Online Voltage Regulation with Minimum Disturbance for Distribution Grid without System Knowledge

Hamad Alduaij, *Student Member, IEEE,* and Yang Weng, *Senior Member, IEEE*

***Abstract*—Distribution systems have limited observability, as they were a passive grid to consume power. Nowadays, increasing distributed energy resources turns individual customers into "generators," and two-way power flow between customers makes the grid prone to power outages. This calls for new control methods with performance guarantees in the presence of limited system information. However, limited system information makes it difficult to employ model-based control, making performance guarantees difficult. To gain information about the model, active learning methods propose to disturb the system consistently to learn the nonlinearity. The exploration process also introduces uncertainty for further outages. To address the issue of frequent perturbation, we propose to disturb the system with decreasing frequency by minimizing exploration. Based on such a proposal, we superposed the design with a physical kernel to embed system non-linearity from power flow equations. These designs lead to a highly robust adaptive online policy, which reduces the perturbation gradually but monotonically based on the optimal control guarantee. For extensive validation, we test our controller on various IEEE test systems, including the 4-bus, 13-bus, 30-bus, and 123-bus grids, with different penetrations of renewables, various set-ups of meters, and diversified regulators. Numerical results show significantly improved voltage control with limited perturbation compared to those of the state-of-the-art data-driven methods.**

***Index Terms*—Distribution grid, Distributed energy resources, Voltage control, Data driven, Minimum disturbance, Optimal control systems, Kernel methods.**

I. Introduction

To combat climate change, distributed energy resources (DERs) are being rapidly installed in the power grid. DERs, such as photovoltaic (PV) panels, provide clean and sustainable energy. However, they also introduce new challenges for the distribution grid, which was passively controlled in the past [1]. For example, the power output of DERs nowadays is highly stochastic due to factors such as weather changes [2]. In addition, DERs result in bidirectional power flow, for which the grid was not designed to accommodate [3].

Therefore, traditional feeder control methods have an increased risk of failure, leading to excessive voltage violation, equipment damage, and cascading outages [2], [3]. For example, a traditional way to control voltage utilizes relays that measure the voltage on a specific bus and adjust a connected capacitor bank by pre-specified settings [4]. Regulation of voltage by these methods is limited, as the relays are localized, uncoordinated, and seasonally adjusted [4]. To achieve coordinated control, rules-based policies based on expert knowledge are used [5]–[7]. Such rules are based on various scenarios for regulators. However, such expert knowledge is expensive and is not universally applicable. When DERs are continuously deployed, this expert knowledge is inaccurate and will quickly become outdated. For more advanced control, physical grid models are needed to help with analytical solutions to controllability [8]–[10]. But unfortunately, these models are often unavailable or outdated on the edge of the system, especially for the secondary distribution grid [11].

An alternative approach to avoid dependency on physical models is to use data-driven algorithms, which do not require system knowledge. Furthermore, an online algorithm is needed for the distribution system because the historical data is often outdated and does not inform which line is energized [12]–[14]. The work of those authors suggests that online probing for distribution grid identification methods are called perturb and observe methods. Although many online data-driven algorithms exist, not all are suitable for voltage regulation. We need a safe data-driven approach for effective voltage regulation that determines the optimal reactive power value based on feedback from observed voltage measurements [15]. While many data-driven algorithms can learn and optimize the system, many do not meet the safety requirements of the unknown power grid. For effective voltage regulation, we need a data-driven approach that determines the optimal reactive power value based on feedback from observed voltage measurements [15]. This need is addressed by reinforcement learning, a subset of machine learning techniques used for data-driven control of systems with unknown dynamics. Using previous interactions, a reinforcement learning agent can learn to take the best action based on error feedback. In power systems, reinforcement learning has been used mainly to perform optimal voltage regulation of known networks. The training is done offline in a simulation, as it involves unconstrained exploration, which would destabilize a real network in online deployment [16]–[18]. Therefore, for unknown networks, the online adaptation of reinforcement learning for voltage regulation in the power domain has practical challenges. The challenges are (1) the lack of a performance guarantee, (2) the requirements to learn the model structure based on historical data [18], [19], which are not always learnable [20], and (3) aggressive probing of reinforcement learning would lead to grid failure in online deployment. [21]. Thus, reinforcement learning has mainly seen applications with offline training, where the agent is trained using a simulated environment or past historical data.

Extremum-seeking, model-free approaches, which also typically employ perturb and observe schemes, such as [22]–[24]

have been used and were able to solve (1), and (2), providing a performance guarantee in the absence of historical data. These methods use gradient search techniques with a probing signal to perform optimized voltage regulation. However, gradient-based approaches have disadvantages, as their performance deteriorates significantly with the presence of noise, which can cause a high disturbance in the system [25]. In addition, while gradient-based approaches can find the global optimum using convex optimization, they do not establish an upper bound on their worst-case performance. Thus, (3) remains unsolved, as they still apply consistent probing, negatively impacting the grid. Thus, the fundamental challenge in online voltage regulation of an unknown distribution grid is the amount of perturbation required to learn a safe control strategy without historical data. Our proposed controller achieves safe control with minimal grid perturbation as follows.

In this paper, we propose to design an online data-driven controller with provable guarantees on its learning capabilities. Recent work in adaptive control theory establishes a theoretical understanding of the probing necessary to learn an unknown stochastic linear system [26]–[28]. To obtain this representation, we derive a linearized relationship between the system voltages and the regulators' reactive powers using an estimate of the inverse Jacobian matrix. Given this is not an exact model, we represent all unmodeled dynamics as a stochastic component. Using the stochastic model, we design an adaptive linear quadratic regulator (LQR) to minimize the voltage deviation. LQR is an optimal control technique that provides optimal control actions to linear systems with quadratic loss. Then, we leverage concepts from [27] to demonstrate that we can learn the optimal gains of the LQR controller using decreasing probing signals, minimizing the disturbance on the grid. The probing signals are designed to decrease in magnitude and frequency. Unlike previous methods, our proposed methods proposed a worst-case bound on the performance in finite time, which is critical for online deployment.

For systems with higher non-linear error where the stochastic linear model may not be suitable, we propose integrating the learning policy with a quadratic kernel representation of the system in the lifted space. Such a design can capture the non-linear relationship of power and voltage with higher accuracy. The kernel model is optimized using a model predictive control (MPC) with a disturbance rejection component to handle the learning error. This approach will require longer probing to be learned, but can offer improved regulation in the presence of significant non-linearity. An outline of the proposed method highlighting the motivation, challenges and contributions is shown in Fig. 1.

The model is extensively validated with multiple IEEE test cases, including IEEE 4-bus, 8-bus, 13-bus, and 123-bus feeders, based on MATLAB using the OpenDSS library to handle unbalanced networks. The power profile was obtained from the National Renewable Energy Laboratory (NREL) synthetic data set and our utility partners. From the numerical validation, the controller achieves optimal voltage regulation with less perturbation and higher performance than reinforcement learning and droop control. The controller also exhibits good performance under different topologies with varying observabilities. In addition, the kernel-based algorithm extension is validated for networks with high nonlinearities.

The rest of this paper is organized as follows. Section II shows how we transform the power system equations into the linear formulation needed to apply the optimal controller of STR. Section III introduces the linear model that is learned to predict voltages based on regulator operations. Section IV provides mathematical analysis of convergence guarantees and details the identification policy for minimum disturbance. Section V shows an extension to handle highly nonlinear cases. Section VI evaluates the performance of our methods, and Section VII concludes the paper.

