Advancing Intelligent DC Microgrids

AI-Enabled Control, Cyber security, and Energy Management

Wan, Yihao

Link to article, DOI:
10.11581/DTU.00000319

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

DTU Wind and Energy Systems
Department of Wind and Energy Systems

Advancing Intelligent DC Microgrids: AI-Enabled Control, Cybersecurity, and Energy Management

Ph.D. thesis

Yihao Wan

December 2023, Kongens Lyngby, Denmark

DANMARKS TEKNISKE UNIVERSITET
Division for Power and Energy Systems
(PES)
DTU Wind and Energy Systems

Advancing Intelligent DC Microgrids: AI-Enabled Control, Cybersecurity, and Energy Management

Yihao Wan

A thesis
submitted in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

December 2023, Kongens Lyngby, Denmark

Advancing Intelligent DC Microgrids: AI-Enabled Control, Cybersecurity, and Energy Management

Author:
Yihao Wan

Supervisor:
Tomislav Dragicevic, Professor
Department of Wind and Energy Systems, Technical University of Denmark, Denmark

Nenad Mijatovic, Associate Professor
Department of Wind and Energy Systems, Technical University of Denmark, Denmark

Ramadhani Kurniawan Subroto, Postdoc
Department of Wind and Energy Systems, Technical University of Denmark, Denmark

Dissertation Examination Committee:
Guangya Yang, Associate Professor (Internal examiner)
Department of Wind and Energy Systems, Technical University of Denmark, Denmark

Amjad Anvari-Moghaddam, Associate Professor
Department of Energy, Aalborg University, Denmark

Dmitri Vinnikov, Professor
Department of Electrical Power Engineering and Mechatronics, Tallinn University of Technology, Estonia

Division of Power and Energy Systems (PES)
DTU Wind and Energy Systems
Technical University of Denmark
Elektrovej, Building 325
2800 Kongens Lyngby, Denmark
Tel: (+45) 4677 5085
Email: communication@windenergy.dtu.dk

Release data: December 2023
Class: Public
Field: Electrical Engineering
Remarks: The dissertation is presented to the Department of Wind and Energy Systems of the Technical University of Denmark in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

ISBN: 000-00-0000-000-0

Summary

Microgrids are regarded as one of the cornerstones of the future smart grid, contributing to the development of zero-carbon cities and diverse energy systems. They consist of various distributed generation resources (DERs), energy storage systems (ESSs), and loads, serving as energy routers to enable the grid to withstand major disasters and enhance the energy security and resilience of the country. DC microgrids are distinguished among different microgrid configurations due to their higher efficiency, compatibility with renewable energy sources, scalability, resilience, cost savings, and alignment with the growing use of DC systems.

In DC microgrids, various modules operate both independently and coordinately to achieve different operational objectives. The hierarchical control framework is recognized as a standardized solution for DC microgrids. The local control at the primary control layer manages voltage and current control, as well as preliminary power sharing among different units. The secondary control layer regulates the voltage and ensures accurate power sharing. The tertiary control layer is responsible for advanced control goals, including energy management, economic dispatch, and power flow control, among others. However, factors such as the complex model, parameter availability and accuracy, dynamic operating conditions, and related aspects pose challenges to the existing DC microgrid control system.

This research aims to revolutionize microgrid operations by incorporating advanced artificial intelligence (AI) technologies into the existing multilayer control framework, addressing the aforementioned challenges. The research encompasses three primary areas: 1) at the local control level, a model-free learning-based controller operating in real-time using standard commercial microprocessors, ensuring optimal converter control and enabling independent unit operations; 2) at the secondary control layer, enhancing the system's cybersecurity by identifying and detecting intelligent cyber-attacks using a data-driven approach; and 3) at the system level, the autonomous energy management of the system with an innovative battery degradation modeling approach under dynamically changing operating conditions. Ultimately, this research endeavors to develop an AI-enabled control system that enhances local control, cybersecurity, and energy management of DC microgrids.

At the primary control layer for local control, the interfacing power converter ensures proper operation at the device level. Conventional advanced controllers, such as finite-control-set model predictive control (FCS-MPC), are extensively applied due to their straightforward and flexible formulation of control objectives and constraints. However, this approach suffers from parameter sensitivity, unmodeled dynamics, and time-consuming optimization of the cost function. To address these issues, a guideline

based on reinforcement learning (RL), aiming to automate weighting factor design in FCS-MPC, is introduced. In addition, a model-free self-learning controller for the converter is proposed, achieving comparable control performance to conventional FCS-MPC without requiring prior system knowledge. Furthermore, a safety framework is proposed to enforce physical limits and enhance the learning efficiency of RL agents. These findings offer opportunities for implementing online RL-based controllers in practical power electronics systems.

Furthermore, at the secondary control layer, the communication-based distributed control, widely adopted for module coordination control, potentially exposes DC microgrids to cyber-attacks. Various advanced cyber-attack detectors have been proposed, yet they face critical challenges when confronted with astute adversaries introducing innovative attack patterns. To address this, RL is first employed to identify the vulnerabilities of the system to cyber-attacks by emulating intelligent attackers, and autonomously generate stealthy attack patterns that bypass conventional cyber-attack detectors. Conversely, a data-driven cyber-attack detector is proposed to enhance the capability of state-of-the-art cyber-attack detectors by identifying these attacks. Consequently, an ML-enabled framework for cyber-attack identification and detection is developed. This development opens up possibilities for iteratively implementing the proposed cybersecurity framework, enabling the identification and detection of a broader spectrum of cyber-attacks in DC microgrids.

Ultimately, at the tertiary control layer, accurate modeling of battery degradation is crucial for the economic operation of DC microgrids to minimize the overall operation cost. Different from the conventional methods that employ a single-stage battery degradation model, this research proposes a multi-stage battery degradation modeling method to accurately capture the varying aging patterns throughout a battery's lifetime, thereby enabling more precise evaluations of capacity losses under diverse operating conditions. In addition, to adapt to varying operating conditions and mitigate system uncertainty, this project explores an RL-based energy management strategy that takes into account battery degradation for the economic and autonomous operation of DC microgrids.

In conclusion, the contributions of this Ph.D. project include the development of an AI-enabled hierarchical control framework, spanning from model-free optimal power converter control at the primary control layer to a data-driven cyber-attack identification and detection scheme at the secondary control layer, and extending to an RL-based energy management strategy incorporating a multi-stage battery degradation model at the tertiary level. The proposed AI-enabled multilayer control framework is expected to significantly advance the development of more intelligent, secure, and efficient DC microgrids.

Resume

Microgrids betragtes som en af hjørnestenene i det fremtidige smart grid, og bidrager til udviklingen af nul-karbon byer og forskellige energisystemer. De består af forskellige distribuerede genereringsressourcer (DERs), energilagringssystemer (ESSs) og belastninger og fungerer som en energirouter, hvilket gør det muligt for nettet at modstå store katastrofer og forbedrer energisikkerheden og modstandsdygtigheden i landet. DC microgrids skiller sig ud blandt forskellige mikronetkonfigurationer på grund af deres højere effektivitet, integration af vedvarende energi, skalérbarhed, modstandsdygtighed, omkostningsbesparelser og overensstemmelse med den stigende anvendelse af DC-systemer.

I DC-mikronet arbejder forskellige moduler uafhængigt og koordineret for at opnå forskellige driftsmål. Den hierarkiske kontrolramme anerkendes som en standardiseret løsning til DC-mikronet. Den lokale kontrol på det primære kontrolniveau håndterer spændings- og strømstyring samt foreløbig fordeling af effekt blandt forskellige enheder. Det sekundære kontrolniveau regulerer spændingen og opnår nøjagtig effektdeling. Det tertiære kontrolniveau er ansvarlig for avancerede kontrolmål såsom energistyring, økonomisk fordeling, strømstyring osv. Men den komplekse model, parameter tilgængelighed og nøjagtighed, dynamiske driftsbetingelser osv. udgør udfordringer for det eksisterende DC-mikronet kontrolsystem.

Denne banebrydende forskning sigter mod at revolutionere mikronetdrift ved at inkorporere avancerede kunstig intelligens (AI) teknologier i det nuværende flerlag kontrolramme for at adressere de ovennævnte udfordringer. Forskningen fokuserer på tre områder: 1) På det lokale kontrolniveau opererer en model-fri controller i realtid ved hjælp af standard kommercielle mikroprocessorer til optimal konverterkontrol for at realisere uafhængig drift af enheder; 2) På det kommunikationsbaserede sekundære kontrolniveau, forbedring af cybersikkerheden i systemet ved at identificere og detektere intelligente cyberangreb med en datadrevet tilgang; og 3) På systemniveauet, økonomisk fordeling af systemet med en innovativ batteridegraderingsmodelmetode i dynamisk skiftende driftsbetingelser for autonome mikronet. Til sidst har denne forskning til formål at udvikle et AI-aktiveret kontrolsystem for at fremme lokal kontrol, cybersikkerhed og energistyring i DC-mikronet.

På det primære kontrolniveau for lokal kontrol sikrer den grænsefladekraftkonverteren korrekt drift på enhedsniveauet. Konventionelle avancerede controllere, f.eks. finite-control-set model predictive control (FCS-MPC), er blevet bredt anvendt på grund af den ligetil og fleksibel formulering af kontrolmål og begrænsninger. Dog lider den af parameterfølsomhed, umodelerede dynamikker og tidskrævende optimal omkostningsfunktionsdesign. For at løse dette, præsenteres først en retningslinje

baseret på reinforcement learning (RL) for at opnå en automatisk vægtfaktordesign for FCS-MPC. Derudover foreslås en model-fri selvlærende prædiktiv controller for konverteren, som opnår sammenlignelig kontrolpræstation til konventionel FCS-MPC uden forudgående kendskab til systemet. Desuden foreslås en sikkerhedsramme for at håndhæve fysiske grænser og samtidig forbedre læringseffektiviteten af RL-agenter. Disse fund åbner muligheder for implementering af den online RL-baserede controller i praktiske elektroniksystemer.

Desuden, på det sekundære kontrolniveau, er den kommunikationsbaserede sekundære controller bredt vedtaget til koordinationskontrol af moduler, hvilket udsætter DC-mikronet for cyberangreb. Forskellige avancerede cyberangrebsdetektorer er foreslået, som stadig står over for en kritisk udfordring, når de konfronteres med snedige modstandere, der introducerer innovative angrebsmønstre. For at løse dette, anvendes RL først til at identificere systemets sårbarheder over for cyberangreb ved at efterligne intelligente angribere og generere nye angrebsmønstre, der autonomt omgår konventionelle cyberangrebsdetektorer. Omvendt foreslås en datadrevet angrebsdetektor til at detektere angrebene og supplere state-of-the-art cyberangrebsdetektorer. På denne måde udvikles en ML-aktiveret ramme for identifikation og detektion af cyberangreb. Det åbner også muligheder for iterativ implementering af det foreslåede cybersikkerhedsramme for at identificere og detektere et bredere spektrum af angreb i DC-mikronet.

Til sidst, på det tertiære kontrolniveau, for økonomisk drift af DC-mikronet, bør nedbrydningen af batterier modelleres nøjagtigt for optimal kontrol for at minimere de samlede driftsomkostninger. Forskellig fra de konventionelle metoder, der bruger en enkelt-trins batteridegraderingsmodel, foreslås en flertrins batteridegraderingsmodelmetode til at fange de varierende aldringsmønstre gennem dets levetid, hvilket muliggør en højere evalueringsnøjagtighed af batterikapacitetstab under forskellige driftsbetingelser. Derudover, for at tilpasse sig varierende driftsbetingelser og mindske systemets usikkerhed, udforskes en RL-baseret energistyringsstrategi under hensyntagen til batteridegradering for økonomisk og autonom drift af DC-mikronet.

Som konklusion er bidragene fra dette ph.d.-projekt forslaget om en AI-aktiveret hierarkisk kontrolramme, der spænder fra model-fri optimal strømomformer kontrol på det primære kontrolniveau til en datadrevet cyberangrebs identifikationsdetekteringsskema på det sekundære kontrolniveau, og udvidende til en RL-baseret energistyringsstrategi med en flertrins batteridegraderingsmodel på det tertiære niveau. Det forventes, at det foreslåede AI-aktiverede flerlagskontrolramme vil fremme udviklingen af mere intelligente, sikre og effektive DC-mikronet.

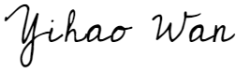
Preface

This thesis was prepared at the Division of Power and Energy Systems, a unit of the Department of Wind and Energy Systems, at the Technical University of Denmark (DTU), in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Engineering.

This dissertation summarizes the work undertaken by the author throughout his Ph.D. project. The project commenced on 1 January 2021 and concluded on 31 December 2023, under the supervision of Professor Tomislav Dragicevic, the co-supervision of Associate Professor Nenad Mijatovic, and the guidance of PostDoc Ramadhani Kurniawan Subroto. A segment of the research was conducted during an external Ph.D. stay at the Department of Electric Power and Energy Systems, KTH Royal Institute of Technology, Sweden, supervised by Assistant Professor Qianwen Xu. The external stay was funded by the Idella Foundation and Otto Mønsteds Fond.

The thesis is composed of five chapters and seven attached scientific papers, five of which have been peer-reviewed and published, whereas the remaining are currently under review.

Kongens Lyngby, 22nd December 2023



Yihao Wan

Acknowledgements

Embarking on this doctoral thesis has been a profound odyssey, which reminds me of the drips and drabs throughout the journey in the past three years. I was confused, struggling, and self-questioning, but I am glad and proud that I went through all of them positively most of the time. This would not have been possible without the help and support of my supervisors, friends, and families.

Foremost, my sincere gratitude extends to my Ph.D. supervisor, Tomislav Dragicevic, whose brilliance and guidance inspire me continuously. During the past years, your thoughtful guidance and insightful discussion have encouraged me to grow professionally and personally. I am also grateful to my PhD co-supervisor, Nenad Mijatovic, for his mentorship, from whom I learned a lot about experimental skills, especially at the beginning of my PhD. I appreciate the help from my co-supervisor, postdoc, Ramadhani Kurniawan Subroto, and Cheng Zhang, for keeping his door open whenever I needed their input. I would also like to express my thanks to Daniel Gebbran, who helped me a lot when I worked in a new field.

I was fortunate to interact with many great researchers at DTU and internationally. I would like to express my gratitude to Qianwen Xu for hosting me for three months during my research stay at KTH, Sweden. Also, I am thankful to my colleagues Pere Izquierdo Gómez, Mohammad Sadegh Orfi Yeganeh, Miguel Lopez, Mohammad Mehdi Mardani, Javiera Quiroz, Jyun Lin, Chang Li, who worked with me throughout my PhD. I am going to remember those days when we did experiments and discussed together for a very long time. I am also glad to collaborate with Ali Jafarian Abianeh from the University of Louisiana, Mateus Kaiss from the Federal University of Paraná. I was fortunate to meet many friends and roommates, Ge Huang, Qingjie Lyu, and Baichen Liu, Haoran Chen, Yi Xu, Dong Ding, and Murat, who made my daily life during my Ph.D.

My undergraduate studies, the first step towards the Ph.D., would not have been possible without my parents' support and unconditional love, Chunlian Wan and Xiaoping Wan. Thank you both for everything. I am grateful to my brother Zhichao Xu for inspiring me when I got confused on my way forward.

Still, the past three years' journey is just a brief chapter in the grand journey of life. Now, I am ready to compose all my emotions and set off again to embrace opportunities and challenges in my next stage.

Acronyms

RES	Renewable energy source
ESS	Energy storage system
DG	Distributed generation
DC	Direct current
EV	Electric vehicle
AC	Alternating current
PI	Proportional-integral
FCS-MPC	Finite control set model predictive control
ANN	Artificial neural network
RL	Reinforcement learning
DQN	Deep Q-network
DDPG	Deep deterministic policy gradient
PPO	Proximal policy optimization
DRL	Deep reinforcement learning
DoD	Depth-of-discharge
SoC	State-of-charge
DER	Distributed energy resource
ML	Machine learning
AI	Artificial intelligence
BEL	Brain emotional learning
VSC	Voltage source converter
MMC	Modular multilevel converter

ADP	Adaptive dynamic programming
TPS	Triple phase shift
DAB	Dual active bridge
UPS	Uninterruptible power supply
MDP	Markov decision process
THD	Total harmonic distortion
MFRL	Model-free reinforcement learning
RMS	Root mean square
FDIA	False data injection attack
DE	Discordant element
MARL	Multi-agent reinforcement learning
DS	Distributed screening
DF	Diverging factor
PRN	Pattern recognition network
BESS	Battery energy storage systems
RCA	Rainflow counting algorithm
SoH	State of health
EoL	End-of-life
BoL	Beginning-of-life
RMSE	Root-mean-square error
FCS	Fast charging station
MPPT	Maximum power point tracking
TD3	Twin delayed DDPG
GAN	Generative adversarial network

Contents

Part I

Report

CHAPTER 1

Introduction

1.1 Background

To mitigate climate change effects and address the gradual depletion of conventional energy sources such as coal, gas, oil, and other fossil fuels, there has been a global shift towards renewable energy resources (RESs), predominantly wind and solar energy. These RESs are integrated into power distribution networks using interfacing power electronic converters. However, due to the stochastic nature of power generation and load demand, the high penetration of RESs in existing distribution networks may lead to reliability and stability issues for the entire network [1]. To address these challenges, a distributed grid configuration comprising various RESs, energy storage systems (ESSs), and local loads organized into distinct entities has emerged [2]. Distributed generation (DG) technology facilitates the integration of RESs into the grid as generation units, reducing transmission losses and promoting an environmentally sustainable power grid. Nonetheless, this requires grid support and cannot operate independently, thereby limiting its practical application flexibility.

Microgrids, defined as self-contained, locally controlled clusters of distributed grid configurations, have gained popularity. Microgrids can operate in either islanded or grid-connected modes, seamlessly transitioning between these modes. They can be installed in industrial areas, official buildings, commercial centers, and residential complexes, ensuring the provision of premium-quality electrical power in islanded mode and supporting power distribution systems in grid-connected mode [3]. Among various microgrid configurations, the direct current (DC) microgrid, illustrated in Fig. 1.1, is regarded as a promising option for future distribution generation systems. Owing to their inherent compatibility with DG resources, DC microgrids possess significant potential to enable intelligent grid applications both efficiently and cost-effectively. In short, the advantages of DC microgrids are as follows [4, 5]:

- Simplified components and integration. High energy efficiency owes to the reduction in the number of conversion stages. Specifically, many RESs, electric vehicles (EVs), ESSs, and other DG sources like fuel cells exhibit a natural DC output. Additionally, modern electronic loads, including computers and light-emitting diodes, operate on DC systems. Consequently, they can be directly connected to a DC bus or via DC/DC converters, eliminating the need for additional AC/DC conversion devices.

- Grid synchronization issues are eliminated. Simpler control and management

than conventional alternating current (AC) microgrids since no reactive power or frequency control is required.

- Higher efficiency and reduced losses. DC microgrids have lower energy losses during transmission and distribution, making them more efficient than AC microgrids due to the absence of skin effects and reactive power losses.

With these advantages, the potential applications of DC microgrids include high-efficiency households, renewable energy parks, hybrid ESSs, EV charging stations, and other terrestrial systems. Moreover, DC microgrids also hold potential for widespread application in transportation electrification, including systems like naval ships, spacecraft, aircraft, submarines, EVs, and more.

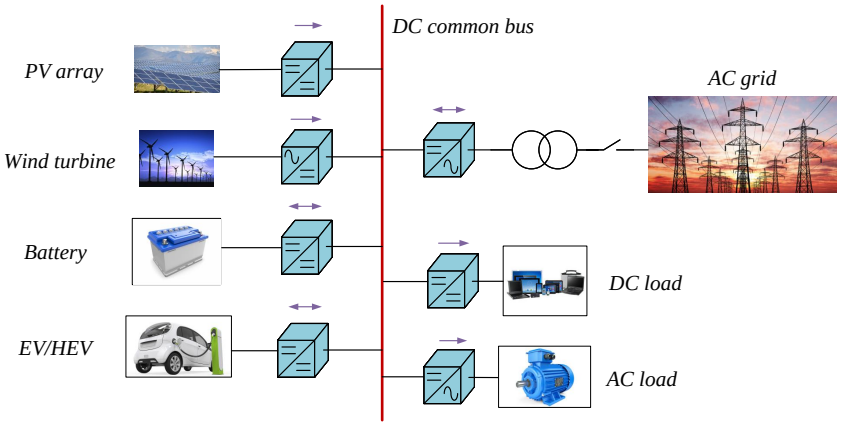


Figure 1.1. Typical DC microgrid configuration.

1.2 Control structure of DC microgrids

Different control objectives should be properly realized to achieve an autonomous DC microgrid operation, including the basic voltage and current control and the advanced objectives such as power sharing among the DGs, power quality control, ancillary services provision, and economic operations. For such multi-objective control in DC microgrids, the hierarchical control structure is acknowledged as a standard solution, distributing control objectives across various layers to enhance system efficiency [6, 7]. This multi-layer control structure is depicted in Fig. 1.2.

Primary control mainly deals with device-level control, including voltage and current regulations and preliminary power sharing control [8], typically implemented locally to enable independent operation of each unit. The secondary control layer is responsible for voltage compensation and further enhances accurate power sharing

beyond the primary control layer. At the system level, the tertiary control layer executes advanced control functions, including energy management and economic dispatch [9]. In general, the multilayer control framework integrates simultaneous local converter control with communication-based coordination control, exemplified by cloud-based platforms [10], wherein the control bandwidth is differentiated by orders of magnitude.

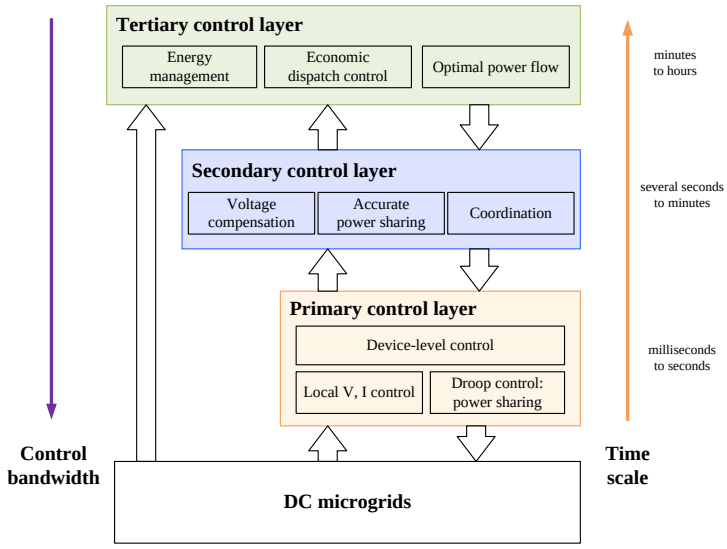


Figure 1.2. Hierarchical control framework for DC microgrids.

Depending on the implementation method of the multilevel control framework in DC microgrids, the control structure can be further categorized into centralized, decentralized, and distributed controls [5]. Centralized secondary control, as illustrated in Fig. 1.3, employs a central controller to regulate voltage based on DC bus voltage information, transmitting the voltage compensation term to the primary controller via a high-bandwidth communication network [11]. In this structure, all data regarding DG units are transmitted to the central controller through the communication network, allowing for comprehensive monitoring of all DG unit operations. However, this approach heavily relies on the central controller, and any failure in the communication link can lead to operational failure in the DC microgrid, compromising its reliability and efficiency. Another challenge associated with centralized control is the increased computational load and storage requirements as the number of DG units grows. Therefore, the control strategy is more suitable for local and small-scale DC microgrids, where the data transmission is limited, allowing centralized control to be achieved with relatively lower computational and communication costs [6, 9].

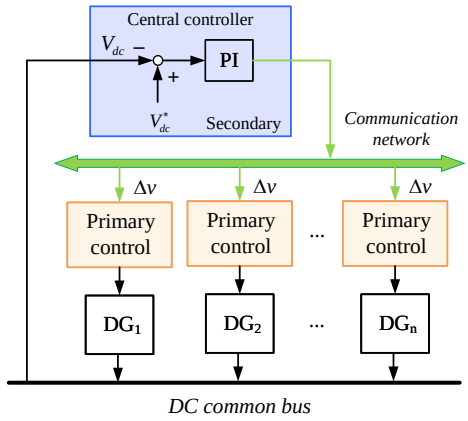


Figure 1.3. Centralized secondary control for DC microgrids.

The decentralized secondary control structure, implemented via the local controller and eliminating the need for a communication network, is depicted in Fig. 1.4. This structure facilitates fast power sharing during transient and steady-state conditions due to reduced real-time communications. Typical decentralized control methods include droop-based control methods and their derivatives [12, 13], distributed DC bus signaling [14], and power line signaling methods [15]. However, this approach lacks coordination among different DG units and scalability for integrating new DG units. In addition, global operations like economic operations and energy management might be less effective under the decentralized control framework.

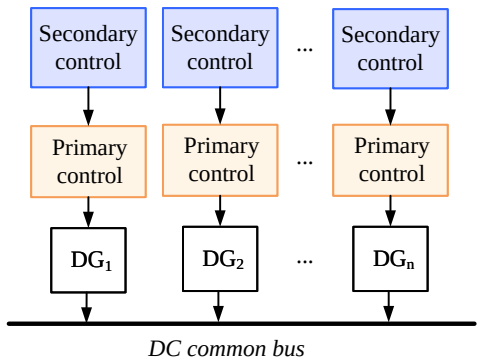


Figure 1.4. Decentralized secondary control for DC microgrids.

The distributed secondary control strategy, as shown in Fig. 1.5, has gained significant attention for its enhanced reliability and flexibility. In general, each DG unit shares information with its neighboring units via the communication network, either

current or voltage, or concurrently both of them [16], to achieve accurate current sharing among the DGs. Compared to centralized and decentralized secondary controls, this strategy offers greater reliability, flexibility, and scalability, thus providing immunity to single points of failure. However, the reliance on a communication network makes it a cyber-physical system vulnerable to cyber-attacks.

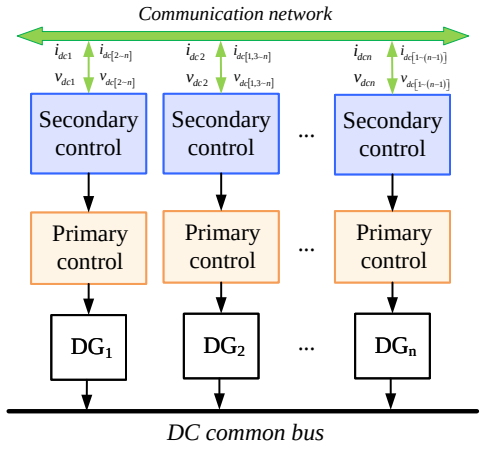


Figure 1.5. Distributed secondary control for DC microgrids.

Finally, the tertiary control layer monitors the global operation of DC microgrids, managing and dispatching power and energy for various objectives, such as power flow control, energy management, and economic operations. For optimal operations of DC microgrids, scheduling, and coordination among different components are critical for an autonomous and efficient microgrid system.

1.2.1 Power converter control at device level

As shown in Fig. 1.1, interfacing power electronic converters play an important role in DC microgrids. To ensure effective local operations and coordination among different units [5], local current and voltage control are essential as inner control loops, with droop control commonly utilized for power sharing among parallel-connected DG units. Proportional-integral (PI) controllers, known for their simplicity, robustness, and effective tracking performance, are commonly employed for local current and voltage control in DC microgrids [11]. Nonetheless, this cascaded-loop control structure is limited by its inherently slow dynamic response, leading to undesirable transient power sharing in conventional local control scenarios. However, recent advancements in the computational capabilities of modern microprocessors have made the application of advanced control strategies, capable of significantly enhancing conventional strategies in DC microgrids, increasingly feasible [17].

To mitigate the limitations of conventional cascaded linear controllers, advanced control strategies have been proposed [9, 17], categorized into model-based and data-driven controls. Among different conventional advanced controllers, finite-control-set model predictive control (FCS-MPC) has been extensively studied due to its fast dynamic response, straightforward and flexible formulation of control objectives and constraints, and the discrete nature of power converters. However, it suffers from parameter sensitivity, unmodeled dynamics, and requires proper weighting factor design in the cost function to achieve desired performances. To mitigate these issues, data-driven controllers, developed by training artificial neural networks (ANN) with data from simulation models or experimental setups, have been proposed for optimal power converter control [18]. Furthermore, reinforcement learning (RL) algorithms, enhancing the exploration and self-learning capabilities of ANN, have gained popularity in power electronics [19]. However, system safety during the learning process of the RL-based controller is one of the major concerns for practical applications, remaining an open question [20]. Consequently, this Ph.D. project will demonstrate two applications of the data-driven RL method at the local control level: the optimization of weighting factors for conventional FCS-MPC and model-free self-learning for optimal converter control.

1.2.2 Cybersecurity at the secondary control layer

Cyber-attacks in energy systems have been escalating [21], as evidenced by the grid blackout in Venezuela during 2019–2020 and the recent cyber-attacks on Ukraine's electricity grid. The cybersecurity of the energy sector has gained increasing attention. For geographically dispersed DC microgrids, distributed control with communication networks is extensively employed to coordinate DGs, making them cyber-physical systems vulnerable to cyber-attacks [22]. Normal operations and the stability of DC microgrids would be compromised if not properly detected.

Different cyber-attack detection methods have been proposed to detect various types of attacks in DC microgrids, which can be divided into model-based and data-driven detection methods. Model-based detection methods are devised according to system characteristics, involving various detection metrics [23, 24], using residual based on system states estimation [25], and different observer-based detection schemes [26]. This approach lacks scalability as it is contingent upon specifically designed frameworks and system configurations. Conversely, the data-driven detection method identifies attacks using system measurements, independent of the system model or parameters. In particular, measured operation data for the healthy DC microgrid are employed to train the data-driven detector [27–29], which shows good performance in detecting different attack types. However, these methods often train NN detectors with data from the healthy system alone, limiting their capability to recognize diverse attacks due to a lack of knowledge about different attack types.

Nonetheless, conventional cyber-attack detection methods have limited adaptability across different systems and may fail to identify more sophisticated attack types.

An intelligent attacker may exploit vulnerabilities of DC microgrids, even under the shield of attack detection methods, by penetrating multiple sensors or communication links in a coordinated manner. In essence, attackers can produce sophisticated attack patterns that evade detection by existing cyber-attack detection methods [30]. Therefore, it is imperative to devise an autonomous and adaptable cybersecurity framework that can detect novel attack patterns, ensuring safe and reliable operation.

1.2.3 Energy management at system level

The system's operation and coordination are monitored at the tertiary control layer, where operational scheduling and economic dispatch for each unit constitute the main control objectives. With the increasing penetration of RESs, including solar systems and wind farms, they can provide green energy and promote low-carbon electricity for the DC microgrids. However, environmental factors, such as geographical location, season, and weather, result in intermittent power generation. In addition, ESSs act as energy buffers for the DC microgrids to mitigate this, and their operational conditions influence their lifespan. All the above should be considered for the scheduling of DC microgrids.

Energy management of DC microgrids for economic operations is intrinsically an optimization-based decision-making process under diverse operating conditions. Conventional scheduling frameworks address the optimization problem under dynamic operating conditions through various methods, such as multistage optimization [31], stochastic programming [32], and MPC-based optimization frameworks [33]. However, they are all model-based optimization strategies that require accurate prediction of system variables such as load demand and electricity prices. More importantly, the increasing integration of DGs in the DC microgrids complicates the system structure, making it difficult to model each component accurately. Data-driven methods with machine learning (ML) techniques have become popular for the energy management of microgrid systems [34]. Various RL algorithms are applied to energy management for the optimal operation of microgrids to address the uncertainty of the system conditions. Deep Q-Network (DQN) [35], double DQN [36] with discrete action space, and deep deterministic policy gradient (DDPG) [37], proximal policy optimization (PPO) [38, 39] with continuous action space, have been extensively explored to manage the system operation, aiming to minimize the operation cost or increase the revenue of the microgrids. However, current deep reinforcement learning (DRL) based energy management frameworks have not comprehensively addressed the cost associated with battery degradation.

In the energy management framework, battery energy storage systems (BESSs) play an important role in dispatching DGs to reduce the operation cost of DC microgrids. It can realize different advanced functions, such as energy arbitrage, peak-shaving, and so on [40]. As the batteries can only sustain limited charging/discharging cycles, accurate battery degradation modeling is critical for optimal scheduling [41]. Different stressing factors contribute to the capacity losses of the battery during op-

erations [42], including temperature, current, depth-of-discharge (DoD), and average state-of-charge (SoC). Therefore, the energy management framework should consider battery degradation under diverse stressing conditions when dispatching DGs to minimize the operation cost of DC microgrids.

1.3 Project motivation and research goals

As the cornerstone of the future smart grid, microgrids are undergoing a transformation, driven by emerging new distributed energy resources (DERs) and advanced techniques. The uncertainties, complex system models, parameter variations and availability, dynamic management, and the like characteristics of the microgrid system pose challenges to the current microgrid control system [19].

As mentioned in previous sections, the hierarchical control framework ensures proper local operations of DGs and coordination among different modules of DC microgrids. Drawing inspiration from recent breakthroughs in deep neural networks—the central workhorses of artificial intelligence (AI)—this research aims to develop an AI-enabled multilayer control framework to enhance the intelligent and secure operations of DC microgrids. In essence, the overall research question for the Ph.D. project can be formulated as

- How can ML be fused into the existing DC microgrid control framework?

To answer this question, we need to look into the existing hierarchical control framework of DC microgrids to fulfill the potential of advanced AI techniques for DC microgrid control systems. At the primary control layer for power converter control, the conventional FCS-MPC suffers from model dependence and requires an autonomous weighting factor tuning method. At the communication-based secondary control layer, cyber-attack threats in DC microgrids are critical, requiring the identification and detection of various attacks to ensure secure and reliable operations. For energy management at the system level, an energy management strategy under varying operating conditions is necessary for efficient and economic dispatch of modules. In addition, accurate evaluation of battery degradation under diverse stressing conditions during scheduling is crucial for an efficient energy management framework.

The goal of the Ph.D. project is to develop an AI-enabled multilayer control framework for DC microgrids at different control levels to address these limitations and concerns. The primary objectives of this dissertation are further articulated with the following research questions:

- How to optimize the cost function design of the FCS-MPC and develop a model-free and self-learning predictive controller for power converters?

- How can the safety of the system be ensured during the training process of a model-free, learning-based optimal controller?

- How to identify the vulnerabilities of the system to cyber-attacks and detect the attacks for a secure communication-based secondary control?
- How to achieve autonomous energy management under dynamically changing conditions with an appropriate battery degradation model?

These questions will be answered in the sequential order of the thesis chapters based on the published journal and conference papers.

1.4 Thesis outline

This thesis sums up the outcomes of the Ph.D. project based on selected journal (J) and conference (C) papers. The dissertation is organized into two main parts: **Report** and **Publications**.

- **Chapter 1: Introduction**
  The state of the art in hierarchical control frameworks is presented, illustrating different applications of AI techniques in DC microgrids. The motivation, research goals, and objectives of the Ph.D. project are comprehensively presented and explained.
- **Chapter 2: Learning-based controller for power electronic converters**
  Guidelines for RL application in automatic weighting factor design for FCS-MPC based on different control objectives and unsupervised model-free learning of optimal power converter control are proposed. In particular, the training of RL agents for controller design and the self-learning of optimal control for power converters are simulated offline. Moreover, a demonstration framework for transferring the online RL methods from simulation to online implementation in a practical system is proposed. In addition, to address the safety concern during the self-learning process of the RL agent, a novel safety-enhanced learning-based control for power converters is proposed by introducing a safe policy into the learning process. The results show that the RL agent converges faster to the optimal controller while ensuring the system's physical limits are not violated during the training process. All the relevant results are also detailed in the chapter.
  Based on publications: J1, J2, C1
- **Chapter 3: Data-driven intelligent attack generation and detection in DC microgrids**
  A data-driven cybersecurity framework has been developed for the autonomous generation and detection of attacks. Firstly, to identify the vulnerabilities of DC microgrids to cyber-attacks, RL is employed to emulate the intelligent attackers to autonomously generate stealthy attack patterns that bypass conventional cyber-attack detection schemes. Conversely, a data-driven attack detector is proposed to detect the attacks, augmenting state-of-the-art cyber-attack

detectors. More importantly, the proposed framework has the potential to be extended to generate and detect a wider spectrum of attacks. The results show that the RL-based attacker can generate stealthy attacks that conventional cyber-attack detection methods fail to detect, and the proposed supplementary data-driven cyber-attack detector can effectively detect the attack.
Based on publications: J3, J4

- **Chapter 4: Battery health-aware DRL-based energy management for DC microgrids**
  A multi-stage battery degradation modeling method is proposed, designed to capture the varying aging patterns throughout its lifetime, thus enabling a more precise estimation of battery capacity losses under diverse operating conditions. In addition, to adapt to varying operating conditions and system uncertainties, a DRL-based energy management strategy is proposed for the economic operation of DC microgrids. The results under different example days and varying operating conditions are presented to verify the effectiveness of the proposed energy management framework.
  Based on publications: J5, C2

- **Chapter 5: Conclusion**
  All the contributions of the Ph.D. project are summed up, and perspectives on future research are elaborated.

1.5 List of publications

The research outcomes of the Ph.D. project have been disseminated in journal and conference publications as listed below. Selected publications are used in the dissemination, as explained previously.

1.5.1 Journal papers

- [**J1**] **Y. Wan**, X. Qian and T. Dragicevic, "Unsupervised learning-based Predictive Control for Power Electronic Converters," in IEEE Transactions on Industrial Electronics, under review.

- [**J2**] **Y. Wan**, X. Qian and T. Dragicevic, "Safety-Enhanced Self-Learning for Optimal Power Converter Control," in IEEE Transactions on Industrial Electronics, under review.

- [**J3**] **Y. Wan** and T. Dragicevic, "Data-driven Cyber-attack Detection of Intelligent Attacks in Islanded DC Microgrids," in IEEE Transactions on Industrial Electronics, vol. 70, no. 4, pp. 4293-4299, April 2023.

- [**J4**] A. J. Abianeh, **Y. Wan**, F. Ferdowsi, N. Mijatovic and T. Dragičević, "Vulnerability Identification and Remediation of FDI Attacks in Islanded DC Microgrids Using Multiagent Reinforcement Learning," in IEEE Transactions on Power Electronics, vol. 37, no. 6, pp. 6359-6370, June 2022.
- [**J5**] **Y. Wan**, D. Gebbran, R. K. Subroto and T. Dragičević, "Optimal Day-ahead Scheduling of Fast EV Charging Station With Multi-stage Battery Degradation Model," in IEEE Transactions on Energy Conversion, early access.
- J. V. Barreras, R. de Castro, **Y. Wan**, and Dragicevic, T, "A consensus algorithm for multi-objective battery balancing. Energies", 2021, 14(14), 4279.

1.5.2 Conference papers

- [**C1**] **Y. Wan**, T. Dragicevic, N. Mijatovic, C. Li and J. Rodriguez, "Reinforcement Learning Based Weighting Factor Design of Model Predictive Control for Power Electronic Converters," 2021 IEEE International Conference on Predictive Control of Electrical Drives and Power Electronics (PRECEDE), Jinan, China, 2021, pp. 738-743.
- [**C2**] **Y. Wan**, D. Gebbran, P. I. Gómez and T. Dragicevic, "Optimal dispatch schedule for a fast EV charging station with account to supplementary battery health degradation," IEEE Transportation Electrification Conference & Expo (ITEC), Anaheim, CA, USA, 2022, pp. 552-556.

CHAPTER 2

Learning-based controller for power electronic converters

2.1 Introduction

In DC microgrids, power electronic converters are critical for integrating different DG units, interfaces between different sources and end users flexibly [8]. Conventional cascaded linear control strategies are commonly employed to control the converters, attributed to their simple structure, robustness, and adaptability. However, due to the inherent nonlinearity characteristics of power electronic systems, conventional linear control structures suffer from slow dynamic response and limited performance for multiple control objectives. As the computation power of processors advances, different advanced control methods are emerging for practical applications [17]. The FCS-MPC has gained increasing attention due to its simple design of control objectives in the cost function, the straightforward inclusion of nonlinear constraints, and flexibility in different converter topologies [43].

Despite those advantages, there are still open research questions about FCS-MPC for application in power converters [44]. The primary issue is the design of the cost function, which involves balancing various control objectives by appropriately selecting the associated weighting factors. The most straightforward way is to perform numerous simulations to select the optimal weighting factors by manual trial and error of different combinations [45], which is rather time-consuming. One of the first endeavors is the branch-and-bound search method [46], which is simple and easy to implement while the method is still too empirical. Heuristic methods, which iteratively tune the weighting factors based on various metaheuristic algorithms, have been proposed [47, 48]. However, the heuristic tuning of weighting factors does not give the most optimal control performance [49] and is also time-consuming for converging to the optimal solutions. An ANN approach has recently been proposed for automatic weighting factor design for voltage source converters [50]. However, this approach requires extensive simulations or experiments for data generation, covering a broad spectrum of operating conditions for parameter sweeping of the weighting factors. An additional exhaustive search step is also required to obtain the optimal weighting factors to optimize the desired control objectives. A model-free method using the brain emotional learning (BEL) algorithm is proposed [51] while it requires tuning of introduced coefficients in the BEL structure.

The second challenge involves dependence on parameters and built-in models to achieve desirable control performance. Various attempts have been made using ML techniques to regulate two-level voltage source converters (VSC) [18] and modular multilevel converters (MMC) [52]. In principle, these methods require training data generated by offline simulations, wherein specific converter models are required to generate data with a specifically designated FCS-MPC. To address this, a model-free adaptive controller that integrates with the FCS-MPC has been proposed [53, 54]. Furthermore, RL has also gained increasing attention for applications in the power electronics field. In [55], an RL-based event-triggered predictive control framework, employing an adaptive dynamic programming (ADP) scheme and an event-triggered mechanism, is utilized to enhance the controller's adaptability amidst unknown nonlinear system dynamics and model variations. In addition, different RL algorithms are employed to optimize the triple phase shift (TPS) modulation for the dual active bridge (DAB) converter and the derivative converter system [56].

Motivated by these, this Ph.D. project aims to explore RL for weighting factor design and achieve model-free predictive control for power converters. The highlights of the methods are **[J1, J2, C1]**:

- An RL-based weighting factor design approach. The approach can autonomously select optimal weighting factors for the FCS-MPC to achieve the desired control objectives.
- An unsupervised model-free RL-based controller for converters. The RL agent can autonomously find the optimal switching strategy by emulating the FCS-MPC, achieving desired control performances for VSC without prior system knowledge.
- A deployment framework for transferring online RL from simulation to practical experimental demonstration. It enables the practical implementation and validation of the proposed methods with RL.

2.2 System model

2.2.1 Converter model

In Fig. 2.1, a two-level VSC interfaces between the AC load and DC source, where the FCS-MPC regulates the AC voltage. The studied converter system is typical in DC microgrids for uninterruptible power supply (UPS) applications of AC loads. For the two-level VSC, the combination of all the switches results in eight switching states in total, which should be selected properly by the controller for desirable control performance. In particular, the predictive controller predicts the converter behavior in the future based on the most recent measurements. A cost function defining different control objectives is designed and employed to evaluate the performance

of the predicted switching states. The optimal switching state with the lowest cost function value will be applied to control the converter. The modeling of the converter system is presented as follows.

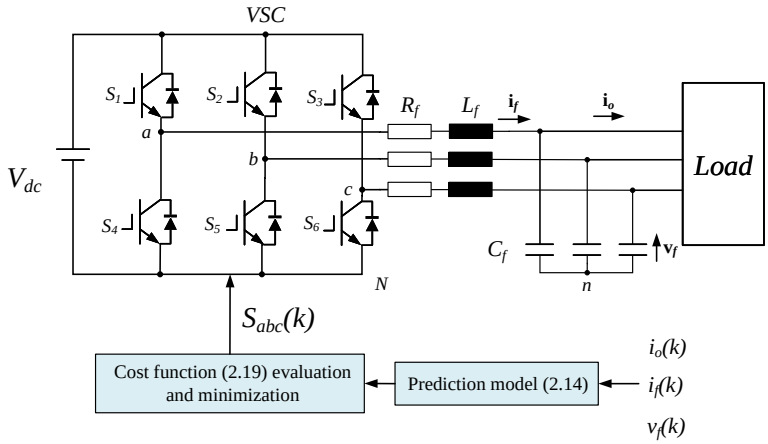


Figure 2.1. Predictive control for VSC. Source: [**J1**]

The switching configurations of the two-level VSC are defined by the complementary gate signals of the phase leg, S_a, S_b, and S_c, represented by

$$S_a = \begin{cases} 0, & S_1 \text{ on}, S_4 \text{ off} \\ 1, & S_1 \text{ off}, S_4 \text{ on} \end{cases} \tag{2.1}$$

$$S_b = \begin{cases} 0, & S_2 \text{ on}, S_5 \text{ off} \\ 1, & S_2 \text{ off}, S_5 \text{ on} \end{cases} \tag{2.2}$$

$$S_c = \begin{cases} 0, & S_3 \text{ on}, S_6 \text{ off} \\ 1, & S_3 \text{ off}, S_6 \text{ on}. \end{cases} \tag{2.3}$$

The output voltage of each phase leg with respect to the point N can be formulated by multiplying the DC link voltage with the associated gating signal as

$$v_{aN} = S_a \cdot v_{dc} \tag{2.4}$$

$$v_{bN} = S_b \cdot v_{dc} \tag{2.5}$$

$$v_{cN} = S_c \cdot v_{dc} \tag{2.6}$$

To obtain the output phase voltage regarding the neutral point n, a common voltage drop v_{nN} across the filter should be subtracted, expressed as

$$v_{nN} = \frac{v_{aN} + v_{bN} + v_{cN}}{3} \tag{2.7}$$

Therefore, the phase voltage across the filter is expressed as

$$v_{an} = v_{aN} - v_{nN} \tag{2.8}$$

$$v_{bn} = v_{bN} - v_{nN} \tag{2.9}$$

$$v_{cn} = v_{cN} - v_{nN}. \tag{2.10}$$

The VSC behavior is modeled in a $\alpha\beta$ frame, where Clark transformation is applied to the filter voltage. In this way, for all the possible combinations, the voltage vector v_i can be obtained, as listed in Table 2.1.

Table 2.1. Switching states and voltage vectors. Source: [**J1**]

S_a	S_b	S_c	Voltage vector $\mathbf{v}_i$
0	0	0	$\mathbf{v}_0 = 0$
1	0	0	$\mathbf{v}_1 = \frac{2}{3}v_{dc}$
1	1	0	$\mathbf{v}_2 = \frac{1}{3}v_{dc} + j\frac{\sqrt{3}}{3}v_{dc}$
0	1	0	$\mathbf{v}_3 = -\frac{1}{3}v_{dc} + j\frac{\sqrt{3}}{3}v_{dc}$
0	1	1	$\mathbf{v}_4 = -\frac{2}{3}v_{dc}$
0	0	1	$\mathbf{v}_5 = -\frac{1}{3}v_{dc} - j\frac{\sqrt{3}}{3}v_{dc}$
1	0	1	$\mathbf{v}_6 = \frac{1}{3}v_{dc} - j\frac{\sqrt{3}}{3}v_{dc}$
1	1	1	$\mathbf{v}_7 = 0$

The VSC model is established based on the dynamics of the LC filter, as

$$L_f \frac{d\mathbf{i}_f}{dt} = \mathbf{v}_i - \mathbf{v}_f - R_f \mathbf{i}_f \tag{2.11}$$

$$C_f \frac{d\mathbf{v}_f}{dt} = \mathbf{i}_f - \mathbf{i}_o \tag{2.12}$$

where L_f and C_f are the filter inductance and capacitance respectively, R_f is the series resistance of L_f.

The equations can be rewritten in state-space form as

$$\frac{d}{dt}\begin{bmatrix}\mathbf{i}_f \\ \mathbf{v}_f\end{bmatrix} = \boldsymbol{A}\begin{bmatrix}\mathbf{i}_f \\ \mathbf{v}_f\end{bmatrix} + \boldsymbol{B}\begin{bmatrix}\mathbf{v}_i \\ \mathbf{i}_o\end{bmatrix} \tag{2.13}$$

where $\boldsymbol{A} = \begin{bmatrix}-\frac{R_f}{L_f} & -\frac{1}{L_f} \\ \frac{1}{C_f} & 0\end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix}\frac{1}{L_f} & 0 \\ 0 & -\frac{1}{C_f}\end{bmatrix}$.

For implementation in digital control, model discretization based on the Euler forward discretization method with a sample time T_s is performed [50]. The discretized model is therefore expressed as

$$\begin{bmatrix}\mathbf{i}_f(k+1) \\ \mathbf{v}_f(k+1)\end{bmatrix} = \boldsymbol{A}_d\begin{bmatrix}\mathbf{i}_f(k) \\ \mathbf{v}_f(k)\end{bmatrix} + \boldsymbol{B}_d\begin{bmatrix}\mathbf{v}_i(k) \\ \mathbf{i}_o(k)\end{bmatrix}. \tag{2.14}$$

where $\boldsymbol{A}_d = e^{\boldsymbol{A}T_s}$ and $\boldsymbol{B}_d = \int_0^{T_s} e^{\boldsymbol{A}\tau}\boldsymbol{B}\, d\tau$.

2.2.2 FCS-MPC for the VSC

As shown in Table 2.1, different switching combinations correspond a voltage vector $\mathbf{v}_i$, consisting of a voltage set of $v_{i\alpha}$ and $v_{i\beta}$. The voltage set will be input to the developed converter model in equation (2.14), of which the output values $v_{f\alpha}$ and $v_{f\beta}$ define the capacitor voltage $\mathbf{v}_f$. In this way, the FCS-MPC aims to find the optimal switching states to regulate the output voltage for tracking the reference AC voltage $\mathbf{v}_f^*$ with minimal deviation. As mentioned, a cost function is employed to evaluate the performance of the selected voltage vector for voltage tracking. Therefore, for a single-step horizon tracking of reference voltage, it is formulated as

$$g = (v_{f\alpha}^* - v_{f\alpha})^2 + (v_{f\beta}^* - v_{f\beta})^2 \tag{2.15}$$

In addition, to improve the steady-state tracking performance, a current reference term is introduced as [57]

$$g_d = (C_f \omega_r v_{f\beta}^* - i_{f\alpha} + i_{o\alpha})^2 + (C_f \omega_r v_{f\alpha}^* + i_{f\beta} - i_{o\beta})^2. \tag{2.16}$$

Moreover, for safe and efficient operation of the system, penalization terms for limiting the inductor current and reducing the switching frequency are employed, expressed below

$$h_{lim} = \begin{cases} 0, & |\bar{i}_f| \leq i_{max} \\ \infty, & |\bar{i}_f| > i_{max} \end{cases} \tag{2.17}$$

$$sw = |\triangle S_a(i)| + |\triangle S_b(i)| + |\triangle S_c(i)| \tag{2.18}$$

where $|\triangle S_j(i)|(j = a, b, c)$ is 1 if the phase leg gate signal switching states change and 0 vice versa.

Therefore, the comprehensive cost function incorporating all the terms can be formulated as

$$g = (v_{f\alpha}^* - v_{f\alpha})^2 + (v_{f\beta}^* - v_{f\beta})^2 + \lambda_d g_d + \lambda_{sw} sw^2 + h_{lim}. \tag{2.19}$$

The two weighting factors, λ_d and λ_{sw}, are associated with different control objectives, which should be appropriately designed to achieve the desired control performance. This performance is typically quantified by metrics such as total harmonic distortion (THD) and average switching frequency f_{sw}.

2.3 Learning-based controller

As demonstrated in equation (2.19), the weighting factors significantly influence control performance and, as such, require fine-tuning. Moreover, a model-free predictive controller that alleviates the model dependence while preserving the desirable control performance characteristic of conventional FCS-MPC should be developed. Due to the model-free and self-learning characteristics, RL is employed to develop the learning-based controller for VSC.

2.3.1 Formulation of RL

The RL paradigm for the learning-based controller is formulated as a Markov Decision Process (MDP), characterized by $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$ for the agent. As shown in Fig. 2.2, $\mathcal{S}$ denotes the state space, $\mathcal{A}$ is the action space, $\mathcal{P}$ represents the probability transition function regarding $s_t \rightarrow s_{t+1}$, R is the reward function for defining the goal of the agent, and γ is the discount factor that balances the immediate reward and long-term reward during the training process.

In particular, by interacting with the environment, at each time step t, based on the incoming environment state, $s_t \in \mathcal{S}$, the agent takes an action, $a_t \in \mathcal{A}$, regarding current policy π. Subsequently, the agent receives a reward $r_t \in R$, and the environment transits to a new state s_{t+1} with respect to the transition function P. The training session continues until the RL agent converges to the optimal policy, as

$$\pi^* \in \underset{\pi}{\operatorname{argmax}}\, G(\pi) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_t(s_t, a_t) \right]. \tag{2.20}$$

where $G(\pi)$ represents the accumulative reward over an infinite time horizon, and $\mathbb{E}[\cdot]$ denotes the expected value of a variable. It indicates the expected return regarding the optimal policy is the greatest among all the policies.

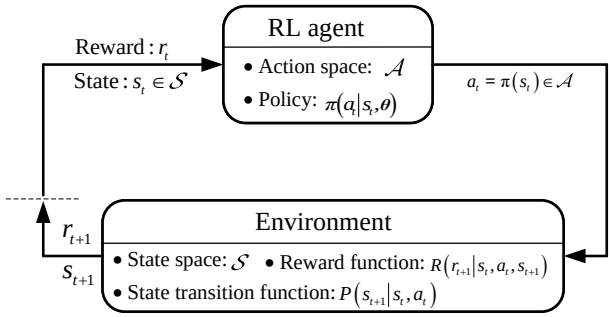


Figure 2.2. The agent environment interaction in an MDP. Source: [**J1**]

In general, value-based and policy gradient algorithms are the two main RL algorithms. Value-based algorithms, e.g., Q-learning and its derivations, learn the Q function that evaluates the state-action pairs and take actions that maximize the Q-value. Policy gradient methods directly learn the policy $\pi(a_t|s_t, \theta)$, using either stochastic or deterministic estimates of policy gradient.

The selection of an appropriate RL algorithm depends on the characteristics of the formulated problem. As mentioned, the RL agent explores and exploits the learning

space to maximize the long-term accumulative reward over an infinite time horizon, expressed as

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}. \tag{2.21}$$

To find the optimal policy $\pi(a_t|s_t)$ in (2.20), different recursive training algorithms can be applied. In off-policy algorithms such as Q-learning, an action-value function $Q_\pi(s_t, a_t)$ is formulated to evaluate the expected return of state-action pairs, as expressed in Bellman equation below [58]

$$\begin{aligned} Q_\pi(s_t,a_t) \leftarrow\ & Q_\pi(s_t, a_t)+ \\ & \alpha[r_{t+1} + \gamma \max_{a'} Q_\pi(s_{t+1}, a') - Q_\pi(s_t, a_t)] \end{aligned} \tag{2.22}$$

where α is the learning rate of the RL agent.

However, the iterative session cannot achieve the desirable performance in high-dimensional real-world applications. To enable a more precise and efficient prediction of the action-value function $Q_\pi(s_t, a_t)$, DQN has been developed for RL agents [59]. In the DQN algorithm, all the sequences $e = (s_t, a_t, r_t, s_{t+1})$ are stored in a R-sized experience memory $\mathcal{M} = \{e_1, ..., e_R\}$. For each training iteration, a minibatch of tuples is randomly selected from $\mathcal{M}$ to train the deep neural network via stochastic gradient descent. In addition, a target network $Q_\pi^{'}$ is used to improve the training stability and convergence speed. The target value is calculated as

$$y_t = r_t + \gamma \max_{a'} Q_\pi^{'}(s_{t+1}, a'|\theta_Q^{'}) \tag{2.23}$$

The network parameter of the Q-network is updated with the loss function below

$$L(\theta_Q) = \mathbb{E}\left[(y_t - Q_\pi(s_t, a_t|\theta_Q))^2\right] \tag{2.24}$$

The target network keeps its separate network parameter $\theta_Q^{'}$ fixed for every T_{target} steps and updates it with the current weights of the Q-network as

$$\theta_Q^{'} \leftarrow \tau\theta_Q + (1-\tau)\theta_Q^{'}. \tag{2.25}$$

where the smoothing factor $\tau \ll 1$.

While DQN is feasible for problems with high-dimensional observation spaces, it can only handle a limited set of discrete actions. As the goal is to select the optimal switching state from limited switching combinations to achieve the desirable control performance, DQN is suitable for the control problem of VSC.

However, in the context of weighting factor design, the values span a broad range within the action space. All the combinations of different values for weighting factors result in an exponential increase of learning space, making the training intractable. Therefore, RL algorithms capable of handling continuous action spaces, such as DDPG [60] with an actor-critic structure, are more suitable for this scenario.

Similarly, the critic network is updated by performing gradient descent on the loss function in (2.24), while the target value is used to train an additional actor network, which maps the state-action function $\mu(s|\theta_\mu)$. Therefore, the target value for the critic network is calculated as

$$y_t = r_t + \gamma Q_\pi^{'}(s_{t+1}, \mu(s_{t+1}|\theta_\mu)|\theta_Q^{'}) \tag{2.26}$$

The actor network parameters θ_μ can be updated as below [61]

$$\nabla_{\theta_\mu} J = \mathbb{E}_{s_t \sim \rho^\beta} \left[\nabla_a Q_\pi(s, a|\theta_Q)|_{s=s_t, a=\mu(s_t)} \nabla_{\theta_\mu} \mu(s|\theta_\mu)|_{s=s_t} \right] \tag{2.27}$$

where $s \sim \rho^\beta$ denotes the state s follows the distribution β. Then the target networks of the critic and actor networks are updated as

$$\theta_Q^{'} \leftarrow \tau\theta_Q + (1-\tau)\theta_Q^{'} \tag{2.28}$$

$$\theta_\mu^{'} \leftarrow \tau\theta_\mu + (1-\tau)\theta_\mu^{'}. \tag{2.29}$$

In addition, a noise model $\mathcal{N}$ decaying by training episodes is usually used to help the agent balance exploration and exploitation better.

In the continuous action space, balancing exploration and exploitation is challenging for the agent. To address this, an exploration policy $\mu^{'}$ is constructed by adding noise to the actor based on a noise process $\mathcal{N}$, as shown below [60]

$$\mu^{'}(s_t) = \mu(s_t|\theta_\mu) + \mathcal{N}_t. \tag{2.30}$$

The added noise decays during the training process, which better helps the agent balance exploration and exploitation.

In summary, DDPG with continuous action space and DQN with discrete action space are employed for weighting factor design and model-free predictive control, respectively.

2.3.2 DDPG-based weighting factor design

Fig. 2.3 demonstrates the overall schematic of the proposed RL-based weighting factor design method. As depicted in Fig. 2.2, a DDPG agent adjusts weighting factors in an environment of an FCS-MPC-regulated VSC. The input states are selected from the voltage and current measurements, and the reward is calculated accordingly. During the training iterations, the actions (weighting factors λ_d and λ_{sw}) generated by the agent are fed into the FCS-MPC, and the corresponding reward is computed based on the control performance of the current transited states. Iteratively, the self-learning RL agents discover the optimal weighting factor design policy to achieve the desired control performance.

It is noteworthy that the proposed RL-based method enables direct acquisition of optimal weighting factor values, bypassing the need for exhaustive search and extensive simulations that cover all potential operating conditions through parameter sweeping, as outlined in [50].

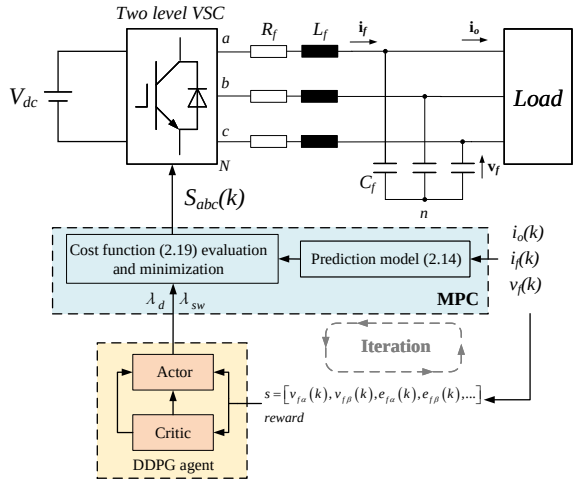


Figure 2.3. Block diagram of RL-based weighting factor tuning of FCS-MPC for VSC. Source: **[J1]**

- **State and action sets**
  To achieve the desired control performance of regulating the AC side voltage to track the reference voltage, the state information is defined as $s = [v_{f\alpha}, v_{f\beta}, e_\alpha, e_\beta, THD]$, where e_α and e_β are formulated as the tracking errors between the reference voltage ($v^*_{f\alpha}$, $v^*_{f\beta}$) and the output voltage ($v_{f\alpha}$, $v_{f\beta}$). The action set consists of weighting factors $a = [\lambda_d, \lambda_{sw}]$, which are tuned for the desirable control performances. In addition, both action spaces are limited to a feasible range [0, 10] for efficient exploration.

- **Reward function design**
  The reward function defines the desired control performances, usually quantified by two performance metrics: THD of the capacitor voltage and average switching frequency f_{sw}. As the control objectives vary under different operating scenarios, the reward function should be specifically designed for these objectives. Two exemplary control objectives are studied: minimizing the THD solely and minimizing the THD while lowering the f_{sw}.

  In the first case study, the reward function for the minimization of the THD is formulated as

  $$r = -THD^2. \tag{2.31}$$

  In this way, the DDPG agent will find the optimal tuning policy for the two weighting factors in the cost function to achieve THD minimization.

  In the second case study, the reward function for lowering THD and average

switching frequency is defined as

$$r = -(k_{thd} \cdot THD^2 + k_{sw} \cdot {f_{sw}}^2). \tag{2.32}$$

where k_{thd} and k_{sw} are the coefficients for THD and f_{sw} respectively; f_{sw} is calculated by averaging the total switching operations of six gate drives within a period [50]. The coefficients are assigned based on the importance of each control objective. In the case study, priority is given more to reducing the THD, e.g., $k_{thd} = 2$ and $k_{sw} = 1$. The training process for the proposed DDPG-based weighting factor design of FCS-MPC is illustrated in **Algorithm 1**.

