Data-Driven Model Predictive Control of DC-to-DC Buck-Boost Converter

KRUPA PRAG, MATTHEW WOOLWAY, AND TURGAY CELIK (Member, IEEE)

1 School of Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg 2000, South Africa
2 Department of Mathematics, Imperial College London, London SW7 2BX, U.K.
3 Faculty of Engineering and the Built Environment, University of Johannesburg, Johannesburg 2000, South Africa
4 School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg 2000, South Africa
5 Wits Institute of Data Science, University of the Witwatersrand, Johannesburg 2000, South Africa

Corresponding authors: Krupa Prag (krupa.prag@gmail.com) and Turgay Celik (celikturgay@gmail.com)

ABSTRACT The data-driven model predictive control (DDMPC) scheme is proposed to obtain fast convergence to a desired reference and to be utilised to mitigate the destabilising effects that a DC-to-DC buck-boost converter (BBC) with an active load experiences. The DDMPC strategy uses the observed state to derive an optimal control policy using a reinforcement learning (RL) algorithm. The employed Proximal Policy Optimisation (PPO) algorithm’s performance is benchmarked against the PI controller. From the simulated results obtained using the MATLAB Simulink solver, the most robust methods for short settling time and stability were the hybrid methods. These methods take advantage of the short settling time provided by the PPO algorithm and the stability provided by the PI controller or the filtering mechanism over the transient time. The source code for this study is available on GitHub to support reproducible research in industrial electronics society.

INDEX TERMS Adaptive control, data-driven model predictive control, DC-to-DC buck-boost converter, proximal policy optimisation, reinforcement learning.

I. INTRODUCTION

The popularity of the data-driven model predictive control (DDMPC) scheme has recently increased due to it being naturally suitable for achieving the objectives in model predictive control (MPC), which can handle non-linear system dynamics and hard constraints, whilst taking performance criterion into account. An attractive characteristic of DDMPC is that an accurate model of the plant is dispensable as it instead utilises the plant’s observational data to learn an optimal policy and make informed predictive decisions using the control feedback mechanism [1]. In contrast, model-based predictive control schemes require accurate modelling of the physical model of the considered plant using first principles, which may either be infeasible or even if these models are available, they may be intractable for controller designs due to their complexity. For the DDMPC scheme, the functions of plant modelling, control design and the optimisation thereof are all encapsulated through learning from the observational data received from the plant [2].

Given the promise of DDMPC, it is introduced to address challenges in controlling a DC-to-DC buck-boost converter. This paper seeks to investigate the DDMPC framework’s performance relative to the steady-state’s settling time and the ability to maintain the steady-state in a DC-to-DC buck-boost converter.

In recent times, DC power systems have predominantly been chosen over AC power systems in various applications, given their reliability and quality. Most modern electronic loads are DC in nature as DC has been a standard choice for microgrid (MG) designs [3]–[6]. Furthermore, there has been considerable attention drawn to the generation of power from renewable energy resources. These systems demand advance control schemes to fully tap their potential of either extracting the maximum power by adjusting the load to match the voltage of the source, such as maximum power point tracking [7], [8], or to maintain a constant supply of power to a passive load [9], [10]. As a result, the study of a buck-boost converter, a type of DC-to-DC converter that regulates the voltage from a source to a load, has gained traction. The load of the buck-boost converter may either be passive or active. These converters are analogous to step-up, and step-down

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
transformers as the desired output voltage are less than or greater than the input or source voltage.

A buck converter steps down the voltage from the source to the load; hence, the magnitude of the output voltage is less than that of the source voltage. A boost converter steps up the voltage from the source to the load; consequently, the magnitude of the output voltage is greater than that of the source voltage. A buck-boost converter has an output voltage that is either less than or greater than the source voltage in magnitude.

A transistor switch, an inductor, and a smoothing capacitor are connected within the basic buck-boost converter circuit to smooth out switching noise into regulated DC voltages. The potential of DC-to-DC buck-boost converters is compromised by the destabilising effects on the circuit, resulting in severe voltage and frequency oscillations [4], [11], [12]. The reason for this instability is caused by the existence of a limit cycle in the switch model. The DC-to-DC buck-boost converter’s formulation details are given in Section II-A.

In the panorama of past to present literature, the buck-boost converter with a passive or a constant power supply (CPL) has been a predominant topic of study compared to that of the buck-boost converter with an active or variable power load (VPL). The motivation behind this study is to support the research of optimising the use of renewable sources of energy for the production of electrical energy in an attempt to save fossil fuel energy resources which effectively translate into a reduction of carbon emissions in the atmosphere. Various researchers have been investigating the aims of the paper. The experimental results section, Section IV, details the experimental procedure and settings used in conducting the experiments, followed by the results and result analysis. Section V is the concluding section.

II. BACKGROUND AND RELATED WORK

This section presents the DC-to-DC buck-boost converter’s formulation details and the general feedback control mechanism in conjunction with traditional PID controller and DDMPC systems, including an overview of the evolution of feedback control systems from MPC to DDMPC, and discussed RL based controllers. A review of related work on applying controller designs to mitigate voltage instability in a buck-boost converter imposed by CPL and a passive load with a fixed resistance is given.

A. DC-TO-DC BUCK-BOOST CONVERTER

A diagram of a DC-to-DC buck-boost circuit comprising of a MOSFET switch, a diode, an inductor $L$, and an output capacitor $C$, and a load $R$ is shown in Fig. 1. The inductor and capacitor’s parallel configuration in the circuit acts as a second-order low pass filter reducing the voltage ripple at the output. The corresponding descriptions of the DC-to-DC buck-boost converter circuit components are tabulated in Table 1.

![DC-to-DC buck-boost converter circuit](image)

**FIGURE 1. DC-to-DC buck-boost converter circuit.**

**TABLE 1. Description of DC-to-DC buck-boost converter circuit components.**

| Component    | Description                                      |
|--------------|--------------------------------------------------|
| MOSFET Switch| The default state of the switch is on, unless some threshold is met. |
| Diode        | Allows flow current with ease in one particular direction. |
| Inductor $L$ | Stores electric current which flows through it. |
| Capacitor $C$| Stores electrical energy in an electric field. |
| Resistor $R$ | Reduces current flow by implementing electrical resistance. |

The output voltage in the buck-boost converter is regulated using pulse width modulation (PWM) pulses, which are given to the gate of the MOSFET switch. The switch-mode of the circuit affects the indirect transfer of energy between the inductor and the output capacitor. The desired output voltage of a buck-boost converter is adjustable based on the duty cycle $D$ of the switching transistor [14]. The duty cycle is the ratio between the pulse width, the elapsed time between the rising and falling edges of a pulse, and the total period of a rectangular waveform. The control input of a buck-boost converter is bounded by zero and one, which are the on and off states of the MOSFET switch in the circuit.

