University of Sannio

Department of Engineering

Master's Degree in

Electronics Engineering for Automation and Sensing

Deep Reinforcement Learning for Adaptive Bidirectional Electric Vehicle Charging Management (Vehicle-to-Grid)

Supervisor:
Prof. Carmela Bernardo

Co-Supervisor:
Dr. Antonio Pepiciello

Candidate:
Angelo Caravella
Student ID 389000016

Academic Year 2024–2025

Contents

Abstract in italian

L'adozione crescente dei **Veicoli Elettrici (EV)** in concomitanza con la sempre maggiore penetrazione di **Fonti di Energia Rinnovabile (RES)** intermittenti, presenta sfide significative alla **stabilità** e all'**efficienza della rete elettrica**. La tecnologia **Vehicle-to-Grid (V2G)** emerge come soluzione fondamentale, trasformando gli EV da carichi passivi a **risorse energetiche flessibili** capaci di fornire vari **servizi di rete**. Questa tesi affronta il complesso **problema di ottimizzazione multi-obiettivo** della gestione intelligente di carica e scarica degli EV, che intrinsecamente implica un equilibrio tra **benefici economici**, **esigenze di mobilità dell'utente**, **preservazione della salute della batteria** e **stabilità della rete** in condizioni stocastiche.
Di fronte alla complessa sfida di ottimizzare la ricarica dei veicoli elettrici (EV) in scenari Vehicle-to-Grid (V2G), un approccio che si limita a un singolo modello di controllo, come il Deep Q-Networks (DQN), risulterebbe inadeguato. La natura del problema, caratterizzata da molteplici obiettivi contrastanti (benefici economici, esigenze dell'utente, salute della batteria, stabilità della rete) e da una profonda incertezza; richiede un'analisi comparativa e rigorosa di un'ampia gamma di strategie di controllo. Per questo motivo, la ricerca si concentra sulla valutazione di un portafoglio diversificato di algoritmi, che include numerosi modelli di Deep Reinforcement Learning (DRL), approcci euristici e il Model Predictive Control (MPC). Questo metodo consente di mappare in modo completo il panorama delle soluzioni, identificando i punti di forza e di debolezza di ciascun approccio in relazione alle diverse sfaccettature del problema V2G.
In conclusione questa lavoro di tesi non si focalizza su un singolo modello, ma adotta un approccio comparativo su larga scala perché:
Non esiste una soluzione unica: La complessità del problema V2G rende improbabile che un solo algoritmo sia ottimale in tutte le condizioni.
Si ricercano i compromessi: L'obiettivo è comprendere i trade-off tra l'efficienza dei dati, la stabilità dell'addestramento, la robustezza all'incertezza e la complessità computazionale delle diverse famiglie di algoritmi.
La validazione è più rigorosa: Confrontare i modelli di DRL non solo tra loro ma anche con benchmark consolidati come le euristiche e l'MPC fornisce una misura molto più credibile del loro reale valore aggiunto.

Abstract

The growing adoption of **Electric Vehicles (EVs)**, combined with the increasing penetration of intermittent **Renewable Energy Sources (RES)**, presents significant challenges to the **stability** and **efficiency** of the power grid[1]. **Vehicle-to-Grid (V2G)** technology emerges as a key solution, transforming EVs from passive loads into **flexible energy resources** capable of providing various **grid services**. This thesis addresses the complex **multi-objective optimization problem** of smart EV charging and discharging, which requires balancing **economic benefits**, **user mobility needs**, **battery health preservation**, and **grid stability** under stochastic conditions.

Given the complexity of optimizing EV charging in V2G scenarios, relying on a single control model is insufficient. The nature of the problem, characterized by multiple conflicting objectives (economic benefits, user needs, battery health, grid stability) and profound uncertainty, demands a rigorous comparative analysis of a wide range of control strategies.

For this reason, the research focuses on evaluating a diverse portfolio of algorithms, including numerous Deep Reinforcement Learning (DRL) models, heuristic approaches, and Model Predictive Control (MPC). This method allows for a complete mapping of the solution landscape, identifying the strengths and weaknesses of each approach in relation to the different facets of the V2G problem.

In short, this thesis adopts a **broad-spectrum comparative approach** for several reasons:

No Single Solution: The complexity of the V2G problem makes it unlikely that a single algorithm can be optimal in all conditions.

Understanding Trade-offs: The goal is to understand the trade-offs between data efficiency, training stability, robustness to uncertainty, and the computational complexity of different algorithm families.

Rigorous Validation: Comparing DRL models not only against each other but also against established benchmarks like heuristics and MPC provides a more credible measure of their true value.

List of Acronyms

Acronym	Description
Artificial Intelligence & Control	
A2C	Advantage Actor-Critic
AC	Actor-Critic
AI	Artificial Intelligence
AL-SAC	Augmented Lagrangian Soft Actor-Critic
ARS	Augmented Random Search
CL	Curriculum Learning

[1] 39.

Acronym	Description
CMDP	Constrained Markov Decision Process
DDPG	Deep Deterministic Policy Gradient
DQN	Deep Q-Networks
DRL	Deep Reinforcement Learning
LQR	Linear Quadratic Regulator
LSTM	Long Short-Term Memory
MARL	Multi-Agent Reinforcement Learning
MDP	Markov Decision Process
MILP	Mixed-Integer Linear Program
MPC	Model Predictive Control
NN	Neural Network
PER	Prioritized Experience Replay
PPO	Proximal Policy Optimization
RL	Reinforcement Learning
SAC	Soft Actor-Critic
TD3	Twin-Delayed Deep Deterministic Policy Gradient
TQC	Truncated Quantile Critics
TRPO	Trust Region Policy Optimization
Electric Vehicles & Charging	
AFAP	As Fast As Possible (Heuristic)
ALAP	As Late As Possible (Heuristic)
CAFA	Charge As Fast As Possible
CALA	Charge As Late As Possible
CPO	Charge Point Operator
EV	Electric Vehicle
G2V	Grid-to-Vehicle
SCP	Scheduled Charging Power
SoC	State of Charge
SoH	State of Health
V2B	Vehicle-to-Building
V2G	Vehicle-to-Grid
V2H	Vehicle-to-Home
V2M	Vehicle-to-Microgrid
V2V	Vehicle-to-Vehicle
VPP	Virtual Power Plant
Power Grid & Energy Markets	
ACE	Area Control Error
ARR	Area Regulation Requirement
DER	Distributed Energy Resources
DR	Demand Response
RES	Renewable Energy Sources
Metrics & Technical Parameters	
DC	Constant Current (charging phase)
CV	Constant Voltage (charging phase)
DoD	Depth of Discharge

Acronym	**Description**
MSE	Mean Square Error
OU	Ornstein-Uhlenbeck (stochastic process)
RMSE	Root Mean Square Error

Chapter 1

Introduction

The shift toward electric mobility is a pivotal element in global strategies for decarbonizing transport. However, the large-scale integration of electric vehicles into existing power grids presents a complex set of challenges and opportunities, which this thesis aims to investigate.