II. Mathematical Modeling for Control in Distribution Grids

For modeling, we assume that there is insufficient knowledge of the system to establish a power flow model of the distribution feeder and its corresponding impedance matrix. However, at least a few network-connected distributed feeder measurements provide readings at regular intervals. At the same time, network-connected distribution regulators are assumed to provide continuous kvar. Such regulators include capacitors, D-STATCOM devices, distributed energy storage devices, and PV inverters. Mathematically, we need to obtain a transition relationship between the input of the regulator and the voltage of the system based on the power flow equation for control purposes. But power flow equations express the relationship between voltage and power in a single time step. Therefore, we convert the power flow equation into a difference equation to relate system voltage with input power using the AC power flow relationship below [29] using the Newton-raphson method.

$$\begin{bmatrix} \Delta\delta_1 \\ \vdots \\ \Delta\delta_M \\ \Delta v_1 \\ \vdots \\ \Delta v_M \end{bmatrix} = \begin{bmatrix} \frac{\partial p_1}{\partial \delta_1} & \cdots & \frac{\partial p_1}{\partial \delta_M} & \frac{\partial p_1}{\partial |v_1|} & \cdots & \frac{\partial p_1}{\partial |v_M|} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial p_M}{\partial \delta_1} & \cdots & \frac{\partial p_M}{\partial \delta_M} & \frac{\partial p_M}{\partial |v_1|} & \cdots & \frac{\partial p_M}{\partial |V_M|} \\ \frac{\partial q_1}{\partial \delta_1} & \cdots & \frac{\partial q_1}{\partial \delta_M} & \frac{\partial q_1}{\partial |v_1|} & \cdots & \frac{\partial q_1}{\partial |v_M|} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial q_M}{\partial \delta_1} & \cdots & \frac{\partial q_M}{\partial \delta_M} & \frac{\partial q_M}{\partial |v_1|} & \cdots & \frac{\partial q_M}{\partial |v_M|} \end{bmatrix}^{-1} \begin{bmatrix} \Delta p_1 \\ \vdots \\ \Delta p_M \\ \Delta q_1 \\ \vdots \\ \Delta q_M \end{bmatrix}, \tag{1}$$

where $\Delta\delta_i$ and Δv_i are the changes in voltage phase angle and magnitude at bus i, Δp_i and Δq_i are the changes of real and reactive powers at bus i, and M is the number of buses. In distribution systems, the goal is to regulate the voltage magnitude; therefore, we only consider the lower half of (1). Furthermore, since we limit ourselves to reactive power control of regulators, the change in active power, Δp, is determined by load changes of customers between the present state and the next state of the feeder. As Δp cannot be controlled, its impact is integrated into the stochastic component. Therefore, (1) is represented as an approximation rather than an equality.

$$\begin{bmatrix} \Delta v_1 \\ \vdots \\ \Delta v_M \end{bmatrix} \approx \begin{bmatrix} \frac{\partial q_1}{\partial |v_1|} & \cdots & \frac{\partial q_1}{\partial |v_M|} \\ \vdots & \ddots & \vdots \\ \frac{\partial q_M}{\partial |v_1|} & \cdots & \frac{\partial q_M}{\partial |v_M|} \end{bmatrix}^{-1} \begin{bmatrix} \Delta q_1 \\ \vdots \\ \Delta q_M \end{bmatrix}. \tag{2}$$

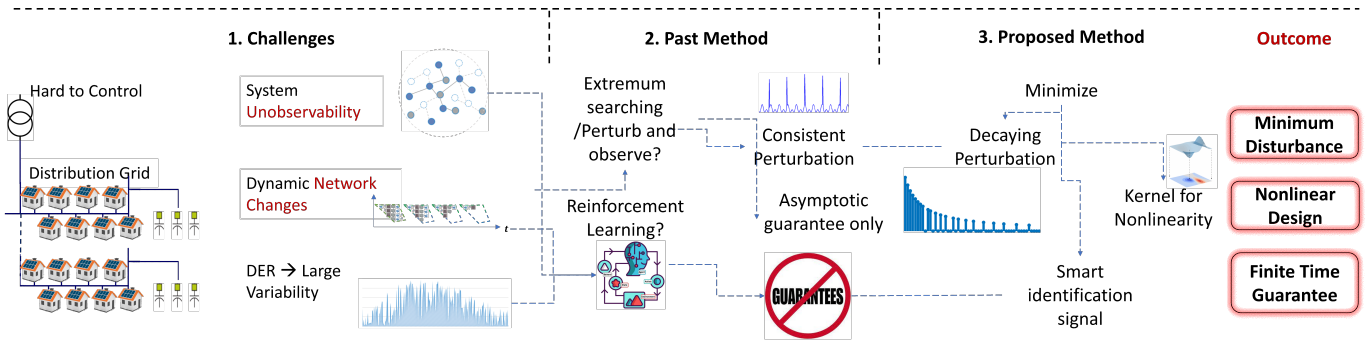

Fig. 1: Summary of challenges and contributions.

We do not have voltage measurements or controls on every bus. So, we use the available measurements and control input to construct an approximate model below with r controller and k voltage measurements.

(2) can be seen as a linearization of the distribution feeder around its operating point [8]. However, linearization of (2) is time-varying, as the operating point of the network can change [8]. We approximate the time-varying linearization Jacobian by an average constant parameter matrix with an additional stochastic component, $E \in \mathcal{R}^{k\times r}$

$$E = \begin{bmatrix} E_{1,1} & \cdots & E_{1,r} \\ \vdots & \ddots & \vdots \\ E_{k,1} & \cdots & E_{k,r} \end{bmatrix} \tag{3}$$

Subsequently, (2) can be approximated as $\Delta \mathbf{v} \approx E\Delta \mathbf{q}$, or

$$\mathbf{v}[n+1] \approx E\mathbf{q}[n+1] - E\mathbf{q}[n] + \mathbf{v}[n], \tag{4}$$

where n is the time index, $\mathbf{v}[n] = [v_1[n], \cdots, v_k[n]]^\top$, and $\mathbf{q}[n] = [q_1[n], \cdots, q_r[n]]^\top$. In our mathematical model above, we have approximation errors in (4). They are caused by approximating the Jacobian near the operating point, the contribution of unmeasured nodes, and the omission of real power. The real power is treated as part of the stochastic component, even when it is measurable. This is because, for many devices, it is an external input that we cannot control. Therefore, we cannot establish any guarantee to learn the resistances of an unknown network. To quantify their impact on our design, we wrap them in the noise term $\mathbf{w}[n] \in \mathcal{R}^k$ to transform the inequality (4) into an equation below.

$$\mathbf{v}[n+1] = E\mathbf{q}[n+1] - E\mathbf{q}[n] + \mathbf{v}[n] + \mathbf{w}[\mathbf{n}], \tag{5}$$

We will show with real data that w can be modeled approximately as a Gaussian random variable with bounded variance. This fact allows us to design a controller with performance guarantees.

III. Optimal Control with Performance Guarantees

This section aims to design a data-driven controller with the following features. The controller should be guaranteed to stabilize a power system with unknown parameters. The controller should have minimal disturbances to the system in the learning process. Finally, the controller should handle systems with higher non-linearity. Therefore, we choose linear-quadratic controller based optimal control theory [28] for our base design and show how to extend the design using a quadratic kernel with an adaptive model predictive control approach to handle large non-linearities [30].

A. State-Space Representation for Linear–Quadratic Regulator

The linear quadratic regulator is based on state-space representation. Therefore, we show how to convert the linear model to the state-space model in the last section. Specifically, state-space representation consists of inputs, outputs, and a system of first-order difference equations. For the distribution grid, since we need to regulate the voltage, we let the output be $\mathbf{v}[n]$ and the control variable be the reactive power $\mathbf{q}[n]$ on the regulator.

$$\mathbf{v}[n+1] = \mathbf{v}[n] + E\mathbf{q}[n+1] - E\mathbf{q}[n] + \mathbf{w}[n], \tag{6}$$

where E is unknown to the controller. To apply the optimal state-feedback control, we need to write the control as a function of n only. For this purpose, we use a change of variables. The key is to write the new control signal $q[n+1]$ in terms of the current time index. $\mathbf{q}[n+1] = \mathbf{q}[n] + L(\mathbf{v}[n] - \mathbf{v_{ref}})$ Consequently, (6) becomes

$$\begin{aligned} \mathbf{v}[n+1] =& \mathbf{v}[n] + E\mathbf{q}[n+1] - E\mathbf{q}[n] + \mathbf{w}[n] \\ \mathbf{q}[n+1] =& \mathbf{q}[n] + L(\mathbf{v}[n] - \mathbf{v_{ref}}) \end{aligned}$$

which can be combined as follows

$$\mathbf{v}[n+1] = \mathbf{v}[n] + E(L(\mathbf{v}[n] - \mathbf{v_{ref}})) + \mathbf{w}[n] \tag{7}$$

which is in the form of first-order difference equation, enabling the design of state-feedback optimal control. The closed-loop system is

$$\mathbf{v}[n+1] = D\mathbf{v}[n] + \mathbf{w}[n], \tag{8}$$

where D is the closed-loop state-feedback matrix such that $D = I + EL$, where L is the control gain, which will be derived in the next section. To guarantee convergence through asymptotic stability, we need to estimate E and design L such that D is stable.