The measured output voltage $V_{out}$ over the load in a circuit, with an inverting converter topology, is of reversed polarity to...
that of the input or source voltage \( V_{in} \) due to how the inductor discharges charge [15]. The output voltage \( V_{out} \) over the load in a buck-boost converter is defined as follows:

\[
V_{out} = -\frac{D}{1-D} \cdot V_{in},
\]

(1)

where \( V_{in} \) is the input or source voltage and \( D \) is the duty ratio.

Comparing buck and boost mode, it is found that the duty ratio \( D \) is greater in boost mode than in buck mode. In boost mode, the switch’s on-state is on for a longer duration than when in buck mode, thus storing more energy in the inductor, which prevents a rapid change in current to be passed to the capacitor. The output voltage is increased when enough energy is built up in the inductor and transferred to the capacitor.

In the buck-boost converter circuit, the flow of charge in the circuit is determined by the MOSFET switch state, and the diode controls the direction of the flow of charge. When the MOSFET switch is in the buck-boost converter is in the on-state in the initial cycle, the circuit is closed. In this state, current flows to only the inductor as the input voltage source is directly connected to the inductor, and the diode prevents current from flowing to the output of the circuit as the diode is reversed biased. Furthermore, while the circuit is closed and the MOSFET switch is in the on-state, the inductor accumulates charge and stores energy in the form of a magnetic field. When the MOSFET switch is in the off-state, the diode will allow current to flow from the inductor to the rest of the components of the circuit [16]. While in this state, the inductor’s polarity is reversed, and the diode is forward biased. The inductor provides the energy stored and works as the source, allowing current to flow from the inductor to the capacitor and the load. In this state, the capacitor now accumulates charge and stores energy.

When the MOSFET switch is in the off-state, the inductor experiences a sudden drop in current, thus inducing a voltage to the output. In this state, the change in the inductor’s current \( i_L \):

\[
\frac{di_L}{dt} = \frac{V_{in}}{L} - i_L \frac{R_{on} + r_L}{L}.
\]

(2)

is the difference between the input voltage \( V_{in} \), and the product of the inductor’s current \( i_L \) and the sum of the resistance when the switch is on \( R_{on} \) and the resistance of the inductor \( r_L \), which is divided by the inductance \( L \) [17]. The change in the capacitor’s voltage \( v_c \):

\[
\frac{dv_c}{dt} = \frac{v_c}{RC},
\]

(3)

is the capacitor’s voltage, and the reciprocal of the product of the resistance \( R \) and the capacitance \( C \).

The state-space representation of the system in the off-state is given by:

\[
\dot{x} = A_1 x + B_1 V_{in},
\]

\[
y = C_1 x,
\]

(4)

where \( A_1 \), \( B_1 \) and \( C_1 \) are the system matrices, \( x \) the state variable, \( \dot{x} \) the derivative of the state variable, and \( y \) the output which are obtained using Eqns. (2)-(3) and defined as:

\[
A_1 = \begin{bmatrix} -\frac{R_{on} + r_L}{L} & 0 \\ 0 & -\frac{1}{RC} \end{bmatrix},
\]

\[
B_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad x = \begin{bmatrix} i_L \\ v_c \end{bmatrix},
\]

\[
C_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad y = V_{out} = v_c.
\]

(5)

In the consecutive cycles, where the MOSFET is in the on-state, the capacitor then supplies energy to the load [17]. The state-space representation for when the MOSFET switch is in the on-state is given by:

\[
\dot{x} = A_2 x + B_2 V_{in},
\]

\[
y = C_2 x.
\]

(6)

where \( A_2 \), \( B_2 \) and \( C_2 \) are the system matrices, \( x \) the state variable, \( \dot{x} \) the derivative of the state variable, and \( y \) the output, which are respectively defined as follows:

\[
A_2 = \begin{bmatrix} \frac{r_L}{RC} & \frac{1}{L} \\ -\frac{1}{C} & \frac{1}{RC} \end{bmatrix},
\]

\[
B_2 = \begin{bmatrix} 0 \\ \frac{1}{L} \end{bmatrix}, \quad x = \begin{bmatrix} i_L \\ v_c \end{bmatrix},
\]

\[
C_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad y = V_{out} = v_c.
\]

(7)

The average matrices \( A \) and \( B \) and \( C \) obtained using Eqn. (5) and Eqn. (6) are given by:

\[
A = DA_1 + (1 - D)A_2
\]

\[
= \begin{bmatrix} -\frac{D(R_{on} + r_L)}{L} & \frac{(1 - D)r_L}{RC} \\ -\frac{D}{L} & \frac{(1 - D)C}{RC} \end{bmatrix},
\]

(8)

\[
B = DB_1 + (1 - D)B_2 = DB_1,
\]

(9)

\[
C = DC_1 + (1 - D)C_2 = C_1 = C_2.
\]

(10)

The steady-state is given by:

\[
X = -A^{-1}BV_{in}
\]

\[
= \begin{bmatrix} \frac{DV_{in}}{(1 - D)^2 R + D(R_{on} + r_L) + \frac{(1 - D)r_L L}{RC}} \\ -\frac{D(D - 1)DV_{in}}{R} \end{bmatrix},
\]

(11)

which is the product of the average state variables, inverse of \( A \) and \( B \), and the input voltage \( V_{in} \).

The transfer function \( G \) from the input voltage \( V_{in} \) to the output capacitor \( v_c \) [17] is given by:

\[
G_{V_{in}}(s) = \frac{\frac{D(D - 1)}{LCs^2 + (1 - D)^2}}{R^2 + \frac{Ls}{C} + (1 - D)^2},
\]

(12)
which is comprises of the duty ratio $D$, inductance $L$, capacitance $C$ and resistance of the load $R$.

The following description can summarise the DC-to-DC buck-boost converter’s operations: the inductor gets charged through the voltage source, and the capacitor powers the load. Hence the supply of energy to the load remains uninterrupted irrespective of the state of the MOSFET switch [16].

**B. FEEDBACK CONTROL SYSTEMS**

A *feedback loop* is a powerful tool used in control systems. It considers the plant’s output and enables the system to iteratively adjust the input into the system to meet the desired output response. A simple feedback loop is illustrated in Fig. 2. Sensors are used to measure the plant’s current state $s_t$, at time $t$. The controller is then fed the current state $s_t$ and the error value $e_t$, which is the difference between the current state and the reference state. Using this information, the controller determines the actuation $a_t$ to be applied to the plant. The actuation applied updates the state of the plant.

![Feedback control system](image)

**FIGURE 2. Feedback control system.**

Control systems that follow the basic feedback control structure are PID controller, MPC systems and DDMPC systems.

**C. PROPORTIONAL INTEGRAL DERIVATIVE CONTROL**

PID controller technology employs a feedback control loop mechanism to reduce the effect that disturbances have on the system, steer the plant towards the desired state and creates well-defined relations between variables in the system [18].