1.0.1 Background and Relevance of Electric Vehicles and Vehicle-to-Grid

The surging Electric Vehicle (EV) market is accelerating a major reconfiguration of modern mobility, promising to lower carbon emissions and foster greater energy efficiency[1]. This evolution is more than a technological trend: it underpins environmental sustainability by reducing dependence on fossil resources, alleviating the impacts of climate change through diminished greenhouse gas emissions, and improving air quality in densely populated areas. However, embedding millions of EVs into existing power systems is a non-trivial task. It can intensify peak demand, place additional stress on transmission and distribution networks, and trigger side effects such as voltage irregularities or higher line losses[2].
It is in this context that the **Vehicle-to-Grid (V2G)** concept emerges as a strategic solution. Through bidirectional power exchange, V2G redefines EVs: no longer passive electrical loads, but mobile and flexible energy assets, able to deliver a spectrum of services to the power system[3]. This potential becomes even more compelling when one considers that, on average, EVs remain parked and unused for nearly 96% of the day, offering an ample time window to actively engage with the grid[4]. A further distinctive benefit lies in the rapid responsiveness of EV batteries, which makes them especially suitable for ancillary services demanding quick interventions, such as frequency regulation[5]. Alongside V2G, other schemes of bidirectional power flow have been proposed, each with its own scope:

1. **Vehicle-to-Home (V2H)**, where an EV sustains household demand during

[1] 27.
[2] 27, 33.
[3] 1.
[4] 18.
[5] 1.

outages or periods of elevated prices, strengthening domestic energy resilience;

2. **Vehicle-to-Building (V2B)**, extending this logic to commercial or industrial facilities, enabling EVs to support load management and improve consumption efficiency; and

3. **Vehicle-to-Vehicle (V2V)**, which allows direct power transfer among EVs, a valuable feature for emergency charging or shared resources.

Taken together, these modalities highlight the versatility of EV batteries as distributed energy units, reinforcing both energy resilience and the transition toward a more sustainable energy ecosystem.

1.0.2 Challenges in EV Integration into the Electricity Grid and the Role of Artificial Intelligence

Modern electricity systems are increasingly shaped by the penetration of intermittent **Renewable Energy Sources (RESs)** such as wind and solar. Their variability causes pronounced swings in power generation and persistent mismatches between supply and demand, which in turn fuels price volatility and complicates dispatch strategies. This continuously puts the stability and economic efficiency of the grid under strain. Managing these fluctuations, while making rapid operational choices to balance the system and minimize costs, has proven difficult for conventional control frameworks[6].

The parallel rise of EV adoption and RES deployment has produced an environment marked by both uncertainty and complexity. In such conditions, traditional approaches are increasingly inadequate, prompting a growing reliance on methods rooted in artificial intelligence—and particularly in **Reinforcement Learning (RL)**. This shift alters the very nature of the grid: from a relatively predictable and centralized infrastructure to one that is decentralized, stochastic, and highly dynamic. Rule-based or deterministic controllers, designed for a past paradigm, are ill-suited to cope with this degree of volatility. The result is a pressing need for more adaptive and intelligent decision-making mechanisms. This transformation extends beyond the simple challenge of absorbing extra load or integrating new generators: it signals a genuine paradigm change towards a *smart grid*[7], where adaptive, real-time, and autonomous operation is no longer optional but vital to preserve efficiency, resilience, and reliability. From this perspective, RL is not just an optimization tool, but an enabling technology for a more cognitive and robust energy infrastructure—one capable of navigating the uncertainties of a decarbonized, electrified future.

Against this backdrop, **Deep Reinforcement Learning (DRL)** has gained attention as an especially powerful approach. Its capacity to derive near-optimal strategies in dynamic and uncertain environments—without requiring a precise model of the system or flawless forecasts—makes DRL particularly well-suited for EV integration and advanced grid management[8].

[6] 27, 24.
[7] 13.
[8] 27, 38.

1.0.3 Objectives and Contributions of the Thesis

This thesis confronts the complex multi-objective optimization problem at the heart of Vehicle-to-Grid (V2G) systems. The overarching objective is to move beyond a purely theoretical analysis by actively developing, testing, and enhancing a high-fidelity simulation architecture. This platform serves as a digital twin to rigorously evaluate and compare advanced control strategies, balancing economic benefits, user mobility needs, battery health, and grid stability under realistic stochastic conditions.
More than a simple review of existing literature, this work focuses on the practical implementation and validation of a V2G simulation framework in Python. This tool is leveraged to demonstrate and explore novel perspectives for training intelligent agents. The main contributions are:

- **Enhancement of a V2G Simulation Architecture:** A key contribution is the systematic testing, validation, and enhancement of the **EV2Gym** simulation framework. This work solidifies its role as a robust and flexible platform for benchmarking control algorithms, ensuring that the models for battery physics, user behavior, and grid dynamics are coherent and realistic for advanced research.

- **Exploration of Novel Reinforcement Learning Perspectives:** The validated simulation environment is used to investigate and implement advanced training methodologies for RL agents. A key focus is placed on techniques like **adaptive reward shaping**, where the reward function dynamically evolves during training to guide the agent towards a more holistic and robust control policy, overcoming the limitations of static reward definitions.

- **Practical Implementation of Advanced Control Paradigms:** The thesis demonstrates the practical transition from a theoretical, offline optimal controller to a realistic, online controller. Specifically, it details the implementation of an **offline MPC using Gurobi**, which acts as a "judge" with perfect foresight, and contrasts it with an **online MPC formulated in PuLP**, designed to operate as a real-time "controller" with limited future information, highlighting the trade-offs and challenges of real-world deployment.

1.0.4 Research Methodology

The research conducted for this thesis was guided by a central question: how can the economic profit of Vehicle-to-Grid (V2G) operations be maximized without imposing undue costs in terms of battery degradation or creating instability by overloading local grid infrastructure? To address this, a systematic research methodology was adopted.