For the distribution feeder, the E matrix represents the network connections and contains $k \times r$ elements, where k

is the number of observed buses, r is the number of regulated capacitors, and n is the current time index. To learn E, rewrite (7) as

$$\Delta \mathbf{v}[n+1] = E\mathbf{q}[n] + \mathbf{w}[n]. \tag{9}$$

Using this relationship, we inject a probing reactive power and measure the output voltage. Then, we perform the least squares estimate (LSE) to obtain E with E_i being the i^{th} row corresponding to the i^{th} bus of the network.

$$E_i[n] = (Q_{i,n}^\top Q_{i,n})^{-1} Q_{i,n}^\top \Delta \mathbf{v}_i[n] \quad \forall i \in \{1, \cdots, k\} \tag{10}$$

where

$$Q_{i,n} = [\mathbf{q}_{i,1}, \cdots, \mathbf{q}_{i,n}]^\top \in \mathbb{R}^{n \times r}. \tag{11}$$

B. Optimal Design for Performance Guarantee

After estimating the system parameter E, we can design the control gain L for optimal performance. When there is no noise, we derive our control policy π, for an optimal controller with known parameters. This policy is designed to minimize the quadratic cost of the voltage deviation output from the reference voltage and the regulation input.

For optimal control, we choose the quadratic cost function

$$c_\pi = \sum_{n=0}^{\infty} \left((\mathbf{v} - \mathbf{v_{ref}})[n]^\top S (\mathbf{v} - \mathbf{v_{ref}})[n] + \mathbf{q}[n]^\top R \mathbf{q}[n] \right), \tag{12}$$

where π is the control policy, S is a positive semi-definite matrix that weighs the voltage deviation cost, and R is a positive definite matrix that weighs the regulator action cost. The goal is to control the linear system that we derived in (7). The optimal control policy π^* is

$$\pi^* : \mathbf{q}[n] = L(E^*)(\mathbf{v}[\mathbf{n}] - \mathbf{v_{ref}}), \tag{13}$$

where L is $r \times n$ matrix corresponding to the gain of the controller, and E^* is a matrix such that $L(E^*)$ is the optimal gain. $L(E^*)$ is obtained by iteratively solving the Riccati equations for $K(E^*)$ [28]. The Riccati equations ensure asymptotically stability for the closed-loop system matrix D and minimize the cost function c, which is the basis for our guarantee.

$$\begin{aligned} K(E) =& S + K(E) - K(E) E \\ & \cdot \left(E^\top K(E) E + R \right)^{-1} E^\top K(E), \end{aligned} \tag{14}$$

$$L(E) = -\left(E^\top K(E) E + R \right)^{-1} E^\top K(E). \tag{15}$$

(13)-(15) provide an optimal control gains for a given E and cost function specified by (12); however, the true system matrix E^* is unknown. Therefore, the controller's performance will depend on its capability to learn the true system dynamics accurately and quickly. The next section provides an identification policy to estimate the system parameters. Then, we state the requirements for our controller to converge to $L(E^*)$ and illustrate that the distribution grid satisfies the requirements. Finally, we provide a combined policy of control and identification with decaying perturbation frequency and magnitude and a bound on the deviation from the optimal gain $L(E^*)$ for a given time index n.

C. Conditions for Strongly Consistent Least Square Estimates

In our design above, we build on the principle of certainty equivalence, which states that the presence of system uncertainty and additive noise does not change the optimal control [28]. Therefore, the control strategy for a known system without noise will be optimal under an unknown noisy system. However, our approximation does have noises $\mathbf{w}[\mathbf{n}]$, which require consistency analysis. Therefore, we show two conditions below for our controller to be consistent [28].

Condition 1. *The system is stabilizable: All uncontrollable buses are stable, and there exists a stabilizing feedback $L \in R^{r \times p}$ such that $|\lambda_{max}(I + EL)| < 1$.*

Condition 1 means that the distribution system is stabilizable, and this is a typical requirement in control theory. For a distribution grid, these buses that cannot be affected by the controllers will not cause voltage collapse. This is practical as the distribution grid is with different protection devices. We can assume that the targeted grid for our control policy is stabilizable with enough equipment for many utilities. For the other places, the proposed method is deployable for the majority of the time except for a few exceptions. This is because utilities have grid upgrades for good control over systems.

Condition 2. *The stochastic component needs to satisfy bias and variance requirements,* $\mathbb{E}(\mathbf{w}[n]) = 0$, $\sup_n \mathbb{E}(||\mathbf{w}[n]||^\alpha) < \infty$ for some $\alpha > 2$.

Condition 2 requires that the stochastic component is unbiased and the variance is finite. This condition comes from the mathematical derivation of the LSE. Intuitively, for estimation to be feasible, the variance of the noise must be finite. So, let us focus on checking if our modeling can ensure $\mathbb{E}(\mathbf{w}) = 0$ or $\mathbb{E}(w_i) = 0$ for the i^{th} bus. Because of our modeling, w_i has three components: measurement error m_i due to sensor quality, feedback delay error f_i for users' load consumption, and the modeling error l_i during the data-driven parameter estimation. If we can show $\mathbb{E}(m_i) = \mathbb{E}(f_i) = \mathbb{E}(l_i) = 0$, $\mathbb{E}(w_i) = 0$ leading to $\mathbb{E}(\mathbf{w}) = 0$.

- For m_i, past work generally assumes the measurement error in distribution feeders to be Gaussian with a zero mean. More recent studies have shown that a combined Gaussian may be a better assumption, but the expected error is still nearly 0 [31].
- For the feedback delay f_i, it is an outcome of the load change in between stages $n-1$ and n. Previous work has shown that this component is Gaussian for shorter intervals [32]. For more evidence, we examined this claim using two data sets. The first data set is loading data from the National Renewable Energy Laboratory , a synthetic load profile representing a typical load curve for a year and simulated in the IEEE-123 load case. The second data set is from the local voltage profile of a sample distribution feeder in Phoenix, Arizona. For the NREL data simulation and the Arizona feeder, the feedback delay errors are on the order of 10^{-6} p.u., which is

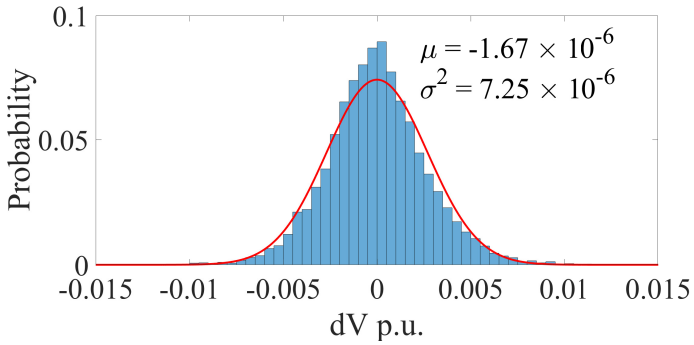

Fig. 2: The change in voltage data from a year of unregulated hourly operation at a measured bus from NREL data. The distribution is numerically symmetric and the mean is zero visually.

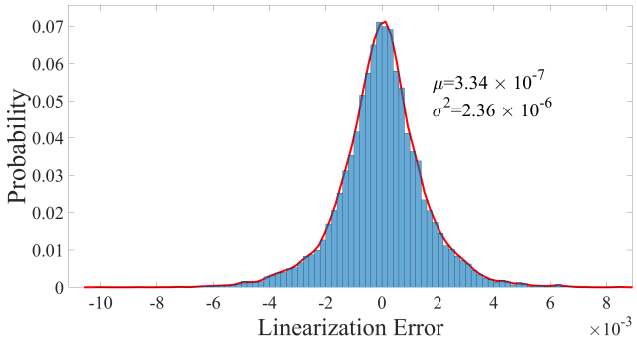

Fig. 3: Histogram of the error of using the linearized model instead of the power flow equations in (1). The error can be well approximated using a zero mean Gaussian distribution.

nearly zero. Here, we show the NREL data in Fig. 2 as an illustration.