A PID controller takes the error at time $t$, $e_t$, as an input. The error is the difference between the measured and the reference value. The output returned by the PID controller is the actuation, $a_t$, which is the action to be applied to the plant or considered system. The control signal or the actuation is equal to the sum of either all or some of the following three terms, where some terms may be zero-valued: the proportional gain $K_p$ multiplied by the magnitude of the error; the integral gain $K_i$ multiplied by the integral of the error; the derivative gain $K_d$ multiplied by the derivative of the error. The generic PID controller shown in Fig. 3 is given by:

$$a_t = K_p e_t + K_i \int_0^t e_i\,dt + K_d \frac{de_t}{dt}$$

(13)

The PID controller returns a control signal $a_t$, which is the sum of the $P$, $I$, and $D$ terms. The characteristics of these terms are:

- **Proportional gain $K_p$:** The control signal proportionally increases with respect to the error to reduce the steady-state process error. This error correction is based on the present steady-state error.
- **Integral gain $K_i$:** The proportional gain $K_p$ may cause oscillation from quick reactions. The integral gain $K_i$ increases the control signal with respect to the past accumulation of the steady-state error.
- **Derivative gain $K_d$:** The derivative gain $K_d$ adds the ability to anticipate future error. Considering the rate of error change, if this change in error increases, this term would add damping to the system to prevent it from overshooting. This term does not affect the steady-state error.

**D. MODEL PREDICTIVE CONTROL**

MPC technique is an advanced feedback control algorithm that uses a model of a plant to forecast behaviours by solving an online optimisation problem to select the most suitable control action, such that the plant being acted upon is driven towards the target. The MPC model is given both the target of the system and the mathematical model of the plant. The number of predicted future steps or the time window into the future, over which control actions are predicted, is known as the prediction horizon. Control actions are computed using a control optimisation algorithm to solve an open-loop optimisation problem over a prediction horizon [19].

A summary of the MPC model design is shown in Fig. 4, which represents an iterative process of updating the calculated control actions over a prediction horizon. The Reference or target, and the predicted control actions are the Inputs to the Plant. Due to Disturbances caused by independent variables, the system may not behave as expected. As a result, the updated state of the physical system, Output, are compared to the model representation of the system, Dynamic Model. The difference between the model and the actual plant is then used to update the calculated control actions to be applied by the MPC Controller. This process is repeated multiple times to get the system acted upon to behave as described by the Reference state.

A challenge associated with solving MPC problems is that it is dependent on the accuracy of the physical system.
realisation (model) of industrial systems that generally have multiple degrees of freedom [20], [21]. Another challenge impeding the performance of MPC systems pertains to the optimal controller, the limitation of the number of recorded measurements of the observable state and the constrained number of actuations available. Given the explosion of data obtained from high fidelity simulations, the accessibility of hardware and faster processing computers that can now be obtained at affordable prices, DDMPC models have been proposed to replace multidimensional model representations and improve the performance of controllers [22].

### E. DATA-DRIVEN MODEL PREDICTIVE CONTROL

DDMPC uses the observation or measured data of the plant with intelligent learning techniques to improve characterising the system model, learning the control policy, determining a combination of sensor placement, and the array of actuations to be made available.

DDMPC is a particular extension of the MPC method that has gained traction, given its efficiency in formulating stochastic MPC problems and autonomously improving a repetitive task's performance. This removes the unrealistic expectation of curating a perfect model of a physical system that incorporates both the system’s complex dynamic characteristics and encapsulates disturbances and uncertainty in the model through the cumbersome process of anticipating and incorporating a discrete number disturbance scenario-based models.

The proposed method by DDMPC is to use data, recent results from past iterations, to improve both safety and performance of the autonomous system by using historical data to represent disturbance variables, that is, the information being propagated through the dynamic system. Two particular challenges faced by controllers that can be addressed using historical data are ensuring recursive feasibility and obtaining optimality despite a short prediction horizon and satisfying input state constraints in the presence of uncertainty. The utilisation of historical data in overcoming challenges associated with solving the stochastic MPC [23]–[26].

The DDMPC scheme learns the optimal control policy using RL is discussed in Section II-E1.

#### 1) REINFORCEMENT LEARNING FOR CONTROL SYSTEMS

RL is a model-free framework that can be used to solve optimal control problems. As per the general feedback control structure described, the controller receives feedback from the plant in the form of a state signal and takes action in response. Similarly, the decision rule is a state feedback control law called the policy in RL [27]. The applied action changes the system’s state, and the latest transition to the updated state is evaluated using a reward function. The objective of the optimal control is to maximise the cumulative reward from each initial state. Given that this is a sequential decision-making process, the problem becomes to maximise the system’s long-term performance.

There exist several powerful model-free RL algorithms. This paper particularly considers a policy gradient method, proximal policy optimisation (PPO) algorithm, which directly optimises policy parameters from the observed data [27].

#### F. CONTROLLER FOR THE DC-TO-DC BUCK-BOOST CONVERTER

The challenge of mitigating voltage instability imposed by a CPL or a passive load with a constant resistance on the DC MGs has been studied and documented in the literature. Control techniques to overcome this challenge have evolved from model-based schemes to model-independent schemes with controller systems ranging from traditional state feedback controllers to, more recently, using machine intelligence techniques.

Common power converter controllers for buck-boost converters with CPL include state-feedback controllers [28], [29], proportional-integral-derivative (PID) control [30]–[32], model predictive control (MPC) [33]–[35] and sliding model control (SMC) [36].

Recent work on voltage regularisation and stabilisation using intelligent controllers include using quadratic D-stable fuzzy controller [37], fuzzy-PID [38], machine learning and reinforcement learning techniques such as Deep Deterministic Policy Gradient (DDPG) [39], Proximal Policy Optimisation (PPO) and Ultralocal Model (UML) control [40], and deep reinforcement learning (DRL) such as Markov Decision Process (MDP), deep Q network (DQN) algorithms [41] and deep deterministic policy gradient based UML [42].

### III. METHODOLOGY

The general framework of the DDMPC scheme, which uses the feedback mechanism as its base structure, takes the observed voltage read over the load or the resistance as the feedback signal. The determined control action is the state of the MOSFET switch in the DC-to-DC BBC converter. The series of control signals sent to the MOSFET switch in the DC-to-DC BBC converter.

The general framework of the DDMPC scheme, which uses the feedback mechanism as its base structure, takes the observed voltage read over the load or the resistance as the feedback signal. The determined control action is the state of the MOSFET switch in the DC-to-DC BBC converter. The series of control signals sent to the MOSFET switch in the DC-to-DC BBC converter.
The DDMPC scheme’s RL policy considered is the PPO algorithm. Furthermore, two hybrid cases are considered.

**A. PROPORTIONAL INTEGRAL CONTROL**

The details of a general PID controller are discussed in Section II-C. For the DC-to-DC buck-boost converter, a PI controller is used to eliminate the steady-state error $e_t$ and reduce the forward gain. The pulse generator generates rectangular wave pulses in a duty cycle. The integrator integrates the proportional gain over the current time step. This value is then subtracted from the output voltage value and then fed into a relay function, allowing its output to switch between two states. The relay function compares the input to a threshold value to determine which corresponding actuation output the controller should return. A summary of the PI controller is given in Fig. 5.