Algorithm 1 DDPG-based weighting factor design for FCS-MPC

Input: Measurements: $[v_{f\alpha}, v_{f\beta}, e_\alpha, e_\beta, THD]$
Output: Weighting factors λ_d and λ_{sw}
Initialize replay buffer $\mathcal{M}$
Initialize critic network Q and actor-network μ with
random weights θ_Q and θ_μ
Initialize target network $Q^{'}$ and actor-network $\mu^{'}$ with weights $\theta_Q^{'} = \theta_Q$, $\theta_\mu^{'} = \theta_\mu$
for episode = 1 to M **do**
  Initialize a random noise $\mathcal{N}$ for action exploration
  Receive initial observation at state s_1
  for iteration = 1 to T **do**
    Select action with current policy and noise
    Excute the action a_t, receive the reward r_t based on (2.31) or (2.32) and new state s_{t+1}
    Store tuple (s_t, a_t, r_t, s_{t+1}) in $\mathcal{M}$
    Randomly sample the mini-batch e_j from $\mathcal{M}$
    Set $y_j = r_j + \gamma Q^{'}(s_{i+1}, \mu^{'}(s_{i+1}|\theta_\mu)|\theta_Q^{'})$
    Update the critic by minimizing (2.24)
    Update the actor with (2.27)
    Update the target networks using (2.25)
  end for
end for

Source: [**J1**]

2.3.3 Unsupervised DQN-based imitation controller of FCS-MPC

A novel unsupervised, model-free RL (MFRL) imitation controller based on DQN is proposed to mitigate the model and parameter dependence of conventional FCS-MPC. This controller is capable of identifying the optimal switching state selection policy, effectively mimicking the strategy of FCS-MPC, thereby achieving the desired

control performance for power electronic converters, without prior system knowledge. Moreover, the proposed application framework has light online computational requirements for practical implementation.

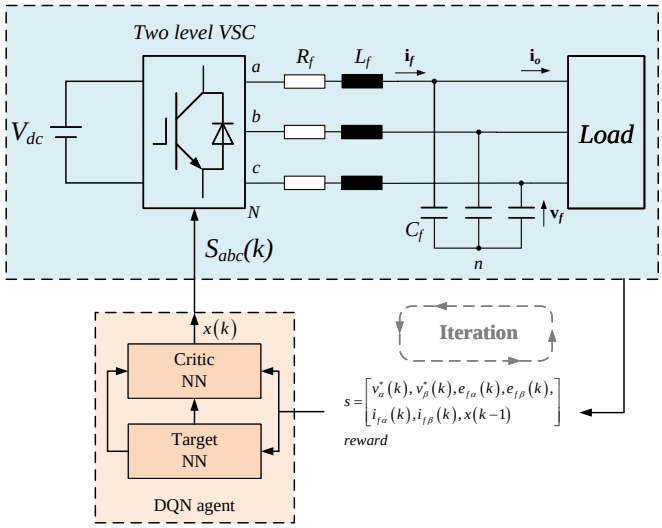


Figure 2.4. Block diagram of unsupervised MFRL imitation controller of MPC-controlled VSC. Source: [**J1**]

The schematic of the proposed DQN-based controller is presented in Fig. 2.4. The agent selects a voltage vector based on the observed system states and rewards received, subsequently employing the corresponding switching states to regulate the VSC. To achieve the desired control performances, the reward function could be designed based on the cost function of FCS-MPC, thereby incentivizing the agent to find the optimal switching strategy.

- **State and action sets**
  Similar to the input measurements of FCS-MPC, the observed states include the reference voltage ($v^*_{f\alpha}$, $v^*_{f\beta}$), voltage tracking error ($\triangle v_{f\alpha}$, $\triangle v_{f\beta}$), filter currents ($i_{f\alpha}$, $i_{f\beta}$) and the previously taken action, voltage vector x_{old}. As for the action space, because the switching states listed in Table 2.1 correspond to each voltage vector, the action space is defined as $x = [1 : 1 : 7]$.

- **Reward function design**
  This study aims to regulate the output voltage to track the AC reference voltage, where the optimal switching states are selected. The reward function should be finite to enable feasible and efficient training of the DQN agent. Therefore,

to approximate the control performance of conventional FCS-MPC, the basic single-step horizon cost function is formulated as the reward function for the DQN agent, as shown below

$$r = -((v^*_{f\alpha} - v_{f\alpha})^2 + (v^*_{f\beta} - v_{f\beta})^2). \tag{2.33}$$

Similarly, the reward function could also be designed to achieve different control objectives, akin to the cost function design of FCS-MPC.

The training process is illustrated in **Algorithm 2**.

Algorithm 2 DQN-based unsupervised imitation controller of FCS-MPC for power converter

Input: $[v^*_{f\alpha}, v^*_{f\beta}, \triangle v_{f\alpha}, \triangle v_{f\beta}, i_{f\alpha}, i_{f\beta}, x_{old}]$
Output: Voltage vector number x

Initialize replay buffer $\mathcal{M}$ to capacity R
Initialize action-value function Q with random weights θ_Q
Initialize target action-value function $Q^{'}$ with weights $\theta^{'}_Q = \theta_Q$
for episode = 1 to M **do**
  Receive initial observation at state s_1
  for iteration = 1 to T **do**
    With probability ϵ, select a random action a_t otherwise select $a_t = \underset{a}{max}\ Q_\pi(s_t, a|\theta_Q)$
    Execute the action a_t, observe the reward r_t calculated with (2.33) and state s_{t+1}
    Store the tuple (s_t, a_t, r_t, s_{t+1}) in $\mathcal{M}$
    Randomly select a mini-batch e_j from $\mathcal{M}$
    Set $y_j = \begin{cases} r_j, \text{if episode terminates at step } j+1 \\ r_j + \gamma max Q^{'}_\pi(s_{j+1}, a^{'}|\theta^{'}_Q), \text{otherwise} \end{cases}$
    Perform gradient descent on (2.24) with respect to network parameter θ_Q
    Every T_{target} steps, reset $\theta^{'}_Q$ to θ_Q
  end for
end for

Source: [**J1**]

2.4 Experimental validation

The optimal weighting factor design and switching state selection policies from the well-trained RL agents are validated experimentally based on the implementation

framework. The experimental setup, matching the simulation parameters listed in Table 2.2 is shown in Fig. 2.5, where a two-level VSC consisting of three IMPERIX PEB 8024 H-bridge power modules is interfaced with a linear load via a LC filter. An IMPERIX B-Box RCP control platform implements the well-trained RL agents.

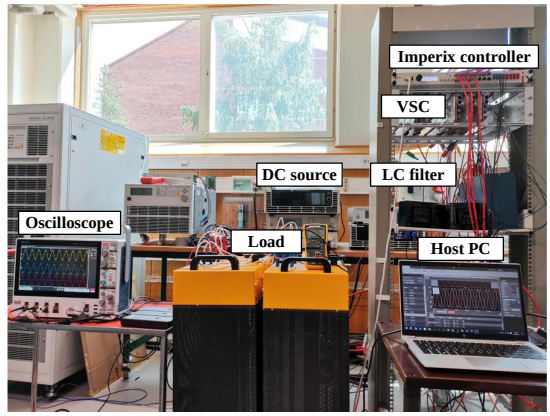


Figure 2.5. Experimental setup of a two-level VSC system. Source: [**J1**]

Table 2.2. Experimental Testbed Parameters

Parameter sets	**Parameter Values**
DC link voltage	$v_{dc} = 520\ V$
Output filter	$L_f = 2.5\ \ mH, R_f = 0.013\ \Omega, C_f = 30\ \mu F$
Reference voltage	$V_r = 200\ V, f_r = 50\ Hz$
Load	$R_{load} = 50\ \Omega$
Sampling time	$T_s = 50\ \mu s$
DDPG	Batch size = 256, Discount Factor = 0.995, Learning rate = 1e-3, Hidden layers/nodes: 2/20
DQN	Batch size = 128, Discount Factor = 0.98, Learning rate = 1e-3, Hidden layers/nodes: 2/20

2.4.1 From simulation to practical implementation of online RL

The proposed framework for RL-aided weighting factor design and MFRL controller is comprised of two main steps. First, the converter system, adhering to testbed parameters listed in Table 2.2, is simulated in the MATLAB/SIMULINK environment, where different RL algorithms are configured for different application scenarios. The RL agents learn autonomously to find the optimal policy through interaction with the system. In particular, the DDPG agent adjusts the weighting factors for the FCS-MPC to achieve the desired control performances. In addition, the DQN agent emulates the conventional FCS-MPC, utilizing the optimal switching selection policy for VSC and achieving the desired control performance without prior system knowl-

edge. In this research, all simulations were conducted in Matlab/Simulink with a PC Intel (R) Core™ i5-9500 CPU at 3.00 GHz and 32 GB of RAM.

Afterward, the optimal policies obtained from RL agents are transferred to the experimental setup for real-time demonstration using the IMPERIX platform. The training process on edge devices using the simulation model can remain consistent when applied to the actual power converter system. During the online implementation, the optimal policy is deployed, wherein the RL agent receives sensor measurements as input states for the actor network and outputs optimal actions. The workflow transferring the online RL from simulation to online demonstration is illustrated in Fig. 2.6.

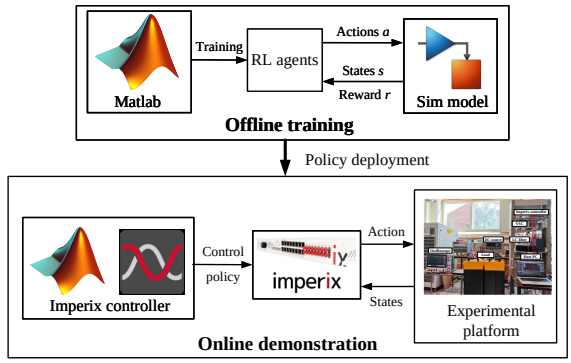


Figure 2.6. Diagram of the RL-based methods development process, including an online demonstration with the practical experimental setup. Source: [**J1**]

The computational complexity is critical for the practical implementation of the proposed approaches. Compared with conventional FCS-MPC, which has cubic complexity due to matrix operations, especially for multi-cell or multi-level converters, the computational burden of the demonstration framework for the proposed RL-based weighting factor design and controller comes from offline training at edge devices, and the time consumption is not critical. Once trained successfully, the RL-based methods involve only light computational matrix manipulation within the neural networks, enjoying linear computational complexity, which also has potential for more complex problems. In summary, regardless of the time-consuming training session, the practical implementation of the proposed RL-based approaches is quite tractable.

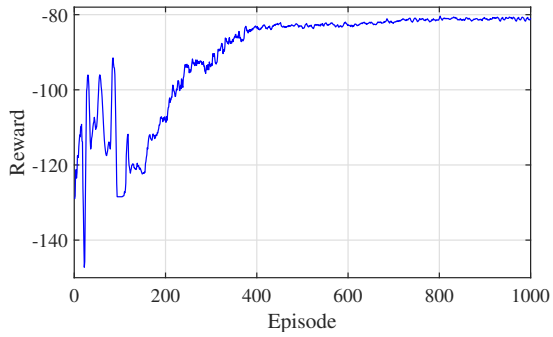
2.4.2 DDPG-based weighting factor design

The proposed RL-based weighting factor design approach is verified with two different case studies, as shown below.

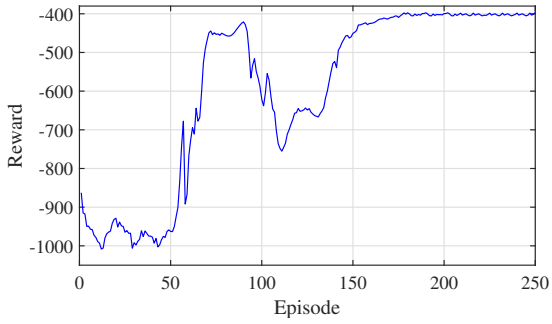
Scenario A: Minimizing the THD. In this scenario, the RL agent is trained using the reward (2.31), enabling it to identify the optimal weighting factors to minimize

the THD. The training result of the accumulated average reward over the episodes is shown in Fig. 2.7(a). At the initial stage, the RL agent keeps exploring the action space, resulting in fluctuations in the accumulated average reward for each training episode. Between episodes 180 and 400, the agent starts to balance exploration and exploitation, gradually approximating the optimal weighting factors. After 400 training episodes, the RL agent starts exploiting the optimal policy, leading to the convergence of the reward curve.

The optimal weighting factor design policy of the well-trained agent for THD minimization is deployed and applied to FCS-MPC-regulated VSC in the experimental setup, yielding the optimal weighting factors $\lambda_d = 1.38, \lambda_{sw} = 2$. The line-to-line capacitor voltage, both without and with the optimal weighting factors, is shown in Fig. 2.8(a) and Fig. 2.8(b), respectively. The resulting THD and average switching frequency with the RL-tuned weighting factors are 2.22% and 4.55 kHz, respectively, compared to 3.37% and 4.95 kHz without the weighting factors.



(a) Scenario A.

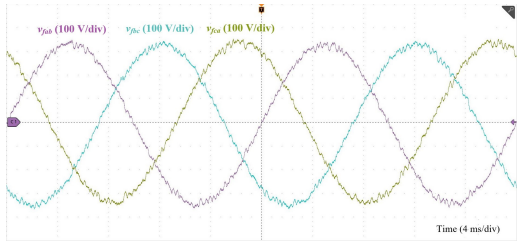


(b) Scenario B.

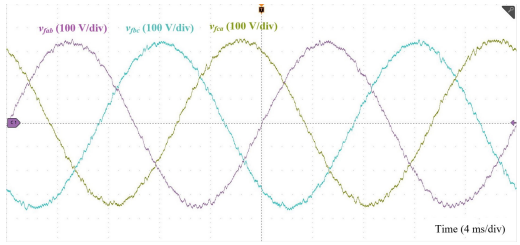
Figure 2.7. Average accumulated reward during the training process. Source: **[J1]**

Scenario B: Minimizing the THD while lowering the f_{sw}. To simultaneously reduce the THD with a lower average switching frequency, the RL agent achieves

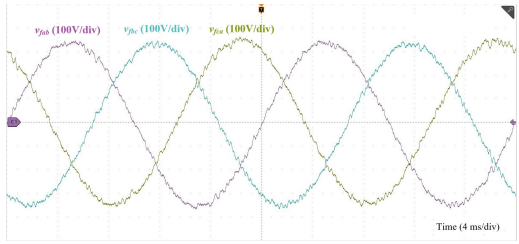
a trade-off between the two control objectives by tuning the weighting factors. The training process is shown in Fig. 2.7(b), illustrating how the agent undergoes exploration, learning, and exploitation phases to identify the optimal weighting factors.



(a) No weighting factor. THD = 3.37%, f_{sw} = 4.95 kHz.



(b) RL-tuned weighting factors, λ_d = 1.38, λ_{sw} = 2. THD = 2.22%, f_{sw} = 4.55 kHz.



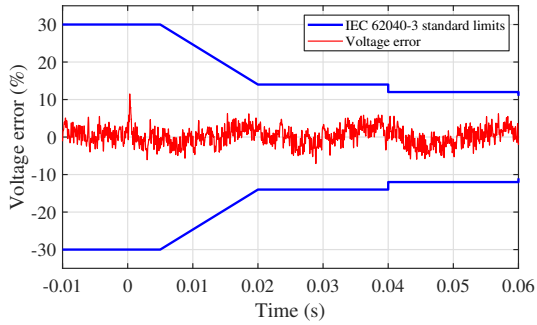
(c) RL-tuned weighting factors, λ_d = 0.75, λ_{sw} = 7.04. THD = 2.61%, f_{sw} = 3.62 kHz.

Figure 2.8. Experimental line-line capacitor voltage waveforms under different conditions. Source: **[J1]**

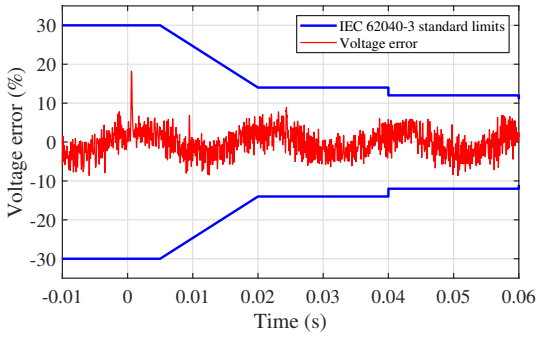
The optimal weighting factor design policy extracted from the RL agent is deployed and validated in a practical FCS-MPC controlled converter system, where the optimal weighting factors are $\lambda_d = 0.75, \lambda_{sw} = 7.04$. The results of the line-to-line capacitor voltage are shown in Fig. 2.8(c), and the calculated THD and average switching frequency with the RL-tuned weighting factors are 2.61% and 3.62 kHz.

These results demonstrate enhanced control performance achieved through the

automatic RL-based weighting factor design method. In addition, a 100% linear load step test is performed to demonstrate the negligible impact of weighting factors on the transient performance of the system. The experimental results are shown in Fig. 2.9. It can be observed that the voltage error in both cases converges immediately when the load step is applied. Thus, the root mean square (RMS) of voltage deviation also converges promptly within the limit. As a result, the voltage deviation complies with the limits specified in the standard IEC 62040-3.



(a) 100% load step with $\lambda_d = 1.38, \lambda_{sw} = 2$.



(b) 100% load step with $\lambda_d = 0.75, \lambda_{sw} = 7.04$.

Figure 2.9. Transient performance under 100% load step with the RL-tuned optimal weighting factors. Source: **[J1]**

2.4.3 MFRL-based optimal controller for VSC

The training process of the DQN agent, aimed at emulating conventional FCS-MPC control for VSC, is depicted in Fig. 2.10. Initially, the DQN agent explores the action space by randomly selecting switching states without any prior system knowledge. This results in the regulated output voltage deviating significantly from the desired reference voltage, causing a sharp drop in the average accumulated reward during

the early training episodes. Following the initial exploration phase, the DQN agent starts to balance between exploration and exploitation. Guided by the rewards from the single-step horizon cost function, the agent learns to mimic the optimal switching strategy of conventional FCS-MPC, correlating input states with optimal switching actions determined by the cost function. Ultimately, the DQN agent exploits the discovered optimal switching policy for regulating the VSC, leading to the convergence of the accumulated average reward.

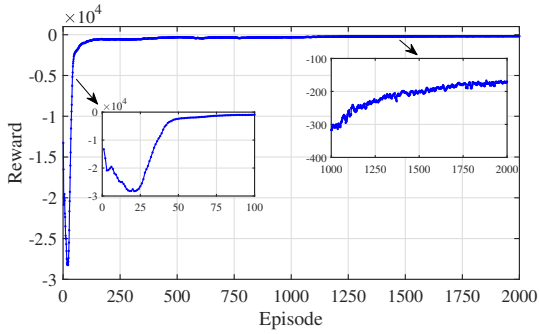


Figure 2.10. Average accumulated reward during the training process. Source: [**J1**]

The optimal control policy is validated experimentally with the proposed deployment framework, and the results are shown in Fig. 2.11. It can be observed that the differences in THD and f_{sw} between the conventional FCS-MPC without weighting factors and the MFRL-based imitation controller are 0.12% and 0.37 kHz, respectively, confirming effective imitation learning of FCS-MPC for VSC. More importantly, the unsupervised DQN-based controller does not require any prior knowledge of the system, unlike conventional FCS-MPC [62] and the supervised ANN-based imitator [18] for power converter control. A reference voltage step is applied to validate the transient performance of the DQN-based controller, as presented in Fig. 2.12. It can be observed that the proposed RL-based controller can still work properly when a reference voltage step is applied, similar to the conventional FCS-MPC.

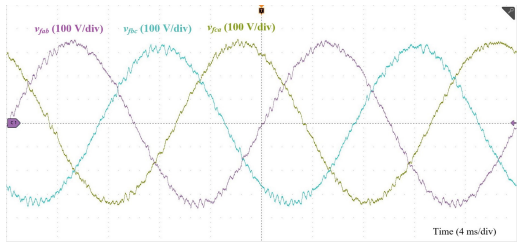
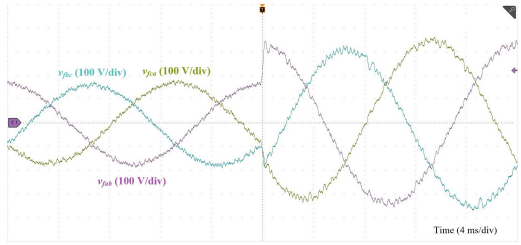
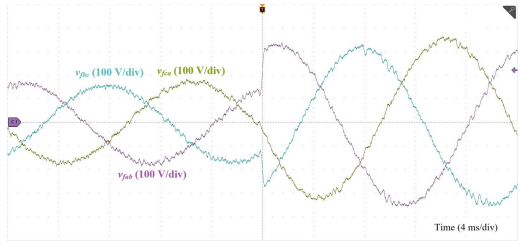


Figure 2.11. MFRL-based unsupervised imitation controller. THD = 3.49%, f_{sw} = 4.58 kHz. Source: [**J1**]



(a) Reference voltage step for FCS-MPC.



(b) Reference voltage step for RL-based controller.

Figure 2.12. Output voltage when applying reference voltage step (100V $\rightarrow$ 200 V). Source: [**J1**]

2.5 Safety-enhanced online self-learning optimal control for power converters

To learn the optimal control for power electronic converters, the RL agent engages in random exploration within the learning space, utilizing available discrete actions. The training process can be time-consuming, with the agent potentially requiring an extended period to discover the optimal control policy. Moreover, system safety is not assured, particularly during the random exploration phase, making online self-learning of optimal power converter controls impractical. To tackle the system safety challenge in training the RL, recent advancements in safe RL frameworks have gained attention [20, 63, 64]. Nonetheless, safe RL-based control for power electronic converters remains an open research question.

Therefore, on top of the DQN-based optimal control strategy for power converters presented in Section 2.4.3, a novel safety framework is introduced to ensure the system's safety. This safe learning framework merges an MPC-based safety filter with the self-learning optimal controller, aiming to achieve control performance on par with conventional FCS-MPC, maintain physical limits, and enhance training efficiency significantly. The schematic of the proposed method is presented in Fig. 2.13.

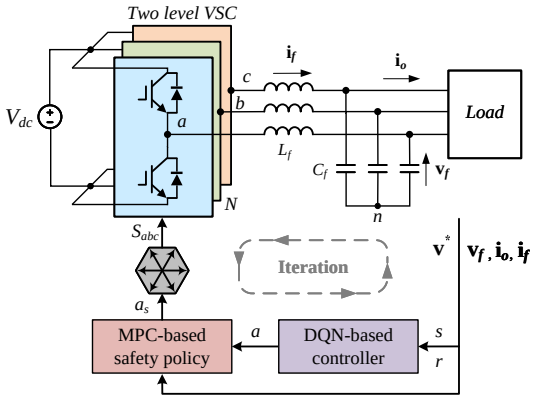


Figure 2.13. Schematic of the safe learning-based control for VSC. Source: [**J2**]

In practice, maintaining the converter-side current within a threshold is essential for ensuring hardware safety. A single-step prediction horizon MPC safety framework is introduced to achieve this during the learning process of the RL agent, as shown in Fig. 2.13. The DQN action is input to the MPC-based safety filter, which assesses potential overcurrent based on the selected switching states, as

$$||i^a_{f\alpha\beta}(k+1)||_2 \leq i_{max}. \tag{2.34}$$

Infeasible actions, corresponding to switching states that cause overcurrent, are excluded and substituted with the MPC-based safety block. At the same time, after certain training episodes, the RL agent gradually converges to the safe learning region, eventually bypassing the safety block. In this way, the safety policy also avoids unnecessary unsafe learning spaces, thereby enhancing the training efficiency of the DQN-based controller. The proposed safe online RL-based controller is detailed in **Algorithm 3**.

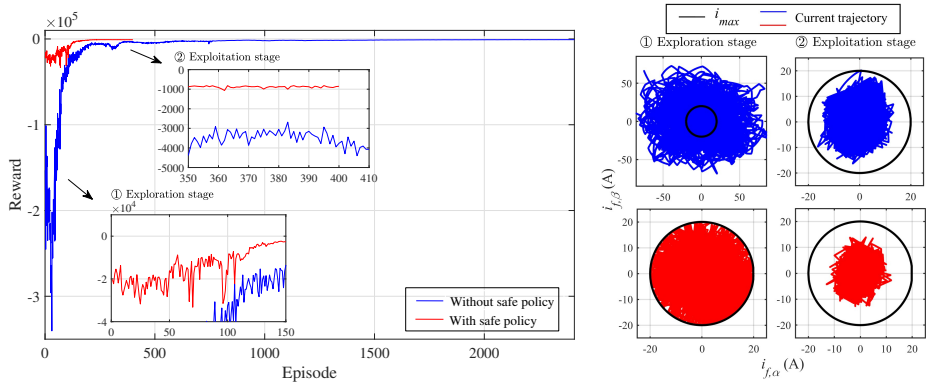
The training process of the safe DQN-based controller for VSC is illustrated in Fig. 2.14. i_{max} is set to 20 A for the safety of system hardware. During initial training episodes, the method with the safe policy learns the optimal switching state selection policy quickly without damaging the converter, as depicted in Fig. 2.14(b), with the inductor current trajectory remaining within limits. In comparison, the agent without the safe policy explores randomly, taking more time to converge to the optimal policy and failing to maintain system limits during initial training.

Similarly, the proposed demonstration framework transfers the safe RL-based controller from the edge device to a practical converter for experimental validation of the optimal control policy. The experimental setup follows the parameters listed in Table 2.2 with the exception of a sample time set to $T_s = 20\,\mu s$. The results, shown in Fig. 2.15, indicate that the proposed safe DQN-based controller attains control performance comparable to that of conventional FCS-MPC.

Algorithm 3 Safe online DQN-based controller

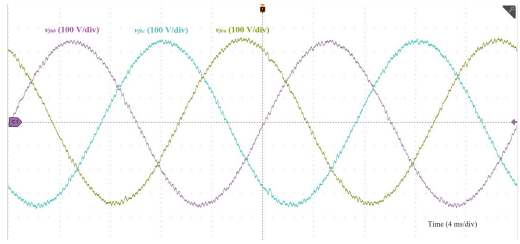
Initialize replay buffer $\mathcal{M}$
Initialize action-value functions Q and $Q^{'}$ with random weights $\theta_Q^{'} = \theta_Q$
for episode = 1 to M **do**
  Receive initial observation at state s_1
  for iteration $t = 1$ to T **do**
    With probability ϵ, select a random action a_t, otherwise $a_t = \underset{a}{max}\ Q_\pi(s_t, a|\theta_Q)$
    Predict filter current $i_{f\alpha\beta}^{a_t}(k+1)$ using (2.13)
    if (2.34) is satisfied **then**
      Output the action a_t
    else
      Employ safe action from safety block: $a_t \leftarrow a_{s,t}$
    end if
    Execute the action a_t, observe the reward r_t calculated with (2.33) and state s_{t+1}
    Store the tuple (s_t, a_t, r_t, s_{t+1}) in $\mathcal{M}$
    Randomly select a mini-batch e_j from $\mathcal{M}$
    Set y_j with (2.23), perform gradient descent on (2.24) regarding the network parameter θ_Q
    Every T_{target} steps, reset $\theta_Q^{'}$ to θ_Q
  end for
end for

Source: [**J2**]

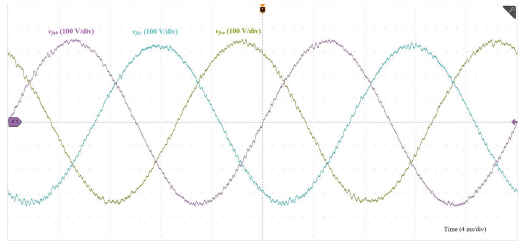


(a) Average accumulated reward **(b)** Converter-side current trajectory

Figure 2.14. Training process of the proposed safety-enhanced self-learning optimal controller. Source: [**J2**]



(a) Conventional FCS-MPC. THD = 1.97%.



(b) Safe DQN-based controller. THD = 2.19%.

Figure 2.15. Performance comparison of different controllers. Source: [**J2**]

In general, parameter variations in the inductor and capacitor may arise due to factors such as aging, temperature, and voltage. Variations in model parameters would deteriorate the control performance of conventional FCS-MPC. In contrast, the proposed safe RL learning-based controller is model-free, and the safe policy framework only functions during the initial exploration stage to ensure the safety of the converter. Once exploration converges within the safe region, the RL-based controller can autonomously update, learn, and identify the optimal switching strategy to achieve the desired control performance. To test the robustness of the safe policy for RL learning-based control, variations in inductance ($\triangle L_f$) and capacitance ($\triangle C_f$) within $\pm$ 30% of their nominal values are implemented.

If the safety framework underestimates the inductance value ($\triangle L_f > 0$), the safe exploration region is more stringently assured, as this reduces the filter current peaks and converter-side reference current, as derived in (2.34). Consequently, the agent converges more rapidly, as the reduction in unsafe learning space minimizes the risk of overcurrent. On the other hand, if the safety framework overestimates the inductance value ($\triangle L_f < 0$), the filter current peaks may exceed the maximum value while still within the physical limits by properly leaving a margin between the predefined maximum current and the physically limited current. Most importantly, current trajectories center on a safe learning space during the learning process. Conversely, the LC filter is indirectly controlled by the inverter voltage due to the cross-coupling effect between the inductor and capacitor. Thus, capacitance uncertainty barely influences the safe exploration of the agent. In summary, the proposed RL-based controller effectively converges to the optimal control policy for the converter, despite

parameter uncertainties.

2.6 Summary

This chapter provides a comprehensive guideline for applying RL to controller design and unsupervised self-learning for the optimal control of power converter controllers. This includes steps from problem formulation and RL agent type selection to reward design and policy deployment in a practical converter testbed. Specifically, a DDPG-based automatic weighting factor tuning method and a DQN-based imitation controller emulating the FCS-MPC are proposed. The proposed methods are model-free, avoiding the need for prior system knowledge while inheriting the superior performance characteristics of conventional FCS-MPC. Moreover, a safety-enhanced self-learning framework for optimal power converter control is proposed, combining computationally light single-step predictive control with the RL-based controller framework. This approach ensures system safety during the training sessions and simultaneously enhances learning efficiency. In addition, a demonstration framework to transfer online RL agents from simulation to a practical setup is proposed, which is used to validate the proposed methods. The results show that the RL-aided weighting factor design for FCS-MPC achieves the desired control performance. In addition, the results also confirm that the RL-based controllers attain control performance comparable to that of conventional FCS-MPC.

Based on publications

- [**J1**] **Y. Wan**, T. Dragicevic and X. Qian, "Unsupervised learning-based Predictive Control for Power Electronic Converters," in IEEE Transactions on Industrial Electronics, under review.
- [**J2**] **Y. Wan**, T. Dragicevic and X. Qian, "Safety-Enhanced Self-Learning for Optimal Power Converter Control," in IEEE Transactions on Industrial Electronics, under review.
- [**C1**] **Y. Wan**, T. Dragicevic, N. Mijatovic, C. Li and J. Rodriguez, "Reinforcement Learning Based Weighting Factor Design of Model Predictive Control for Power Electronic Converters," IEEE PRECEDE, Shandong, China, 2021, pp. 738-743.

CHAPTER 3

Data-driven intelligent attack generation and detection in DC microgrids

3.1 Introduction

In DC microgrids, different DG units are scattered across geographical areas, which need to be coordinated for efficient system operations. Various units are connected to DC microgrids exclusively via interfacing converters, which should work in harmony with properly developed control strategies. In particular, the control framework in DC microgrids consists of local control for independent units and coordinated control among the units, where accurate power sharing and voltage regulation should be achieved.

Among different control strategies in DC microgrids, distributed control is prevalent primarily due to its notable advantages in scalability, reliability, and efficiency [65]. The communication network for distributed control makes the DC microgrid a cyber-physical system that is susceptible to cyber-attacks [66]. The primary concern of the chapter is the most common attack type, false data injection attack (FDIA), which manipulates the information in sensors or communication links. If not detected and mitigated in a timely manner, the DC microgrids could become unstable.

Different detection approaches are proposed to address this, which can be categorized into model-based and model-free detection methods [67]. Model-based methods suffer from discrepancies between the established model and real-world power electronic systems. Among model-free methods, data-driven detection methods utilize system measurements for attack detection, mitigating model dependency. Various NN-based detection methods are proposed for different FDIAs based on signal estimation with various machine learning algorithms [29, 68, 69]. However, the data is collected from the system operating in a healthy mode, and they do not explicitly incorporate knowledge of various attack types, resulting in their inability to detect various types of FDIAs. Another data-driven detection method is proposed to detect stealthy cyber-attacks in DC microgrids [27]. In addition, a metric-based detection method is proposed, introducing a discordant element (DE) term, which is calculated based on the input currents of neighboring units [23]. Similarly, other metrics are pro-

posed to detect attacks in [70, 71]. Nonetheless, even these cutting-edge index-based detection techniques face a critical challenge when confronted with astute adversaries introducing innovative attack patterns. This encompasses a broader spectrum of both coordinated and uncoordinated attacks, achieved by penetrating sensors and communication links across multiple nodes in various ways. Such intelligent and deceptive behaviors can be emulated using RL.

Therefore, in this chapter, a novel data-driven intelligent cyber-attack generation and detection framework is developed. RL is first employed to identify the vulnerability of DC microgrids under the protection of the DE-based detection scheme by generating novel and stealthy attack patterns. The RL attacker can autonomously learn to find intelligent attacks that can bypass metric-based detection approaches. Conversely, a data-driven cyber-attack detection method is proposed to complement the metric-based detection approach so that RL-based attacks can also be detected. Moreover, the proposed cyber-attack generation and detection framework is experimentally validated. Finally, insight is provided to identify and detect a broader spectrum of attacks in DC microgrids with the proposed framework.

3.2 Distributed controlled DC microgrids

A cyber-physical DC microgrid, shown in Fig. 3.1, is studied in this chapter. As mentioned, different units connect to the DC microgrid via interfacing converters, where primary and secondary controls are employed to achieve accurate current sharing and voltage regulation.

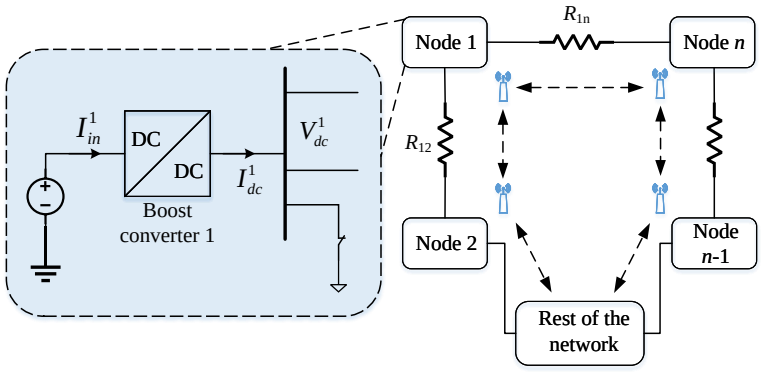


Figure 3.1. Cyber-physical DC microgrid with n nodes.

Using the general rule of consensus to apply distributed control over the term ϕ between the n interconnected DC-DC power generation units, the resultant regulatory

term (u_i) at node i can be defined as:

$$u_i = \sum_{j=1}^{m} a_{ij}(\phi_j - \phi_i), \quad where \;\; a_{ij} = \begin{cases} > 0, \; if(x_i, x_j) \in \boldsymbol{M} \\ 0, \; else \end{cases} \tag{3.1}$$

where a_{ij} denotes the connection between nodes in the adjacency matrix $\boldsymbol{A} = [a_{ij}]_{n \times n}$, $\boldsymbol{M}$ represents the underlying communication topology, and x_i and x_j are the local and neighboring nodes.

Denoting the received data matrices as $\boldsymbol{Y_{in}}$, balanced Laplacian matrices of $\boldsymbol{L} = \boldsymbol{Y_{in}} - \boldsymbol{A}$ are obtained when both $\boldsymbol{Y_{in}}$ and $\boldsymbol{A}$ are identical, and the elements of matrix $\boldsymbol{L}$ are:

$$l_{ij} = \begin{cases} deg(n_i), \; i = j \\ -1, \;\; i \neq j \\ 0, \quad otherwise \end{cases} \tag{3.2}$$

where $deg(n_i)$ denotes the degree of agent i.

Under distributed regulation, the synchronization of the distributed state variables $\boldsymbol{X}$ between neighboring agents can be achieved using $\dot{\boldsymbol{X}} = -\boldsymbol{L}\boldsymbol{X}$, where, on a properly connected network, convergence to the constant value c is ensured, and this can be represented by $lim_{t \to \infty} x_i(t) = c$.

To enable distributed secondary regulation over the voltage and current terms at node i in an autonomous DC microgrid, it is necessary to modify the local voltage setpoint $V^i_{dc_{ref}}$ as follows:

$$V^i_{dc_{ref}} = V_{dc_{ref}} + \triangle V^i_1 + \triangle V^i_2 \tag{3.3}$$

where $V_{dc_{ref}}$ denotes the global reference voltage, and $\triangle V^i_1$ and $\triangle V^i_2$ are the resultant regulatory terms from the secondary voltage and current controllers, respectively, which can be formulated as

$$\triangle V^i_1 = (K^V_p + \frac{K^V_i}{s}) \cdot (V_{dc_{ref}} - u^V_i) \tag{3.4}$$

$$\triangle V^i_2 = (K^I_p + \frac{K^I_i}{s}) \cdot (I_{dc_{ref}} - u^I_i). \tag{3.5}$$

where u^V_i and u^I_i are the consensus terms for voltage and current, K^V_p and K^V_i, K^I_p and K^I_i are the proportional and integral gains for PI controllers G_v and G_c, respectively, regarding the voltage and current loops in the primary control layer, and $I_{dc_{ref}}$ is the global reference current, set as 0 for proportionate current sharing.

As shown in Fig. 3.2, $\overline{V}^i_{dc}$ is the average voltage for node i, which is updated based on the consensus law, as

$$\overline{V}^i_{dc}(t) = V^i_{dc}(t) + \int_0^t \sum_{j \in M_i} a_{ij}(\overline{V}^j_{dc}(\tau) - \overline{V}^i_{dc}(\tau))d\tau. \tag{3.6}$$

where M_i is the set of neighbors of node i. Similarly, the calculation of the consensus current $\bar{I}_{dc}^{i}$ can be expressed as

$$\bar{I}_{dc}^{i} = \sum_{j \in M_i} a_{ij} \left(\frac{I_{dc}^{j}}{I_{max}^{j}} - \frac{I_{dc}^{i}}{I_{max}^{i}} \right). \tag{3.7}$$

where $\bar{I}_{dc}^{i}$ is the consensus current, I_{dc}^{i} denotes the output current of node i, I_{dc}^{j} denotes the distributed current from neighboring nodes for node i, I_{max}^{i} denotes the maximum output current allowed for the ith converter, which is the same for all four converters in our case study.

The overall distributed control structure is presented below.

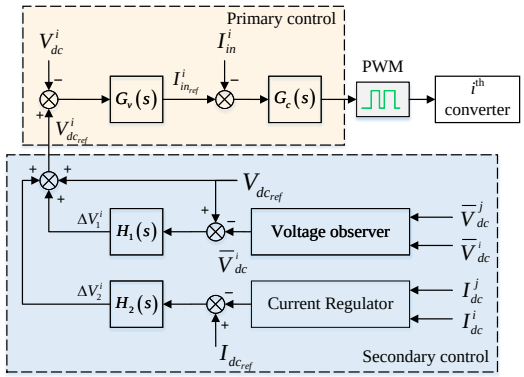


Figure 3.2. Schematic of the distributed secondary control layer.

3.3 Cyber-attack generation and detection in DC microgrids

3.3.1 Cyber-physical DC Microgrids protected with metric-based detection scheme

As mentioned, FDIAs occur when false signals are injected into the sensors or communication links to compromise the normal operations of DC microgrids. Thus, the attacked signals are expressed as

$$\boldsymbol{I_A} = \begin{cases} I_a^i = I_{dc}^i + k_i I_{fi}, & \text{sensor attack} \\ I_a^{ij} = I_{dc}^j + k_{ij} I_{fj}, & \text{link attack} \end{cases} \tag{3.8}$$

where $\boldsymbol{I_A}$ denotes the attacked signal vector for multiple nodes, I_a^i is the output current from node i, and I_{dc}^i is the actual current sensor measurement for node i.

I_a^{ij} denotes the communication current for neighboring nodes i and j, I_{fi} and I_{fj} represent the false signals, which penetrate into the corresponding points by the attacker; their presences are represented by unity value for the coefficients k_i and k_{ij}, or 0 vice versa.

Different attack patterns are created by the attackers in various ways, which could be broadly categorized into stealthy and destabilization attacks [23]. In particular, the stealthy attack occurs when the attacker injects false signals into the sensors and communication links coordinately, where the system seems to operate normally as

$$\lim_{t\to+\infty} \bar{V}_{dc}^{i}(t) = V_{dc_{ref}}, \lim_{t\to+\infty} \bar{I}_{dc}^{i} = 0. \tag{3.9}$$

For the destabilization attack, the system stability is compromised due to the false signals, where the system is regulated at a different voltage level without current sharing, expressed as

$$\lim_{t\to+\infty} \bar{V}_{dc}^{i}(t) = V_{dc_{ref}}^{a}, \lim_{t\to+\infty} \bar{I}_{dc}^{i} \neq 0. \tag{3.10}$$

Among different detection schemes, the discordant detection algorithm has exhibited excellent performance in detecting both types of FDIAs [23]. In general, the detection method rests on the synchrony between neighboring reference current terms due to the discordance between the compromised node and healthy neighboring nodes. The metric for attack detection is expressed as

$$DE_i = l_i[\sum_{j\in M_i}(I_{in_{ref}}^{j} - I_{in_{ref}}^{i})][\sum_{j\epsilon M_i}(I_{in_{ref}}^{j} + I_{in_{ref}}^{i})] \tag{3.11}$$

where DE_i denotes the formulated DEs for node i, $I_{in_{ref}}$ is the input reference current from the outer voltage control loop in the primary control layer, l_i is a positive coefficient to increase/decrease the value of DE_i.

According to [23], due to the discordance between the attacked nodes and healthy nodes, a positive increase of the DE value above zero indicates whether the node is attacked or not. In a practical system, considering sensor noise, attacks are detected by comparing the DE value with a threshold value for the DE term obtained under normal operation, which is expressed below.

$$DE_i = \begin{cases} < DE_{min}, & if\ k_i \& k_{ij} = 0 \\ > DE_{min}, & if\ k_i \| k_{ij} \neq 0. \end{cases} \tag{3.12}$$

3.3.2 Multi-agent Reinforcement Learning (MARL) for FDIA generation

To automatically uncover the vulnerabilities of metric-based detection approaches, the autonomous coordinated attack pattern generation process could also be modeled as an MDP. While a similar approach can target all index-based detection schemes for different types of attacks, such as distributed screening (DS) [72] and diverging

factor (DF) [73], more details are explained in the study for the discordant algorithm integrated into an autonomous DC microgrid with n interconnected nodes, as shown in Fig. 3.3. Due to the distributed control structure in DC microgrids, the MARL attacker is implemented in a distributed manner.

For distributed controlled DC microgrids, the false signals in the local current sensor measurements I^i_{dc} or distributed current terms I^j_{dc} would lead to an offset in the term in (3.3), deviating the reference voltage $V^i_{dc_{ref}}$. As a result, input current references $I_{in_{ref}}$ change for the nodes, and the corresponding DE value also changes. As is expressed in equation (3.12), any evident increase in the value above the threshold will reflect an attack on the current counterparts of the ith node. Thus, an intelligent attacker must harmonize and synchronize this offset for local nodes by generating stealthy attack patterns to bypass the conventional DE-based detection method.

In addition, in distributed cooperative control of the system, shown in Fig. 3.3, any deviation in current distributed terms I^j_{dc} or local terms I^i_{dc} would create an offset in term $\triangle V^i_2$, deviating the voltage setpoint $V^i_{dc_{ref}}$ from the secondary control layer to the local control layer. Thus, the intelligent attacker can harmonize and synchronize this offset between the neighboring nodes by targeting multiple nodes, and such a stealthy attack for the discordant detection algorithm is generated. This way, a sophisticated attack that the conventional DE-based detection algorithm cannot detect is generated. This intelligent attacker for stealthy attack generation could be emulated by MARL.

In a DC microgrid shown in Fig. 3.1, for each node with m incoming links, the attacker agents are defined as $\{CA_1, ..., CA_{m+1}\}$. Since the agents generate attack signals based on variations of DE value, the input observations of each agent thus

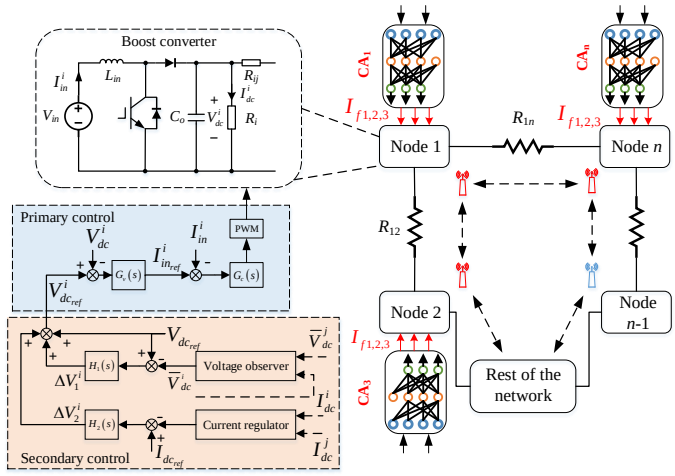


Figure 3.3. Configuration of DC microgrid under MARL-based cyber-attack. Source: [**J3**]

incorporate the DE value and the integral of neighboring nodes, represented by $\boldsymbol{S_i^t} = \{\{DE_1^t, ..., DE_{m+1}^t\}, \{\int DE_1^t, ..., \int DE_{m+1}^t\}\}$. The actions of each attacker agent are intrusion signals either for local sensors or communication links, $\boldsymbol{A_i^t} = \{I_{f1}^t, ..., I_{fm}^t\}$, where I_{f1}^t is the attack signal on the local sensor, and the rest are on incoming communication links.

The reward function design is critical for stealthy FDIA generation. The RL attacker aims to explore different attack patterns to suppress the DE value and bypass the DE-based detection method. In addition, DE variations would be induced during loading conditions, which should also be minimized so that the generated attacks could still minimize the DE value to a satisfactorily low level. Therefore, the general reward function is designed as

$$\begin{aligned} r_i^t = &- (k_{DE} \sum_{i=1}^{m+1} (DE_i^t)^2 + k_{\dot{DE}} \sum_{i=1}^{m+1} (\dot{DE_i^t})^2 \\ &+ k_{\dot{I}_f} \sum_{i=1}^{m} (\dot{I}_{fi}^{t-1})^2) + r_{dis}^t \end{aligned} \tag{3.13}$$

where k_{DE} and $k_{\dot{DE}}$ are the coefficients for the summation of DE values for all the neighboring nodes and their derivatives, respectively. $k_{\dot{I}_f}$ is the coefficient for the summation of the derivative of attack actions taken in the last time step $t-1$, which is tuned to minimize the variations of the generated attack signals, especially when a desirable stealthy attack is generated. In addition, a discrete reward r_{dis}^t is designed for the agents, expressed as

$$r_{dis}^t = -(k_1 \cdot r_1^t + k_2 \cdot r_2^t) \tag{3.14}$$

$$r_1^t = |\sum_{j=1}^{m} (I^{t-1}{}_{fj} - I^{t-1}{}_{fi})| < I_{f_{min}} \tag{3.15}$$

$$r_2^t = \Big((DE_1^t | ... | DE_i^t) > DE_{max} \Big) \Big|_{i=1}^{m+1}. \tag{3.16}$$

The term r_1^t indicates the attack signals on sensor and communication links are canceling each other due to the calculation of the consensus current, which should be avoided as this attack pattern will be minimized or mitigated within the system. k_1 is the coefficient to ensure the presence of the minimum non-canceling destabilizing cyber-attack, denoted by $I_{f_{min}}$, with the generated attack signals at time $t-1$. In addition, the reward term r_2^t is designed to further penalize the excessive DE value during the training stage, where a coefficient k_2 is assigned. DE_{max} is the upper threshold for discordant terms. It is noteworthy that $I_{f_{min}}$ is chosen considering a trade-off between the ramp slope for the destabilizing phenomenon under the generated attacks and the minimum DE threshold DE_{min}.

Therefore, during the training process, the RL agents will learn autonomously to generate destabilization FDIAs to minimize the DE value so that the attacks remain undetected by the DE-based detection method.

DQN is employed due to its computationally efficient characteristics [59]. The attacker agents are trained centrally and implemented in a distributed manner at each node. The whole training process is presented as follows:

Algorithm 4 Multi-agent RL-based FDIA

Input: DE and $\int DE$ of neighboring nodes
Output: Attack signals $\{I_{f1}^t, ..., I_{fm}^t\}$

Initialize replay buffer $\mathcal{D}$ to capacity R
Initialize action-value function Q with random weights θ_i^Q
Initialize target action-value function $\hat{Q}$ with weights $\theta_i^{Q'} = \theta_i^Q$
for episode = 1 to M **do**
  Receive initial observation at state s_i^1
  for iteration = 1 to T **do**
    For each agent i, select and execute action a_i^t with respect to policy $\pi(a_i^t|s_i^t)$, receive the reward r_i^t calculated with (3.13) and transition into state s_i^{t+1}
    Store tuple $(s_i^t, a_i^t, r_i^t, s_i^{t+1})$ in the $\mathcal{D}$
    for each agent $i = 1$ to m **do**
      Randomly select the mini-batch e from $\mathcal{D}$
      Set $y_j = \begin{cases} r_j, \text{if episode terminates at step } j+1 \\ r_j + \gamma max Q'_\pi(s_{j+1}, a'|\theta'_Q), \text{otherwise} \end{cases}$
      Perform gradient descent on (2.24) with (2.23) regarding the network parameter θ_i^Q
    end for
    Update the target network using (2.28)
  end for
end for

Source: [**J3**]

3.3.3 Data-driven cyber-attack detector

As the DE-based detection method fails to detect the sophisticated attack patterns generated by the MARL-based attackers, a data-driven detection method is proposed as a supplementary attack detector. The proposed data-driven cyber-attack detection framework is presented in Fig. 3.4. The proposed method aims to detect the attacks and identify the attacked nodes by extracting the mapping relationship between the input measurements and the target labels, consisting of offline model training and online cyber-attack detection steps.

The RL-based attacker can exploit vulnerabilities of the DE-based detection method by generating novel attack patterns. Therefore, the proposed supplementary data-driven detector aims to complement the DE-based detection method under RL-based

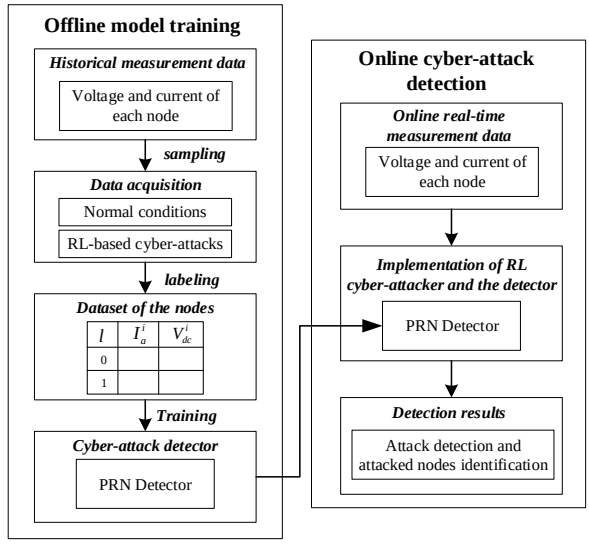


Figure 3.4. Data-driven cyber-attack detection method. Source: **[J3]**

attacks by detecting and identifying the attacked node, formulated as a classification problem.

As mentioned, the measurements under both healthy and attacked conditions are employed for training the data-driven attack detector. To relieve the computational burden of online implementation, a pattern recognition network (PRN) for classifying the inputs into target classes is employed. Similarly, different machine learning models or artificial neural networks could also be employed in the same way to detect RL-based attacks in DC microgrids. The structure of the PRN is shown in Fig. 3.5, where the corresponding mathematical description is expressed as

$$Y = F(X_{in}) = f_{out}[f_{hid}(b_{hid} + W_{hid}X_{in})W_{out} + b_{out}]. \tag{3.17}$$

where f, W, and b denote the activation function, weight matrix, and bias matrix, respectively, $X_{in} = \{x_1, ..., x_n\}$ represents the input vector, and the subscripts *hid* and *out* denote the hidden layer and output layer.

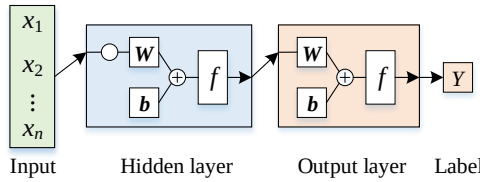


Figure 3.5. Structure of PRN. Source: **[J3]**

3.4 Experimental results

To validate the proposed RL-based attack generation and data-driven attack detection framework, a DC microgrid with $n = 4$ nodes in Fig. 3.1 is constructed and controlled via the distributed control method, shown in Fig. 3.6. Each node consists of a DC source and an interfacing boost converter, interconnected via tie lines. The proposed methods are implemented using the dSPACE MicroLabBox DS1202, where a computer is employed as the real-time control interface. The DC microgrid and controller parameters are listed in Table 3.1.

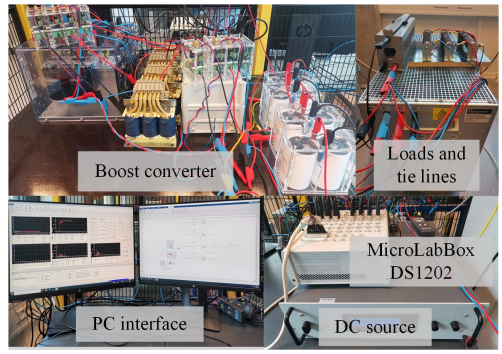


Figure 3.6. Experimental setup of a DC microgrid with four nodes. Source: **[J3]**

Table 3.1. Experimental Setup Parameters. Source: **[J3]**

Parameter sets	Values
Converter	$L_{in} = 0.86\ mH, C_o = 1.1\ mF, f_s = 10\ kHz, I_{rated} = 32\ A$ Loads: $R_1 = R_2 = R_3 = R_4 = 30.6\ \Omega$ Tie lines: $R_{12} = R_{23} = R_{34} = 0.5\ \Omega, R_{14} = 0\ \Omega$
Controller	$V_{in} = 48\ V, V_{dc_{ref}} = 60\ V, I_{dc_{ref}} = 0$ Primary layer: $K_p^V = 1, K_V^i = 20, K_p^I = 2.4, K_i^I = 10$ Secondary layer: $K_p^I = 0.12, K_i^I = 0.15$

As mentioned, to generate novel attack patterns that remain stealthy to the conventional DE-based detection method, the RL-based attackers target the current sensor signals of neighboring nodes (nodes 1, 2, and 4). The training process is performed in the Matlab/Simulink environment with a run-time period of 5 s for each episode. Then, the trained agents are implemented in the experimental test bed using dSPACE MicroLabBox DS1202 with a sampling time of 50 μs, where primary and secondary controllers are also deployed. In terms of communication delays, the primary to secondary delay t_{p-s} and the cyber-attack output delay t_{CA} are set at 5 ms and 40 ms, respectively, and the secondary to secondary delay is set with $80-100\ ms$.

As for the data-driven attack detector, operation data collected under healthy DC microgrid and attacked operation scenarios through the computer interface are

employed for training, where corresponding labels are assigned to the operating conditions, respectively. 80% of the randomly divided dataset was used to train the PRN, and 10% was used for validation and testing, respectively. A confusion matrix in Fig. 3.7 presents the performance of the PRN. The classification accuracy of the detector for classifying the inputs to the target labels is about 98.3%. In addition, to evaluate the employed machine learning model (PRN), classical 10-fold cross-validation is carried out, where the collected dataset for training and testing is randomly and repeatedly assigned [74]. The average classification accuracy is about 98.5%, which verifies the effectiveness of deploying PRN for attack detection.

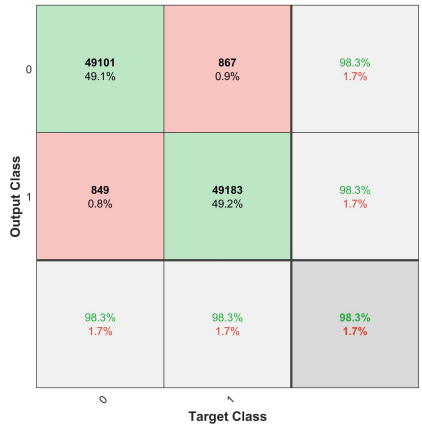


Figure 3.7. Training results of the PRN for cyber-attack detection (1:Attacked; 0:Healthy). Source: **[J3]**

3.4.1 DC microgrids protected by DE-based detection method

To investigate the detection performance of the conventional DE-based detection method [23], conventional attack patterns, including both deceptive and destabilizing FDIAs, are initiated in the DC microgrid system, shown in Fig. 3.8. The corresponding attack points among the communication links and local sensors for all the nodes are displayed in red. At around $t = 15$ s, conventional stealthy attacks penetrate nodes 1 and 2, targeting the local sensors and outgoing communication links coordinately. As presented in (3.9), the control objectives are still satisfied, and the system operates similarly to a load step as all the current converges to around 2.5 A. At around $t = 55$ s, destabilization attacks targeting node sensors are performed on nodes 1 and 2, resulting in a deviation from the control objectives, as in (3.10). The destabilizing phenomenon becomes evident as the voltage drops by 0.15 V/s to around 57 V at $t = 75$ s when the attack is removed. The conventional DE-based detection method could properly detect both types of attacks, as shown in Fig. 3.8.

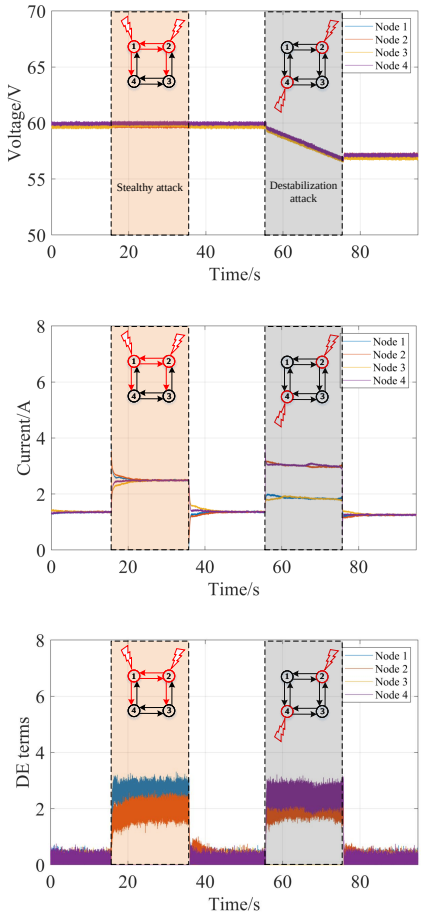


Figure 3.8. Performance of the DE-based detection method under conventional attack patterns. Source: **[J4].**

3.4.2 MARL intelligent attack generation

To bypass the DE-based detection method, the RL-based attacker generates novel attack patterns by targeting multiple nodes, as shown in Fig. 3.9. With the initial loading condition, around 5 A is proportionally shared among the four nodes, as presented in Fig. 3.9. At around $t = 16$ s, the RL-based intelligent attackers inject false signals into the local sensors of three neighboring nodes 1, 2, and 4, lasting around 30 seconds. As a result of such an attack pattern, the current sharing for the compromised and non-compromised nodes is deceptively regulated at around 1.9 A. A destabilizing phenomenon in the system caused by the attacks is noticeable in

the DC bus voltage, which drops by 0.04 V/s. It can also be observed that the DE metrics for all the nodes are curtailed to a low level, similar to or below the threshold under normal conditions, indicating the failure of the conventional DE-based detection method. The performance of the RL-based attacks is still maintained under load step-up and reversal applied at node 1.

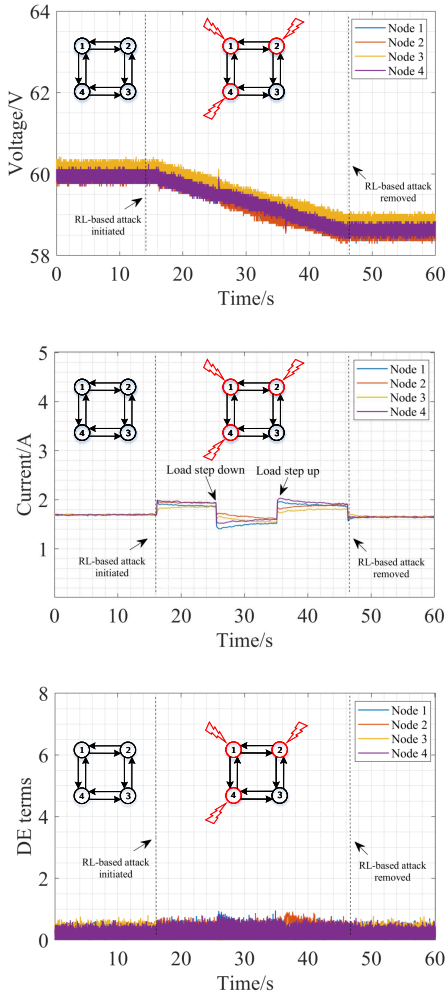


Figure 3.9. Performance of the MARL on the generation of stealthy destabilizing FDIAs.

3.4.3 Supplementary data-driven attack detector

A more intuitive illustration of the proposed data-driven detection method for the RL-based attacks is presented in Fig. 3.10. As shown in the diagram, the DE-based detection method can detect certain types of attacks (the grey area in the left diagram). The RL-based attacker explores the attacks and identifies the vulnerability of the DE-protected DC microgrid by generating attacks that bypass the DE-based detection method (the orange area). The data-driven detector is proposed to complement the DE-based detection method in detecting RL-based attacks (the grey area in the right diagram), addressing this issue. Therefore, the new detection mechanism, where the DE-based and proposed data-driven detection methods complement each other, can detect more attacks, including the RL-based attack.

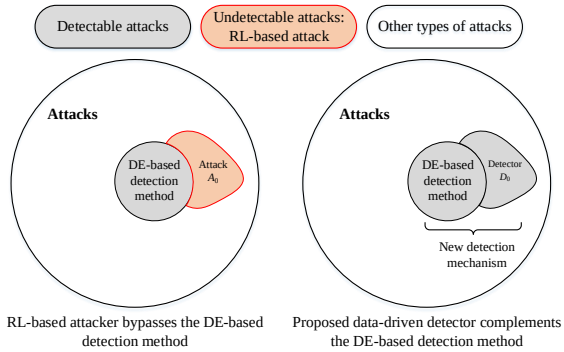


Figure 3.10. Unified picture of the RL-based attacker and data-driven detector for attack generation and detection.

Experimental results are shown in Fig. 3.11. The system operates with distributed control, where the voltage is regulated at a reference voltage of 60 V, and current sharing is achieved among the converters. When the RL-based FDIA is initiated at around $t = 11$ s, a 0.3 A current deviation occurs on the compromised node, and the current of the non-compromised node rises to around 1.75 A. The destabilizing phenomenon is evident as the output voltage ramps down by 0.05 V/s. Meanwhile, the detection metrics DE are maintained at a low level, indicating stealthy RL-based attacks to the DE-based detection method.