The initial phase consisted of an extensive literature review, primarily leveraging Google Scholar as a search engine for academic papers. This review focused on identifying the state-of-the-art in V2G control strategies, with a particular emphasis on identifying the common trade-offs between profitability and physical constraints.

Following the literature review, this work transitioned to an empirical phase centered on the simulation framework detailed in Chapter 3. The development and testing of various control agents were not performed in a vacuum. The significance of the obtained results was continuously evaluated through a multi-faceted validation process, which included:

- **Comparative Analysis:** Benchmarking the performance of novel agents against the results reported by other researchers in the field.
- **Heuristic Evaluation:** Assessing the outcomes based on the experience and intuition developed over the course of the study, ensuring that the results were not only quantitatively strong but also qualitatively sound.
- **Academic Grounding:** Constantly relating the empirical results back to the foundational knowledge acquired during the literature review to understand the underlying reasons for an agent's success or failure.

This iterative process of exploration, implementation, and critical evaluation forms the methodological backbone of the contributions presented in this thesis.

1.0.5 Thesis Structure

The remainder of this thesis is organized as follows:

- **Chapter 2: Overview of Optimal Management of EV Charging and Discharging** provides foundational knowledge on V2G technology, the complex multi-objective nature of EV charging optimization, and presents a comprehensive review of state-of-the-art research approaches.

- **Chapter 3: The V2G Simulation Framework: A Digital Twin for V2G Research** details the architecture and core models of the simulation environment. This chapter describes the enhancements made to the framework, establishing it as the central experimental platform for implementing and evaluating the control agents analyzed in this work.

- **Chapter 4: Experimental Campaign and Results Analysis** This chapter presents the results of the comparative analysis between the different control strategies (DRL, MPC, heuristics). It analyzes the performance of novel training techniques and discusses the implications of the findings.

- **Bibliography** lists all cited references.

Chapter 2

State of the Art in Optimal V2G Management

2.1 The V2G Imperative: A Foundation of Europe's Green Transition

Europe finds itself at the confluence of two unprecedented transformations: the wholesale electrification of transport and a fundamental restructuring of energy systems. These are not merely aspirational goals but constitute binding legal obligations enshrined in the **European Green Deal** and its comprehensive **"Fit for 55"** legislative framework[1]. The policy architecture demands a 55% reduction in net greenhouse gas emissions by 2030, necessitating the rapid elimination of internal combustion engines alongside a dramatic expansion of renewable energy capacity, as outlined in the revised **Renewable Energy Directive** (RED III). Electric Vehicles (EVs) occupy a central position in this transition, simultaneously driving decarbonisation efforts while presenting complex challenges for grid stability and management.

	RED II (2018)	RED III (2023)
RENEWABLE ENERGY TARGET	32% by 2030	42.5% by 2030 (45% aspirational target)
TRANSPORT SECTOR TARGET	14% renewable energy	29% renewable energy, 14.5% GHG intensity reduction in transport
GHG SAVINGS THRESHOLD FOR BIOFUELS	50–65% depending on installation date	70% (existing), 80% (new installations)
MASS BALANCE TRACEABILITY	Encouraged	Mandatory
ENFORCEABILITY	Partially voluntary or indicative	Legally binding and auditable
CHAIN OF CUSTODY SYSTEMS	Not required	Required across the entire value chain
ALIGNMENT WITH OTHER EU LEGISLATION	Limited	Integrated with ETS, CBAM, and EUDR frameworks

CIRCULARISE

Figure 2.1: RED 3

The initial response to mass EV adoption was characterised by considerable concern

[1] 8.

within the power sector. The prospect of millions of new electric vehicles was perceived primarily through the lens of risk—vast, temporally correlated loads threatening to overwhelm local distribution networks during peak evening hours. This perspective has undergone a fundamental reassessment. EVs are now recognised not as burdens to be managed, but as essential infrastructure for achieving Europe's energy transition objectives. This conceptual shift finds its most concrete expression in **Vehicle-to-Grid (V2G)** technology, which fundamentally reimagines the role of electric vehicles within the energy system. V2G technology transforms what were previously passive, unidirectional energy consumers into active, distributed, and smart grid resources. The underlying opportunity is both elegant and substantial: private vehicles spend approximately 96% of their operational lifetime parked and connected[2], representing an enormous, geographically distributed, and currently underutilised repository of mobile energy storage capacity. The transformative potential of V2G becomes apparent when individual vehicles are coordinated through centrally managed aggregation. While a single EV's contribution remains modest, a carefully orchestrated fleet can function as a unified entity—a **Virtual Power Plant (VPP)**. These software-defined power plants aggregate the collective capacity of numerous distributed energy resources, delivering grid services at scales and reliability levels comparable to conventional generation facilities. The rapid response characteristics of contemporary battery inverters, operating at millisecond timescales, enable these aggregated fleets to provide a comprehensive range of critical grid services. This capability extends beyond mere utility; it represents a prerequisite for maintaining stability in grids increasingly dependent on the variable and non-dispatchable output of wind and solar generation, thereby enabling the technical and economic viability of the EU's ambitious renewable energy targets[3]. The grid services enabled by V2G technology form the technical foundation for the smart, resilient, and decarbonised electricity system required for Europe's energy future:

- **Frequency Regulation:** Grid stability fundamentally depends on maintaining precise equilibrium between electricity supply and demand, manifested as stable grid frequency (50 Hz across European networks). Frequency deviations signal supply-demand imbalances that, if uncorrected, can trigger cascading failures across interconnected systems. V2G fleets, leveraging their rapid response capabilities, can participate directly in ancillary service markets including Frequency Containment Reserve (FCR) and automatic Frequency Restoration Reserve (aFRR). These systems can inject or absorb power within seconds of frequency deviations, providing immediate counteraction to imbalances and preventing system-wide failures or blackouts[4].

- **Demand Response and Peak Shaving:** Through intelligent temporal shifting of charging activities to off-peak periods—when energy is abundant and inexpensive—combined with strategic discharging during peak demand when energy becomes scarce and costly, V2G systems can effectively flatten daily

[2] 18.
[3] 39.
[4] 1, 43.

load profiles. This approach directly addresses the "duck curve" phenomenon associated with high solar photovoltaic penetration. Such load profile management reduces dependence on expensive and carbon-intensive "peaker" plants, typically gas or diesel turbine facilities, while potentially deferring or eliminating requirements for costly transmission and distribution infrastructure upgrades[5].

