

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

Qiuling Yang, Gang Wang, *Member, IEEE*, Alireza Sadeghi,
Georgios B. Giannakis, *Fellow, IEEE*, and Jian Sun, *Member, IEEE*

Abstract—Modern distribution grids are currently being challenged by frequent and sizable voltage fluctuations, due mainly to the increasing deployment of electric vehicles and renewable generators. Existing approaches to maintaining bus voltage magnitudes within the desired region can cope with either traditional utility-owned devices (e.g., shunt capacitors), or contemporary smart inverters that come with distributed generation units (e.g., photovoltaic plants). The discrete on-off commitment of capacitor units is often configured on an hourly or daily basis, yet smart inverters can be controlled within milliseconds, thus challenging joint control of these two types of assets. In this context, a novel two-timescale voltage regulation scheme is developed for distribution grids by judiciously coupling data-driven with physics-based optimization. On a faster timescale, say every second, the optimal setpoints of smart inverters are obtained by minimizing instantaneous bus voltage deviations from their nominal values, based on either the exact alternating current power flow model or a linear approximant of it; whereas, on the slower timescale (e.g., every hour), shunt capacitors are configured to minimize the long-term discounted voltage deviations using a deep reinforcement learning algorithm. Extensive numerical tests on a real-world 47-bus distribution network as well as the IEEE 123-bus test feeder using real data corroborate the effectiveness of the novel scheme.

Index terms— Two timescales, voltage control, inverters, capacitors, deep reinforcement learning.

I. INTRODUCTION

Frequent and sizable voltage fluctuations caused by the growing deployment of electric vehicles, demand response programs, and renewable energy sources, challenge modern distribution grids. Electric utilities are currently experiencing major issues related to the unprecedented levels of load peaks as well as renewable penetration. For instance, a solar farm connected at the end of a long distribution feeder in a rural area can cause voltage excursions along the feeder, while the apparent power capability of a substation transformer is strained by frequent reverse power flows. Moreover, over-voltage happens

during midday when photovoltaic (PV) generation peaks and load demand is relatively low; whereas voltage sags occur mostly overnight due to low PV generation even when load demand is high [1]. This motivates why voltage regulation, the task of maintaining bus voltage magnitudes within desirable ranges, is critical in modern distribution grids.

Early approaches to regulating the voltages at a residential level have mainly relied on utility-owned devices, including load-tap-changing transformers, voltage regulators, and capacitor banks, to name a few. They offer a convenient means of controlling reactive power, through which the voltage profile at their terminal buses as well as at other buses can be regulated [2, p. 678]. Obtaining the optimal configuration for these devices entails solving mixed-integer programs, which are NP-hard in general. To optimize the tap positions, a semi-definite relaxation heuristic was used in [3], [4]. Control rules based on heuristics were developed in [5], [1]. However, these approaches can be computationally demanding, and do not guarantee optimal performance. A batch reinforcement learning (RL) scheme based on linear function approximation was lately advocated in [6].

Another characteristic inherent to utility-owned equipment is their limited life cycle, which prompts control on a daily or even monthly basis. Such configurations have been effective in traditional distribution grids without (or with low) renewable generation, and with slowly varying load. Yet, as distributed generation grows in residential networks nowadays [7], [8], rapid voltage fluctuations occur frequently. According to a recent landmark bill, California mandated 50% of its electricity to be powered by renewable resources by 2025 and 60% by 2030. The power generated by a solar panel can vary by 15% of its nameplate rating within one-minute intervals [9]. Voltage control would entail more frequent switching actions, and further installation of control devices.

Smart power inverters on the other hand, come with contemporary distributed generation units, such as PV panels, and wind turbines. Embedded with computing and communication units, these can be commanded to adjust reactive power output within seconds, and in a continuously-valued fashion. Indeed, engaging smart inverters in reactive power control has recently emerged as a promising solution [10]. Computing the optimal setpoints for inverters' reactive power output is an instance of the optimal power flow task, which is non-convex [11]. To deal with the renewable uncertainty as well as other communication issues (e.g., delay and packet loss), stochastic, online, decentralized, and localized reactive control schemes have been advocated [10], [12], [13], [9], [14], [15], [16].

Manuscript received April 19, 2019; revised May 28, and August 29, 2019; accepted October 31, 2019. The work of Q. Yang and J. Sun was supported in part by the National Natural Science Foundation of China under Grants 61522303, 61720106011, and 61621063. Q. Yang was also supported by the China Scholarship Council. The work of G. Wang, A. Sadeghi, and G. B. Giannakis was supported by National Science Foundation under Grants 1509040, 1711471, and 1901134.

Q. Yang and J. Sun are with the State Key Lab of Intelligent Control and Decision of Complex Systems, School of Automation, Beijing Institute of Technology, Beijing 100081, China (e-mail: yang6726@umn.edu, sunjian@bit.edu.cn). G. Wang, A. Sadeghi, and G. B. Giannakis are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA (e-mail: gangwang@umn.edu, sadeghi@umn.edu, georgios@umn.edu).

RL refers to a collection of tools for solving Markovian decision processes (MDPs), especially when the underlying transition mechanism is unknown [17]. In settings involving high-dimensional, continuous action and/or state spaces however, it is well known that conventional RL approaches suffer from the so-called ‘curse of dimensionality,’ which limits their impact in practice [18]. Deep neural networks (DNNs) can address the curse of dimensionality in the high-dimensional and continuous state space by providing compact low-dimensional representations of high-dimensional inputs [19]. Wedding deep learning with RL (using a DNN to approximate the action-value function), deep (D) RL has offered artificial agents with human-level performance across diverse application domains [18], [20]. (D)RL algorithms have also shown great potential in several challenging power systems control and monitoring tasks [21], [22], [6], [23], [24], [25], and load control [26], [27]. A batch RL scheme using linear function approximation was developed for voltage regulation in distribution systems [6]. For voltage control of transmission networks, DRL was recently investigated to adjust generator voltage setpoints [21]. A shortcoming of the mentioned (D)RL voltage control schemes is their inability to cope with the curse of dimensionality in action space. Moreover, *joint control* of both utility-owned devices and emerging power inverters has not been fully investigated. In addition, the discrete variables describing the on-off operation of capacitors and slow timescale associated with changing capacitor statuses, compared with those of fast-responding inverters further challenges voltage regulation. As a consequence, current capacitor decisions have a long-standing influence on future inverter setpoints. The other way around, current inverter setpoints also affect future commitment of capacitors through the aggregate cost. Indeed, this two-way long-term interaction is difficult to model and cope with.

In this context, voltage control is dealt with in the present paper using shunt capacitors and smart inverters. Preliminary results were presented in [28]. A novel two-timescale solution combining first principles based on physical models and data-driven advances is put forth. On the slow timescale (e.g., hourly or daily basis), the optimal configuration (corresponding to the discrete on-off commitment) of capacitors is formulated as a Markov decision process, by carefully defining state, action, and cost according to the available control variables in the grid. The solution of this MDP is approached by means of a DRL algorithm. This framework leverages the merits of the so-termed *target network* and *experience replay*, which can remove the correlation among the sequence of observations, to make the DRL stable and tractable. On the other hand, the setpoints of the inverters’ reactive power output, are computed by minimizing the instantaneous voltage deviation using the exact or approximate grid models on the fast timescale (e.g., every few seconds).

Compared with past works, our contributions can be summarized as follows.

- c1)** *Joint control of two types of assets.* A hybrid data- and physics-driven approach to managing both utility-owned equipment as well as smart inverters;
- c2)** *Slow-timescale learning.* Modeling demand and genera-

tion as Markovian processes, optimal capacitor settings are learned from data using DRL;

- c3)** *Fast-timescale optimization.* Using exact or approximate grid models, the optimal setpoints for inverters are found relying on the most recent slow-timescale solution; and,
- c4)** *Curse of dimensionality in action space.* Introducing hyper deep Q -network to handle the curse of dimensionality emerging due to large number of capacitors.

II. VOLTAGE CONTROL IN TWO TIMESCALES

In this section, we describe the system model, and formulate the two-timescale voltage regulation problem.