Meanwhile, as shown in Fig. 3.11(d), the labels of all the nodes are scaled with their index and a factor of 0.2. The well-trained PRN detector singles out the compromised nodes 1, 2, and 4, revealing attacks at nodes 1, 2, and 4. When the attack is removed at around $t = 36$ s, the DC microgrid operates normally at a new point, where the control objectives are satisfied, and the labels for all the nodes are maintained at 0.

In this way, the data-driven detection method can supplement the DE-based detection method in detecting novel attack patterns generated by intelligent attackers.

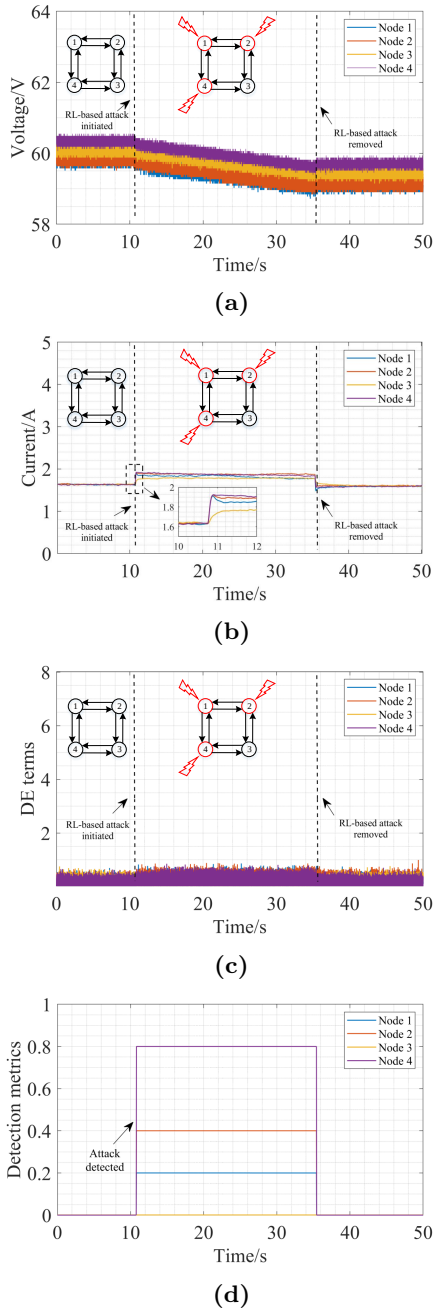


Figure 3.11. Experimental validation of the proposed data-driven detection method for the DC microgrid. (a) Output voltage, (b) Output current, (c) Performance of conventional DE-based detection method, and (d) Performance of proposed data-driven detection method. Source: [**J3**]

Based on the above analysis, it is clear that the proposed data-driven detection method aims to complement the DE-based detection method, not to act as an alternative. Therefore, when these two methods complement each other, the presented detection mechanism will always be more effective than a solely DE-based detector.

Moreover, to detect other types of attacks, we can design the RL-based attacker and the proposed data-driven detector to compete with each other in an iterative way to detect other types of attacks. Eventually, the new detection mechanism can detect more attacks than solely DE-based detection schemes.

3.5 Insights: iterative design for novel attack patterns generation and detection

The RL-based attacker can also learn to avoid new types of detection mechanisms because we can design RL attackers iteratively. The flowchart in Fig. 3.12 below shows the iterative design of the RL-based attacker and the training data-driven detector, which is also demonstrated with a unified illustrative diagram in Fig. 3.13.

3.5.1 Initialization

As presented in Fig. 3.10, with the system model protected by the DE-based detection method, the RL-based attacker generates a sophisticated attack A_0 that avoids the DE-based detector. The operation data under the attack A_0 and normal operation are used to train the NN-based detector D_0. Then, the proposed data-driven detector D_0 is employed to complement the DE-based detector in detecting the RL-based attacks A_0.

To enable the RL-based attacker to explore autonomously and randomly in all other attacks, a generic reward function r for training the RL attacker can be designed based on the reward function for the DE-based detection method, expressed below.

$$r = k_{_{DE}} \cdot r_{_{DE}} - k_{_{NN}} \cdot r_{_{NN}} \tag{3.18}$$

where $r_{_{DE}}$ denotes the reward function designed for DE terms (equation (3.13)), which is negative, $k_{_{DE}}$ is the corresponding coefficient, $r_{_{NN}}$ denotes the successful detection of the attacks with the previous data-driven detector D_{i-1}, $k_{_{NN}}$ is the coefficient to penalize detection of the RL-based attacks in training stage.

3.5.2 Iterative training process

For each iteration i, the following two steps are implemented:

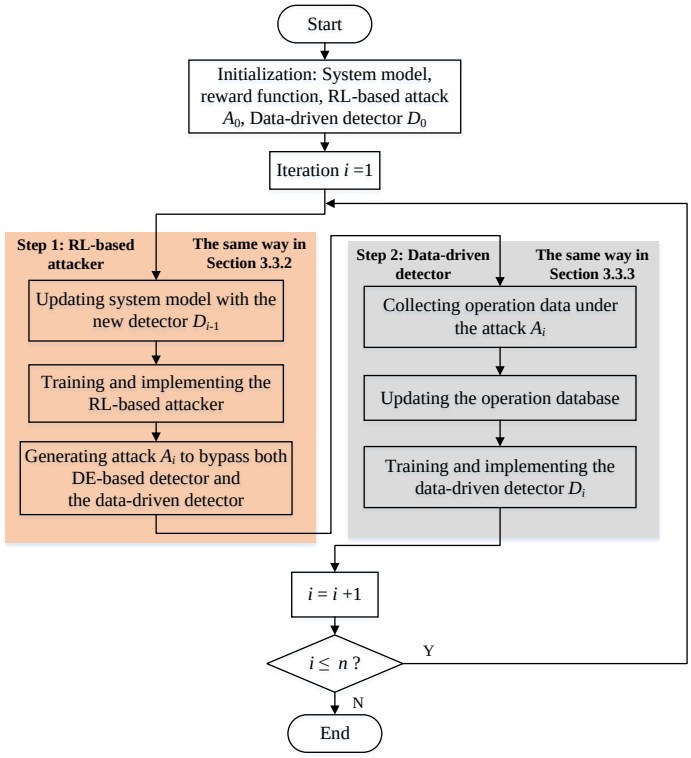


Figure 3.12. The flowchart of the iterative design of the RL-based attacker against the new detection mechanism.

Step 1: RL-based attacker.
As presented in Section 3.3.2, by updating the system model with the up-to-date data-driven detector, the RL-based attacker will randomly explore all other attack types and generate attack A_i to bypass both the DE-based detector and the data-driven detector D_{i-1}, corresponding to the orange region in the left diagram of Fig. 3.13.
Step 2: Data-driven detector.
In the same way presented in Section 3.3.3, by collecting the data under the attack A_i, the database of operation data is updated. It is then used to train the new data-driven detector D_i, which can detect both attacks A_{i-1} and A_i, corresponding to the additional grey region in the right diagram of Fig. 3.13. Then, repeat steps 1 and 2 until the stopping criteria are satisfied. The new detection mechanism, where the DE-based detection method and data-driven detector complement each other, is obtained.

As illustrated in Fig. 3.13, more attacks, denoted with grey color, can eventually be detected.

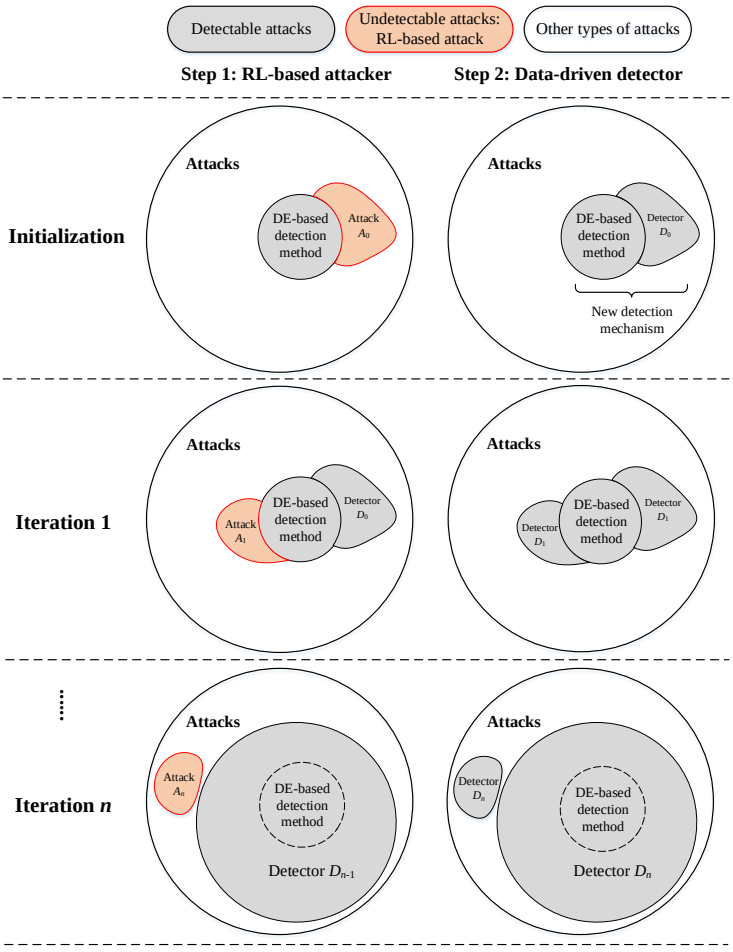


Figure 3.13. Unified illustration of the iterative process of RL-based attacker against the data-driven detector.

3.5.3 Discussion on practical implementation

For larger systems, where the RL attacker becomes more successful due to its capability of dealing with complex environments, the MARL attacker could still generate novel attack patterns, making it scalable in larger systems with more nodes. In addition, the attacker agents could be trained centrally while being implemented in a distributed manner. Therefore, MARL-based attackers can be created and trained in a larger system to generate a broader range of cyber-attacks. More importantly, as

the data-driven detector is trained with historical operation data, it can supplement conventional metric-based detection methods to detect more attack patterns.

For the computational requirements for implementation, as mentioned, due to the distributed control structure in DC microgrids, the distributed MARL-based attack is scalable in larger systems with more nodes, and the proposed data-driven detection method could still be applied with the data collected during system operation. Given this, the computational requirement will not increase as the attack and NN-based detectors are implemented independently in each microprocessor of the nodes in larger systems. The overall computational requirement is the same for each microprocessor. Therefore, the MARL-based attack and proposed data-driven detection method could be implemented in larger systems without increasing computational requirements, although more microprocessors would be needed.

3.6 Summary

This chapter first presents a MARL-based attacker to uncover vulnerabilities in the existing cyber-attack detection methods. The performance of the intelligent attackers is verified by locating the penetrable spots in a sampled DE-based detection method. With such an intelligent attack mechanism, the RL-based attackers generate novel stealthy destabilizing FDIAs in a DC microgrid protected by the DE-based detection method. To address this, a data-driven cyber-attack detector is proposed to supplement the DE-based detection method. The operation data from healthy and attacked DC microgrids are collected to train the PRN cyber-attack detectors. Finally, both the MARL-based attackers and data-driven detectors are implemented in an experimental setup to verify their performances.

It is observed that despite the effectiveness of the discordant scheme in detecting conventional deceptive and destabilizing FDIAs, the DE-based detection method fails to identify more coordinated FDIAs generated by the MARL-based attackers. Moreover, the data-driven detector can single out the attacked nodes, showing the effectiveness of complementing the conventional DE-based cyber-attack detection approach. This way, the proposed data-driven cybersecurity framework could generate and detect more attack patterns.

Based on publications

- [**J3**] **Y. Wan** and T. Dragicevic, “Data-driven Cyber-attack Detection of Intelligent Attacks in Islanded DC Microgrids,” in IEEE Transactions on Industrial Electronics, vol. 70, no. 4, pp. 4293-4299, April 2023.
- [**J4**] A. J. Abianeh, **Y. Wan**, F. Ferdowsi, N. Mijatovic and T. Dragičević, “Vulnerability Identification and Remediation of FDI Attacks in Islanded DC Microgrids Using Multiagent Reinforcement Learning,” in IEEE Transactions on Power Electronics, vol. 37, no. 6, pp. 6359-6370, June 2022.

CHAPTER 4

Battery health-aware DRL-based energy management for DC microgrids

4.1 Introduction

As mentioned, energy management for DC microgrids functions above the primary and secondary control layers for advanced control objectives, such as optimal power flow control and economical operation. Due to the intermittent nature of RESs and loads, an efficient energy management strategy for controlling and coordinating DGs is necessary. As an indispensable element of DC microgrids, BESSs have high power and energy density, flexible location selection, and fast response speed. In addition, it can charge and discharge power from the grid, functioning as a prosumer for both power consumption and supply, making it suitable for addressing the aforementioned issues [75, 76]. Moreover, it can provide significant operational benefits, such as peak shaving for the load [40] and energy arbitrage for economic operations [31].

To achieve these objectives, conventional techniques such as linear programming, mixed-integer linear programming, and dynamic programming have been proposed [77, 78]. However, these methods suffer from the curse of dimensionality and are susceptible to becoming trapped in local optima due to system uncertainties regarding load demand and electricity prices, limiting their flexibility and scalability [79]. Metaheuristic algorithms and their hybrids have been proposed to address the aforementioned issues [31, 80, 81]. However, these methods cannot adapt to varying operating conditions and need to optimize the scheduling from scratch whenever the operating conditions change over time. More importantly, all the aforementioned methods depend on the precise formulation of the system model. However, with the continuous integration of more DGs, it is challenging to build detailed mathematical models for all the components inside DC microgrids, posing challenges to the current energy management framework.

Recently, with the development of AI, learning-based energy management strategies have advanced, mitigating the need for a precise and explicit mathematical system

model and adapting to varying conditions due to their model-free capability. Different NN-based energy management frameworks have been proposed, leveraging the nonlinear mapping capability to realize optimal system control based on historical operation data [81–83]. However, NN-based scheduling methods lack exploration capabilities for dynamically changing environmental conditions and are usually trained based on historical operation data. Therefore, DRL methods have gained increasing attention due to their self-learning and model-free characteristics. Deep Q-learning, DQN, DDPG, and PPO are deployed to achieve autonomous operation of DC microgrids [39, 84].

Further, BESSs have a limited operational lifetime, where repeated charging and discharging operations result in battery degradation, influenced by various stress factors, including DoD, average SoC, charging/discharging current, and ambient temperature [85]. Researchers have endeavored in different ways to formulate the battery degradation cost for the conventional model-based energy management framework:

- Battery operational factors. The simplest way is to introduce battery operational terms into the objective function or constraints, such as cycle number [86], SoC [87], DoD [88], C-rate [89], and power term [90]. The positive side of such methods is that they simplify the scheduling problem, which can be solved using mathematical programming methods such as linear or quadratic programming. Moreover, actual battery degradation cannot be explicitly evaluated with these operational terms, where assigned weighting factors need to be properly selected [91]. Therefore, the method would result in suboptimal operations and higher operation costs in the long run.

- Stress factor-based battery degradation model. Various battery degradation models for the energy management framework have been proposed. A combined factor-based battery degradation model, validated only with early degradation data, is employed to evaluate the battery capacity losses in the energy management structure for DC microgrids [92]. Rainflow counting algorithm (RCA) is also employed to evaluate capacity losses during charging/discharging operations [93]. However, RCA lacks the analytical equation, making it difficult to implement explicitly in the optimization-based energy management framework [40, 85].

- ML-based degradation models. The data-driven model can accurately capture battery degradation under varying conditions by training a neural network model based on sufficient experimental data [94]. However, they either model for the evaluation of the state of health (SoH) without considering stress factors [95, 96] or increase computational burden due to the introduction of highly nonlinear and nonconvex NN-based degradation models [97].

However, all the aforementioned battery degradation formulation methods are constructed on a single-stage battery degradation model that only considers one degradation mechanism over the course of the battery's lifetime. This is typically unrealistic

because the aging mechanism under diverse stressful conditions would exhibit multiple patterns over the lifespan. Not acknowledging this in the scheduling strategy would result in suboptimal operations. Moreover, most studies regarding DRL-based methods only consider power exchange and generation costs while ignoring or simplifying battery usage costs during the agent training process.

In this chapter, solutions to the challenges of the current energy management framework are discussed in two aspects:

- A novel multi-stage battery degradation modeling method to accurately capture varying aging patterns across its lifespan is proposed, which is adaptively integrated into the energy management framework for scheduling the system.
- A battery health-aware DRL-based energy management framework for DC microgrids to optimize system operations based on the battery degradation model.

This chapter includes the content and results of **Paper J5, C2**.

4.2 Modeling of battery degradation

The battery degradation process should be modeled appropriately to evaluate its usage cost for the energy management framework. For the short-term operational planning problem, calendar aging is independent of operations and contributes negligibly to battery capacity losses. Therefore, battery cycle degradation is the main focus of the planning framework. Furthermore, to demonstrate the proposed multi-stage degradation modeling method, a generic cycle life model for lithium-ion batteries from [98] is adopted as an example. In addition, open-source experimental battery cycle aging test data are employed for model parameter identification [99], covering battery cycle aging data under diverse stressing conditions.

4.2.1 Single-stage battery degradation model

Based on the availability of experimental data under various cycling conditions, DoD, current rate, and temperature are considered for the battery degradation model. The temperature stress model is built based on the Arrhenius equation [100]. Moreover, charging and discharging currents are assumed to contribute equally to battery aging. According to Wöhler's approximation and Miner's rule, the contribution of each cycle to battery capacity loss is evaluated by subjecting the amplitude of stress factors to the maximum yield strength of the material [101]. In this way, the battery degradation model for each stress factor can be formulated, and the combined stress factor model θ is expressed as

$$\theta(t) = \theta_{DoD}(t) \cdot \theta_{I_{ch/dis}}(t) \cdot \theta_T(t), \tag{4.1}$$

$$\theta_{DoD}(t) = \left(\frac{DoD(t)}{DoD_{ref}}\right)^{\frac{1}{\alpha}}, \tag{4.2}$$

$$\theta_{I_{ch/dis}}(t) = (\frac{I_{ch/dis}(t)}{I_{ref}})^{\frac{1}{\beta}}, \tag{4.3}$$

$$\theta_T(t) = exp\left[-\psi\left(\frac{1}{T_a(t)} - \frac{1}{,}T_{ref}\right)\right] \tag{4.4}$$

where α and β are the stress exponents for DoD and current rate, respectively. ψ is the Arrhenius rate constant, DoD_{ref} and I_{ref} denote the reference DoD and reference current, and T_a and T_{ref} represent the ambient and reference temperatures.

The maximum number of cycles to end-of-life (EoL) is calculated as

$$N_c(t) = \frac{N_{cref}}{\theta(t)} = N_{cref} \cdot \left(\frac{DoD(t)}{DoD_{ref}}\right)^{-\frac{1}{\alpha}} \cdot \left(\frac{I_{ch/dis}(t)}{I_{ref}}\right)^{-\frac{1}{\beta}} \cdot exp\left[-\psi\left(\frac{1}{T_{ref}} - \frac{1}{T_a(t)}\right)\right], \tag{4.5}$$

where N_{cref} denotes the number of cycles to EoL for the reference cycle condition, and N_c represents the number of cycles to EoL for corresponding stressing conditions.

The battery cycle degradation process, considering the non-linear aging characteristics, is formulated based on beginning-of-life (BoL) and EoL as

$$Q(t) = Q_{BoL} - \varphi(t)^{\xi} \cdot (Q_{BoL} - Q_{EoL}), \tag{4.6}$$

where Q represents the battery cycle life, and Q_{BoL} and Q_{EoL} denote the battery capacity at BoL and EoL, respectively. Q_{EoL} is assumed to be 80% of Q_{BoL}. φ represents the resultant capacity losses for the corresponding cycles, and ξ is the aging exponent, calculated as

$$\xi = \frac{ln(\frac{Q_{BoL}-Q_5}{Q_{BoL}-Q_{EoL}})}{ln(\frac{N_{ref}}{N_{cref}})}, \tag{4.7}$$

where Q_5 denotes 5% capacity loss of Q_{BoL} at the reference cycling condition with the number of cycles N_{ref}.

The relevant parameters are calculated as follows.

$$N_{ci} = \frac{N_{cref} \cdot N_i}{N_{ref}}, \tag{4.8}$$

where N_{ci} and N_i represent the number of cycles to EoL and Q_5 for testing condition $i \in \{2, ..., n\}$. The exponent for each stress model is calculated as

$$\alpha = -\frac{ln(\frac{DoD_i}{DoD_{ref}})}{ln(\frac{N_{ci}}{N_{cref}})}, \tag{4.9}$$

$$\beta = -\frac{ln(\frac{I_i}{I_{ref}})}{ln(\frac{N_{ci}}{N_{cref}})}, \tag{4.10}$$

$$\psi = \frac{ln(\frac{N_{ci}}{N_{cj}})}{\frac{1}{T_i} - \frac{1}{T_{ref}}}. \tag{4.11}$$

The testing conditions for model parameterization are presented in Table 4.1, and the identified model parameters are presented in Table 4.2. By substituting these parameters into the equations from (4.1) to (4.4), the single-stage battery degradation model is constructed.

Table 4.1. Cycling conditions for parameter identification. Source: **[J5]**

Conditions	1*	2	3	4
DoD	100%	60%	100%	100%
C-rate	0.5C	0.5C	2C	2C
Temperature	25℃	25℃	25℃	35℃

* Reference condition

Table 4.2. Parameter identification for single-stage battery degradation model. Source: **[J5]**

Inputs for parameters identification				
N_1	N_2	N_3	N_4	N_{c1}
22	256	18	23	513
Model parameters identification				
N_{cref}	α	β	ψ	ξ
513	0.2081	6.9083	-21.4482	0.4402

4.2.2 Multi-stage battery degradation model

The multi-stage battery degradation model is constructed using aging test data throughout the battery's lifetime under specific cycling conditions. As depicted in Fig. 4.1, the battery aging rate changes adaptively at different battery states, where the varying degradation patterns under diverse stressing conditions within different SoH ranges across the whole lifetime could be captured by the proposed multi-stage battery degradation modeling method. In this chapter, the battery aging process from BoL to EoL is evenly divided into three stages, corresponding to the SoH range as stage 1: 100%-93.3%, stage 2: 93.3%-86.6%, and stage 3: 86.6%-80% SoH.

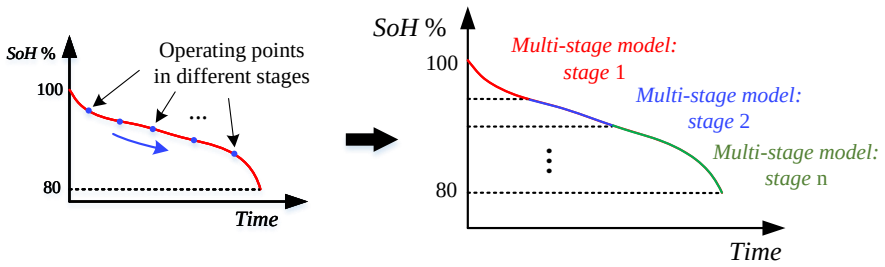


Figure 4.1. Battery operating stages and the multi-stage degradation model. Source: **[J5]**.

In particular, for each battery operational stage, the degradation model is parameterized in the same way presented in Section 4.2.1, where the experimental degradation data for different SoH ranges will be employed. More separation of the stages for the battery aging process would improve the model accuracy and lead to more economical operations, which will be presented in the following sections. The model parameters for each stage are presented in Tables 4.3-4.5.

Table 4.3. Multi-stage battery degradation model: Stage 1. Source: **[J5]**.

Inputs for parameters identification				
N_1	N_2	N_3	N_4	N_{c1}
36	378	34	39	513
Model parameters identification				
N_{cref}	α	β	ψ	ξ
513	0.2172	24.2535	-12.0051	0.3952

Table 4.4. Multi-stage battery degradation model: Stage 2. Source: **[J5]**.

Inputs for parameters identification				
N_1	N_2	N_3	N_4	N_{c1}
231	1586	201	280	513
Model parameters identification				
N_{cref}	α	β	ψ	ξ
513	0.2652	9.9653	-29.0049	0.4470

Table 4.5. Multi-stage battery degradation model: Stage 3. Source: **[J5].**

Inputs for parameters identification				
N_1	N_2	N_3	N_4	N_{c1}
513	3628	562	727	513
Model parameters identification				
N_{cref}	α	β	ψ	ξ
513	0.2611	-15.1963	-22.5247	0.5066

In this way, different single-stage battery degradation models could be extended to multi-stage versions with the proposed multi-stage degradation modeling method. Additional aging test data independent of those used for parameterization are employed to verify the accuracy of the proposed multi-stage model. Two experimental datasets with cycling conditions of 60% DoD, 3C current rate, and 25 ℃, and 100% DoD, 1C current rate, and 15 ℃ are adopted, of which the experimental data at different stages of SoH are used for model validation. The root-mean-square error (RMSE) of the two degradation modeling methods regarding the experimental test data at different stages is presented in Table 4.6. Despite the close model accuracy at the initial aging stage, the multi-stage battery degradation model outperforms the single-stage degradation model across the whole battery lifespan.

Table 4.6. Model performance under different tests at different stages. Source: **[J5].**

Tests	Stages	Single-stage model	Multi-stage model
Test 5	Stage 1	0.78%	0.50%
	Stage 2	2.21%	0.42%
	Stage 3	3.15%	1.10%
Test 6	Stage 1	1.00%	0.78%
	Stage 2	2.37%	0.75%
	Stage 3	2.86%	1.28%

The combined stress model is parameterized with experimental data under specific cycling conditions, which is expected to yield reasonable estimation accuracy of battery under diverse stressing conditions by providing general trends of battery degradation [98].

In summary, the single-stage degradation model only accounts for an exclusive degradation pattern, which varies with degradation stages, resulting in inaccurate degradation estimation. In contrast, the multi-stage degradation modeling method constructs the aging model employing experimental data across the lifespan for dif-

ferent operational stages. In this sense, the multi-stage degradation model could be considered a piecewise linearization of the real battery degradation process, which can capture multiple degrading patterns at different stages, thus improving the accuracy of capacity loss estimation. Subsequently, the models at different stages could be adaptively integrated into the energy management framework to minimize the operation cost, which will be presented in the following sections.

4.3 Energy management structure

In this project, the energy management system determines the day-ahead optimal dispatch strategy for efficient operation of the DC microgrids. The diagram of the system is illustrated in Fig. 4.2, where a typical DC microgrid system consists of a PV system, a battery energy storage system, and a fast charging station (FCS) for EVs. The EV demand profile is emulated based on charging events over the course of the day and the charging profile of each EV, which is described with a normal distribution function [102]. The spot market price and PV generation profile could also be predicted based on historical prices, which is not the main focus of the thesis. In addition, battery usage is accounted for in the scheduling model based on the proposed battery degradation modeling method, which assigns a cost coefficient for converting it into degradation cost. Due to the improvement of heating, ventilation, and air conditioning systems for the stationary battery system [103], the temperature can be independently controlled regardless of the operational planning. Therefore, the cell temperature is maintained at the reference temperature of 25 °C with the battery thermal management system. The scheduling model is formulated below.

4.3.1 Objective function

The objective of the energy management framework for the DC microgrids is to minimize the overall operation cost, consisting of electricity purchase and battery degradation costs, which is expressed as

$$\min_{P_t^{ch}, P_t^{dis}} \sum_{t=1}^{T} (P_t^{in} \cdot p_t - P_t^{out} \cdot p_t) + \lambda_k \cdot C_t, \tag{4.12}$$

where P_t^{in} and P_t^{out} denote the power flow between the grid and DC microgrids, P_t^{ch} and P_t^{dis} are the operational commands for the BESS, p_t is the time-of-use (ToU) electricity price, C_t represents the battery capacity losses, assigned with a constant battery replacement cost coefficient λ_k.

In particular, the dispatch strategies P_t^{ch} and P_t^{dis} at each discrete time step t will be optimized within a finite time horizon (i.e., $t \in \mathcal{T} = \{1, 2, ..., T\}$). The first term in the objective equation (4.12) is the electricity purchase cost, which could be minimized via energy arbitrage, exploiting temporal price differentials on the spot

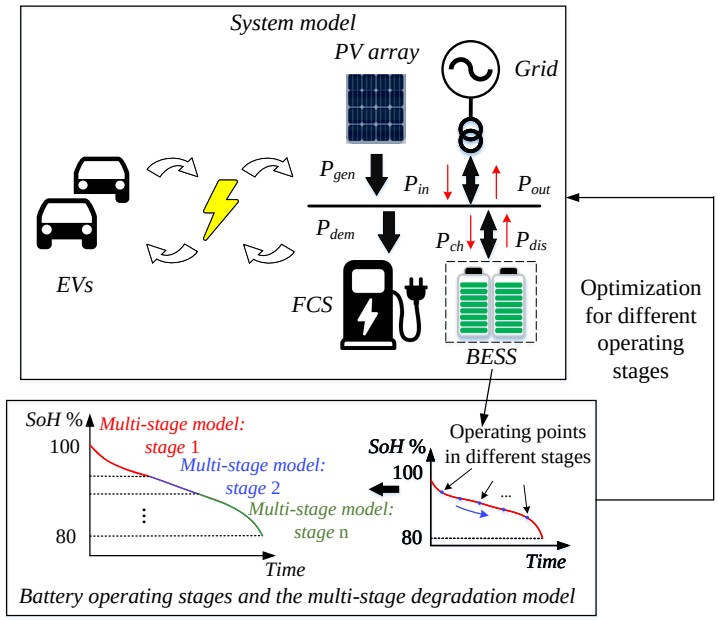


Figure 4.2. Energy management framework of DC microgrids with BESS. Source: **[J5].**

market. The second term is associated with battery usage cost, evaluated based on the battery degradation model during repeated charging and discharging operations.

4.3.2 Operation constraints

4.3.2.1 Energy balance

The energy balance of the DC microgrid system should be maintained among different modules, where the BESS acts as a buffer between the supplier and demand side. Therefore, the equality constraint is formulated as

$$P_t^{in} - P_t^{out} + P_t^{pv} = P_t^{ch} - P_t^{dis} + P_{dem}^t, \tag{4.13}$$

where P_t^{pv} is the power generation of the PV system under the maximum power point tracking (MPPT) control, P_{dem} is the load demand for the vehicles from the FCS.

4.3.2.2 Battery operation

The SoC evolution of the battery is formulated in (4.14). In addition, constraint (4.15) is introduced to ensure an operational SoC for each day with an expected

value SoC_{end}. Moreover, the SoC has a safe battery operating range. Therefore, the battery operation constraints are represented as

$$SoC_t = SoC_{t-1} + \frac{\tau}{E_{bat}}(\eta_{ch} \cdot P_t^{ch} - \frac{P_t^{dis}}{\eta_{dis}}), \tag{4.14}$$

$$SoC_{t=T} = SoC_{end}, \tag{4.15}$$

$$SoC_{min} \leq SoC_t \leq SoC_{max}, \tag{4.16}$$

where SoC_t denotes the SoC of battery at time t, τ denotes the time step, E_{bat} is the capacity of battery, η_{ch} and η_{dis} are the charging and discharging efficiency respectively, SoC_{min} and SoC_{max} are the SoC range for battery operation.

4.3.2.3 Power limits

The battery power ratings limit the maximum charging and discharging power as follows:

$$0 \leq P_t^{ch}, P_t^{dis} \leq P_{max}, \tag{4.17}$$

where P_{max} denotes the battery's maximum charging and discharging power.

In addition, the transmission power between the grid and DC microgrids should be restricted for both systems' stability and power quality based on the maximum power of the inverter P_{max}^{AC}. Therefore, the power limit for the power flow between the grid and the DC microgrid is expressed as

$$-P_{max}^{AC} \leq P_t^{in} - P_t^{out} \leq P_{max}^{AC}, \tag{4.18}$$

4.3.3 Cost for battery degradation

To explicitly consider battery degradation in the energy management framework, the battery degradation model should be discretized for real-time implementation. As the aging index represents the resultant capacity losses for certain cycles over a finite time horizon, the battery degradation could be formulated as the cumulative aging index over a certain period regarding real-time DoD, current, and temperature, expressed as

$$\begin{aligned} Q_{loss} = \sum_{t=1}^{T} \frac{\theta_t}{N_{cref}} = & \sum_{t=1}^{T} \frac{1}{N_{cref}} \cdot (\frac{DoD_t}{DoD_{ref}})^{\frac{1}{\alpha}} \cdot (\frac{I_t^{ch/dis}}{I_{ref}})^{\frac{1}{\beta}} \cdot \\ & \cdot exp\left[-\psi\left(\frac{1}{T_a} - \frac{1}{T_{ref}}\right)\right], \end{aligned} \tag{4.19}$$

In addition, an incomplete half cycle or full cycle may happen in each time step and the whole future optimization window. Based on the approximation in [104] and the $DoD_{ref} = 100\%$, $I_{ref} = 0.5C$, $T_a = 25$℃, and $T_{ref} = 25$℃, substituting the C_t in (4.12) with the formulated term Q_{loss} obtains

$$C_t = \sum_{t=1}^{T} \frac{(\triangle DoD_t)^{\frac{1}{\alpha}} \cdot (2I_t^{ch/dis})^{\frac{1}{\beta}}}{N_{cref}}. \tag{4.20}$$

Thus, with the identified battery degradation model parameters, the energy management framework is constructed with both the single-stage and multi-stage degradation models, respectively [1].

Notably, the proposed multi-stage battery degradation modeling method is scalable to a variety of battery degradation models, including empirical and semi-empirical models. The proposed optimization framework, along with the multi-stage modeling method, can similarly be employed to enhance economic scheduling strategies throughout the entire battery lifetime.

The battery usage cost under various stressing conditions can be obtained with an accurate estimation of battery degradation. The autonomous energy management framework for the DC microgrid, based on DRL, can be developed to optimally schedule the battery, reducing the overall operation cost, including the battery degradation cost.

4.4 DRL for energy management

As presented in the previous sections, the energy management framework for economic operations is an optimization-based decision-making process, which can be formulated as an MDP. The primary focus of the energy management framework is to coordinate the operation of the units of microgrids to leverage BESS to minimize the overall operation cost. Various RL algorithms with a continuous action space are deployed, including DDPG, PPO, and Twin Delayed DDPG (TD3).

4.4.1 State space

The information provided by the DC microgrid, as shown in Fig. 4.2, is essential for the DRL agent to make optimal decisions at each time step. Particularly, the system states that impact the control objectives, namely minimizing the operation costs, are electricity prices p_t, instant load demand P_t^{dem}, and battery state SoC_t. Consequently, the environmental information at each time step t forms the state space, expressed as

$$s_t = [t, P_t^{pv}, P_t^{dem}, p_t, SoC_t], \tag{4.21}$$

The state space is enumerated over the optimization horizon in T time steps for day-ahead optimization.

[1] We use a linear approximation for calculating the DoD term as in [C2]. Incorporating the RCA for cycle depth $\triangle DoD$ in an optimization problem [85, 105] is still an open research question. Our approach is similar to other works employing a DoD linearization method [106–108].

4.4.2 Safe Action Control

The actions taken by the agent for the energy management framework are the dispatch strategies for the battery, namely, P_t^{ch} and P_t^{dis}. A battery charging ratio coefficient α, bounded within [-1, 1], is employed for simplification. The positive and negative values of α denote charging and discharging the battery, respectively. Additionally, the agent explores randomly without guidance through trial and error, which may result in overcharging or overdischarging the batteries. Therefore, a safety action control is introduced to enforce battery operating constraints and guide action selection during optimization, ensuring practicality for online autonomous energy management.

Specifically, the PV generation first meets the load demand, with any surplus power used to charge the BESS, and any remaining power exported to the grid. In scenarios where PV generation cannot meet the load demand, the BESS and the grid supply the load, balancing the electricity purchase price and the battery usage cost.

For the time step t with battery state SoC_t, the upward and downward available regions for the battery operation are expressed as

$$SoC_t^{up} = SoC_{max} - SoC_t, \tag{4.22}$$

$$SoC_t^{dn} = SoC_t - SoC_{min}, \tag{4.23}$$

Consequently, based on (4.14), the maximum battery charging/discharging power is derived. Thus, the maximum possible battery charging ratio is derived as

$$\alpha_{max} = \begin{cases} \dfrac{SoC_t^{up} \cdot E_{bat}}{P_{max} \cdot \tau \cdot \eta_{ch}}, & if\ \alpha_t \geq 0 \\ \dfrac{SoC_t^{dn} \cdot E_{bat}}{P_{max} \cdot \tau / \eta_{ch}}, & if\ \alpha_t < 0 \end{cases} \tag{4.24}$$

To limit the charging ratio coefficient within the derived feasible interval, a clip function is used, returning

$$\alpha_t = min(max(\alpha, \alpha_{min}), \alpha_{max}), \tag{4.25}$$

where α_{min} and α_{max}, the inputs to the clip function, represent the feasible bounds for α_t. Consequently, the battery operating power is derived as

$$P_t^{Bat} = P_{max} \cdot \alpha_t, \tag{4.26}$$

Actions taken during the online training process are constrained to ensure the safety of BESSs. However, clipping the action space does not adequately guide the RL agent in taking feasible actions. Consequently, to guide the learning of RL agents, a soft constraint is introduced by adding a penalty term when the battery SoC is calculated outside the operational range, as elaborated below.

4.4.3 Reward design

The reward function design aligns with the control objectives of minimizing operation costs, which include electricity purchase and battery degradation costs. Battery degradation cost is formulated based on a real-time evaluation of battery degradation, considering various stress factors, as presented in (4.20). Consequently, the reward representing the cost of energy exchange with the grid and battery usage is formulated as

$$r_1 = -(P_t^{grid} \cdot p_t + \lambda_k \cdot C_t), \tag{4.27}$$

where $P_t^{grid} = P_t^{dem} - P_t^{pv} - P_t^{Bat}$.

Notably, the immediate and battery-related reward enables the RL agent to make decisions to minimize the overall operation cost. This approach contrasts with conventional methods, which often lead to suboptimal scheduling due to oversimplified or absent formulations of battery degradation in the reward function.

Additionally, to guide the agents towards safe actions that prevent overcharging or overdischarging the battery, soft penalties are employed. These penalties increase as electricity prices rise.

$$r_2 = \begin{cases} -\left(|\alpha_t| - |\alpha_{max}|\right) \cdot E_{bat} \cdot p_t, & if\ |\alpha_t| \geq |\alpha_{max}| \\ 0, & else \end{cases}, \tag{4.28}$$

Consequently, the reward function for training the RL agents is formulated as

$$r_t = k_1 \cdot r_1 + k_2 \cdot r_2. \tag{4.29}$$

where k_1 and k_2 are the scaling factors used to balance the trade-off between cost minimization and the penalty for overcharging/overdischarging the battery, aiding in the convergence of the RL agents.

4.5 Results analysis

4.5.1 Scheduling with multi-stage battery degradation model

The performance of the proposed multi-stage battery degradation modeling method is validated by comparing it to conventional single-stage degradation counterparts across different battery operating stages. The adaptive implementation framework is depicted in Fig. 4.3. Specifically, various case studies representing distinct operating stages are conducted. In these studies, both single-stage and multi-stage battery degradation models are respectively integrated into the scheduling model. The optimization problem is implemented in Python using the Pyomo modeling interface [109], and is solved for the global optimum by the multistart solver based on IPOPT [110].

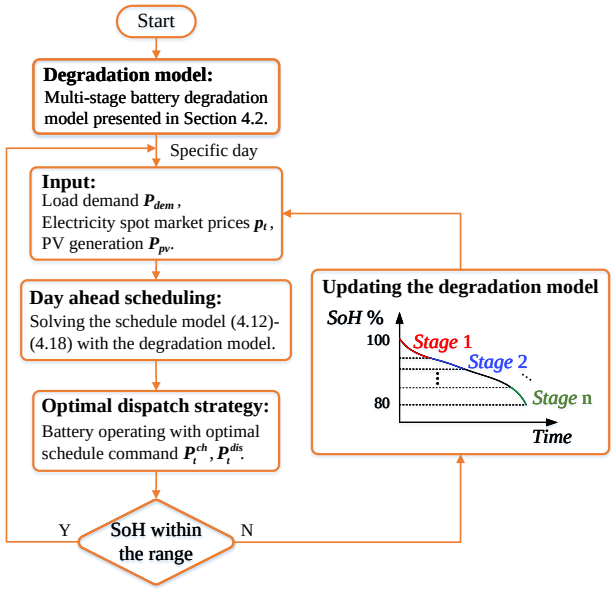


Figure 4.3. Flowchart of the proposed adaptive optimization framework with the multi-stage battery degradation model. Source: **[J5].**

The nominal capacity of the BESS in the case studies is $E_{bat} = 460$ kWh. The SoC range is (0.1, 0.9), with both the initial SoC and SoC_{end} being 0.1. The round-trip battery efficiency is $\eta_{ch} = \eta_{dis} = 0.92$, and the cost coefficient is $\lambda_k = 3560$ DKK/kWh. The power limits for the battery and grid are set as P_{max}= 460 kW and P_{max}^{AC}= 10 MW, respectively. For the day-ahead scheduling, the battery operation horizon is $\tau = 1$ h, thus $T = 24$, resulting in 48 decision variables. Owing to its superior model accuracy, the actual battery degradation cost in each case study is evaluated using the multi-stage degradation model.

Case 1: operating stage 1

Assuming the battery operates within stage 1, the battery degradation model is listed in Table 4.3. The load demand, PV generation profiles, and electricity prices of the selected example day are illustrated in Fig. 4.4(a)-(b). The scheduling results of the BESS are presented in Fig. 4.4(c), while the corresponding battery SoC profile is depicted in Fig. 4.4(d).

It can be observed that the battery charges during 1:00–3:00 and 14:00–15:00 when electricity prices are at their lowest, and discharges primarily at 8:00 and during 19:00–20:00 when electricity prices peak within the period. This indicates that the optimal dispatch strategies involve charging the battery during low-price periods and discharging it for energy supply during high-price periods. Additionally, surplus

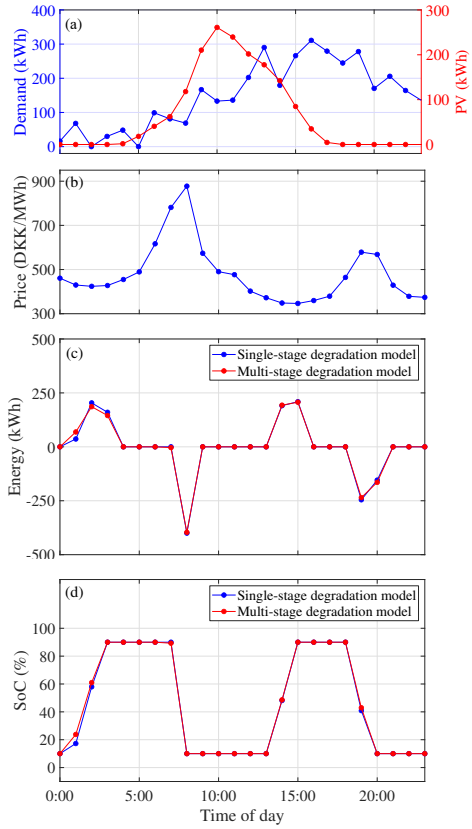


Figure 4.4. Optimal dispatch strategies over a 24-hour period in stage 1. From top to bottom, the panels show (a) Load profile, (b) Spot market prices and ambient temperature, (c) Battery operation profile, and (d) Battery SoC level. Source: **[J5].**

energy from the PV and BESS is sold back to the grid during high-price periods. Generally, the dispatch strategies shift the load demand away from high electricity prices and operate the battery with consideration of its degradation, thereby minimizing the overall operation cost.

Despite the slightly higher battery operation command in the single-stage model, scheduling using both degradation models results in similar battery SoC profiles, leading to slightly higher accumulated battery degradation over one day, as presented in Fig. 4.7(a). Furthermore, numerical results in Table 4.8 reveal that energy arbitrage revenue from the battery using the single-stage degradation model is higher than that of the multi-stage counterpart. Concurrently, the battery capacity loss is higher, resulting in increased overall operation costs. In summary, the results align closely with the model accuracy of the two degradation modeling methods in the first stage.

Case 2: operating stage 2

When the battery operates in stage 2, the parameters listed in Table 4.4 update the multi-stage battery degradation model. The load demand, PV generation, and electricity prices within the day are depicted in Fig. 4.5(a) and Fig. 4.5(b). The scheduling results of the battery and the resultant SoC profile, showing noticeable differences between the energy management outcomes of the two degradation models, are presented in Fig. 4.5(c)-(d). Similar to the stage 1 strategy, the battery is charged when prices bottom out and discharged when prices peak. However, Fig. 4.5(c) reveals that with the multi-stage degradation model, the operator uses the battery more efficiently, charging it at 16:00 and discharging at 19:00. This strategy not only reduces battery degradation but also shifts load consumption away from high-price periods. Additionally, during 0:00–13:00, the battery operates with lower charging/discharging commands, leading to reduced degradation at the cost of a slight decrease in energy arbitrage revenue.

The daily accumulated battery degradation, demonstrating greater capacity losses with the single-stage degradation model, is presented in Fig. 4.7(b). Moreover, the numerical results for case 2 in Table 4.8 indicate approximately a 2.5% reduction in overall operation costs using the multi-stage battery degradation model compared to the single-stage counterpart.

Case 3: operating stage 3

In stage 3, the model parameters listed in Table 4.5 are utilized to update the multi-stage battery degradation model within the scheduling model. The operation results, including demand, PV generation, and electricity market prices as depicted in Fig. 4.6(a)-(b), are illustrated in Fig. 4.6(c). Due to high electricity prices from 6:00 to 18:00, the operator initially charges the battery between 1:00 and 3:00 when prices are low, then discharges it at 7:00 to meet the increased load demand, which cannot be covered by PV generation alone. However, with the multi-stage model, the battery is discharged less, effectively balancing degradation costs and energy arbitrage revenue. This contrasts with the single-stage model, where more intense discharge leads to sharper increases in battery degradation. During periods of high electricity prices, the operator manages the battery to capitalize on the temporal differences in electricity prices, thereby generating revenue. Scheduling using the multi-stage degradation model results in decreased amplitudes of charging and discharging power compared to the single-stage model, consequently leading to reduced battery degradation, as illustrated in Fig. 4.7(c). Additionally, numerical results in Table 4.8 indicate a 1.8% reduction in overall operation costs with the multi-stage degradation model compared to the single-stage model, albeit with a lower energy arbitrage revenue.

Another multi-stage modeling method, as described in [95], is integrated into the adaptive energy management framework for a more comprehensive comparison with the proposed multi-stage degradation modeling method. The models and their parameters are detailed in Table 4.7.

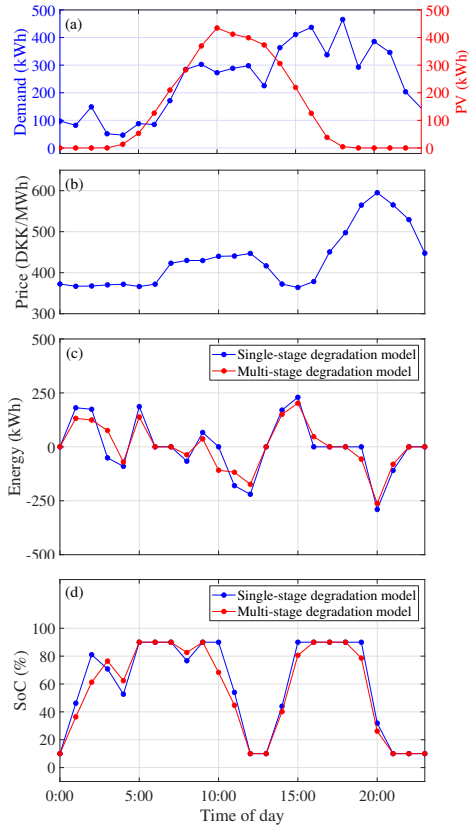


Figure 4.5. Optimal dispatch strategies over a 24-hour period in stage 2. From top to bottom, the panels show (a) Load profile, (b) Spot market prices and ambient temperature, (c) Battery operation profile, and (d) Battery SoC level. Source: **[J5].**

Table 4.7. Model parameterization

Model	Multi-stage models
$Q_{loss} = a * I^b + c$	Stage 1: a=2.36e-5, b=3.119, c=5.569e-5 Stage 2: a=6.954e-6, b=3.092, c=1.348e-5 Stage 3: a=3.778e-6, b=3.077, c=6.938e-6

The multi-stage degradation modeling method from [95] has not been explored for scheduling problems. Nevertheless, this method considers only a single current stress factor and lacks the flexibility to incorporate multiple stress factors. The scheduling results are presented in Table 4.8 for comparison. It is observed that the operator underestimates battery degradation, leading to more intense and frequent charging and discharging of the battery. This results in higher energy arbitrage revenue but

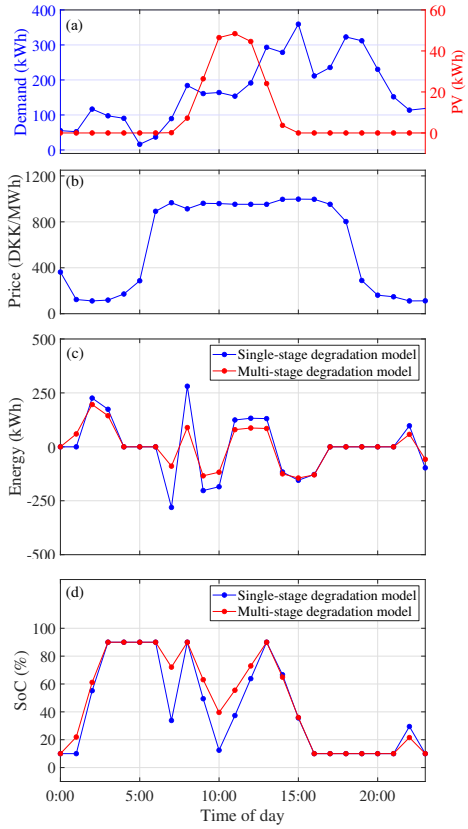


Figure 4.6. Optimal dispatch strategies over a 24-hour period in stage 3. From top to bottom, the panels show (a) Load profile, (b) Spot market prices and ambient temperature, (c) Battery operation profile, and (d) Battery SoC level.

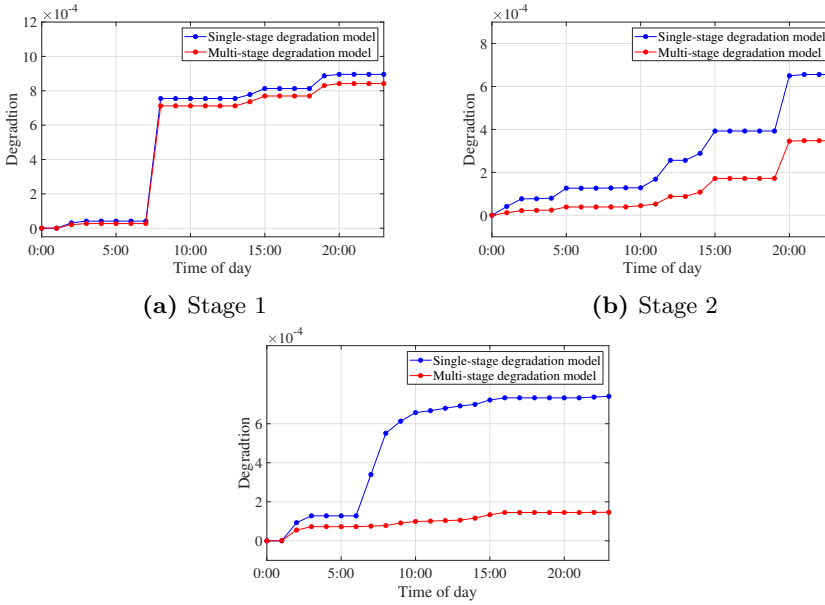
at the expense of increased overall operation costs.

Notably, the proposed multi-stage battery degradation modeling method is scalable to various conventional battery degradation models. This scalability allows operators to accurately capture the varying aging patterns across the battery's lifespan under different stressing conditions, enabling more precise estimation of battery degradation costs and the formulation of optimal dispatch strategies to minimize overall operation costs.

Another area of concern is the sensitivity analysis related to the number of stages into which the battery's operational lifetime is divided within the energy management framework. In principle, the more stages into which the battery lifetime is divided, the higher the accuracy of the multi-stage battery degradation model under varying stress conditions, leading to more effective energy management strategies. To investigate the

Table 4.8. Comparison of the optimal scheduling results with different degradation models.

Cases	Case 1			Case 2			Case 3		
	[95]	Single-stage model	Multi-stage model	[95]	Single-stage model	Multi-stage model	[95]	Single-stage model	Multi-stage model
Battery degradation (%)	0.20	0.090	0.084	0.56	0.066	0.035	0.49	0.074	0.015
Energy arbitrage revenue (DKK)	521.63	483.55	482.95	309.28	290.58	288.29	593.78	537.71	525.95
Operation cost (DKK)	477.81	419.59	415.59	1377.13	974.32	950.26	2419.32	2119.01	2080.02



(a) Stage 1 **(b)** Stage 2

(c) Stage 3

Figure 4.7. Comparison of the cumulative battery degradation for two models in each stage. Source: **[J5].**

impact of the number of stages on scheduling results, the battery lifetime is divided into various stages. In each stage, the corresponding multi-stage degradation models are parameterized similarly and adaptively incorporated into the energy management framework.

To showcase the scheduling results using multi-stage battery degradation models, the results of the single-stage model are used as a baseline, against which the reduced operation cost for each multi-stage model is calculated. For simplification, all the multi-stage models are implemented to obtain the dispatch strategies for the same day. The model's relative error is calculated following the method presented in Section 4.2.2. The results of the average reduced operation cost for the selected day regarding the number of divided stages are shown in Fig. 4.8. It is observed that model

accuracy improves marginally with an increasing number of segments, and the average reduced operation cost stabilizes at around 27 DKK. This indicates that there are no significant benefits to further dividing the stages.

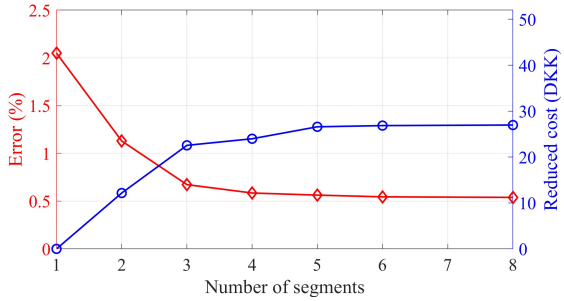


Figure 4.8. Model accuracy and average reduced operation cost using the proposed optimization framework based on multi-stage degradation model regarding different number of segments for battery's lifetime. (Blue line: Reduced cost; Red line: Model relative error.)

The results demonstrate that the proposed multi-stage battery degradation modeling method can effectively estimate battery degradation and be integrated into the energy management framework to optimize system operation and minimize operation costs. As mentioned, the model-free DRL-based energy management framework can learn to operate the system autonomously, where the reward function used to encourage the RL agent to achieve desired objectives should be appropriately formulated. Given that the battery has limited charging/discharging cycles and degrades during operations, it is crucial to factor the battery degradation cost into the energy management framework. This innovative multi-stage battery degradation modeling method allows for the formulation of operation costs while considering battery degradation. Consequently, the DRL-based energy management framework aims to minimize operation costs while accounting for battery degradation.

4.5.2 DRL-based energy management framework

4.5.2.1 Training process of DRL-based energy management framework

The training process utilizes historical data from 2022, specifically from January 1 to December 31, sourced from Energinet [111], as the training dataset. The testing data is obtained from various months in 2023. This approach enables RL agents to learn and optimize system operations using seasonal and weekly PV generation and demand data. The training process for various DRL algorithms is depicted in Fig. 4.9. Parameters k_1 and k_2 in the reward function (4.29) were set as 0.00125 and 0.025,

respectively. In addition, the initial battery capacity is randomly selected within the SoC range [0.1, 0.9].

It is observed that the reward rapidly increases after approximately 50 episodes of training, a result of the agents' random exploration stage. As training progresses, RL agents learn to take feasible and economical actions under various operating conditions, leading to fluctuations in rewards. Eventually, all agents converge to a similar performance level, with the PPO agent achieving the highest reward among the RL algorithms.

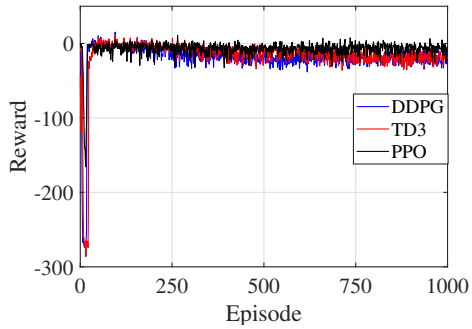


Figure 4.9. Training process of different DRL algorithms for energy management of DC microgrids.

4.5.2.2 Scheduling results

Various example days throughout the year are selected to test the performance of different RL agents in energy management for DC microgrids under varying and unwitnessed operating conditions. Additionally, given the demonstrated superior performance of the PPO algorithms in the training process, a version of the PPO algorithm trained without considering battery degradation was also implemented to assess the impact of battery degradation on DRL energy management. The baseline method, providing global optimal solutions, is the energy management framework presented in Section 4.3. As mentioned, the deterministic model-based day-ahead optimization framework relies on precise system model formulations and accurate predictions of stochastic variables (electricity prices, PV generation, and load demand), whereas DRL-based methods manage system operations with currently available information. Therefore, with comprehensive system knowledge, the operational costs of the baseline method across different days serve as a reference for evaluating the performance of DRL-based methods.

The resultant operational costs of the DRL methods exhibit similar performance to the optimal scheduling achieved by the model-based nonlinear programming method. Additionally, the table indicates that the PPO-based energy management framework

outperforms others in most case studies. Furthermore, omitting battery degradation in the DRL energy management framework leads to higher operational costs.

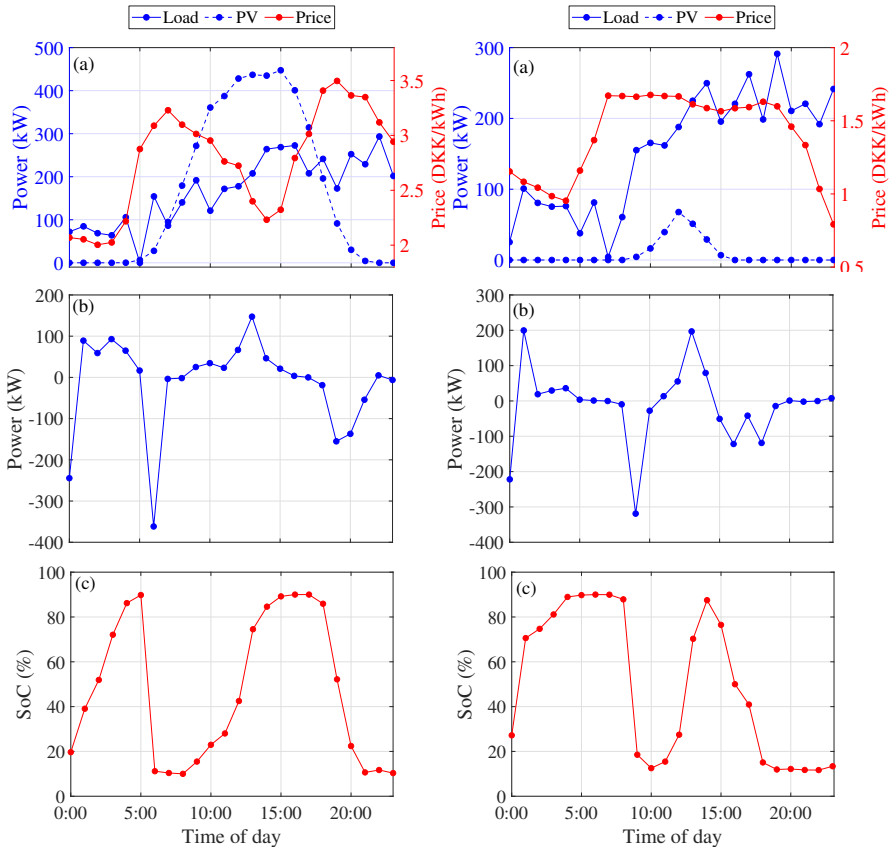
Table 4.9. Performance of different DRL algorithms for energy management of DC microgrids. (Unit: DKK)

Cases	PPO	DDPG	TD3	No Q_{loss}	Baseline*
1	**2379.71**	2611.42	2685.25	2401.47	2047.07
2	1617.56	**1608.33**	1675.00	1689.47	1149.71
3	**-1333.88**	-1115.22	-1103.22	-1139.14	-1490.79
4	**740.70**	753.04	781.84	745.32	661.79
5	**26.91**	89.92	78.31	73.00	-147.69
6	**-1028.29**	-886.60	-932.27	-923.37	-1348.23
7	**10262.67**	10597.85	10526.17	10531.63	9651.45
8	2293.87	**2259.11**	2456.47	2385.97	1990.20
9	**6276.16**	6370.06	6419.89	6398.20	5936.30
10	**4477.51**	4578.09	4705.18	4568.60	4347.27

The scheduling results of the PPO-based DRL energy management for typical summer and winter days are presented in Fig. 4.10. DRL operators can learn optimized dispatch strategies for the BESS by charging it during periods of lower electricity prices and discharging it when prices increase. This is applicable for both summer and winter, accounting for different PV generation profiles and price variations.

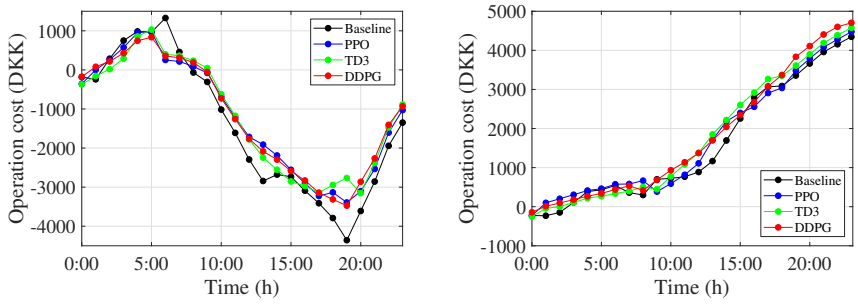
The cumulative operational costs over typical days in both summer and winter of various DRL-based methods are compared to the baseline method for typical days, as depicted in Fig. 4.11. It is observed that the operational costs of the DRL-based energy management framework with different RL algorithms are close to those of the model-based method, and the PPO has superior performance.

In addition, the impact of battery degradation on the DRL-based energy management framework is presented in Fig. 4.12, where the multistage, single-stage degradation model, and no degradation model scenarios are compared under the same operating conditions. Not considering battery degradation leads to increased overall operational costs in the DRL-based energy management framework. Moreover, with the proposed multi-stage battery degradation modeling method, the operation cost is also reduced compared to the DRL with the single-stage degradation model counterpart.