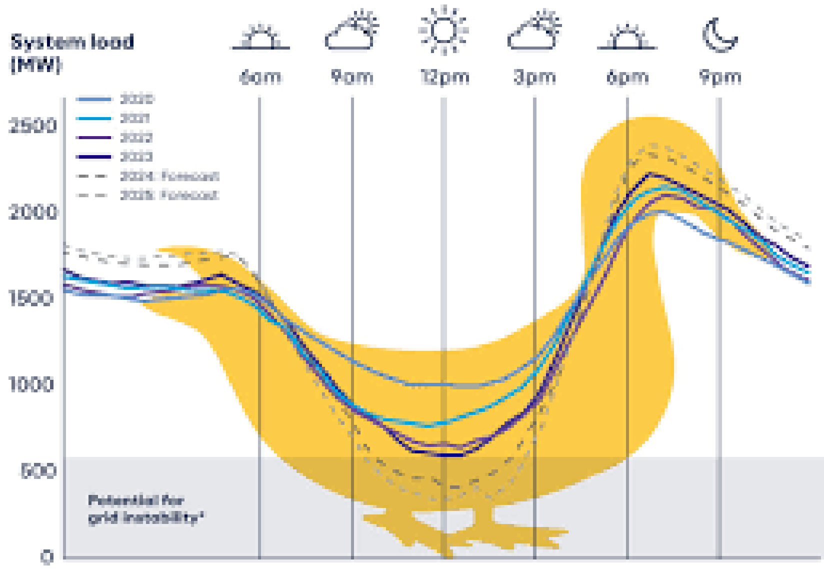


Figure 2.2: Duck curve

- **Renewable Energy Integration:** The most strategically significant contribution of V2G lies in addressing renewable energy intermittency challenges. V2G fleets function as large-scale energy buffers, absorbing excess solar and wind generation during periods of high production and low demand that would otherwise require curtailment. This stored renewable energy can subsequently be released during periods of low generation, such as evening hours or periods of limited wind availability. This mechanism directly increases the utilisation efficiency of renewable energy resources, supporting the integration objectives outlined in RED III while enhancing overall energy system efficiency[6].

This vision is being actively incorporated into European legal frameworks and technical standards. The transformative **Alternative Fuels Infrastructure Regulation (AFIR, EU 2023/1804)** now requires that new public charging infrastructure incorporate smart charging capabilities and, critically, bidirectional power flow capacity. The technical implementation of these requirements is supported by standards such as **ISO 15118-20**, which provides detailed specifications for "Vehicle-to-Grid Communication Interface" (V2GCI) protocols, enabling secure, interoperable bidirectional power transfer. With mandatory implementation scheduled for 2027,

[5] 27, 32.
[6] 15, 46.

the necessary technological and regulatory infrastructure is being systematically established. This regulatory push is supported by significant pilot initiatives including the **'SCALE'** and **'V2G Balearic Islands'** projects, which are conducting comprehensive testing of the technology's technical performance and economic viability under real-world operating conditions.
Despite this regulatory progress, substantial obstacles to widespread V2G deployment persist. These challenges span technical, economic, and social dimensions:

- **Market and Economic Barriers:** A coherent, pan-European framework for compensating EV owners for grid service provision remains undeveloped. While value streams exist within energy markets, accessing these markets involves significant complexity. Critical issues such as **"double taxation"** of electricity—where energy faces taxation both during charging and discharging phases—create substantial economic disincentives that require resolution through coordinated policy intervention.

- **Regulatory and Grid Access Challenges:** The recognition of EV fleets as qualified flexibility resources varies significantly across national electricity markets. Standardised procedures for grid interconnection, aggregator certification, secure data exchange protocols, and baseline calculation methodologies remain under development, creating a fragmented operational environment that complicates commercial deployment.

- **Technical and Consumer Adoption Barriers:** Consumer concerns regarding accelerated **battery degradation** and its implications for vehicle warranty coverage represent primary obstacles to participation. Additionally, the current installed base of EVs and charging infrastructure lacks universal bidirectional capability, although this limitation is being addressed through new vehicle platforms and evolving charging standards.

The fundamental challenge addressed by this thesis extends beyond simply enabling V2G technology to encompass its *intelligent orchestration.* This requires developing control strategies sophisticated enough to operate within emerging regulatory frameworks, navigate economic uncertainties, accommodate diverse user preferences, and overcome existing technical constraints. The objective is to unlock the substantial potential of EVs as fundamental components of Europe's energy transition while ensuring system reliability, economic viability, and user acceptance.

2.2 The Optimizer's Trilemma: Navigating a Stochastic World

While the potential of V2G technology is substantial, the management of distributed vehicular assets presents a complex control challenge. Economic viability drives aggregator decisions, yet a narrow focus on profitability alone proves insufficient for sustainable operations. Effective V2G management requires balancing three competing objectives that frequently conflict with one another. This challenge can be framed as the "V2G Optimizer's Trilemma": the concurrent pursuit of **economic profitability**, preservation of **battery longevity**, and maintenance of **user convenience**. Rather than representing a straightforward, static trade-off, this constitutes a dynamic, multi-objective optimisation challenge characterised by **stochasticity** and **uncertainty** arising from multiple, interconnected sources[7]:

- **Market Volatility:** Wholesale electricity prices exhibit significant variability driven by unpredictable supply variations (such as sudden reductions in wind generation capacity) and demand fluctuations (including heat-driven increases in cooling demand). Effective control systems must respond to these price signals dynamically and in real-time.
- **Renewable Intermittency:** Co-located solar and wind generation exhibit inherently variable output patterns with limited predictability. Controllers must coordinate EV fleet operations to capture available generation during surplus periods without compromising other operational objectives.
- **Human Behaviour:** Perhaps the most challenging uncertainty source involves EV owner patterns. Arrival times, departure schedules, and required state of charge (SoC) at departure lack deterministic characteristics. Emergency departures or unexpected schedule changes represent hard, non-negotiable constraints that intelligent systems must accommodate to preserve user trust and satisfaction.

This dynamic, uncertain, and multifaceted operational environment renders static, rule-based control approaches (such as "charge when price falls below threshold X, discharge when exceeding threshold Y") inadequate and brittle. More sophisticated and adaptive methodologies are required—approaches capable of learning from operational experience and making optimal decisions under conditions of significant uncertainty. Reinforcement Learning excels in precisely this domain, providing a framework for developing control policies that demonstrate robustness, adaptability, and scalability.

2.3 The Economics of V2G: Navigating a Spectrum of Value Streams

The economic viability of Vehicle-to-Grid services constitutes the fundamental driver for investment and widespread adoption. This viability depends on aggregators' ability to capture value across diverse electricity markets. While energy arbitrage represents the most intuitive revenue stream, a comprehensive analysis reveals a broader and potentially more lucrative economic landscape.

