further, the target network is not trained, but its parameters $\boldsymbol{\theta}^{\mathrm{Tar}}$ are reset to $\boldsymbol{\theta}$ periodically, say every B training iterations of the DQN. Consider a least-squares loss for the experience $e(\tau')$

$$\mathcal{L}^{\mathrm{Tar}}(\boldsymbol{\theta}_\tau; e(\tau')) := \frac{1}{2}\Big[c(\boldsymbol{s}(\tau'-1),\boldsymbol{a}(\tau')) + \gamma \min_{\boldsymbol{a}'} Q^{\mathrm{Tar}}(\boldsymbol{s}(\tau),\boldsymbol{a}';\boldsymbol{\theta}^{\mathrm{Tar}}_{\tau'}) - Q(\boldsymbol{s}(\tau'-1),\boldsymbol{a}(\tau');\boldsymbol{\theta}_\tau)\Big]^2. \quad (11)$$

Upon taking expectation with respect to all sources of randomness generating this experience, we have the following expected loss

$$\mathcal{L}^{\mathrm{Tar}}(\boldsymbol{\theta}_\tau; \mathcal{R}(\tau))) := \mathbb{E}_{e(\tau')}\, \mathcal{L}^{\mathrm{Tar}}(\boldsymbol{\theta}_\tau; e(\tau')). \quad (12)$$

In practice however, as no distributional knowledge is available, one has to approximate the expected loss with some empirical loss. To that end, we draw a mini-batch of size M_τ experiences uniformly at random from the buffer $\mathcal{R}(\tau)$, whose indices are collected in $\mathcal{M}_\tau$, i.e., $\{e(\tau')\}_{\tau'\in\mathcal{M}_\tau} \sim U(\mathcal{R}(\tau))$. Upon computing for each of these sampled experiences an output using the target network, we define the following empirical loss

$$\mathcal{L}^{\mathrm{Tar}}(\boldsymbol{\theta}_\tau; \mathcal{M}_\tau) := \frac{1}{2M_\tau} \sum_{\tau'\in\mathcal{M}_\tau} \Big[c(\boldsymbol{s}(\tau'-1),\boldsymbol{a}(\tau')) + \gamma \min_{\boldsymbol{a}'} Q^{\mathrm{Tar}}(\boldsymbol{s}(\tau'),\boldsymbol{a}';\boldsymbol{\theta}^{\mathrm{Tar}}_{\tau}) - Q(\boldsymbol{s}(\tau'-1),\boldsymbol{a}(\tau');\boldsymbol{\theta}_\tau)\Big]^2. \quad (13)$$

In words, the weight parameters of the DQN are updated using stochastic gradient descent (SGD) over the empirical loss $\mathcal{L}^{\mathrm{Tar}}(\boldsymbol{\theta}_\tau; \mathcal{M}_\tau)$, as follows

$$\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \beta_\tau \nabla \mathcal{L}^{\mathrm{Tar}}(\boldsymbol{\theta}_\tau; \mathcal{M}_\tau). \quad (14)$$

where $\beta_\tau > 0$ is a preselected learning rate, and $\nabla\mathcal{L}(\boldsymbol{\theta})$ denotes the (sub-)gradient. The target network and experience replay scheme result in stable updates when training a DQN in an unsupervised fashion. The novel voltage regulation scheme is summarized in Alg. 1.

Algorithm 1 Two-timescale voltage regulation scheme.

Initialize: weight $\boldsymbol{\theta}_0$ randomly; weight of target network $\boldsymbol{\theta}_0^{\text{Tar}} = \boldsymbol{\theta}_0$; replay buffer $\mathcal{R}$; and initial state $\boldsymbol{s}(0)$.
for $\tau = 1, 2, \ldots$ **do**
  Take action $\boldsymbol{a}(\tau)$ through exploration-exploitation

$$\boldsymbol{a}(\tau) = \begin{cases} \text{random} \quad \boldsymbol{a} \in \mathcal{A} & \text{w.p. } \epsilon_\tau \\ \arg\min_{\boldsymbol{a}'} Q(\boldsymbol{s}(\tau-1), \boldsymbol{a}'; \boldsymbol{\theta}_\tau) & \text{w.p. } 1-\epsilon_\tau \end{cases}$$