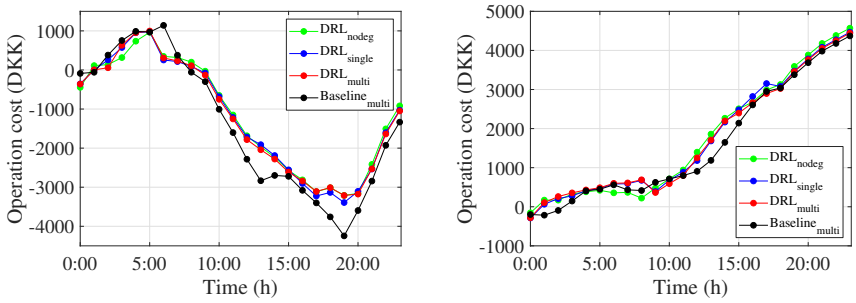
(a) Summer **(b)** Winter

Figure 4.10. Scheduling results of PPO-based energy management framework for DC microgrids.



(a) Summer **(b)** Winter

Figure 4.11. Comparison of the cumulative operation cost over one day for different DRL algorithms.



(a) Summer **(b)** Winter

Figure 4.12. Comparison of the cumulative operation cost over one day for different models.

4.6 Summary

This chapter initially proposes an innovative multi-stage battery degradation modeling method and integrates it into an adaptive energy management framework to minimize operational costs. Compared to conventional single-stage models, the proposed multi-stage degradation modeling method captures the varying aging patterns across the BESS lifespan, leading to a more accurate battery degradation cost evaluation for optimization in different operational stages. Case studies across different stages verify the effectiveness of the proposed method, achieving a reduction of approximately 1.9% in overall operational costs compared to the single-stage degradation model. In addition, as the number of divided stages in the battery lifespan increases, model error steadily decreases to about 0.5%, and the average reduced operational cost rises to approximately 27 DKK.

Furthermore, the proposed battery degradation model is utilized to formulate a real-time immediate reward function for training RL agents, enabling a battery health-aware autonomous DRL-based energy management framework for DC microgrids. The proposed DRL-based energy management framework adapts to varying operating conditions, optimizing dispatch strategies with current system information without the need for precise predictions of demand, prices, and PV generation. The results from various case studies confirm the comparable performance of the proposed method to the model-based approach, which relies on perfect day-ahead predictions of uncertain variables. Moreover, by considering battery degradation, the operational costs associated with the DRL-based energy management frameworks are reduced.

Based on publications

- [**J5**] **Y. Wan**, D. Gebbran, R. K. Subroto and T. Dragičević, "Optimal Day-ahead Scheduling of Fast EV Charging Station With Multi-stage Battery Degra-

dation Model," in IEEE Transactions on Energy Conversion, early access.

- **[C2] Y. Wan**, D. Gebbran, P. I. Gómez, and T. Dragicevic, "Optimal dispatch schedule for a fast EV charging station with account for supplementary battery health degradation," IEEE Transportation Electrification Conference & Expo (ITEC), Anaheim, CA, USA, 2022, pp. 552–556.

CHAPTER 5

Conclusion

5.1 Summary

The Ph.D. project explored the potential of AI techniques to advance DC microgrids for more intelligent, secure, and efficient operations by fusing them into standard hierarchical control structures. Some proposed methods extend beyond DC microgrids and can be applied to other power electronics systems with appropriate modifications, addressing the limitations of conventional control approaches.

The primary advantages of RL-based control methods come from their self-learning and model-free characteristics. In particular, hierarchical control, recognized as a standard control structure, has been widely applied in DC microgrids. Under the multilayer control framework, different control objectives are realized in various layers. In the primary control layer, the power converter control ensures proper operation at the device level. Conventional advanced controllers, such as FCS-MPC, are extensively applied due to their straightforward and flexible formulation of control objectives and constraints. However, the performance of the FCS-MPC relies on the design of the cost function, which needs proper tuning of the weighting factors for desired performance. Moreover, it suffers from parameter sensitiveness and unmodeled dynamics. Therefore, guidelines based on the DDPG algorithm are first introduced to achieve automatic weighting factor design for FCS-MPC, thereby improving desired control performance. Furthermore, the MFRL controller is proposed to emulate the FCS-MPC to select the optimal switching state, requiring no prior system information or mathematical models of the system, which achieves satisfactory control performance comparable to the FCS-MPC. Moreover, a safety framework is proposed to enhance the learning efficiency of the MFRL controller while ensuring the safety of the system throughout the training process, facilitating practical implementation. Both simulated and experimental results confirm the effectiveness of these methods. This finding presents opportunities to apply the online MFRL-based controller in practical power electronics systems.

For the coordination of multiple geographically separated units in DC microgrids, the communication-based distributed control strategy dominates due to its fault resilience, scalability, and enhanced reliability while exposing the system to cyber-attacks. Various cutting-edge cyber-attack detectors are proposed, yet they face critical challenges when confronted with astute adversaries introducing innovative attack patterns. To address these, RL is first employed to emulate intelligent attackers, generating novel attack patterns that autonomously bypass conventional cyber-

attack detectors. Subsequently, a data-driven cyber-attack detector using only system measurements is proposed to detect RL-based intelligent attacks. Consequently, the data-driven cyber-attack detector can complement state-of-the-art detection methods, enhancing the protection of DC microgrids. This approach also opens possibilities for an iterative design method using the proposed attack generation and detection framework, capable of identifying and detecting a broader spectrum of attacks in DC microgrids.

The energy management framework at system-level control aims to achieve economical operation, while the uncertain distributed energy resources and difficulties in modeling a precise DC microgrid system pose challenges to the existing energy management framework. The DRL-based energy management framework can mitigate these challenges and optimize operations using only current system measurements, achieving autonomous operation of DC microgrids. Additionally, a multi-stage battery degradation modeling method is proposed to accurately capture the battery degradation process throughout its lifespan, formulating an immediate and real-time reward for training the RL agent. This model-free DRL-based energy management framework optimizes DC microgrid operations while considering battery degradation, thereby reducing overall operational costs, including battery degradation costs.

In summary, this Ph.D. project initiates an interesting collaboration between DC microgrids and computer science. The primary objective is to develop intelligent control systems that overcome the challenges of the current hierarchical control structure, integrating them with AI technologies for a more intelligent, secure, and efficient DC microgrid.

5.2 Main contributions

- **Unsupervised model-free learning-based local control for interfacing power electronic converters in DC microgrids**
  The proposed approach addresses two aspects of conventional FCS-MPC for power converter control: controller design and model-free predictive control. With the DDPG agent, the RL-based method for optimizing the weighting factors in the cost function of FCS-MPC offers a model-free and automatic weighting factor design process. Furthermore, DQN-based unsupervised learning control provides the opportunity to achieve performance akin to FCS-MPC without requiring system information, effectively mitigating parameter variations and unmodeled dynamics in conventional FCS-MPC. A demonstration framework is proposed to transfer online RL methods from offline simulation to online implementation. Experimental results confirm the effectiveness of the proposed methods.

- **Safety-enhanced self-learning for optimal converter control**
  Implementing a safety policy for the DQN-based controller ensures the system's safety during the online training phase. The proposed MPC-based safety block

poses strict constraints on the DQN-based controller, which guarantees the physical limit of the converter is not violated. More importantly, it also narrows the exploration region for the DQN agent, improving the convergence speed. Both simulation and experimental results confirm the effectiveness of the safe learning strategy and control performance of the controller.

- **ML-enabled cyber-attack identification and detection framework for DC microgrids**
  Employing a MARL approach to emulate intelligent attackers enables the generation of novel and sophisticated attack patterns that remain undetected by conventional metric-based methods. Conversely, a data-driven cyber-attack detector is proposed to complement the failed detection scheme to detect and identify the attacked nodes with only system measurements under identified intelligent attacks. Iteratively, the proposed cybersecurity framework is expected to identify and detect a broader range of attacks, promoting more secure DC microgrids.

- **DRL-based Energy management framework for DC microgrids considering battery degradation**
  The proposed battery health-aware DRL-based energy management framework for DC microgrids optimizes energy flows between the main grid and microgrid by strategically dispatching the battery to minimize overall costs. In addition, a multi-stage battery degradation modeling method is proposed to capture the varying aging patterns across the battery's lifetime, help formulate a more accurate battery degradation cost for the energy management framework, and thus obtain optimal scheduling strategies. Additionally, it aids in designing an immediate and real-time reward function for the DRL-based energy management framework, enabling the learning of optimal battery scheduling strategies under complex operating conditions, considering battery degradation. Numerical results show the effectiveness of the proposed multi-stage battery degradation modeling method and DRL-based energy management framework.

5.3 Perspectives: future research

The presented outcomes in the thesis open many possibilities for future research in DC microgrids and other power electronics systems:

- The learning-based control can also be applied to more complex power converters, such as multi-level converters. Further investigation is needed into the capability of the proposed method to reduce the computational burden associated with the increasing prediction horizon and switching states, a major drawback of conventional FCS-MPC.

- Implementing the proposed safety-enhanced self-learning control in real-time directly on the practical experimental setups in further development adds more practical significance.
- Adapting the generative adversarial network (GAN) concept to the proposed cyber-attack generation and detection framework for DC microgrids. The primary challenge lies in incentivizing the agent to spontaneously explore beyond the already-explored regions to discover multiple optimal policies that can bypass the proposed cyber-attack detection mechanism.
- Exploring resilient control strategies to mitigate the impact of cyber-attacks on the system, based on the employed ML methods, could follow the detection of such attacks. This approach could also achieve a model-free, self-learning mechanism that adapts to time-varying or varied attack scenarios by leveraging the characteristics of ML algorithms.
- Extending the safety-enhanced RL-based method to the energy management framework enforces system constraints during the training process, enhancing feasibility for online implementation in practical systems.
- Similarly, the concept of self-learning optimal power converter control can be implemented in the DRL-based energy management framework. Instead of learning through trial and error, DRL can be trained to imitate the scheduling strategies of conventional model-based methods, thereby achieving optimal energy management.

Bibliography

[1] Y. Zhang, N. Gatsis, and G. B. Giannakis, "Robust energy management for microgrids with high-penetration renewables," *IEEE transactions on sustainable energy*, vol. 4, no. 4, pp. 944–953, 2013.

[2] M. A. Hossain, H. R. Pota, M. J. Hossain, and F. Blaabjerg, "Evolution of microgrids with converter-interfaced generations: Challenges and opportunities," *International Journal of Electrical Power & Energy Systems*, vol. 109, pp. 160–186, 2019.

[3] T. Dragicevic, P. Wheeler, and F. Blaabjerg, *DC distribution systems and microgrids*. Institution of Engineering and Technology, 2018.

[4] H. Kakigano, Y. Miura, and T. Ise, "Low-voltage bipolar-type dc microgrid for super high quality distribution," *IEEE transactions on power electronics*, vol. 25, no. 12, pp. 3066–3075, 2010.

[5] T. Dragičević, X. Lu, J. C. Vasquez, and J. M. Guerrero, "Dc microgrids—part i: A review of control strategies and stabilization techniques," *IEEE Transactions on power electronics*, vol. 31, no. 7, pp. 4876–4891, 2015.

[6] L. Meng, Q. Shafiee, G. F. Trecate, H. Karimi, D. Fulwani, X. Lu, and J. M. Guerrero, "Review on control of dc microgrids and multiple microgrid clusters," *IEEE journal of emerging and selected topics in power electronics*, vol. 5, no. 3, pp. 928–948, 2017.

[7] Q. Shafiee, T. Dragičević, J. C. Vasquez, and J. M. Guerrero, "Hierarchical control for multiple dc-microgrids clusters," *IEEE transactions on energy conversion*, vol. 29, no. 4, pp. 922–933, 2014.

[8] J. Hu, Y. Shan, K. W. Cheng, and S. Islam, "Overview of power converter control in microgrids—challenges, advances, and future trends," *IEEE Transactions on Power Electronics*, vol. 37, no. 8, pp. 9907–9922, 2022.

[9] Z. Zhang, O. Babayomi, T. Dragicevic, R. Heydari, C. Garcia, J. Rodriguez, and R. Kennel, "Advances and opportunities in the model predictive control of microgrids: Part i–primary layer," *International Journal of Electrical Power & Energy Systems*, vol. 134, p. 107411, 2022.

[10] D. Gebbran, A. Barragán-Moreno, P. I. Gómez, R. K. Subroto, M. M. Mardani, M. López, J. Quiroz, and T. Dragičević, "Cloud and edge computing for smart management of power electronic converter fleets: A key connective fabric to enable the green transition," *IEEE Industrial Electronics Magazine*, 2022.

[11] J. M. Guerrero, J. C. Vasquez, J. Matas, L. G. De Vicuña, and M. Castilla, "Hierarchical control of droop-controlled ac and dc microgrids—a general approach toward standardization," *IEEE Transactions on industrial electronics*, vol. 58, no. 1, pp. 158–172, 2010.

[12] P. Zhao, Z. Liu, and J. Liu, "An adaptive discrete piecewise droop control in dc microgrids," *IEEE Transactions on Smart Grid*, 2023.

[13] Y. Gao, S. Wang, H. Hussaini, T. Yang, T. Dragičević, S. Bozhko, P. Wheeler, and S. Vazquez, "Inverse application of artificial intelligence for the control of power converters," *IEEE transactions on power electronics*, vol. 38, no. 2, pp. 1535–1548, 2022.

[14] F. Li, Z. Lin, H. Xu, and R. Wang, "A review of dc bus signalling control methods in dc microgrids," in *2022 IEEE International Power Electronics and Application Conference and Exposition (PEAC)*, pp. 1286–1291, IEEE, 2022.

[15] H.-J. Choi and J.-H. Jung, "Enhanced power line communication strategy for dc microgrids using switching frequency modulation of power converters," *IEEE Transactions on Power Electronics*, vol. 32, no. 6, pp. 4140–4144, 2017.

[16] E. Espina, J. Llanos, C. Burgos-Mellado, R. Cardenas-Dobson, M. Martinez-Gomez, and D. Saez, "Distributed control strategies for microgrids: An overview," *IEEE Access*, vol. 8, pp. 193412–193448, 2020.

[17] T. Dragičević, S. Vazquez, and P. Wheeler, "Advanced control methods for power converters in dg systems and microgrids," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 7, pp. 5847–5862, 2020.

[18] M. Novak and T. Dragicevic, "Supervised imitation learning of finite-set model predictive control systems for power electronics," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 2, pp. 1717–1723, 2020.

[19] B. She, F. Li, H. Cui, J. Zhang, and R. Bo, "Fusion of microgrid control with model-free reinforcement learning: review and vision," *IEEE Transactions on Smart Grid*, 2022.

[20] D. Weber, M. Schenke, and O. Wallscheid, "Safe reinforcement learning-based control in power electronic systems," in *2023 International Conference on Future Energy Solutions (FES)*, pp. 1–6, IEEE, 2023.

[21] P. and Beyond, "Energy sector: More cyber attacks in 2022 than ever before." https://www.power-and-beyond.com/energy-sector-more-cyber-attacks-in-2022-than-ever-before-a-a53dfeb9e1a85d8a0710a010c7a7e7d3/, accessed 2023-12-15.

[22] S. K. Mazumder, A. Kulkarni, S. Sahoo, F. Blaabjerg, H. A. Mantooth, J. C. Balda, Y. Zhao, J. A. Ramos-Ruiz, P. N. Enjeti, P. Kumar, *et al.*, "A review of current research trends in power-electronic innovations in cyber–physical systems," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 9, no. 5, pp. 5146–5163, 2021.

[23] S. Sahoo, J. C.-H. Peng, A. Devakumar, S. Mishra, and T. Dragičević, "On detection of false data in cooperative dc microgrids—a discordant element approach," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 8, pp. 6562–6571, 2019.

[24] S. Tan, P. Xie, J. M. Guerrero, and J. C. Vasquez, "False data injection cyber-attacks detection for multiple dc microgrid clusters," *Applied Energy*, vol. 310, p. 118425, 2022.

[25] S. Madichetty and S. Mishra, "Cyber attack detection and correction mechanisms in a distributed dc microgrid," *IEEE Transactions on Power Electronics*, vol. 37, no. 2, pp. 1476–1485, 2021.

[26] M. Liu, C. Zhao, J. Xia, R. Deng, P. Cheng, and J. Chen, "Pddl: Proactive distributed detection and localization against stealthy deception attacks in dc microgrids," *IEEE Transactions on Smart Grid*, vol. 14, no. 1, pp. 714–731, 2022.

[27] A. Takiddin, S. Rath, M. Ismail, and S. Sahoo, "Data-driven detection of stealth cyber-attacks in dc microgrids," *IEEE Systems Journal*, vol. 16, no. 4, pp. 6097–6106, 2022.

[28] A. Abazari, M. Zadsar, M. Ghafouri, R. Atallah, and C. Assi, "A data mining/anfis and adaptive control for detection and mitigation of attacks on dc mgs," *IEEE Transactions on Smart Grid*, vol. 14, no. 3, pp. 2406–2422, 2022.

[29] M. R. Habibi, H. R. Baghaee, T. Dragičević, and F. Blaabjerg, "Detection of false data injection cyber-attacks in dc microgrids based on recurrent neural networks," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 9, no. 5, pp. 5294–5310, 2020.

[30] N. I. Haque, M. H. Shahriar, M. G. Dastgir, A. Debnath, I. Parvez, A. Sarwat, and M. A. Rahman, "Machine learning in generation, detection, and mitigation of cyberattacks in smart grid: A survey," *arXiv preprint arXiv:2010.00661*, 2020.

[31] S. Wang, M. Du, L. Lu, W. Xing, K. Sun, and M. Ouyang, "Multilevel energy management of a dc microgrid based on virtual-battery model considering voltage regulation and economic optimization," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 9, no. 3, pp. 2881–2895, 2020.

[32] G. Fathi, N. Ghadimi, A. Akbarimajd, and A. B. Dehkordi, "Stochastic-based energy management of dc microgrids," in *Risk-Based Energy Management*, pp. 31–47, Elsevier, 2020.

[33] S. Batiyah, R. Sharma, S. Abdelwahed, and N. Zohrabi, "An mpc-based power management of standalone dc microgrid with energy storage," *International Journal of Electrical Power & Energy Systems*, vol. 120, p. 105949, 2020.

[34] E. Mohammadi, M. Alizadeh, M. Asgarimoghaddam, X. Wang, and M. G. Simões, "A review on application of artificial intelligence techniques in microgrids," *IEEE Journal of Emerging and Selected Topics in Industrial Electronics*, vol. 3, no. 4, pp. 878–890, 2022.

[35] J. Cao, D. Harrold, Z. Fan, T. Morstyn, D. Healey, and K. Li, "Deep reinforcement learning-based energy storage arbitrage with accurate lithium-ion battery degradation model," *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4513–4521, 2020.

[36] V.-H. Bui, A. Hussain, and H.-M. Kim, "Double deep q-learning-based distributed operation of battery energy storage system considering uncertainties," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 457–469, 2019.

[37] B. Zhang, W. Hu, J. Li, D. Cao, R. Huang, Q. Huang, Z. Chen, and F. Blaabjerg, "Dynamic energy conversion and management strategy for an integrated electricity and natural gas system with renewable energy: Deep reinforcement learning approach," *Energy conversion and management*, vol. 220, p. 113063, 2020.

[38] B. Huang and J. Wang, "Deep-reinforcement-learning-based capacity scheduling for pv-battery storage system," *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2272–2283, 2020.

[39] H. Shengren, E. M. Salazar, P. P. Vergara, and P. Palensky, "Performance comparison of deep rl algorithms for energy systems optimal scheduling," in *2022 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, pp. 1–6, IEEE, 2022.

[40] S. F. Schneider, P. Novák, and T. Kober, "Rechargeable batteries for simultaneous demand peak shaving and price arbitrage business," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 1, pp. 148–157, 2020.

[41] Y. Yang, S. Bremner, C. Menictas, and M. Kay, "Modelling and optimal energy management for battery energy storage systems in renewable energy systems: A review," *Renewable and Sustainable Energy Reviews*, vol. 167, p. 112671, 2022.

[42] F. Calero, C. A. Cañizares, K. Bhattacharya, C. Anierobi, I. Calero, M. F. Z. de Souza, M. Farrokhabadi, N. S. Guzman, W. Mendieta, D. Peralta, *et al.*, "A review of modeling and applications of energy storage systems in power grids," *Proceedings of the IEEE*, 2022.

[43] S. Vazquez, J. Rodriguez, M. Rivera, L. G. Franquelo, and M. Norambuena, "Model predictive control for power converters and drives: Advances and trends," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 2, pp. 935–947, 2016.

[44] P. Karamanakos, E. Liegmann, T. Geyer, and R. Kennel, "Model predictive control of power electronic systems: Methods, results, and challenges," *IEEE Open Journal of Industry Applications*, vol. 1, pp. 95–114, 2020.

[45] R. Vargas, P. Cortés, U. Ammann, J. Rodríguez, and J. Pontt, "Predictive control of a three-phase neutral-point-clamped inverter," *IEEE Transactions on Industrial Electronics*, vol. 54, no. 5, pp. 2697–2705, 2007.

[46] P. Cortés, S. Kouro, B. La Rocca, R. Vargas, J. Rodríguez, J. I. León, S. Vazquez, and L. G. Franquelo, "Guidelines for weighting factors design in model predictive control of power converters and drives," in *2009 IEEE International Conference on Industrial Technology*, pp. 1–7, IEEE, 2009.

[47] L. Hu, W. Lei, J. Zhao, and X. Sun, "Optimal weighting factor design of finite control set model predictive control based on multiobjective ant colony optimization," *IEEE Transactions on Industrial Electronics*, 2023.

[48] P. R. U. Guazzelli, W. C. de Andrade Pereira, C. M. R. de Oliveira, A. G. de Castro, and M. L. de Aguiar, "Weighting factors optimization of predictive torque control of induction motor by multiobjective genetic algorithm," *IEEE Transactions on Power Electronics*, vol. 34, no. 7, pp. 6628–6638, 2018.

[49] O. Babayomi, Y. Zhang, Y. Wang, Z. Li, Z. Zhang, *et al.*, "A comparative study on weighting factor design techniques for the model predictive control of power electronics and energy conversion systems," 2021.

[50] T. Dragičević and M. Novak, "Weighting factor design in model predictive control of power electronic converters: An artificial neural network approach," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 11, pp. 8870–8880, 2018.

[51] M. S. O. Yeganeh, A. Oshnoei, S. Peyghami, N. Mijatovic, T. Dragicevic, and F. Blaabjerg, "Weighting factor design for fs-mpc in vscs: A brain emotional

learning-based approach," in *2022 24th European Conference on Power Electronics and Applications (EPE'22 ECCE Europe)*, pp. 01–09, IEEE, 2022.

[52] S. Wang, T. Dragicevic, G. F. Gontijo, S. K. Chaudhary, and R. Teodorescu, "Machine learning emulation of model predictive control for modular multilevel converters," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 11, pp. 11628–11634, 2020.

[53] W. Wu, L. Qiu, J. Rodriguez, X. Liu, J. Ma, and Y. Fang, "Data-driven finite control-set model predictive control for modular multilevel converter," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 11, no. 1, pp. 523–531, 2022.

[54] W. Wu, L. Qiu, X. Liu, F. Guo, J. Rodriguez, J. Ma, and Y. Fang, "Data-driven iterative learning predictive control for power converters," *IEEE Transactions on Power Electronics*, vol. 37, no. 12, pp. 14028–14033, 2022.

[55] X. Liu, L. Qiu, Y. Fang, and J. Rodríguez, "Reinforcement learning-based event-triggered fcs-mpc for power converters," *IEEE Transactions on Industrial Electronics*, 2023.

[56] Y. Zeng, J. Pou, C. Sun, S. Mukherjee, X. Xu, A. K. Gupta, and J. Dong, "Autonomous input voltage sharing control and triple phase shift modulation method for isop-dab converter in dc microgrid: A multiagent deep reinforcement learning-based method," *IEEE Transactions on Power Electronics*, vol. 38, no. 3, pp. 2985–3000, 2022.

[57] T. Dragičević, "Model predictive control of power converters for robust and fast operation of ac microgrids," *IEEE Transactions on Power Electronics*, vol. 33, no. 7, pp. 6304–6317, 2017.

[58] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, pp. 279–292, 1992.

[59] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[60] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[61] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *International conference on machine learning*, pp. 387–395, Pmlr, 2014.

[62] P. Cortés, G. Ortiz, J. I. Yuz, J. Rodríguez, S. Vazquez, and L. G. Franquelo, "Model predictive control of an inverter with output *lc* filter for ups applications," *IEEE Transactions on industrial electronics*, vol. 56, no. 6, pp. 1875–1883, 2009.

[63] Y. Xia, Y. Xu, Y. Wang, S. Mondal, S. Dasgupta, A. K. Gupta, and G. M. Gupta, "A safe policy learning-based method for decentralized and economic frequency control in isolated networked-microgrid systems," *IEEE Transactions on Sustainable Energy*, vol. 13, no. 4, pp. 1982–1993, 2022.

[64] P. Chen, S. Liu, X. Wang, and I. Kamwa, "Physics-shielded multi-agent deep reinforcement learning for safe active voltage control with photovoltaic/battery energy storage systems," *IEEE Transactions on Smart Grid*, 2022.

[65] M. Yazdanian and A. Mehrizi-Sani, "Distributed control techniques in microgrids," *IEEE transactions on smart grid*, vol. 5, no. 6, pp. 2901–2909, 2014.

[66] S. Sahoo, S. Mishra, J. C.-H. Peng, and T. Dragičević, "A stealth cyber-attack detection strategy for dc microgrids," *IEEE Transactions on Power Electronics*, vol. 34, no. 8, pp. 8162–8174, 2018.

[67] A. S. Musleh, G. Chen, and Z. Y. Dong, "A survey on the detection algorithms for false data injection attacks in smart grids," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2218–2234, 2019.

[68] M. R. Habibi, S. Sahoo, S. Rivera, T. Dragičević, and F. Blaabjerg, "Decentralized coordinated cyberattack detection and mitigation strategy in dc microgrids based on artificial neural networks," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 9, no. 4, pp. 4629–4638, 2021.

[69] A. Basati, N. Bazmohammadi, J. M. Guerrero, and J. C. Vasquez, "Real-time estimation in cyber attack detection and mitigation framework for dc microgrids," in *2023 23rd International Scientific Conference on Electric Power Engineering (EPE)*, pp. 1–6, IEEE, 2023.

[70] J. Zhang, S. Sahoo, J. C.-H. Peng, and F. Blaabjerg, "Mitigating concurrent false data injection attacks in cooperative dc microgrids," *IEEE Transactions on Power Electronics*, vol. 36, no. 8, pp. 9637–9647, 2021.

[71] S. Sahoo, T. Dragičević, and F. Blaabjerg, "An event-driven resilient control strategy for dc microgrids," *IEEE Transactions on Power Electronics*, vol. 35, no. 12, pp. 13714–13724, 2020.

[72] S. Sahoo, J. C.-H. Peng, S. Mishra, and T. Dragičević, "Distributed screening of hijacking attacks in dc microgrids," *IEEE Transactions on Power Electronics*, vol. 35, no. 7, pp. 7574–7582, 2019.

[73] S. Sahoo, T. Dragičević, and F. Blaabjerg, "Multilayer resilience paradigm against cyber attacks in dc microgrids," *IEEE Transactions on Power Electronics*, vol. 36, no. 3, pp. 2522–2532, 2020.

[74] G. James, D. Witten, T. Hastie, R. Tibshirani, *et al.*, *An introduction to statistical learning*, vol. 112. Springer, 2013.

[75] S. Paul, A. P. Nath, and Z. H. Rather, "A multi-objective planning framework for coordinated generation from offshore wind farm and battery energy storage system," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 4, pp. 2087–2097, 2019.

[76] M. F. Zia, M. Nasir, E. Elbouchikhi, M. Benbouzid, J. C. Vasquez, and J. M. Guerrero, "Energy management system for a hybrid pv-wind-tidal-battery-based islanded dc microgrid: Modeling and experimental validation," *Renewable and Sustainable Energy Reviews*, vol. 159, p. 112093, 2022.

[77] M. F. Zia, E. Elbouchikhi, and M. Benbouzid, "Microgrids energy management systems: A critical review on methods, solutions, and prospects," *Applied energy*, vol. 222, pp. 1033–1055, 2018.

[78] G. S. Thirunavukkarasu, M. Seyedmahmoudian, E. Jamei, B. Horan, S. Mekhilef, and A. Stojcevski, "Role of optimization techniques in microgrid energy management systems—a review," *Energy Strategy Reviews*, vol. 43, p. 100899, 2022.

[79] M. Beaudin and H. Zareipour, "Home energy management systems: A review of modelling and complexity," *Renewable and sustainable energy reviews*, vol. 45, pp. 318–335, 2015.

[80] S. Ferahtia, A. Djeroui, H. Rezk, A. Houari, S. Zeghlache, and M. Machmoum, "Optimal control and implementation of energy management strategy for a dc microgrid," *Energy*, vol. 238, p. 121777, 2022.

[81] M. G. Abdolrasol, R. Mohamed, M. A. Hannan, A. Q. Al-Shetwi, M. Mansor, and F. Blaabjerg, "Artificial neural network based particle swarm optimization for microgrid optimal energy scheduling," *IEEE Transactions on Power Electronics*, vol. 36, no. 11, pp. 12151–12157, 2021.

[82] O. Han, T. Ding, C. Mu, Y. Huang, X. Zhang, and Z. Ma, "Multi-time scale optimal dispatch for the wind power integrated system with demand response of data centers based on neural network-based model predictive control," *IEEE Transactions on Industry Applications*, 2023.

[83] X. Fang and J. Khazaei, "A two-stage deep learning approach for solving microgrid economic dispatch," *IEEE Systems Journal*, 2023.

[84] L. Yu, S. Qin, M. Zhang, C. Shen, T. Jiang, and X. Guan, "A review of deep reinforcement learning for smart building energy management," *IEEE Internet of Things Journal*, vol. 8, no. 15, pp. 12046–12063, 2021.

[85] B. Xu, J. Zhao, T. Zheng, E. Litvinov, and D. S. Kirschen, "Factoring the cycle aging cost of batteries participating in electricity markets," *IEEE Transactions on Power Systems*, vol. 33, no. 2, pp. 2248–2259, 2017.

[86] M. Alramlawi and P. Li, "Design optimization of a residential pv-battery microgrid with a detailed battery lifetime estimation model," *IEEE Transactions on Industry Applications*, vol. 56, no. 2, 2020.

[87] S. Choi and S.-W. Min, "Optimal scheduling and operation of the ess for prosumer market environment in grid-connected industrial complex," *IEEE Transactions on Industry Applications*, vol. 54, no. 3, pp. 1949–1957, 2018.

[88] U. R. Nair, M. Sandelic, A. Sangwongwanich, T. Dragičević, R. Costa-Castello, and F. Blaabjerg, "An analysis of multi objective energy scheduling in pv-bess system under prediction uncertainty," *IEEE Transactions on Energy Conversion*, vol. 36, no. 3, pp. 2276–2286, 2021.

[89] Z. Qiu, W. Zhang, S. Lu, C. Li, J. Wang, K. Meng, and Z. Dong, "Charging-rate-based battery energy storage system in wind farm and battery storage cooperation bidding problem," *CSEE Journal of Power and Energy Systems*, vol. 8, no. 3, pp. 659–668, 2022.

[90] M. Elkazaz, M. Sumner, and D. Thomas, "Energy management system for hybrid pv-wind-battery microgrid using convex programming, model predictive and rolling horizon predictive control with experimental validation," *International Journal of Electrical Power & Energy Systems*, vol. 115, p. 105483, 2020.

[91] J. Faraji, A. Ketabi, and H. Hashemi-Dezaki, "Optimization of the scheduling and operation of prosumers considering the loss of life costs of battery storage systems," *Journal of Energy Storage*, vol. 31, p. 101655, 2020.

[92] S. Wang, D. Guo, X. Han, L. Lu, K. Sun, W. Li, D. U. Sauer, and M. Ouyang, "Impact of battery degradation models on energy management of a grid-connected dc microgrid," *Energy*, vol. 207, p. 118228, 2020.

[93] K. Abdulla, J. De Hoog, V. Muenzel, F. Suits, K. Steer, A. Wirth, and S. Halgamuge, "Optimal operation of energy storage systems considering forecasts and battery degradation," *IEEE Transactions on Smart Grid*, vol. 9, no. 3, pp. 2086–2096, 2016.

[94] M. Lucu, E. Martinez-Laserna, I. Gandiaga, K. Liu, H. Camblong, W. D. Widanage, and J. Marco, "Data-driven nonparametric li-ion battery ageing model aiming at learning from real operation data-part b: Cycling operation," *Journal of Energy Storage*, vol. 30, p. 101410, 2020.

[95] Y. Gao, J. Jiang, C. Zhang, W. Zhang, Z. Ma, and Y. Jiang, "Lithium-ion battery aging mechanisms and life model under different charging stresses," *Journal of Power Sources*, vol. 356, pp. 103–114, 2017.

[96] Y. Qin, C. Yuen, X. Yin, and B. Huang, "A transferable multistage model with cycling discrepancy learning for lithium-ion battery state of health estimation," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 1933–1946, 2022.

[97] C. Zhao and X. Li, "Microgrid optimal energy scheduling considering neural network based battery degradation," *IEEE Transactions on Power Systems*, 2023.

[98] S. N. Motapon, E. Lachance, L.-A. Dessaint, and K. Al-Haddad, "A generic cycle life model for lithium-ion batteries based on fatigue theory and equivalent cycle counting," *IEEE Open Journal of the Industrial Electronics Society*, vol. 1, pp. 207–217, 2020.

[99] S. N. Laboratories, "Battery archive." `https://batteryarchive.org`, accessed 2023-06-15.

[100] K. Smith, M. Earleywine, E. Wood, J. Neubauer, and A. Pesaran, "Comparison of plug-in hybrid electric vehicle battery life across geographies and drive-cycles," tech. rep., National Renewable Energy Lab.(NREL), Golden, CO (United States), 2012.

[101] I. Laresgoiti, S. Käbitz, M. Ecker, and D. U. Sauer, "Modeling mechanical degradation in lithium ion batteries during cycling: Solid electrolyte interphase fracture," *Journal of Power Sources*, vol. 300, pp. 112–122, 2015.

[102] J. Lin, D. Gebbran, and T. Dragičević, "Surrogate-assisted combinatorial optimization of ev fast charging stations," *IEEE Transactions on Transportation Electrification*, 2023.

[103] Y. Lin, Y.-W. Chen, and J.-T. Yang, "Optimized thermal management of a battery energy-storage system (bess) inspired by air-cooling inefficiency factor of data centers," *International Journal of Heat and Mass Transfer*, vol. 200, p. 123388, 2023.

[104] D. Tran and A. M. Khambadkone, "Energy management for lifetime extension of energy storage system in micro-grid applications," *IEEE Transactions on Smart Grid*, vol. 4, no. 3, pp. 1289–1296, 2013.

[105] J.-O. Lee and Y.-S. Kim, "Novel battery degradation cost formulation for optimal scheduling of battery energy storage systems," *International Journal of Electrical Power & Energy Systems*, vol. 137, p. 107795, 2022.

[106] W. Chen, J. Qiu, J. Zhao, Q. Chai, and Z. Y. Dong, "Bargaining game-based profit allocation of virtual power plant in frequency regulation market considering battery cycle life," *IEEE Transactions on Smart Grid*, vol. 12, no. 4, pp. 2913–2928, 2021.

[107] C. Liu, X. Wang, X. Wu, and J. Guo, "Economic scheduling model of microgrid considering the lifetime of batteries," *IET Generation, Transmission & Distribution*, vol. 11, no. 3, pp. 759–767, 2017.

[108] Y. Qin, H. Hua, and J. Cao, "Stochastic optimal control scheme for battery lifetime extension in islanded microgrid via a novel modeling approach," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 4467–4475, 2018.

[109] W. E. Hart, J.-P. Watson, and D. L. Woodruff, "Pyomo: modeling and solving mathematical programs in python," *Mathematical Programming Computation*, vol. 3, pp. 219–260, 2011.

[110] L. T. Biegler and V. M. Zavala, "Large-scale nonlinear programming using ipopt: An integrating framework for enterprise-wide dynamic optimization," *Computers & Chemical Engineering*, vol. 33, no. 3, pp. 575–582, 2009.

[111] Energinet, "ENERGI DATA SERVICE." `https://www.energidataservice.dk/`, accessed 2023-12-05.

Part II

Publications

Collection of relevant publications

- [**J1**] **Y. Wan**, T. Dragicevic and X. Qian "Unsupervised learning-based Predictive Control for Power Electronic Converters," in IEEE Transactions on Industrial Electronics, under review.

- [**J2**] **Y. Wan**, T. Dragicevic and X. Qian "Safety-Enhanced Self-Learning for Optimal Power Converter Control," in IEEE Transactions on Industrial Electronics, under review.

- [**J3**] **Y. Wan** and T. Dragicevic, "Data-driven Cyber-attack Detection of Intelligent Attacks in Islanded DC Microgrids," in IEEE Transactions on Industrial Electronics, vol. 70, no. 4, pp. 4293-4299, April 2023.

- [**J4**] A. J. Abianeh, **Y. Wan**, F. Ferdowsi, N. Mijatovic and T. Dragičević, "Vulnerability Identification and Remediation of FDI Attacks in Islanded DC Microgrids Using Multiagent Reinforcement Learning," in IEEE Transactions on Power Electronics, vol. 37, no. 6, pp. 6359-6370, June 2022.

- [**J5**] **Y. Wan**, D. Gebbran, R. K. Subroto and T. Dragičević, "Optimal Day-ahead Scheduling of Fast EV Charging Station With Multi-stage Battery Degradation Model," in IEEE Transactions on Energy Conversion, early access.

- [**C1**] **Y. Wan**, T. Dragicevic, N. Mijatovic, C. Li and J. Rodriguez, "Reinforcement Learning Based Weighting Factor Design of Model Predictive Control for Power Electronic Converters," 2021 IEEE International Conference on Predictive Control of Electrical Drives and Power Electronics (PRECEDE), Jinan, China, 2021, pp. 738-743.

- [**C2**] **Y. Wan**, D. Gebbran, P. I. Gómez and T. Dragicevic, "Optimal dispatch schedule for a fast EV charging station with account to supplementary battery health degradation," IEEE Transportation Electrification Conference & Expo (ITEC), Anaheim, CA, USA, 2022, pp. 552-556.

[J1] Unsupervised learning-based Predictive Control for Power Electronic Converters

Authors:
Yihao Wan, Qianwen Xu, Tomislav Dragicevic

Submitted to:
IEEE Transactions on Industrial Electronics

Status:
Under review

Unsupervised learning-based Predictive Control for Power Electronic Converters

Yihao Wan, *Student Member, IEEE,* Qianwen Xu, *Member, IEEE,* and Tomislav Dragičević, *Senior Member, IEEE*

***Abstract*—Finite-set model predictive control (FS-MPC) appears to be a promising and effective control method for power electronic converters. Conventional FS-MPC suffers from the time-consuming process of weighting factor selection, which significantly impacts control performance. In addition, another ongoing challenge of FS-MPC is the dependence on the prediction model for desirable control performance. To overcome the above issues, we propose to apply reinforcement learning (RL) to FS-MPC for power converters. The RL algorithm is first employed for the automatic weighting factor design of the FS-MPC, aiming to minimize the total harmonic distortion (THD) or reduce the average switching frequency. Furthermore, by formulating the incentive for the RL agent with the cost function of the predictive algorithm, the agent learns autonomously to find the optimal switching policy for the power converter by imitating the predictive controller without prior knowledge of the system model. Finally, a deployment framework that allows for experimental validation of the proposed RL-based methods on a practical FS-MPC regulated stand-alone converter configuration is presented. Two exemplary control objectives are demonstrated to show the effectiveness of the proposed RL-aided weighting factor tuning method. Moreover, the results show a good match between the model-free RL-based imitation controller and the FS-MPC performance.**

***Index Terms*—Finite-set model predictive control (FS-MPC), reinforcement learning (RL), unsupervised imitation controller, voltage source converter (VSC), and weighting factor design.**

I. Introduction

POWER electronic converters are one of the critical enabling technologies in DC microgrids, which interface DC microgrids with renewable energy sources, energy storage systems, electric vehicles, and AC systems [1]. Conventional linear control structures employ cascaded linear loops to control the converters, which have simple and analytical control synthesis. However, the slow transient performance, lack of flexibility for multi-objective controls, etc., limit its applications. Different advanced control methods have been proposed to address the obstacles of conventional cascaded linear control structures [2], among which model predictive control (MPC) stands out due to its simple inclusion of control objectives and straightforward design.

In general, MPC relies on the converter model to predict the future behavior of output voltage for all the possible switching configurations, and a cost function is used as the criterion for selecting the optimal switching state of the following sampling interval to achieve desired control performances [3]. Despite its advantages, there are still challenges with the MPC for converter control. The first is weighting factor design in the cost function for desirable control performance and balancing different control objectives. Conventionally, time-consuming simulations for optimal weighting factor selection are conducted by manual trial-and-error with different combinations of the parameters [4]. One of the pioneering proposals to relieve this is the branch-and-bound search method [5], while it is still too empirical. In [6], an artificial neural network (ANN) approach is proposed for automatic weighting factor design. However, it requires numerous simulations using the model to generate the training data, covering a wide range of conditions by parameter sweeping. In addition, an exhaustive search procedure to optimize weighting factors for different objectives is necessary. In [7], a model-free brain emotional learning (BEL) based weighting factor method is proposed. However, the method introduces more coefficients required for tuning within the BEL method, increasing the complexity of controller design instead.

Another challenge is the high reliance on the converter model for the predictive controller to achieve optimized performance [8], and the high computation burden for online implementation, especially when the model order is high, e.g., multicell or multilevel converters. Even though different efforts have been carried out to mitigate these with the data-based methods, e.g., supervised NN-based controller for voltage source converter (VSC) [9], [10], ML-based controller for a modular multilevel converter (MMC) [11], and 5-level flying capacitor converter [12], etc., in principle, those methods require data generated by offline simulations based on the converter models in a supervised way to train the NN imitator. To solve this, in [13], [14], a model-free data-driven predictive current control framework is proposed by synthesizing a model-free adaptive control (MFAC) with the FS-MPC for MMC and a three-level neutral-point-clamped converter. However, the MFAC employs a non-linear function to map the discrete-time dynamics of the controller, which could be further improved by the NN. In addition, computational costs are increased due to the introduction of several hyperparameters in the MFAC for fine-tuning.

Recently, reinforcement learning (RL) has gained much attention for applications in the power electronics field due to its model-free and self-learning characteristics. In [15], the RL is employed to tune the weights of the neuro-fuzzy controller for the active and reactive power control of distributed generators in microgrids. In [16], a self-adaptive active disturbance rejection control with RL-based controller coefficients tuning method is proposed to mitigate the instability issues caused by constant power load for DC/DC converters. Similarly, Cui et al. [17] proposed generating the duty ratio with RL to stabilize the DC/DC converter. In addition, different RL

Yihao Wan and Tomislav Dragičević are with the Department of Wind and Energy Systems, Technical University of Denmark, Copenhagen, Denmark (e-mails: wanyh@dtu.dk, tomdr@dtu.dk).

Qianwen Xu is with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden (email: qianwenx@kth.se).

agents are applied to a dual active bridge converter to find the optimal phase shift angle of the triple phase shift modulation to minimize the reactive power [18], power losses [19], and current stress [20]. In [21], RL is integrated into the predictive control framework to improve the robustness of the control to different operating conditions by generating an optimal feedback control reference for the predictive controller.

Motivated by the advancements of RL, this paper proposes to leverage the RL algorithms to address the above-mentioned challenges of FS-MPC for power converters. Particularly, an RL-based automatic weighting factor design method is proposed for the FS-MPC. By defining different objectives, the RL agent will learn autonomously by interacting with the system to find the optimal weighting factor tuning policy for FS-MPC to achieve desired control performance. It directly outputs the optimal weighting factors without an additional optimization stage. Furthermore, an unsupervised RL-based controller is proposed to imitate the conventional FS-MPC for power converters. By properly configuring the RL agent, it can find the optimal switching strategy by emulating the switching state selection of FS-MPC to achieve the desired control performance. In addition, the methods are model-free, which can be directly implemented with the input measurements. Data generation covering a wide range of operating conditions using the system model is thus not necessary. The main contributions of the paper are summarized below:

- A RL-based automatic weighting factor design method for FS-MPC. The proposed method can automatically select optimal weighting factors for the predictive controller to achieve desired control objectives.
- A model-free RL-based imitation controller for converters. The RL agent can autonomously find the optimal switching strategy by emulating the FS-MPC to achieve desired control performance for VSC without the requirement of model information.
- A deployment framework for transferring online RL from simulation to practical experimental demonstration. The RL agents are trained offline and implemented online, avoiding time-consuming and storage-intensive online computation during training. In this way, the proposed methods can be validated experimentally.

The rest of the paper is structured as follows. Section II introduces the model of the studied MPC-regulated system. Section III introduces the proposed RL applications in MPC controller design and imitation control. Section IV presents the implementation framework and experimental results for different case studies. Section V concludes the paper.

II. System model

A. Converter model

In Fig. 1, a two-level VSC interfaces between the load and DC source, where the FS-MPC is used to regulate it for the uninterruptible power supply (UPS) application. The FS-MPC predicts the converter behaviors in the future for all the possible switching combinations based on the prediction model, starting from the most recent measurements. A cost function denoting the desired control objective evaluates the performance of those predictions. The optimal switching policy for current states, i.e., with the least cost function value, will be applied to the converter.

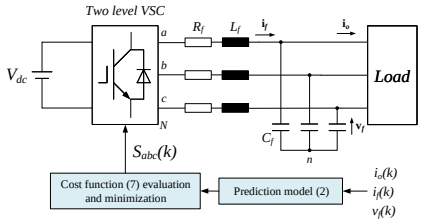


Fig. 1. Predictive control for VSC.

The VSC is modeled in a stationary $\alpha\beta$ reference frame. Thus, amplitude-invariant Clarke transformation is employed to convert all the three-phase filter current $\mathbf{i}_f$, load current $\mathbf{i}_o$, and output capacitor voltage $\mathbf{v}_f$ into the corresponding $\alpha\beta$ frame from abc frame. In addition, the upper arm switching states $S = \{S_a, S_b, S_c\}$ for the three-phase legs are modeled as "ON" for "1" and "OFF" for "0". Therefore, considering all the possible combinations of switching states for the three phases, there are eight switching configurations. The phase voltage of the VSC can be derived by multiplying DC link voltage and associated gating signal, as $v_{aN} = S_a \cdot v_{dc}$, $v_{bN} = S_b \cdot v_{dc}$, and $v_{cN} = S_c \cdot v_{dc}$. Therefore, the output voltage vector of the inverter can be expressed as [3]

$$\mathbf{v} = \frac{2}{3}(v_{aN} + \boldsymbol{a} v_{bN} + \boldsymbol{a}^2 v_{cN}) \tag{1}$$

where $\boldsymbol{a} = e^{j(2\pi/3)}$, representing the 120° phase displacement between the phases. The LC filter model can be expressed with the differential equations of the inductance current and capacitor voltage in state-space representation, as

$$\frac{d}{dt}\begin{bmatrix}\mathbf{i}_f \\ \mathbf{v}_f\end{bmatrix} = \boldsymbol{A}\begin{bmatrix}\mathbf{i}_f \\ \mathbf{v}_f\end{bmatrix} + \boldsymbol{B}\begin{bmatrix}\mathbf{v}_i \\ \mathbf{i}_o\end{bmatrix} \tag{2}$$

where $\boldsymbol{A} = \begin{bmatrix}-\frac{R_f}{L_f} & -\frac{1}{L_f} \\ -\frac{1}{C_f} & 0\end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix}-\frac{1}{L_f} & 0 \\ 0 & -\frac{1}{C_f}\end{bmatrix}$.

For implementing the digital control, the model is discretized with a sampling time T_s based on Euler forward discretization method [6].

B. Cost function

Considering the computational complexity of the multistep prediction horizon, switching frequency, and topology of the two-level VSC, a single-step horizon is used for the case study in the paper. The cost function evaluating the switching strategy is expressed as [8]

$$g = (v_\alpha^* - v_{f\alpha})^2 + (v_\beta^* - v_{f\beta})^2 \tag{3}$$

where v_α^* and v_β^* are the reference voltage.

An additional current reference term introduced to improve the steady-state performance is expressed as [22]

$$g_d = (C_f \omega_r v_\beta^* - i_{f\alpha} + i_{o\alpha})^2 + (C_f \omega_r v_\alpha^* + i_{f\beta} - i_{o\beta})^2 \tag{4}$$

Moreover, penalization terms to reduce the switching frequency and limit inductor current are employed, as below

$$sw = |\triangle S_a(i)| + |\triangle S_b(i)| + |\triangle S_c(i)| \tag{5}$$

$$h_{lim} = \begin{cases} 0, & |\bar{i}_f| \leq i_{max} \\ \infty, & |\bar{i}_f| > i_{max} \end{cases} \quad (6)$$

where $|\triangle S_j(i)|(j = a, b, c)$ is 1 if switching states change in the phase leg at time i and 0 vice versa.

Therefore, the comprehensive cost function incorporating all the terms can be formulated as

$$g = (v_\alpha^* - v_{f\alpha})^2 + (v_\beta^* - v_{f\beta})^2 + \lambda_d g_d + \lambda_{sw} sw^2 + h_{lim} \quad (7)$$

It can be observed that two weighting factors assigned to different control objectives, i.e., λ_d and λ_{sw}, should be selected properly to achieve the desired control performance. Moreover, the FS-MPC relies heavily on the prediction model to predict the future behavior of the converter. How to mitigate the reliance on the system model is still an open research problem. In the following sections, the paper proposes to employ the RL to solve the above issues.

III. Model-free reinforcement learning methods for FS-MPC in converter control

In this section, model-free reinforcement learning (MFRL) algorithms are applied to the FS-MPC for converter control. The application framework involves problem formulation, configuring the RL algorithms according to the case studies, and implementation. We first present preliminaries of RL algorithms. Then the proposed RL-based automated weighting factor design is presented for expected control performance. Finally, we propose RL-based unsupervised limitation learning to achieve optimal performance and fast online implementation without the requirement of model information.

A. Formulation of RL

RL paradigm is formulated as a Markov Decision Process (MDP) characterized with the tuple $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$ for the agent, where $\mathcal{S}$ is the state space, $\mathcal{A}$ represents the action space, $\mathcal{P}$ denotes the state transition function for characterizing the distribution over states regarding $s_t \rightarrow s_{t+1}$, R is the reward function for evaluating the goal of the problem, and $\gamma \in [0, 1]$ is the discount factor for balancing the agent learning with immediate reward and long term reward. As shown in Fig. 2, the RL agent interacts with the environment iteratively. At each time step, the RL agent receives the environment state, $s_t \in \mathcal{S}$, and on that basis selects an action, $a_t \in \mathcal{A}$, according to the policy π. Then at the time step t+1, the agent receives a reward $r_t \in R$ and the environment proceeds to a new state s_{t+1} based on the transition function P. The iteration terminates until the RL agent finds the optimal policy as

$$\pi^* \in \underset{\pi}{\operatorname{argmax}} G(\pi) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_t(s_t, a_t) \right] \quad (8)$$

where $G(\pi)$ represents the accumulative reward over an infinite time horizon, $\mathbb{E}[\cdot]$ denotes the expected value of a variable. The optimal policy guarantees that the expected return for all the states and actions is greater than other policies.

There are two main RL algorithms, i.e., value-based and policy gradient algorithms. Value-based algorithms, e.g., Q-learning and its derivations, learn the Q function that evaluates the state-action pairs and take actions that maximize the

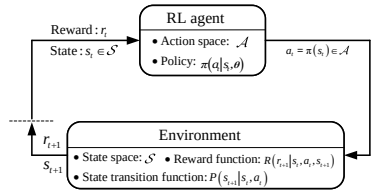


Fig. 2. The agent environment interaction in an MDP.

Q-value. Policy gradient methods directly learn the policy $\pi(a_t|s_t, \theta)$, using either stochastic or deterministic estimates of policy gradient. Based on the characteristics of the formulated problem, the RL agents should be appropriately selected.

Deep deterministic policy gradient (DDPG) with continuous actions space and deep Q-learning network (DQN) with discrete action space are two basic paradigms of MFRL algorithms mentioned above, which are employed in this studies to demonstrate how the MFRL methods are applied to controller design and converter control respectively.

B. DDPG-based weighting factor design for FS-MPC

Fig. 3 presents the overall schematic of the proposed automatic weighting factor design approach. The environment consists of an MPC-regulated VSC for the UPS application. The measurements are selected as the states for the agent, along with a calculated reward based on the current status of the system. The RL agent will output the λ_d and λ_{sw} to the FS-MPC, and the system transits to a new state by executing the actions. Iteratively, the RL agent learns autonomously to find the optimal weighting factors tuning policy for FS-MPC. Eventually, the optimal weighting factors for desired control performance are obtained.

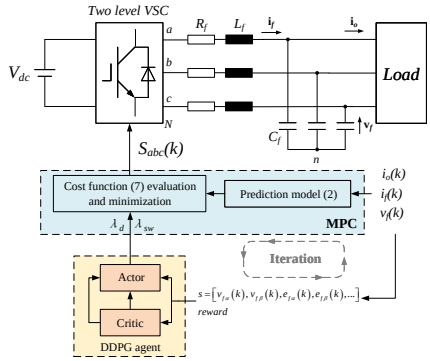


Fig. 3. Block diagram of RL-based weighting factor tuning of FS-MPC for VSC.

1) RL agents: As mentioned in Section III-A, the weighting factor design problem is defined as a Markov Decision Process (MDP). The RL agent explores and exploits the action space to maximize the long-term accumulative reward over an infinite time horizon, which is expressed as

$$G_t = \sum_{k=0}^{\infty} \gamma^n r_{t+k+1} \quad (9)$$

In order to find the optimal policy $\pi(a_t|s_t)$ to maximize the accumulative reward in (9), different recursive training algorithms can be applied. In the off-policy algorithms such as Q-learning, the action-value function $Q_\pi(s_t, a_t)$ is used to evaluate the expected return for action a_t with respect to the state s_t, which is expressed as Bellman equation below [23]

$$\begin{aligned} Q_\pi(s_t,a_t) \leftarrow Q_\pi(s_t, a_t)+ \\ \alpha[r_{t+1} + \gamma max Q_\pi(s_{t+1}, a_{t+1}) - Q^\pi(s_t, a_t)] \end{aligned} \tag{10}$$

where α is the learning rate of the RL agent.

However, the iterative approach cannot achieve desirable performance in high-dimensional real-world applications. For a more precise and efficient prediction of the Q value for each state-action pair, a deep neural network is employed [24]. In the DQN, transition sequences $e = (s_t, a_t, r_t, s_{t+1})$ are stored in a R-sized experience memory $\mathcal{M} = \{e_1, ..., e_R\}$. During the training session, a minibatch of the tuples is randomly selected from $\mathcal{M}$ to train the NN via stochastic gradient descent. In addition, to enhance the stability and convergence of training Q-network with network parameter θ, a target network $Q_\pi^{'}$ is used. The target value is calculated as

$$y_t = r_t + \gamma Q_\pi^{'}(s_t^{'}, a_t^{'}|\theta_Q^{'}) \tag{11}$$

The loss function for updating the Q-network parameter is

$$L(\theta_Q) = \mathbb{E}\left[(y_t - Q_\pi(s_t, a_t|\theta_Q))^2\right] \tag{12}$$

The target network keeps its separate network parameter $\theta^{'}$ fixed for every T_{target} steps and is updated with the current weights of the Q-network as

$$\theta_Q^{'} \leftarrow \tau\theta + (1-\tau)\theta_Q^{'} \tag{13}$$

where the smoothing factor $\tau \ll 1$.

While DQN is feasible for problems with high-dimensional observation spaces, it can only handle a limited set of actions. In the scenarios of weighting factor design, the weighting factor values cover a wide range within the action space, making the number of actions increase exponentially and the training intractable. Therefore, the RL agents capable of handling continuous action space using actor-critic networks, such as DDPG [25], are more suitable and employed. Similar to the DQN, the critic network is updated by minimizing the loss function in (12), another actor function $\mu(s|\theta_\mu)$ specifies the current policy by mapping states to the specific action. The actor network is updated with the equation below [26]

$$\nabla_{\theta_\mu} J = \mathbb{E}_{s_t\sim\rho^\beta}\left[\nabla_{\theta_\mu}(s|\theta_\mu)|s = s_t \nabla_a Q_\pi(s, a|\theta_Q)|_{s=s_t, a=\mu(s_t)}\right] \tag{14}$$

where $s \sim \rho^\beta$ represents the state s following the distribution β. Then the target networks of the critic and actor networks are updated with (13). In addition, a noise model $\mathcal{N}$ decaying by training episodes is usually used to help the agent balance exploration and exploitation better [25].

2) State and action sets: As shown in Fig. 3, the DDPG agent receives states from the MPC-controlled VSC system and tunes the weighting factors of the MPC controller. As the goal is to achieve the desired control performance, therefore, the state is defined as $s = [v_{f\alpha}, v_{f\beta}, e_\alpha, e_\beta, THD]$, where e_α and e_β are the reference voltage tracking error. The action set is defined as $a = [\lambda_d, \lambda_{sw}]$. In addition, for more efficient training, the action exploration space is limited within a feasible range [0, 10] for both actions.

3) Reward function design: The reward is formulated according to desired control performances, minimizing the THD and reducing the average switching frequency. Therefore, the reward function should be properly designed for a specific control objective. Two exemplary goals are studied, i.e., minimization of the THD exclusively and simultaneously decreasing the switching frequency f_{sw}.

For the first case study, the reward function is defined as

$$r = -THD^2 \tag{15}$$

Therefore, the reward will incentivize the RL agent to tune the two weighting factors of the FS-MPC to minimize the THD.

For the second case study, the reward function is defined as

$$r = -(k_{thd} \cdot THD^2 + k_{sw} \cdot {f_{sw}}^2) \tag{16}$$

where k_{thd} and k_{sw} are the coefficients for the THD and f_{sw} respectively. The f_{sw} is calculated by averaging the total SW from equation (5) within a certain operating time [6]. The coefficients could be selected according to the importance of the control objectives. In the case study, more importance is given to reducing the THD, thus $k_{thd} = 2$ and $k_{sw} = 1$. The training process of DDPG for FS-MPC controller design is described in **Algorithm 1**.

Algorithm 1 DDPG-based controller design for FS-MPC

Input: Measurements: $[v_{f\alpha}, v_{f\beta}, e_\alpha, e_\beta, THD]$
Output: Weighting factors λ_d and λ_{sw}

Initialize replay buffer $\mathcal{M}$
Initialize critic network Q and actor-network μ with random weights θ_Q and θ_μ
Initialize target network $Q^{'}$ and actor-network $\mu^{'}$ with weights $\theta_Q^{'} = \theta_Q$, $\theta_\mu^{'} = \theta_\mu$
for episode = 1 to M **do**
  Initialize a random noise $\mathcal{N}$ for action exploration
  Receive initial observation at state s_1
  for iteration = 1 to T **do**
    Select action with current policy and noise
    Excute the action a_t, observe the reward r_t based on (15) or (16) and new state s_{t+1}
    Store tuple (s_t, a_t, r_t, s_{t+1}) in $\mathcal{M}$
    Randomly sample the mini-batch e_j from $\mathcal{M}$
    Set $y_j = r_j + \gamma Q^{'}(s_{i+1}, \mu^{'}(s_{i+1}|\theta_\mu^i)|\theta_Q^{'})$
    Update the critic by minimizing (12)
    Update the actor with (14)
    Update the target networks using (13)
  end for
end for

C. DQN-based unsupervised imitation controller of FS-MPC

As shown in Fig. 4, to address the reliance on the system model for conventional FS-MPC, we propose a novel unsupervised model-free RL-based imitation controller to emulate the FS-MPC to achieve desired control performance for power

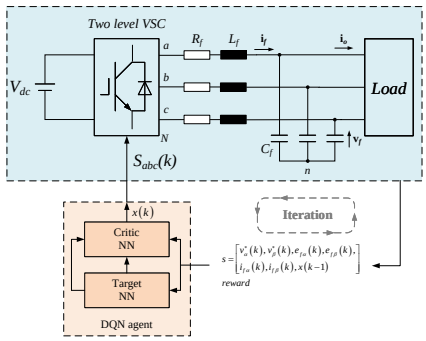


Fig. 4. Block diagram of unsupervised MFRL-based imitation controller of MPC-controlled VSC.

electronic converters with light online computation requirement and without prior model information.

Fig. 4 presents the RL-based controller for VSC. Due to the limited discrete switching states listed in Table **??**, the DQN agent is more suitable for the case study. The DQN agent will observe the system states and receive the reward for the executed action, i.e., the voltage vector corresponding to the switching states x. By properly designing the reward function as incentives for training RL agents based on the cost function of FS-MPC, the RL agent will learn autonomously to find the optimal switching strategy for the converter to achieve desirable performance similar to the FS-MPC.

1) State and action sets: As shown in Fig. 4, the observed states include reference voltage (v_α^*, v_β^*), output voltage deviations ($\triangle v_{f\alpha}$, $\triangle v_{f\beta}$), the filter currents ($i_{f\alpha}$, $i_{f\beta}$) and the voltage vector x_{old} of previous sampling period. The action space consists of all the possible voltage vectors. Due to the repeated voltage vector for two switching states, the action space is defined as $x = [1 : 1 : 7]$, corresponding to the switching states.

2) Reward function design: It is noteworthy that different control objectives could be realized by fine-tuning the reward function, similar to the cost function design for FS-MPC. In this study, an exemplary control strategy for regulating AC voltage for VSC is presented. The reward function of RL should be finite for feasible and efficient training. Therefore, to mimic the optimal switching state selection strategy of FCS-MPC, the reward function is formulated based on the single-step horizon cost function, expressed as

$$r = -[(v_\alpha^* - v_{f\alpha})^2 + (v_\beta^* - v_{f\beta})^2] \tag{17}$$

The training process is shown in **Algorithm 2**.