A. System model

Consider a distribution grid of $N + 1$ buses rooted at the substation bus indexed by $i = 0$, whose buses are collected into $\mathcal{N}_0 := \{0\} \cup \mathcal{N}$, and lines into $\mathcal{L} := \{1, \dots, N\}$. For all $i \in \mathcal{N}$ (i.e., without substation bus), let v_i denote their squared voltage magnitude, and $p_i + jq_i$ their complex power injected. For brevity, collect all nodal quantities into column vectors $\mathbf{v}, \mathbf{p}, \mathbf{q}$. Active power injection is split into its generation p_i^g and consumption p_i^c as $p_i := p_i^g - p_i^c$; likewise, reactive power injection is $q_i := q_i^g - q_i^c$. In distribution grids, it holds that $p_i^g = p_i^c = q_i^c = 0$ and $q_i^g > 0$ if bus i has a capacitor; while $p_i^g = q_i^g = 0$ if bus i is a purely load bus; and $p_i^g \geq 0$, $q_i^c \geq 0$, $p_i^c \geq 0$ if bus i is equipped with a DG. Let us stack generation and consumption components into vectors $\mathbf{p}^g, \mathbf{q}^g, \mathbf{p}^c$, and $\mathbf{q}^c$ accordingly. Predictions of active power consumption and solar generation $(\mathbf{p}^c, \mathbf{q}^c, \mathbf{p}^g)$ can be obtained through the hourly and real-time market (see e.g., [10]), or by running load demand (solar generation) prediction algorithms [29].

As mentioned earlier, there are two types of assets in modern distribution grids that can be engaged in reactive power control; that is, utility-owned equipment featuring discrete actions and limited lifespan, as well as smart inverters controllable within seconds and in a continuously-valued fashion. As the aggregate load varies in a relatively slow way, traditional devices have been sufficient for providing voltage support; while fast-responding solutions using inverters become indispensable with the increase of uncertain renewable penetration. In this context, the present work focuses on voltage regulation by capitalizing on the reactive control capabilities of both capacitors and inverters, while our framework can also account for other reactive power control devices. To this end, we divide every day into $N_{\bar{T}}$ intervals indexed by $\tau = 1, \dots, N_{\bar{T}}$. Each of these $N_{\bar{T}}$ intervals is further partitioned into N_T time slots which are indexed by $t = 1, \dots, N_T$, as illustrated in Fig. 1. 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 $\mathbf{p}^g(\tau, t)$, $\mathbf{p}^c(\tau, t)$, and $\mathbf{q}^c(\tau, t)$ remain the same within each t -slot, but may change from slot t to $t + 1$.

Suppose there are N_a shunt capacitors installed in the grid, whose bus indices are collected in $\mathcal{N}_a$, and are in one-to-one

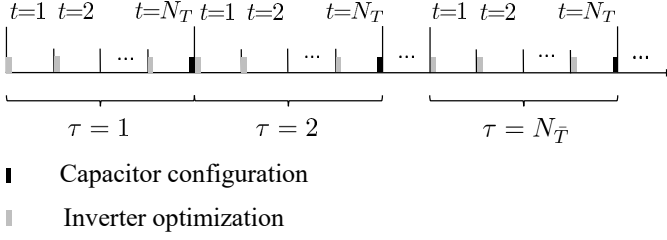


Fig. 1: Two-timescale partitioning of a day for joint capacitor and inverter control.

correspondence with entries of $\mathcal{K} := \{1, \dots, N_a\}$ (a simple renumbering). Assume that every bus is equipped with either a shunt capacitor or a smart inverter, but not both. The remaining buses, after removing entries in $\mathcal{N}_a$ from $\mathcal{N}$, collected in $\mathcal{N}_r$, are assumed equipped with inverters. This assumption is made without loss of generality as one can simply set the upper and lower bounds on the reactive output to zero at buses having no inverters installed.

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

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

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 in 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 $\hat{\mathbf{y}}(\tau)$.

On the other hand, the reactive power $q_{r, i}^g(\tau, t)$ generated by inverter i is adjusted on the fast timescale (every t), and it is constrained by $|q_{r, i}^g(\tau, t)| \leq \sqrt{(\bar{s}_i)^2 - (p_i^g(\tau, t))^2}$, where $\bar{s}_i$ is the power capability of inverter i . Traditionally, inverter i is designed as $\bar{s}_i = \bar{p}_i^g$, where $\bar{p}_i^g$ is the active power capacity of the renewable generation unit installed at bus i . However, when maximum output is reached, i.e., $p_i^g(\tau, t) = \bar{p}_i^g$, no reactive power can be provided. To address this, oversized inverters' nameplate capacity has been advocated such that $\bar{s}_i > \bar{p}_i^g$ [10]. For instance, choosing $\bar{s}_i = 1.08\bar{p}_i^g$ and limiting $q_{r, i}^g(\tau, t)$ to $\sqrt{(\bar{s}_i)^2 - (\bar{p}_i^g)^2}$ instead of $\sqrt{(\bar{s}_i)^2 - (p_i^g(\tau, t))^2}$, the reactive power compensation provided by inverter i is $|q_{r, i}^g(\tau, t)| \leq 0.4\bar{p}_i^g$, regardless of the instantaneous PV output $p_i^g(\tau, t)$ [10]. As such, $q_{r, i}^g(\tau, t)$ generated by inverter i is constrained as

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

B. Two-timescale voltage regulation formulation

Given two-timescale load consumption and generation that we model as Markovian processes [30], the task of voltage regulation is to find the optimal reactive power support per slot by configuring capacitors in every interval and adjusting inverter outputs in every slot, such that the *long-term* average voltage deviation is minimized. As voltage magnitudes $\mathbf{v}(\tau, t)$ depend solely on the control variables $\mathbf{q}^g(\tau, t)$, they are expressed as implicit functions of $\mathbf{q}^g(\tau, t)$, yielding $\mathbf{v}_{\tau, t}(\mathbf{q}^g(\tau, t))$, whose

actual function forms for postulated grid models will be given Section III. The novel two-timescale voltage control scheme entails solving the following stochastic optimization problem

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

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

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

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

for some discount factor $\gamma \in (0, 1)$, where the expectation is taken over the joint distribution of $(\mathbf{p}^c(\tau, t), \mathbf{q}^c(\tau, t), \mathbf{p}^g(\tau, t))$ across all intervals and slots. Clearly, the optimization problem (3) involves infinitely many variables $\{\mathbf{q}^g(\tau, t)\}$ and $\{\hat{\mathbf{y}}(\tau)\}$, which are coupled across time via the cost function and the constraint (3b). Moreover, discrete variables $\hat{\mathbf{y}}(\tau) \in \{0, 1\}^{N_a}$ render problem (3) nonconvex and generally *NP-hard*. Last but not least, it is a multi-stage optimization, whose decisions are not all made at the same stage, and must also account for the power variability during real-time operation. In words, tackling (3) exactly is challenging.

Instead, our goal is to design algorithms that sequentially observe predictions $\{(\mathbf{p}^c(\tau, t), \mathbf{q}^c(\tau, t), \mathbf{q}^g(\tau, t))\}$, and solve near optimally problem (3). The assumption is that, although no distributional knowledge of those stochastic processes involved is given, their realizations can be made available in real time, by means of e.g., accurate forecasting methods [29]. In this sense, the physics governing the electric power system will be utilized together with data to solve (3) in real time. Specifically, on the slow timescale, say at the end of each interval $\tau - 1$, the optimal on-off capacitor decisions $\hat{\mathbf{y}}(\tau)$ will be set through a DRL algorithm that can learn from the predictions collected within the current interval $\tau - 1$; while, on the fast timescale, namely at the beginning of each slot t within interval τ , our two-stage control scheme will compute the optimal setpoints for inverters, by minimizing the instantaneous bus voltage deviations while respecting physical constraints, given the current on-off commitment of capacitor units $\hat{\mathbf{y}}(\tau)$ found at the very end of interval $(\tau - 1)$. These two timescales are detailed in Sections III and IV, respectively.

III. FAST-TIMESCALE OPTIMIZATION OF INVERTERS

As alluded earlier, the actual forms of $\mathbf{v}_{\tau, t}(\mathbf{q}^g(\tau, t))$ will be specified in this section, relying on the exact AC model or a linearized approximant of it. Leveraging convex relaxation to deal with the nonconvexity, the considered AC model yields a second-order cone program (SOCP), whereas the linearized one leads to a linearly constrained quadratic program. In contrast, the latter offers an approximate yet computationally more affordable alternative to the former. Selecting between these two models relies on affordable computational capabilities.

A. Branch flow model

Due to the radial structure of distribution grids, every non-root bus $i \in \mathcal{N}$ has a unique parent bus termed π_i . The two are joined through the i -th distribution line represented by

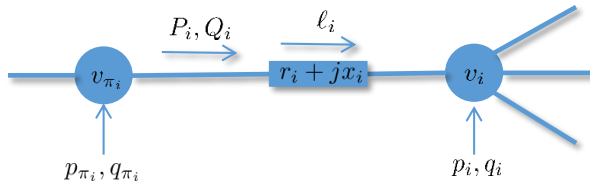


Fig. 2: Bus i is connected to its unique parent π_i via line i .