2.3.1 Beyond Arbitrage: A Spectrum of Value Streams

V2G-enabled fleets can participate across multiple markets simultaneously, enabling revenue stream stacking. The primary categories include:

[7] 42.

- **Energy Markets (Arbitrage):** This strategy involves purchasing energy during low-price periods (typically overnight) and selling during high-price periods (such as evening peaks). Profitability correlates directly with **Day-Ahead** and **Intraday** market price volatility. Despite conceptual simplicity, this market presents challenges due to retail-wholesale price spreads.
- **Ancillary Services Markets:** These markets procure services essential for maintaining grid stability and security. V2G technology demonstrates exceptional suitability for these applications due to rapid response capabilities. Key products include:
  - **Frequency Response:** As previously discussed, this involves rapid power injection or absorption for frequency stabilisation. Compensation typically combines availability payments (capacity-based in €/MW/h) with activation payments (energy-based). This often provides more reliable revenue streams than pure arbitrage operations.
  - **Voltage Support:** Involves injecting or absorbing reactive power to maintain local voltage levels within acceptable operational ranges.
- **Capacity Markets:** These markets compensate participants for committing to power generation or load reduction availability during system stress periods. EV Virtual Power Plants can bid aggregated capacity into these markets, securing steady, long-term revenue streams based on availability commitments, regardless of actual activation frequency.

Truly optimal V2G strategies must therefore co-optimise across these diverse markets, significantly increasing control problem complexity.

2.3.2 Sources for Energy Price Data

Access to reliable, real-time, and historical market data remains crucial for both control agent training in simulation environments and real-world deployment. Key public sources for European market data include:

- **ENTSO-E Transparency Platform:** The European Network of Transmission System Operators for Electricity maintains a mandatory, open-access platform serving as a comprehensive repository of pan-European electricity market data. This includes harmonised day-ahead prices, load forecasts, and generation data, serving as the primary source for academic research through both web portal access and free RESTful API services.
- **National Transmission System Operators (TSOs):** Many national TSOs (including Terna in Italy, National Grid in the UK, and RTE in France) publish detailed market data covering real-time frequency and imbalance prices for their respective jurisdictions.
- **Power Exchanges:** Exchanges such as **EPEX SPOT** and **Nord Pool** constitute actual trading venues. While they represent direct price data sources, comprehensive real-time access typically requires commercial subscription services.

2.3.3 Buying vs. Selling: The Critical Retail-Wholesale Spread

A critical yet frequently overlooked aspect of V2G economics involves the distinction between EV owner charging costs and aggregator grid sales revenue.

- **Selling Price (V2G Revenue):** When EVs provide energy to the grid, revenue calculation bases on **wholesale prices** (such as day-ahead spot prices). These prices reflect pure marginal energy costs at specific times.
- **Buying Price (Charging Cost):** End consumer EV charging costs reflect **retail prices**, significantly exceeding wholesale prices due to numerous non-energy components, termed "non-commodity costs":
  - Base wholesale energy costs
  - **Grid Tariffs:** Charges for high-voltage transmission and low-voltage distribution network usage
  - **Taxes and Levies:** National or regional taxation including VAT and environmental levies applied to electricity consumption
  - **Supplier Margin:** Retail energy provider profit margins

This substantial gap between retail purchasing prices and wholesale selling prices constitutes the "retail-wholesale spread," creating the primary opportunity for profitable energy arbitrage. For instance, with wholesale prices at €50/MWh and retail prices at €250/MWh, arbitrage operations achieve profitability only when selling prices exceed the full €250/MWh acquisition cost, not merely the €50/MWh wholesale component. Successful control strategies must account for these price differentials to enable economically rational decision-making.

A further perspective on this issue is provided by Parisio et al.[8] (2014), who develop a model predictive control (MPC) framework for microgrid operation. Their formulation explicitly considers the decision of when to buy from or sell to the utility grid, under time-varying spot prices and operational constraints. Importantly, the model prevents physically and economically unrealistic behaviors such as simultaneous buying and selling, and accounts for the real cost of storage and generation. This reinforces the notion that optimal energy management strategies must capture the full set of economic signals—including retail charges, network tariffs, and non-commodity costs—rather than relying solely on wholesale price arbitrage. In the V2G context, this highlights the need for predictive, multi-constraint optimization frameworks capable of managing battery limitations, retail–wholesale spreads, and market participation simultaneously in order to ensure profitability.

[8] 29.

2.4 Modelling the V2G Ecosystem

Before examining control algorithms, establishing clear, high-fidelity models of system core components becomes essential: the electric vehicle as a controllable cyber-physical asset, and the operational environment or "scenario." The interaction between these elements defines V2G optimisation task boundaries and objectives.

2.4.1 The Grid-Interactive EV as a Controllable Asset

From a power grid perspective, electric vehicles represent sophisticated mobile energy storage devices. For V2G applications, EVs can be characterised through several key state variables and parameters:
The Battery: The core grid asset component, defined by **nominal energy capacity** (in kWh), current **State of Charge (SoC)**, and **State of Health (SoH)** representing degradation over time. Operation is constrained by **power limits** (in kW) dictating maximum charge or discharge rates, and charging/discharging **efficiencies** accounting for energy losses.
The On-Board Charger (OBC): For AC charging applications, the OBC converts grid alternating current to battery direct current. Power rating often constitutes the primary bottleneck for both charging and V2G power output.
Communication Interface: V2G participation requires vehicle-charging station (EVSE) communication capabilities. This is governed by standards including **ISO 15118** and protocols such as the **Open Charge Point Protocol (OCPP)**, enabling secure information exchange required for smart and bidirectional power flow operations.
Combined with vehicle availability patterns—arrival and departure times plus user energy requirements—these characteristics transform EVs from simple loads into fully dispatchable grid resources.

2.5 A New Paradigm for Control: Reinforcement Learning

To address the complexities of uncertainty, multi-objective trade-offs, and dynamic systems, this work employs Reinforcement Learning (RL), a machine learning paradigm focused on learning optimal sequential decision-making policies through trial and error. Unlike traditional optimal control methods that require explicit and perfect models of the environment, RL learns control policies directly from interactions, offering robustness to uncertainty and unmodeled dynamics [44, 30, 45].