Algorithm 2 DQN-based unsupervised imitation controller of FS-MPC for power converter

Input: $[v_\alpha^*, v_\beta^*, \triangle v_{f\alpha}, \triangle v_{f\beta}, i_{f\alpha}, i_{f\beta}, x_{old}]$
Output: Voltage vector number x

Initialize replay buffer $\mathcal{M}$ to capacity R
Initialize action-value function Q with random weights θ_Q
Initialize target action-value function $Q^{'}$ with weights $\theta_Q^{'} = \theta$
for episode = 1 to M **do**
  Receive initial observation at state s_1
  for iteration = 1 to T **do**
    With probability ϵ select a random action a_t otherwise select $a_t = \max_a Q_\pi(s_t, a|\theta)$
    Execute the action a_t, observe the reward r_t calculated with (17) and state s_{t+1}
    Store tuple (s_t, a_t, r_t, s_{t+1}) in the $\mathcal{M}$
    Randomly select a mini-batch e_j from $\mathcal{M}$
    Set $y_j = \begin{cases} r_j, \text{if episode terminates at step } j+1 \\ r_j + \gamma max Q_\pi^{'}(s_{j+1}, a^{'}|\theta_Q^{'}), \text{otherwise} \end{cases}$
    Perform gradient descent on (12) regarding the network parameter θ
    Every T_{target} steps reset $\theta_Q^{'}$=θ_Q
  end for
end for

IV. Experimental Results and Discussion

In this section, the proposed RL-aided weighting factor design method of FS-MPC and the MFRL-based controller for power converters are validated experimentally. As shown in Fig. 5, the two-level VSC consisting of three IMPERIX PEB 8024 H-bridge power modules is interfaced with a linear load via a LC filter, where $V_{dc} = 520\,V$, $L = 2.5\,mH$, $C = 30\mu F$, the load is 50 Ω, the reference voltage is 200 V with frequency 50 Hz, and the sampling time $T_s = 50\,\mu s$. An IMPERIX B-Box RCP control platform is used to implement the optimal weighting factor design and switching states selection policies from the well-trained RL agents.

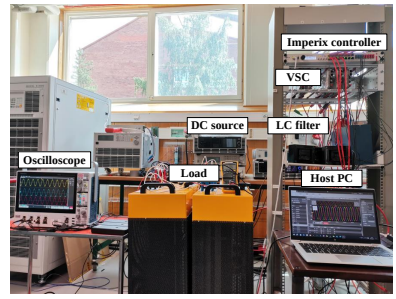


Fig. 5. Experimental setup of a two-level VSC system.

A. From simulation to practical implementation of online RL

The development of RL-based controller design and imitation controller can be split into two main steps. First, the RL agent is trained in MATLAB/SIMULINK with a model matching testbed parameters. As presented in Section III, by properly configuring the RL agents respectively for different case studies, the RL agents learn autonomously through interacting with the system. In particular, the DDPG agent tunes the optimal weighting factors of the FS-MPC for different control objectives, and the DQN agent emulates the optimal switching strategy of conventional FCS-MPC for VSC to achieve desired control performance without the prediction model. In this paper, all the simulations have been performed

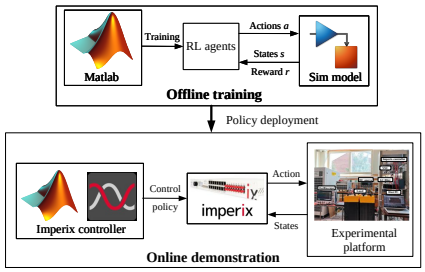


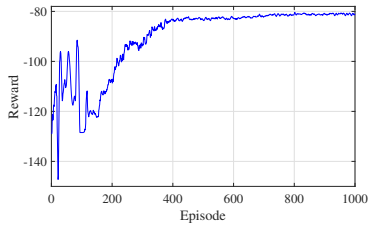
Fig. 6. Diagram of the RL-based methods development process, including an online demonstration with the practical experimental setup.

in Matlab/Simulink with a PC Intel(R) Core(TM) i5-9500 CPU at 3.00GHz and 32 GB RAM.

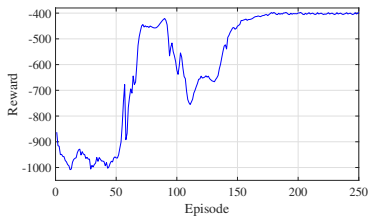
Afterward, the optimal policies from the well-trained online RL agents are transferred from simulation to practical setup using the IMPERIX platform. The training on the edge devices with the simulation model can stay the same with the real power converter system. In the practical implementation process, the optimal policy is deployed by configuring the corresponding measurements as the input for the RL agents, where the actions for achieving optimal control performance are output. The workflow of the above is presented in Fig. 6.

B. DDPG-based weighting factor design for FS-MPC

Different case studies are conducted to verify the proposed RL-based weighting factor design method. The training results are shown in the figures below.



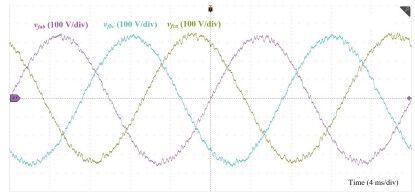
(a) Scenario A.



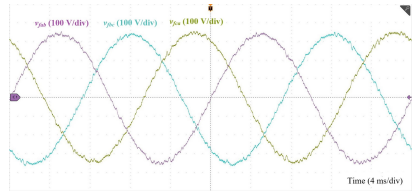
(b) Scenario B.

Fig. 7. Average accumulated reward during training process.

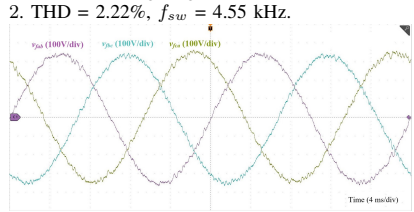
Scenario A: Minimization of THD. In this scenario, the reward in (15) is designed for training the agent, which encourages it to find the optimal weighting factors for minimization of the THD. The training process regarding the accumulated average reward over the training episodes is presented in Fig. 7a. It shows that the RL agent keeps exploring the action space during the initial sessions, resulting in a fluctuation of the accumulated average reward. During episodes 180-400, the agent learns to converge to the optimal weighting factors by balancing the exploration and exploitation process, and the reward increases. After around 400 training episodes, the agents exploit and converge to the optimal policy, where the weighting factors for minimizing the THD are obtained.



(a) No weighting factor. THD = 3.37%, f_{sw} = 4.95 kHz.



(b) RL-tuned weighting factors, $\lambda_d = 1.38, \lambda_{sw} = 2$. THD = 2.22%, f_{sw} = 4.55 kHz.



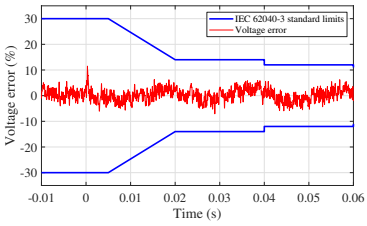
(c) RL-tuned weighting factors, $\lambda_d = 0.75, \lambda_{sw} = 7.04$. THD = 2.61%, f_{sw} = 3.62 kHz.

Fig. 8. Experimental line-line capacitor voltage waveforms under different conditions.

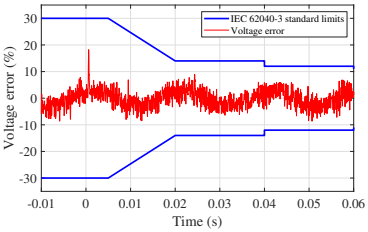
The weighting factor design policy from the well-trained agent for THD minimization is deployed and applied with the demonstration framework, where the obtained optimal weighting factors are $\lambda_d = 1.38, \lambda_{sw} = 2$. The line-to-line capacitor voltage without and with the optimal weighting factors is shown in Fig. 8a and Fig. 8b, respectively. The calculated THD and average switching frequency is 2.22%, 4.55 kHz with the RL-tuned weighting factors compared with the 3.37%, 4.95 kHz of the weightless counterpart.

Scenario B: Minimization of THD and f_{sw} simultaneously. To achieve the trade-off between the two control objectives in this case study, the RL agent learns to tune the weighting factors to minimize the THD with a low average switching frequency. The training process is presented in Fig. 7b, where the agent also experiences the exploration, learning, and exploitation stages to find the optimal weighting factors.

The optimal weighting factors design policy for FS-MPC



(a) 100% load step with $\lambda_d = 1.38, \lambda_{sw} = 2$.



(b) 100% load step with $\lambda_d = 0.75, \lambda_{sw} = 7.04$.

Fig. 9. Transient performance under 100% load step with the RL-tuned optimal weighting factors.

from the RL agent is deployed and validated experimentally, where the optimal weighting factors are $\lambda_d = 0.75, \lambda_{sw} = 7.04$. The results of the line-to-line capacitor voltage are shown in Fig. 8c, and the THD and average switching frequency regarding the RL-tuned weighting factors is 2.61%, 3.62 kHz.

The above results verify that the control performance is improved based on different desired objectives with the RL-tuned weighting factors. Furthermore, 100% linear load step tests for both cases are conducted to illustrate the marginal influence of weighting factors on the system's transient performance, which are presented in Fig. 9. The voltage error in both case studies converges quickly when there is a load step, resulting in quick convergence of the root mean square (RMS) of voltage deviation within the limit. Thus, the voltage deviation complies with the limits indicated in the standard IEC 62040-3.

C. DQN-based unsupervised imitation controller for VSC

The training process of the DQN agent learning the FS-MPC controller for VSC is shown in Fig. 10. At the initial training session, the agent explores the learning space without any prior system information by selecting the switching states randomly, resulting in a sharp drop in average accumulated reward due to the significant deviation of regulated capacitor voltage from the reference voltage. Afterward, the DQN agent starts to balance the exploration and exploitation process, where the designed reward function helps the agent learn to select the optimal switching states based on the input measurements and previous switching actions. In this way, the accumulated average reward increases and converges gradually, indicating the DQN agent exploits and finds the optimal switching state selection policy, emulating the FS-MPC.

The optimal control policy from the well-trained DQN is validated experimentally using the deployment framework, and the results are presented in Fig. 11. The difference in THD

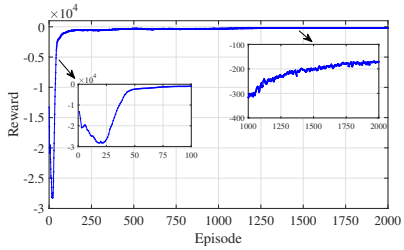


Fig. 10. Average accumulated reward during training process.

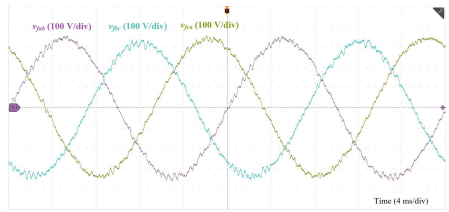
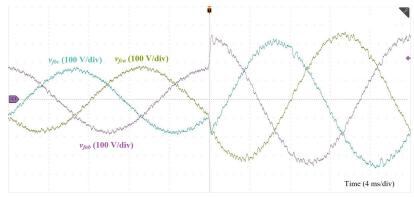


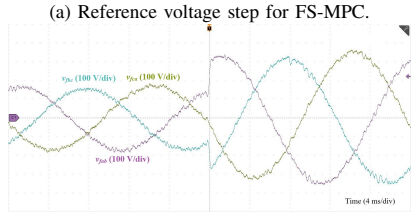
Fig. 11. MFRL-based unsupervised imitation controller. THD = 3.49%, f_{sw} = 4.58 kHz.

and f_{sw} between the conventional FS-MPC without weighting factor and MFRL-based imitation controller are 0.12% and 0.37 kHz, respectively, confirming an effective learning of the FS-MPC control for VSC. More importantly, the DQN-based controller does not require any prior knowledge of the system and is trained in an unsupervised way without explicit labels, compared with conventional FS-MPC [8] and supervised ANN-based imitator [9] for power converter control.

To validate the DQN-based controller's transient performance, a reference voltage step is applied. It shows that the RL-based controller achieves comparable performance to the conventional FS-MPC.



(a) Reference voltage step for FS-MPC.



(b) Reference voltage step for RL-based controller.

Fig. 12. Output voltage when applying reference voltage step (100V $\rightarrow$ 200 V).

D. Computation complexity

The computational complexity is crucial for implementing the proposed approaches in practical applications. The MPC

controller has cubic complexity considering the need for matrix multiplication and inversion, especially for multi-cell or multi-level converters within the control horizon where the number of switching states increases. In comparison, the vast majority of the computing cost of the proposed RL-aided weighting factor design and MFRL-based imitation controller comes from the offline training stage at edge devices, and the time consumption is not critical. Once trained successfully, the RL-based methods involve only computational light matrix manipulation within the neural networks, which enjoy linear computational complexity.

More importantly, with a simple deep neural network structure, the RL can map the nonlinear relationship between the input and output for most problems without constructing more hidden layers, thus not inducing more computational burden. There is potential for the proposed RL methods to solve more complex problems, e.g., more weighting factors and multi-level or multi-cell converter control. Therefore, despite the time-consuming training process, the practical implementation of the proposed RL-based control strategy is quite tractable.

V. Conclusions

This paper provides a guideline for applying the RL to the controller design and unsupervised imitation controller for power converters, starting from the problem formulation and RL agent type selection to the reward design and policy deployment in a practical converter testbed. In particular, a DDPG-based automatic weighting factor tuning method and a DQN-based imitation controller emulating the optimal switching states selection of FS-MPC are proposed. The proposed methods are model-free and thus require no prior knowledge of the system, improving or inheriting the good performance of conventional FS-MPC. In addition, with the proposed deployment framework for transferring the online RL agents from simulation to practical implementation, the proposed methods are validated experimentally. The results show that the RL-aided weighting factor design for FS-MPC achieves the desired control performance. In addition, the results also confirm that the RL-based imitation controller achieves a comparable control performance to conventional FS-MPC for VSC.

Findings in this paper open interesting topics for the future development of the proposed method. The RL-based automatic controller design method is also applicable to tuning the coefficients of other controllers. In addition, due to the inherent nonlinear mapping capability of employed deep neural networks in RL, the proposed MFRL-based imitation controller may still work well without adding hidden layers. Future research could explore the possibilities of the imitation controller for algorithms for multi-level converters or modular multi-level topologies, where the computational burden increases exponentially due to the increasing number of switching states.

References

[1] T. Dragičević, X. Lu, J. C. Vasquez, and J. M. Guerrero, "Dc microgrids—part i: A review of control strategies and stabilization techniques," *IEEE Transactions on power electronics*, vol. 31, no. 7, pp. 4876–4891, 2015.

[2] T. Dragičević, S. Vazquez, and P. Wheeler, "Advanced control methods for power converters in dg systems and microgrids," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 7, pp. 5847–5862, 2020.

[3] J. Rodriguez and P. Cortes, *Predictive control of power converters and electrical drives*. John Wiley & Sons, 2012.

[4] R. Vargas, P. Cortés, U. Ammann, J. Rodríguez, and J. Pontt, "Predictive control of a three-phase neutral-point-clamped inverter," *IEEE Transactions on Industrial Electronics*, vol. 54, no. 5, pp. 2697–2705, 2007.

[5] P. Karamanakos and T. Geyer, "Guidelines for the design of finite control set model predictive controllers," *IEEE Transactions on Power Electronics*, vol. 35, no. 7, pp. 7434–7450, 2019.

[6] T. Dragičević and M. Novak, "Weighting factor design in model predictive control of power electronic converters: An artificial neural network approach," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 11, pp. 8870–8880, 2018.

[7] M. S. O. Yeganeh, A. Oshnoei, N. Mijatovic, T. Dragicevic, and F. Blaabjerg, "Intelligent secondary control of islanded ac microgrids: A brain emotional learning-based approach," *IEEE Transactions on Industrial Electronics*, 2022.

[8] P. Cortés, G. Ortiz, J. I. Yuz, J. Rodríguez, S. Vazquez, and L. G. Franquelo, "Model predictive control of an inverter with output lc filter for ups applications," *IEEE Transactions on industrial electronics*, vol. 56, no. 6, pp. 1875–1883, 2009.

[9] M. Novak and T. Dragicevic, "Supervised imitation learning of finite-set model predictive control systems for power electronics," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 2, pp. 1717–1723, 2020.

[10] E. Zafra, *et al.*, "Computationally efficient sphere decoding algorithm based on artificial neural networks for long-horizon fcs-mpc," *IEEE Transactions on Industrial Electronics*, 2023.

[11] S. Wang, T. Dragicevic, G. F. Gontijo, S. K. Chaudhary, and R. Teodorescu, "Machine learning emulation of model predictive control for modular multilevel converters," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 11, pp. 11 628–11 634, 2020.

[12] D. Wang, *et al.*, "Model predictive control using artificial neural network for power converters," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 4, pp. 3689–3699, 2021.

[13] W. Wu, L. Qiu, J. Rodriguez, X. Liu, J. Ma, and Y. Fang, "Data-driven finite control-set model predictive control for modular multilevel converter," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, 2022.

[14] W. Wu, *et al.*, "Data-driven iterative learning predictive control for power converters," *IEEE Transactions on Power Electronics*, vol. 37, no. 12, pp. 14 028–14 033, 2022.

[15] S. S. Khorramabadi and A. Bakhshai, "Intelligent control of grid-connected microgrids: An adaptive critic-based approach," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 3, no. 2, pp. 493–504, 2014.

[16] M. Gheisarnejad and M. H. Khooban, "Iot-based dc/dc deep learning power converter control: Real-time implementation," *IEEE Transactions on Power Electronics*, vol. 35, no. 12, pp. 13 621–13 630, 2020.

[17] C. Cui, N. Yan, B. Huangfu, T. Yang, and C. Zhang, "Voltage regulation of dc-dc buck converters feeding cpls via deep reinforcement learning," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, no. 3, pp. 1777–1781, 2021.

[18] Y. Tang, *et al.*, "Artificial intelligence-aided minimum reactive power control for the dab converter based on harmonic analysis method," *IEEE Transactions on Power Electronics*, vol. 36, no. 9, pp. 9704–9710, 2021.

[19] Y. Tang, *et al.*, "Deep reinforcement learning aided variable-frequency triple phase shift control for dual active bridge converter," *IEEE Transactions on Industrial Electronics*, 2022.

[20] Y. Zeng, *et al.*, "Autonomous input voltage sharing control and triple phase shift modulation method for isop-dab converter in dc microgrid: A multiagent deep reinforcement learning-based method," *IEEE Transactions on Power Electronics*, vol. 38, no. 3, pp. 2985–3000, 2022.

[21] X. Liu, L. Qiu, Y. Fang, and J. Rodríguez, "Reinforcement learning-based event-triggered fcs-mpc for power converters," *IEEE Transactions on Industrial Electronics*, 2023.

[22] T. Dragičević, "Model predictive control of power converters for robust and fast operation of ac microgrids," *IEEE Transactions on Power Electronics*, vol. 33, no. 7, pp. 6304–6317, 2017.

[23] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, pp. 279–292, 1992.

[24] V. Mnih, *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[25] T. P. Lillicrap, *et al.*, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.
[26] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *International conference on machine learning*. Pmlr, 2014, pp. 387–395.

Yihao Wan (S'21) received the B.S. degree in electrical engineering from the Wuhan University of Technology, Wuhan, China, in 2017 and the M.S. degree in electrical engineering from Chongqing University, Chongqing, China, in 2020. He is currently pursuing the Ph.D. degree in electrical engineering with Technical University of Denmark. His research interests include advanced control, cyber security, and application of artificial intelligence, and machine learning in power electronic systems.

Qianwen Xu (S'14-M'18)) received the B.Sc. degree from Tianjin University, Tianjin, China, in 2014, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2018, both in electrical engineering.

From 2018 to 2020, she was a Postdoc Research Fellow with Aalborg University, Aalborg, Denmark, a Visiting Researcher with Imperial College London, London, U.K., and a Wallenberg-NTU Presidential Postdoc Fellow with Nanyang Technological University, Singapore. She is currently an Assistant Professor with the Department of Electric Power and Energy Systems, KTH Royal Institute of Technology, Stockholm, Sweden. Her research interests include advanced control, optimization, and AI application for microgrid and smart grid.

Dr. Xu is the Vice Chair for the IEEE Power and Energy Society and Power Electronics Society, Sweden Chapter, and an Associate Editor for the IEEE TRANSACTIONS ON SMART GRID, IEEE TRANSACTIONS ON TRANSPORTATION ELECTRIFICATION, and IEEE Journal of Emerging and Selected Topics in Power Electronics. She was a recipient of Humboldt Research Fellowship, Excellent Doctorate Research Work, Best Paper Award in IEEE PEDG 2020, Nordic Energy Award 2022, etc.

Tomislav Dragičević (S'09-M'13-SM'17) received the M.Sc. and the industrial Ph.D. degrees in Electrical Engineering from the Faculty of Electrical Engineering, University of Zagreb, Croatia, in 2009 and 2013, respectively. From 2013 until 2016 he has been a Postdoctoral researcher at Aalborg University, Denmark. From 2016 until 2020 he was an Associate Professor at Aalborg University, Denmark. From 2020 he is a Professor at the Technical University of Denmark. He made a guest professor stay at Nottingham University, UK during spring/summer of 2018. His research interest is application of advanced control, optimization and artificial intelligence inspired techniques to provide innovative and effective solutions to emerging challenges in design, control and cyber-security of power electronics intensive electrical distributions systems and microgrids. He has authored and co-authored more than 250 technical publications (more than 120 of them are published in international journals, mostly in IEEE), 8 book chapters and a book in the field.

He serves as an Associate Editor in the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, in IEEE TRANSACTIONS ON POWER ELECTRONICS, in IEEE Emerging and Selected Topics in Power Electronics and in IEEE Industrial Electronics Magazine. Dr. Dragičević is a recipient of the Končar prize for the best industrial PhD thesis in Croatia, a Robert Mayer Energy Conservation award, and he is a winner of an Alexander von Humboldt fellowship for experienced researchers.

[J2] Safety-Enhanced Self-Learning for Optimal Power Converter Control

Authors:
Yihao Wan, Qianwen Xu, Tomislav Dragicevic

Submitted to:
IEEE Transactions on Industrial Electronics

Status:
Under review

Safety-Enhanced Self-Learning for Optimal Power Converter Control

Yihao Wan, *Student Member, IEEE,* Qianwen Xu, *Member, IEEE,* and Tomislav Dragičević, *Senior Member, IEEE*

***Abstract*—Data-driven learning-based control methods such as reinforcement learning (RL) have become increasingly popular with recent proliferation of the machine learning paradigm. These methods address the parameter sensitiveness and unmodeled dynamics in model-based controllers, such as finite control-set model predictive control. RL agents are typically utilized in simulation environments, where they are allowed to explore multiple "unsafe" actions during the learning process. However, this type of learning is not applicable to online self-learning of controllers in physical power converters, because unsafe actions would damage them. To address this, this letter proposes a safe online RL-based control framework to autonomously find the optimal switching strategy for the power converters, while ensuring system safety during the entire self-learning process. The proposed safe online RL-based control is validated in a practical testbed on a two-level voltage source converter system, and the results confirm the effectiveness of the proposed method.**

***Index Terms*—Finite control-set model predictive control (FCS-MPC), learning-based control, power converters, reinforcement learning (RL), safety policy.**

I. Introduction

REINFORCEMENT learning (RL) has gained increasing attention for applications in the power electronics field due to its model-free and self-learning characteristics. Compared with conventional cascaded linear control strategies that suffer from slow dynamics and finite control-set model predictive control (FCS-MPC) that has fast response and straightforward design but rely heavily on accurate parameters and the established model [1], the learning-based controllers can simultaneously mitigate the dependence on the precise system model and achieve fast and accurate control performance.

Different RL-based controllers have been proposed for converters. In [2], an RL-based controller is incorporated into the sliding mode observer to stabilize the buck-boost converter and improve the voltage regulation under constant power load. In addition, different RL algorithms are used to optimize the shift angle of the triple phase shift modulation in dual active bridge converters to minimize reactive power [3], power losses [4], and current stress [5].

However, a pivotal challenge in RL lies in striking the balance between exploration and exploitation, often leading to protracted convergence time for RL agents in finding the optimal policy. Moreover, the physical limitations are not ensured during training sessions, especially in the initial stages involving random exploration, thereby limiting its practical application within the domain of online self-learning in power electronics. Recent advancements have embraced the safe RL paradigm to address these issues [6]–[8]. Existing safety frameworks for RL predominantly target optimization and power systems, accomplished either through direct action evaluation against constraints [6], training a safety model correlating actions with constraint functions [7], or substituting unsafe actions with those derived from physics-related functions [8], etc. Nevertheless, there appear to be research gaps concerning safe learning-based control for power converters.

Motivated by these, this letter proposes a novel safety-enhanced self-learning approach for optimal power converter control. The proposed safe learning framework combines model-based and data-driven learning-based approaches by incorporating a computationally efficient MPC-based safety policy into the learning-based controller. The proposed safe learning-based controller can achieve comparable control performance to the conventional FCS-MPC, guarantee system safety throughout the learning process, and significantly improve learning efficiency.

Yihao Wan and Tomislav Dragičević are with the Department of Wind and Energy Systems, Technical University of Denmark, Copenhagen, Denmark (e-mails: wanyh@dtu.dk, tomdr@dtu.dk).

Qianwen Xu is with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden (email: qianwenx@kth.se).

II. Unsupervised Online Safe Reinforcement Learning Based Controller

The studied system is shown in Fig. 1, where a two-level voltage source converter (VSC) interfaces the load and DC source. In particular, eight possible switching combinations align with different voltage vectors. The conventional FCS-MPC predicts the converter behaviors for those switching combinations based on a model discretized with a sampling time T_s, shown below, where a cost function is used to select the optimal switching states [9].

$$\begin{bmatrix}\mathbf{i}_f(k+1)\\ \mathbf{v}_f(k+1)\end{bmatrix}=\begin{bmatrix}-\frac{R_f T_s}{L_f} & -\frac{T_s}{L_f}\\ \frac{T_s}{C_f} & 0\end{bmatrix}\begin{bmatrix}\mathbf{i}_f(k)\\ \mathbf{v}_f(k)\end{bmatrix}+\begin{bmatrix}\frac{T_s}{L_f} & 0\\ 0 & -\frac{T_s}{C_f}\end{bmatrix}\begin{bmatrix}\mathbf{v}_i(k)\\ \mathbf{i}_o(k)\end{bmatrix} \tag{1}$$

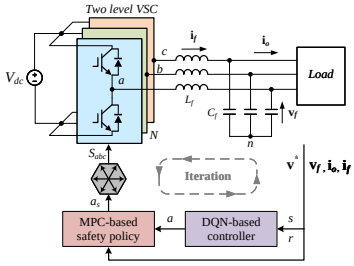


Fig. 1. Schematic of the safe learning-based control for VSC.

A. Problem formulation for the RL

RL paradigm for learning optimal converter control is formulated as a Markov Decision Process (MDP). Due to the limited number of switching combinations, the DQN algorithm with a discrete action space is employed. The proposed RL-based controller aims to find the optimal switching state selection policy and regulate the VSC by merely interacting with the converter system. As shown in Fig. 1, at each discrete

time step t, the DQN agent receives the measurements s_t, and takes an action a_t according to the policy π. At time step t+1, the system shifts to a new state s_{t+1}, and the agent receives a reward r_t for the transition. The agent takes actions according to the maximum Q value with ϵ-greedy strategy, described as

$$a_t = \begin{cases} \arg\max_a Q_\pi(s_t, a), & \text{with probability } \epsilon; \\ \text{a random action}, & \text{otherwise.} \end{cases} \quad (2)$$

In DQN, a deep neural network is employed to evaluate the Q value for each state-action pair. During the training session, transition sequences (s_t, a_t, r_t, s_{t+1}) are stored in a replay memory, where a minibatch of the tuples is randomly selected to train the neural network via stochastic gradient descent. In addition, to enhance the stability and convergence of training the network with parameter θ, a target network Q'_π synchronizing its separate network parameter θ' with the θ for fixed time step is introduced. The Q value is updated as

$$y_j = \begin{cases} r_j, \text{if episode terminates at step } j+1; \\ r_j + \gamma \max Q'_\pi(s_{j+1}, a' | \theta'), \text{otherwise.} \end{cases} \quad (3)$$

The loss function for updating the Q-network parameter is

$$L(\theta) = \mathbb{E}\left[(y_t - Q_\pi(s_t, a_t|\theta))^2\right] \quad (4)$$

B. RL-based controller design

Using a reward formulated to incentivize the RL agent, the proposed RL-based controller learns autonomously to select the optimal switching state based on the input measurements, emulating the switching state selection strategy of FCS-MPC.

1) State and action sets: To achieve similar control performance to the FCS-MPC, the input states include reference voltage (v^*_α, v^*_β), capacitor voltage deviations ($\triangle v_{f\alpha}$, $\triangle v_{f\beta}$), filter currents ($i_{f\alpha}$, $i_{f\beta}$), and the previous action, i.e., the voltage vector $x(k-1)$ from the previous sampling period. The action space consists of the number of voltage vectors as $[1:1:7]$, matching the switching combinations, respectively.

2) Reward function design: The reward function should be finite for feasible and efficient training. To emulate the optimal switching states selection strategy in FCS-MPC and regulate the AC voltage, the reward function is thus formulated as

$$r = -[(v^*_\alpha - v_{f\alpha})^2 + (v^*_\beta - v_{f\beta})^2] \quad (5)$$

C. Safety policy for the online RL

Practically, the converter-side current should always be limited within an allowable range to ensure hardware safety. Therefore, the actions taken by the DQN agent should be considered for the sake of the system's safety. To achieve this, a safety framework based on a computational light single-step prediction illustrated in Fig. 1 is proposed.

Specifically, at each training step, the action taken by the RL agent from (2) is input to the MPC-based safety block, which performs one-step prediction for the current $i_f(k+1)$ based on equations in (1). It will also evaluate whether the corresponding switching states would lead to overcurrent as

$$||i^a_{f\alpha\beta}(k+1)||_2 \leq i_{max} \quad (6)$$

The unsafe actions causing overcurrent will be abandoned. Instead, safe switching states from the MPC-based safety block will be used to ensure the safe operation of the system at all times. Notably, the safety framework is not to find optimal actions but to guide the RL agent to take safe actions. Meanwhile, the agent will gradually converge within the safe learning space, bypassing the safety block. In this way, the safety framework also narrows the learning spaces by avoiding unsafe actions, improving the learning efficiency. In short, the proposed safety-enhanced self-learning optimal controller combines a computationally light single-step prediction to exclude unsafe actions and narrow the learning space with an RL-based learning framework.

III. Experimental Results and Discussion

This section presents experimental validations for the proposed safe online RL-based control framework. The experimental setup in Fig. 2 aligns with the configuration in Fig. 1. The DC link voltage is 520 V, the LC filter is 2.5 mH and 30 μF, the load is 50 Ω, the reference voltage is 200 V with frequency 50 Hz, the sampling time $T_s = 20\,\mu s$, and i_{max} is set as 20 A. In particular, the DQN agent is trained via an edge device, where a simulation model matching the testbed parameters is built in MATLAB/SIMULINK. The training in simulation on the edge device can stay the same as the real converter system, where the agent could also be trained online safely with the proposed method. Afterward, the online safe RL-based controller for VSC is transferred from edge devices to the practical setup using the IMPERIX control platform.

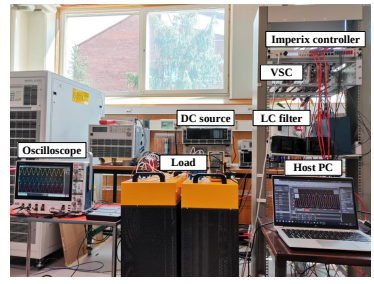
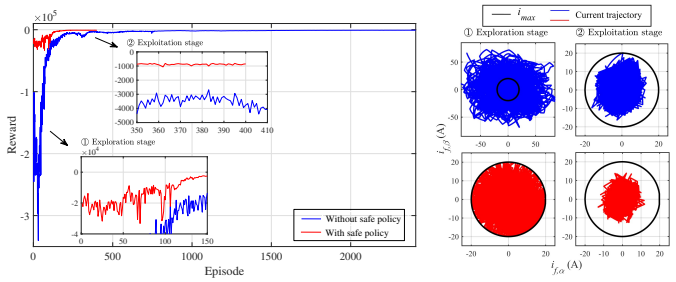


Fig. 2. Experimental setup of a two-level VSC system.

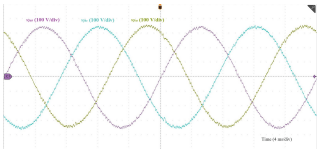
The training process of the DQN-based controller is shown in Fig. 3. Starting from exploration, it can be observed from Fig. 3a that the agent without the safe policy takes actions randomly, resulting in significant deviations of the output voltage from the reference voltage, causing a sharp drop in the average accumulated reward. In addition, the converter-side current trajectory shown in Fig. 3b exceeds the maximum current, which would damage the converter. As the training episodes progress, the RL agent learns and converges within the safe learning region. Conversely, the proposed safe RL agent explores the safe learning space to constrain the current within the physical limit throughout the training process. The accumulated reward is much higher, and the agent quickly converges to the optimal policy. In this way, the accumulated average reward for the two training methods converges to the same level, and the RL agent learns the optimal switching strategy for the VSC, emulating the conventional FCS-MPC.

The safety-enhanced self-learning framework for optimal converter control is transferred from the edge device to a practical converter to experimentally validate the obtained optimal

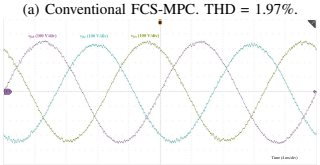


(a) Average accumulated reward (b) Converter-side current trajectory

Fig. 3. Training process of the proposed safety-enhanced self-learning optimal controller.



(a) Conventional FCS-MPC. THD = 1.97%.



(b) Safe DQN-based controller. THD = 2.19%.

Fig. 4. Performance of different controllers.

control policy with the deployment framework. The results shown in Fig. 4 verify that the safety-enhanced self-learning optimal controller achieves desirable control performance with a THD of 2.19%, comparable to the THD of 1.97% for FCS-MPC with precise system parameters and established models.

Moreover, sensitivity analysis regarding the parameter variations is performed. The variations in the model parameters would deteriorate the performance of FCS-MPC. In contrast, the proposed safe RL learning-based controller is model-free, and the proposed safe policy framework only functions when unsafe actions occur. Once converging within the safe learning region, as presented in Fig. 3b, the RL agent can find the optimal control strategy safely. A variation of inductance ($\triangle L_f$) and capacitance ($\triangle C_f$) within $\pm$ 30% the nominal value is implemented to investigate the robustness of the safety-enhanced self-learning optimal control framework.

If the safety framework underestimates the inductance value ($\triangle L_f > 0$), the safe exploration region is more tightly guaranteed as it reduces the current filter peaks and the reference current. On the other hand, if the safety framework overestimates the inductance value ($\triangle L_f < 0$), the filter current peaks may exceed the maximum value while still within the limits by properly leaving a margin between the predefined maximum current and physical limited current. As the training advances, the RL agent steadily converges within the designated safe region, where the current trajectories also become centered on the safe region. In either case, the agent still finds the optimal control policy. Conversely, the capacitor is indirectly controlled by the inverter voltage due to the cross-coupling effect between the inductor and capacitor. Thus, the capacitance uncertainty barely influences the safe exploration of the RL agent. In summary, despite the parameter uncertainty, the proposed RL-based controller can still find the optimal switching policy for the converter.

IV. Conclusions

This paper proposes a safety-enhanced self-learning approach for optimal power converter control by introducing computational light single-step predictive control to the learning-based control framework. The proposed safe policy can guarantee the system's safety and reduce unnecessary exploration regions, thus also improving training efficiency. A deployment framework for transferring the online safe RL-based controller from simulation on edge devices to practical implementation in an experimental setup is also demonstrated. The experimental results confirm that the RL-based controller achieves a satisfactory control performance for VSC comparable to the conventional FCS-MPC. Future work will include self-learning control in real-time directly on the edge devices of a practical experimental setup.

References

[1] T. Dragičević, S. Vazquez, and P. Wheeler, "Advanced control methods for power converters in dg systems and microgrids," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 7, pp. 5847–5862, 2020.

[2] 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 Transactions on Industrial Electronics*, vol. 68, no. 7, pp. 6379–6384, 2020.

[3] Y. Tang, *et al.*, "Artificial intelligence-aided minimum reactive power control for the dab converter based on harmonic analysis method," *IEEE Transactions on Power Electronics*, vol. 36, no. 9, pp. 9704–9710, 2021.

[4] Y. Tang, *et al.*, "Deep reinforcement learning aided variable-frequency triple phase shift control for dual active bridge converter," *IEEE Transactions on Industrial Electronics*, 2022.

[5] Y. Zeng, *et al.*, "Autonomous input voltage sharing control and triple phase shift modulation method for isop-dab converter in dc microgrid: A multiagent deep reinforcement learning-based method," *IEEE Transactions on Power Electronics*, vol. 38, no. 3, pp. 2985–3000, 2022.

[6] D. Weber, M. Schenke, and O. Wallscheid, "Safe reinforcement learning-based control in power electronic systems," in *2023 International Conference on Future Energy Solutions*. IEEE, 2023, pp. 1–6.

[7] Y. Xia, *et al.*, "A safe policy learning-based method for decentralized and economic frequency control in isolated networked-microgrid systems," *IEEE Transactions on Sustainable Energy*, vol. 13, no. 4, pp. 1982–1993, 2022.

[8] P. Chen, S. Liu, X. Wang, and I. Kamwa, "Physics-shielded multi-agent deep reinforcement learning for safe active voltage control with photovoltaic/battery energy storage systems," *IEEE Transactions on Smart Grid*, 2022.

[9] T. Dragičević, "Model predictive control of power converters for robust and fast operation of ac microgrids," *IEEE Transactions on Power Electronics*, vol. 33, no. 7, pp. 6304–6317, 2017.

[J3] Data-driven Cyber-attack Detection of Intelligent Attacks in Islanded DC Microgrids

Authors:
Yihao Wan, Tomislav Dragicevic

Submitted to:
IEEE Transactions on Industrial Electronics

Status:
Published.
Digital Object Identifier: 10.1109/TIE.2022.3176301

Data-driven Cyber-attack Detection of Intelligent Attacks in Islanded DC Microgrids

Wan, Yihao; Dragicevic, Tomislav

Published in:
IEEE Transactions on Industrial Electronics

Link to article, DOI:
10.1109/TIE.2022.3176301

Publication date:
2022

Document Version
Peer reviewed version

Data-driven Cyber-attack Detection of Intelligent Attacks in Islanded DC Microgrids

Yihao Wan, *Student Member, IEEE*, Tomislav Dragičević, *Senior Member, IEEE*

***Abstract*—In this paper, a data-driven cyber-attack detection method for islanded DC microgrids is proposed. Data is collected by monitoring the behavior of an intelligent attacker who is able to bypass conventional cyber-attack detection algorithms and disrupt the operation of the system. Reinforcement learning (RL) algorithm emulates the actions of such intelligent attacker, who exploits the vulnerability of index-based cyber-attack detection methods, such as discordant detection algorithm. The data is then used to train a neural network based detector that complements the conventional method with additional capability to detect a larger set of possible attacks. Through experiments, the effectiveness of the proposed method is validated.**

***Index Terms*—DC microgrids, data-driven cyber-attack detection, reinforcement learning, discordant detection algorithm, neural network based detector.**

I. Introduction

DC microgrids facilitate smart grid applications in an efficient and cost-effective way due to the natural matching with different distributed generation resources [1]. For the control of microgrid, distributed control has become popular as it offers better scalability, reliability, and efficiency compared with centralized control, which also suffers from the single point of failure [2]. As the distributed control rests on the communication network, it makes the DC microgrids cyber-physical systems, which are vulnerable to cyber-attacks. Among different types of cyber-attacks [3], [4], the most common are false data injection attacks (FDIAs). FDIAs alter the system states by injecting data into the sensors or communication links and disrupt the operation of the system [5]. Such attacks can destabilize the DC microgrids if not detected and mitigated properly.

To mitigate the vulnerability of DC microgrids, different cyber-attack detection methods are proposed, which could be broadly classified into model-based and model-free methods [6]. The model-based detection methods rely on the accuracy of the system model, which is challenging to implement in practical applications due to its unavoidable mismatch with the complex real-world power electronic systems. On the other hand, the model-free methods utilize the measurements without prior knowledge of the system. In [7], based on the estimated outputs of the system using an artificial neural network, the stealthy FDIA detection method is proposed. However, since the data used to train the given neural network based detector is generated from the healthy model of the microgrid, it does not explicitly incorporate knowledge about different types of attacks. As a result, the method is not able to detect some types of attacks, such as destabilization attacks. Signal temporal logic (STL) is proposed to detect FDIAs by monitoring the output voltage and current with defined specifications [8] while the performance under stealthy attack is not verified. A discordant detection algorithm is proposed to detect both destabilization and stealthy attacks by calculating the discordant element (DE) term [9]. In addition, similar methods are proposed in [10]–[12] utilizing different indices for attack detection. However, these index-based detection methods will still fail when intelligent attackers introduce novel attack patterns and a wider range of coordinated or uncoordinated attacks by injecting false signals into the sensors, communication links of multiple nodes, or concurrently both of them [11]. Such intelligent and deceptive behaviors can be emulated via reinforcement learning.

Of different forms of machine learning, reinforcement learning is the learning paradigm closest to the human learning process as it can learn through experience by exploring and exploiting the dynamic and unknown environment [13]. RL can model an intelligent agent to take sequential optimal actions without or with limited knowledge of the environment, which makes it particularly adaptable and feasible in real-time systems. Therefore, reinforcement learning demonstrates excellent suitability for application in cyber-security areas, where cyber-attacks become increasingly sophisticated [14]–[16]. Based on this, the reinforcement learning agent is able to serve as an intelligent attacker, who exploits the vulnerability of DC microgrids system protected with conventional discordant detection scheme by generating novel attack patterns. To generate sophisticated cyber-attacks in DC microgrids, in reinforcement learning algorithm, deep neural networks that represent the attacker are trained over many system rollouts and autonomously discover the deficiency of the index-based cyber-attack detection method in DC microgrids. The deep neural network afterward interacts with the real system, injects false signals into multiple nodes coordinately, nullifies the indices that are used for detection, such as discordant elements in [9], and crafts stealthy attacks that can bypass conventional cyber-attack detection methods.

To solve the aforementioned issue, in this letter, a data-driven cyber-attack detection method for DC microgrids is proposed. Particularly, the RL-based intelligent cyber-attacker can uncover the deficiencies of the DE-based detection algorithm but could also expose other index-based detection methods if trained in such a way. In view of this, the proposed data-driven cyber-attack detection method is to complement

Yihao Wan and Tomislav Dragičević are with the Department of Electrical Engineering, Technical University of Denmark, Copenhagen, Denmark (e-mails: wanyh@elektro.dtu.dk, tomdr@elektro.dtu.dk).

the conventional index-based detection methods by detecting the attacks and the attacked nodes under the RL-based cyber-attacks.

Unlike model-based cyber-attack detection methods, in the proposed data-driven cyber-attack detection method, only the current and voltage measurements are used. Moreover, the historical operation data is directly collected from a DC microgrid experimental testbed, which allows higher training precision compared to the previous data-based method that used simulation models for this purpose. Finally, to the best of the authors' knowledge, this paper is the first attempt to use the data from the system exposed to intelligent cyber-attacks to train the attack detector, while existing methods use data from normally operated microgrids. In real-time applications, the RL-based attacker will be implemented to generate sophisticated attacks to bypass the DE-based detection method, such a data-driven attack detector is then utilized to complement the DE-based detection method for detecting the attacks and identifying the attacked nodes in DC microgrids. The neural network based cyber-attack detectors are implemented in an experimental setup to verify the performance of the proposed method.

II. RL-Based FDIA on Cooperative Control Based DC Microgrids

A. Cooperative Control of DC Microgrids with Discordant Detection Algorithm

The configuration of the DC microgrid is shown in Fig. 1, where *n* DC sources connected through dc-dc converters are linked via communication networks and form the cyber-physical network. The main objectives of the control are voltage regulation and proportional current sharing. The output voltage of each node is regulated by the primary control layer. The current sharing causes voltage error, which is compensated by the distributed secondary control [17].

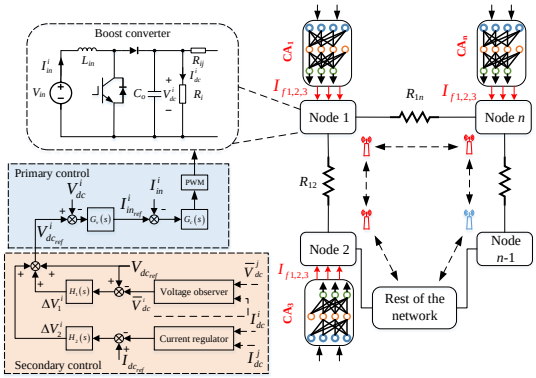


Fig. 1. Configuration of DC microgrid under multi-agent RL-based cyber-attack.

The distributed control rests on the communications between neighboring nodes. To achieve the control objectives, in the secondary control layer, two voltage correction terms $\triangle V_1^i$ and $\triangle V_2^i$ are calculated for the reference voltage of the primary control layer to regulate the output voltage of each node [17]. The reference voltage can thus be expressed as

$$V_{dc_{ref}}^i = V_{dc_{ref}} + \triangle V_1^i + \triangle V_2^i \tag{1}$$

where $V_{dc_{ref}}$ is the global reference voltage for all the nodes and $I_{dc_{ref}}$ is set to 0 to achieve current sharing.

Based on the distributed control in a fully-connected cyber network in DC microgrids, the control objectives will converge as follows [17]

$$\lim_{t\to+\infty} \bar{V}_{dc}^i(t) = V_{dc_{ref}}, \lim_{t\to+\infty} \bar{I}_{dc}^i = 0 \tag{2}$$

where $\bar{V}_{dc}^i$ and $\bar{I}_{dc}^i$ are the estimated average voltage and the normalized current regulation input for node i respectively.

False signals could be injected into the sensor measurements, the distributed signals from the neighboring nodes, or both of them. The attacked signals can be expressed as

$$\boldsymbol{I_A} = \begin{cases} I_a^i = I_{dc}^i + k_i I_{fi}, & \text{sensor attack} \\ I_a^{ij} = I_{dc}^i + k_{ij} I_{fj}, & \text{link attack} \end{cases} \tag{3}$$

where $\boldsymbol{I_A}$ denotes the attacked vector at multiple nodes, I_a^i is the output current from the node *i*, and I_{dc}^i is the real measured value from the sensors, I_a^{ij} is the distributed current received from neighboring node *j* for node *i*, I_{fi} and I_{fj} are the injected false signals from the attacker, k_i and k_{ij} are the corresponding coefficients for the attack signals, of which $k_i = 1$ or $k_{ij} = 1$ denotes the presence of an attack in the corresponding signals, or 0 otherwise.

For different uncoordinated attacks, the control objectives will not converge as in equation (2). These events could destabilize the DC microgrids depending on the attack intensity. On the other hand, the attack that results in the control inputs converging as per equation (2) is considered as coordinated attack, which is generally formed by attacking sensors and communication links concurrently. Among different detection schemes, discordant detection algorithm, based on the synchrony between the neighboring reference current terms of a node, has shown excellent performance in detecting both types of cyber-attacks [9]. In the distributed control, the control objectives will converge as per (2), and the input current I_{in}^i will always converge to zero under no attacks, thus the input current reference $I_{in_{ref}}^i$ quantities will also achieve consensus among themselves. However, when the attack occurs, for the attacked node, the secondary layer will maloperate due to the compromised current information, which causes different operation of the outer voltage loop, thus the current reference for the compromised node goes discordant with the remaining healthy nodes. The index for attack detection is calculated as

$$DE_i = l_i[\sum_{j\epsilon M_i}(I_{in_{ref}}^j - I_{in_{ref}}^i)][\sum_{j\epsilon M_i}(I_{in_{ref}}^j + I_{in_{ref}}^i)] \tag{4}$$

$$DE_i = \begin{cases} < DE_{min}, & if\ k_i\ \&\ k_{ij} = 0 \\ > DE_{min}, & if\ k_i \| k_{ij} \neq 0 \end{cases} \tag{5}$$

where DE_i denotes the discordant term of node *i*, $I_{in_{ref}}^i$ and $I_{in_{ref}}^j$ are the input reference current from the outer voltage control loop for node *i* and the neighboring node *j* respectively,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2022.3176301, IEEE Transactions on Industrial Electronics

M_i denotes the set of neighbors of node i, l_i is a positive coefficient to increase/decrease the value of DE_i. According to [9], based on (5), the attacks on the ith node can be determined by comparing the positive value of DE_i terms with the minimum threshold, which is obtained in normal operation.

B. Multi-agent Reinforcement Learning Based FDIA

To compromise the conventional detection method, multi-agent reinforcement learning is utilized. The autonomous attack generation process can be modeled as a Markov decision process (MDP), which consists of state space S, action space A, state transition probability function P and reward function R [13]. At each time step t, the agent i observes the state s_i^t from the environment, takes action a_i^t based on the policy $\pi(a_i^t|s_i^t)$ and receives a reward r_i^t. The policy $\pi(a_i^t|s_i^t)$ maps the state s_i^t to a probability distribution of action a_i^t. For the next time step, a new state s_i^{t+1} is formed. The cumulative discounted reward could be expressed as

$$G_i^t = \sum_{k=0}^{\infty} \gamma_i^k r_i^{t+k} \tag{6}$$

where $\gamma \in [0,1]$ is the discounting factor.

In order to generate FDIA in DC microgrids to bypass the DE-based detection scheme, the target of the RL agents is to suppress the DE terms by injecting false signals into the sensors or communication links of multiple nodes. As is expressed in equations (4) and (5), any evident increase of the value will reflect an attack in the current counterparts of ith node. In addition, in the distributed cooperative control of the system, shown in Fig. 1, any deviation on current distributed terms I_{dc}^j or local terms I_{dc}^i would create an offset on term $\triangle V_2^i$ which deviates the voltage set-point $V_{dc_{ref}}^i$ from secondary control layer to the local control layer. This will lead to the deviation of the corresponding input reference current $I_{in_{ref}}^i$ and $I_{in_{ref}}^j$, which will change the value of DE terms. Thus, the RL-based intelligent attackers can harmonize and synchronize this offset between the neighboring nodes by attacking multiple nodes. In this way, the sophisticated attack that the conventional DE-based detection algorithm cannot detect is generated.

Particularly, in a DC microgrid shown in Fig. 1, for a node with m incoming links, the list of cyber-attack agents is defined as $\{CA_1, ..., CA_{m+1}\}$. The observations of each agent are $\boldsymbol{S_i^t} = \{\{DE_1^t, ..., DE_{m+1}^t\}, \{\int DE_1^t, ..., \int DE_{m+1}^t\}\}$ of neighboring nodes. The corresponding actions of each agent are $\boldsymbol{A_i^t} = \{I_{f1}^t, ..., I_{fm}^t\}$, where I_{f1}^t is the attack signals on local sensor and the rest are on incoming communication links from the neighboring nodes. The reward function is defined as

$$\begin{aligned} r_i^t = &-(k_{DE} \sum_{i=1}^{m+1} (DE_i^t)^2 + k_{\dot{DE}} \sum_{i=1}^{m+1} (\dot{DE}_i^t)^2 \\ &+ k_{\dot{I}_f} \sum_{i=1}^{m} (\dot{I}_{fi}^{t-1})^2) + r_{dis}^t \end{aligned} \tag{7}$$

where k_{DE} and $k_{\dot{DE}}$ are the coefficients for the summation of DE terms and the corresponding derivatives, respectively, which are adjusted to minimize the discordant terms and their variations. $k_{\dot{I}_f}$ is the coefficient for the summation of derivative of attack actions taken in the last time step $t-1$, which is tuned to minimize the variations of the generated attack signals especially when desirable stealthy attack performance is obtained. To help with the convergence of training the agents, a negative discrete reward term r_{dis}^t is introduced, which is expressed as below

$$r_{dis}^t = -(k_1 \cdot r_1^t + k_2 \cdot r_2^t) \tag{8}$$

$$r_1^t = |\sum_{j=1}^{m} (I^{t-1}{}_{fj} - I^{t-1}{}_{fi})| < I_{f_{min}} \tag{9}$$

$$r_2^t = \Big((DE_1^t | ... | DE_i^t) > DE_{max} \Big) \Big|_{i=1}^{m+1} \tag{10}$$

where the term r_1^t indicates the intrusion terms on sensor and cyber links are canceling each other according to the calculation of the consensus current [17], k_1 is the coefficient to ensure the presence of the minimum non-canceling destabilizing cyber-attack as denoted by $I_{f_{min}}$ with the overall output actions, i.e. the generated attack signals, in time $t-1$, and k_2 is the coefficient for penalizing the detection of excessive value of DE terms during the training stage. The value of $I_{f_{min}}$ is chosen with a trade-off between the slope of the ramp for the destabilizing phenomenon under the generated attacks and the minimum discordant terms threshold, DE_{max} is the upper threshold of discordant terms. Therefore, during the training process, the RL agents will learn autonomously to produce destabilization FDIAs to minimize the DE value. Thus, the attacks remain undetected by the discordant detection method.

The goal of the agents is to learn the policy $\pi(a_i^t|s_i^t)$ to maximize the reward r_i^t and thus to maximize the discounted reward G_i^t. For a specific policy π, the action-value function $Q^\pi(s_i^t, a_i^t)$ is used in reinforcement learning algorithm to describe the expected return with the action a_i^t with respect to the state s_i^t, which is estimated based on the Bellman equation as [18]

$$\begin{aligned} Q^\pi(s_i^t, a_i^t) \leftarrow\ & Q^\pi(s_i^t, a_i^t) + \\ & \alpha_i [r_i^{t+1} + \gamma_i max Q^\pi(s_i^{t+1}, a_i^{t+1}) - Q^\pi(s_i^t, a_i^t)] \end{aligned} \tag{11}$$

where α_i is the learning rate of agent i.

Deep Q-Network (DQN) is employed due to its computationally efficient characteristics [19]. To stabilize the training process, each experience tuple $e = (s_i^t, a_i^t, r_i^t, s_i^{t+1})$ of agent i at each time step is first stored in an R-sized experience memory $\mathcal{D} = \{e_i^1, ..., e_i^R\}$. In each time step of training process, a minibatch of the tuples are randomly selected from R. Afterwards, as the DQN agent has a Q-network, which approximates the action-value function $Q_i^\pi(s_i^t, a_i^t|\theta^Q)$ with weights θ^Q. In addition, to enhance the convergence of Q-network, a target network $\hat{Q}_i^\pi(s_i^t, a_i^t|\theta_i^{Q'})$ is used. The weights of the Q-network are optimized in the training process based on the loss function as below

$$L(\theta_i^Q) = \mathbb{E}[(Q_i^\pi(s_i^t, a_i^t|\theta_i^Q) - y^t)^2] \tag{12}$$

where

$$y^t = r_i^t + \gamma \hat{Q}_i^\pi(s_i^{'}, a_i^{'}|\theta_i^{Q'}) \tag{13}$$

The parameters of the target network are updated as

$$\theta_i^{Q'} \leftarrow \tau\theta_i^{Q} + (1-\tau)\theta_i^{Q'} \tag{14}$$

where the smoothing factor $\tau \ll 1$.

The whole training process is shown as follows.

Algorithm 1 Multi-agent RL-based FDIA

Input: DE and $\int DE$ of neighboring nodes
Output: Attack signals $\{I_{f1}^t, ..., I_{fm}^t\}$

Initialize replay buffer $\mathcal{D}$ to capacity R
Initialize action-value function Q with random weights θ_i^Q
Initialize target action-value function $\hat{Q}$ with weights $\theta_i^{Q'} = \theta_i^Q$
for episode = 1 to M **do**
  Receive initial observation at state s_i^1
  for iteration = 1 to T **do**
    For each agent i, select and execute action a_i^t with respect to policy $\pi(a_i^t|s_i^t)$, receive the reward r_i^t calculated with (7) and transition into state s_i^{t+1}
    Store tuple $(s_i^t, a_i^t, r_i^t, s_i^{t+1})$ in the $\mathcal{D}$
    for each agent i = 1 to m **do**
      Randomly select the mini-batch e from $\mathcal{D}$
      Set $y^t = \begin{cases} r_i^t, \text{if episode terminates at step } t+1 \\ r_i^t + \gamma_i max\hat{Q}_i^{\pi}(s_i^{'}, a_i^{'}|\theta_i^{Q'}), \text{otherwise} \end{cases}$
      Perform gradient descent on (12) with (13) regarding the network parameter θ_i^Q
    end for
    Update the target network using (14)
  end for
end for

III. PROPOSED DATA-DRIVEN CYBER-ATTACK DETECTION METHOD

We propose a data-driven detection method to complement the conventional DE-based detection method for the DC microgrids under multi-agent RL-based cyber-attacks. The proposed data-driven cyber-attack detection framework is illustrated in Fig. 2. The basis of the proposed method is to detect the attacks and identify the attacked nodes by extracting the mapping relationship between the inputs and the target labels, which consists of offline model training and online cyber-attack detection.

As the RL-based cyber-attacks can bypass the discordant detection algorithm, we aim to achieve attack detection as well as attacked nodes identification, which could be considered as a classification model. Since we use the measurement data of the system under both normal and attack conditions, considering the computational burden of the artificial neural network in the real-time application, we use the pattern recognition network (PRN), a type of feedforward neural network (FNN) for classifying the inputs to target classes. It is noteworthy that the paper is not focused on comparing the performance of different machine learning models or artificial neural networks for cyber-attack detection but to apply them to detect the cyber-attacks as well as identify the attacked nodes when the intelligent attacks occur in DC microgrids. The utilized

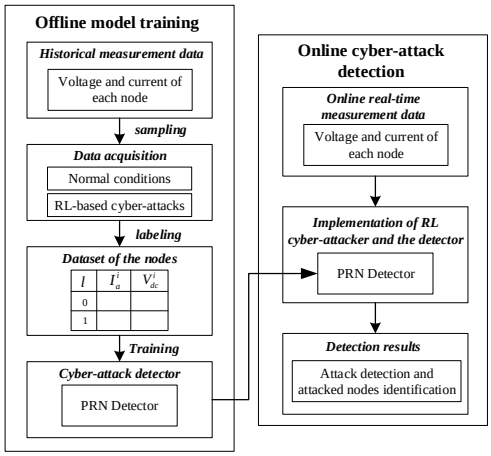


Fig. 2. Proposed data-driven cyber-attack detection method.

PRN is constructed with input, output, and single hidden layer as shown in Fig. 3, and the mathematical description is as follows.

$$Y = F(X_{in}) = f_{out}[f_{hid}(b_{hid} + W_{hid}X_{in})W_{out} + b_{out}] \tag{15}$$

where f, W, and b denote the activation function, weight matrix, and bias matrix respectively, $X_{in} = \{x_1, ..., x_n\}$ represents the input vector, the subscripts hid and out denote the hidden layer and output layer.

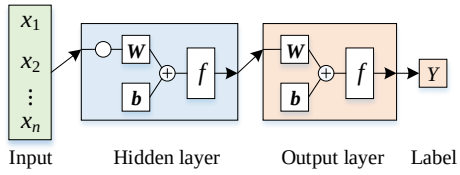


Fig. 3. Structure of PRN.

In the offline training process, to collect the data for training the neural network based cyber-attack detector, the historical operational data, consisting of measurements for I_a, V_{dc} in the system under both normal and RL-based cyber-attack conditions for each node, is collected. By labeling the attacked nodes as 1 and the unaffected nodes as 0, the dataset for all the nodes is generated. Then the data is randomly split into training dataset, validation dataset, and test dataset. During the offline training process, the weights and biases of each layer are optimized. Subsequently, a well-trained cyber-attack detector with prior knowledge of multi-agent RL-based cyber-attack is attained.

In the online cyber-attack detection process, during the real-time system operation, the measured voltage V_{dc} and current I_a are input to the trained PRN based cyber-attack detector. The PRN based cyber-attack detector is implemented as a classifier for each node to extract the labels of corresponding

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2022.3176301, IEEE Transactions on Industrial Electronics

input measurements, which will indicate whether the system is under attack and which nodes are attacked.

IV. EXPERIMENTAL RESULTS

To validate the performance of the proposed data-driven cyber-attack detection method, an experimental setup of the DC microgrid with n = 4 nodes in Fig. 1 is implemented, shown in Fig. 4. The parameters of the system and controller are listed in Table I. The trained RL agents target the current sensor signals on three neighboring nodes 1, 2 and 4, and produce stealthy destabilization FDIAs on DE terms. The experimental testbed consists of dSPACE MicroLabBox DS1202 and a computer as the real-time control interface. For both the normal condition and RL-based attack condition, operation data was collected from the computer interface with a sampling time of 0.2 ms for 10 s respectively, which means the dataset comprising a total of 100000 samples of I_a, V_{dc}, and the corresponding labels for each node. The PRN is trained based on the dataset, where 80% randomly divided data was used to train the neural network, and 10% was used for validation and testing, respectively. The training was carried out on Intel (R) Core (TM) i5-10210U 1.60GHz processor with 8.00 GB RAM. And the training time for a PRN is around 8 s. The performance of the training can be observed in the confusion matrix, which is shown in Fig. 5. The classification accuracy of the trained PRNs for correctly classifying the inputs to the target labels is about 98.3%. In order to evaluate the performance of the used PRN, classical 10-fold cross-validation is carried out, where the collected dataset for training and testing is randomly and repeatedly assigned [20]. The average classification accuracy is about 98.5%, which verifies the effectiveness of the employed PRN for attack detection. The RL agents are applied via dSPACE with a sampling time of 50 μs. Then, the trained cyber-attack detectors are implemented in the experimental testbed to verify its performance in the DC microgrid under the multi-agent RL-based cyber-attack.

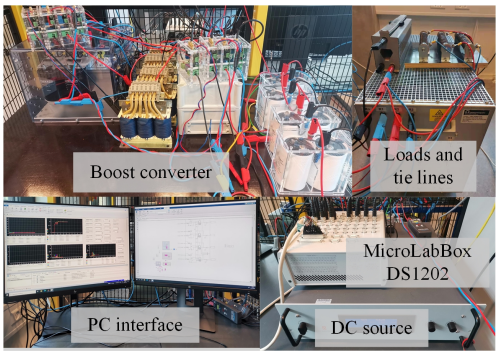


Fig. 4. Experimental setup of a DC microgrid with 4 nodes.

The experimental results are shown in Fig. 6. The RL agents target the sensor signals to produce destabilizing FDIAs on three neighboring nodes 1, 2, and 4. It can be observed when

TABLE I. Experimental Setup Parameters

Parameter sets	Values
Converter	$L_{in} = 0.86\ mH, C_o = 1.1\ mF, f_s = 10\ kHz, I_{rated} = 32\ A$ Loads: $R_1 = R_2 = R_3 = R_4 = 30.6\ \Omega$ Tie lines: $R_{12} = R_{23} = R_{34} = 0.5\ \Omega, R_{14} = 0\ \Omega$
Controller	$V_{in} = 48\ V, V_{dc_{ref}} = 60\ V, I_{dc_{ref}} = 0$ Primary layer: $K_{pV} = 1, K_{iV} = 20, K_{pI} = 2.4, K_{iI} = 10$ Secondary layer: $K_p^I = 0.12, K_i^I = 0.15$

Fig. 5. Training results of the PRN for cyber-attack detection. (1:Attacked; 0:Healthy)

the RL-based attack is initiated, the agents generate the attacks on multiple nodes, and the system deviates from the normal operating condition. Moreover, from the DE terms shown in Fig. 6(c), it is evident that the RL algorithm successfully generates the sophisticated attacks that remain stealthy to the DE-based detection method as their values resemble the normal conditions counterparts and are maintained in their lower permissible range.