$(\pi_i, i) \in \mathcal{L}$ having impedance $r_i + jx_i$. Let $P_i(\tau, t) + jQ_i(\tau, t)$ stand for the complex power flowing from buses π_i to i seen at the ‘front’ end at time slot t of interval τ , as depicted in Fig. 2. Throughout this section, the interval index τ will be dropped when it is clear from the context.

With further ℓ_i denoting the squared current magnitude on line $i \in \mathcal{L}$, the celebrated *branch flow model* is described by the following equations for all buses $i \in \mathcal{N}$, and for all t within every interval τ [31], [32]

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

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

$$v_i(t) = v_{\pi_i}(t) - 2(r_i P_i(t) + x_i Q_i(t)) + (r_i^2 + x_i^2) \ell_i(t) \quad (4c)$$

$$\ell_i(t) = \frac{P_i^2(t) + Q_i^2(t)}{v_{\pi_i}(t)} \quad (4d)$$

where we have ignored the dependence on τ for brevity, and χ_i denotes the set of all children buses for bus i .

Clearly, the set of equations in (4d) is quadratic in $P_i(t)$ and $Q_i(t)$, yielding a nonconvex set. To address this challenge, consider relaxing the equalities (4d) into inequalities (a.k.a. hyperbolic relaxation, see e.g., [11])

$$P_i^2(t) + Q_i^2(t) \leq v_{\pi_i}(t) \ell_i(t), \quad \forall i \in \mathcal{N}, t \quad (5)$$

which can be equivalently rewritten as the following second-order cone constraints

$$\left\| \begin{pmatrix} 2P_i(t) \\ 2Q_i(t) \\ \ell_i(t) - v_{\pi_i}(t) \end{pmatrix} \right\| \leq v_{\pi_i}(t) + \ell_i(t), \quad \forall i \in \mathcal{N}. \quad (6)$$

Equations (4a)-(4c) and (6) now define a convex feasible set. The procedure of leveraging this relaxed set (instead of the nonconvex one) is known as SOCP relaxation [32]. Interestingly, it has been shown that under certain conditions, SOCP relaxation is exact in the sense that the set of inequalities (6) holds with equalities at the optimum [33].

Given the capacitor configuration $\hat{\mathbf{y}}(\tau)$ found at the end of the last interval $\tau - 1$, under the aforementioned relaxed grid model, the voltage regulation on the fast timescale based on the exact AC model can be described as follows

$$\begin{aligned} & \underset{\mathbf{v}(t), \mathbf{q}_r^g(t), \mathbf{P}(t), \mathbf{Q}(t)}{\text{minimize}} && \|\mathbf{v}(t) - v_0 \mathbf{1}\|^2 \\ & \text{subject to} && (4a) - (4d) \end{aligned} \quad (7a)$$

$$q_i^g(t) = \hat{y}_{k_i}(\tau) q_{a,k_i}^g, \quad \forall i \in \mathcal{N}_a \quad (7b)$$

$$q_i^g(t) = q_{r,i}^g(t), \quad \forall i \in \mathcal{N}_r \quad (7c)$$

$$|q_{r,i}^g(t)| \leq \bar{q}_i^g, \quad \forall i \in \mathcal{N}_r \quad (7d)$$

which is readily a convex SOCP and can be efficiently solved by off-the-shelf convex programming toolboxes. The optimal setpoints of smart inverters for the exact AC model are found as the q_r^g -minimizer of (7).

However, solving SOCPs could be computationally demanding when dealing with relatively large-scale distribution grids, say of several hundred buses. Trading off modeling accuracy for computational efficiency, our next instantiation of the fast-timescale voltage control relies on an approximate grid model.

B. Linearized power flow model

As line current magnitudes $\{\ell_i\}$ are relatively small compared to line flows, the last term in (4a)-(4c) can be ignored yielding the next set of linear equations for all i, t [34]

$$p_i(t) = \sum_{j \in \chi_i} P_j(t) - P_i(t) \quad (8a)$$

$$q_i(t) = \sum_{j \in \chi_i} Q_j(t) - Q_i(t) \quad (8b)$$

$$v_i(t) = v_{\pi_i}(t) - 2(r_i P_i(t) + x_i Q_i(t)) \quad (8c)$$

which is known as the linearized distribution flow model. In this fashion, all squared voltage magnitudes $\mathbf{v}(t)$ can be expressed as linear functions of $\mathbf{q}^g(t)$.

Adopting the approximate model (8), the optimal setpoints of inverters can be found by solving the following optimization problem per slot t in interval τ , provided $\hat{\mathbf{y}}(\tau)$ is available from the last interval on the slow timescale

$$\begin{aligned} & \underset{\mathbf{v}(t), \mathbf{q}_r^g(t), \mathbf{P}(t), \mathbf{Q}(t)}{\text{minimize}} && \|\mathbf{v}(t) - v_0 \mathbf{1}\|^2 \\ & \text{subject to} && (8a) - (8c) \end{aligned} \quad (9a)$$

$$q_i^g(t) = \hat{y}_{k_i}(\tau) q_{a,k_i}^g, \quad \forall i \in \mathcal{N}_a \quad (9b)$$

$$q_i^g(t) = q_{r,i}^g(t), \quad \forall i \in \mathcal{N}_r \quad (9c)$$

$$|q_{r,i}^g(t)| \leq \bar{q}_i^g, \quad \forall i \in \mathcal{N}_r. \quad (9d)$$

As all constraints are linear and the cost is quadratic, (9) constitutes a standard convex quadratic program. As such, it can be solved efficiently by e.g., primal-dual algorithms, or off-the-shelf convex programming solvers, whose implementation details are skipped due to space limitations.

IV. SLOW-TIMESCALE CAPACITOR RECONFIGURATION

Here we deal with reconfiguration of shunt capacitors on the slow timescale. This amounts to determining their on-off status for the ensuing interval. Past approaches to solving the resultant integer-valued optimization were heuristic, or, relied on semidefinite programming relaxation. They do not guarantee optimality, while they also incur high computational and storage complexities. We take a different route by drawing from advances in artificial intelligence, to develop data-driven solutions that could near optimally learn, track, as well as adapt to unknown generation and consumption dynamics.

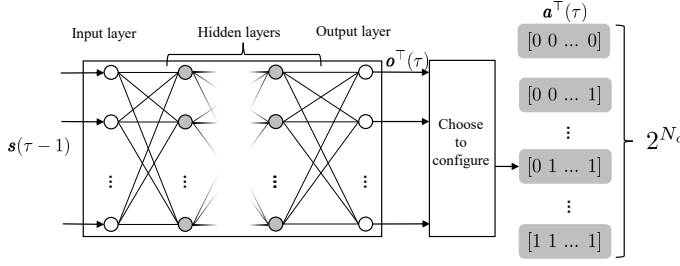


Fig. 3: Deep Q-network

A. A data-driven solution

Clearly from (7b)–(9b), the capacitor decisions $\hat{\mathbf{y}}(\tau)$ made at the end of interval $\tau - 1$ (slow-timescale learning) influence inverters' setpoints during the entire interval τ (fast-timescale optimization). The other way around, inverters' regulation on voltages influences the capacitor commitment for the next interval. This two-way between the capacitor configuration and the optimal setpoints of inverters motivates our RL formulation. Dealing with learning policy functions in an environment with action-dependent dynamically evolving states and costs, RL seeks a policy function (of states) to draw actions from, in order to minimize the average cumulative cost [17].

Modeling load demand and renewable generation as Markovian processes, the optimal configuration of capacitors can be formulated as an MDP, which can be efficiently solved through RL algorithms. 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 $\mathbf{s} \in \mathcal{S}$ and $\mathbf{a} \in \mathcal{A}$, $c = (c(\mathbf{s}, \mathbf{a}))_{\mathbf{s} \in \mathcal{S}, \mathbf{a} \in \mathcal{A}}$ are the real-valued instantaneous costs after the system operator takes an action $\mathbf{a}$ at state $\mathbf{s}$; 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 , giving rise to an action vector $\mathbf{a}(\tau) = \mathbf{y}(\tau)$ per interval τ . The set of binary action vectors constitutes the action space $\mathcal{A}$, whose cardinality is exponential in the number of capacitors, meaning $|\mathcal{A}| = 2^{N_a}$.