2.5.1 The Language of Learning: Markov Decision Processes (MDPs)

The mathematical foundation of RL is based on the **Markov Decision Process (MDP)**, a formal framework for modeling decision-making in stochastic environments. An MDP is defined by the tuple (S, A, p, R, γ), where, in the V2G context:

- S: The **state space**. A comprehensive snapshot of the world at any given moment, including the state of charge (SoC) of all connected electric vehicles (EVs), current electricity prices, time of day, and other relevant variables.
- A: The **action space**. The set of all possible agent decisions. In V2G applications, this typically represents a continuous space indicating charging or discharging power for each EV.
- $p(s', r|s, a)$: The **transition dynamics model**. This function defines the probability of transitioning to new state s' and receiving reward r, given current state s and action a. In model-free RL, this function remains unknown to the agent.
- R: The **reward function**. A scalar feedback signal indicating action quality. This represents the most critical design element, implicitly defining agent objectives.
- γ: The **discount factor** ($0 \leq \gamma < 1$). This value determines the present value of future rewards, balancing immediate gratification against long-term gains.

This framework relies on the **Markov Property**, assuming future independence from the past given the present. This enables agents to make decisions based solely on current states, without requiring complete historical information of previous states and actions [30].

The **Markov property** is **causal**, in the sense that the future state depends *only* on the current state and not on past states:

$$P(S_{t+1} \mid S_t, S_{t-1}, \ldots, S_0) = P(S_{t+1} \mid S_t) \tag{2.1}$$

- **Causal:** The future state depends solely on the current state and action.

- **Non-causal:** This would imply that the future state could depend on "future" information or arbitrary past states, which is not allowed in Markov processes.

Thus, the Markov property encapsulates a **causal model of system evolution**, relying only on the present state without needing to store the entire history. This causality is fundamental in RL because it allows agents to make decisions based solely on the current state, greatly simplifying learning and computation.

2.5.2 The Bellman Equations

The agent's objective involves learning a **policy**, $\pi(a|s)$, representing a strategy for action selection to maximize expected discounted cumulative reward. To achieve this, it learns **value functions** estimating long-term value. The relationship between state values and successor state values is defined by the **Bellman equations**. The **Bellman expectation equation** for the state-value function $v_\pi(s)$ expresses state value under policy π as the sum of immediate reward and discounted next state value:

$$v_\pi(s) = \mathbb{E}_\pi \left[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s\right] \tag{2.2}$$

The ultimate objective involves finding the **optimal policy**, π_*, with corresponding optimal value function, $v_*(s)$. The **Bellman optimality equation** states that optimal policy state value must equal the expected return for the best action from that state:

$$v_*(s) = \max_a \mathbb{E} \left[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a\right] \tag{2.3}$$

RL algorithms can be viewed as iterative methods for solving this equation and thereby discovering optimal policies [44].

2.5.3 Actor-Critic Architectures

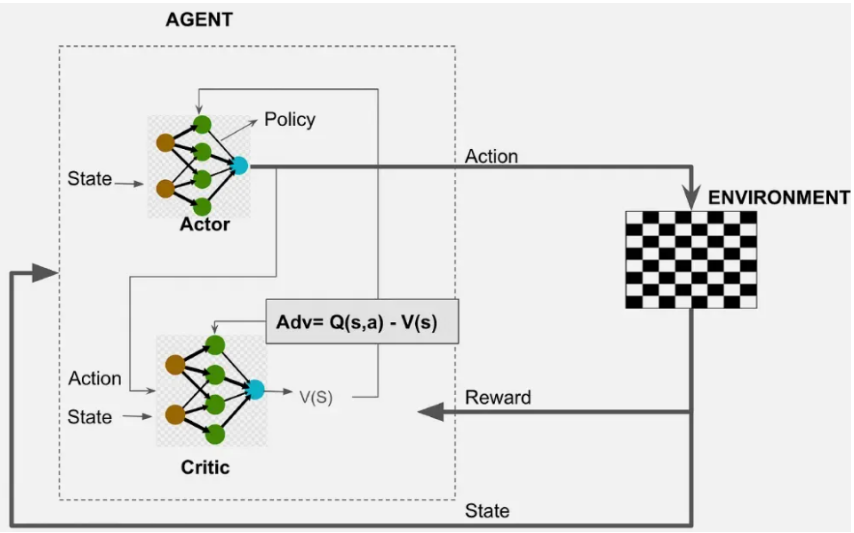


Figure 2.3: Actor critic

The **Actor-Critic** architecture provides a powerful and widely adopted method for solving RL problems, particularly in continuous action spaces. It employs two distinct neural networks:

- **The Critic**: Learns a value function (typically the action-value function, $q_\pi(s, a)$). Its role involves evaluating actor decisions, providing supervisory signals in the form of Temporal Difference (TD) errors.
- **The Actor**: Represents the policy, $\pi(a|s)$. It receives current state inputs and outputs actions. It utilizes critic feedback—TD error—to update parameters via policy gradients, improving strategies over time.

This architecture suits V2G applications well because it can directly learn policies over continuous action spaces, enabling fine-grained power control. Agent behavior is shaped entirely by received reward signals. Designing these signals to align agent objectives with desired multi-faceted goals represents a critical discipline known as reward engineering [44, 45].

2.6 Reward Engineering: Shaping Agent Behavior

The reward function constitutes perhaps the most crucial element in any Reinforcement Learning system, serving as the primary mechanism through which designers communicate desired behaviors to learning agents. Poor reward design can result in unintended and suboptimal behaviors, even when agents successfully maximise their assigned objectives. The field of reward engineering has emerged as a fundamental aspect of modern RL, proving essential for effective algorithm application in complex, real-world scenarios[9]. The V2G challenge, characterised by multiple competing objectives—profit maximisation, user satisfaction assurance, battery health preservation, and grid stability maintenance—presents particularly demanding reward design requirements. This work investigates several sophisticated techniques for effective learning guidance.

2.6.1 Potential-Based Reward Shaping (PBRS)

Potential-Based Reward Shaping (PBRS) represents one of the most theoretically robust methods for reward augmentation[10]. The approach involves supplementing the environment's intrinsic reward, $R(s, a, s')$, with a shaping term, $F(s, s')$. The resulting shaped reward R' is defined as:

$$R'(s, a, s') = R(s, a, s') + F(s, s')$$

The shaping term itself derives from the difference between a potential function, Φ, evaluated at successive states:

$$F(s, s') = \gamma\Phi(s') - \Phi(s)$$

[9] 14.
[10] 26.

where γ represents the discount factor. A key property of PBRS, as highlighted by Ng et al.[11], is **policy invariance**: adding a potential-based shaping reward does not change the optimal policy of the underlying MDP. In their paper, Ng et al. explicitly state that ""**for any MDP and any potential function Φ, the optimal policy for the shaped reward function R' is identical to the optimal policy for the original reward** R"". This ensures that while the agent may experience denser feedback and faster learning, the long-term optimal behavior remains exactly the same.
By providing intermediate guidance through the shaping term, PBRS effectively accelerates learning without compromising the correctness of the learned policy.