  Evaluate $\boldsymbol{c}(\boldsymbol{s}(\tau-1), \boldsymbol{a}(\tau))$ using (7).
  for $t = 1, 2, ..., N_T$ **do**
    Compute $\boldsymbol{q}^g(\tau, t)$ using (5).
  end for
  Update $\boldsymbol{s}(\tau)$.
  Save $(\boldsymbol{s}(\tau-1), \boldsymbol{a}(\tau), c(\boldsymbol{s}(\tau-1), \boldsymbol{a}(\tau)), \boldsymbol{s}(\tau))$ into $\mathcal{R}(\tau)$.
  Randomly sample M_τ experiences from $\mathcal{R}(\tau)$.
  Form the mini-batch loss $\mathcal{L}^{\text{Tar}}(\boldsymbol{\theta}_\tau; \mathcal{M}_\tau)$ using (13).
  Update $\boldsymbol{\theta}_{\tau+1}$ using (14).
  if $\text{mod}(\tau, B) = 0$ **then**
    Update the target network $\boldsymbol{\theta}_\tau^{\text{Tar}} = \boldsymbol{\theta}_\tau$.
  end if
end for

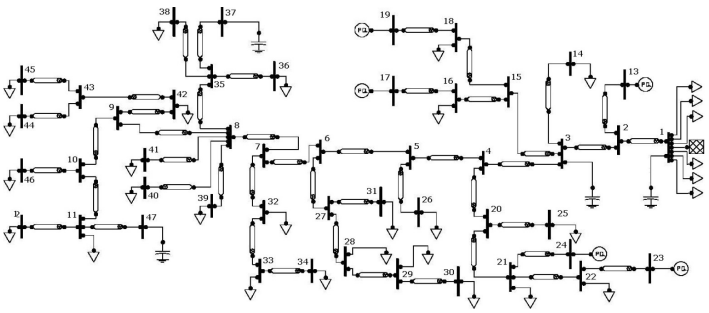

Fig. 4. Schematic diagram of the 47-bus industrial distribution feeder.

V. Numerical Tests

The two-timescale voltage regulation scheme presented in Alg. 1 is numerically examined using the Southern California Edison 47-bus distribution feeder [9], depicted in Fig. 4. Real consumption and solar generation data were obtained from the Smart* project collected on August 24, 2011 [5], and preprocessed by following the procedure detailed in [8]. This distribution feeder integrates with four shunt capacitors and five PVs. As one capacitor is installed at the substation whose voltage magnitude v_0 is regulated to 1 through a voltage regulation transformer, it is excluded from our control paradigm. The rest three capacitors are installed on buses 3, 37, and 47, with capacities 120, 180, and 180 kVar, respectively, while the five large PV plants are located on buses 2, 16, 18, 21, and 22, with capacities 300, 80, 300, 400, and 200 kW, respectively. For our test to match the availability of real data, every slot t was taken as a minute, while every interval τ was 5 minutes. A power factor of 0.8 was assumed for all loads. The DQN used a fully connected feed-forward neural network of 3 layers, which was found sufficient. The rectified linear unit (ReLU) activation functions were used in the hidden layers while the logistic sigmoid functions were at the output layer [21]. The

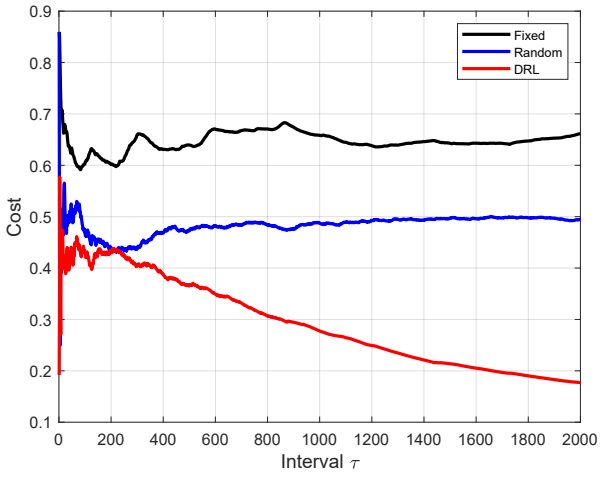

Fig. 5. Time-averaged instantaneous costs incurred by the three schemes.