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

The action is decided according to the configuration policy π that is a function of the most recent state $\mathbf{s}(\tau - 1)$, given as

$$\mathbf{a}(\tau) = \pi(\mathbf{s}(\tau - 1)). \quad (10)$$

Cost function c . The cost on the slow timescale is

$$c(\mathbf{s}(\tau - 1), \mathbf{a}(\tau)) = \sum_{t=1}^{N_T} \|\mathbf{v}_{\tau,t}(\mathbf{q}^g(\tau, t)) - \mathbf{v}_0\|^2. \quad (11)$$

Set of transition probability matrices $\mathcal{P}$. While being at a state $\mathbf{s} \in \mathcal{S}$ upon taking an action $\mathbf{a}$, the system moves to a new state $\mathbf{s}' \in \mathcal{S}$ probabilistically. Let $P_{\mathbf{s}\mathbf{s}'}^{\mathbf{a}}$ denote the transition

probability matrix from state $\mathbf{s}$ to the next state $\mathbf{s}'$ under a given action $\mathbf{a}$. Evidently, it holds that $\mathcal{P} := \{P_{\mathbf{s}\mathbf{s}'}^{\mathbf{a}} | \forall \mathbf{a} \in \mathcal{A}\}$.

Discount factor γ . 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 so-termed action-value function under the control policy π is defined as

$$Q_{\pi}(\mathbf{s}(\tau - 1), \mathbf{a}(\tau)) := \mathbb{E} \left[\sum_{\tau'=\tau}^{\infty} \gamma^{\tau'-\tau} c(\mathbf{s}(\tau' - 1), \mathbf{a}(\tau')) \middle| \pi, \mathbf{s}(\tau - 1), \mathbf{a}(\tau) \right] \quad (12)$$

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

$$Q_{\pi^*}(\mathbf{s}, \mathbf{a}) = \mathbb{E}[c(\mathbf{s}, \mathbf{a})] + \gamma \sum_{\mathbf{s}' \in \mathcal{S}} P_{\mathbf{s}\mathbf{s}'}^{\mathbf{a}} \min_{\mathbf{a}' \in \mathcal{A}} Q_{\pi^*}(\mathbf{s}', \mathbf{a}'). \quad (13)$$

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

$$\pi^*(\mathbf{s}) = \arg \min_{\mathbf{a}} Q_{\pi^*}(\mathbf{s}, \mathbf{a}). \quad (14)$$

It is clear from (13) that if all transition probabilities $\{P_{\mathbf{s}\mathbf{s}'}^{\mathbf{a}}\}$ were available, we can derive $Q_{\pi^*}(\mathbf{s}, \mathbf{a})$, and subsequently the optimal policy π^* from (14). Nonetheless, obtaining those transition probabilities is impractical in practical distribution systems. This calls for approaches that aim directly at π^* , without assuming any knowledge of $\{P_{\mathbf{s}\mathbf{s}'}^{\mathbf{a}}\}$.

One celebrated approach of this kind is Q-learning, which can learn π^* by approximating $Q_{\pi^*}(\mathbf{s}, \mathbf{a})$ ‘on-the-fly’ [17, p. 107]. Due to its high-dimensional continuous state space $\mathcal{S}$ however, Q-learning is not applicable for the problem at hand. This motivates function approximation based Q-learning schemes that can deal with continuous state domains.

B. A deep reinforcement learning approach

DQN offers a NN function approximator of the Q-function, chosen to be e.g., a fully connected feed-forward NN, or a convolutional NN, depending on the application [18]. It takes as input the state vector, to generate at its output Q-values for all possible actions (one for each). As demonstrated in [18], such a NN indeed enables learning the Q-values of all state-action pairs, from just a few observations obtained by interacting with the environment. Hence, it effectively addresses the challenge brought by the ‘curse of dimensionality’ [18]. Inspired by this, we employ a feed-forward NN to approximate the Q-function in our setting. Specifically, our DNN consists of L fully connected hidden layers with ReLU activation functions, depicted in Fig. 3. At the input layer, each neuron is fed with one entry of the state vector $\mathbf{s}(\tau - 1)$, which, after passing through L ReLU layers, outputs a vector $\mathbf{o}(\tau) \in \mathbb{R}^{2^{N_a}}$, whose elements predict the Q-values for all possible actions (i.e., capacitor configurations). Since each

output unit corresponds to a particular configuration of all N_a capacitors, there is a total of 2^{N_a} neurons at the output layer. For ease of exposition, let us collect all weight parameters of this DQN into a vector θ which parameterizes the input-output relationship as $\mathbf{o}(\tau) = Q_\pi(\mathbf{s}(\tau-1), \mathbf{a}(\tau); \theta)$ (c.f. (12)). At the end of a given interval $\tau-1$, upon passing the state vector $\mathbf{s}(\tau-1)$ through this DQN, the corresponding predicted Q -values $\mathbf{o}(\tau)$ for all possible actions become available at the output. Based on these predicted values, the system operator selects the action having the smallest predicted Q -value to be in effect over the next interval.

Intuitively, the weights θ should be chosen such that the DQN outputs match well the actual Q -values with input any state vector. Toward this objective, the popular stochastic gradient descent (SGD) method is employed to update θ ‘on the fly’ [18]. At the end of a given interval τ , precisely when i) the system operator has made decision $\mathbf{a}(\tau)$, ii) the grid has completed the transition from the state $\mathbf{s}(\tau-1)$ to a new state $\mathbf{s}(\tau)$, and, iii) the network has incurred and revealed cost $c(\mathbf{s}(\tau-1), \mathbf{a}(\tau))$, we perform a SGD update based on the current estimate θ_τ to yield $\theta_{\tau+1}$. The so-termed temporal-difference learning [17] confirms that a sample approximation of the optimal cost-to-go from interval τ is given by $c(\mathbf{s}(\tau-1), \mathbf{a}(\tau)) + \gamma \min_{\mathbf{a}' \in \mathcal{A}} Q_\pi(\mathbf{s}(\tau), \mathbf{a}'; \theta_\tau)$, where $c(\mathbf{s}(\tau-1), \mathbf{a}(\tau))$ is the instantaneous cost observed, and $\min_{\mathbf{a}'} Q_\pi(\mathbf{s}(\tau), \mathbf{a}'; \theta_\tau)$ represents the smallest possible predicted cost-to-go from state $\mathbf{s}(\tau)$, which can be computed through our DQN with weights θ_τ , and is discounted by factor γ . In words, the target value $c(\mathbf{s}(\tau-1), \mathbf{a}(\tau)) + \gamma \min_{\mathbf{a}' \in \mathcal{A}} Q_\pi(\mathbf{s}(\tau), \mathbf{a}'; \theta_\tau)$ is readily available at the end of interval $\tau-1$. Adopting the ℓ_2 -norm error criterion, a meaningful approach to tuning the weights θ entails minimizing the following loss function

$$\mathcal{L}(\theta) := \left[c(\mathbf{s}(\tau-1), \mathbf{a}(\tau)) + \gamma \min_{\mathbf{a}' \in \mathcal{A}} Q_\pi(\mathbf{s}(\tau), \mathbf{a}'; \theta_\tau) - Q_\pi(\mathbf{s}(\tau-1), \mathbf{a}(\tau); \theta) \right]^2 \quad (15)$$

for which the SGD update is given by

$$\theta_{\tau+1} = \theta_\tau - \beta_\tau \nabla \mathcal{L}(\theta)|_{\theta_\tau} \quad (16)$$

where $\beta_\tau > 0$ is a preselected learning rate, and $\nabla \mathcal{L}(\theta)$ denotes the (sub-)gradient.

However, due to the compositional structure of DNNs, the update (16) does not work well in practice. In fact, the resultant DQN oftentimes does not provide a stable result; see e.g., [35]. To bypass these hurdles, several modifications have been introduced. In this work, we adopt the *target network* and *experience replay* [18]. To this aim, let us define an experience $e(\tau') := (\mathbf{s}(\tau'-1), \mathbf{a}(\tau'), c(\mathbf{s}(\tau'-1), \mathbf{a}(\tau')), \mathbf{s}(\tau'))$, to be a tuple of state, action, cost, and the next state. Consider also having a replay buffer $\mathcal{R}(\tau)$ on-the-fly, which stores the most recent $R > 0$ experiences visited by the agent. For instance, the replay buffer at any interval $\tau \geq R$ is $\mathcal{R}(\tau) := \{e(\tau-R+1), \dots, e(\tau)\}$. Furthermore, as another effective remedy to stabilizing the DQN updates, we replicate the DQN to create a second DNN, commonly referred to as the *target network*, whose weight parameters are concatenated in the vector θ^{Tar} . It is worth highlighting that this target network

Algorithm 1 Two-timescale voltage regulation scheme.