![Figure 5. PI controller for DC-to-DC buck-boost converter.](image)

The applied PI controller’s design uses only the output voltage value from the DC-to-DC buck-boost converter and does not consider the dynamics of the model, hence making this a model-free PI controller implementation.

**B. PROXIMAL POLICY OPTIMISATION ALGORITHM**

PPO algorithm is a model-free, online, or on-policy RL method employed in the DDMPC scheme. The algorithm entails using small batches of experiences from interacting with the environment to update the decision making policy. Iteratively, once the policy is updated, past experiences are discarded, and a new batch is generated to update the policy.

The PPO algorithm is a class of policy gradient training methods that try to reduce the gradient estimations’ variance towards better policies, causing consistent progress and ensuring that the policy does not drastically change from the previous policy or go down irrecoverable paths [40].

The PPO algorithm alternates between sampling data through interacting with the environment and optimising a clipped surrogate objective function which employs stochastic gradient ascent [43]. The stability of training the agent is improved by utilising a clipped surrogate objective function and limiting the size of the policy change at each iteration [44].

The PPO algorithm maintains two function approximators, the actor and the critic networks.

- The actor network maps action choices directly to the observed state. At any particular time, $t$, the actor takes the observed state $s$ and returns the probability of taking action $a$ in the action space when in this state. However, this function does not measure how good an action is compared to the other available actions; hence, the critic network is employed to critique the actions returned by the actor network.

- The critic takes the observed state $s_t$ and the action $a_t$ returned by the actor network as an input and returns the discounted long-term reward’s corresponding expectation. The critic network is trained to predict the value function shown by Eqn. (17), which measures how good it is to be in a specific state $s_t$.

The actor-critic network aims to maximise the surrogate objective function:

$$L(\theta) = \mathbb{E}[\min\left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)],$$

which is an expectation function of the advantage function by $\hat{A}$, policy parameters $\theta$, and the probability ratio $r_t(\theta)$ which is defined as:

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta\text{old}}(a_t|s_t)}.$$

which is the ratio between the current policy and the policy based on past experiences. The comprehensive definition of the probability ratio is: a ratio between the probability taking an action $a$ when in state $s$ at time $t$, given the policy parameters $\pi_\theta$, and the probability of taking action $a$ when in state $s$ at time $t$ using the past or old policy parameters $\theta_{\text{old}}$ from the previous epoch.

The general algorithmic structure of the PPO algorithm [40] is as follows:

1. For the PPO algorithm, the first step is to initialise the parameters of the actor-critic network.
2. The next step is to generate $N$ experiences:

   \[\{s_1, a_1, r_1\}, \{s_2, a_2, r_2\}, \ldots, \{s_N, a_N, r_N\}\]

   the sequence of experiences consist of a tuple of the state-action pair and their corresponding reward value.
3. Calculate the action-value function and the advantage function for each time instance $t$.

- For each instance, the action-value function and the advantage functions are computed at each time step $t$. The action-value function is defined as the expected return of starting at state $s$ and taking action $a$ following the policy $\pi$ is given by:

$$Q^\pi(s, a) = \sum_t E_{\pi_\theta} [R(s_t, a_t) | s, a],$$

where the function is the sum of the expected rewards given the corresponding state-action pair.

- The value function is the expected return of how good it is to be in a particular state, is shown by:

$$V^\pi(s) = \sum_t E_{\pi_\theta} [R(s_t, a_t) | s].$$
where the function is the sum of the expected rewards given the state. The Advantage function, given by:

$$A^\pi (s, a) = Q^\pi (s, a) - V^\pi (s), \quad (18)$$

is the difference between the action-value function \(Q\) and the value function \(V\).

4) Over \(K\) epochs, learn from the mini-batch experiences.
   - Randomly sample a set of \(M\) experiences to form part of the mini-batch which is used to estimate the gradient.
   - The critic network’s parameters can be updated using critic loss function \(L_c\):

$$L_c (\theta_c) = \frac{1}{M} \sum_{t=1}^{M} (Q^\pi (s, a) - V(s(\theta_c)))^2, \quad (19)$$

which minimises the loss over the sampled mini-batch.
   - The actor network’s parameters are updated by:

$$L_a (\theta_a) = -\frac{1}{M} \sum_{t=1}^{M} \min \left( r_t (\theta) \hat{A}_t, \right.$$

$$\left. \text{clip}(r_t (\theta), 1-\epsilon, 1+\epsilon) \hat{A}_t \right \}, \quad (20)$$

which minimises the loss \(L_a\) over the sampled mini-batch.

5) Repeat steps (2) through to (4) until the terminating criterion is met.

Sampling actions train the PPO based RL agent according to the updated stochastic policy; hence it is considered a stochastic policy trained in an on-policy manner. During the initial stage of training, the state-action space is explored through randomly selecting actions. As policy training progresses, the policy becomes less random, and the update rule exploits actions found to yield higher rewards.

During training, the PPO agent estimates and associated probabilities of taking each action in the action space. An action is randomly selected based on the probability distribution over actions. The actor and critic properties are updated after training over multiple epochs, using mini-batches, as the PPO agent interacts with the environment. The PPO agent aims to train the coefficient of the actor-critic neural networks’ coefficient to reduce the error \(e\) between the desired output \(V_{\text{ref}}\) with the actual value \(V_{\text{out}}\).

### C. HYBRID APPROACHES

The hybrid approach uses the PPO algorithm, discussed in Section III-B, with either a PI controller or a filter which are discussed in Section III-C1 and Section III-C2, respectively.

1) HYBRID I

This hybrid approach applies the PPO algorithm until a stopping condition is met, which is when the output voltage value is greater than or equal to the reference voltage in magnitude,

and then employs the PI controller, as discussed in Section III-A, to determine the actions to be applied to the buck-boost converter.

2) HYBRID II

This hybrid approach conditionally utilises a filtering mechanism with the PPO algorithm. The PPO algorithms are solely used to determine the actions to be applied until the reference voltage is reached, then the filtering mechanism is used to filter the pulse signal dictated by the PPO algorithm. The filter sends a 0 pulse to the buck-boost converter if the stipulated conditional statement is violated, else applies the signal dictated by the PPO algorithm. The reason for applying a 0 pulse if the measured output voltage is greater than the reference voltage is that when the MOSFET switch is open, the inductor dissipates current to the capacitor, which powers the fixed load, thus reducing the stored energy in the inductor and capacitor.

### IV. RESULTS

The DC-to-DC buck-boost converter model’s performance with an active load or fixed resistance, detailed in Section II-A, is analysed using the different applied control techniques, which are discussed in Section III. The procedure of experimentally testing the applied control techniques outlined in Section IV-A, the corresponding experimental results, quantitative results and result analysis are reported for three different cases in Section IV-C. The three cases are for the different reference voltages used in the model, which are 30V, 80V and 110V.

#### A. EXPERIMENTAL PROCEDURE

The setup of the DC-to-DC buck-boost converter and the procedure followed to experimentally compare the performance of the four applied control methods are described in this section.