At the initial stage within 11 s, the system operates normally with well-tuned distributed control, where about 6 A current are shared proportionately among the four converters and output voltage for each converter converges to around 60 V. At around t = 11 s, the RL-based FDIA is initiated, which results in a 0.3 A current deviation on the compromised node and the unaffected node also experience a current rise to about 1.75 A, as shown in Fig. 6(b). It could also be observed in Fig. 6(a) that the output voltage of all the converters decreases by 0.05 V/s.

The performances of the conventional discordant detection method and proposed data-driven cyber-attack detection method are shown in Fig. 6(c) and (d). When the attack is initiated, it is observed in Fig. 6(c) that the DE terms for the compromised nodes are suppressed to the normal condition counterparts and within their minimum threshold. Neither of the attacked nodes is manifested with evident higher index values than the normal condition, according to the detection criteria in equation (5), clearly indicating that the conventional discordant detection method fails to detect the attacks and the attacked nodes under the RL-based cyber-attacks. For the proposed method, when the attack occurs, as shown in Fig. 6(d), with the labels scaled with the corresponding node index

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2022.3176301, IEEE Transactions on Industrial Electronics

and 0.2 scaling factor, the well-trained PRN detectors signal out the compromised nodes 1, 2 and 4, while the output label for the node 3 is kept at 0, revealing that the cyber-attack occurs and the node 1, 2 and 4 are attacked. When the attack is removed, at around t = 36 s, the system works normally at a new operating point and the detection metrics for all the four nodes are maintained at 0. From the above analysis, it can be concluded that the conventional discordant detection scheme is ineffective under the RL-based cyber-attack, and the proposed data-driven neural network based detector can detect the attacks and identify the attacked nodes for DC microgrids under the RL-based cyber-attacks.

V. Discussion

Due to the system being regulated under the distributed cooperative control structure, the distributed multi-agent reinforcement learning algorithm can be applied to generate sophisticated attacks in larger systems. And the proposed data-driven detection method can still be employed to detect the attacks as long as the real operational data is collected. Moreover, in larger systems, as the distributed control is implemented independently in each microprocessor, the computational burden on the implementation of the proposed data-driven method and RL-based cyber-attack will not increase. Therefore, the RL-based cyber-attack and proposed data-driven detection method are scalable in larger systems with more nodes.

In addition, we can also design the RL-based attacker and train the proposed data-driven detector in an iterative way. In each iteration, by updating the model of the system protected with the new detection mechanism, i.e. the DE-based detection method and the up-to-date data-driven detector complementing each other, the RL-based attacker will explore in all the other attack types and exploit the vulnerability of the new detection mechanism, thus generating sophisticated attacks to bypass the new detection mechanism. In turn, by updating the database of the operation data under the new attacks, the data-driven cyber-attack detector is trained and implemented to detect the new attacks. Eventually, more attacks can be detected with the new detection mechanism.

To sum up, the proposed data-driven cyber-attack detection method can complement the DE-based detection method to detect the RL-based intelligent attacks and identify the attacked nodes which the conventional DE-based detection method fails to detect. Compared with the method in [7], which collects simulation data based on the healthy model of microgrids and can only detect a certain type of attack, the proposed method collects real data from system operation under both normal and attack conditions, which enable it to detect a wider range of cyber-attacks with high precision. Moreover, as the RL-based intelligent attacker can also learn to bypass other index-based detection methods [10]–[12], in the same way, the proposed data-driven attack detection method can be employed to complement other conventional cyber-attack detection methods.

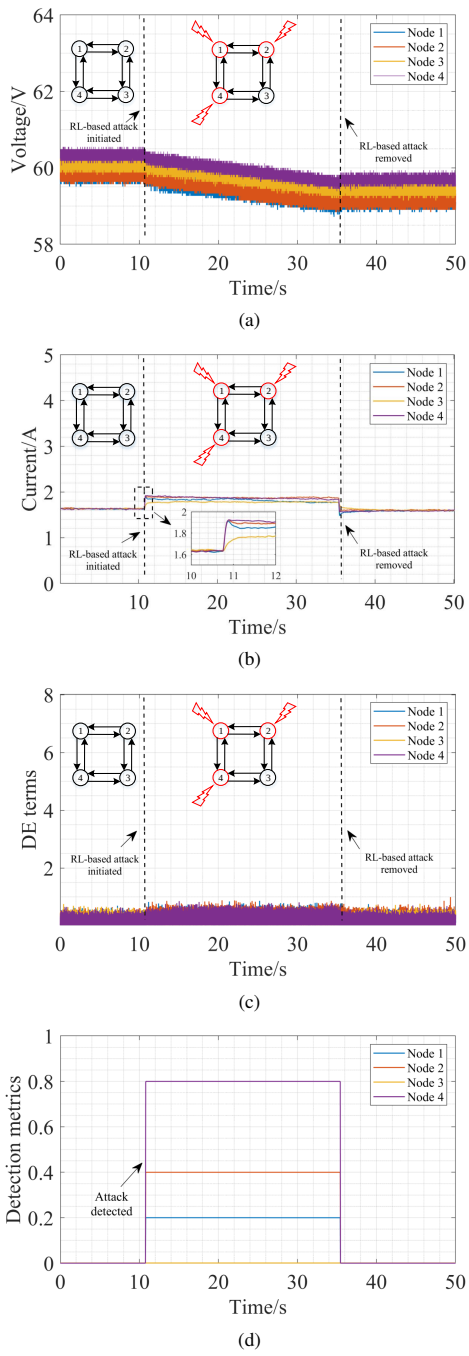


Fig. 6. Experimental validation of the proposed data-driven detection method for the DC microgrid. (a) Output voltage, (b) Output current, (c) Performance of conventional DE-based detection method, and (d) Performance of proposed data-driven detection method.

VI. Conclusions

This paper proposes a data-driven cyber-attack detection method to complement the DE-based detection method for DC microgrids under multi-agent RL-based cyber-attack. In particular, the multi-agent RL algorithm is employed to generate sophisticated attacks against the conventional DE-based detection method. The dataset of the DC microgrids operating under both normal and cyber-attack conditions is collected for offline training of the PRN based cyber-attack detectors. Then, the well-trained neural network based cyber-attack detectors are implemented in an experimental testbed to verify the performance of the proposed data-driven method. The experimental results show that the RL-based attacks remain undetected by the DE-based detection method as DE indices are maintained within their minimal permissible range, and the proposed data-driven detector works as a complementary detection scheme, detects the attacks and attacked nodes successfully. Moreover, the proposed detection mechanism could also be employed to complement other conventional cyber-attack detection approaches when they fail under the intelligent attacks.

References

[1] T. Dragičević, X. Lu, J. C. Vasquez, and J. M. Guerrero, "Dc microgrids part i: A review of control strategies and stabilization techniques," *IEEE Transactions on power electronics*, vol. 31, no. 7, pp. 4876–4891, 2015.

[2] M. Yazdanian and A. Mehrizi-Sani, "Distributed control techniques in microgrids," *IEEE Transactions on Smart Grid*, vol. 5, no. 6, pp. 2901–2909, 2014.

[3] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Transactions on Information and System Security (TISSEC)*, vol. 14, no. 1, pp. 1–33, 2011.

[4] O. A. Beg, T. T. Johnson, and A. Davoudi, "Detection of false-data injection attacks in cyber-physical dc microgrids," *IEEE Transactions on industrial informatics*, vol. 13, no. 5, pp. 2693–2703, 2017.

[5] A. Karimi, A. Ahmadi, Z. Shahbazi, H. Bevrani, and Q. Shafiee, "On the impact of cyber-attacks on distributed secondary control of dc microgrids," in *2020 10th Smart Grid Conference (SGC)*, pp. 1–6. IEEE, 2020.

[6] A. S. Musleh, G. Chen, and Z. Y. Dong, "A survey on the detection algorithms for false data injection attacks in smart grids," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2218–2234, 2019.

[7] M. R. Habibi, S. Sahoo, S. Rivera, T. Dragičević, and F. Blaabjerg, "Decentralized coordinated cyber-attack detection and mitigation strategy in dc microgrids based on artificial neural networks," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, 2021.

[8] O. A. Beg, L. V. Nguyen, T. T. Johnson, and A. Davoudi, "Signal temporal logic-based attack detection in dc microgrids," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3585–3595, 2018.

[9] S. Sahoo, J. C.-H. Peng, A. Devakumar, S. Mishra, and T. Dragičević, "On detection of false data in cooperative dc microgrids:a discordant element approach," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 8, pp. 6562–6571, 2019.

[10] J. Zhang, S. Sahoo, J. C.-H. Peng, and F. Blaabjerg, "Mitigating concurrent false data injection attacks in cooperative dc microgrids," *IEEE Transactions on Power Electronics*, vol. 36, no. 8, pp. 9637–9647, 2021.

[11] D. Shi, P. Lin, Y. Wang, C.-C. Chu, Y. Xu, and P. Wang, "Deception attack detection of isolated dc microgrids under consensus-based distributed voltage control architecture," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 11, no. 1, pp. 155–167, 2021.

[12] S. Sahoo, T. Dragičević, and F. Blaabjerg, "An event-driven resilient control strategy for dc microgrids," *IEEE Transactions on Power Electronics*, vol. 35, no. 12, pp. 13 714–13 724, 2020.

[13] C. Wang, J. Wang, Y. Shen, and X. Zhang, "Autonomous navigation of uavs in large-scale complex environments: A deep reinforcement learning approach," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 2124–2136, 2019.

[14] Y. Chen, S. Huang, F. Liu, Z. Wang, and X. Sun, "Evaluation of reinforcement learning-based false data injection attack to automatic voltage control," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 2158–2169, 2018.

[15] Z. Zhang, D. Zhang, and R. C. Qiu, "Deep reinforcement learning for power system applications: An overview," *CSEE Journal of Power and Energy Systems*, vol. 6, no. 1, pp. 213–225, 2019.

[16] Z. Ni and S. Paul, "A multistage game in smart grid security: A reinforcement learning solution," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2684–2695, 2019.

[17] V. Nasirian, S. Moayedi, A. Davoudi, and F. L. Lewis, "Distributed cooperative control of dc microgrids," *IEEE Transactions on Power Electronics*, vol. 30, no. 4, pp. 2288–2303, 2014.

[18] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[19] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[20] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. Springer, 2013.

[J4] Data-driven Cyber-attack Detection of Intelligent Attacks in Islanded DC Microgrids

Authors:
Ali Jafarian Abianeh, Yihao Wan, Farzad Ferdowsi, Nenad Mijatovic, Tomislav Dragicevic

Submitted to:
IEEE Transactions on Power Electronics

Status:
Published.
Digital Object Identifier: 10.1109/TPEL.2021.3132028

Vulnerability Identification and Remediation of FDI Attacks in Islanded DC Microgrids Using Multi-agent Reinforcement Learning

Abianeh, Ali Jafarian ; Wan, Yihao; Ferdowsi, Farzad; Mijatovic, Nenad; Dragicevic, Tomislav

Published in:
IEEE Transactions on Power Electronics

Link to article, DOI:
10.1109/TPEL.2021.3132028

Publication date:
2022

Document Version
Peer reviewed version

Vulnerability Identification and Remediation of FDI Attacks in Islanded DC Microgrids Using Multi-agent Reinforcement Learning

Ali Jafarian Abianeh, *Student Member, IEEE,* Yihao Wan, *Student Member, IEEE,* Farzad Ferdowsi, *Senior Member, IEEE,* Nenad Mijatovic, *Senior Member, IEEE* and Tomislav Dragičević, *Senior Member, IEEE*

***Abstract*—This paper proposes a novel approach to uncover deficiencies of the existing cyber-attack detection schemes and thereby to serve as a foundation for establishing more reliable cybersecure solutions, with particular application in DC microgrids. For this purpose, a multi-agent deep Reinforcement Learning (RL) based algorithm is proposed to automatically discover the vulnerable spots on the conventional index-based cyberattack detection schemes, and automatically generate coordinated stealthy destabilizing False Data Injection (FDI) attacks on cyber-protected islanded DC microgrids. To enable a continuous action space for the trained RL agents and enhance the algorithm's precision and convergence rate, Deep Deterministic Policy Gradient (DDPG) is incorporated. Using this approach, susceptibility of a state-of-the-art detection scheme to several different coordinated FDI attacks on the distributed communication links is identified. The proposed algorithm is also enhanced with a sniffing feature to enable maintaining the stealthy attacks even under the sudden disconnection of any of the compromised links. To address the discovered deficiencies within the index-based detection scheme, a complementary multi-agent RL detection algorithm using Deep Q-Network (DQN) is integrated, which provides a more reliable overall identification performance. Taking into account the communication delays and load changes, the effectiveness of the proposed algorithm is verified by the experimental tests.**

***Index Terms*—Distributed Control, DC Microgrid, Cybersecurity, False Data Injection, Reinforcement Learning.**

I. Introduction

DC microgrids have recently received a wide range of attention and growing popularity in power generation systems, as they provide an efficient way for integration of renewable energy systems, energy storage units and electrical power loads [1]. Using a hierarchical control structure with a combination of both primary and secondary control layers, the voltage regulation at the output terminals and current sharing among generation units are deployed in such systems [2]. Conventional approach for the secondary control schemes were formed on the basis of the centralised control, where a single control block was in charge of receiving secondary signals and dispatching the voltage regulatory terms to all downstream primary control units based on the underlying control objectives. However, this approach makes the system vulnerable to the single point of failure. To overcome this problem, distributed control algorithms have been developed, where the secondary control command signals are generated at the place of each node based on the received distributed signals from the neighboring agents and a consensus rule of operation. Owing to the dense integration of communication links among the neighboring agents and local-to-secondary control layers for each agent, DC microgrids are highly prone to malicious cyber-attacks. Such intrusions can highly deteriorate the system's performance and even result in unstable conditions and protective circuits tripping under severe cases. Different types of cyberattacks on DC microgrids and their detrimental impacts are studied in the literature including False Data Injection (FDI) [3], Denial Of Service (DOS) [4], Hijacking [5] and Man In the Middle (MIM) [6] attacks.

Compared with other forms of cyberattacks, FDI is known as one of the most challenging types for proper detection and it can occur in different forms. Destabilizing FDIs can make the microgrid unstable with only a minimal uncoordinated penetration level. On the other hand, deceptive FDIs can produce deviations from optimal operating points without loss of regulation through more coordinated attacks [7]. The latter can be only generated with a limited set of coordinated intrusions and is effectively detectable by the existing identification algorithms [7], [8]. However, it is highly critical to ensure the reliable performance of the reported detection schemes against all possible forms of destabilizing FDIs, where any detection failure in this regard can result in protective circuit tripping or damage to power converters. In efforts to effectively address the aforementioned destabilizing or deceptive attacks, multiple FDI detection and mitigation algorithms are reported in different research works. Such schemes can be generally categorized into the model-dependent and model-independent methods.

For model-dependent schemes, researchers have incorporated adaptive control concepts [9], or sliding mode observers [10]. However, complexity and precision of such algorithms are highly dependent on the order size of the underlying model. They are also prone to instability under system parameter uncertainties or presence of multiple coordinated cyberattacks. Model-free FDI mitigation methods based on the distributed observers are also reported in the literature, where adaptive distributed terms [11], or sliding mode observer based distributed terms [12] are employed to rectify the FDI adverse

Ali Jafarian Abianeh and Farzad Ferdowsi are with the Electrical and Computer Engineering Department of the University of Louisiana at Lafayette, Louisiana, USA (emails: ali.jafarian-abianeh1@louisiana.edu, farzad.ferdowsi@louisiana.edu).

Yihao Wan, Nenad Mijatovic and Tomislav Dragičević are with the Electrical Engineering Department of Technical University of Denmark, Copenhagen, Denmark (e-mails: wanyh@elektro.dtu.dk, nm@elektro.dtu.dk, tomdr@elektro.dtu.dk).

impacts. However, such schemes are highly reliant on the secure transmission of the extra distributed signals, which can be themselves targeted by the FDIs. In addition, their dynamic performance are significantly deteriorated with the communication delays, and their load switching response is also adversely impacted by the integrated distributed terms.

FDI detection algorithms based on the supervised learning have also been investigated [13], [14], but their performance is greatly impacted by the quality of collected labeled dataset, and they are also prone to the over-fitting phenomenon. To mitigate these challenges, model-free FDI detection schemes, based on the system physical observations, are recently employed by researchers. A signal-temporal-logic approach is used in [15] to detect a sawtooth form of FDI, and an exponential data integrity index on the distributed current signals is employed in [16] to signal out the compromised links. However, the performance of these schemes against deceptive attacks are not studied. A discordant detection algorithm is proposed by [8] for detection of both deceptive and destabilizing FDI attacks on secondary regulation of current signals. This scheme is formed on the basis of monitoring the synchrony of the resultant current references. A similar approach is also utilized by [17] and [18] to develop event-driven cyber-attack detection and mitigation algorithms against FDI attacks. Despite the promising performance of such model-free FDI detection schemes, their performance against more systematic FDIs is still not guaranteed. Thus, it is crucial to explore the susceptibilities within the existing algorithms, and accordingly apply the proper modifications.

Reinforcement Learning (RL) algorithms have recently received an enormous attention in the cyber-physical systems [19], as it is known as the closest form to the human learning compared with other types of intelligent algorithms. However, only very limited number of research works have explored RL application to cyber-security in microgrids and smart grids [20]. Using a combination of SARSA RL on the learning phase and Q-Learning RL for FDI detection are reported in [21] and [22]. However, Q-learning based algorithms do not provide an efficient solution for real-world applications where deep learning based RLs are more desired. Despite some reported RL based FDI detection schemes, the great potential in such learning methods is still not well realized for vulnerability exploration and exploitation in the existing cyberattack detection schemes and developing effective complementary mitigation. In a recent research study [23], the application of a Temporal Difference (TD) RL actor-critic based method is studied for intervening the cost optimization in the tertiary control layer of microgrids.

In this paper, deep RL algorithms are proposed to autonomously discover the vulnerabilities of the index based cyberattack detection methods commonly used in distributed control of DC microgrids, and provide complementary solutions. Using a multi-agent RL approach with Deep Deterministic Policy Gradient (DDPG) agents for exploring the FDI cyberattack continuous space action, a more precise identification of the vulnerable spots is attained. The proposed method explores the detection algorithm susceptibilities against stealthy destabilizing FDIs on distributed links in a way that indices remain minimized to the normal operating condition. Then, a multi-agent RL DQN based scheme is proposed to supplement the identified detection weaknesses and operates in conjunction with it. The performance of the proposed scheme is verified using an experimental testbed against one of the state-of-the-art model-free FDI detection methods [8]. To the best of authors' knowledge, this paper proposes the first multi-agent deep RL based schemes for cybersecurity issues in the secondary control layer of microgrids. Thus, the main contributions of this paper can be summarized as follows:

- A novel approach for automatic discovery of the vulnerabilities within the existing cyberattack detection algorithms is proposed using the reinforcement learning concept. This method enables both wide-range and targeted exploration of the penetrable spots for all the index-based cyberattack detection schemes, and provides foundations for their effective mitigation.
- The proposed deficiency identification scheme is implemented using the multi-agent DDPG RL agents. This multi-agent configuration facilitates its effective integration into the existing distributed cyberattack detection schemes and alleviates the impacts of communication delays on its performance compared with a centralised approach. In addition, utilization of the DPPG agents enables a continuous space action for finer exploration of the susceptibilities to cyberattacks.
- Using the identified deficiencies in the model-free index based detection schemes, a complementary multi-agent RL DQN based cyberattack identification algorithm is proposed, which signals out coordinated attacks undetected by the fundamental scheme.

The rest of this paper is organized as follows; in Section II, distributed control for DC microgrids and discordant detection algorithm are discussed. The proposed multi-agent RL based algorithms are also presented in Section III. Experimental results are provided in Section IV, and Section V concludes this research study.

II. Distributed control of DC microgrids with Discordant Cyberattack Detection

An autonomous DC microgrid with the topology shown in Fig. 1 is considered. In this system, each DC source is connected to the common DC bus through a DC-DC converter, which is regulated with cascaded voltage and current controllers at the primary layer. Distributed secondary regulators are also integrated with the primary controllers to enable transmission and sharing the distributed terms of $\phi_n = \{\bar{V}_{dc_n}, I_{dc_n}\}$ for n distributed agents based on the underlying communication topology. In this case, $\bar{V}_{dc_n}$ denotes the estimated average voltage, and I_{dc_n} is the per unit value for the output current. For the adjacency matrix $\boldsymbol{A} = [a_{ij}]$ with the dimension of $n \times n$, the resultant consensus secondary term (u_i) at the place of node i can be then represented by:

$$u_i = \sum_{j=1}^{n} a_{ij}(\phi_j - \phi_i), \;\; where \;\; a_{ij} = \begin{cases} > 0, \; if(x_i, x_j) \, \epsilon \, \boldsymbol{G} \\ 0, \; else \end{cases} \tag{1}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPEL.2021.3132028, IEEE Transactions on Power Electronics

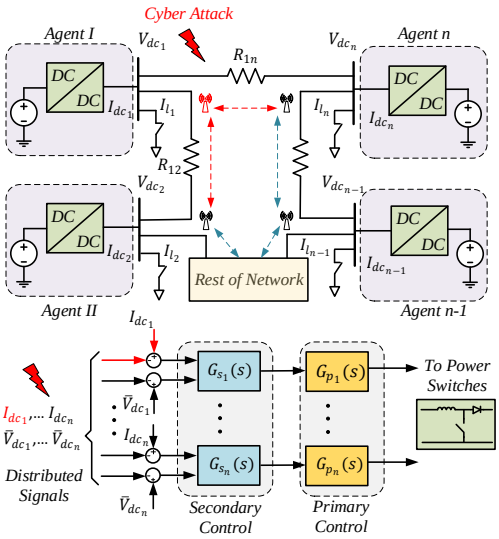


Fig. 1. General block diagram for distributed control of DC microgrid with N agents under cyber-attacks.

where a_{ij} denotes the interconnection between all nodes, $\boldsymbol{G}$ represents the existing communication topology, and x_i and x_j are the secondary signals for local and neighboring nodes.

To implement secondary voltage and current sharing between the neighboring agents, it is required to modify the primary voltage setpoint for node i as follows:

$$V^*_{dc_i} = V_{dc_{ref}} + \triangle V_{1_i} + \triangle V_{2_i} \tag{2}$$

where $V_{dc_{ref}}$ is the global reference voltage for all agents and $V^*_{dc_i}$ is the voltage reference to the primary controller at node i. $\triangle V_{1_i}$ and $\triangle V_{2_i}$ also represent the resultant regulatory terms from the secondary voltage and current controllers at node i, respectively, and can be formulated with:

$$\triangle V_{1_i} = (K^V_p + \frac{K^V_i}{s}) \cdot (V_{dc_{ref}} - u^V_i) \tag{3}$$

$$\triangle V_{2_i} = (K^I_p + \frac{K^I_i}{s}) \cdot (I_{dc_{ref}} - u^I_i) \tag{4}$$

where $I_{dc_{ref}}$ is the global current setpoint, u^V_i and u^I_i are the consensus terms for voltage and current, and K^V_p, K^V_i, K^I_p, K^I_i represent the proportional and integral gains for voltage and current PI controllers, respectively. For a proportionate current sharing among the neighboring agents, $I_{dc_{ref}}$ is set by zero.

Due to the droop concept for the interconnected DC power sources, the secondary sharing algorithms are usually applied to either the voltage or current distributed terms. Since this paper is focused on discovering the vulnerabilities in the index based detection algorithms applied on the current sharing [8], only FDI attacks on secondary current control are investigated. For secondary current regulation schemes, FDI cyberattacks can be applied through an offsetting term either on sensor or neighboring communication links, as formulated by (5):

$$\boldsymbol{X^{FDI}} = \begin{cases} x^{FDI}_i = x_i + k_i x_f, \ for\ sensor\ attack \\ x^{FDI}_{ij} = x_{ij} + k_{ij} x_f, for\ link\ attack \end{cases} \tag{5}$$

where $\boldsymbol{X^{FDI}}$ denotes the matrices of compromised communication signals, x^{FDI}_i and x^{FDI}_{ij} are the FDI manipulated signals for sensor and neighboring links, k_i and k_{ij} are the FDI scaling factors for sensor and link communications, respectively, while x_f is the fundamental FDI intrusion term.

The FDI cyberattacks can be generated in the form of destabilizing or deceptive attacks through coordinated or uncoordinated intrusions. As a result, a discordance effect is experienced on the input current signals, which forms the basis for the discordant FDI detection algorithm [8], as formulated in (6). In this algorithm, the resultant impacts from any forms of FDIs including attacks on sensors, communication links or concurrent ones are monitored through the deviations introduced on the input current signal references for the neighboring agents. As also represented by (7), the presence of cyberattack on the distributed control block for node i is detected when the positive term DE_i has a value greater than its minimum threshold DE_{min}. This value is chosen by considering the resultant DE_i values under normal operating conditions in the presence of other possible contributing factors such as underlying controllers' performance and existing communication delays. In addition, a time delay triggering process is employed before applying the counteracting measure in order to account for load disturbances. However, for the normal operating conditions, the discordant value should never consistently retain a value greater than its lower threshold.

$$DE_i = M_i \cdot [\sum_{j \epsilon G_i} (I^*_{j_{in}} - I^*_{i_{in}})][\sum_{j \epsilon G_i} (I^*_{j_{in}} + I^*_{i_{in}})] \tag{6}$$

$$DE_i = \begin{cases} < DE_{min} : \ \ for\ k_i\ \&\ k_{ij} = 0 \\ > DE_{min} : \ \ for\ k_i \| k_{ij} \neq 0 \end{cases} \tag{7}$$

where DE_i is the discordant term at node i, $DE_{min} > 0$ is its minimum threshold, $I^*_{i_{in}}$ and $I^*_{j_{in}}$ are the reference input current values at local node i, and neighboring nodes j, M_i is the scaling factor for discordant term, and G_i represents the communication graph to the neighboring agents.

III. Proposed Multi-Agent RL Schemes to Unveil Susceptibilities and Complement Detection

A. Multi-Agent Deep Reinforcement Learning

For development of multi-agent RL algorithms in an observable environment, the problem is defined as a Markov Decision Process (MDP) characterised with the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma \rangle$ for each agent in the agent set $\mathcal{N} = \{N_1, .., N_m\}$. Where $\mathcal{S} \epsilon \mathbb{R}^n$ represents the finite set of states, $\mathcal{A} \epsilon \mathbb{R}^m$ denotes the finite set of actions, $\mathcal{T}$ is the state transition function that represents the probability of state transition $s^t \rightarrow s^{t+1}$ by taking the action a^t and receiving the immediate reward of r^t, r is the reward function, and $\gamma \ \epsilon \ [0, 1]$ is the discount factor. In each time step, the RL agents observe the current system state s^t and take the action a^t based on the selected policy $\pi(a^t|s^t)$. This

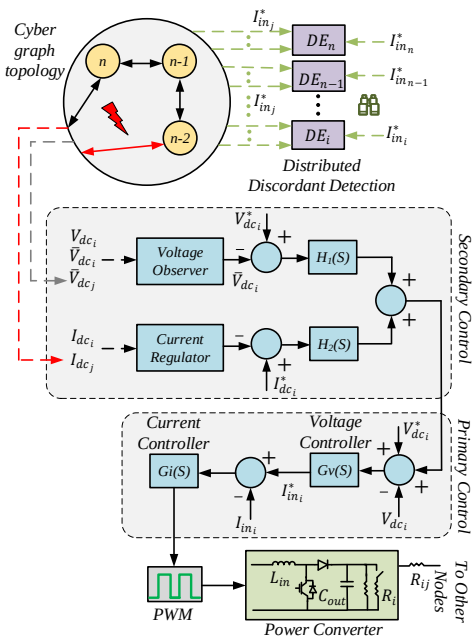


Fig. 2. Block diagram for discordant cyberattack detection and the hierarchical control structure in DC microgrid.

taken action results in receiving an immediate reward r^t, and its transition into the new state s^{t+1}. The term accumulative reward over an infinite time horizon for each RL agent i can be also represented by:

$$\Gamma_i^t = \sum_{n=0}^{\infty} \gamma_i^n r_i^{t+n+1} \tag{8}$$

In order to maximize the accumulative reward in (8), different recursive training algorithms can be applied. In the off-policy based algorithms such as Q-learning, the Bellman iterative equation in (9) with the learning rate α_i is employed to estimate the action-value function Q^π:

$$\begin{aligned} Q^\pi(s_i^t,a_i^t) \leftarrow Q^\pi(s_i^t,a_i^t) + \\ \alpha_i[r_i^{t+1} + \gamma_i max Q^\pi(s_i^{t+1},a_i^{t+1}) - Q^\pi(s_i^t,a_i^t)] \end{aligned} \tag{9}$$

However, this iterative approach does not provide a feasible performance in a high dimensional real-world application. To enable a more precise prediction of the action-value function Q_i^π for each pair of state-action, DQN is employed for the RL agents [24]. In DQN algorithm, first a random mini-batch of $\mathcal{S}$ samples (s^j, a^j, r^j, s'^j) from the replay buffer $\mathcal{D}$ is chosen for each agent i. Then, the critic network is adjusted by trying to predict the return value with minimizing the following loss function:

$$L(\theta_i) = \frac{1}{S}\sum_j (y^j - Q_i^\pi(s^j, a_1^j, ..., a_m^j))^2 \tag{10}$$

where y^j is set with:

$$y^j = r_i^j + \gamma_i Q_i^{\pi'}(s'^j, a'_1, ..., a'_m)|_{a'_i=\pi'_i(s^j)} \tag{11}$$

While DQN RL agents can efficiently meet the algorithm objectives in some applications where the limited set of discrete actions are adequate to interact with the environment, RL agents with the capability of continuous space action using actor-critic networks such as DDPG, are unavoidable for more complex environments [25]. Similar to the DQN, the critic network is adjusted by minimizing the loss function in (10), but the actions are decided based on the adjusted actor network with minimizing the loss function in (12) to acquire the optimal policy parameter θ:

$$\nabla_{\theta_i} J = \frac{1}{S}\sum_j \nabla_{\theta_i}\pi_{\theta i}(a^j|s^j)\nabla_{a_i} Q_i^\pi(s^j, a_1^j, ..., a_m^j)|_{a_i=\pi_{\theta i}(s^j)} \tag{12}$$

Then the target network parameters for both actor and critic networks are updated with (13):

$$\theta'_i \leftarrow \tau\theta_i + (1-\tau)\theta'_i \tag{13}$$

where $\tau \ll 1$.

B. Multi-Agent RL DDPG to Uncover Cyberattack Detection Deficiencies

In order to automatically discover the vulnerabilities within an index-based cyberattack detection scheme, the problem of coordinated cyberattack generation can be formulated as a Markov Decision Process (MDP). With specifying a continuous action space for cyberattack exploration using the DDPG RL agent, the task of generating stealthy FDI attacks can be accomplished by properly rewarding the low detection indices in presence of intrusions. The application of this approach to the discordant detection scheme, as an existing well-established identification algorithm, is explained in this section, while the similar approach can be applied to all other index based detection schemes. Due to the lower vulnerabilities and exposure of node links to cyberattacks, the proposed algorithm only explores the undetectable FDI attacks on the neighboring links. However, it can be easily reconfigured for nodes and combinative attacks as well. In this section, a multi-agent DDPG reinforcement learning based FDI attack generation scheme is proposed. This approach enables modular integration of the proposed distributed RL cyberattack on the more densely connected or expanded networks, lessens the impact of transmission delays from distributed agents to a centralised attacking unit, and ensures optimal complexity level for each trained agent in terms of possible output actions and convergence effort.

In a DC microgrid with m incoming distributed communication links, the agents list is set with $\boldsymbol{N} = \{CA_1, ..., CA_{m+1}\}$ in which CA denotes the RL based cyberattack generation agents. Considering the existing inter-dependency of the neighboring discordant terms, the observations list for each agent at the instance t is defined as $\boldsymbol{O_{CA}^t} = \{\boldsymbol{O_1^t}, \boldsymbol{O_2^t}\}$, where $\boldsymbol{O_1^t} = \{DE_1^t, ..., DE_{m+1}^t\}$ is devoted to the neighboring

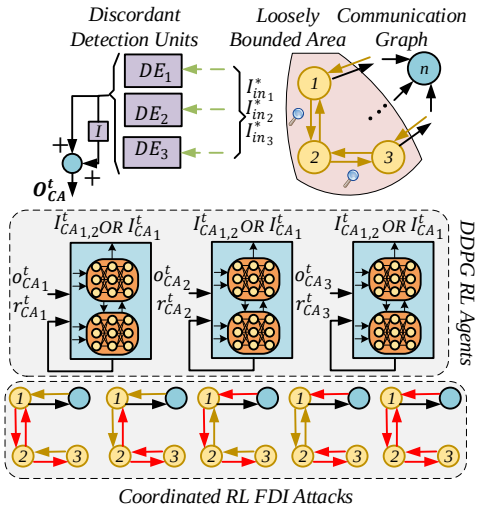


Fig. 3. Proposed multi-agent DDPG RL-based FDI attacks against DC microgrid equipped with discordant detection algorithm- Sample network.

discordant terms, and $\boldsymbol{O_2^t} = \{\int DE_1^t, ..., \int DE_{m+1}^t\}$ represents their associated integration. With respect to the received observations, the action set $\boldsymbol{A_{CA}^t} = \{I_{CA_1}^t, ..., I_{CA_m}^t\}$ is generated at instance t, where I_{CA}^t is the FDI intrusion term applied on the incoming communication link m at moment t, and $\boldsymbol{A_{CA}^t} \epsilon \boldsymbol{A_{CA}}$ where $\boldsymbol{A_{CA}}$ represents the finite set of available actions to CA agent. The corresponding algorithm steps are also presented by Algorithm 1. The reward function for each agent at time t is also defined by (14), which is characterised with both continuous and discrete reward terms:

$$r_{CA}^t = -(k_{DE}\sum_{i=1}^{m+1}(DE_i^t)^2 + k_{\dot{DE}}\sum_{i=1}^{m+1}(\dot{DE}_i^t)^2 + k_{\dot{I}_{CA}}\sum_{j=1}^{m}(\dot{I}_{CA_j}^{t-1}))^2 + r_{d_p}^t - r_{d_n}^t \tag{14}$$

where k_{DE}, $k_{\dot{DE}}$ and $k_{\dot{I}_{CA}}$ are the reward coefficients for summation of neighboring discordant terms, their corresponding derivatives, and derivatives of cyberattack actions taken in $t-1$, respectively. $\dot{DE}_i^t$ and $\dot{I}_{CA_j}^{t-1}$ also denote the derivatives for discordant terms at time t and cyberattack action signals at time $t-1$, respectively. $r_{d_p}^t$ and $r_{d_n}^t$ also represent the positive and negative discrete reward terms considered to ensure effective training for desired destabilizing conditions, penalizing the excessive DE observation values, while rewarding the stealthy destabilizing attacks. These discrete reward terms also facilitate the convergence process during the training stage.

For the continuous reward terms, k_{DE} is tuned to ensure minimized discordant observations, $k_{\dot{DE}}$ is applied to reflect the impact of excessive transient modes from disturbances such as load changes or cyberattack variations for enhanced stealthy performance, and $k_{\dot{I}_{CA}}$ is in charge of minimizing variations on the generated cyberattack terms especially while the desired objectives are met. However, if more dynamic cyberattack steps are desired, this coefficient can be adjusted with lower values. The expansion of discrete reward terms are also represented by (15)-(21):

$$r_{d_p}^t = K_{d_p}(\mathcal{E}^t \,\&\, \mathcal{P}^t) \tag{15}$$

$$r_{d_n}^t = K_{d_{n1}}\left(\overline{\mathcal{E}^t} \,\&\, (\mathcal{G}^t|\mathcal{H}^t)\right) + K_{d_{n2}}\mathcal{F}^t \tag{16}$$

$$\mathcal{P}^t = (|\sum_{j=1}^{m}(I^{t-1}{}_{CA_j} - I^{t-1}{}_{CA_i})| > I_{CA_{min}}) \tag{17}$$

$$\mathcal{E}^t = \left((DE_1^t \,\&\, ... \,\&\, DE_i^t) < DE_{min}^t\right)\Big|_{i=1}^{m+1} \tag{18}$$

$$\mathcal{F}^t = \left((DE_1^t|...|DE_i^t) > DE_{max}^t\right)\Big|_{i=1}^{m+1} \tag{19}$$

$$\mathcal{G}^t = (I_{1_{in}}^*|...|I_{i_{in}}^* > I_{max}^*)\Big|_{i=1}^{m+1} \tag{20}$$

$$\mathcal{H}^t = (I_{1_{in}}^*|...|I_{i_{in}}^* < I_{min}^*)\Big|_{i=1}^{m+1} \tag{21}$$

where K_{d_p} is the positive discrete reward coefficient for stealthy destabilizing condition, $K_{d_{n1}}$ and $K_{d_{n2}}$ are the negative discrete reward coefficients for non-stealthy destabilizing condition and excessive discordant term detection, respectively. $\mathcal{P}^t$ denotes the presence of significant non-canceling intrusion terms on the overall action outputs from time $t-1$. $\mathcal{E}^t$ and $\mathcal{F}^t$ also represent the acceptable and excessive discordant term detection, and $\mathcal{G}^t$ and $\mathcal{H}^t$ also represent occurrence of outbounded $I_{i_{in}}^*$ when it hits the upper and lower limits, respectively. In terms of the specified threshold values, $I_{CA_{min}}$ is the minimum overall intrusion action term, DE_{max}^t and DE_{min}^t are the upper and lower thresholds for discordant terms, I_{max}^* and I_{min}^* are the upper and lower bounds to the observed neighboring terms I_{in}^*, respectively.

In terms of the threshold value selection for these discrete reward terms, DE_{min}^t is selected based on the normal operation condition and common system disturbances, $I_{CA_{min}}$ is chosen with a trade-off between the desired ramp up/down slope for the destabilizing phenomenon, and the incorporated minimum discordant threshold. While I_{max}^* is chosen with respect to the protective circuit tripping thresholds, I_{min}^* is adjusted with zero for a dc microgrid with purely resistive loads, or with a safe value for a system which has constant power loads. In terms of negative discrete reward coefficients, $K_{d_{n2}} > K_{d_{n1}}$ is chosen to further penalize occurrence of non-stealthy attack conditions. In addition, the positive discrete reward coefficient of K_{d_p} is adjusted with a value with respect to the observed convergence performance.

In order to minimize the intrusion level requirements for the proposed algorithm, it can target the node with the weakest bonding level to its neighboring agents and also have the m action signals for each agent combined to a single action. The only drawback to merging action signals is that it becomes more vulnerable to detection if any of the targeted communication links is disconnected. To ensure a dynamic stealthy destabilizing FDI attack performance under

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPEL.2021.3132028, IEEE Transactions on Power Electronics

this circumstance, the algorithm can be enhanced with a sniffer on the compromised link data transmissions. Using this sniffing feature, any disconnection on the compromised links can be detected and surpassed by switching to other operational incoming communication links connected to the same node.

Algorithm 1 Multi-Agent DDPG to Unveil Cyberattack Susceptibilities

Initialize weights of actor and critic networks, replay buffer $\mathcal{D}$, and target networks.
for episode = 1 to M **do**
  Receive initial process observation at state s_1^t.
  for iteration = 1 to T **do**
    For each agent i, select and execute action a_i^t with respect to policy $\pi(s_i^t|a_i^t)$, receive the reward r_i^t calculated with (14) and transition into state s_i^{t+1}.
    Store tuple $(s_i^t, a_i^t, r_i^t, s_i^{t+1})$ in the $\mathcal{D}$.
    $s_i^t \leftarrow s_i^{t+1}$
    for each agent i = 1 to m **do**
      Randomly select the mini-batch $\mathcal{S}$ from $\mathcal{D}$.
      Set y^j according to (11).
      Update the actor and critic networks with (12) and (10), respectively.
    end for
    Update the target network using (13).
  end for
end for

C. Complementary Cyberattack Detection with Multi-Agent RL DQN

In order to overcome the inefficacy of the discordant scheme on detecting a group of coordinated link FDI attacks, as discovered by the proposed multi-agent DDPG algorithm, a complementary multi-agent DQN FDI detection algorithm is proposed. This algorithm is trained with the recorded dataset from undetected FDI intrusion vectors. This complementary feature is activated if the discordant scheme does not reflect any irregularities and its associated DE terms remain minimized. In this case, the corresponding list of RL agents are set with $\boldsymbol{N} = \{RD_1, ..., RD_n\}$, where RD represents the RL based detection agent operated in parallel with the distributed discordant unit for each of the n nodes. The list of observation signals for each agent is also defined as $\boldsymbol{O_D^t} = \{IE_s^t, I\dot{E}_s^t\}$, where $IE_s^t = \sum_{j=1}^{m} |I_j - I_i|$ is the accumulative distributed current error term and $I\dot{E}_s^t$ denotes its corresponding derivatives for m incoming distributed communication links connected to the node i. The list of discrete actions for each agent is also selected with $\boldsymbol{R_D^t} = \{0, 1\}$, where 1 represents detection of FDI intrusion presence and 0 is signaled out under normal operating condition. It should be also noted that the observations to the DQN agents are only enabled if the corresponding discordant terms are within the normal operating range with some delay to avoid overlapped or false detection.

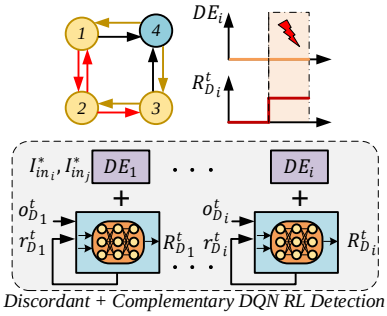


Fig. 4. Proposed complementary multi-agent DQN RL detection scheme to mitigate the detection failure on discordant algorithm.

The reward function r_D^t for each agent is also formulated by (22). Since the proposed RL DQN scheme is designed to operate as a FDI detection scheme, similar to the discordant method, and generate only discrete action signals, the reward function only includes the discrete reward terms. It is basically formed with two negative and one positive reward terms using the detection action signal from the previous time step $R_{D_i}^{t-1}$, FDI presence on any of the incoming distributed signals $F_{D_i}^t$, and their corresponding derivatives as represented by $\dot{R}_{D_i}^{t-1}$ and $\dot{F}_{D_i}^t$, respectively. The first negative reward term ensures minimal delay on intrusion state detection, and the second negative discrete reward term is in charge of minimizing excessive action signal alterations. The third term is devoted to rewarding proper detection of FDIs with minimal delay over one complete episode period. The inclusion of $F_{D_i}^t$ as a part of reward equation not only enables more effective rewarding formulation but also ensures the improved algorithm convergence. It should be also noted that after the training process in offline mode, only the observation signals are fed into the RL agents. The algorithm steps for the proposed complementary multi-agent RL DQN detection scheme are also presented by Algorithm 2.

$$\begin{aligned} r_D^t = &- K_{d_{n1}}(R_{D_i}^{t-1} \oplus F_{D_i}^t) - K_{d_{n2}}|\dot{R}_{D_i}^{t-1}| \\ &+ K_{d_p}(0 \leq \mathcal{B}^t \leq \mathcal{B}_d)u(t - T_p + T_s) \end{aligned} \tag{22}$$

where $K_{d_{n1}}$ and $K_{d_{n2}}$ are the negative reward coefficients for proper detection and action variations, respectively, and K_{d_p} is the positive reward coefficient for desired detection performance over an episode time period of T_p and with the step time of T_s. Considering the impacts associated with each of these reward terms, the reward coefficients should be adjusted in a way that condition $K_{d_{n2}} < K_{d_{n1}} \ll K_{d_p}$ is satisfied. Also, $u(.)$ is the step function and the term $\mathcal{B}^t$ is represented by (23), which implies a desired detection performance over the complete episode period with p steps. $\mathcal{B}_d$ is also an integer constant value chosen based on the desired delay performance on detection and with respect to the observation input delays.

$$\mathcal{B}^t = (\sum_{k=1}^{p}(R_{D_{ik}}^{t-1} \oplus F_{D_{ik}}^{t}) - \sum_{k=1}^{p}\dot{F}_{D_{ik}}^{t}) \quad (23)$$

Algorithm 2 Multi-Agent DQN to Complement Cyberattack Detection

1: Initialize replay buffer $\mathcal{D}$, and action-value function Q with random weights.
2: **for** episode = 1 to M **do**
3: Receive initial process observation at state s_1^t.
4: **for** iteration = 1 to T **do**
5: For each agent i, select and execute action a_i^t with respect to policy $\pi(s_i^t|a_i^t)$, receive the reward r_i^t calculated with (22) and transition into state s_i^{t+1}.
6: Store tuple $(s_i^t, a_i^t, r_i^t, s_i^{t+1})$ in the $\mathcal{D}$.
7: $s_i^t \leftarrow s_i^{t+1}$
8: **for** each agent i = 1 to m **do**
9: Randomly select the mini-batch $\mathcal{S}$ from $\mathcal{D}$.
10: Set y^j according to (11).
11: Perform gradient descent on (10) with (11).
12: **end for**
13: Update the target network using (13).
14: **end for**
15: **end for**

IV. Experimental Results

In order to verify the performance of the proposed multi-agent DDPG RL-based system on discovering the vulnerabilities in the cyberattack detection scheme, and generating stealthy destabilizing FDI intrusions against the discordant algorithm, an autonomous DC microgrid configuration, as previously depicted in Fig. 1, with $n = 4$ power generation units is considered. The system electrical and control parameters for both primary and secondary control layers are also presented in Table. I. The effectiveness of the proposed multi-agent RL DQN on complementing the discordant detection algorithm and mitigating its vulnerability to coordinated FDIs is also experimented. The training process is carried out in the Matlab/Simulink environment, and the algorithm verification is performed using the experimental setup shown in Fig. 5 by means of dSPACE MicroLabBox DS1202, where only control parameters are slightly modified to ensure similar controller performance under both conditions.

TABLE I. Experimental Testbed Parameters

Parameter Sets	Parameter Values
Plant	$R_{12} = R_{23} = R_{34} = 0.5\ \Omega, R_{14} = 0\ \Omega$
Converter	$L_{in} = 0.86\ mH, C_{out} = 1.1\ mF, f_s = 10\ khz, I_{rated} = 32\ A$
Controller	$V_{in} = 48\ V, V_{dc_{ref}} = 60\ V, I_{dc_{ref}} = 0, M_i = 2$ $G_p(s): K_{pV} = 1, K_{iV} = 20, K_{pI} = 2.4, K_{iI} = 10$ $G_s(s): K_p^I = 0.15, K_i^I = 0.06$
Load	$R_1 = R_2 = R_3 = R_4 = 30.6\ \Omega, R_1 = 30.6 \rightarrow 65.7\ \Omega$

Three DDPG RL agents are structured similar to Fig. 3, to enable proper exploration of cyberattacks on the neighboring communication links and generate a group of different coordinated FDI attacks between the agents on nodes 1, 2, and 4. Moreover, for complementing discordant detection units, a DQN RL agent is integrated for each node. The parameters

TABLE II. Hyperparamaters for DDPG RL Agents and DQN RL Agents

Hyperparameters	RL DDPG	RL DQN
Batch Size	512	64
Discount Factor	Actor: 0.9995 Critic: 0.9995	0.99
Learning Rates	Actor: 10^{-4} Critic: $5x10^{-4}$	$1x10^{-3}$
Hidden Layers/Nodes	Actor: 2/2048 Critic: 2/1024	2/512

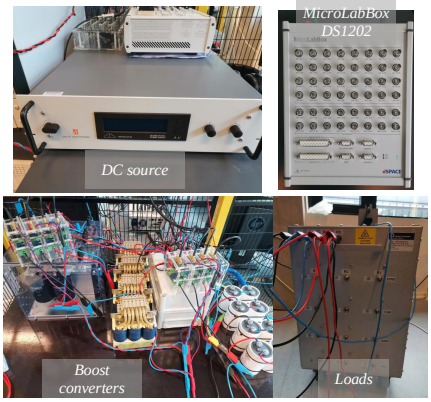


Fig. 5. Experimental Setup

for DDPG RL agents as well as DQN RL agents are also provided in Table II. Using the previously discussed reward function and observation vector for each agent, the multi-agent FDI generation unit and complementary detection unit are trained through the simulation model for a run-time period of 6 seconds for each training episode. Both primary and secondary controllers as well as MA RL units are implemented inside the dSPACE controller with the sampling time of 100 μs. In order to account for the communication delays, the primary to secondary delay term t_{p-s}, and cyberattack output delay term t_{CA} values are set with 5 ms and 40 ms, respectively, and 70-100 ms delay on the secondary to secondary communication link delays t_{s-s}, are considered in the test cases.

Scenario A: The performance of the discordant cyberattack detection algorithm under load switchings and both conventional deceptive and destabilizing FDI attacks on the distributed current signals is shown in Fig. 6, where the associated output voltages, distributed currents and discordant terms are displayed. For this experiment, $t_{s-s} = 100\ ms$ is set and the default delay value for t_{p-s} is maintained. In this case, subsequent load step-up and step-down are first applied at $t = 21\ s$ and $t = 41\ s$, respectively, where a proper convergence among the distributed current signals to $I = 1.55\ A$ and $I = 1.25\ A$ within less than 8 seconds are resulted, as shown in Fig. 6b. A slight dwell on the discordant terms under the load transient conditions in Fig. 6c is also noteworthy, where their convergence rates and peak values are a function of underlying current controllers performance. At

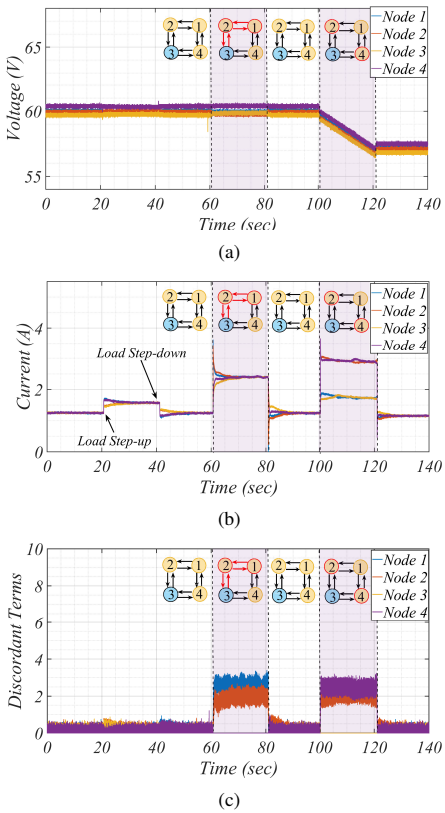


Fig. 6. Performance of the discordant cyberattack detection algorithm under load steps and both conventional deceptive and destabilizing FDI attacks with $t_{s-s} = 100\ ms$.

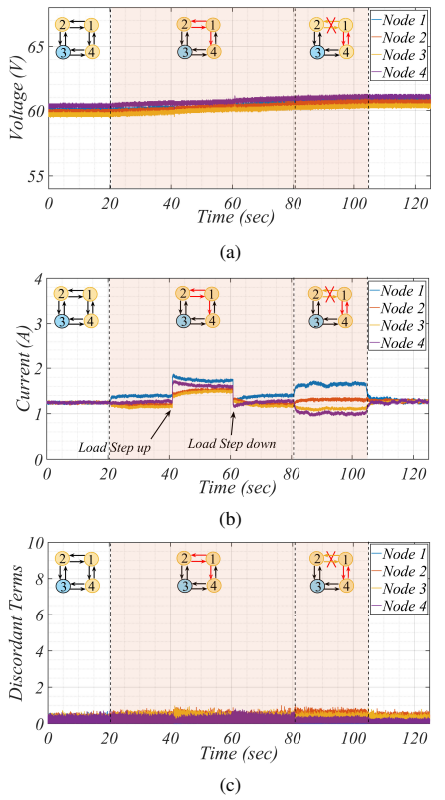


Fig. 7. Performance of the proposed Multi-agent DDPG RL unit on generation of stealthy destabilizing FDI attacks under load steps and sudden compromised link disconnection- two compromised links and $t_{s-s} = 100\ ms$.

$t = 62\ s$, conventional deceptive FDI attacks are introduced into the nodes 1 and 2 with 2.2 A and 2 A, respectively, which target the associated sensors and outgoing distributed terms. The impact of such an attack is reflected in the similar manner to the load step-ups as all current signals converge to a value of 2.5 A, and this attack is properly detected with the significant increase on the associated discordant terms. Effectiveness of the discordant algorithm on identifying the destabilizing attack is also verified where it detects the intrusion vector 2.2 A and 2 A on sensors for agent 2 and 4, which initiated at $t = 100\ s$, with discordant values greater than 2 which is significantly distinctive from its normal operating condition. Destabilizing phenomenon is also further evident for that period with the voltage ramp down to a value of about 57 V, as shown in Fig. 6a ,where it is stopped after removal of attack at $t = 120\ s$.

Scenario B: The effectiveness of the proposed multi-agent RL algorithm on deceiving discordant detection algorithm and generating stealthy destabilizing FDI attacks on this DC microgrid is tested using a specific attack configuration where only two communication links are compromised, as shown in Fig. 7. In this case, $t_{s-s} = 100\ ms$ is applied to the distributed terms and the default delay values for t_{p-s} and t_{CA} are used. While the same initial loading condition is maintained, starting at $t = 20\ s$, the connection link between nodes 1 and 2 is fully compromised and only the incoming signal to node 3 is manipulated. It is observed that despite about 0.35 A deviations on node 1 from the original consensus setpoint and some deviations on current signals for nodes 2 and 3, as shown in Fig. 7b, the associated discordant terms depicted in Fig. 7c fail to properly detect such intrusions. By applying subsequent load step-up and step-down at about $t = 40\ s$ and $t = 60\ s$, respectively, similar convergence performance to the normal operating condition is observed, and the coordinated stealthy attack is remained hidden to the identification algorithm. While sniffing tool is utilized to monitor the availability of connection links between the compromised agents, a sudden communication link disconnection is applied at about $t = 80\ s$. A reaction delay of 60 ms is then considered for re-configuring the intelligent FDI attack to the link compromise between nodes 1 and 3. It is observed that significant deviations up to $0.5A$ are introduced between agents 1, 3, and 4 current signals. However, discordant algorithm is not able to signal

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPEL.2021.3132028, IEEE Transactions on Power Electronics

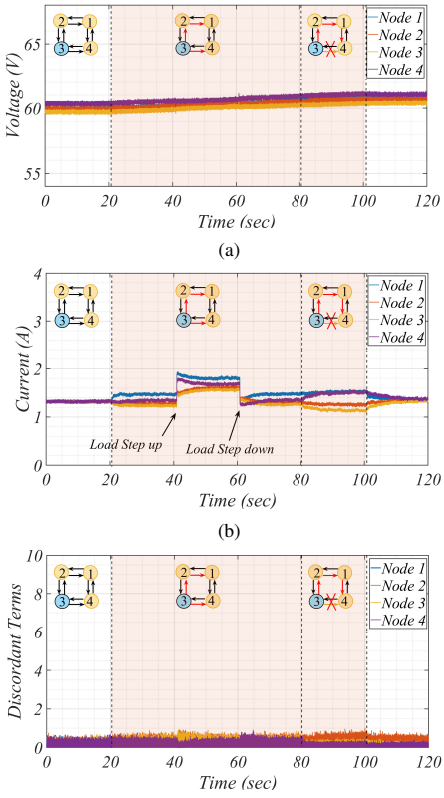


Fig. 8. Performance of the proposed Multi-agent DDPG RL unit on generation of stealthy destabilizing FDI attacks under load steps and sudden compromised link disconnection- only one transmission pathway is impacted on each of the three compromised links and $t_{s-s} = 70\ ms$.

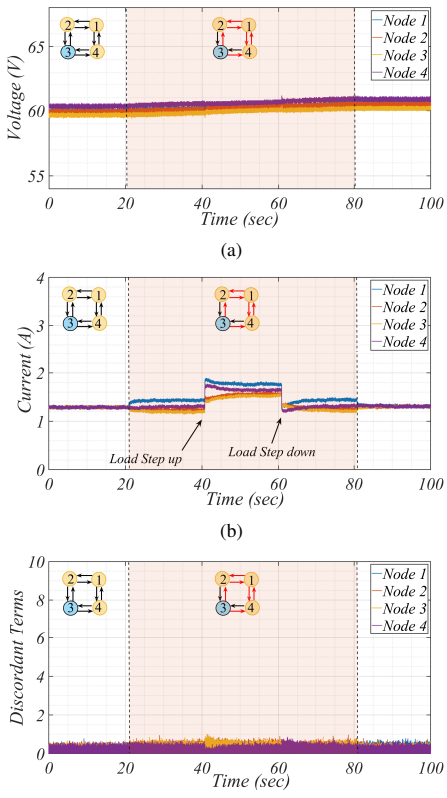


Fig. 9. Performance of the proposed Multi-agent DDPG RL unit on generation of stealthy destabilizing FDI attacks under load steps - two transmission pathways are impacted on each of the three compromised links and $t_{s-s} = 90\ ms$.

out the compromised agents. By keeping all discordant terms to their minimal level values, a voltage ramp-up as a function of interlinking impedances between the nodes is resulted and maintained over the whole RL FDI attack duration, where in this case introduced about $1\ V$ increment on all agents, as depicted in Fig. 7a. This destabilizing condition can deteriorate the regulation performance, and even lead to protective circuits tripping, and erroneous communication link disconnections, especially if it remains over a longer time period.

Scenario C: With a modified RL FDI cyberattack configuration in the distributed control layer with lower transmission delay of $t_{s-s} = 70\ ms$, the attacks are concentrated on three distinctive communication links between the neighboring agents where for each only one communication pathway is impacted, as shown in Fig. 8. It is observed that after applying such a coordinated intrusion at $t = 20\ s$, a similar destabilizing phenomenon occurs where distributed current signals experience deviations from the desired setpoint value, as shown in Fig. 8b, and this effect is maintained after consequent load stepping incidents. At $t = 80\ s$, a sudden link disconnection between nodes 2 and 4 is introduced, where its detection with sniffing tool and re-configuring the attack to the alternative incoming link is applied with the delay of $50\ ms$. Compared with the former attack configuration, lower deviation on distributed term from node 1 is observed which is attributed to its stronger communication bonding under the existing condition. In this case, the other major difference is the negative error introduced on agent 4 with respect to significant positive error in the former test scenario. As a result of introduced deviations, a similar destabilizing voltage ramp-up by about $1V$ over the intrusion period is resulted, as shown in Fig. 8a. From the discordant signals in Fig. 8c, it is also evident that the attacks remained undetected as their values resemble the normal operating conditions.

Scenario D: In this scenario, a more widespread coordinated attack is launched against the three neighboring nodes which impacts three out of four available communication links,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPEL.2021.3132028, IEEE Transactions on Power Electronics

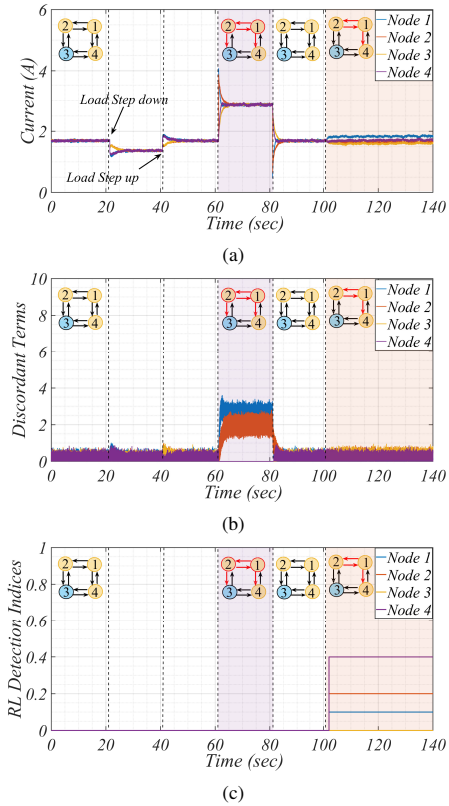


Fig. 10. Performance of the proposed Multi-agent DDPG RL cyberattack generation and multi-agent RL DQN complementary detection units under load steps, $t_{s-s} = 50\ ms$ and $t_{CA} = 80\ ms$.

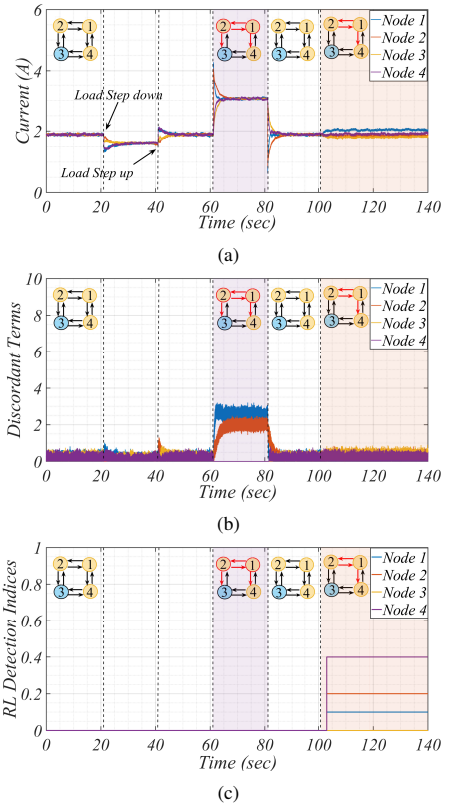


Fig. 11. Performance of the proposed Multi-agent DDPG RL cyberattack generation and multi-agent RL DQN complementary detection units under load steps, $t_{s-s} = 150\ ms$ and $t_{CA} = 80\ ms$.

as depicted in Fig. 9. Unlike the previous coordinated RL attacks which used the merged action signals from the DDPG RL agents, both output actions are incorporated to generate such a stealthy destabilizing attack against the DC microgrid protected with discordant algorithm. It is observed that despite the high level of penetration by the attacker and its persistence for about 60 seconds, the intrusion still remains stealthy as discordant terms does not reflect any distinctive value than their minimal values on the normal operating condition, as depicted in Fig. 9c. It is also observed that such attacks can produce the similar destabilizing impact if it is used either in the merged or independent mode to target at least three incoming communication links for three neighboring agents. In this case, slightly lower voltage ramp up, by about $0.2\ V$, for the destabilizing duration is resulted in Fig. 9a, which is mainly attributed to lower duration of the applied coordinated attacks.