- 1: **Initialize:** θ_0 randomly; weight of the target network $\theta_0^{\text{Tar}} = \theta_0$; replay buffer $\mathcal{R}$; and the initial state $\mathbf{s}(0)$.
 - 2: **for** $\tau = 1, 2, \dots$ **do**
 - 3: Take action $\mathbf{a}(\tau)$ through exploration-exploitation

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

where $\epsilon_\tau = \max\{1 - 0.1 \times \lfloor \tau/50 \rfloor, 0\}$.
 - 4: Evaluate $c(\mathbf{s}(\tau-1), \mathbf{a}(\tau))$ using (11).
 - 5: **for** $t = 1, 2, \dots, N_T$ **do**
 - 6: Compute $q^g(\tau, t)$ using (7) or (9).
 - 7: **end for**
 - 8: Update $\mathbf{s}(\tau)$.
 - 9: Save $(\mathbf{s}(\tau-1), \mathbf{a}(\tau), c(\mathbf{s}(\tau-1), \mathbf{a}(\tau)), \mathbf{s}(\tau))$ into $\mathcal{R}(\tau)$.
 - 10: Randomly sample M_τ experiences from $\mathcal{R}(\tau)$.
 - 11: Form the mini-batch loss $\mathcal{L}^{\text{Tar}}(\theta_\tau; \mathcal{M}_\tau)$ using (19).
 - 12: Update $\theta_{\tau+1}$ using (20).
 - 13: **if** $\text{mod}(\tau, B) = 0$ **then**
 - 14: Update the target network $\theta_\tau^{\text{Tar}} = \theta_\tau$.
 - 15: **end if**
 - 16: **end for**
-

is not trained, but its parameters θ^{Tar} are only periodically reset to estimates of θ , say every B training iterations of the DQN. Consider now the temporal-difference loss for some randomly drawn experience $e(\tau')$ from $\mathcal{R}(\tau)$ at interval τ

$$\begin{aligned} \mathcal{L}^{\text{Tar}}(\theta_\tau; e(\tau')) &:= \frac{1}{2} \left[c(\mathbf{s}(\tau'-1), \mathbf{a}(\tau')) \right. \\ &\quad \left. + \gamma \min_{\mathbf{a}'} Q^{\text{Tar}}(\mathbf{s}(\tau), \mathbf{a}'; \theta_\tau^{\text{Tar}}) - Q(\mathbf{s}(\tau'-1), \mathbf{a}(\tau'); \theta_\tau) \right]^2. \end{aligned} \quad (17)$$

Upon taking expectation with respect to all sources of randomness generating this experience, we arrive at

$$\mathcal{L}^{\text{Tar}}(\theta_\tau; \mathcal{R}(\tau)) := \mathbb{E}_{e(\tau')} \mathcal{L}^{\text{Tar}}(\theta_\tau; e(\tau')). \quad (18)$$

In practice however, the underlying transition probabilities are unknown, which challenges evaluating and hence minimizing $\mathcal{L}^{\text{Tar}}(\theta_\tau; \mathcal{R}(\tau))$ exactly. A commonly adopted alternative is to approximate the expected loss with an empirical loss over a few samples (that is, experiences here). To this end, we draw a mini-batch of M_τ experiences uniformly at random from the replay buffer $\mathcal{R}(\tau)$, whose indices are collected in the set $\mathcal{M}_\tau$, i.e., $\{e(\tau')\}_{\tau' \in \mathcal{M}_\tau} \sim U(\mathcal{R}(\tau))$. Upon computing for each of those sampled experiences an output using the target network with parameters θ_τ^{Tar} , the empirical loss is

$$\begin{aligned} \mathcal{L}^{\text{Tar}}(\theta_\tau; \mathcal{M}_\tau) &:= \frac{1}{2M_\tau} \sum_{\tau' \in \mathcal{M}_\tau} \left[c(\mathbf{s}(\tau'-1), \mathbf{a}(\tau')) \right. \\ &\quad \left. + \gamma \min_{\mathbf{a}'} Q^{\text{Tar}}(\mathbf{s}(\tau'), \mathbf{a}'; \theta_\tau^{\text{Tar}}) - Q(\mathbf{s}(\tau'-1), \mathbf{a}(\tau'); \theta_\tau) \right]^2. \end{aligned} \quad (19)$$

In a nutshell, the weight parameter vector θ_τ of the DQN is efficiently updated ‘on-the-fly’ using SGD over the empirical loss $\mathcal{L}^{\text{Tar}}(\theta_\tau; \mathcal{M}_\tau)$, with iterates given by

$$\theta_{\tau+1} = \theta_\tau - \beta_\tau \nabla \mathcal{L}^{\text{Tar}}(\theta_\tau; \mathcal{M}_\tau). \quad (20)$$

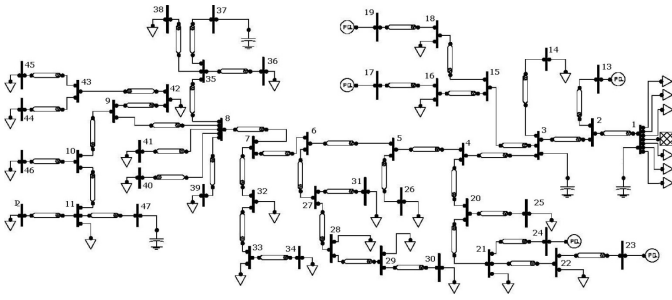


Fig. 4: Schematic diagram of the 47-bus industrial distribution feeder. Bus 1 is the substation, and the 6 loads connected to it model other feeders on this substation.

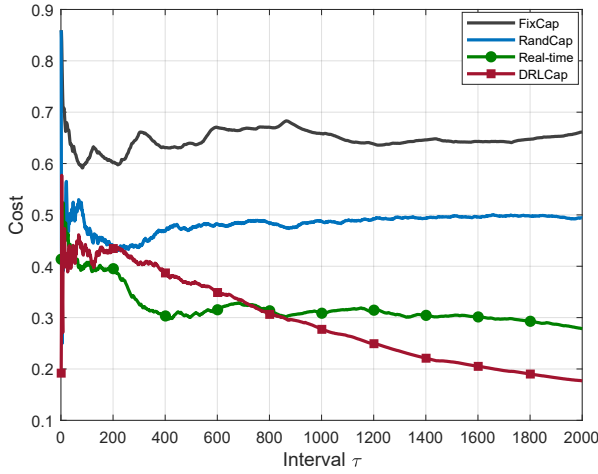


Fig. 5: Time-averaged instantaneous costs incurred by the four voltage control schemes.

Incorporating *target network* and *experience replay* remedies for stable DRL, our proposed two-timescale voltage regulation scheme is summarized in Alg. 1.

V. NUMERICAL TESTS

In this section, numerical tests on a real-world 47-bus distribution feeder as well as the IEEE 123-bus benchmark system are provided to showcase the performance of our proposed DRL-based voltage control scheme (cf. presented in Alg. 1). As has already been shown in previous works (e.g., [10], [13], [32]), the linearized distribution flow model approximates the exact AC model very well; hence, numerical results based on the linearized model were only reported here.

The first experiment entails the Southern California Edison 47-bus distribution feeder [11], which is depicted in Fig. 4. This feeder is integrated with four shunt capacitors as well as five smart inverters. As the voltage magnitude v_0 of the substation bus is regulated to be a constant (1 in all our tests) through a voltage transformer, the capacitor at the substation was excluded from our control. Thus, a total of three shunt capacitors along with five smart inverters embedded with large PV plants were engaged in voltage regulation. The rest three capacitors are installed on buses 3, 37, and 47, with capacities 120, 180, and 180 kVar, respectively, while the five large

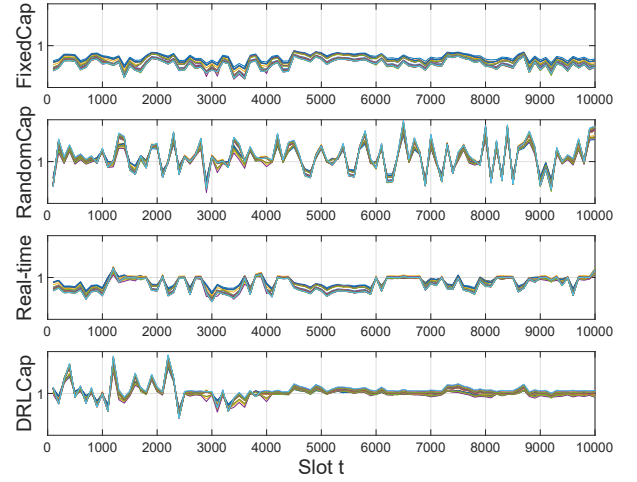


Fig. 6: Voltage magnitude profiles obtained by the four voltage control schemes over the simulation period of 10,000 slots.