2.6.2 Dynamic and Adaptive Rewards

In contrast to PBRS, which provides static reward bonuses, dynamic or adaptive reward functions can evolve throughout the training process. This approach proves particularly valuable for complex problems where the relative importance of different objectives may shift as agent competency develops. For instance, an agent might initially receive rewards simply for maintaining EV charge levels, but as training progresses, the reward function could adapt to incorporate penalties for grid overloads or incentives for V2G service provision[12]. This enables agents to master different problem aspects sequentially rather than simultaneously.

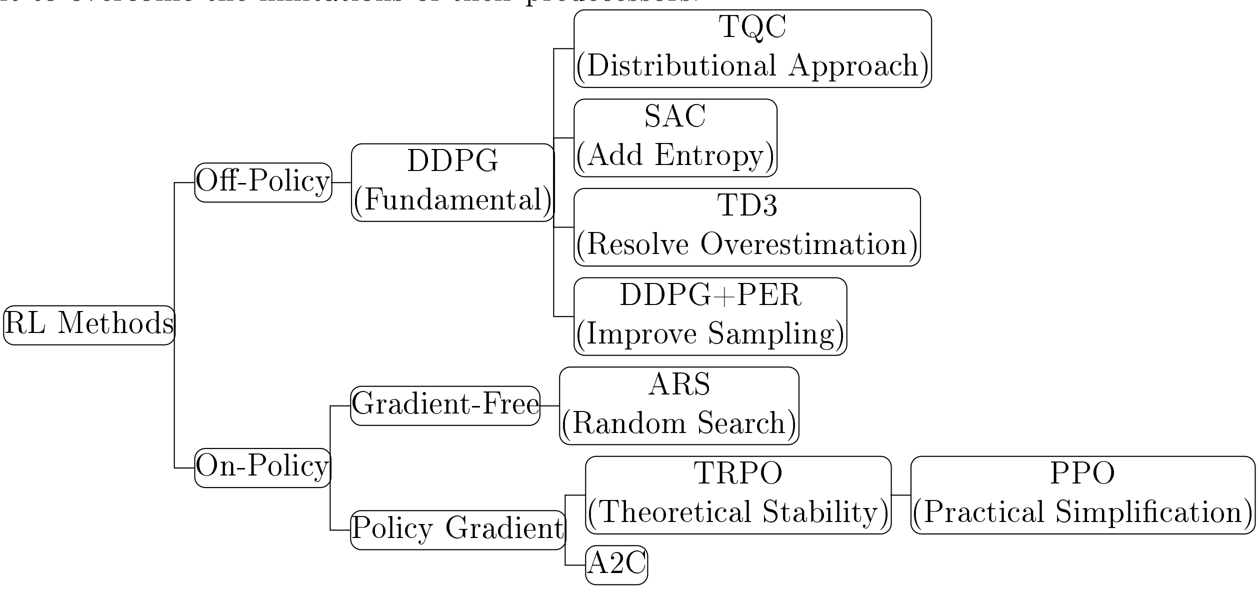
2.6.3 Curriculum Learning

While technically representing a training paradigm rather than a direct reward modification technique, Curriculum Learning (CL) maintains high relevance to reward engineering practices. CL involves training agents on task sequences that progressively increase in difficulty. Within the V2G context, agents might initially train in simplified scenarios featuring few EVs and stable pricing conditions. Upon mastering these fundamentals, they advance to more complex environments incorporating larger EV fleets, volatile pricing, and hard operational constraints. This structured learning approach prevents agents from becoming overwhelmed by full problem complexity from the outset, potentially leading to more robust and generalisable policies.

[11] 26.
[12] 41.

2.7 The Rise of Deep Reinforcement Learning for V2G Control

The convergence of RL with the substantial representational capabilities of deep neural networks has given rise to **Deep Reinforcement Learning (DRL)**, which currently represents the leading edge of V2G control research. The development of DRL algorithms has yielded a comprehensive toolkit, primarily divided into two main families: off-policy and on-policy methods, each exhibiting distinct operational characteristics. The following figure illustrates the evolutionary relationships between the algorithms that will be discussed, highlighting how newer ideas were built to overcome the limitations of their predecessors.



2.7.1 Off-Policy Methods: Data-Efficient Learning from Experience

Off-policy algorithms distinguish themselves through their capacity to learn optimal policies from data generated by different, often more exploratory, behavioral policies. This enables them to reuse historical experiences stored in large *replay buffers*, breaking temporal correlation between experiences and achieving high sample efficiency.

Deep Deterministic Policy Gradient (DDPG) A foundational algorithm that successfully extended Deep Q-Networks (DQN) to high-dimensional, continuous action spaces, DDPG represented a significant breakthrough for control problems including V2G[13]. It employs an actor-critic architecture where the actor deterministically maps states to actions. However, practical applications often encounter substantial training instability and systematic vulnerability to **overestimation bias**, where critic networks systematically overestimate Q-values, resulting in suboptimal

[13] 19.

policy learning[14].

Twin Delayed DDPG (TD3) Developed as a direct successor to address DDPG's instabilities, TD3 introduces three key innovations that have become standard in contemporary DRL[15]. It incorporates (1) clipped double Q-learning, utilizing paired critic networks and selecting minimum estimates to mitigate overestimation bias; (2) delayed policy updates, updating actors less frequently than critics for enhanced stability; and (3) target policy smoothing, adding noise to target actions for learning regularization. These enhancements establish TD3 as a more robust and reliable baseline for complex V2G applications[16].

Soft Actor-Critic (SAC) SAC represents a state-of-the-art off-policy algorithm for continuous control, recognized for superior sample efficiency and stability[17]. Its fundamental innovation involves the **maximum entropy framework**. The agent's objective is modified to maximize both expected reward and policy entropy. This entropy bonus encourages agents to act as randomly as possible while maintaining task success, promoting broader exploration, improved noise robustness, and reduced risk of poor local optima convergence[18].

Truncated Quantile Critics (TQC) TQC addresses overestimation bias through a distributional RL approach[19]. Rather than learning single expected returns (Q-values), its critic learns complete return probability distributions using quantile regression. By learning multiple return distribution quantiles and truncating the most optimistic quantile estimates before averaging, it provides a more principled and effective method for removing primary overestimation bias sources, frequently achieving superior performance.