Scenario E: In this case, the performance of the overall combined detection scheme is verified under load steps, conventional deceptive FDI attacks and DDPG RL based stealthy destabilizing FDI attacks. This test scenario is carried out for the cyberattack delay of $80\ ms$ and two different distributed communication delays of $50\ ms$ and $150\ ms$, as depicted in Fig. 10 and Fig. 11, respectively. From the obtained results, it is evident that the discordant method is only capable of detecting the conventional FDI attacks, as applied during 60-80 s, and the corresponding indices remain minimized under load switching and DDPG RL FDIs for both distributed delay conditions. However, the proposed RL DQN detection algorithm properly signals out the non-cooperative nodes within 2-3 seconds after launching stealthy RL attacks. This delay on detection is mainly attributed to the filtered observation signals, communication delays, as well as the chosen $1\ s$ timestep for DQN agents. For enhanced visibility of indices, they are scaled with the corresponding node index and 0.1 scaling factor. Also, it is noteworthy that the performance of the proposed RL DDPG agents are not impacted by the distributed communication delays, where permissible low levels on discordant terms are well maintained during the interval 100-140 for both delay conditions.

V. Conclusions

This paper proposes a multi-agent deep reinforcement learning based algorithm to exploit the vulnerabilities in the existing cyberattack detection methods, which basically provides the foundations for their effective mitigation. The effectiveness of the proposed algorithm is verified by locating the penetrable spots on a sample cyberattack detection algorithm. Using this approach, stealthy destabilizing cyberattacks are launched on the distributed control layer in a DC microgrid protected with the discordant detection algorithm. It is observed that despite the effectiveness of the discordant scheme on detection of the conventional deceptive and destabilizing FDI attacks, it fails to identify more coordinated FDI attacks generated by the proposed scheme. Using the proposed reward function, the training algorithm is reinforced to introduce distributed destabilizing terms into the neighboring communication links in a way that remains hidden to the discordant observers. To overcome the discordant method failure on proper detection of such coordinated stealthy FDIs, a complementary RL DQN detection algorithm is proposed. This hybrid detection approach enables enhancing the reliability of all such index based detection algorithms against the autonomously detected FDI susceptibilities with the aim of reaching a comprehensive cybersecure solution.

Acknowledgment

This work was partially supported by Louisiana Board of Regents under grant number: LEQSF(2021-24)-RD-B-06.

References

[1] A. Jafarian Abianeh and F. Ferdowsi, "Sliding mode control enabled hybrid energy storage system integrated into islanded dc microgrids with pulsing loads," *Sustainable Cities and Society*, 2021.

[2] T. Dragičević, X. Lu, J. C. Vasquez, and J. M. Guerrero, "Dc microgrids—part i: A review of control strategies and stabilization techniques," *IEEE Transactions on Power Electronics*, vol. 31, no. 7, pp. 4876–4891, 2015.

[3] O. A. Beg, T. T. Johnson, and A. Davoudi, "Detection of false-data injection attacks in cyber-physical dc microgrids," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 5, pp. 2693–2703, 2017.

[4] J. Liu, X. Lu, and J. Wang, "Resilience analysis of dc microgrids under denial of service threats," *IEEE Transactions on Power Systems*, vol. 34, no. 4, pp. 3199–3208, 2019.

[5] S. Sahoo, J. C.-H. Peng, S. Mishra, and T. Dragičević, "Distributed screening of hijacking attacks in dc microgrids," *IEEE Transactions on Power Electronics*, vol. 35, no. 7, pp. 7574–7582, 2019.

[6] S. Sahoo, T. Dragičević, and F. Blaabjerg, "Multilayer resilience paradigm against cyber attacks in dc microgrids," *IEEE Transactions on Power Electronics*, vol. 36, no. 3, pp. 2522–2532, 2020.

[7] D. Shi, P. Lin, Y. Wang, C.-C. Chu, Y. Xu, and P. Wang, "Deception attack detection of isolated dc microgrids under consensus-based distributed voltage control architecture," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 11, no. 1, pp. 155–167, 2021.

[8] S. Sahoo, J. C.-H. Peng, A. Devakumar, S. Mishra, and T. Dragičević, "On detection of false data in cooperative dc microgrids—a discordant element approach," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 8, pp. 6562–6571, 2019.

[9] X.-K. Liu, C. Wen, Q. Xu, and Y.-W. Wang, "Resilient control and analysis for dc microgrid system under dos and impulsive fdi attacks," *IEEE Transactions on Smart Grid*, 2021.

[10] A. Cecilia, S. Sahoo, T. Dragičević, R. Costa-Castelló, and F. Blaabjerg, "Detection and mitigation of false data in cooperative dc microgrids with unknown constant power loads," *IEEE Transactions on Power Electronics*, vol. 36, no. 8, pp. 9565–9577, 2021.

[11] M. S. Sadabadi, S. Sahoo, and F. Blaabjerg, "Stability oriented design of cyber attack resilient controllers for cooperative dc microgrids," *IEEE Transactions on Power Electronics*, 2021.

[12] Y. Jiang, Y. Yang, S.-C. Tan, and S. Y. Hui, "Distributed sliding mode observer-based secondary control for dc microgrids under cyber-attacks," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 11, no. 1, pp. 144–154, 2020.

[13] M. R. Habibi, H. R. Baghaee, T. Dragičević, F. Blaabjerg, *et al.*, "Detection of false data injection cyber-attacks in dc microgrids based on recurrent neural networks," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, 2020.

[14] M. R. Habibi, S. Sahoo, S. Rivera, T. Dragičević, and F. Blaabjerg, "Decentralized coordinated cyber-attack detection and mitigation strategy in dc microgrids based on artificial neural networks," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, 2021.

[15] O. A. Beg, L. V. Nguyen, T. T. Johnson, and A. Davoudi, "Signal temporal logic-based attack detection in dc microgrids," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3585–3595, 2018.

[16] S. Jena, N. P. Padhy, and J. M. Guerrero, "Cyber-resilient cooperative control of dc microgrid clusters," *IEEE Systems Journal*, 2021.

[17] S. Sahoo, T. Dragičević, and F. Blaabjerg, "Resilient operation of heterogeneous sources in cooperative dc microgrids," *IEEE Transactions on Power Electronics*, vol. 35, no. 12, pp. 12 601–12 605, 2020.

[18] S. Sahoo, T. Dragicevic, and F. Blaabjerg, "An event-driven resilient control strategy for dc microgrids," *IEEE Transactions on Power Electronics*, vol. 35, no. 12, pp. 13 714–13 724, 2020.

[19] T. T. Nguyen and V. J. Reddi, "Deep reinforcement learning for cyber security," *arXiv preprint arXiv:1906.05799*, 2019.

[20] D. Zhang, X. Han, and C. Deng, "Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE Journal of Power and Energy Systems*, vol. 4, no. 3, pp. 362–370, 2018.

[21] M. N. Kurt, O. Ogundijo, C. Li, and X. Wang, "Online cyber-attack detection in smart grid: A reinforcement learning approach," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5174–5185, 2018.

[22] W. Jiang, W. Yang, J. Zhou, W. Ding, Y. Luo, and Y. Liu, "Reinforcement learning based detection for state estimation under false data injection," *IEEE Access*, 2021.

[23] C. Neal, H. Dagdougui, A. Lodi, and J. M. Fernandez, "Reinforcement learning based penetration testing of a microgrid control algorithm," in *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2021, pp. 0038–0044.

[24] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, *et al.*, "Value-decomposition networks for cooperative multi-agent learning," *arXiv preprint arXiv:1706.05296*, 2017.

[25] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *arXiv preprint arXiv:1706.02275*, 2017.

Ali Jafarian Abianeh (S'19) received his M. Eng. degree in Electrical Engineering from University of Malaya, Kuala Lumpur, Malaysia, in 2010. He is currently working toward his Ph.D. degree at Department of Electrical and Computer Engineering at University of Louisiana at Lafayette, USA. He developed some solid professional expertise through several years of working in the industry as a power electronics engineer with the main focus on electric motor drives, and grid-tied power converters. His current research interests include application of advanced control algorithms and machine learning techniques to AC/DC microgrids, power converters, motor drive control, distributed control, fault tolerant control algorithms and cybersecurity.

Yihao Wan (S'21) received the B.S. degree in electrical engineering from the Wuhan University of Technology, Wuhan, China, in 2017 and the M.S. degree in electrical engineering from Chongqing University, Chongqing, China, in 2020. He is currently pursuing the Ph.D. degree in electrical engineering with Technical University of Denmark. His current research interests include application of artificial intelligence in power electronics and power systems.

Farzad Ferdowsi (S'13–M'17-SM'20) is an Assistant Professor at University of Louisiana at Lafayette. He is with the Electrical and Computer Engineering Dept. Prior to joining UL Lafayette, Farzad worked as a research associate at the Center for Energy Studies at Louisiana State University. He received his Ph.D. from Florida State University in 2016. His research interests include power system stability and control and application of power electronic-based components in power systems.

Nenad Mijatovic after obtaining his Dipl.Ing. education in Electrical Power Engineering at University of Belgrade, Serbia in 2007, was enrolled as a doctoral candidate at Technical University of Denmark. He received his Ph.D. degree from Technical University of Denmark for his work on technical feasibility of novel machines and drives for wind industry. Upon completion of his PhD, he continued work within the field of wind turbine direct-drive concepts as an Industrial PostDoc. Dr. N. Mijatovic currently holds position of Associate Professor at Technical University of Denmark where he is in charge of managing research projects and education related to the field of electrical machines and drives, power electronic converters, motion control, application of energy storage and general applications of low frequency electromagnetism and large scale application of superconductivity with main focus on emerging eMobility and renewable energy generation.

He is a member of IEEE since 2008 and senior member of IEEE since 2018 and his field of interest and research includes novel electrical machine drives/actuator designs, operation, control and diagnostic of electromagnetic assemblies, advance control of drives and grid connected power electronics, energy storage and eMobility.

Tomislav Dragičević (S'09-M'13-SM'17) received the M.Sc. and the industrial Ph.D. degrees in Electrical Engineering from the Faculty of Electrical Engineering, University of Zagreb, Croatia, in 2009 and 2013, respectively. From 2013 until 2016 he has been a Postdoctoral researcher at Aalborg University, Denmark. From 2016 until 2020 he was an Associate Professor at Aalborg University, Denmark. From 2020 he is a Professor at the Technical University of Denmark. He made a guest professor stay at Nottingham University, UK during spring/summer of 2018. His research interest is application of advanced control, optimization and artificial intelligence inspired techniques to provide innovative and effective solutions to emerging challenges in design, control and cyber-security of power electronics intensive electrical distributions systems and microgrids. He has authored and co-authored more than 250 technical publications (more than 120 of them are published in international journals, mostly in IEEE), 8 book chapters and a book in the field.

He serves as an Associate Editor in the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, in IEEE TRANSACTIONS ON POWER ELECTRONICS, in IEEE Emerging and Selected Topics in Power Electronics and in IEEE Industrial Electronics Magazine. Dr. Dragičević is a recipient of the Končar prize for the best industrial PhD thesis in Croatia, a Robert Mayer Energy Conservation award, and he is a winner of an Alexander von Humboldt fellowship for experienced researchers.

[J5] Optimal Day-ahead Scheduling of Fast EV Charging Station With Multi-stage Battery Degradation Model

Authors:
Yihao Wan, Daniel Gebbran, Ramadhani Kurniawan Subroto, Tomislav Dragicevic

Submitted to:
IEEE Transactions on Energy Conversion

Status:
Published.
Digital Object Identifier: 10.1109/TEC.2023.3335661

Optimal Day-ahead Scheduling of Fast EV Charging Station With Multi-stage Battery Degradation Model

Wan, Yihao; Gebbran, Daniel; Subroto, Ramadhani Kurniawan; Dragicevic, Tomislav

Published in:
IEEE Transactions on Energy Conversion

Link to article, DOI:
10.1109/TEC.2023.3335661

Publication date:
2023

Document Version
Peer reviewed version

Optimal Day-ahead Scheduling of Fast EV Charging Station With Multi-stage Battery Degradation Model

Yihao Wan, *Student Member, IEEE,* Daniel Gebbran, *Member, IEEE,* Ramadhani Kurniawan Subroto, *Member, IEEE,* and Tomislav Dragičević, *Senior Member, IEEE*

***Abstract*—The paper proposes a day-ahead scheduling framework with a novel multi-stage battery degradation modeling method for an electric vehicle (EV) fast charging station (FCS) equipped with a battery energy storage system (BESS). Unlike previous studies, which employ a single battery degradation model to represent the aging process, this paper proposes a novel multi-stage battery degradation modeling method to capture the degradation process across the whole lifespan accurately. Subsequently, the multi-stage model is explicitly integrated into the proposed adaptive optimization framework in a computationally tractable way, thus having important practical implications in the field. The paper provides case studies to demonstrate the effectiveness of the proposed modeling method on a selected cycle aging model in reducing the operation cost of FCS with BESS operating in different stages. As a result, the overall operation cost with the multi-stage model is around 2.1% on average lower than the single-stage model counterpart. In addition, results show that with the increasing number of divided stages, the model error decreases and becomes stable, while the reduced operation cost compared with the single-stage model increases and saturates. Finally, we apply the multi-stage framework considering other conventional degradation models to show the superiority of the proposed method.**

***Index Terms*—Battery degradation, energy storage, fast charging station, optimization, operation cost.**

I. Introduction

With the increasing transportation electrification to reduce Greenhouse Gas (GHG) emissions, electric vehicles (EVs) are gaining a significant market share [1]. Although EVs are becoming more popular, one of the major bottlenecks for the large-scale replacement of Internal Combustion Engine (ICE) vehicles is the lack of fast-charging infrastructures, particularly on the highways between cities and rural districts. Fast charging stations (FCSs) are important infrastructures that can provide a quick recharging experience for EVs similar to the experience of refueling ICE vehicles [2]. However, due to the impulsive nature of the load profile in FCSs, they may cause adverse impacts on the stability of the grid, especially for those installed in rural areas with weak grid connections [3], [4]. In addition, the installation of FCSs may require upgrading the electrical infrastructure (e.g., installation of new transformers and power lines), resulting in high installation costs for both the charge point operator (CPO) and the distribution system operator (DSO), who usually transfers its share of the costs to the CPO in the form of grid-connection fee [5].

To partially mitigate the above issues, battery energy storage systems (BESSs) can be integrated into FCSs, acting as a buffer between the grid and the EVs [6], [7]. BESSs can not only partially mitigate the high cost involving grid connection fees but can also provide premium grid services such as frequency containment reserve (FCR) [5]. In addition, the BESS can help reduce the utility charges for the FCS by doing energy price arbitrage [8].

One of the most commonly researched applications of BESSs is energy arbitrage (i.e., charging at low prices during off-peak intervals and selling the stored energy during the peak load at high price intervals), where the operators take advantage of the electricity spot price differences. The key challenge with arbitrage is obtaining an economic dispatch strategy to achieve a trade-off among the arbitrage revenue, battery degradation, and energy consumption costs for minimal operation cost. Therefore, battery usage should also be explicitly considered in the optimization problem because more usage implies more losses and degradation [9].

A. Background

To formulate the battery degradation cost in the scheduling framework, the simplest way is introducing operational factors of battery into the objective function or constraints such as the number of cycles [10], state-of-charge (SoC) [11], depth-of-discharge (DoD) [12], current rate [13], charging/discharging power [14], [15], etc. Although those methods make the scheduling problem possible to be solved with common mathematical programming methods such as linear programming (LP) or quadratic programming (QP). However, since battery degradation behavior is highly non-linear [16], these terms do not explicitly represent the battery capacity loss throughout its operational lifetime, which may lead to overuse or underuse of the battery. In addition, the optimal scheduling is highly dependent on the selected weighting coefficients for the introduced operational factors, which are not straightforward to select [17], [18]. Therefore, though using the simple term to account for the battery usage reduces the computational complexity of the operational planning problem, it doesn't explicitly represent the actual battery degradation which would result in suboptimal operations and incur higher costs in the long term.

To accurately account for battery usage in scheduling problems, different degradation models have been previously proposed. In our previous work [19], a DoD-based degradation model is implemented by discretizing the SoC of the battery during operation, which reduced the small charge/discharge cycles, thus reducing the battery capacity loss while maintaining near optimal revenue from grid services. However, the battery charging/discharging current also contributes to the degradation,

Yihao Wan, Ramadhani Kurniawan Subroto and Tomislav Dragičević are with the Electrical Engineering Department of Technical University of Denmark, Copenhagen, Denmark; Daniel Gebbran is with Equilibrium Energy, California, USA (e-mails: wanyh@dtu.dk, daniel.gebbran@equilibriumenergy.com, rkusu@dtu.dk, tomdr@dtu.dk).

and the same cycle depth for low current and high current could be induced if the charge time is sufficient. In [20], a combined factor-based battery degradation model based on different stress-based models is proposed, while the model accuracy is only validated in the early aging stage. The optimal operation of BESS based on a multi-factor battery cycling model is proposed in [21], which employs the rainflow counting algorithm (RCA) for capacity loss estimation. RCA algorithm is widely used in the context of fatigue analysis and damage estimation [22]. However, due to the lack of an analytical formula for RCA, it is hard to implement it explicitly in an optimization problem [23]. Different ways of solving this are proposed, such as using problem-specific solvers for the nonlinear RCA term [24], approximation [25], [26], etc.

In any case, all the aforementioned methods are based on a single-stage battery degradation model, considering only a single degradation manner across the whole battery lifetime, which is usually not realistic due to the nonlinear and complex battery degradation process. The degradation behavior under various stressing conditions may experience multiple patterns over the lifespan [27].

Given the complex aging mechanism across the lifespan under diverse stress situations, the degradation pattern is constant within a single stage but varies among other stages. Therefore, a single battery degradation model cannot represent the dynamic and complex aging process over the entire battery lifetime. Some attempt to model the battery aging process in different degradation stages. The multi-stage battery degradation modeling method in [27] describes the varying charging stress at different battery states. However, it is not employed in practical scheduling problems, and the model also lacks flexibility or adaptability for more stress variables. In [28], a multi-stage model with cycling discrepancy learning for SoH estimation is proposed. It estimates the degrading trend in different stages. Similarly, a one-shot battery degradation trajectory prediction method based on deep learning is proposed [29]. However, the two methods are used to project the cycle life, and the stress conditions are not considered. Thus, they cannot be applied explicitly to operational planning problems. In [30], the aging mechanism and SoH prediction across the total lifespan are investigated based on three stages with a neural network (NN). However, the model is developed based on electrochemical impedance spectroscopy (EIS), which cannot be directly employed due to the unavailability of detailed cell conditions during the real-time planning stages. Moreover, the scheduling problem with an NN-based degradation model is highly nonlinear and nonconvex, which makes it difficult to solve. Another work in [31] proposes to estimate the battery SoH and remaining useful life (RUL) at different aging states with a linear aging model based on a moving window. Nonetheless, it depends on the extraction and fitting of the health indicator. The model is not explicitly related to the stress variables but to the cycling patterns under specific conditions, which is not applicable to operational planning problems. In summary, the works above reveal that battery degradation patterns change in different battery states. Though they are proposed to estimate the battery SoH or RUL at different stages, most models are not feasible for operational planning problems as the battery degradation is not explicitly formulated with different stress variables.

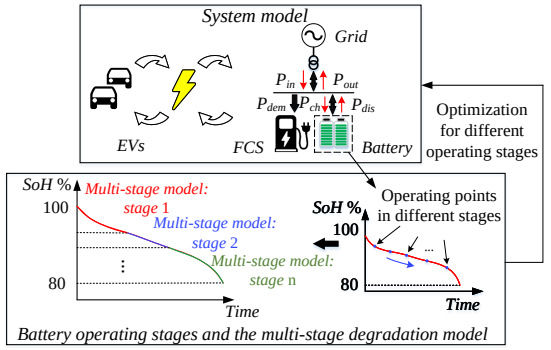


Fig. 1: Scheduling framework with the proposed multi-stage degradation model.

B. Literature gap

In general, the above-mentioned methods utilize a single-stage battery degradation model for the scheduling problem throughout the battery lifespan. The degradation pattern within a single stage would change and differ from other stages due to the complicated and coupling aging mechanism, resulting in a nonlinear battery aging process. As a result, no single model can accurately capture battery degradation across its entire lifespan. In addition, the scheduling strategy should also be adaptively changed according to the varying degradation patterns within different operating stages. Operating a BESS without acknowledging this fact results in suboptimal operations, leading to an increase in overall operation costs.

C. Contributions

This paper aims to bridge the aforementioned gaps by proposing a multi-stage battery degradation modeling method and integrating the multi-stage models explicitly into an optimization framework, as shown in Fig. 1. In particular, we focus on the day-ahead scheduling optimization with the multi-stage battery degradation model, where the battery cycle degradation process should be properly modeled. The proposed multi-stage modeling method is demonstrated on a generic cycle aging model selected in [32]. Different stressing factors are considered in the degradation model, developed respectively for different operational stages. The models are then adaptively implemented in the scheduling model with respect to the battery operating stages. The main contributions of the paper are summarized below:

- A novel multi-stage battery degradation modeling method. The proposed multi-stage degradation modeling method can accurately project the nonlinear battery degradation process throughout its lifespan by capturing the varying aging patterns. In particular, by segmenting the battery lifetime into several stages according to the SoH ranges, the degradation model considering the stress factors is

developed for each stage, respectively. The proposed approach is applicable to various conventional single-stage battery degradation models.

- An adaptive scheduling framework with the multi-stage degradation model. A novel optimization framework for the day-ahead scheduling problem is proposed by explicitly accounting for the multi-stage battery degradation model in a computationally tractable way for real-time implementation. The effectiveness of the proposed optimization framework is demonstrated with case studies by adaptively integrating the multi-stage battery degradation model into the scheduling model at different stages, while the computational burden is explicitly measured to prove the feasibility of real-time implementation.
- Analysis regarding the number of divided stages and validations on various models. The performance of the proposed multi-stage modeling method is further investigated by increasing the number of divided stages. The reduced operation cost with the proposed optimization framework based on different multi-stage degradation models regarding the single-stage model counterpart is compared. Moreover, the multi-stage framework is applied to other conventional single-stage degradation models and compared with a multi-stage modeling approach to show the superiority of the proposed method.

The rest of the paper is structured as follows. Section II introduces the proposed multi-stage modeling concept based on a specific degradation model. Section III introduces the mathematical modeling of the FCS with a BESS, combining the scheduling model of the system with the proposed multi-stage battery degradation model. Section IV compares the scheduling results of the single- and multi-stage battery degradation model based on the case studies of example days for different operating stages. Section V concludes the paper.

II. Modeling of Battery Degradation

To formulate the aging cost for the scheduling problems, the battery degradation process consisting of calendar and cycle aging should be properly modeled. However, the study does not consider battery calendar aging due to its independence from operations and negligible influence on short-term operational planning problems. Therefore, the battery cycle degradation model is developed for evaluating the capacity loss during operations in this section. To demonstrate the proposed multi-stage degradation modeling framework, a generic cycle life model for lithium-ion batteries from [32] is employed as the fundamental model. In addition, to identify the parameters of the models, battery experimental cycling test data in [33] is utilized, which provides battery cycle aging test data under various conditions (e.g., different DoD levels, current rates, and temperatures).

A. Single-stage battery degradation model

The degradation model considers multiple stress factors, DoD, current rate, and temperature. Moreover, charging and discharging currents are assumed to contribute equally to capacity loss. The temperature stress is formulated with Arrhenius equation [34]. In addition, according to Wöhler's approximation and Miner's rule, the battery degradation per cycle is calculated as the stress amplitude subjected to the peak stress the material can sustain. By analogy, based on the assumption of the same degrading contribution for charging and discharging current, the combined stress factor model is formulated as [34], [35]

$$\theta(t) = \theta_{DoD}(t) \cdot \theta_{I_{ch/dis}}(t) \cdot \theta_T(t), \tag{1}$$

$$\theta_{DoD}(t) = (\frac{DoD(t)}{DoD_{ref}})^{\frac{1}{\alpha}}, \tag{2}$$

$$\theta_{I_{ch/dis}}(t) = (\frac{I_{ch/dis}(t)}{I_{ref}})^{\frac{1}{\beta}}, \tag{3}$$

$$\theta_T(t) = exp\left[-\psi\left(\frac{1}{T_a(t)} - \frac{1}{T_{ref}}\right)\right] \tag{4}$$

where θ represents the combined stressing factor model, α and β are the stress exponents for DoD and current rate, respectively. ψ is the Arrhenius rate constant, DoD_{ref} and I_{ref} are the stress amplitude for the reference DoD and reference current, T_a and T_{ref} are the ambient and reference temperature respectively.

The maximum number of cycles to EoL regarding the stress condition is expressed as

$$\begin{aligned} N_c(t) = \frac{N_{cref}}{\theta(t)} =& N_{cref} \cdot \left(\frac{DoD(t)}{DoD_{ref}}\right)^{-\frac{1}{\alpha}} \cdot \left(\frac{I_{ch/dis}(t)}{I_{ref}}\right)^{-\frac{1}{\beta}} \\ & \cdot exp\left[-\psi\left(\frac{1}{T_{ref}} - \frac{1}{T_a(t)}\right)\right], \end{aligned} \tag{5}$$

where N_{cref} is the number of cycles to EoL for the reference cycle condition, N_c is the number of cycles to EoL for the resultant stress factors.

The battery capacity degradation over its lifetime, considering the non-linear aging characteristics, is expressed as

$$Q(t) = Q_{BoL} - \varphi(t)^{\xi} \cdot (Q_{BoL} - Q_{EoL}), \tag{6}$$

where Q denotes the battery capacity, Q_{BoL} and Q_{EoL} denote the battery capacity at BoL and EoL, respectively. Q_{EoL} is assumed to be 80% of Q_{BoL}. φ is the aging index, representing the contribution of the cycles to the aging of the battery, while ξ is the aging exponent of capacity, expressed below.

$$\xi = \frac{ln(\frac{Q_{BoL}-Q_5}{Q_{BoL}-Q_{EoL}})}{ln(\frac{N_{ref}}{N_{cref}})}, \tag{7}$$

where Q_5 denotes 5% capacity losses of Q_{BoL}, N_{ref} is the number of cycles to Q_5 for reference testing condition.

The model parameters are estimated below.

$$N_{ci} = \frac{N_{cref} \cdot N_i}{N_{ref}}, \tag{8}$$

where N_{ci} and N_i denote the number of cycles to EoL and Q_5 in cycling test condition $i \in \{2, ..., n\}$ for parameters identification. The stress exponents are formulated below.

$$\alpha = -\frac{ln(\frac{DoD_i}{DoD_{ref}})}{ln(\frac{N_{ci}}{N_{cref}})}, \tag{9}$$

$$\beta = -\frac{ln(\frac{I_i}{I_{ref}})}{ln(\frac{N_{ci}}{N_{cref}})}, \tag{10}$$

$$\psi = \frac{ln(\frac{N_{ci}}{N_{cj}})}{\frac{1}{T_i} - \frac{1}{T_{ref}}}. \quad (11)$$

The test conditions for identifying the model parameters are listed in Table I. Finally, we assign the identified parameters from Table II into (1)-(3), and the single-stage battery degradation model is obtained.

TABLE I: Cycling conditions for parameter identification.

Conditions	1*	2	3	4
DoD	100%	60%	100%	100%
C-rate	0.5C	0.5C	2C	2C
Temperature	25°C	25°C	25°C	35°C

* Reference condition

TABLE II: Parameter identification for single-stage battery degradation model.

Inputs for parameters identification				
N_1	N_2	N_3	N_4	N_{c1}
22	256	18	23	513
Model parameters identification				
N_{cref}	α	β	ψ	ξ
513	0.2081	6.9083	-21.4482	0.4402

B. Multi-stage battery degradation model

The multi-stage battery degradation model is developed based on the experimental data of battery degradation throughout its lifetime. In particular, as the degrading rate changes adaptively in different states, the multi-stage degradation modeling method allows the model to capture the different degradation patterns under various stress conditions within different SoH ranges of BESS. This effect is depicted in the bottom left image in Fig. 1. In this paper, the battery degrading process until it reaches EoL with 80% SoH is evenly divided into three stages (i.e., stage 1: 100%-93.3%, stage 2: 93.3%-86.6% and stage 3: 86.6%-80% SoH), where the battery degradation model is parameterized respectively. More divided stages increase the accuracy of the degradation model and more economical operations. The performance of the model with the different number of divided stages will be discussed in the following sections in terms of operation costs with the scheduling results and model estimation error.

In particular, to identify the model parameters for each stage, in the same way presented in Section II.A, the experimental test data corresponding to each stage (i.e., different SoH range), is employed [36]. The model parameters for each stage are listed in Table III-V.

To validate the performance of the proposed multi-stage battery degradation model, additional cycling test datasets that are different from those used for parameter identification are employed. Table VI shows the root-mean-square error (RMSE) of the two models based on the experimental data during different battery operating stages when the battery is cycled at 60% DoD, 3C current rate and 25 °C, and 100% DoD, 1C current rate and 15 °C, respectively. Despite the close performance between the two models in the initial degradation stage, the multi-stage battery degradation model outperforms the single battery degradation model in terms of estimation accuracy for all the stages across the battery lifespan.

It is noteworthy that the combined stress model is parameterized with experimental data under specific cycling conditions. The formulated model is expected to yield reasonable evaluations of battery degradation under different stressing conditions by providing general trends of battery degradation [32] (e.g., the highest current rate condition in the available experimental data used for parameterizing and validating the model is 3 C, and it is expected to evaluate the impact of current rates higher than 3 C on battery degradation. The validation of the current stress model can be found in the Appendix.).

TABLE III: Multi-stage battery degradation model: Stage 1.

Inputs for parameters identification				
N_1	N_2	N_3	N_4	N_{c1}
36	378	34	39	513
Model parameters identification				
N_{cref}	α	β	ψ	ξ
513	0.2172	24.2535	-12.0051	0.3952

TABLE IV: Multi-stage battery degradation model: Stage 2.

Inputs for parameters identification				
N_1	N_2	N_3	N_4	N_{c1}
231	1586	201	280	513
Model parameters identification				
N_{cref}	α	β	ψ	ξ
513	0.2652	9.9653	-29.0049	0.4470

TABLE V: Multi-stage battery degradation model: Stage 3.

Inputs for parameters identification				
N_1	N_2	N_3	N_4	N_{c1}
513	3628	562	727	513
Model parameters identification				
N_{cref}	α	β	ψ	ξ
513	0.2611	-15.1963	-22.5247	0.5066

TABLE VI: Model performance under different tests at different stages.

Tests	Stages	Single-stage model	Multi-stage model
Test 5	Stage 1	0.78%	0.50%
	Stage 2	2.21%	0.42%
	Stage 3	3.15%	1.10%
Test 6	Stage 1	1.00%	0.78%
	Stage 2	2.37%	0.75%
	Stage 3	2.86%	1.28%

Remark 1. *The single-stage degradation model only accounts for a single degradation pattern, which differs between different*

stages, resulting in increased degradation estimation error. The multi-stage model is built based on the experimental data of battery aging from BoL to EoL by separating the lifespan into different stages. Therefore, the multi-stage model could be considered a piecewise linearization of the real battery degradation process. In this way, the multi-stage battery degradation model can capture multiple degradation patterns at different stages so that the model accuracy is improved. Afterward, the models built for each stage could be adaptively integrated into the scheduling model to achieve economic dispatch strategies throughout the battery lifetime, which will be elaborated on in the following sections.

III. System Modeling and Problem Formulation

In this work, we have considered a model determining the day-ahead optimal dispatch strategy for the FCS system. The demand profile from the electric vehicles is obtained through load prediction, and the electricity prices are known in advance from the supplier. Battery usage is taken into consideration in the scheduling model by incorporating the battery degradation model as a penalty term, with a cost coefficient converted into a battery usage cost. The optimal scheduling strategy for specific loads, prices and ambient temperature during a day for balancing the energy arbitrage revenue, battery usage, and energy consumption cost could be obtained. The formulation of the optimization problem is as below.

A. Objective function

The objective function (12) is formulated to optimize the dispatch strategies, P_t^{ch} and P_t^{dis} at each discrete time step t, within a finite time horizon (i.e., $t \in \mathcal{T} = \{1, 2, ..., T\}$) for minimum daily operation cost, achieving the best trade-off among energy arbitrage, load supply and battery degradation cost. In equation (12), the first term yields the power supply cost, which could be minimized by exploiting temporal price differentials on the electricity spot market. Meanwhile, the second term accounts for the cost associated with battery degradation. The objective function which reflects this is expressed below

$$\min_{P_t^{ch}, P_t^{dis}} \sum_{t=1}^{T} (P_t^{in} - P_t^{out}) \cdot p_t + \lambda_k \cdot C_t, \tag{12}$$

where P_t^{in} and P_t^{out} are the energy flow between the grid and FCS, P_t^{ch} and P_t^{dis} are the charging and discharging power commands for the BESS, p_t is the time-of-use (ToU) electricity spot market price, C_t denotes the battery capacity losses, and a cost coefficient λ_k is assigned. In addition, the cost coefficient λ_k represents the replacement cost of the battery, assigned with a constant value across the battery's operational lifespan. Therefore, the objective function aims to minimize the overall operation cost of the system while considering the battery degradation.

The scheduling problem is solved concerning the battery charging power P_t^{ch} and discharging power P_t^{dis}, the energy flow P_t^{in} (grid to the FCS) and P_t^{out} (FCS to the grid) between the grid and system, the operation strategy of the battery DoD_t and $I_t^{ch/dis}$, and the resultant capacity losses C_t. The distinction between the power terms is also depicted in Fig. 1.

B. Operation constraints

1) Energy balance: The energy balance is achieved by utilizing the BESS as a buffer between the grid and FCS, which is formulated as

$$P_t^{in} - P_t^{out} = P_t^{ch} - P_t^{dis} + P_{dem}, \tag{13}$$

where P_{dem} denotes the load demand of the vehicles.

2) Battery operation: The evolution of the battery during the operation is formulated in (14). In addition, to ensure the state-of-charge (SoC) at the end of the time horizon equals an expected value SoC_{end}, constraint (15) is introduced. Moreover, the SoC has a safe battery operating range. Based on the above analysis, the battery operation constraints are shown below

$$SoC_t = SoC_{t-1} + \frac{\tau}{E_{bat}} (\eta_{ch} \cdot P_t^{ch} - \frac{P_t^{dis}}{\eta_{dis}}), \tag{14}$$

$$SoC_{t=T} = SoC_{end}, \tag{15}$$

$$SoC_{min} \leq SoC_t \leq SoC_{max}, \tag{16}$$

where SoC_t denotes the SoC of battery at time t, τ denotes the time step, E_{bat} is the capacity of battery, η_{ch} and η_{dis} are the charging and discharging efficiency respectively, SoC_{min} and SoC_{max} are the SoC range for battery operation.

3) Power limits: During the operation, the charging and discharging power is limited within the battery power ratings, as shown below:

$$0 \leq P_t^{ch} \leq P_{max}, \tag{17}$$

$$0 \leq P_t^{dis} \leq P_{max}, \tag{18}$$

where P_{max} is the maximum battery charging/discharging power.

In addition, the power transmission between the FCS and the grid should be restricted to ensure stability

$$-P_{gmax} \leq P_t^{in} - P_t^{out} \leq P_{gmax}, \tag{19}$$

where P_{gmax} is maximum energy flow into the FCS.

C. Cost for battery degradation

The real-time implementation of the degradation model is important to consider battery degradation in the scheduling problem explicitly. The aging index represents the contribution of the cycles to the aging of the battery. As the battery degradation model could be considered discretized over a finite time horizon, the battery degradation could be formulated by the cumulative aging index over a certain period at different DoD levels, current rates, and temperatures, which is expressed below

$$Q_{loss} = \sum_{t=1}^{T} \frac{\theta_t}{N_{cref}} = \sum_{t=1}^{T} \frac{1}{N_{cref}} \cdot (\frac{DoD_t}{DoD_{ref}})^{\frac{1}{\alpha}} \cdot (\frac{I_t^{ch/dis}}{I_{ref}})^{\frac{1}{\beta}} \cdot \cdot exp\left[-\psi\left(\frac{1}{T_a} - \frac{1}{T_{ref}}\right)\right], \tag{20}$$

In addition, an incomplete half cycle or full cycle may happen in each time step and the whole future optimization window, as the DoD_{ref} = 100%, $I_{ref} = 0.5C$, and $T_{ref} = 25$°C, based on the approximation in [37], replacing the C_t in the objective function (12) with the battery degrading equation as below

$$C_t = \sum_{t=1}^{T} \frac{(\triangle DoD_t)^{\frac{1}{\alpha}} \cdot (2I_t^{ch/dis})^{\frac{1}{\beta}} \cdot exp\left[-\psi\left(\frac{1}{T_a} - \frac{1}{25}\right)\right]}{N_{cref}}. \tag{21}$$

In general, by replacing the battery usage term C_t in equation (21) with the corresponding model parameters, the scheduling model incorporating the single-stage and multi-stage degradation models are built respectively [1].

Remark 2. *The proposed multi-stage degradation modeling method is independent of various empirical/semi-empirical degradation models. The proposed optimization framework with the proposed multi-stage battery degradation model concept can still be employed in the same way for other battery degradation models, which could be integrated to improve the economic dispatch strategies throughout the battery lifespan.*

IV. Case Studies

To validate the performance of the proposed multi-stage battery degradation model, the proposed framework is implemented in an FCS operation problem and compared with the conventional single-stage battery degradation counterparts in different battery operating stages. Specifically, for different case studies (i.e., different battery operating stages), both the single-stage and the corresponding multi-stage battery degradation models are implemented in the scheduling framework as presented in Section III by replacing the battery cost penalty term with the corresponding model. Meanwhile, due to the superior performance of the multi-stage degradation model, the actual battery degradation with the single model is evaluated with the multi-stage model for each stage and utilized to calculate the actual operation cost. Assuming the battery operates in three stages for the three example days across a year, as shown in Fig. 2, the optimal scheduling strategies are obtained, which are compared respectively with the single degradation model counterparts. The optimization problem is implemented in Python with Pyomo as a modeling interface [41] and solved by IPOPT [42].

The nominal energy of the battery is E_{bat} = 1 MWh, and the battery operation range is within SoC_{min} = 0 and SoC_{max} = 1, the initial and ending SoC is 0.1, the round trip charge/discharge efficiency is set as $\eta_{ch} = \eta_{dis}$ = 0.92 and the cost coefficient λ_k = 3560 DKK/kWh (478.55 EUR/kWh). The power limit P_{max} is fixed to be 1 MW, and P_{gmax} is set to be 10 MW. Assume the battery operation horizon is for one day (24-hour period) and the time interval is 30 min, thus T = 48 and yields 96 decision variables. To compare the proposed multi-stage degradation model with the single-stage degradation model, numerical results corresponding to the optimal dispatch strategies for each case study are presented in Table VIII. In addition, Fig. 6 presents the cumulative battery capacity degradation at different time points throughout the day.

A. Case 1: Operating stage 1

Assume the battery operating point of the selected day is in stage 1, where the degradation model parameters for the stage

[1]We use a linear approximation for calculating the DoD term as we did before [19]. Incorporating the RCA for cycle depth $\triangle DoD$ in an optimization problem [23], [25] is still an open research question, and our approach is similar to other works employing a DoD linearization method [38]–[40].

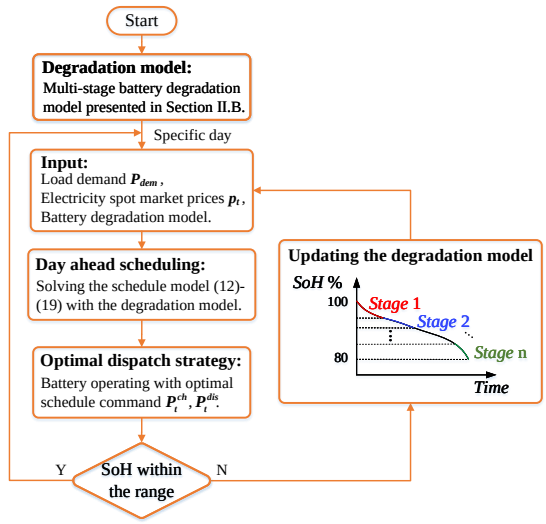


Fig. 2: Flowchart of the proposed adaptive optimization framework with the multi-stage battery degradation model.

are shown in Table III. The load demand, electricity spot prices, and temperature for the selected day are shown in Fig. 3(a)-(b). The battery operation profile (i.e., optimal dispatch strategies) with the two degradation models is shown in Fig. 3(c). In this case, battery charging operations are most desirable during 2:00–3:30 and 14:00–15:30 when the electricity prices are the lowest, and discharging operations are mostly performed during 8:00–8:30 and 19:00–20:30 when the electricity prices are the highest. It indicates that the optimal dispatch strategies for the battery are to keep the battery charged during a low-price period and discharge it to supply energy to the load during a high-price period. In this way, by shifting the load consumption away from peak prices, the overall operation cost could be reduced, and the impact of the oscillating load demand could also be mitigated.

Despite the slightly higher charging/discharging power command for the single-stage degradation model compared with the multi-stage degradation model, it can be observed that the optimal operation strategies, as well as the battery SoC shown in the bottom subfigure with the two degradation models, are quite similar. Fig. 6(a) shows that the accumulated battery capacity loss at the end of the day with the single degradation model is slightly larger than the multi-stage degradation model. In addition, based on the numerical results shown in Table VIII, the energy arbitrage revenue with the single-stage degradation model is slightly higher than the multi-stage degradation model counterparts, but at the cost of more battery capacity losses, resulting in higher overall operation costs. In summary, the results above align with the model performance comparison for the first stage, where the battery degradation estimation accuracy for the two models is close.

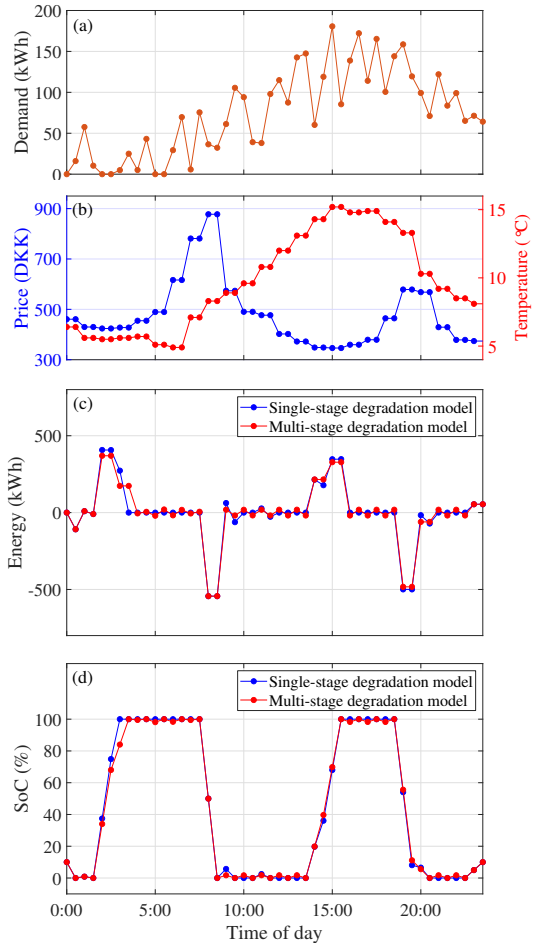


Fig. 3: Optimal dispatch strategies over a 24-hour period in stage 1. From top to bottom, the panels show (a) Load profile, (b) Spot market prices and ambient temperature, (c) Battery operation profile, and (d) Battery SoC level.

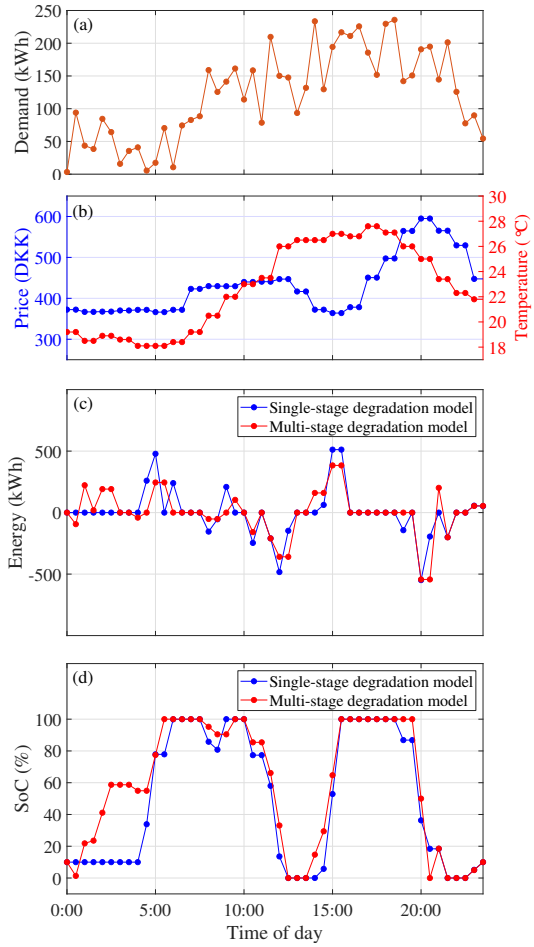


Fig. 4: Optimal dispatch strategies over a 24-hour period in stage 2. From top to bottom, the panels show (a) Load profile, (b) Spot market prices and ambient temperature, (c) Battery operation profile, and (d) Battery SoC level.

B. Case 2: Operating stage 2

When the battery operates in stage 2, the model parameters in Table IV are utilized. The load demand, spot market prices, and temperature for the example day in stage 2 are shown in Fig. 4(a) and Fig. 4(b). The battery operation profile and the corresponding SoC profile with the two degradation models are shown in the bottom two subfigures, where the optimal dispatch strategies are quite different for the two models. For both models, the battery is charged at around 5:00 and between 14:00 and 15:30 during low price periods and discharged for power supply during 10:00–12:30 and 19:30–20:30 when the electricity prices peak. It could also be observed that the battery remains uncharged rather than supplying energy to the load when the load demand is high and oscillates continuously due to the increasing electricity prices and high ambient temperature during 16:00–19:30. It indicates a higher degradation cost for the period compared to the energy arbitrage revenue. With the multi-stage model, the battery also operates during 0:00–4:00 to mitigate power fluctuation when the electricity price and ambient temperature are low.

The resultant battery capacity losses process in Fig. 6(b) shows that the battery capacity losses with the single-stage degradation model are larger than the multi-stage degradation model counterpart. Based on the numerical results of case 2 in Table VIII, compared with the single-stage battery degradation model, the scheduling with the multi-stage degradation model is more desirable and results in around 1.7% reduction in the overall operation costs, which are 2032.24 and 1997.14 DKK, respectively, for the two models.

C. Case 3: Operating stage 3

In stage 3, the model parameters in Table V are employed. The load demand, electricity spot price, and temperature during the day are shown in Fig. 5(a)-(b). As the electricity price keeps at a high level during the daytime between 6:00 and 17:00, the battery is first charged from 2:00 to 4:00 during the low price period and discharged from 7:00 to 7:30 when the demand increases to achieve price arbitrage. The battery is also charged slowly at around 8:00 and during 11:00–13:30 when the electricity drops, and discharged gradually during 9:00–10:30 and 14:00–16:30 to supply the heavy load demand for energy arbitrage. During the decreasing load demand and the high-temperature period between 17:00 and 21:30, the battery maintains a steady state without operations as the electricity prices decrease to a low level. Compared with the single-stage degradation model, the amplitude of charging and discharging power commands for the multi-stage model is reduced, resulting in reduced battery degradation at the cost of decreasing energy arbitrage revenue, as shown in Fig. 6(c) and Table VIII.

According to the numerical results in Table VIII, the scheduling strategies with the single-stage degradation model cause more capacity losses, though they achieve slightly higher energy arbitrage revenue than the multi-stage degradation model counterpart. As a result, the overall operation cost for scheduling with the multi-stage model is around 3.3% lower than the single-stage model counterpart.

D. Discussion on the number of divided operating stages for battery throughout lifetime

In principle, with the increasing number of divided stages over the battery lifetime, the accuracy of the multi-stage battery degradation model improves, and more desirable battery scheduling strategies could be obtained with the proposed optimization framework. To investigate the influence of the number of divided stages on the scheduling strategies, in the same way presented in Section II, by separating the battery degradation into different stages, the different multi-stage degradation models are implemented into the scheduling model.

To demonstrate the scheduling results of different multi-stage degradation models, the reduced operation cost for each model compared with the results of the single-stage model is calculated. For simplification, the single-stage model and the different multi-stage models are both implemented for the FCS operational scheduling framework on a single day, irrespective of the different divided stages. The results of the average reduced operation cost of the single day for the multi-stage degradation models with the different number of divided stages are shown in Fig. 7. As shown in Fig. 7, the improvement in model accuracy becomes marginal with the increasing number of divided stages. At the same time, it can be observed that the increase in reduced operation cost becomes stable at around 38 DKK on average after around five divided stages, at which point there are no evident benefits to further considering more divided stages.

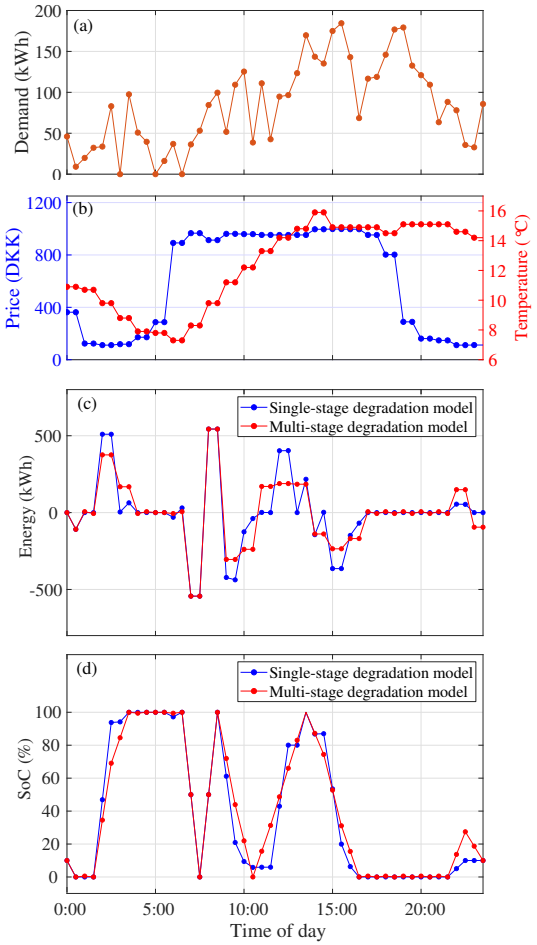
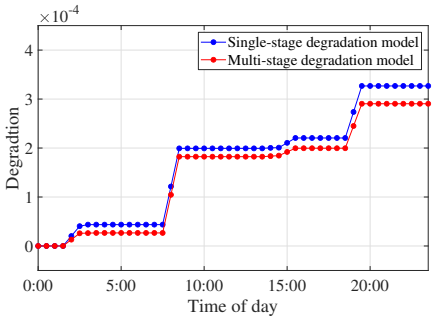


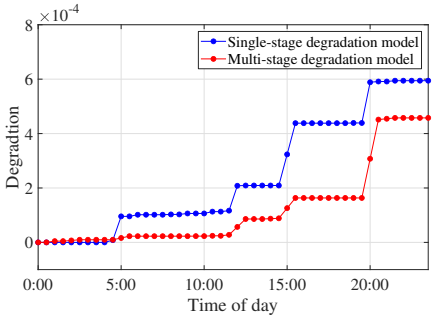
Fig. 5: Optimal dispatch strategies over a 24-hour period in stage 3. From top to bottom, the panels show (a) Load profile, (b) Spot market prices and ambient temperature, (c) Battery operation profile, and (d) Battery SoC level.

E. Application of the multi-stage framework considering other battery degradation models

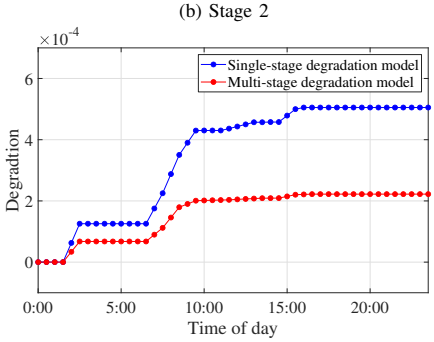
To further validate the performance of the proposed multi-stage modeling method, we also applied the adaptive optimization framework with two battery degradation models in [23] and [27]. The model in [23] demonstrates the applicability of our proposed multi-stage modeling method in another conventional model for the scheduling framework. To identify the parameters of the model, the battery cycling test data at different DoD levels are employed, and the parameters for different models are listed in Table VII. Details of the parameterization method can be found in [35]. Meanwhile, we also compare our method with the conventional multi-stage degradation model in [27] by applying it to the proposed scheduling framework. To parameterize the model, the battery aging tests at different C-



(a) Stage 1



(b) Stage 2



(c) Stage 3

Fig. 6: Comparison of the cumulative battery degradation for two models in each stage.

rates are employed. The models and corresponding parameters are listed in Table VII.

TABLE VII: Model parameterization

Model	Single-stage model	Multi-stage models
$Q_{loss} = a * DoD^b$	a=3.817e-4, b=3.226	Stage 1: a=1.989e-3, b=4.226 Stage 2: a=4.072e-4, b=3.497 Stage 3: a=2.002e-4, b=2.788
$Q_{loss} = a * I^b + c$	\	Stage 1: a=2.36e-5, b=3.119, c=5.569e-5 Stage 2: a=6.954e-6, b=3.092, c=1.348e-5 Stage 3: a=3.778e-6, b=3.077, c=6.938e-6

As mentioned, the multi-stage modeling method in [27] is not explored for the scheduling problem, and it considers a single current stress factor, which also lacks the flexibility of incorporating multiple factors into the degradation model. The scheduling results with the multi-stage modeling method are added to Table VIII for better comparison. It can be observed that the conventional multi-stage model underestimates the battery degradation at different stages. Thus, the battery usage is increased, resulting in higher energy arbitrage revenue while also inducing higher overall operation cost compared with the proposed multi-stage modeling method counterpart.

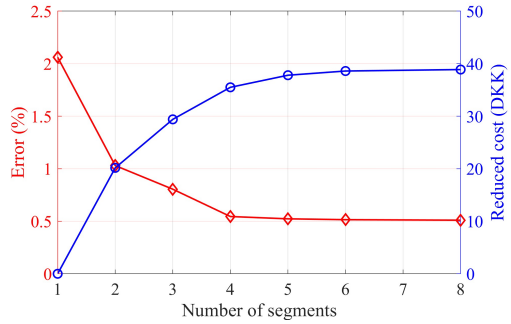


Fig. 7: Model accuracy and average reduced operation cost using the proposed optimization framework based on multi-stage degradation model regarding different number of segments for battery's lifetime. (Blue line: Reduced cost; Red line: Model error.)

In addition, we also demonstrate the applicability of the proposed multi-stage modeling method on another battery degradation model from [23]. The multi-stage models are implemented into the scheduling model in the same way presented in Fig. 2, and the quantitative comparison between the single battery degradation model and the derived multi-stage models at different stages is shown in Table IX.

It can be inferred from the model parameters that the battery degrading pattern varies at different stages, and the battery degrades faster with higher DoD at stage 1 while stabilizing at stages 2 and 3. At stage 1, the single-stage model underestimates the battery degradation, and even though higher energy arbitrage revenue is obtained the overall operation cost is higher compared with the multi-stage model counterpart. For the rest two stages, the single-stage model overestimates the battery degradation. Thus, the energy arbitrage revenue is reduced, and the overall operation cost is higher compared with the scheduling results of multi-stage model counterparts. The operation cost of the proposed method is reduced by 3.4% on average compared with the single-stage model counterpart.

In summary, the proposed multi-stage degradation modeling method is independent of different conventional battery degradation models, which could be employed to capture the varying degradation patterns to project the battery degradation across its whole lifetime more accurately, thus obtaining optimal scheduling strategies for minimizing the operation cost.

TABLE VIII: Comparison of the optimal scheduling results with different degradation models.

Cases	Case 1			Case 2			Case 3		
	[27]	Single-stage model	Multi-stage model	[27]	Single-stage model	Multi-stage model	[27]	Single-stage model	Multi-stage model
Battery degradation (%)	0.0467	0.0327	0.0290	0.134	0.0594	0.0458	0.0750	0.0505	0.0222
Energy arbitrage revenue (DKK)	1215.01	1213.12	1212.42	706.79	685.75	697.48	1420.74	1419.78	1419.08
Operation cost (DKK)	481.48	459.45	453.94	2138.27	2032.24	1997.14	1492.93	1452.07	1404.42

TABLE IX: Performance of the proposed multi-stage degradation modeling method based on battery degradation model in [23].

Cases	Case 1		Case 2		Case 3	
	Single-stage model	Multi-stage model	Single-stage model	Multi-stage model	Single-stage model	Multi-stage model
Battery degradation (%)	0.0563	0.0365	0.0369	0.0340	0.0261	0.0287
Energy arbitrage revenue (DKK)	1214.43	1211.40	697.64	702.82	1338.33	1419.67
Operation cost (DKK)	498.43	467.73	1981.62	1971.67	1457.38	1406.65

F. Discussion on the practical implementation

The average computation time of the proposed optimization framework for the case study is only around 5.02 s, carried out on a PC with an Intel (R) Core (TM) i5-10210U 1.60 GHz processor and 8.00 GB of RAM. To further investigate the computation time of the formulated problem, the optimization problem is solved under various conditions. Moreover, the average computation time to obtain the optimal charging and discharging decisions is around 5.65 s. As the scheduling problem is formulated for the day-ahead stage, the computation time is more than acceptable for day-ahead optimization, where updates of schedule will typically be calculated once every several minutes or more. Furthermore, suppose more ancillary services are incorporated into the optimization problem, such as frequency regulation, which has a smaller time step. In that case, the scheduling model should be properly formulated to satisfy the computation time for the specific service.

In addition, the proposed multi-stage degradation modeling method is applicable for other practical systems equipped with BESS by improving the accuracy of assessing the battery degradation across its entire lifetime. Moreover, the scheduling with the multi-stage models can also be deployed as part of the cloud-based real-time control platform presented in [43].

In summary, the proposed optimization framework with the multi-stage battery degradation model could be implemented in practical applications.

V. Conclusion

In this paper, a novel adaptive optimization framework with a multi-stage battery degradation modeling method is proposed. The multi-stage degradation modeling method is proposed to capture the varying battery aging patterns across its entire lifetime, thus achieving an accurate evaluation of the aging cost for the scheduling problems in different battery operational stages. Case studies for the scheduling problem at different stages by applying the multi-stage modeling method to a generic cycle degradation model, resulting in around 2.1% reduction in the overall operation cost compared with the single-stage model counterpart. Additionally, as the number of segments for battery lifespan increases, the model error decreases and stabilizes at roughly 0.5% with enough separated segments for battery degradation, while the reduced operation cost increases and saturates. The superiority of the proposed multi-stage framework is further validated based on other battery degradation models. The computation time for the day-ahead scheduling problem is around 5 s, which makes the proposed optimization framework practical for real-time applications.

More importantly, the proposed multi-stage degradation modeling method is independent of various battery chemistry types and empirical/semi-empirical degradation models. Moreover, for other systems equipped with BESSs, by properly formulating the schedule model, the proposed optimization framework with the multi-stage degradation model could still be employed to obtain the economic operations. Results in the study provide a guide for battery scheduling in different SoH stages, concluding that the multi-stage modeling method in the proposed optimization framework achieves more cost-effective dispatch strategies. Future work could be mitigating the requirement of the experimental aging test across the whole lifespan for parameterizing the multi-stage model, which is time-consuming. Furthermore, more ancillary services provided by the battery at different operating stages could be investigated from an economic perspective.

Appendix

Impact of C-rate on battery degradation

The model performance under cycling conditions of different current rates is validated based on the available test data. The validation results, quantified in root-mean-square error, are presented in Table X.

In addition, Section II shows that the current stress model is parameterized with experimental tests cycled at a current rate up to 2 C. In summary, the current rate condition of the experimental data used for parameter identification and model validation is up to 2 C and 3 C, respectively. In this way, the

TABLE X: Model performance under different tests at different stages.

Testing conditions	Stages	Single-stage model	Multi-stage model
1C, 100% DoD, 25 °C	Stage 1	1.53%	0.64%
	Stage 2	1.61%	1.14%
	Stage 3	1.79%	0.41%
2C, 100% DoD, 25 °C	Stage 1	1.37%	0.70%
	Stage 2	2.03%	1.22%
	Stage 3	1.31%	1.25%
3C, 100% DoD, 25 °C	Stage 1	1.34%	0.75%
	Stage 2	1.62%	1.03%
	Stage 3	2.33%	1.47%

built model is expected to predict the general trend of battery degradation under different current rates above 3 C.

The above results confirm the effectiveness of the proposed multi-stage degradation modeling method for the current stress factor. More importantly, the results in Table VI offer a more comprehensive validation of various stress factors on battery degradation, highlighting the superior performance of the proposed multi-stage battery degradation modeling method.

REFERENCES

[1] I. G. E. Outlook, "Accelerating ambitions despite the pandemic," *International Energy Agency: Paris, France*, 2021.

[2] P. Morrissey, P. Weldon, and M. O'Mahony, "Future standard and fast charging infrastructure planning: An analysis of electric vehicle charging behaviour," *Energy Policy*, vol. 89, pp. 257–270, 2016.

[3] M. M. Mahfouz and M. R. Iravani, "Grid-integration of battery-enabled dc fast charging station for electric vehicles," *IEEE Transactions on Energy Conversion*, vol. 35, no. 1, pp. 375–385, 2019.

[4] M. R. Khalid, M. S. Alam, A. Sarwar, and M. J. Asghar, "A comprehensive review on electric vehicles charging infrastructures and their impacts on power-quality of the utility grid," *ETransportation*, vol. 1, p. 100006, 2019.

[5] M. A. H. Rafi and J. Bauman, "A comprehensive review of dc fast-charging stations with energy storage: Architectures, power converters, and analysis," *IEEE Transactions on Transportation Electrification*, vol. 7, no. 2, pp. 345–368, 2020.

[6] C. Zheng, W. Li, and Q. Liang, "An energy management strategy of hybrid energy storage systems for electric vehicle applications," *IEEE Transactions on Sustainable Energy*, vol. 9, no. 4, pp. 1880–1888, 2018.

[7] Energiforskning, "Heart super high efficiency fast ev charger with energy storage and grid support functionality: A heart of the future green transport," 2021. [Online]. Available: https://energiforskning.dk/projekter/heart-super-high-efficiency-fast-ev-charger-with-energy-storage-and-grid-support

[8] M. Yang, L. Zhang, Z. Zhao, and L. Wang, "Comprehensive benefits analysis of electric vehicle charging station integrated photovoltaic and energy storage," *Journal of Cleaner Production*, vol. 302, p. 126967, 2021.

[9] F. Wankmüller, P. R. Thimmapuram, K. G. Gallagher, and A. Botterud, "Impact of battery degradation on energy arbitrage revenue of grid-level energy storage," *Journal of Energy Storage*, vol. 10, pp. 56–66, 2017.