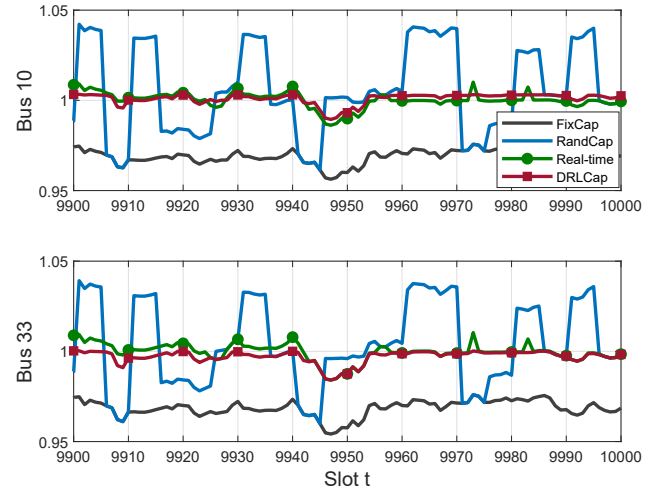


Fig. 7: Voltage magnitude profiles obtained by the four voltage control schemes at buses 10 and 33 from slot 9,900 to 10,000.

PV plants are located on buses 2, 16, 18, 21, and 22, with capacities 300, 80, 300, 400, and 200 kW, respectively. To test our scheme in a realistic setting, real consumption as well as solar generation data were obtained from the Smart* project collected on August 24, 2011 [36], which were first preprocessed by following the procedure described in our precursor work [10].

In our tests, to match the availability of real data, each slot t was set to a minute, and each interval τ was set to five minutes. A power factor of 0.8 was assumed for all loads. The DQN used here consists of three fully connected layers, which has 44 and 12 units in the first and second hidden layers, respectively. Although simple, it was found sufficient for the task at hand. ReLU activation functions ($\sigma(x) = \max(x, 0)$) were employed in the hidden layers, and logistic sigmoid functions $s(x) = 1/(1 + e^{-x})$ were used at the output layer.

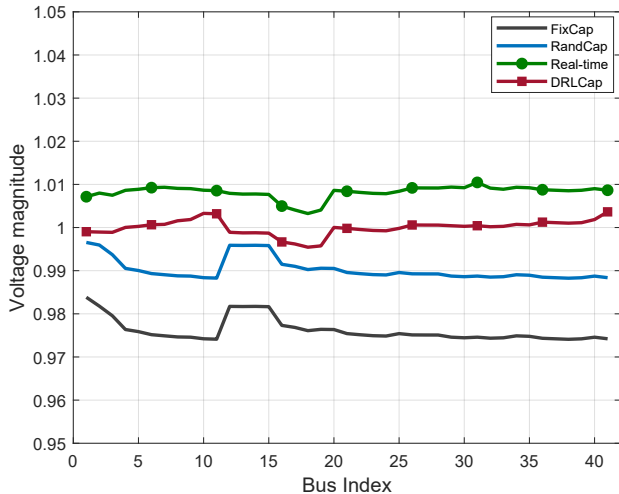


Fig. 8: Voltage magnitude profiles at all buses at slot 9,900 obtained by the four voltage control schemes.

To assess the performance of our proposed scheme, we have simulated three capacitor configuration policies as baselines, that include a fixed capacitor configuration (FixCap), a random capacitor configuration (RandCap), and an (impractical) ‘real-time’ policy. Specifically, the FixCap uses a fixed capacitor configuration throughout, and the RandCap implements random actions to configure the capacitors on every slow time interval; both of which compute the inverter setpoints by solving (9) per slot t . The impractical Real-time scheme however, optimizes over inverters and capacitors on a single-timescale, namely at every slot – hence justifying its ‘real-time’ characterization. To carry out this optimization task, first the binary constraints $y_{k_i}(t) \in \{0, 1\}$ are relaxed to box ones $y_{k_i}(t) \in [0, 1]$, the resulting convex program is solved using an off-the-shelf routine [37], which is followed by a standard rounding step to recover binary solutions for capacitor configurations [38].

In the first experiment, the DRL-based capacitor configuration (DRLCap) voltage control approach was examined. The replay buffer size was set to $R = 10$, the discount factor $\gamma = 0.99$, the mini-batch size $M_\tau = 10$, and the exploration-exploitation parameter $\epsilon_\tau = \max\{1 - 0.1 \times \lfloor \tau/50 \rfloor, 0\}$. During training, the target network was updated every $B = 5$ iterations. The time-averaged instantaneous costs

$$\frac{1}{\tau} \sum_{i=1}^{\tau} c(\mathbf{s}(i-1), \mathbf{a}(i))$$

incurred by the four schemes over the first $1 \leq \tau \leq 2,000$ intervals are plotted in Fig. 5. Evidently, the proposed scheme attains a lower cost than FixCap, RandCap, and Real-time after a short period of learning and interacting with the environment. Even though the real-time scheme optimizes both capacitor configurations and inverter setpoints per slot t , its suboptimal performance in this case arises from the gap between the convexified problem and the original nonconvex counterpart. Fig. 6 presents the voltage magnitude profiles for all buses regulated by the four schemes sampled at every 100 slots.

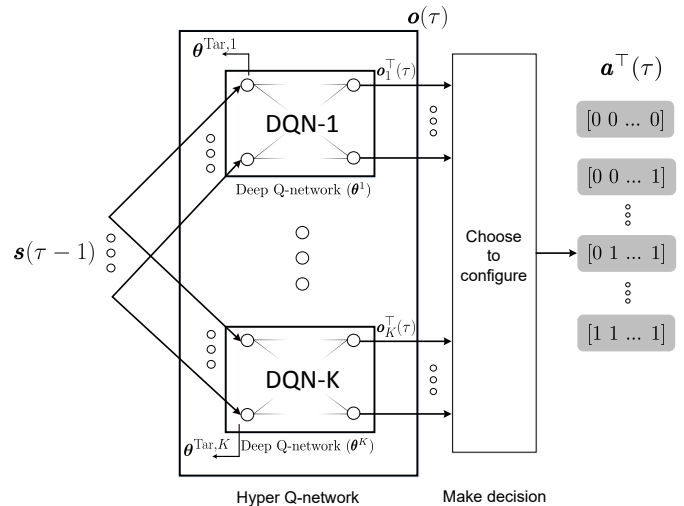


Fig. 9: Hyper deep Q -network for capacitor configuration.

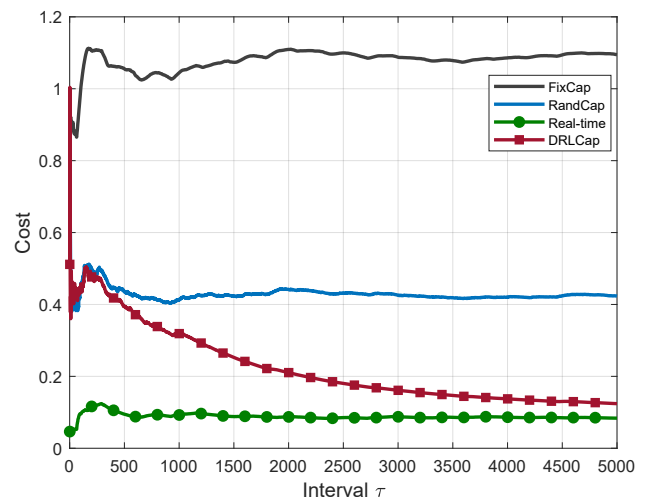


Fig. 10: Time-averaged instantaneous costs incurred by the four approaches on the IEEE 123-bus feeder.

Again, after a short period ($\sim 4,500$ slots) of training through interacting with the environment, our DRLCap voltage control scheme quickly learns a stable and (near-) optimal policy. In addition, voltage magnitude profiles regulated by FixCap, RandCap, Real-time, and DRLCap at buses 10 and 33 from slot 9,900 to 10,000 are shown in Fig. 7, while the voltage magnitude profiles at all buses at slot 9,900 are presented in Fig. 8. Curves showcase the effectiveness of our DRLCap scheme in smoothing voltage fluctuations incurred due to large solar generation as well as heavy load demand.