1) DC-TO-DC BUCK-BOOST CONVERTER

The model of a DC-to-DC buck-boost converter with a passive load used is as per Fig. 1. The corresponding parameters of the circuit are tabulated in Table 2.

| Parameters | Definition | Value |
|------------|------------|-------|
| \(V_{\text{in}}\) | Converter Input Voltage | 48V |
| \(L\) | Inductance | 10\(\mu\)F |
| \(C\) | Capacitance | 40mF |
| \(R\) | Resistance | 100Ω |
| \(V_{\text{ref}}\) | Nominal Bus Voltage | 30/80/110V |

The model was constructed using the computation engine in MATLAB® and Simulink® R2021a. The motivation for using the simulated model rather than a state-space model is...
to take advantage of the native matrix computation engine in MATLAB/Simulink [15].

2) PI CONTROLLER
The PI controller uses the output voltage from the buck-boost converter and the reference voltage value to determine the action signal pulse to be applied to the model. Details of the PI controller are given in Section III-A, and the corresponding parameters used are detailed in this section. The PI controller’s corresponding parameters were selected after performing a grid search. The corresponding optimised PI controller parameters utilised in the experiments are tabulated in Table 3. The MATLAB/Simulink solver used for the PI controller is ODE23 stiff/TR-BDF2).

| Parameter          | Value     |
|--------------------|-----------|
| Gain               | 100       |
| Sample Period      | 1E−5s     |
| Integrator         | 350kHls   |
| Pulse Width Percent | 50%       |

3) PPO
The DDMPC scheme’s RL based controller employs the PPO algorithm. The corresponding details of the PPO algorithm are delineated in Section III-B. The PPO RL agent parameters are consistent for the three different reference voltage cases; 30V, 80V and 110V. The duration of the simulation was 0.3s, with a sample time of 1E−5s.

The actor-critic network architecture is built using three completely connected hidden layers for both the actor network and the critic network. Each of these hidden layers is built using 256 neurons. The non-linear mapping function used in both these networks is a rectified linear unit (ReLU). The output layer of the actor network employs a softmax activation function. The parameters of the PPO algorithm and the neural networks are tabulated in Table 4.

| Parameter                   | Value     |
|-----------------------------|-----------|
| Episodes                    | 1024      |
| Learning Rate of Actor NN   | 0.05      |
| Mini-Batch Size             | 8         |
| Learning Rate of Critic NN  | 0.05      |
| Discount Factor             | 0.995     |
| Experience Horizon Length   | 10        |
| Clip Factor                 | 0.2       |
| Entropy Loss Weight         | 0.05      |
| GAE Factor                  | 0.98      |
| Epochs                      | 5         |

In the PPO implementation, the error value calculated at each sample time instance \( t \) is given by:

\[
e_t = V_{ref} - \text{abs}(V_{out}),
\]

which is the difference between the reference voltage value and the magnitude of the output voltage value. The absolute value of the measured output voltage is used in both the reward and error value calculation; as the output voltage is reversed in polarity to that of the input voltage, as discussed in Section II-A.

At each sample time step \( t \), a vector representing the state \( s_t \) is constructed. The PPO agent measures and calculates the following parameters of the DC-to-DC buck-boost converter model which forms the state \( s_t \): the output voltage \( V_{out} \), the error value \( e_t \) and change in error \( \frac{de}{dt} \), thus the state vector is represented as \( s_t = \{V_{out}, e_t, \frac{de}{dt}\} \). Eqn. (22) shows how the change in value error is calculated:

\[
\frac{de}{dt} = \frac{e_{t-1} - e_t}{t - (t-1)}.
\] (22)

The training of the PPO RL agent uses a fixed number of sample steps \( T \) unless the termination criterion is met. Should the output voltage value exceed \( V_{ref} \) is greater than the upper bound \( u_B \), the training for that episode is terminated. During the training of the PPO agent, at each sample time step \( t \), the PPO RL agent takes the current state \( s_t \) and the awarded reward value \( r_t \) as inputs. The reward function is given in Algorithm 1.

Algorithm 1: PPO Reward Function

Input: \( V_{ref}, V_{out}, e_t, e_{t-1}, \epsilon, V_{ref} \) Reached

1: \( u_B = V_{ref} \cdot (1 + \epsilon) \) \n   \( \triangleright \) Upper Bound
2: \( l_B = V_{ref} \cdot (1 - \epsilon) \) \n   \( \triangleright \) Lower Bound
3: if \( (V_{ref} \text{ Reached} \equiv \text{True } \& \ V_{out} < l_B) \text{ or } (V_{out} > u_B) \)
4: \( r_t = -1 \)
5: else
6: \( r_t = -\frac{1}{\text{abs}(e_t)} \)
7: \( \text{end if} \)

4) HYBRID I
This hybrid approach uses the PPO RL agent with a PI controller, as described in Section III-C1. The parameters used for the RL agent and the PI controller are given in Section IV-A2 and Section IV-A3, respectively. The PI controller is implemented to determine the action to be applied to the buck-boost converter once the magnitude of the absolute voltage exceeds that of the reference voltage, \( \text{abs}(V_{out}) > V_{ref} \).

5) HYBRID II
This hybrid approach uses the PPO RL agent with filter, as described in Section III-C2. The parameters used for the RL agent are given in Section IV-A3. The filter mechanism is applied if the following condition is violated: if the absolute output voltage is greater than the reference voltage \( \text{abs}(V_{out}) > V_{ref} \).

The simulation time for the PI controller is 3s and, for the PPO, Hybrid I and Hybrid II, each of the conducted
experiments are simulated for 0.3s with a fixed sample time of 1E−5s. The corresponding Simulink models and code utilised can be found on GitHub. All experiments were conducted using an AMD RYZEN 3770x @3.6GHz CPU.

B. SETTLING TIME
The settling time is the time elapsed from the instantaneous step to when the outputs of the considered dynamical control system remain within a specified error range. The error range used is 2% of the reference voltage value. Thus, the \( \epsilon \) value used in the reward function, Algorithm 1 is 0.02. In Fig. 6 two time periods of interest for a dynamical control system are highlighted; these are settling time and transient time. All the responses or the observed states of a control system from the end of the simulation duration to the first data point value that does not fall between the error band or the range of accepted values with respect to a reference value make up the transient time. The time taken to reach the transient time is known as the settling time.

![Image of response time of dynamical control system.](image)

**FIGURE 6.** Response time of dynamical control system.

C. EXPERIMENTAL RESULTS
The experimental results for the four different control techniques applied are presented in this section, followed by the analysis of the results obtained.

The results presented in Table 5, record the settling time and the error values for both the entire duration of the simulation time and for the transient duration after the settling time period. For each of the attributes in the results table, the average of the validation experiment values is recorded.

The lowest obtained value for each corresponding quantitative measures is highlighted for the respected reference voltage cases.

The quantitative measurements used to evaluate the applied control techniques’ performance are the real elapsed time, settling time, MSE, mean absolute error (MAE) and integral absolute error (IAE), the standard deviation \( \sigma \) is calculated for each of these attributes, respectively. These results are tabulated in Table 5.