[10] M. Alramlawi and P. Li, "Design optimization of a residential pv-battery microgrid with a detailed battery lifetime estimation model," *IEEE Transactions on Industry Applications*, vol. 56, no. 2, 2020.

[11] S. Choi and S.-W. Min, "Optimal scheduling and operation of the ess for prosumer market environment in grid-connected industrial complex," *IEEE Transactions on Industry Applications*, vol. 54, no. 3, pp. 1949–1957, 2018.

[12] U. R. Nair, M. Sandelic, A. Sangwongwanich, T. Dragičević, R. Costa-Castelló, and F. Blaabjerg, "An analysis of multi objective energy scheduling in pv-bess system under prediction uncertainty," *IEEE Transactions on Energy Conversion*, vol. 36, no. 3, pp. 2276–2286, 2021.

[13] J. Shen and A. Khaligh, "A supervisory energy management control strategy in a battery/ultracapacitor hybrid energy storage system," *IEEE Transactions on Transportation Electrification*, vol. 1, no. 3, pp. 223–231, 2015.

[14] M. Elkazaz, M. Sumner, and D. Thomas, "Energy management system for hybrid pv-wind-battery microgrid using convex programming, model predictive and rolling horizon predictive control with experimental validation," *International Journal of Electrical Power & Energy Systems*, vol. 115, p. 105483, 2020.

[15] F. Fan, Y. Xu, R. Zhang, and T. Wan, "Whole-lifetime coordinated service strategy for battery energy storage system considering multi-stage battery aging characteristics," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 3, pp. 689–699, 2021.

[16] A. Maheshwari, N. G. Paterakis, M. Santarelli, and M. Gibescu, "Optimizing the operation of energy storage using a non-linear lithium-ion battery degradation model," *Applied Energy*, vol. 261, p. 114360, 2020.

[17] J. Faraji, A. Ketabi, and H. Hashemi-Dezaki, "Optimization of the scheduling and operation of prosumers considering the loss of life costs of battery storage systems," *Journal of Energy Storage*, vol. 31, p. 101655, 2020.

[18] A. Merabet, A. Al-Durra, and E. F. El-Saadany, "Energy management system for optimal cost and storage utilization of renewable hybrid energy microgrid," *Energy Conversion and Management*, vol. 252, p. 115116, 2022.

[19] Y. Wan, D. Gebbran, and T. Dragičević, "Optimal dispatch schedule for a fast ev charging station with account to supplementary battery health degradation," *arXiv preprint arXiv:2203.08029*, 2022.

[20] S. Wang, D. Guo, X. Han, L. Lu, K. Sun, W. Li, D. U. Sauer, and M. Ouyang, "Impact of battery degradation models on energy management of a grid-connected dc microgrid," *Energy*, vol. 207, p. 118228, 2020.

[21] K. Abdulla, J. De Hoog, V. Muenzel, F. Suits, K. Steer, A. Wirth, and S. Halgamuge, "Optimal operation of energy storage systems considering forecasts and battery degradation," *IEEE Transactions on Smart Grid*, vol. 9, no. 3, pp. 2086–2096, 2018.

[22] A. Barragán-Moreno, P. I. Gomez, and T. Dragičević, "Enhancement of stress cycle-counting algorithms for li-ion batteries by means of fuzzy logic," in *2022 IEEE Transportation Electrification Conference & Expo (ITEC)*. IEEE, 2022, pp. 981–985.

[23] B. Xu, J. Zhao, T. Zheng, E. Litvinov, and D. S. Kirschen, "Factoring the cycle aging cost of batteries participating in electricity markets," *IEEE Transactions on Power Systems*, vol. 33, no. 2, pp. 2248–2259, 2017.

[24] S. F. Schneider, P. Novák, and T. Kober, "Rechargeable batteries for simultaneous demand peak shaving and price arbitrage business," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 1, pp. 148–157, 2020.

[25] J.-O. Lee and Y.-S. Kim, "Novel battery degradation cost formulation for optimal scheduling of battery energy storage systems," *International Journal of Electrical Power & Energy Systems*, vol. 137, p. 107795, 2022.

[26] X. Ke, N. Lu, and C. Jin, "Control and size energy storage systems for managing energy imbalance of variable generation resources," *IEEE Transactions on Sustainable Energy*, vol. 6, no. 1, pp. 70–78, 2014.

[27] Y. Gao, J. Jiang, C. Zhang, W. Zhang, Z. Ma, and Y. Jiang, "Lithium-ion battery aging mechanisms and life model under different charging stresses," *Journal of Power Sources*, vol. 356, pp. 103–114, 2017.

[28] Y. Qin, C. Yuen, X. Yin, and B. Huang, "A transferable multi-stage model with cycling discrepancy learning for lithium-ion battery state of health estimation," *IEEE Transactions on Industrial Informatics*, 2022.

[29] W. Li, N. Sengupta, P. Dechent, D. Howey, A. Annaswamy, and D. U. Sauer, "One-shot battery degradation trajectory prediction with deep learning," *Journal of Power Sources*, vol. 506, p. 230024, 2021.

[30] J. Liu, Q. Duan, K. Qi, Y. Liu, J. Sun, Z. Wang, and Q. Wang, "Capacity fading mechanisms and state of health prediction of commercial lithium-ion battery in total lifespan," *Journal of Energy Storage*, vol. 46, p. 103910, 2022.

[31] R. Xiong, Y. Zhang, J. Wang, H. He, S. Peng, and M. Pecht, "Lithium-ion battery health prognosis based on a real battery management system used in electric vehicles," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 5, pp. 4110–4121, 2018.

[32] S. N. Motapon, E. Lachance, L.-A. Dessaint, and K. Al-Haddad, "A generic cycle life model for lithium-ion batteries based on fatigue theory and equivalent cycle counting," *IEEE Open Journal of the Industrial Electronics Society*, vol. 1, pp. 207–217, 2020.

[33] S. N. Laboratories, "Battery archive," https://batteryarchive.org, accessed 2023-06-15.

[34] K. Smith, M. Earleywine, E. Wood, J. Neubauer, and A. Pesaran, "Comparison of plug-in hybrid electric vehicle battery life across

geographies and drive-cycles," National Renewable Energy Lab.(NREL), Golden, CO (United States), Tech. Rep., 2012.

[35] I. Laresgoiti, S. Käbitz, M. Ecker, and D. U. Sauer, "Modeling mechanical degradation in lithium ion batteries during cycling: Solid electrolyte interphase fracture," *Journal of Power Sources*, vol. 300, pp. 112–122, 2015.

[36] Y. Preger, H. M. Barkholtz, A. Fresquez, D. L. Campbell, B. W. Juba, J. Romàn-Kustas, S. R. Ferreira, and B. Chalamala, "Degradation of commercial lithium-ion cells as a function of chemistry and cycling conditions," *Journal of The Electrochemical Society*, vol. 167, no. 12, p. 120532, 2020.

[37] D. Tran and A. M. Khambadkone, "Energy management for lifetime extension of energy storage system in micro-grid applications," *IEEE Transactions on Smart Grid*, vol. 4, no. 3, pp. 1289–1296, 2013.

[38] W. Chen, J. Qiu, J. Zhao, Q. Chai, and Z. Y. Dong, "Bargaining game-based profit allocation of virtual power plant in frequency regulation market considering battery cycle life," *IEEE Transactions on Smart Grid*, vol. 12, no. 4, pp. 2913–2928, 2021.

[39] C. Liu, X. Wang, X. Wu, and J. Guo, "Economic scheduling model of microgrid considering the lifetime of batteries," *IET Generation, Transmission & Distribution*, vol. 11, no. 3, pp. 759–767, 2017.

[40] Y. Qin, H. Hua, and J. Cao, "Stochastic optimal control scheme for battery lifetime extension in islanded microgrid via a novel modeling approach," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 4467–4475, 2018.

[41] W. E. Hart, J.-P. Watson, and D. L. Woodruff, "Pyomo: modeling and solving mathematical programs in python," *Mathematical Programming Computation*, vol. 3, no. 3, pp. 219–260, 2011.

[42] L. T. Biegler and V. M. Zavala, "Large-scale nonlinear programming using ipopt: An integrating framework for enterprise-wide dynamic optimization," *Computers & Chemical Engineering*, vol. 33, no. 3, pp. 575–582, 2009.

[43] D. Gebbran, A. Barragán-Moreno, P. I. Gomez, K. Subroto, M. M. Mohammad, M. López, J. Quiroz, and T. Dragičević, "Cloud & edge computing for smart management of power electronic converter fleets: A key connective fabric to enable the green transition," *IEEE Industrial Electronics Magazine*, 2022.

[C1] Reinforcement Learning Based Weighting Factor Design of Model Predictive Control for Power Electronic Converters

Authors:
Yihao Wan, Tomislav Dragicevic, Nenad Mijatovic, Chang Li, Jose Rodriguez

Submitted to:
The 6th IEEE International Conference on Predictive Control of Electrical Drives and Power Electronic (PRECEDE)

Status:
Published.
Digital Object Identifier: 10.1109/PRECEDE51386.2021.9680964

Reinforcement Learning Based Weighting Factor Design of Model Predictive Control for Power Electronic Converters

Yihao Wan
Department of Electrical Engineering
Technical university of Denmark
Kongens Lyngby, Denmark
wanyh@elektro.dtu.dk

Tomislav Dragicevic
Department of Electrical Engineering
Technical university of Denmark
Kongens Lyngby, Denmark
tomdr@elektro.dtu.dk

Nenad Mijatovic
Department of Electrical Engineering
Technical university of Denmark
Kongens Lyngby, Denmark
nm@elektro.dtu.dk

Chang Li
Department of Electrical Engineering
Technical university of Denmark
Kongens Lyngby, Denmark
changli@dtu.dk

Jose Rodriguez
Department of Engineering
Universidad Andres Bello
Santiago, Chile
jose.rodriguez@unab.cl

2021 IEEE International Conference on Predictive Control of Electrical Drives and Power Electronics (PRECEDE) | 978-1-6654-2557-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/PRECEDE51386.2021.9680964

***Abstract*—Weighting factor design is one of the challenges for finite-set model predictive control (FS-MPC) controlled power electronic converters, which plays an important role in the balance of control objectives in the cost function to achieve desired performance. This paper investigates the application of reinforcement learning algorithm for the weighting factor design for FS-MPC regulated voltage source converter in uninterrupted power supply (UPS) system. The deep deterministic policy gradient (DDPG) agent is employed to learn the optimal weighting factor design policy. The reinforcement learning (RL) agent is trained in the system and the weighting factor is optimized based on reward calculation with the interactions between the agent and environment. The key performance metric, total harmonic distortion (THD), is incorporated in the reward function. Effectiveness of the proposed reinforcement learning based weighting factor design method is validated by simulations.**

Keywords—weighting factor, finite-set model predictive control (FS-MPC), power electronic converters, deep deterministic policy gradient (DDPG).

I. Introduction

The power converters play an important role in the power system operation, and among different kinds of topologies, voltage source converters (VSC) are one of the most popular converters in many power conversion application, such as uninterruptible power supply (UPS) system [1]. Different advanced control approaches are proposed to overcome the limitations of classical linear control methods [2]. Model predictive control (MPC) has been widely applied in power electronics and drives. It outperforms traditional linear control methods by multi-objective optimization and system constraints management capability. Among different kinds of MPC, the finite-set MPC (FS-MPC) becomes popular due to the simple implementation. It employs the discrete model of the converter to predict future behavior for all possible switching combinations based on the latest measurements. The performance of the converter is defined through a cost function which incorporates the control objectives. Then the predictions are evaluated according to the cost function, and the switching states with the least value will be applied to the converter.

In addition, different control objectives are balanced with weighting factors of each term in the cost function. Therefore, how to optimally select the weighting factor is still an open question for the FS-MPC. A simple way to select the weighting factor is the branch and bound search [3], while the method is too empirical. Genetic algorithms are utilized to obtain the suitable weighting factors, which is offline and new sets of simulations is performed for each design objectives [4]. In [5], an automated approach based on artificial neural networks (ANN) was proposed to aid the weighting factor design. However, a large number of offline simulation or experiments are required to obtain desired performance. Other attempts are focused on the simplification of the FS-MPC by removing the weighting factors [6-7]. However, there are still priority coefficients in the cost functions which will affect the behavior of the system.

In order to implement automatic online weighting factor design, in this paper, reinforcement learning algorithm is employed. Recently, the application of reinforcement learning algorithm in power electronic system has been attached great importance with the rapid development of artificial intelligence [8-10]. Reinforcement learning involves an agent learning the optimal action through trial-and-error interactions with dynamic and unknown environment. A reward function is defined to evaluate the performance of the actions during the training process.

In this paper, a weighting factor design method based on RL algorithm is proposed. A RL algorithm, DDPG, is used to train the agent for optimal weighting factor design of FS-MPC controlled UPS system, which is able to improve the system performance by taking optimal actions in an infinite continuous action space to the MPC controller. A proper reward function incorporating THD and reference voltage tracking error is designed to help improve the behavior of the system. The agent will output optimal weighting factor to the FS-MPC based on the measurements after the optimal policy is obtained through online training. Based on this, optimal weighting factor can be obtained by the trained RL agent.

II. System Model

A. Converter model

In the UPS system, shown in Fig. 1, a two level VSC is used. It is modeled in the stationary α-β reference frame, where the filter voltage for all possible 8 gating signal combinations is transformed. Therefore, all the three phase variables in the system will be transformed into α-β reference frame based on the amplitude-invariant Clarke

transformation. The gating signals S_a, S_b, and S_c determines the voltage of each phase leg by $v_{aN} = S_a \cdot v_{dc}$, $v_{bN} = S_b \cdot v_{dc}$, $v_{cN} = S_c \cdot v_{dc}$ respectively, where v_{dc} is the DC source voltage and N is the negative terminal of DC link. Considering all the possible combination of S_a, S_b, and S_c, eight voltage vectors are obtained, shown in Table I.

TABLE I. SWITCHING STATES AND COMPLEX VOLTAGE VECTORS

S_a	S_b	S_c	Voltage vector v_i
0	0	0	$v_0 = 0$
1	0	0	$v_1 = \frac{2}{3}v_{dc}$
1	1	0	$v_2 = \frac{1}{3}v_{dc} + j\frac{\sqrt{3}}{3}v_{dc}$
0	1	0	$v_3 = -\frac{1}{3}v_{dc} + j\frac{\sqrt{3}}{3}v_{dc}$
0	1	1	$v_4 = -\frac{2}{3}v_{dc}$
0	0	1	$v_5 = -\frac{1}{3}v_{dc} - j\frac{\sqrt{3}}{3}v_{dc}$
1	0	1	$v_6 = \frac{1}{3}v_{dc} - j\frac{\sqrt{3}}{3}v_{dc}$
1	1	1	$v_7 = 0$

The voltage vector v_i can be applied to LC filter and the differential equations of the inductor current i_f and capacitor voltage v_f are expressed as follows

$$L_f \frac{di_f}{dt} = v_i - v_f - R_f i_f \tag{1}$$

$$C_f \frac{dv_f}{dt} = i_f - i_o \tag{2}$$

where L_f and C_f are the filter inductance and capacitance respectively, R_f is the series resistance of L_f and i_o is the output current.

The equations can be rewritten in state-space representation as

$$\frac{d}{dt}\begin{bmatrix} i_f \\ v_f \end{bmatrix} = \mathbf{A}\begin{bmatrix} i_f \\ v_f \end{bmatrix} + \mathbf{B}\begin{bmatrix} v_i \\ i_o \end{bmatrix} \tag{3}$$

where $\mathbf{A} = \begin{bmatrix} -\frac{R_f}{L_f} & -\frac{1}{L_f} \\ -\frac{1}{C_f} & 0 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} -\frac{1}{L_f} & 0 \\ 0 & -\frac{1}{C_f} \end{bmatrix}$.

B. Cost function

To implement the digital control of FS-MPC to the converter, the zero-order hold discretization method is utilized to estimate the future states of voltages and currents of the system. And the single-step horizon cost function is employed, which can be expressed as [11]

$$\begin{aligned} \mathrm{g} &= \left(v_{f\alpha}^* - v_{f\alpha}\right)^2 + \left(v_{f\beta}^* - v_{f\beta}\right)^2 + \lambda_d g_d + h_{lim} \\ \mathrm{g}_d &= \left(C_f \omega_f v_{f\beta}^* - i_{f\alpha} + i_{o\alpha}\right)^2 + \left(C_f \omega_f v_{f\alpha}^* - i_{f\beta} + i_{o\beta}\right)^2 \\ h_{lim} &= \begin{cases} 0, \text{ if } |\bar{i}_f| \le i_{max} \\ \infty, \text{ if } |\bar{i}_f| > i_{max} \end{cases} \end{aligned} \tag{4}$$

where $v_f^* = v_{f\alpha}^* + v_{f\beta}^*$ is the voltage reference, $\omega_f = 2\pi f$ is the reference frequency and λ_d is the weighting factor of the additional current reference term g_d, which is introduced to improve the steady state performance [11]. h_{lim} is the current limiting term.

It can be observed that single weighting factor λ_d should be selected. The weighting factor has influence on the performance of the system, which should be properly designed. And THD of the capacitor voltage is an important metric to quantify the performance of the system.

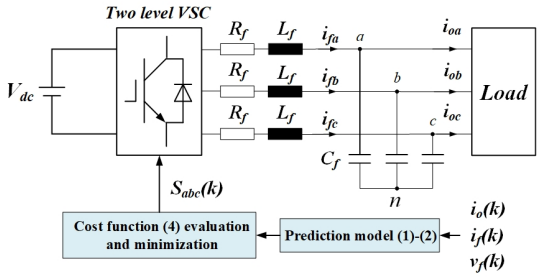


Fig. 1. Structure of the FS-MPC controlled VSC for UPS applications.

III. PROPOSED REINFORCEMENT LEARNING BASED WEIGHTING FACTOR DESIGN METHOD

A. Deep Deterministic Policy Gradient (DDPG)

In a standard reinforcement learning paradigm, it consists of an agent interacting with stochastic environment sequentially to learn the optimal policy. In this way, the weighting factor design problem can be modelled as a Markov decision process (MDP), which is composed of state space S, action space A, state transition probability function P, reward function R [12]. At each time step t, the agent will perceive an observation s_t in the environment, and take an action a_t based on the policy $\pi(a_t|s_t)$ and receive a reward r_t. The policy $\pi(a_t|s_t)$ maps the state s_t to a probability distribution of action a_t. After that, a new state s_{t+1} will be devised. The cumulative discounted reward, which can be expressed as

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k} \tag{5}$$

where $\gamma \in [0,1]$ is the discounting factor.

As the goal of the agent is to learn the policy that maximizes the reward r_t and thus maximize the discounted reward R_t from the initial state, the objective function can be expressed as

$$J = \mathbb{E}_{s\sim\rho^\pi, a\sim\pi^\theta}[R_1] \tag{6}$$

For a specific policy π, the value-function is used in reinforcement learning algorithm to describe the expected return R_t with the action a_t with respect to the state s_t, which is

$$Q^\pi(s_t, a_t) = \mathbb{E}_\pi[R_t|s_t, a_t] \tag{7}$$

Employing the recursive relationship, called as the Bellman equation, the action-value function can be expressed as

$$Q^\pi(s_t, a_t) = \mathbb{E}_\pi[r(s_t, a_t) + \gamma \mathbb{E}_\pi Q^\pi(s_{t+1}, a_{t+1})] \tag{8}$$

In the context of RL algorithm, DDPG algorithm has good performance in solving the problems with continuous dynamics [13-14]. DDPG algorithm has the actor-critic structure, where the critic network approximates the action value function $Q(s_t, a_t|\theta^Q)$ with weight θ^Q and actor network defines the current policy $\mu(s_t, \theta^\mu)$ with the state and corresponding action [15]. The weights of the critic network

is optimized during training process based on the loss function L, which is

$$L(\theta^Q) = E_{(s,a)}[Q(s_t, a_t|\theta^Q) - \mathrm{y_t}]^2 \tag{9}$$

where

$$y_t = r_t + \gamma Q(s_{t+1}, \mu(s_t|\theta^\mu)|\theta^Q) \tag{10}$$

The weight θ^μ of actor network is updated using the policy gradient as follows

$$\begin{gathered} \nabla_{\theta^\mu} J \approx \mathbb{E}_{s_t\sim\rho^\beta}\left[G_a G_\mu\right] \\ G_a = \nabla Q(s_t, a_t|\theta^Q) \\ G_\mu = \nabla_{\theta^\mu}\mu(s_t|\theta^\mu) \end{gathered} \tag{11}$$

where $a_t = \mu(s_t|\theta^\mu)$, ρ is the discounted distribution and β represents the current policy. G_a denotes the gradient of the critic output regarding the action computed by the actor network, and G_μ denotes the gradient of the actor output regarding the actor parameters. Both of them are evaluated with observation s_t.

To enhance the stability and reliability of the DDPG agent during the training process, two distinct networks, i.e. critic target $Q'(s, a|\theta^{Q'})$ and actor target $\mu'(s|\theta^{\mu'})$, are introduced in the critic network and actor network respectively. The soft weights $\theta^{Q'}$ and $\theta^{\mu'}$ are updated as follows [13]

$$\begin{aligned} \theta^{Q'} &\leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'} \\ \theta^{\mu'} &\leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'} \end{aligned} \tag{12}$$

where τ <<1. It represents the smoothing factor, with which the target parameters are updated at each time step.

In addition, an Ornstein-Uhlenbeck (OU) action noise model is employed to improve the exploration of the agent [13], and the actor output action is expressed as

$$a_t = \mu(s_t, \theta^\mu) + \mathcal{N} \tag{13}$$

where $\mathcal{N}$ is the exploration noise.

B. Proposed approach

In this paper, the DDPG algorithm is employed to design the weighting factor for the MPC controller in the UPS system. As is shown in Fig.2, the environment states consist of reference voltage $v_{f\alpha}^*, v_{f\beta}^*$ and capacitor voltage $v_{f\alpha}, v_{f\beta}$. Also, THD is an important metric to quantify the system behavior. Therefore, the state S is defined as $s_t = \left[v_{f\alpha}^*, v_{f\beta}^*, v_{f\alpha}, v_{f\beta}, e_\alpha, e_\beta, THD\right]$, the action space A is defined as $a_t = [\lambda_d]$, where e_α, e_β is the voltage error between reference voltage and capacitor voltage. The training data is measured and sampled randomly in the UPS system. And the action of the agent is the weighting factor which will be input to the MPC controller for each training time step.

The reward function is one of the most important part in the reinforcement learning algorithm, which should be properly designed to achieve desired performance for the system. In the UPS system shown in Fig.2, the goal is to output the standard sinusoidal waveform with respect to the reference voltage, the THD of capacitor voltage is one of the most important metrics to quantify the performance of the controller. Therefore, THD of capacitor voltage is incorporated in the reward function. In addition, voltage error between the reference voltage and output capacitor voltage is also introduced in the reward function to reduce the amplitude of error voltage and improve the tracking accuracy of the MPC controller. In the training process of the agents, the agents may take actions that results in the performance being far from the desired. Therefore, to assist the training of DDPG agent, a penalty term is introduced in the reward function based on whether the system has desired performance with the output weighting factor for a certain time period by comparing the real-time THD with the preset value, if not, current training episode will be stopped and it will move on to a new training episode. The formula of the reward function is as below

$$\begin{aligned} r = -\Big\{k_1 \cdot \left[\left(v_{f\alpha}^* - v_{f\alpha}\right)^2 + \left(v_{f\beta}^* - v_{f\beta}\right)^2\right] + \\ k_2 \cdot THD^2 + 100a\Big\} \end{aligned} \tag{14}$$

where a is a flag that equals to 1 when the simulation is terminated early, implying current training episode is not desired and a new training episode will be initialized. Due to the different order of magnitude for the THD value and voltage tracking error, the coefficients for the two terms are set as k_1 = 1e-5 and k_2 = 100.

Based on the reward calculation after each training episode, the action will be optimized to enable the agent to obtain optimal reward, thus minimizing the THD of capacitor voltage. The episode number is set to 500 and the training time step can be obtained with the simulation period and sampling time, which is 0.08s and 20e-6s respectively. Therefore, the time step of each training episode is 4000. The size of random experience mini-batch M is set to 512. The pseudo-code of the DDPG algorithm training process is shown below.

TABLE II. TRAINING PROCESS OF DDPG ALGORITHM

Algorithm: DDPG algorithm

Input: Environment $\left[v_{f\alpha}^*, v_{f\beta}^*, v_{f\alpha}, v_{f\beta}, e_\alpha, e_\beta, THD\right]$

Output: Weighting factor λ_d

1: Randomly initialize actor network $\mu(s_t, \theta^\mu)$ and critic network $Q(s_t, a_t|\theta^Q)$ with weights θ^μ and θ^Q respectively

2: Initialize target network μ' and Q' with weights $\theta^{Q'} \leftarrow \theta^Q, \theta^{\mu'} \leftarrow \theta^\mu$

3: **for** episode = 1 to 500 **do**

4: Begin with an random noise $\mathcal{N}$ of Ornstein-Uhelnbeck action noise model for exploration

5: Obtain current observation s_t

6: **for** t =1 to 4000 **do**

7: Apply action $a_t = \mu(s_t, \theta^\mu) + \mathcal{N}$ to the system

8: Observe the new state s_{t+1} and the reward r_t

9: Store experience (s_t, a_t, r_t, s_{t+1}) in experience buffer R

10: Sample a random mini-batch of M experiences (s_t, a_t, r_t, s_{t+1}) from experience buffer

11: If s_{t+1} is a terminal state, set the y_t to r_t. Otherwise, set

$$y_t = r_t + \gamma Q'(s_{t+1}, \mu'(s_{t+1}|\theta^\mu)|\theta^{Q'})$$

12: Update critic network by minimizing the loss $L(\theta^Q)$

$$L(\theta^Q) = \frac{1}{M}\sum_{i=1}^{M}[Q(s_i, a_i|\theta^Q) - \mathrm{y_i}]^2$$

13: Update actor network using following sampled policy gradient

$$\nabla_{\theta^\mu} J \approx \frac{1}{M}\sum_{i=1}^{M}\left[\nabla Q(s_t, a_t|\theta^Q)|_{a_t=\mu(s_t|\theta^\mu)}\nabla_{\theta^\mu}\mu(s_t|\theta^\mu)\right]$$

14: Update target actor and critic network using

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$$
$$\theta^{\mu'} = \tau\theta^\mu + (1-\tau)\theta^{\mu'}$$

15: **end for**

16: **end for**

As is shown in the table, the observations for DDPG agent from environments states $[v^*_{f\alpha}, v^*_{f\beta}, v_{f\alpha}, v_{f\beta}, e_\alpha, e_\beta, THD]$ is used. In the training process, the training data, state S, is sampled randomly within the operation range of the FS-MPC regulated UPS system shown in Fig.2. After each training episode, the reward will be calculated based on equation (14) and the neural network will be updated and the agent will take new actions based on the reward. Eventually, the optimal policy of continuous action space in DDPG algorithm is learned with a deep neural network. In this way, after training session is finished according to the termination criteria for calculated reward value, the trained agent will be applied to the weighting factor design for the system online. With the corresponding inputs from the system states, which is the same as the inputs during training, the deep neural network in the agent will enable it to output the optimal weighting factor for MPC controller with respect to the inputs from the environment.

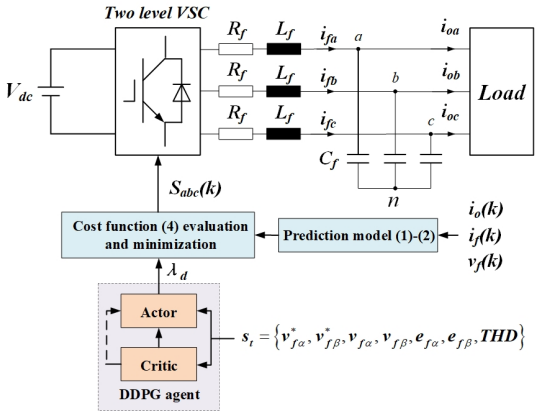


Fig. 2. Diagram of reinforcement learning based weighting factor design for FS-MPC controlled two-level VSC

IV. Results and Discussion

To validate the proposed reinforcement learning based weighting factor design method, simulation based on the UPS system in Fig.1 is implemented and the parameters are shown in Table 3.

TABLE III. System Parameters

Parameters	Value
DC-link voltage	V_{dc} = 520 V
LC-filter	L_f = 2.2 mH, R_f = 0.1 Ω, C_f = 15 μF
Reference voltage	V_r = 200V, f_r = 50 Hz
Nominal load	R_{load} = 30.6 Ω

A. Simulation results

The training process of DDPG agent is shown in Fig.7, where x-axis and y-axis denote the episode number and episode reward respectively, showing that the long term reward converges after about 100 training episodes. It can be observed from Fig.7 that the training process of the DDPG agent can be divided into three stages, i.e. the exploration, learning and convergence stage. During the exploration stage, the action value, was randomly selected and stored in the experience buffer, which results in a wide range of adjustment.

When the buffer was filled, the DDPG agent started the learning process, where the parameters of the actor and critic network are updated according to equation (12) to maximize the cumulative reward. After around 30 training episodes, the episode reward value tends to converge, which implies that the RL agent has successfully learned the policy and the optimal weighting factor for the MPC controller is obtained. The average reward researches the threshold and the training process terminated.

Then the trained agent will be applied to the system for weighting factor tuning. The weighting factor is designed and optimized by the reinforcement learning agent. During the simulation, for each sample time, the environment states, $s_t = [v^*_{f\alpha}, v^*_{f\beta}, v_{f\alpha}, v_{f\beta}, e_\alpha, e_\beta, THD]$, is measured and calculated, which will be input to the DDPG agent. And with the learned optimal policy, the agent will take actions based on the input data, thus the optimal weighting factor is obtained and output to the MPC controller.

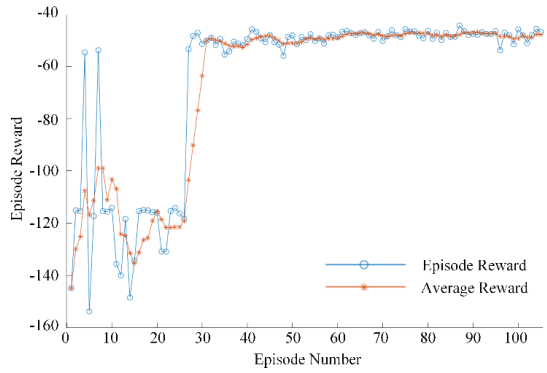
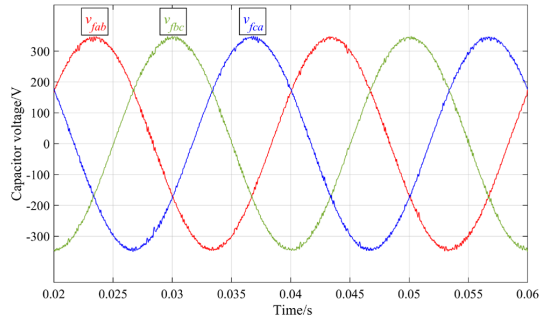
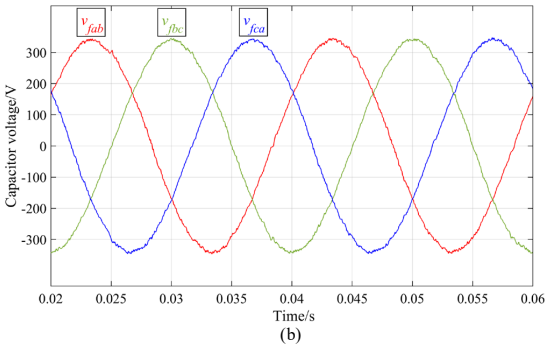


Fig. 3. Training process of DDPG agent

The capacitor voltage waveforms for the selected weighting factor λ_d = 5, 0 and RL tuned weighting factor are shown in Fig.4. And the THD of capacitor voltage is around 2.12%, 1.54% and 1.11% respectively. In addition, as the voltage error between capacitor voltage and reference voltage is introduced in the reward function, the voltage tracking error waveforms are shown in Fig.5. The improved performance proves the effectiveness of the proposed RL-based weighting factor design method. In summary, with the RL tuned weighting factor, in the FS-MPC regulated UPS system, the THD of capacitor voltage is improved and the amplitude of reference voltage tracking error is reduced.



(a)



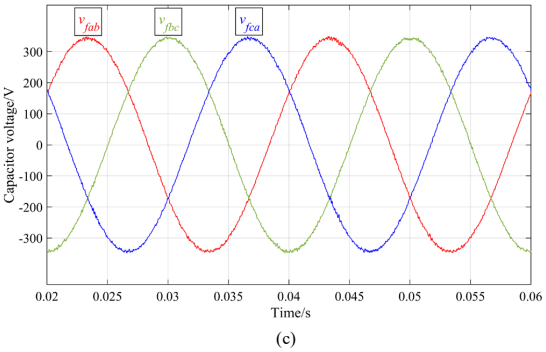
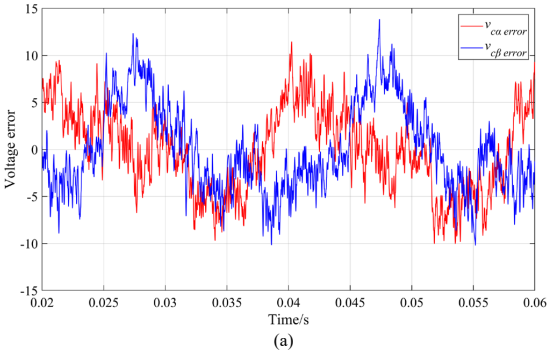
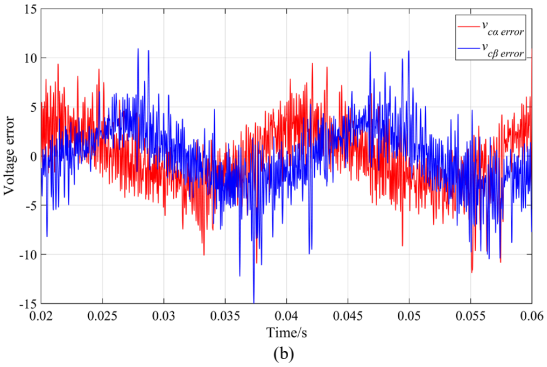


Fig. 4. Capacitor voltage waveforms. (a) With weighting factor $\lambda_d = 5$, THD = 2.12%. (b) With weighting factor $\lambda_d = 0$, THD = 1.54%. (c) With RL tuned weighting factor THD = 1.11%.





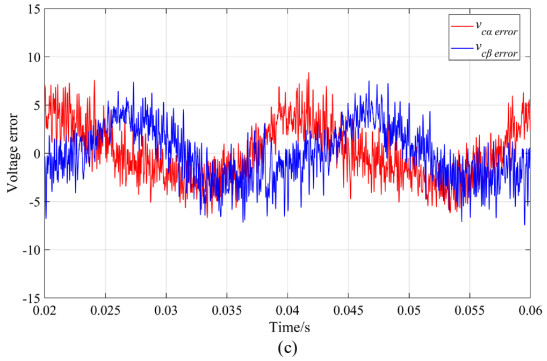


Fig. 5. Voltage reference tracking error waveform in α-β frame. (a) With weighting factor $\lambda_d = 5$. (b) With weighting factor $\lambda_d = 0$. (c) With RL tuned weighting factor.

V. Conclusion and Future Work

In this paper, a novel reinforcement learning based weighting factor design method for FS-MPC algorithm is proposed. The effectiveness of the proposed method is validated in a configuration of FS-MPC controlled two level VSC for UPS applications. The proposed method is also applicable to other MPC controlled converters. For further research, the RL weighting factor design method will be applied to more complex situations, e.g. multiple weighting factors for more complex cost function. Also, the proposed RL-based weighting factor design method will be implemented as online tuning method in practical experimental setups for further validation and the computational burden in the practical setup will be also considered.

Acknowledgment

The authors would like to thank the Chinese Scholarship Council for its foundation for the research in Technical University of Denmark, the support from Energy Technology Development and Demonstration Program (EUDP) (ACTION project, grant No. 56537).

References

[1] P. Cortes, G. Ortiz, J. I. Yuz, J. Rodriguez, S. Vazquez, and L. G.Franquelo, “Model predictive control of an inverter with output lc filter for UPS applications,” IEEE Trans. Ind. Electron., vol. 56, no. 6,pp. 1875–1883, Jun. 2009.

[2] Dragicevic, Tomislav, Sergio Vazquez, and Patrick Wheeler. "Advanced control methods for power converters in dg systems and microgrids." *IEEE Transactions on Industrial Electronics* (2020).

[3] P. Cortes, S. Kouro, B. L. Rocca, R. Vargas, J. Rodriguez, J. I. Leon, S. Vazquez, and L. G. Franquelo, “Guidelines for weighting factors design in model predictive control of power converters and drives,” in 2009 IEEE International Conference on Industrial Technology, Feb 2009, pp. 1–7.

[4] P. Zanchetta, “Heuristic multi-objective optimization for cost function weights selection in finite states model predictive control,” in 2011 Work- shop on Predictive Control of Electrical Drives and Power Electronics, Oct 2011, pp. 70–75.

[5] Dragičević T, Novak M. Weighting factor design in model predictive control of power electronic converters: An artificial neural network approach[J]. IEEE Transactions on Industrial Electronics, 2018, 66(11): 8870-8880.

[6] Fuentes, Esteban, César A. Silva, and Ralph M. Kennel. "MPC implementation of a quasi-time-optimal speed control for a PMSM

drive, with inner modulated-FS-MPC torque control." IEEE Transactions on Industrial Electronics 63.6 (2016): 3897-3905.

[7] Z. Lu, R. Zhang, L. Hu, L. Gan, J. Lin, and P. Gong, "Model predictivecontrol of induction motor based on amplitude-phase motion equation," IET Power Electron., vol. 12, no. 9, pp. 2400–2406, Aug. 2019.

[8] Hajihosseini, Mojtaba, et al. "DC/DC power converter control-based deep machine learning techniques: Real-time implementation." *IEEE Transactions on Power Electronics* 35.10 (2020): 9971-9977.

[9] Wei, Chun, et al. "Reinforcement-learning-based intelligent maximum power point tracking control for wind energy conversion systems." IEEE Transactions on Industrial Electronics 62.10 (2015): 6360-6370.

[10] Tang, Yuanhong, et al. "Artificial Intelligence-Aided Minimum Reactive Power Control for the DAB Converter Based on Harmonic Analysis Method." IEEE Transactions on Power Electronics (2021).

[11] T. Dragicevic, "Model predictive control of power converters for robust and fast operation of AC microgrids," IEEE Trans. Power Electron., vol. 33, no. 7, pp. 6304–6317, July 2018.

[12] C. Wang, J. Wang, Y. Shen, and X. Zhang, "Autonomous Navigation of UAVs in Large-Scale Complex Environments: A Deep Reinforcement Learning Approach," IEEE Transactions on Vehicular Technology, vol. 68, pp. 2124-2136, 2019.

[13] Lillicrap T P, Hunt J J, Pritzel A, et al. "Continuous control with deep reinforcement learning". arXiv preprint arXiv:1509.02971, 2015.

[14] Cao, Di, et al. "Reinforcement learning and its applications in modern power and energy systems: A review." Journal of Modern Power Systems and Clean Energy 8.6 (2020): 1029-1042.

[15] Silver, David, et al. "Deterministic policy gradient algorithms." International conference on machine learning. PMLR, 2014

[C2] Optimal dispatch schedule for a fast EV charging station with account to supplementary battery health degradation

Authors:
Yihao Wan, Daniel Gebbran, Pere Izquierdo Gomez, Tomislav Dragicevic

Submitted to:
2022 IEEE/AIAA Transportation Electrification Conference and Electric Aircraft Technologies Symposium (ITEC+EATS)

Status:
Published.
Digital Object Identifier: 10.1109/ITEC53557.2022.9813864

2022 IEEE/AIAA Transportation Electrification Conference and Electric Aircraft Technologies Symposium (ITEC+EATS) | 978-1-6654-0560-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ITEC53557.2022.9813864

Optimal dispatch schedule for a fast EV charging station with account to supplementary battery health degradation

1st Yihao Wan
Department of Wind and Energy Systems
Technical University of Denmark
Kongens Lyngby, Denmark
wanyh@dtu.dk

2rd Daniel Gebbran
Department of Wind and Energy Systems
Technical University of Denmark
Kongens Lyngby, Denmark
dgeco@dtu.dk

3th Pere Izquierdo Gómez
Department of Wind and Energy Systems
Technical University of Denmark
Kongens Lyngby, Denmark
pizgo@dtu.dk

4nd Tomislav Dragicevic
Department of Wind and Energy Systems
Technical University of Denmark
Kongens Lyngby, Denmark
tomdr@dtu.dk

***Abstract*—This paper investigates the usage of battery storage systems in a fast charging station (FCS) for participation in energy markets and charging electrical vehicles (EVs) simultaneously. In particular, we focus on optimizing the scheduling strategies to reduce the overall operational cost of the system over its lifetime by combining the model of battery degradation and energy arbitrage. We implement the battery degradation as a penalty term within an energy arbitrage model and show that the battery degradation plays an important role in the optimal energy dispatch scheduling of the FCS system. In this case study, with different penalty coefficients for the battery degradation penalty term, it is found that including the penalty of battery usage in the scheduling model will reduce the number of small charging/discharging cycles, thereby prolonging the battery lifetime, while maintaining near optimal revenue from grid services.**

***Index Terms*—Fast charging station, energy storage, battery degradation, operational cost**

I. Introduction

With the increasing electrification of transportation for reducing Greenhouse Gas (GHG) emissions, the electric vehicle (EV) market is taking off [1]. Although EVs are becoming more popular, one of the major bottlenecks for the large-scale replacement of Internal Combustion Engine (ICE) is the lack of fast charging infrastructures, particularly on the highways in between cities and rural districts. The fast charging stations are important infrastructures, especially for long-range driving. They can be installed along the highways, providing a refueling experience of EVs similar to the experience of refueling ICE vehicles [2]. However, the load profile of the FCSs is impulsive in nature, which causes adverse impacts on the stability of the grid [3]. Moreover, the FCSs are often installed in rural areas with weak grid connections, which requires upgrading electrical grid, resulting in higher installation costs.

To mitigate the impact of FCS on the grid, battery energy storage systems (BESSs) can be employed [3], [4]. BESSs play an increasingly important role in FCSs due to its participation in the followings: (i) energy price arbitrage, (ii) energy demand and supply, (iii) power supply and demand, (iv) other ancillary services, such as peak shaving, frequency regulation, etc. In particular, a combination of (ii) and (iii) for energy demand and energy supply sides can ensure power balance even under fluctuating load [5], [6]. Due to the fact that batteries have limited number of charge and discharge cycles, batteries degrade accumulatively during the operations, which should be accounted for in the operational planning problems. The battery degradation is analyzed based on the operation of the system and considered in cost analysis in [7], however, the battery degradation is not directly accounted for in the dispatch scheduling strategies. Choi *et al.* [8] proposed to limit the battery usage with a state-of-charge (SoC) term to avoid the accelerated aging of the BESS while the term is not correlated with the actual degradation of battery. In [9], [10], a coordinated charging and discharging strategy for a fast charging station is proposed to optimize the economic benefits while only the charging power is considered for the battery life expenditure cost, which doesn't reflect the actual capacity loss for the battery usage, leading to over usage or under usage of the battery over its lifetime. To solve this, Schneider *et al.* [11] proposed to implement the rainflow counting algorithm (RCA) in the scheduling model for dispatch optimization. However, due to the RCA having no analytical mathematical expression, the scheduling model doesn't explicitly include the battery degradation in the optimization problem but its effects are only analyzed post mortem. To simplify the RCA in the operation optimization problem, by approximating the battery degradation for complete and incomplete cycles with a capacity

loss function, Lee $et\ al.$ [12] proposed to formulate battery degradation in the objective function with an auxiliary SoC profile.

In summary, previous researches introduce an additional term (e.g. charging power, SoC change) to penalize the battery operation thus the battery lifetime is prolonged, while this is not directly correlated with the battery degradation model, which will lead to overuse or under use of the battery over its lifetime based on the choice of the coefficient for the formulated penalty term, or formulate the battery degradation cost term with an approximation function. To solve the aforementioned issues, we propose to implement a DoD-based degradation model which considers full charge and discharge cycles during the scheduling optimization.

In this paper, particularly, we focus on minimizing the operational cost of FCS by combining the energy arbitrage model and the battery degradation model. In particular, we integrate the approximation of battery degradation based on state-of-charge (SoC) change as a penalty term directly in the optimization problem, which is also correlated with the actual battery degradation, thereby filling the current research gap. The FCS will charge the EVs and the BESS participates in the electricity market for energy arbitrage and it also mitigates the impulse of load demand. In this way, a shorter investment payback period and long-term profits of the FCS can be realized.

II. Modeling approach

In this section, we model the operation of battery storage systems in the FCS and degradation of batteries. As we try to investigate the impact of battery degradation on the optimal control of charging/discharging for the battery to minimize the overall operational cost of the system, we combine the energy arbitrage model with the battery degradation model. The demand profile is obtained from the predicted load demand in the FCS system and the electricity prices of Denmark are known a day in advance. Based on the load demand, the offering strategy will be determined, and the battery storage system will participate in the day-ahead energy market to buy energy in cheaper periods to keep it state-of-charge. More importantly, the energy arbitrage will also be considered if proved profitable to shift the energy consumption of the charging station away from the peaks of energy markets. Therefore, the dispatching problem in the paper is to determine the optimal scheduling strategy for the FCS based on the spot market prices and demand. Thus, in the scheduling model, the main part is the daily operation cost minimization based on the load profile and the operation of the battery storage system, the other part is the battery lifetime losses due to the operations. The two parts will be combined to build the final scheduling model to balance the energy arbitrage revenue, energy consumption cost and battery degradation cost. The details of the model are elaborated in the following sections.

A. Battery operations in an FCS

Considering the operations within a finite time horizon $t \in \mathcal{T} = \{1, 2, ..., T\}$, the goal is to optimize the charge and discharge operations in the FCS at each time step t to minimize the overall operational cost for daily scheduling. The scheduling model for an FCS with a battery storage system within a day is presented in (1). In the objective function of (1a), the overall operation cost J_t of the FCS system is composed of energy consumption cost with an additional battery degradation penalty cost and the energy arbitrage revenue. It is minimized by determining the set of charging and discharging variables $\boldsymbol{P_t^{ch}}$ and $\boldsymbol{P_t^{dis}}$. The operation cost from energy consumption and energy arbitrage is implemented by multiplying the electricity spot market price p_t and the energy flow P_t^{in}, P_t^{out} from and to the grid, as formulated in the first term of the objective function. The penalty term $\triangle c$ for the battery degradation is associated with specific battery degradation model, which will be elaborated in the next subsection. The scaling factor a_k [DKK/kWh] is applied to the degradation penalty term for the operational cost of battery usage. Constraint (1b) defines the energy balance between the supply, demand side and battery energy flow. The constraints for battery operation in the FCS system are presented by (1c)-(1g), where (1c) describes the evolution of battery SoC over the time horizon with the variables P_t^{ch} and P_t^{dis}, (1d) ensures the SoC at the end is equal to an expected value SoC_{end}, (1e) limits the SoC within predefined battery SoC range, (1f) and (1g) represent the power limits for charging and discharging. C_{bat} denotes the battery capacity. The η_{ch} and η_{dis} are the battery charging and discharging efficiencies. τ represents the time resolution for optimization.

$$
\begin{aligned}
J_t = \min_{P_t^{ch}, P_t^{dis}} \quad & \sum_{t=1}^{T} (P_t^{in} - P_t^{out}) \cdot p_t + a_k \cdot \triangle c && \text{(1a)}\\
\text{s.t.} \quad & P_t^{in} - P_t^{out} = P_t^{ch} - P_t^{dis} + P_{dem}, && \text{(1b)}\\
& SOC_t = SOC_{t-1} + \frac{\tau}{C_{bat}} (\eta_{ch} P_t^{ch} - P_t^{dis}/\eta_{dis}), && \text{(1c)}\\
& SOC_{t=T} = SOC_{end}, && \text{(1d)}\\
& SOC_{min} \leq SOC_t \leq SOC_{max}, && \text{(1e)}\\
& 0 \leq P_t^{ch} \leq P_{max}, && \text{(1f)}\\
& 0 \leq P_t^{dis} \leq P_{max}. && \text{(1g)}
\end{aligned}
$$

B. Battery degradation cost

To account for the battery degradation in the energy arbitrage model, we utilized a DOD-based Peukert Lifetime Energy Throughput (PLET) model for estimating the battery capacity loss [13]. The PLET model involves the battery cycle with DOD based on Peukert's Law, where lifetime energy throughput is expressed as

$$C_{PLET}^{life} \triangleq n DOD^{k_p} \tag{2}$$

where n denotes the total number of cycles for charging and discharging within the lifetime of the battery, k_p denotes the Peukert Lifetime constant between 1.1 and 1.3.

In the PLET battery degradation model, the total energy throughput C_{PLET}^{life} obtained from experiment or datasheet is assumed to be constant for each specific DOD. Therefore, with the discretization of the equation (2), the accumulated battery degradation based on PLET model for time period T could be estimated as

$$Q_{loss}^{PLET} = \frac{1}{C_{PLET}^{life}} \sum_{t=1}^{T} (\triangle DOD(t))^{k_p} \tag{3}$$

where Q_{loss}^{PLET} represents the total capacity loss, $\triangle DOD(t)$ denotes the SoC change within the time interval t.

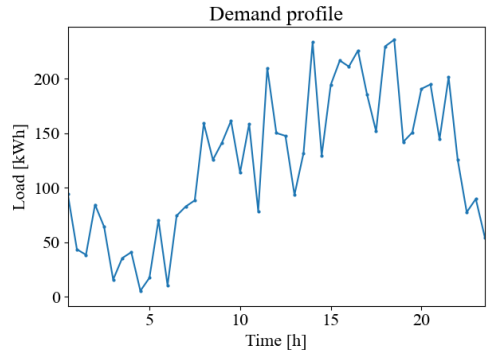
As we propose to integrate the battery degradation model in the scheduling optimization, the calculation of the total capacity loss for the battery during the optimization is needed. Therefore, the degradation model could be reformulated as

$$L_{loss}^{PLET} = \sum_{t=1}^{T} |SoC(t) - SoC(t-1)|^{k_p} \tag{4}$$

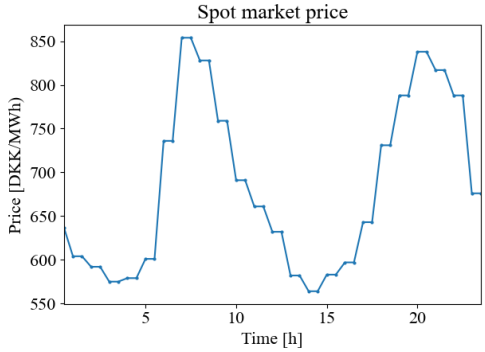
Finally, by assigning the L_{loss}^{PLET} to the $\triangle c$ in (1), the FCS scheduling model is built.

III. Case Study

In the case study, the proposed scheduling model was verified and the results were obtained based on the day-ahead spot market prices from an electricity service provider in Denmark and the predicted load data, shown in Fig. 1. For the parameters of batteries utilized in the case study, the C_{PLET}^{life} is 12500 and the $k_p = 1.15$. The maximum energy flow $P_{max} = 1$ MW. The SoC limit for the battery operation is set as SoC_{min}=0, SoC_{max}=1. The round trip charge efficiency is set as η_{ch}=η_{dis}=0.92. The time interval is 30 min in this study, therefore $T = 48$ and yields 96 decision variables. Problem (1) is implemented in Python with Pyomo as a modeling interfacing [14], and solved by IPOPT with MUMPS linear solver [15]. The problem is solved on a rolling horizon basis, i.e. every time step a new solution is derived.



(a) Load profile of FCS system.



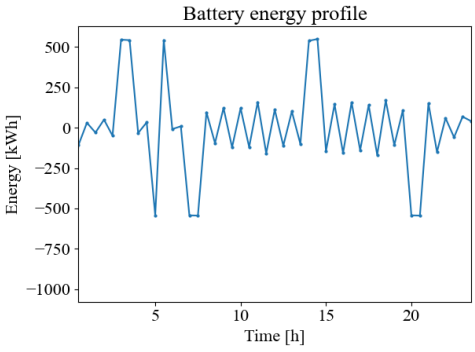
(b) Day-ahead spot market prices for a day.

Fig. 1. Predicted load profile and day-ahead electricity prices.

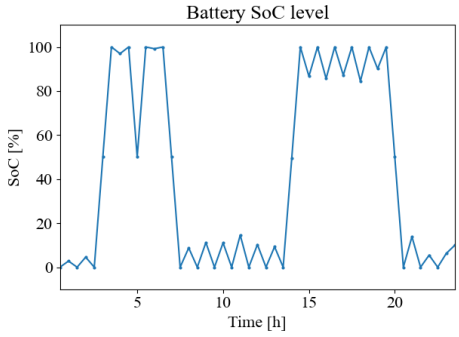
As we mentioned, there is an additional economic cost for battery usage, corresponding to the impacts of operation in long-term degradation, which is assigned by a scaling factor a_k [DKK/kWh]. To investigate the influence of battery usage on the dispatch scheduling strategy, different empirical values of the coefficient a_k are assigned. The optimal value can be further determined by comparing the optimization results with a finite set of values. When the coefficient $a_k = 0$ [DKK/kWh], which indicates the battery usage is not accounted for in the operational planning. Thus, to minimize the overall operational cost of FCS, based on the load demand and electricity prices, the battery storage system participates fully in energy arbitrage by purchasing and storing more energy during lower price periods and shifting the bulky energy consumption of FCS away from peak prices, as shown in Fig. 2.

The total energy arbitrage revenue is 1233.9 DKK. To evaluate the remaining lifetime of the battery under the operations, the yearly capacity loss is simplified based on the daily capacity loss. The battery reaches end-of-life (EOL) when the capacity is reduced by 20%. When the scheduling optimization doesn't consider the battery degradation for operation, the estimated lifetime of the battery is 11.9 years. In addition, due to the oscillation of load profile, shown in Fig. 1(a), the battery storage system will charge and discharge repeatedly to ensure energy balance. However, according to the DOD-based battery degradation model, such operations will accelerate the degradation of the battery due to the cycle aging due to the frequent change of the SoC. Therefore, by assigning a value to the coefficient a_k for the degradation penalty term, the penalization of battery usage is realized. And with different values for the coefficient, the optimal scheduling strategies will be influenced.

The coefficient is then assigned with another empirical value $a_k = 1$ [DKK/kWh]. The resulting battery operations are shown in Fig. 3. It can be observed from the battery energy profile that the battery storage system still participates in energy arbitrage and helps reducing the consumption cost of FCS during peak prices period, while the repeating charging and discharging operations are reduced during the load oscillation period, thus SoC profile of the battery becomes stable, thus

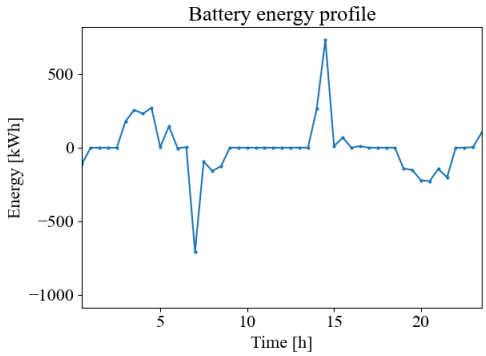


(a) Battery operation.

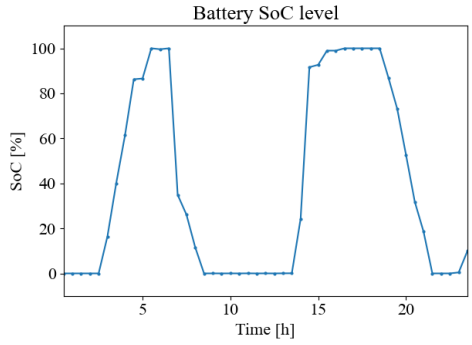


(b) Battery SoC dynamics.

Fig. 2. Optimal dispatch schedule without battery usage penalty ($a_k = 0$ DKK/kWh)



(a) Battery operation.



(b) Battery SoC dynamics.

Fig. 3. Optimal dispatch schedule when a_k=1 DKK/kWh.

the small charge and discharge cycles are reduced. In addition, the energy arbitrage revenue is reduced a bit to 1196.8 DKK. It indicates the direct energy purchase from the grid for the FCS consumption when the spot market prices are relatively low is preferred due to the penalty of battery usage. And the lifetime of battery is calculated in the same way mentioned before, which is prolonged to around 16.5 years. Moreover, as is shown in Fig. 4, when the penalty coefficient a_k increases to 10 [DKK/kWh], the battery usage is reduced further and the optimal dispatch schedule is significantly different. In this case, battery charging would be most desirable at around 3 am, 2 pm when the electricity spot market prices are lowest, and discharging at 7 am, 8 pm when the spot market prices are highest. It indicates that the optimal dispatch strategy for the battery operation is to keep the battery state-of-charge and supply to the load when the electricity price is highest. While for the other time periods, the load demand is supplied mainly by the grid and the impact of the load fluctuation will be imposed directly on the grid. The energy arbitrage revenue is 1233.8 DKK which is almost the same as that without battery degradation. And the battery lifetime is prolonged to around 16.5 years. In other words, the battery degradation is reduced at the expense of more impacts of load fluctuations on the grid.

IV. Conclusion

In this paper, an operation scheduling model for an FCS with battery storage system incorporating battery degradation is proposed. The model optimizes the battery dispatch schedules for energy arbitrage and shifts the energy consumption away from the peak prices, thus minimizing the overall operational cost while considering battery degradation. This is realized by including the degradation model based on the SoC change into the scheduling model. By assigning different values of coefficient for the penalty of battery usage, the scheduling strategies will be influenced. The case study demonstrated that adding a degradation penalty in the objective function of the scheduling model leads to a trade-off among battery degradation, energy arbitrage and consumption cost. This provides the foundation for the assessment of cost-effectiveness

of battery storage system investment in FCS for long-term operation.

REFERENCES

[1] I. G. E. Outlook, "Accelerating ambitions despite the pandemic," *International Energy Agency: Paris, France*, 2021.

[2] P. Morrissey, P. Weldon, and M. O'Mahony, "Future standard and fast charging infrastructure planning: An analysis of electric vehicle charging behaviour," *Energy Policy*, vol. 89, pp. 257–270, 2016.

[3] M. M. Mahfouz and M. R. Iravani, "Grid-integration of battery-enabled dc fast charging station for electric vehicles," *IEEE Transactions on Energy Conversion*, vol. 35, no. 1, pp. 375–385, 2019.

[4] D. Sbordone, I. Bertini, B. Di Pietra, M. C. Falvo, A. Genovese, and L. Martirano, "Ev fast charging stations and energy storage technologies: A real implementation in the smart micro grid paradigm," *Electric Power Systems Research*, vol. 120, pp. 96–108, 2015.

[5] X. Han, Y. Liang, Y. Ai, and J. Li, "Economic evaluation of a pv combined energy storage charging station based on cost estimation of second-use batteries," *Energy*, vol. 165, pp. 326–339, 2018.

[6] H. Ding, Z. Hu, and Y. Song, "Value of the energy storage system in an electric bus fast charging station," *Applied Energy*, vol. 157, pp. 630–639, 2015.

[7] L. Richard and M. Petit, "Fast charging station with battery storage system for ev: Grid services and battery degradation," in *2018 IEEE International Energy Conference (ENERGYCON)*. IEEE, 2018, pp. 1–6.

[8] S. Choi and S.-W. Min, "Optimal scheduling and operation of the ess for prosumer market environment in grid-connected industrial complex," *IEEE Transactions on Industry Applications*, vol. 54, no. 3, pp. 1949–1957, 2018.

[9] H. Chen, Z. Hu, H. Zhang, and H. Luo, "Coordinated charging and discharging strategies for plug-in electric bus fast charging station with energy storage system," *IET Generation, Transmission & Distribution*, vol. 12, no. 9, pp. 2019–2028, 2018.

[10] J. Faraji, A. Ketabi, and H. Hashemi-Dezaki, "Optimization of the scheduling and operation of prosumers considering the loss of life costs of battery storage systems," *Journal of Energy Storage*, vol. 31, p. 101655, 2020.

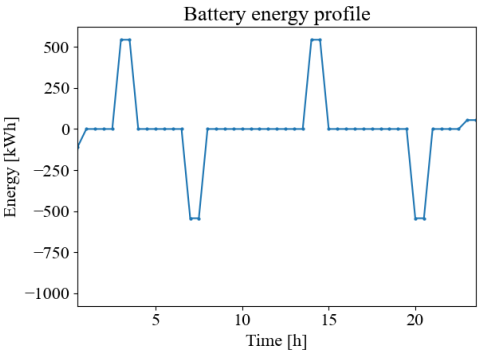
[11] S. F. Schneider, P. Novák, and T. Kober, "Rechargeable batteries for simultaneous demand peak shaving and price arbitrage business," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 1, pp. 148–157, 2020.

[12] J.-O. Lee and Y.-S. Kim, "Novel battery degradation cost formulation for optimal scheduling of battery energy storage systems," *International Journal of Electrical Power & Energy Systems*, vol. 137, p. 107795, 2022.

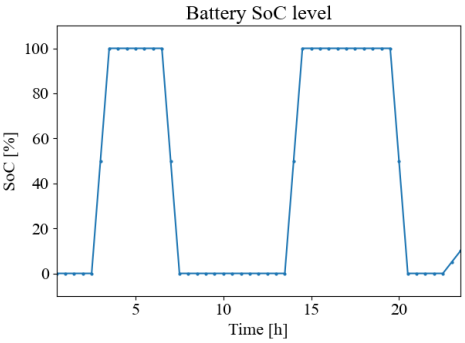
[13] D. Tran and A. M. Khambadkone, "Energy management for lifetime extension of energy storage system in micro-grid applications," *IEEE Transactions on Smart Grid*, vol. 4, no. 3, pp. 1289–1296, 2013.

[14] W. E. Hart, J.-P. Watson, and D. L. Woodruff, "Pyomo: modeling and solving mathematical programs in python," *Mathematical Programming Computation*, vol. 3, no. 3, pp. 219–260, 2011.

[15] L. T. Biegler and V. M. Zavala, "Large-scale nonlinear programming using ipopt: An integrating framework for enterprise-wide dynamic optimization," *Computers & Chemical Engineering*, vol. 33, no. 3, pp. 575–582, 2009.



(a) Battery operation.



(b) Battery SoC dynamics.

Fig. 4. Optimal dispatch schedule when a_k=10 DKK/kWh.

Division for Power and Energy Systems (PES)
Department of Wind and Energy Systems
Technical University of Denmark
Elektrovej, Building 325
2800 Kgs. Lyngby
Denmark

https://wind.dtu.dk
Tel: (+45) 4677 5085
Email: communication@windenergy.dtu.dk