- For l_i, it arises from assuming constants for the sensitivity matrix, the omission of real power, and estimation errors in the LSE. Within an operation point, a constant matrix estimate leads to negligible error, which has been shown in simplified power flow solutions [29]. Although it is not possible to prove that the error is exactly zero, [33] proves that with sufficient persistent perturbation, the estimation error is minimized, allowing the LSE to converge to true values. To further validate this claim, using the NREL data set, we compute $l_i = (\mathbf{v}[n+1] - E\mathbf{q}[n+1] - E\mathbf{q}[n] + \mathbf{v}[n])$, where $\mathbf{v}[n+1]$ is obtained from the exact nonlinear expression in (1). Then, we fit a histogram to show that the error can be represented as a zero-mean Gaussian random variable as seen in Fig. 3.

Based on conditions 1 and 2, Theorem 1 shows that the parameter estimation of E will converge to the true parameter, leading to both system convergence and optimality of our controller.

Theorem 1. *If the system is stabilizable as in Condition* 1*, and its stochastic components satisfy Condition* 2*, then given the control policy of* (13) *, the LSE for bus i at stage n is consistent, and $L(E)$ will converge to $L(E^*)$.*

Proof. See Appendix 1. □

Thus, we can state that the LSE is consistent and that the controller will stabilize the system, guaranteeing convergence to the optimal control $L(E^*)$.

Proposition 1. *(Optimal policy [26]–[28]) if the linear system in* (7) *is stabilizable,* (14) *has a unique solution, and π^* defined in* (13) *is an optimal regulator.*

This proposition assures that any stabilizable linear system has a unique optimal solution given by (14). Thus, we guarantee stability by finding a consistent estimate of E^*, which yields $L(E^*)$ that can stabilize the system.

IV. Reduce Disturbance but Keep Performance Guarantee

A. Trade-off between Exploration and Exploitation

Theorem 1 establishes an asymptotic convergence guarantee. Persistent perturbation is deployed to identify E for better control. However, the system operators want to limit the perturbation, as the disturbance may lead to outages in the network. To capture the trade-off, we need a metric that captures both the transient degradation of a perturbation and the steady-state improvement it causes. For this purpose, we adopt regret, describing the deviation between our identified policy π, and the optimal policy π^*:

$$r^{\pi} = c_{\pi}[n] - c_{\pi^*}[n], \tag{16}$$

where π^* knows E^*, and the instantaneous regret is computed for a single hour. To measure how well a policy regulates the measured outputs over time, we need to define total regret as the running sum of the instantaneous regret r^{π}, such that $\mathcal{R}^{\pi} = \sum_{i=0}^{n-1} r^{\pi}[i]$. Using this metric, we can investigate the rate at which our estimate converges to the optimal control. To establish a bound on the regret, we examine the behavior of the identification policy with two stages [28]:

- Find an initial estimate $L(\hat{E})$.
- Apply an asymptotically consistent adaptive policy to regulate and estimate the voltage.

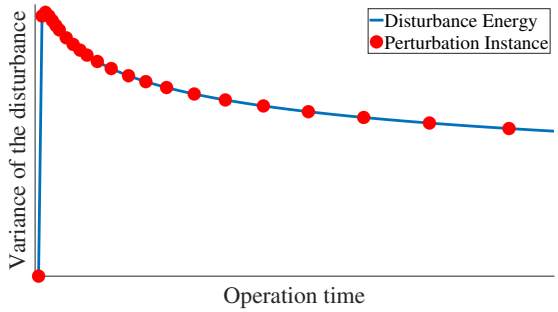

Fig. 4: Perturbation magnitude and frequency over time for a sample γ decaying policy .

The initial estimate $L(\hat{E})$ can be obtained by applying k linearly independent random probing inputs to obtain $\hat{E}$ as an estimate of E and subsequently $L(\hat{E})$ as an estimate of $L(E)$ using (15). During the first k instances, the policy will randomly set the capacitor states $\mathbf{q}[\mathbf{n}]$. After an initial

LSE is obtained, the second stage is initiated with persistent perturbation. In this stage, the controller will either explore the system to improve the estimation or exploit using optimal regulation for each operation hour. During exploration, a random perturbation is added to improve the estimation at each $n = \lfloor \gamma_i \rfloor$, so that γ_i is the perturbation instance. During exploitation, we apply (15) without perturbation. The faster γ_i grows, the more spaced out the instances and the less we perturb the system. Fig. 4 visualizes the perturbation process. The red dots indicate the instances in which the perturbation is added to the control output. Both the frequency and the magnitude of the perturbation decrease over time. Note that the plot shown for the figure is illustrative; in the next section, we propose a minimum perturbation policy with guarantees on the worst-case regret.

B. Regret-Based Policy for Asymptotic Convergence

In previous iterations of the STR, during the exploration process, the controller uses a random signal. This leads to sub-optimal regulation during the exploration. Instead, we can apply the optimal control signal with a small superimposed random signal with a constrained design. This ensures that we are still regulating during the exploration steps.

Algorithm 1: Input Perturbation Algorithm for Unknown Distribution Voltage Control

Input $\gamma > 1$ and $\Sigma > 0$.
Let r be the number of input capacitors.
Obtain initial $\hat{E}_0$ and $L(\hat{E}_0)$ by generating r randomly generated $\mathbf{q}$ then using (10) and (14).
for $m \in (r+1, \cdots, \infty)$ **do**
  if $n = \lfloor \gamma^m \rfloor + r$ **then**
    $\mathbf{q}[n] = L(\hat{E}_n)(\mathbf{v}[n] - \mathbf{v_{ref}}) + \tilde{\mathbf{q}}[n]$, where $\tilde{\mathbf{q}}[n]$ is drawn according to (17).
    Update $\hat{E}$ by (10).
  else
    $\mathbf{q}[n] = L(\hat{E}_n)(\mathbf{v}[n] - \mathbf{v_{ref}})$.
  end
end

In the algorithm, $\tilde{\mathbf{q}}$ is a vector of independent zero-mean Gaussian variables satisfying:

$$\underline{C} < n^{-2}\gamma^{n/2}|\lambda_{min}(\Sigma)| < n^{-2}\gamma^{n/2}\overline{q} < \overline{C}, \qquad (17)$$

where $\overline{q}$ is the norm of $\tilde{\mathbf{q}}$ and Σ is the covariance of $\tilde{\mathbf{q}}$. The given constraints on $\tilde{\mathbf{q}}$ ensures fast convergence. The following equation, from Theorem 1, and Corollary 1 in [27], characterizes the convergence of the regret of this algorithm. The corollary allows us to state the following:

Theorem 2. *(IP rates) Suppose π is the adaptive regulator of Algorithm* 1. *Let $E[\Re_n(\pi)]$ be expected regret at time n. Then we have*

$$\lim_{n\to\infty} \frac{\sum_{i=0}^{n-1} r^{\pi}[i]}{n} = 0, sup_{n\geq n_0} E[\Re_n(\pi)] \leq C, \qquad (18)$$

where C, n_0 are constants and the mathematical formulation to compute them can be found in the Appendix of [27]. The expected value of the cumulative regret is finite. This can intuitively be seen as the regret bound is guaranteed for $n > n_0$ where n_0 is finite. We note previous methods such as the extremum seeking method and other decreasing perturbation methods [8], [28] do not provide a worse case bound for performance in finite time. Therefore, the method is claimed to asymptotically converge; grid stability is not guaranteed during the early deployment phase. On the other hand, the guarantee given by Theorem 2 is for both asymptotic and finite time. A flow chart summarizing the operation is shown in Fig. 5. In the figure, we perform system identification in stage 1 to obtain an initial estimate of the system matrix. Then, in stage 2, we follow a decaying perturbation approach using our new constrained perturbation signal.