The relationship between time and the output voltage by the buck-boost converter when employing various controllers are presented for the three different reference voltage values, respectively, in Fig. 7.

![Image of buck-boost converter output voltage.](image)

**FIGURE 7.** Buck-Boost converter output voltage for the employed control techniques for reference voltages: (a) 30V, (b) 80V, (c) 110V.

1) PI CONTROLLER
The PI controller’s settling time is proportional to the magnitude of the reference voltage. It is observed that the average settling time is greater than that of the respective reference values for all three cases. The PI controller can be seen as a
TABLE 5. Average quantitative measurements of the applied control techniques for the DC-to-DC buck-boost converter using Algorithm 1.

| Voltage | PID | PPO | Hybrid I | Hybrid II |
|---------|-----|-----|---------|----------|
| 30 V    |     |     |         |          |
| MSE     | 54.9851 | 22.1677 | 4.5075  | 24.3725  |
|         | 9.6574  | 0.8495  | 24.6018  |           |
| MAE     | 2.5224  | 2.9109  | 0.7676  | 1.2874   |
|         | 0.7388  | 0.0753  | 0.8172   |           |
| IAE     | 9.3102  | 0.5241  | 0.1993  | 0.3327   |
|         | 0.2357  | 0.0226  | 0.2480   |           |
| Settling Time (s) | 0.455 | 0.2221 | 0.0504 | 0.2162 |
| percentage | 1.0 | 0.0217 | 1.0 | 0.9583 |
| σ       | -0.0622 | 0.0305  | 0.0608  |           |
| MSE     | 0.0091  | 0.1346  | 0.0173  | 0.1543   |
|         | 0.0453  | 0.0012  | 0.0653  | 0.0078   |
| MAE     | 0.0948  | 0.3465  | 0.1094  | 0.3479   |
|         | 0.0703  | 0.0037  | 0.1199  |           |
| IAE     | 0.2476  | 0.0228  | 0.0391  | 0.0229   |
|         | 0.012   | 0.0029  | 0.0002  | 0.0092   |
| Elapsed Time (s) | 45.65 | 285.72 | 294.48 | 285.53 |

For the entire simulation duration.

For the remaining time of simulation duration after settling time.

stable control technique, based on the numerical output voltage values, the plots illustrating the relationship between time and the output voltage, and the MSE, MAE, and IAE values for the transient time. A disadvantage of the PI controller is that the settling time is the longest compared to the other employed control techniques. The PI controller stabilises the output voltage at a value less than that of the desired voltage; hence it does not necessarily guarantee the lowest error values over the transient time.

2) PPO

The PPO algorithm is found to have the shortest settling time for when the converter is used for the boost cases. The variance and standard deviation values for the corresponding error values are considered when discussing stability. It is found that these values, for the entire simulation duration, are lower than that when the PI or hybrid approaches are used for the boost cases, which can be attributed to the short settling time for these cases. However, if only transient time is considered, the PPO quantitative measurements are not the lowest indicating that this is not the most robust method in terms of stability. From the relationship between time and output voltage illustrated in Fig 7, it can be seen that the PI controller resembles a parabolic decay rate, whilst for the PPO algorithm, exponential decay is seen for the settling time duration.

It is highlighted that the PPO algorithm does not always converge to the desired reference voltage. Hence a 100% of the experiments do not fall within the settling time, as seen for the case when the converter is used for a buck case.

3) HYBRID I

This hybrid approach uses the PPO algorithm and employs the PI controller to determine the actions to be applied to the buck-boost converter once the magnitude of the output voltage reaches or exceeds that of the desired voltage. The shortcoming of the PI controller is that it stabilises to a lower absolute output voltage than the reference voltage; this features for the boost mode instances using this hybrid technique, which was seen for the vanilla PI control method. Taking advantage of the short settling time provided PPO and the stability provided by the PI controller, the error values are significantly smaller than that of both the individual implementations of the PPO and PI controller for the 30V case; for both the entire simulation duration and for that after the settling time, making it a robust method for when the converter is used in buck mode.

4) HYBRID II

Combining the PPO algorithm with a filter mechanism in this hybrid approach, it has been found that this method’s variance and standard deviation values for corresponding error values are generally less than that of the PPO algorithm and Hybrid I for the boost cases. This indicates that this approach is the most robust control method with respect to stability, as the mean settling time values, output voltage value and corresponding quantitative tabulated values substantiate the performance of this control technique.

5) REWARD FUNCTION

In [40] the reward function used is the same as Algorithm 1, whilst in [39] Algorithm 2, both these methods, respectively, have been applied without the conditional statements to similar DC-DC converters.

The results obtained using this alternative reward function for the PPO algorithm is tabulated in Table 6. The results of the applied PPO algorithm to the DC-to-DC buck-boost converter using the reward function defined in Algorithm 1 are tabulated in Table 5, and the results when Algorithm 2 is used is tabulated in Table 6. From these results it is found that the PPO used with the reward function defined in Algorithm 1, has lower settling time and quantitative measurements in comparison to when Algorithm 2 is used, hence was the reward function employed.
Algorithm 2 PPO Updated Reward Function

Input: $V_{\text{ref}}$, $V_{\text{out}}$, $e_{t}$, $e_{t-1}$, $\epsilon$, $V_{\text{ref Reached}}$

1: $u_B = V_{\text{ref}} \cdot (1 + \epsilon)$ \quad \triangleright \text{Upper Bound}
2: $l_B = V_{\text{ref}} \cdot (1 - \epsilon)$ \quad \triangleright \text{Lower Bound}
3: if ($V_{\text{ref Reached}} == \text{True} \& V_{\text{out}} < l_B$) or ($V_{\text{out}} > u_B$) then
4: \hspace{1em} $r_t = -1$
5: else
6: \hspace{1em} $r_t = \frac{1}{\text{abs}(e_{t})^2}$
7: \hspace{1em} end if

TABLE 6. Average quantitative measurements of the applied control techniques for the DC-to-DC buck-boost converter using Algorithm 2.

| PPO | 30V | 80V | 110V |
|-----|-----|-----|-----|
| MSE\(^1\) (V) value | 11.8695 | 366.7747 | 1062.447 |
| MSE\(^1\) (V) | 0.8594 | 78.6585 | 121.3693 |
| MAE\(^1\) (V) | 1.7657 | 10.6412 | 21.9291 |
| MAE\(^1\) | 0.2345 | 1.9567 | 1.6257 |
| IAE\(^1\) (V) | 0.4831 | 2.4841 | 3.5085 |
| Setting Time (s) | 0.0728 | 0.5229 | 0.5431 |
| percentage | 0.0301 | 0.0693 | 0.0587 |
| IAE\(^1\) (V) | 0.2045 | 1.4914 | 2.3692 |
| MSE\(^1\) (V) | 0.059 | 0.4348 | 0.7601 |
| MSE\(^1\) (V) | 0.4401 | 1.1919 | 1.4927 |
| MSE\(^1\) | 0.0731 | 0.1935 | 0.2569 |
| IAE\(^1\) (V) | 0.0129 | 0.1005 | 0.1538 |
| IAE\(^1\) | 0.0107 | 0.0698 | 0.0801 |

\(^1\) For the entire simulation duration.
\(^2\) For the remaining time of simulation duration after settling time.