To deal with distribution systems having a moderately large number of capacitors, we further advocate a hyper deep Q -network implementation, that endows our DRL-based scheme with scalability. The idea here is to first split the total number 2^{N_a} of Q -value predictions $\mathbf{o}(\tau) \in \mathbb{R}^{2^{N_a}}$ at the output layer into K smaller groups, each of which is of the same size $2^{N_a}/K$ and is to be predicted by a small-size DQN. This ev-

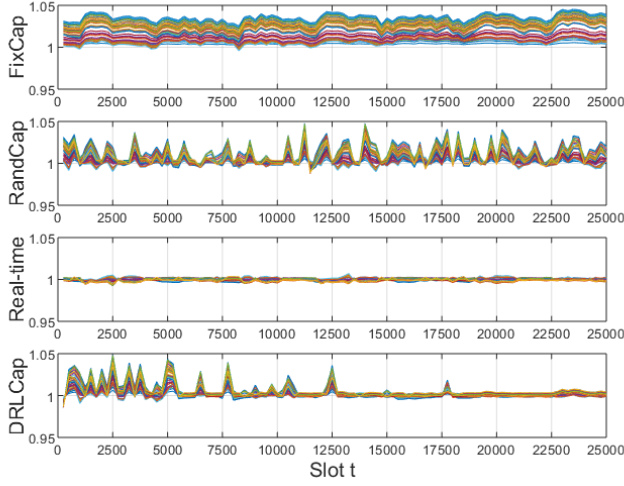


Fig. 11: Voltage magnitude profiles at all buses over the simulation period of 25,000 slots on the IEEE 123-bus feeder.

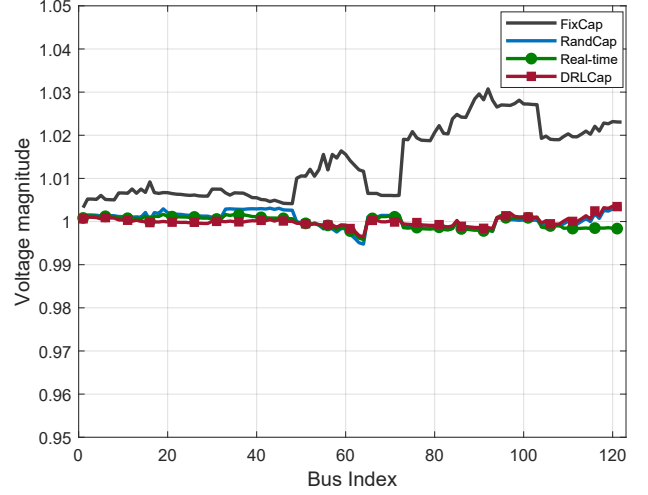


Fig. 13: Voltage magnitude profiles at all buses on slot 24,900 obtained by four approaches on the IEEE 123-bus feeder.

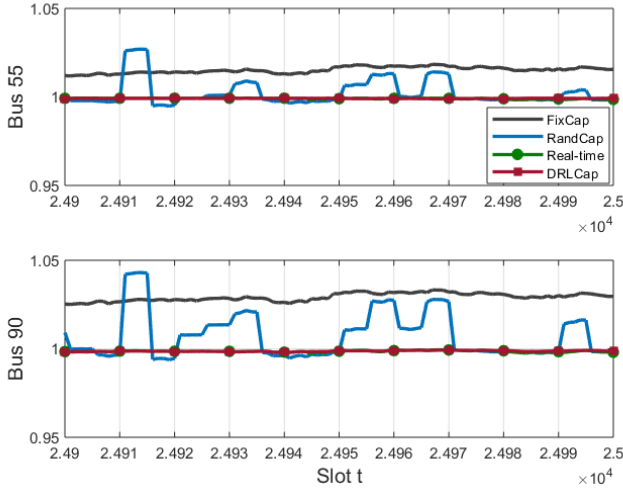


Fig. 12: Voltage magnitude profiles at buses 55 and 90 from slot 24,900 to 25,000 obtained by the four approaches on the IEEE 123-bus feeder.

idently yields the representation $\mathbf{o}(\tau) := [\mathbf{o}_1^\top(\tau), \dots, \mathbf{o}_K^\top(\tau)]^\top$, where $\mathbf{o}_k(\tau) \in \mathbb{R}^{2^{N_a}/K}$ for $k = 1, \dots, K$. By running K DQNs in parallel along with their corresponding target networks, each DQN- k generates predicted Q -values $\mathbf{o}_k(\tau)$ for the subset of actions corresponding to k th group. Note that all DQNs are fed with the same state vector $\mathbf{s}(\tau - 1)$; see also Fig. 9 for an illustration.

To examine the scalability and performance of this hyper Q -network implementation, additional tests using the IEEE 123-bus test feeder with 9 shunt capacitors were performed. Again, the capacitor at bus 1 was excluded from the control, rendering a total number of $2^8 = 256$ actions (capacitor configurations). Renewable (PV) units are located on buses 47, 49, 63, 73, 104, 108, 113, with capacities 100, 16, 70, 20, 20, 30, and 10 k, respectively. The 8 shunt capacitors are installed on buses

3, 20, 44, 93, 96, 98, 100, and 114, with capacities 50, 80, 100, 100, 100, 100, 100, and 60 kVar. In this experiment, we used a total of $K = 64$ equal-sized DQNs to form the hyper Q -network, where each DQN implemented a fully connected 3-layer feed-forward neural network, with ReLU activation functions in the hidden layers, and sigmoid functions at the output. The replay buffer size was set to $R = 50$, the batch size to $M_\tau = 8$, and the target network updating period to $B = 10$. The time-averaged instantaneous costs obtained over a simulation period of 5,000 intervals is plotted in Fig. 10. Moreover, voltage magnitude profiles of all buses over the simulation period of 25,000 slots sampled at every 100 slots under the four schemes are plotted in Fig. 11; voltage magnitude profiles at buses 55 and 90 from slot 24,900 to 25,000 are shown in Fig. 12; and, voltage magnitude profiles at all buses on slot 24,900 are depicted in Fig. 13. Evidently, the hyper deep Q -network based DRL scheme smooths out the voltage fluctuations after a certain period ($\sim 7,000$ slots) of learning, while effectively handling the curse of dimensionality in the control (action) space. Evidently from Figs. 10 and 13, both the time-averaged immediate cost as well as the voltage profiles of DRLCap converge to those of the impractical ‘real-time’ scheme (which jointly optimizes inverter setpoints and capacitor configurations per slot).

VI. CONCLUSIONS

In this work, joint control of traditional utility-owned equipment and contemporary smart inverters for voltage regulation through reactive power provision was investigated. To account for the different response times of those assets, a two-timescale approach to minimizing bus voltage deviations from their nominal values was put forth, by combining physics- and data-driven stochastic optimization. Load consumption and active power generation dynamics were modeled as MDPs. On a fast timescale, the setpoints of smart inverters were found by minimizing the instantaneous bus voltage deviations, while on a slower timescale, the capacitor banks were configured to

minimize the long-term expected voltage deviations using a deep reinforcement learning algorithm. The developed two-timescale voltage regulation scheme was found efficient and easy to implement in practice, through extensive numerical tests on real-world distribution systems using real solar and consumption data. This work also opens up several interesting directions for future research, including deep reinforcement learning for real-time optimal power flow as well as unit commitment.