V. Integrate Quadratic Kernel for Learning System Nonlinearity

The learning-based method can be further boosted because we have knowledge of the functionals used in the non-linear power flow equations. This is specifically the case where the variation of active and reactive power in the load profile is too large to be approximated with the linear model. In such a case, we can extend our representation to higher dimensions to capture the non-linearities for better prediction and regulation. There are various methods to represent a non-linear system, such as neural networks, regression trees, and kernels [34]. Our goal is to find a representation where a mathematical basis for the convergence guarantee can be provided. Representations such as neural networks or decision trees do not provide an analytical relationship with their models. Without an analytical model, the stability and convergence of a controller cannot be guaranteed [35]. Also, data-driven models typically require a data-driven controller design, which has good disturbance rejection compared to classical control techniques. Furthermore, machine learning algorithms typically have poor early training performance [21], making online training infeasible. However, Hofmann's work shows that the deterministic kernel approach provides an explicit relationship between the input and output.

Specifically, we choose the quadratic kernel because it best fits the physical nature of the power flow equation [29]. How to apply a quadratic kernel to the linear representation in (10)? We observe that the notation for the single-output voltage estimates of (10) describes the inputs with the vector $\mathbf{q}_i$ and the output as $\mathbf{v}_i$. Therefore, we can transform the linear inputs into their corresponding quadratic inputs $\phi(\mathbf{q}_i)$ using the quadratic kernel for bus i.

$$\phi_h(\mathbf{q}_i) = \Big[q_{i,1}^2, \cdots, q_{i,r}^2, \sqrt{2}q_{i,1}q_{i,2}, \cdots, \sqrt{2}cq_{i,1}, \sqrt{2}q_{i,2}q_{i,3}, \cdots, \sqrt{2}q_{i,r-1}q_{i,r}, \sqrt{2}cq_{i,r}, c\Big], \qquad (19)$$

where r is the number of regulated capacitors. The quadratic kernel captures the relationship between power and voltage. The corresponding parameters for the quadratic kernel φ^h are constants that correspond to the quadratic inputs. These parameters can be approximated so that the voltage output $y[n+1]_i$ can be predicted such that

$$v_i[n+1] = \varphi_i^{h\,\top}\phi_h(v_i[n], \mathbf{q}_i[n]) + \mathbf{w}_i. \qquad (20)$$

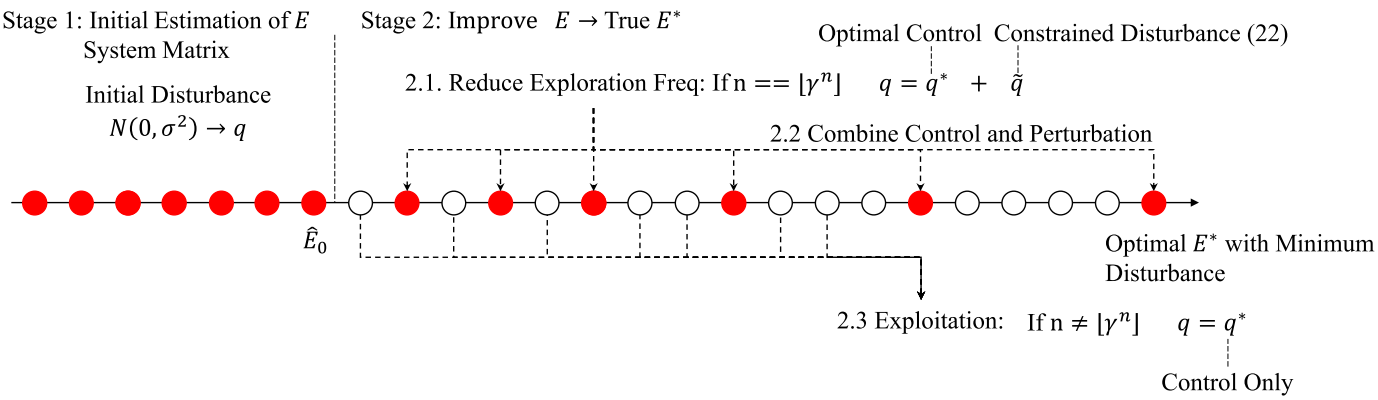

Fig. 5: Conceptual flow of the operation.

As this equation is non-linear, we cannot use (14) to find the control. Thus, we apply a suitable non-linear control method for our system. While there are many non-linear control approaches that can be found, e.g., the Lyapunov-based approach and sliding mode control [36], those methods require extensive hyper-parameter tuning under disturbance [37]. As our approach is based on perturbation, we choose model predictive control to reject disturbances and modeling errors. Therefore, we implement a model predictive control with disturbance rejection [30]. In MPC solutions, the cost function is optimized for a specific time window called a rollout horizon. We find the control action for each time window p by solving a constrained optimization problem.

$$\begin{aligned}\min_{\mathbf{q}} c_\pi[n] = \sum_{m=0}^{p} \Big[&(\mathbf{v}[n+m] - \mathbf{v_{ref}})^\top S(\mathbf{v}[n+m] - \mathbf{v_{ref}}) \\ &+\mathbf{q}[n+m]^\top R\mathbf{q}[n+m] + L(\mathbf{v}[n+m] \\ &-\hat{\mathbf{v}}[n+m])\Big], \end{aligned} \tag{21}$$

subject to the model given in (20).With this formulation, we obtain a kernel-based implementation of the controller.

VI. Numerical Validation

This section uses various case studies to examine our controller's performance and investigate the disturbance on the grid. Specifically, we use various benchmark methods, including the RL type method, e.g., Deep Policy Gradient (DDPG) algorithm and droop control method. The implementation of such a method for an unknown network is detailed in subsection VI-A. Then, we assess our convergence guarantee under different loading scenarios and using multiple load cases. Furthermore, kernel implementation is demonstrated. For an accurate simulation of the distribution grid, we use the OpenDSS MATLAB interface to simulate power flow, and the load profile data are obtained from the National Renewable Energy Laboratory data set. For diversity, we employ IEEE 4-bus, 13-bus, 30-bus, 123-bus, etc. Additional validations are with realistic feeder data from a local utility partner in Phoenix, Arizona. The grid is considered to have a high penetration of renewables.

To implement the controller, we need to select some hyper-parameters, including the state cost S, the control cost R, the perturbation energy σ, and the perturbation rate γ. The choice of S and R balances the regulation and operating cost based on the system operators' needs. Therefore, S and R should be selected to achieve the maximum regulation without exceeding the rated value of the regulators. The choices of Σ and γ impact the learning of our controller and the disturbance on the grid. Ideally, we want to select the minimum value of γ, Σ to learn the model. Based on these considerations, we conducted extensive simulations. The performances are similar, so we use the following setup for consistency: The controller was implemented with $S = I$, $R = 1000I$, where I is the identity matrix. Also, we let $\gamma = 1.4$, and $\Sigma = 0.001I$.

A. Optimal Regulation through LQR based Design

For comparison, we use droop control, where the reactive power of each regulator is determined by the voltage measurement at that regulator. Furthermore, the performance of the proposed controller has been compared with that of a DDPG-based RL controller. DDPG was chosen for comparison because it utilizes continuous action and state spaces, similar to feedback control. The DDPG is composed of two neural networks. The actor-network maps the measured state to action, and the critic network estimates the expected reward given an action-state pair. In the implementation, the action space is the MVAR compensation for each capacitor, and the state space is the voltage magnitude. Furthermore, we implemented a multi for the actor and critic networks. The reward function is the negative cost function used in adaptive LQR control.

We simulate both our controller and the droop controller for the $123-$bus system in Fig. 6. The figure shows the change in the average nodal voltage of the 123 nodes in time for the proposed controller (red), reinforcement learning (purple), and droop controller (blue). We observe that our design outperforms traditional droop control in average voltage regulation and learns faster than reinforcement learning, achieving optimal regulation.