6) SENSITIVITY TO NOISE

The robustness of the applied controllers to the BBC can be evaluated based on the controller’s performance when experiencing noise. Additive White Gaussian Noise (AWGN) is an added linear noise model applied to the transmitted signal, which has a uniform power across the frequency band for the output signal and has a Gaussian distribution with respect to time. The AWGN channel model is represented by the outputs $H_k$ at discreet time-steps $k$. The sum of the input $F_k$ and noise $G_k$ is the value of $H_k$:

$$H_k = F_k + G_k,$$  \hspace{1em} (23)

where $G_k$ is independently and identically distributed from a zero-mean normal distribution with variance $\sigma$, that is $G_k \sim \mathcal{N}(0, \sigma^2)$. AWGN is added to the transmitted signal to measure and compare the controller’s performance when experiencing such an impairment. The measurement parameter, signal-to-noise ratio (SNR), compares the power of the desired information signal to the power of the undesired signal or background noise, which is denoted as:

$$\text{SNR} = \frac{P_{\text{signal}}}{P_{\text{noise}}}. \hspace{1em} (24)$$

To compare the performance of the controllers an AWGN with a range of SNRs have been applied to the output measured voltage of the DC-to-DC BBC. Table 7 records the quantitative measurement when applying the PI controller with the AWGN model.

Comparing the results in Table 7 to Table 5, it can be seen that as the SNR decreases, both the average error and settling time increases, as a result of the noise signal increasing. The settling time is used to decide the threshold of SNR of the PI controller. The settling time of the PI controller without AWGN remains unchanged for when AWGN with SNR is

\(101912\)
TABLE 8. Average quantitative measurements of the applied control techniques with AWGN for the DC-to-DC buck-boost converter with reference voltages: Table (8a) 30 V, Table (8b) 80 V and Table (8c) 110 V.

| Technique | PI | Hybrid I | Hybrid II |
|-----------|----|----------|-----------|
| PPO       |     |          |           |
| MSE       | 3.1267 | 0.8736   | 0.9307    |
| MAE       | 0.5285 | 0.3217   | 0.3143    |
| SAR       | 0.2032 | 0.2897   | 0.2872    |
| Setting Time | 0.0422 | 0.0246   | 0.0402    |
| percentage | 1.0 | 1.0 | 1.0 |
| SIGMA      | 0.0861 | 0.0028   | 0.0618    |
| MAE        | 0.0416 | 0.0194   | 0.0216    |
| SAR        | 0.0849 | 0.0084   | 0.0112    |
| SIGMA      | 0.1735 | 0.0226   | 0.0446    |
| MAE        | 0.0255 | 0.0050   | 0.0258    |
| SAR        | 0.0151 | 0.0080   | 0.0171    |

For the remaining time of simulation duration, after settling time.

Thus the results indicate a bias towards the signal value, this is a result of the error band being calculated relative to the reference voltage value, the cases with a lower reference voltage are more sensitive to noise. Comparing the performance of the applied controllers with the AWGN model, it can be seen that the PPO and Hybrid II controllers are the most robust when considering their quantitative measurements, particularly the percentage value of episodes that converged and the error values for the settled time duration.

In summary, applying the DDMPC scheme with the PPO algorithm has been found to have a shorter settling time for the boost cases and Hybrid I compared to the PI controller. However, it is found that the hybrid approaches are the most robust in the light of both setting time and stability, as they take advantage of the short settling time provided by the PPO algorithm and the stability ensured by the PI controller and the filtering mechanism. Given that the literature does not document the performance of buck-boost converters with an active load or VPL, there is no direct comparison to previous work in this regard. However, considering similar work [40] where the buck-boost converter has a CPL, and the PI controller is tuned using the PPO algorithm, comparing the error values and the inferred settling time, it is found that the hybrid approaches’ performance is comparable. Furthermore, from observing the impact of the reward function, we find that investigating the impact of the employed reward function and optimising the reward function does hold promise in improving the quality of the results found using the DDMPC techniques. With respect to the robustness of the controllers when experiencing noise, the PPO and Hybrid II controllers were found to be most robust.

V. CONCLUSION

The popularity of renewable energy plants and the increasing number of electronic applications, which are DC in nature, make the study of DC-to-DC buck-boost converters with active loads nascent. The buck-boost converter input converts an input voltage to the desired lower reference output voltage when in buck mode, and to the desired reference with a greater output voltage magnitude, in boost mode. The quality of these converters is based on the settling time to reach the reference voltage and the ability of the controller to maintain a constant output voltage. The impact of the reward function on controllers using the PPO algorithm opens up interesting lines of follow-up research for future development, as well as applying and testing the robustness of the discussed control methods on physical BBC prototype. DDMPC techniques have been considered to improve the quality of these converters. The applied control techniques’ performance to the buck-boost converter was evaluated based on the applied control technique’s short settling time and stability. The PI controller’s performance was used as a benchmark to compare the performance of the vanilla DDMPC technique using the PPO algorithm. The PPO algorithm was found to provide a short settling time to reach the reference voltage and outperformed the PI controller in this respect.

Table 8 records the results of the PPO controller with AWGN when applied to a BBC for respective reference voltages. When the Hybrid I controller with AWGN of 25 dB SNR was applied to a BBC with a reference voltage of 30 V, the results obtained indicated that the simulation was terminated as the stopping conditions described in Algorithm 1 were met before reaching the full simulation duration of 0.3s, greater than or equal to 30 dB is applied. However, when the SNR is set to 25 dB, the settling time exceeds that of the PI controller with no AWGN.
Taking advantage of the short settling time of the PPO method and the stability provided by the PI controller and the filtering mechanism, merit was found in the hybrid techniques as their performance surpass that of the PI controller with respect to settling time and the PPO algorithm with respect stability. Furthermore, the PPO and Hybrid II controllers were found to perform comparatively to the controllers without noise when AWGN was applied to the feedback signal, thus in general the PPO and Hybrid II controllers have merit with respect to short settling time, stability and sensitivity to noise.

REFERENCES

[1] J. Berberich, J. Kohler, M. A. Muller, and F. Allgower, “Data-driven model predictive control with stability and robustness guarantees,” IEEE Trans. Control Syst. Technol., vol. 26, no. 4, pp. 1112–1125, Apr. 2018.

[2] Z. Hou, H. Gao, and F. Lewis, “Data-driven control and learning systems,” IEEE Trans. Ind. Electron., vol. 66, no. 3, pp. 1702–1711, Apr. 2019.

[3] D. Marx, P. Magne, B. Nahid-Mobarakeh, S. Pierfederici, and B. Davat, “Large signal stability analysis tools in DC power systems with constant power loads and variable power loads—A review,” IEEE Trans. Power Electron., vol. 27, no. 4, pp. 1700–1707, Apr. 2012.