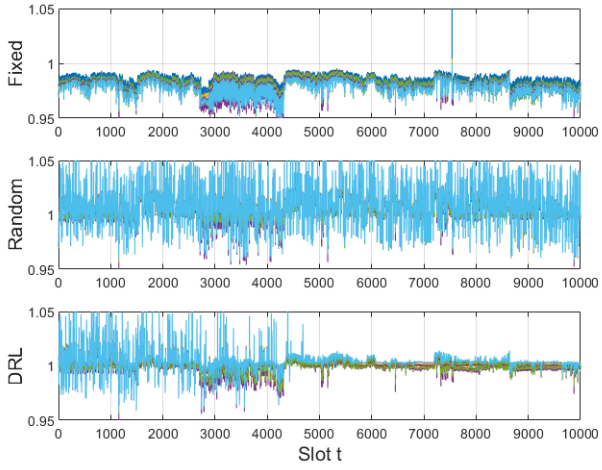

Fig. 6. Voltage magnitude profiles obtained by the three schemes over the simulation period of 10,000 slots.

replay buffer size was $R = 10$, the discount factor $\gamma = 0.99$, and the mini-batch size $\mathcal{M}_\tau = 10$. The target network was updated every $B = 5$ iterations.

To benchmark the performance of our scheme, we adopted a fixed and a randomly switching capacitor configuration policies as baselines. Both schemes compute the optimal inverter setpoints by solving (5) on a fast timescale, while the former employs a fixed capacitor configuration, and the latter switches its capacitor commitment randomly per slow-timescale interval. The time-averaged instantaneous costs $(1/\tau)\sum_{i=1}^{\tau} c(\boldsymbol{s}(i-1), \boldsymbol{a}(i))$ incurred by the three schemes over the first 2,000 intervals are plotted in Fig. 5. Clearly, the novel scheme achieved a lower cost than the other two after a short period of learning. Voltage magnitude profiles at all buses regulated by the three schemes are presented in Fig. 6. Again, after a short period ($\sim 4,500$ slots) of training by interacting with the environment, our DRL-base voltage regulation scheme quickly learns a stable and (near-) optimal policy. In addition, voltage

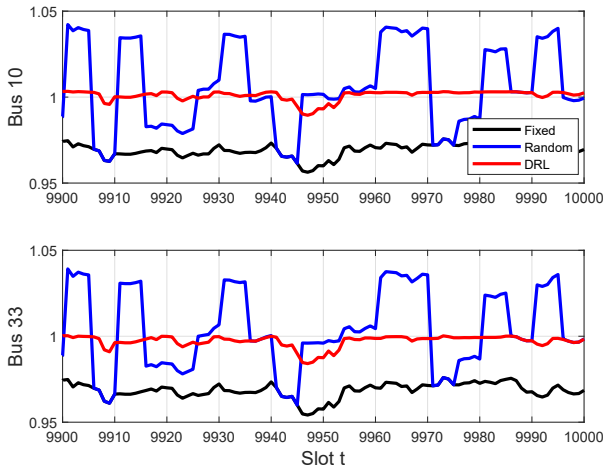

Fig. 7. Voltage magnitude profiles obtained by the three schemes at buses 10 and 33 from slot $9,900$ to slot $10,000$.

magnitude profiles regulated by three schemes at buses 10 and 33 from slot $9,900$ to $10,000$ are presented in Fig. 7. Curves showcase the effectiveness of the novel scheme in smoothing voltage fluctuations due to high solar generation as well as heavy load demand.

VI. Conclusions

This paper put forward a two-timescale voltage regulation scheme for residential distribution networks, by means of joint control of smart power inverters and shunt capacitors that are readily available in contemporary distribution grids. In particular, a linearized distribution flow model was adopted to determine the optimal setpoints for power inverters on a fast timescale (say, e.g., on a minute basis), while the optimal configurations of shunt capacitors were obtained on the fly using a deep reinforcement learning algorithm on a slow timescale (e.g., per hour). The proposed scheme was shown to be efficient in practice and easy to implement, through numerical tests on a real-world distribution feeder using real solar and consumption data.

References

[1] P. M. Carvalho, P. F. Correia, and L. A. Ferreira, "Distributed reactive power generation control for voltage rise mitigation in distribution networks," *IEEE Trans. Power Syst.*, vol. 23, no. 2, pp. 766–772, May 2008.

[2] W. H. Kersting, *Distribution System Modeling and Analysis*. New York, NY, USA: CRC press, 2006.