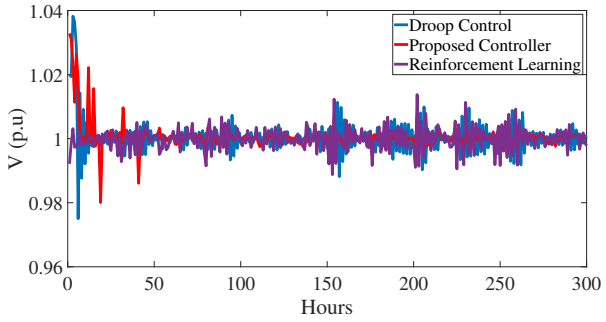

Fig. 6: Average node voltage performance benchmark against droop control and reinforcement learning.

B. Minimized Grid Disturbance with Decaying Perturbation

This simulation benchmarks against traditional LQR control with no persistent perturbation after stage 1. Fig. 7 shows the behavior of average nodal voltage in the early and late stages of operation. The figure shows that our controller has a worse initial performance due to perturbation, while having a better performance after the parameter estimate improves. This can be explained by the convergence of the parameter with the true value, which does not occur in stage 1, as shown in Fig. 8. In the figure, we show how a component of $E_{i,j}$ updates, as the x-axis represents the parameter updates defined by γ, and the y-axis is the parameter value. Furthermore, we validate our perturbation decay rate is optimal by comparing the perturbation policy per (17) against the implementation of constant perturbation in Fig. 9.

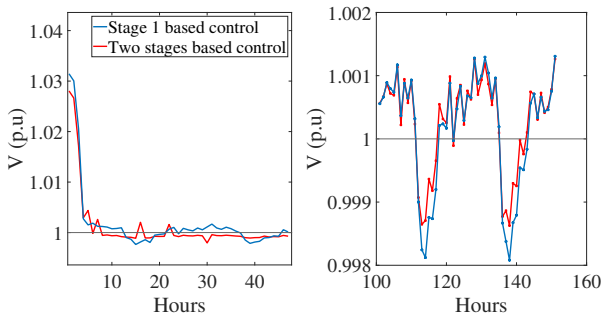

Fig. 7: Impact of persistent perturbation on the average node voltage.

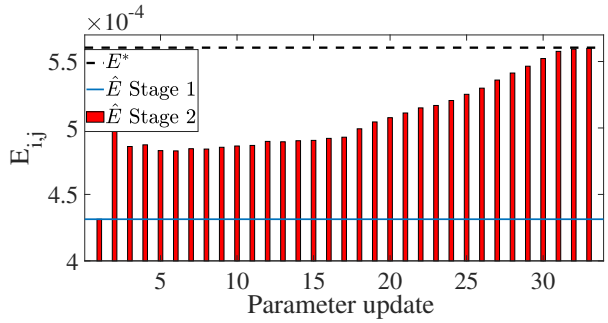

Fig. 8: Parameter convergence through perturbation.

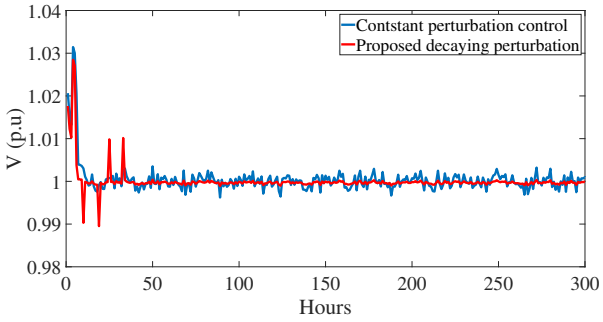

Fig. 9: Diminishing perturbation vs. constant perturbation.

We observe that constant perturbation has a consistently negative impact on the grid's voltage while the proposed controller's impact is minimal and diminishes in time. We have evaluated a typical response and the convergence rate of the model for a particular case, but a power network has many configurations. Therefore, the persistent decaying perturbation improves the performance of our controller and has a minimal impact on the voltage. Next, we examine three representative control outputs at different buses in Fig. 10. In this figure, we plot the reactive power that the regulator produces over time. The patterns include constant, oscillatory, and zero control. Furthermore, all three control patterns converge after the initial perturbation.

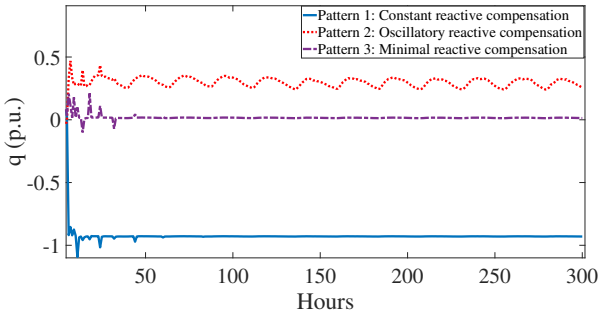

Fig. 10: Reactive power controller output at different buses to highlight different control patterns.

C. Robust Stability via LQR Design

To validate the convergence guarantee and the robustness of the control, we implemented the controller using multiple measurements and regulator configurations. Then, we simulate other test cases to verify consistent performance. Using the IEEE-123 test case, twenty-five simulations were performed. Each simulation represents a different regulator and voltage measurement availability. In each case, the availability is increased by 20% of the total available buses by assigning more measured and regulated buses.

Fig. 11 shows the controller performance as the percentage of the measured and controlled buses vary. The x-axis shows the percentage of controlled buses and the y-axis is the regret. Each line shows a percentage for different ratios of measured buses. Both controlled and measured buses are chosen randomly. The figure indicates that, while the performance

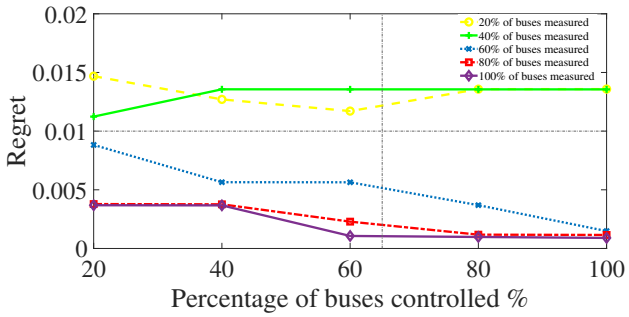

Fig. 11: Change of regret as Percentage of measured buses vary.

may vary on different penetrations, the controller converges to a finite regret, which shows the combined performance of voltage regulation and input cost.

Furthermore, we tested the controller on different grid designs to show robustness, including the IEEE 4-bus, 13-bus, and 30-bus cases. Also, we tested the controller with the CLW-13 feeder network configuration. The CLW-13 is a local distribution feeder in Phoenix, Arizona, operated by the Arizona Public Service (APS) utility. We obtain a simplified model of their feeder and load profiles and use them to test our controller. The simulations can be seen in Fig. 12. In the figure, the y-axis represents the mean absolute voltage deviation for all nodes. We see that, while the behavior may vary as the topology and loading data vary, the voltage deviation of the measured buses is always minimized. Thus, the previous results confirmed our performance guarantee for numerous test cases and loading scenarios. Next, we study the sensitivity of our controller to varying disturbance levels.

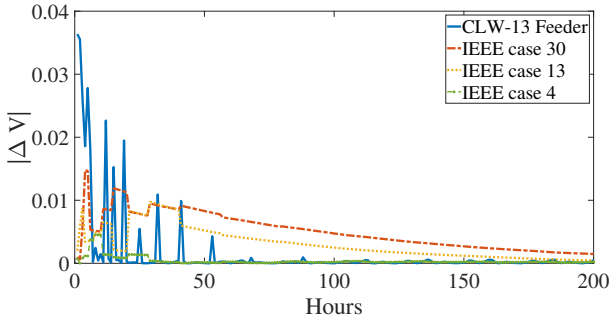

Fig. 12: Average absolute voltage deviation for four different cases.

D. Validate the Conditions of the Proposed Method via Sensitivity Analysis

Here, we study our controller's sensitivity to the disturbance level. We vary the disturbance by changing the variation of the real power, which is a primary source of disturbance to the system. In Fig. 13, we plot a box plot of the voltage deviation against the variance of the total system noise w. We observe that our controller becomes unstable beyond a critical disturbance level, and the error grows exponentially. We repeat a similar experiment by changing the estimation bias by adding an error vector to the optimal gain as shown in Fig. 14. We observe a decrease in the controller's performance as the estimation bias increases; however, the controller is still robust to small estimation errors. Therefore, we notice that the controller's stability is not limited by the topology but by the disturbance of the system validating, as mentioned in Condition 2. Next, the kernel control performance is studied.