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] P. Kundur, N. J. Balu, and M. G. Lauby, *Power System Stability and Control*. Duisburg, Germany: McGraw-hill New York, May 1994.
- [3] 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.
- [4] M. Bazrafshan, N. Gatsis, and H. Zhu, "Optimal tap selection of step-voltage regulators in Multi-phase distribution networks," in *Proc. of Power Syst. Comput. Conf.*, Dublin, Ireland, Jun. 11–15 2018.
- [5] D. A. Tziouvaras, P. McLaren, G. Alexander, D. Dawson, J. Esztergalyos, C. Fromen, M. Glinkowski, I. Hasenwinkle, M. Kezunovic, L. Kojovic *et al.*, "Mathematical models for current, voltage, and coupling capacitor voltage transformers," *IEEE Trans. Power Del.*, vol. 15, no. 1, pp. 62–72, Jan. 2000.
- [6] H. Xu, A. D. Domínguez-García, and P. W. Sauer, "Optimal tap setting of voltage regulation transformers using batch reinforcement learning," *arXiv:1807.10997v2*, 2018.
- [7] W. Su, J. Wang, and J. Roh, "Stochastic energy scheduling in microgrids with intermittent renewable energy resources," *IEEE Trans. Smart Grid*, vol. 5, no. 4, pp. 1876–1883, July 2014.
- [8] A. Ipakchi and F. Albuyeh, "Grid of the future," *IEEE Power Energy Mag.*, vol. 7, no. 2, pp. 52–62, Feb. 2009.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] Y. Zhang, M. Hong, E. Dall'Anese, S. V. Dhople, and Z. Xu, "Distributed controllers seeking AC optimal power flow solutions using ADMM," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 4525–4537, Sept. 2018.
- [17] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT press, 2018.
- [18] 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.
- [19] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*. Cambridge, MA, USA: MIT press, 2016.
- [20] A. Sadeghi, G. Wang, and G. B. Giannakis, "Deep reinforcement learning for adaptive caching in hierarchical content delivery networks," *IEEE Trans. Cogn. Commun. Netw.*, to appear, 2019.
- [21] R. Diao, Z. Wang, D. Shi, Q. Chang, J. Duan, and X. Zhang, "Autonomous voltage control for grid operation using deep reinforcement learning," in *Proc. of PESGM*, Atlanta, GA, Aug. 4–8, 2019, pp. 1–5.
- [22] D. Ernst, M. Glavic, and L. Wehenkel, "Power systems stability control: Reinforcement learning framework," *IEEE Trans. Power Syst.*, vol. 19, no. 1, pp. 427–435, Feb. 2004.
- [23] A. S. Zamzam, B. Yang, and N. D. Sidiropoulos, "Energy storage management via deep Q-networks," in *Proc. of PESGM*, Atlanta, GA, Aug. 4–8, 2019, pp. 1–7.
- [24] Z. Yan and Y. Xu, "Data-driven load frequency control for stochastic power systems: A deep reinforcement learning method with continuous action search," *IEEE Trans. Power Syst.*, vol. 34, no. 2, pp. 1653–1656, Nov. 2018.
- [25] R. Lu, S. H. Hong, and M. Yu, "Demand response for home energy management using reinforcement learning and artificial neural network," *IEEE Trans. Smart Grid*, Apr. 2019.
- [26] B. J. Claessens, P. Vrancx, and F. Ruelens, "Convolutional neural networks for automatic state-time extraction in reinforcement learning applied to residential load control," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3259–3269, July 2018.
- [27] J. Duan, H. Xu, and W. Liu, "Q-learning-based damping control of wide-area power systems under cyber uncertainties," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6408–6418, Nov. 2018.
- [28] Q. Yang, G. Wang, A. Sadeghi, G. B. Giannakis, and J. Sun, "Two-timescale voltage regulation in distribution grids using deep reinforcement learning," in *Proc. of SmartGridComm*, Beijing, CN, Oct. 21–23, 2019, pp. 1–6.
- [29] L. Zhang, G. Wang, and G. B. Giannakis, "Real-time power system state estimation and forecasting via deep unrolled neural networks," *IEEE Trans. Signal Process.*, vol. 67, no. 15, pp. 4069–4077, Aug. 2019.
- [30] 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.
- [31] M. Baran and F. F. Wu, "Optimal sizing of capacitors placed on a radial distribution system," *IEEE Trans. Power Del.*, vol. 4, no. 1, pp. 735–743, Jan. 1989.
- [32] S. H. Low, "Convex relaxation of optimal power flow—Part II: Exactness," *IEEE Trans. Control Netw. Syst.*, vol. 1, no. 2, pp. 177–189, May 2014.
- [33] L. Gan, N. Li, U. Topcu, and S. H. Low, "Exact convex relaxation of optimal power flow in radial networks," *IEEE Trans. on Autom. Control*, vol. 60, no. 1, pp. 72–87, Jan. 2015.
- [34] 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.
- [35] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv:1609.08144*, 2016.
- [36] 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.
- [37] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," 2014.
- [38] M. E. Baran and F. F. Wu, "Optimal capacitor placement on radial distribution systems," *IEEE Trans. Power Del.*, vol. 4, no. 1, pp. 725–734, Jan. 1989.



Qiuling Yang received the B.S. degree in automation and the M.Eng. degree in control science and engineering from Yanshan University, Hebei, China, in 2012 and 2015, respectively. Since 2017, she has been working toward her Ph.D. degree in the School of Automation at the Beijing Institute of Technology, Beijing, China. She is currently a visiting student in the Digital Technology Center at the University of Minnesota, Minneapolis, MN, USA. Her research interests include multi-agent systems, machine learning, and optimization for smart power grids.



Gang Wang (M'18) received the B.Eng. degree in electrical engineering and automation from the Beijing Institute of Technology, Beijing, China, in 2011, and the Ph.D. degree in electrical engineering from the University of Minnesota, Minneapolis, MN, USA, in 2018.

He is currently a Post-Doctoral Associate with the Department of Electrical and Computer Engineering, University of Minnesota. His research interests focus on the areas of statistical signal processing, optimization, and deep learning with applications to

data science and smart grids. He was a recipient of the National Scholarship in 2013, the Guo Rui Scholarship in 2015, and the Innovation Scholarship (First Place in 2017), all from China, as well as the Best Conference Papers at the 2017 European Signal Processing Conference and 2019 IEEE Power and Energy Society General Meeting. He is currently on the editorial board of *Signal Processing*.



Jian Sun received the bachelor's degree from the Department of Automation and Electric Engineering, Jilin Institute of Technology, Changchun, China, in 2001, the master's degree from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences (CAS), Changchun, China, in 2004, and the Ph.D. degree from the Institute of Automation, CAS, Beijing, China, in 2007.

He was a Research Fellow with the Faculty of Advanced Technology, University of Glamorgan, Pontypridd, U.K., from 2008 to 2009. He was a Post-

Doctoral Research Fellow with the Beijing Institute of Technology, Beijing, from 2007 to 2010. In 2010, he joined the School of Automation, Beijing Institute of Technology, where he has been a Professor since 2013. His current research interests include networked control systems, time-delay systems, and security of cyber-physical systems. Dr. Sun is an Editorial Board Member of the *IEEE Transactions on Systems, Man and Cybernetics: Systems*, the *Journal of Systems Science & Complexity*, and *Acta Automatica Sinica*.



Alireza Sadeghi received the B.Sc. degree (Hons.) from the Iran University of Science and Technology, Tehran, Iran, in 2012, and the M.Sc. degree from the University of Tehran, in 2015, both in Electrical Engineering. In 2015, he was a visiting scholar at the University of Padova, Padova, Italy. Since January 2016, he has been pursuing the Ph.D. degree in the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, USA. His research interests include machine learning, optimization, and signal processing with applications

to networking. He was the recipient of an ADC Fellowship awarded by the Digital Technology Center of the University of Minnesota, and student travel awards from the IEEE Communications Society and the National Science Foundation.



Georgios B. Giannakis (F'97) received the Diploma degree in electrical engineering from the National Technical University of Athens, Athens, Greece, in 1981. From 1982 to 1986, he was with the University of the Southern California, Los Angeles, CA, USA, where he received the M.Sc. degree in electrical engineering in 1983, the M.Sc. degree in mathematics, and the Ph.D. degree in electrical engineering both in 1986.

He was with the University of Virginia from 1987 to 1998 and since 1999, he has been a Professor with the University of Minnesota, Minneapolis, MN, USA, where he holds an Endowed Chair in Wireless Telecommunications, University of Minnesota McKnight Presidential Chair in Electrical and Computer Engineering, and serves as the Director of the Digital Technology Center. His general interests include the areas of communications, networking, and statistical signal processing—subjects on which he has published more than 450 journal papers, 750 conference papers, 25 book chapters, two edited books, and two research monographs (h-index 143). Current research focuses on learning from big data, wireless cognitive radios, and network science with applications to social, brain, and power networks with renewables. He is the (co-)inventor of 32 patents issued, and the (co-)recipient of nine best paper awards from the IEEE Signal Processing (SP) and Communications (Com) Societies, including the G. Marconi Prize Paper Award in Wireless Communications. He is also a recipient of the Technical Achievement Awards from the SP Society (2000), from the EURASIP (2005), the IEEE ComSoc Education Award (2019), a Young Faculty Teaching Award, the G. W. Taylor Award for Distinguished Research from the University of Minnesota, and the IEEE Fourier Technical Field Award (inaugural recipient in 2015). He is a Fellow of the EURASIP, and has served the IEEE in a number of posts, including that of a Distinguished Lecturer for the IEEE-SP Society.