[3] C. Masters, "Voltage rise: the big issue when connecting embedded generation to long 11 kV overhead lines," *Power Eng. J.*, vol. 16, no. 1, pp. 5–12, Feb. 2002.

[4] B. A. Robbins, H. Zhu, and A. D. Domínguez-García, "Optimal tap setting of voltage regulation transformers in unbalanced distribution systems," *IEEE Trans. Power Syst.*, vol. 31, no. 1, pp. 256–267, Feb. 2016.

[5] S. Barker, A. Mishra, D. Irwin, E. Cecchet, P. Shenoy, and J. Albrecht, "Smart*: An open data set and tools for enabling research in sustainable homes," *SustKDD*, vol. 111, no. 112, p. 108, Aug. 2012.

[6] G. Wang, V. Kekatos, A. J. Conejo, and G. B. Giannakis, "Ergodic energy management leveraging resource variability in distribution grids," *IEEE Trans. Power Syst.*, vol. 31, no. 6, pp. 4765–4775, Nov. 2016.

[7] A. Ipakchi and F. Albuyeh, "Grid of the future," *IEEE Power Energy Mag.*, vol. 7, no. 2, pp. 52–62, Feb. 2009.

[8] V. Kekatos, G. Wang, A. J. Conejo, and G. B. Giannakis, "Stochastic reactive power management in microgrids with renewables," *IEEE Trans. Power Syst.*, vol. 30, no. 6, pp. 3386–3395, Dec. 2015.

[9] M. Farivar, C. R. Clarke, S. H. Low, and K. M. Chandy, "Inverter VAR control for distribution systems with renewables," in *Proc. IEEE SmartGridComm.*, Brussels, Belgium, Oct. 2011, pp. 457–462.

[10] H. Zhu and H. J. Liu, "Fast local voltage control under limited reactive power: Optimality and stability analysis," *IEEE Trans. Power Syst.*, vol. 31, no. 5, pp. 3794–3803, Dec. 2016.

[11] V. Kekatos, L. Zhang, G. B. Giannakis, and R. Baldick, "Voltage regulation algorithms for multiphase power distribution grids," *IEEE Trans. Power Syst.*, vol. 31, no. 5, pp. 3913–3923, Sep. 2016.

[12] G. Wang, G. B. Giannakis, J. Chen, and J. Sun, "Distribution system state estimation: An overview of recent developments," *Front. Inform. Technol. Electron. Eng.*, vol. 20, no. 1, pp. 4–17, Jan. 2019.

[13] W. Lin, R. Thomas, and E. Bitar, "Real-time voltage regulation in distribution systems via decentralized PV inverter control," in *Proc. Annual Hawaii Intl. Conf. System Sciences*, Waikoloa Village, Hawaii, Jan. 2-6, 2018.

[14] M. Bazrafshan, N. Gatsis, and H. Zhu, "Optimal power flow with step-voltage regulators in multi-phase distribution networks," *IEEE Trans. Power Syst.*, pp. 1–1, 2019.

[15] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, Feb. 2015.

[16] J. A. Carta, P. Ramirez, and S. Velazquez, "A review of wind speed probability distributions used in wind energy analysis: Case studies in the Canary Islands," *Renew. Sust. Energ. Rev.*, vol. 13, no. 5, pp. 933–955, Jun. 2009.

[17] M. E. Baran and F. F. Wu, "Network reconfiguration in distribution systems for loss reduction and load balancing," *IEEE Trans. Power Del.*, vol. 4, no. 2, pp. 1401–1407, Apr. 1989.

[18] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT press, 2018.

[19] A. Sadeghi, G. Wang, and G. B. Giannakis, "Deep reinforcement learning for adaptive caching in hierarchical content delivery networks," *IEEE Trans. Cogn. Commun. Netw.*, 2019 (to appear).

[20] Q. Yang, G. Wang, A. Sadeghi, G. B. Giannakis, and J. Sun, "Deep reinforcement learning for voltage regulation in distribution grids," *IEEE Trans. Smart Grid*, 2019 (submitted).

[21] G. Wang, G. B. Giannakis, and J. Chen, "Learning ReLU networks on linearly separable data: Algorithm, optimality, and generalization," *IEEE Trans. Signal Process.*, vol. 67, no. 9, pp. 2357–2370, May 2019.