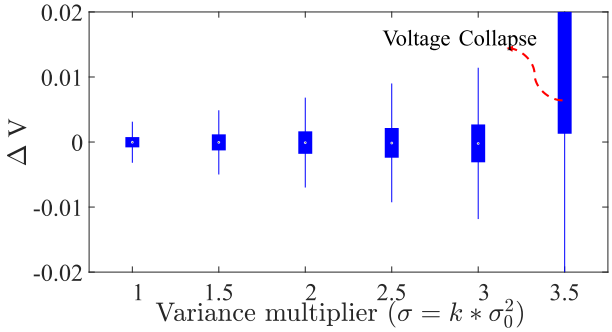

Fig. 13: Stability of controller with increasing disturbance variance.

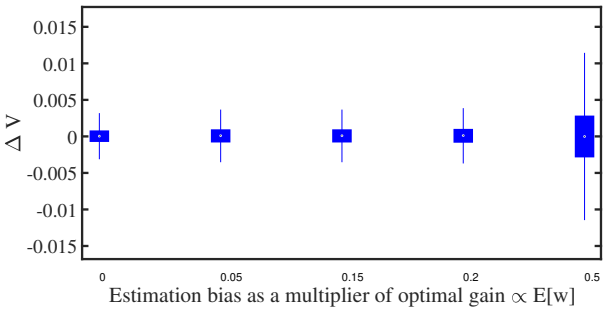

Fig. 14: Stability of controller with increasing disturbance bias.

E. Improved Performance via Kernel Implementation of Nonlinear Design

The kernel implementation is important for large disturbances inducing highly nonlinear behavior. To benchmark the kernel design in such a scenario, the performance of the IP, DDPG control, and kernel-based algorithms is shown in Fig. 15. Notice that the DDPG algorithm experiences voltage spikes and exhibits poorer performance, while the kernel has the best performance. We see that the regret has negative values, which means the kernel implementation outperforms π^*, as defined in Section III. That is the optimal policy that can be achieved by a linear controller that knows the true parameters of the system.

VII. Conclusion

This paper proposes an algorithm to regulate an unknown distribution system online in the presence of stochastic elements such as DERs with a specified convergence rate. Unlike previous work, we provide a policy with provable performance guarantees and minimal grid disturbance. Additionally, our

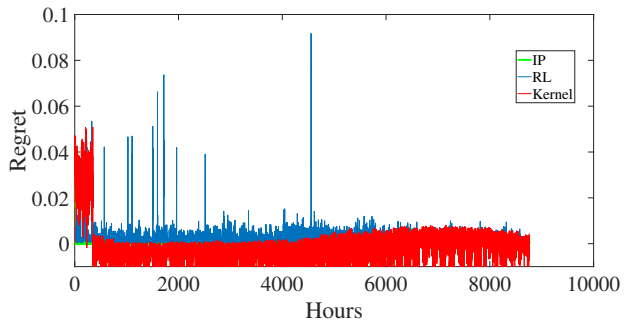

Fig. 15: Regret of kernel algorithm in comparison to the linear controllers and RL.

method contains a theoretical closed-loop stability guarantee, which is uncommon for data-driven methods, e.g., reinforcement learning. Using optimal control theory, an optimal adaptive LQR controller was designed for power systems, and we provide an extension to the kernel representation to deal with non-linearity for better performance. The results show improved voltage regulation capability compared to reinforcement learning and local droop control. Our decaying perturbation scheme exhibits lower disturbance than consistent perturbation while still learning the optimal gains. We demonstrated the robustness of the controller by showing the voltage regulation capability after imposing additional uncertainty on the system and varying grid topology. Our result also validates the adequacy of our quadratic kernel extension to handle cases with profiles with large fluctuations leading to an increased linearization error. Future research work can improve the kernel extension by employing different disturbance rejection methods, such as the attractive ellipsoid method, and obtaining worst-case regret bounds for nonlinear identification.

REFERENCES

[1] R. E. Fehr, *Industrial power distribution*, 2nd ed., ser. IEEE Press Series on Power and Energy Systems. Wiley, 2016.

[2] A. Woyte *et al.*, "Voltage fluctuations on distribution level introduced by photovoltaic systems," *IEEE Transactions on Energy Conversion*, vol. 21, no. 1, pp. 202–209, 2006.

[3] M. E. Baran *et al.*, "Accommodating high pv penetration on distribution feeders," *IEEE Transactions on Smart Grid*, vol. 3, no. 2, pp. 1039–1046, 2012.

[4] C. R. Sarimuthu *et al.*, "A review on voltage control methods using on-load tap changer transformers for networks with renewable energy sources," *Renew. Sust. Energ. Rev.*, vol. 62, pp. 1154–1161, 2016.

[5] R. Walling *et al.*, "Summary of distributed resources impact on power delivery systems," *IEEE Transactions on Power Delivery*, vol. 23, no. 3, pp. 1636–1644, 2008.

[6] Y. P. Agalgaonkar, B. C. Pal, and R. A. Jabr, "Distribution voltage control considering the impact of pv generation on tap changers and autonomous regulators," *IEEE Transactions on Power Systems*, vol. 29, no. 1, pp. 182–192, 2013.

[7] P. Brady, C. Dai, and Y. Baghzouz, "Need to revise switched capacitor controls on feeders with distributed generation," in *IEEE PES Transmission and Distribution Conference and Exposition*, vol. 2, 2003, pp. 590–594.

[8] A. Bernstein and E. Dall'Anese, "Bi-level dynamic optimization with feedback," in *IEEE Global Conference on Signal and Information Processing*, 2017.

[9] X. Zhou *et al.*, "Accelerated voltage regulation in multi-phase distribution networks based on hierarchical distributed algorithm," *IEEE Transactions on Power Systems*, vol. 35, no. 3, pp. 2047–2058, 2019.

[10] Fazio, Anna Rita Di and Risi, Chiara and Russo, Mario and Santis, Michele De, "Coordinated Optimization for Zone-Based Voltage Control in Distribution Grids," *IEEE Transactions on Industry Applications*, vol. 58, no. 1, pp. 173–184, 2022 .

[11] W. Warwick *et al.*, "Electricity distribution system baseline report," *Pacific Northwest National Laboratory*, vol. 25178, 2016.

[12] G. Cavraro and V. Kekatos, "Graph algorithms for topology identification using power grid probing," *IEEE Control Systems Letters*, vol. 2, no. 4, pp. 689–694, 2018.

[13] ——, "Inverter probing for power distribution network topology processing," *IEEE Transactions on Control of Network Systems*, vol. 6, no. 3, pp. 980–992, 2019.

[14] S. Taheri, V. Kekatos, and G. Cavraro, "An milp approach for distribution grid topology identification using inverter probing," in *2019 IEEE Milan PowerTech*, 2019, pp. 1–6.

[15] S. Karagiannopoulos, P. Aristidou, and G. Hug, "Data-Driven Local Control Design for Active Distribution Grids Using Off-Line Optimal Power Flow and Machine Learning Techniques," *IEEE Transactions on Smart Grid*, vol. 10, no. 6, pp. 6461–6471, 2019.

[16] R. E. Helou, D. Kalathil, and L. Xie, "Fully Decentralized Reinforcement Learning-based Control of Photovoltaics in Distribution Grids for Joint Provision of Real and Reactive Power," *IEEE Open Access Journal of Power and Energy*, vol. 8, pp. 175–185, 2020.

[17] R. Diao *et al.*, "Autonomous Voltage Control for Grid Operation Using Deep Reinforcement Learning," *IEEE Power and Energy Society General Meeting*, 2019.

[18] J. Duan *et al.*, "Deep-Reinforcement-Learning-Based Autonomous Voltage Control for Power Grid Operations," *IEEE Transactions on Power Systems*, vol. 35, no. 1, pp. 814–817, 2020.