[4] S. Huddly and J. D. Skufca, “Amplitude death solutions for stabilization of DC microgrids with instantaneous constant-power loads,” IEEE Trans. Power Electron., vol. 28, no. 1, pp. 247–253, Jan. 2013.

[5] S. Singh, N. Rathore, and D. Fulwani, “Mitigation of negative impedance instabilities in a DC/DC buck-boost converter with composite load,” J. Power Electron., vol. 16, no. 3, pp. 1046–1055, May 2016.

[6] Q. Xu, C. Zhang, C. Wen, and P. Wang, “A novel composite nonlinear controller for stabilization of constant power load in DC microgrid,” IEEE Trans. Smart Grid, vol. 10, no. 1, pp. 752–761, Jan. 2019.

[7] V. C. Kotak and P. T. Balsara, “State-space modelling and control using state space approach,” in Proc. IEEE Conf. Energy Convers., vol. 3297, no. 12, pp. 6115–6125, 2007. [Online]. Available: https://www.ijareeie.com

[8] R. F. Coelho, F. Concer, and D. C. Martins, “A study of the basic DC–DC converters applied in maximum power point tracking,” in Proc. Brazilian Power Electron. Conf., Sep. 2009, pp. 673–678.

[9] N. A. Mohamed, Q. Guan, and M. Rashed, “Control of DC–DC converter for interfacing supercapacitors energy storage to DC micro grids,” in Proc. IEEE Int. Conf. Electr. Syst. Air., Railway, Ship Propuls. Road Vehicles Int. Transpn. Electrific. Conf. (ESARs-ITEC), Nov. 2018, pp. 1–8.

[10] R. Lu, M. R. M. Tavakoli, and T. Sato, “A control method using predicted output voltage for digitally controlled DC/DC converter,” in Proc. Int. Conf. Electr. Control Instrum. Eng. (ICECIE), Nov. 2019, pp. 1–7.

[11] R. D. Bhagya and R. M. Patel, “PWM based double loop PI control of a bidirectional DC–DC converter in a standalone PV/battery DC power system,” in Proc. IEEE 16th India Council Int. Conf. (INDICON), Dec. 2019, pp. 1–4.

[12] T. Kobaku, R. Jeyasenthil, S. Sahoo, R. Ramchand, and T. Dragicevic, “Quantitative feedback design-based robust PI control of voltage mode controlled DC–DC boost converter,” IEEE Trans. Circuits Syst. I, Exp. Briefs, vol. 68, no. 1, pp. 286–290, Jan. 2021.

[13] Q. Xu, Y. Yan, C. Zhang, T. Dragicevic, and F. Blaabjerg, “An offset-free composite model predictive control strategy for DC/DC buck converter feeding constant power loads,” IEEE Trans. Power Electron., vol. 35, no. 5, pp. 5331–5342, May 2020.

[14] N. Bouchich, A. Moufli, N. Bennis, and S. E. Hani, “A constrained MPC approach applied to buck DC–DC converter for greenhouse powered by photovoltaic source,” in Proc. Int. Conf. Electr. Inf. Technol. (ICETI), Mar. 2020, pp. 1–6.

[15] Z. Zhou, L. Zhang, Z. Liu, Q. Chen, R. Long, and H. Su, “Model predictive control for the receiving-side DC–DC converter of dynamic wireless power transfer,” IEEE Trans. Power Electron., vol. 35, no. 9, pp. 8895–8897, Sep. 2020.

[16] J. Fan, S. Li, J. Wang, and Z. Wang, “A GPI based sliding mode control method for boost DC–DC converter,” in Proc. IEEE Int. Conf. Ind. Technol. (ICIT), Mar. 2016, pp. 1826–1831.

[17] M. M. Mardani, N. Vafamand, M. H. Khooban, T. Dragicevic, and F. Blaabjerg, “Design of quadratic d-stable fuzzy controller for DC microgrids with multiple CPLS,” IEEE Trans. Ind. Electron., vol. 66, no. 6, pp. 4805–4812, Jun. 2019.

[18] R. F. Bastos, C. R. Aguier, A. F. Q. Gonçalves, and R. Q. Machado, “An intelligent control system used to improve energy production from alternative sources with DC/DC integration,” IEEE Trans. Smart Grid, vol. 5, no. 5, pp. 2486–2495, Sep. 2014.

[19] M. Ghiesarnejad, H. Farsizadeh, M. R. Tavana, and M. H. Khooban, “A novel deep learning controller for DC–DC buck-boost converters in wireless power transfer feeding CPLS,” IEEE Trans. Ind. Electron., vol. 68, no. 7, pp. 6379–6384, Jul. 2021.
[40] M. Hajibosseini, M. Andalibi, M. Gheisarnejad, H. Farsizadeh, and M.-H. Khooban, “DC/DC power converter control-based deep machine learning techniques: Real-time implementation,” *IEEE Trans. Power Electron.*, vol. 35, no. 10, pp. 9971–9977, Oct. 2020.

[41] C. Cui, N. Yan, and C. Zhang, “An intelligent control strategy for buck DC–DC converter via deep reinforcement learning,” 2020, arXiv:2008.04542. [Online]. Available: http://arxiv.org/abs/2008.04542.

[42] M. Gheisarnejad, H. Farsizadeh, M.-R. Tavana, and M. H. Khooban, “A novel deep learning controller for DC/DC buck-boost converters in wireless power transfer feeding CPLs,” *IEEE Trans. Ind. Electron.*, vol. 68, no. 7, pp. 6379–6384, Jul. 2021.

[43] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” 2017, arXiv:1707.06347. [Online]. Available: http://arxiv.org/abs/1707.06347.

[44] B. Liu, Q. Cai, Z. Yang, and Z. Wang, “Neural proximal/trust region policy optimization attains globally optimal policy,” 2019, arXiv:1906.10306. [Online]. Available: http://arxiv.org/abs/1906.10306.

**KRUPA PRAG** is currently a postgraduate student at the University of the Witwatersrand, Johannesburg, South Africa. She is an Associate Lecturer with the School of Computer Science and Applied Mathematics, University of the Witwatersrand. Her research interests include optimization, optimal control theory, and computational intelligence.

**MATTHEW WOOLWAY** received the Ph.D. degree from the University of the Witwatersrand, Johannesburg, South Africa, in 2018. He is currently a Teaching Fellow in applied mathematics with the Department of Mathematics, Imperial College London, and a Research Associate with the Faculty of Engineering and the Built Environment, University of Johannesburg. His research interests include computational intelligence, artificial intelligence, and optimization.

**TURGAY CELIK** (Member, IEEE) received the second Ph.D. degree from the University of Warwick, Coventry, U.K., in 2011. He is currently a Professor of digital transformation and the Director of the Wits Institute of Data Science, University of the Witwatersrand, Johannesburg, South Africa. His research interests include signal and image processing, computer vision, machine intelligence, robotics, data science and engineering, and remote sensing. He is an Associate Editor of the *ELL (IET), IEEE ACCESS, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (GRSL), IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS), and SIVP* (Springer).

* * *