[19] G. Dulac-Arnold, D. Mankowitz, and T. Hester, "Challenges of real-world reinforcement learning," *arXiv preprint arXiv:1904.12901*, 2019.

[20] H. Li *et al.*, "Distribution grid impedance topology estimation with limited or no micro-pmus," *International Journal of Electrical Power Energy Systems*, vol. 129, p. 106794, 2021.

[21] G. Dulac-Arnold *et al.*, "Challenges of real-world reinforcement learning: definitions, benchmarks and analysis," *Machine Learning*, pp. 1–50, 2021.

[22] D. B. Arnold *et al.*, "Model-free optimal control of var resources in distribution systems: An extremum seeking approach," *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 3583–3593, 2016.

[23] H. Nazaripouya *et al.*, "Real-time model-free coordination of active and reactive powers of distributed energy resources to improve voltage regulation in distribution systems," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 3, pp. 1483–1494, 2020.

[24] Y. Huo *et al.*, "Data-driven coordinated voltage control method of distribution networks with high dg penetration," *IEEE Transactions on Power Systems*, vol. 38, no. 2, pp. 1543–1557, 2023 .

[25] X.-S. Yang, "Optimization algorithms," *Introduction to Algorithms for Data Mining and Machine Learning*, pp. 45–65, 2019.

[26] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Finite Time Adaptive Stabilization of LQ Systems," *IEEE Transactions on Automatic Control*, vol. 64, no. 8, pp. 3498–3505, 2018.

[27] M. K. Shirani Faradonbeh, A. Tewari, and G. Michailidis, "Input perturbations for adaptive control and learning," *Automatica*, vol. 117, p. 108950, 2020.

[28] M. K. Shirani Faradonbeh, A. Tewari, and G. Michailidis, "On adaptive Linear–Quadratic regulators," *Automatica*, vol. 117, p. 108982, 2020.

[29] J. Yu, Y. Weng, and R. Rajagopal, "Robust mapping rule estimation for power flow analysis in distribution grids," in *North American Power Symposium (NAPS)*, 2017.

[30] M. Mohammadkhani, F. Bayat, and A. A. Jalali, "Robust Output Feedback Model Predictive Control: A Stochastic Approach," *Asian Journal of Control*, vol. 19, no. 6, pp. 2085–2096, 2017.

[31] S. Wang *et al.*, "Assessing gaussian assumption of pmu measurement error using field data," *IEEE Transactions on Power Delivery*, vol. 33, no. 6, pp. 3233–3236, 2017.

[32] Y. Liao *et al.*, "Urban mv and lv distribution grid topology estimation via group lasso," *IEEE Transactions on Power Systems*, vol. 34, no. 1, pp. 12–27, 2019.

[33] T. L. Lai, "Asymptotically efficient adaptive control in stochastic regression models," *Advances in Applied Mathematics*, vol. 7, no. 1, pp. 23–45, 1986.

[34] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *Annals of Statistics*, vol. 36, no. 3, pp. 1171–1220, 2008.

[35] Z.-S. Hou and Z. Wang, "From model-based control to data-driven control: Survey, classification and perspective," *Information Sciences*, vol. 235, pp. 3–35, 2013.

[36] Z. Wu *et al.*, "Control Lyapunov-Barrier function-based model predictive control of nonlinear systems," *Automatica*, vol. 109, p. 108508, 2019.
[37] J. Hu *et al.*, "A survey on sliding mode control for networked control systems," *International Journal of Systems Science*, vol. 52, no. 6, pp. 1129–1147, 2021.

APPENDIX A

A. Proof of Theorem 1

Proof. First, we define preliminary concepts for our proof. Let Θ be the space containing all possible values of E to which the LSE can converge. We will show that E is a zero-dimensional affine subspace that contains E^*, where $L(E^*)$. Next, we introduce the following lemma.

Lemma 1. *For a controller with sublinear regret, Θ identified by LSE forms an affine subspace Θ_0 with a translation of E^* and of dimension $dim(\Theta_0) = 0$.*

Using Lemma 1, we have $dim(\Theta^*) = 0$ implying $\Theta = \{E^*\}$, which proves the consistency of estimating E^*. Thus, Lemma 1 guarantees that E converges to E^*, and by Proposition 1 we can uniquely determine $L(E^*)$. □

B. Proof of Lemma 1

Proof. First, we note that $D = I + EL$; thus, the estimate of E in (10) is equivalent to the estimation of D in (8). For an arbitrary E, based on condition 2, $D(E)$ is an unbiased estimate of $D(E^*)$. Then let $N(E)$ form the space of all E so that the closed-loop matrix $D(E)$ is equal to D^*.

$$N(E) := \{E \in R^{r\times p} : D(E) = D(E^*)\}.$$

Also, let $U(E)$ form the space for all E such that the feedback matrices $L(E)$ are equal to $L(E^*)$.

$$U(E) := \{E \in R^{r\times p} : L(E) = L(E^*)\}.$$

The subspace $S(E)$ forms all possible values of E such that $L(E)$ is the optimal regulator. To prove consistency, we show that $U(E) \cap N(E) = E^*$. For an arbitrary $E \in N(E)$, we have $D^* = I + E^*(L(E^*))$.

Let $L = L(E^*) + \epsilon\tilde{L}$, where $\tilde{L} \in R^{r\times p}$, such that L is a linear feedback matrix, which stabilizes the closed-loop dynamics system. Applying L to the system yields the following Lyapunov equation [26].

$$P(\epsilon) - (I + EL)^\top P(\epsilon)(I + EL) = S + L^\top RL,$$

where P is the solution to the Lyapunov equation. For $\epsilon = 0$, P(0)=$K(E^*)$, where $K(E^*)$ in (14), and the regret of the optimal policy, $\mathcal{R}_\pi = tr(P(\epsilon)C)$. Then we take the derivative of the Lyapunov equation, which leads to the following.

$$\frac{d}{d\epsilon}\tilde{L} - D^\top \frac{d}{d\epsilon}\tilde{L}D = \tilde{L}^\top Z + Z^\top \tilde{L}, \tag{22}$$

where $Z = RL(E^*) + E^\top K(E^*)D^*$. Now, as per the definition of U, for $E \in U(E)$, $\tilde{L}(E)$ must be optimal linear feedback control achieving the optimal regret. This implies $\frac{d\mathcal{R}_\pi}{dw} = tr(\delta\tilde{L})C = 0$. leading to $Z = 0$, which implies

$$D_0^\top K_0(B - E^*) = -L^\top(E^*)R - D_0^\top K_0 E^*. \tag{23}$$

$D^\top K_0$ and $(E - E^*)$ form an affine system. Now, (23) is a necessary condition for $E \in \Theta$. Suppose that B satisfies (23), then the previous results show that $E \in N(E)$. Thus, it suffices to show that $E \in S(E)$. That is given (23), then E solves the optimal Riccati equation (14), and $K(E) = K(E^*)$. Let $Y = E - E^*$, $H = E^\top K(E^*)$, $M = E^\top K(E^*)E + R$, and $S = E^\top K(E^*)Y + (E^\top K(E^*)Y)^\top + Y^\top EY$. We will show that for the following system defined by E, $K(E^*)$ corresponding to the optimal solution based on the true parameters is a solution. To show $E \in S(E)$, let T be defined as a solution of (14). Then we have,

$$T = S + K(E^*) - K(E^*)(E^\top K(E^*)E + R)^{-1}EK(E^*).$$

Then substituting the values of YH, M, S. We obtain $T = K(E^*)$ which solves the Riccati equation and is unique per Proposition 1. Hence, $K(E) = K(E^*)$, which implies $E \in S(E)$. Next, we show that the dimension of the affine subspace formed by (23) is equal to zero, which implies that it contains a single point only proving consistency.

$$dim(\Theta) = (p - rank(K(E^*)D^*))r$$

Since S is positive definite, $rank(K(E^*)) = p$. We show that the rank of $D^* = p$. Using (14), $M = E^\top K(E^*)E + R$, we rewrite $D^* = (I_p - E^*M^{-1}E^\top K(E^*))$, which has a rank of p due to the positive definiteness of R. Thus, $dim(\Theta) = (p - p)r = 0$ completing the proof. □