Enhancement: Prioritized Experience Replay (PER) This represents not a standalone algorithm but a crucial orthogonal modification for off-policy methods. Standard replay buffers sample past transitions uniformly. PER samples transitions based on their "importance," typically proportional to TD error magnitude[20]. This focuses learning processes on surprising or informative experiences, significantly accelerating convergence and improving final performance.

2.7.2 On-Policy Methods: Stability through Cautious Updates

On-policy methods learn exclusively from data generated by current policies being optimized. Once data is used for updates, it is discarded. While this makes

[14] 27, 1.
[15] 11.
[16] 20, 42.
[17] 12.
[18] 21.
[19] 17.
[20] 34.

them inherently less sample-efficient than off-policy counterparts, their updates often demonstrate greater stability and reduced divergence risk.

Advantage Actor-Critic (A2C/A3C) A2C represents a foundational synchronous, on-policy Actor-Critic algorithm. Its powerful extension, **Asynchronous Advantage Actor-Critic (A3C)**, was a landmark contribution demonstrating parallelism benefits[21]. A3C utilizes multiple parallel workers, each with individual model and environment copies. These workers interact with their environments independently, and collected gradients update a global model asynchronously. This decorrelates data streams and provides powerful stabilizing effects on learning processes.

Trust Region Policy Optimization (TRPO) TRPO was the first algorithm to formalize policy update size constraints for guaranteed monotonic policy improvement[22]. It maximizes a "surrogate" objective function subject to policy change constraints, measured by Kullback-Leibler (KL) divergence. This creates a "trust region" within which new policies are guaranteed to improve upon previous ones, preventing catastrophic updates that can permanently derail learning. However, implementation complexity arises from second-order optimization requirements.

Proximal Policy Optimization (PPO) PPO achieves TRPO's stability benefits and reliable performance using only first-order optimization, making it far simpler to implement and more broadly applicable[23]. It employs a novel **clipping** mechanism in its objective function to discourage large policy updates that would move new policies too far from previous ones, effectively creating "soft" trust regions. Due to its excellent balance of performance, stability, and implementation simplicity, PPO has become a default choice for many on-policy applications.

2.7.3 Gradient-Free Methods: An Alternative Path

Augmented Random Search (ARS) As a counterpoint to dominant gradient-based methods, ARS represents a gradient-free approach that optimizes policies directly in parameter space[24]. It operates by exploring random policy parameter perturbations and updating central policy parameter vectors based on observed performance of these perturbations. While often less sample-efficient for complex, high-dimensional problems, its extreme simplicity, scalability, and robustness to noisy rewards can make it competitive in certain domains.

[21] 25.
[22] 36.
[23] 35.
[24] 22.

2.8 The Model-Based Benchmark: Model Predictive Control (MPC)

While DRL offers powerful model-free approaches, academic rigor requires benchmarking against its most robust model-based counterpart: **Model Predictive Control (MPC)**. MPC originated in the 1970s within the process control industry, with pioneering contributions by Richalet et al.[25] and Cutler and Ramaker[26]. The theoretical foundations for stability and optimality were subsequently established rigorously by researchers including Mayne and Rawlings[27]. At its core, MPC represents a proactive, forward-looking strategy that employs explicit mathematical system models to predict future evolution. At each control step, it solves finite-horizon optimal control problems to determine optimal control action sequences. Its primary strength, explaining widespread industrial adoption, lies in its inherent capability to proactively handle complex system dynamics and operational constraints[28].

2.8.1 Implicit MPC: Online Optimization

The most common formulation involves **Implicit MPC**, where constrained optimization problems are solved online at each control step. For linear systems with quadratic costs, this typically constitutes a Quadratic Programme (QP). The controller's objective involves finding future control input sequences $U = [u_{t|t}, ..., u_{t+N-1|t}]$ that minimize cost function J over prediction horizon N.

$$\min_U J(x_t, U) = \sum_{k=0}^{N-1} (x_{t+k|t}^T Q x_{t+k|t} + u_{t+k|t}^T R u_{t+k|t}) + x_{t+N|t}^T P x_{t+N|t} \tag{2.4}$$

where $x_{t+k|t}$ represents predicted state at future step k based on time t information, and Q, R, and P are weighting matrices defining trade-offs between state deviation and control effort. This optimization is subject to system dynamics models and operational constraints:

$$x_{t+k+1|t} = Ax_{t+k|t} + Bu_{t+k|t} \tag{2.5}$$

$$x_{min} \leq x_{t+k|t} \leq x_{max} \tag{2.6}$$

$$u_{min} \leq u_{t+k|t} \leq u_{max} \tag{2.7}$$

At each time step t, this complete problem is solved, but only the first action of the optimal sequence, $u^*_{t|t}$, is applied to the system. The process then repeats at the next time step, $t + 1$, using new system state measurements. This feedback mechanism, known as a *receding horizon* strategy, makes MPC robust to disturbances and model mismatch.

[7] [5]

[25] 31.

[26] 9.

[27] 23.

[28] 24.

2.8.2 Explicit MPC: Offline Pre-computation

For systems with fast dynamics or limited online computational capacity, **Explicit MPC** offers an alternative approach. Here, constrained optimization problems are solved offline for all possible initial states within given ranges, using multi-parametric programming. The result is not an online algorithm, but a pre-computed, explicit control law, $\pi(x_t)$, which is a **piecewise affine function** of state vector x_t.

$$u^*(x_t) = F_i x_t + g_i \quad \text{if } x_t \in \mathcal{X}_i \tag{2.8}$$

The state space is partitioned into convex polyhedral region sets $\mathcal{X}_i$, with unique affine control laws defined for each region. Online operation then reduces to fast lookup operations to identify which region current state x_t occupies and applying the corresponding simple control law. This trades very high offline computational burden and significant memory requirements for extremely fast online execution.

[3] [5]

2.9 A Comparative Perspective on Control Methodologies

While DRL represents the cutting edge, contextualizing it within the broader control strategy landscape remains crucial for academic rigor. The choice between learning-based, model-free approaches and optimization-based, model-based approaches represents fundamental philosophical and practical trade-offs.

Table 2.1: Comparative Analysis: DRL vs. Model Predictive Control (MPC) for V2G

Aspect	**Deep Reinforcement Learning (DRL)**	**Model Predictive Control (MPC)**
Paradigm	Model-Free, learning-based. Learns an optimal policy (a reactive function) via trial-and-error interaction with the environment.	Model-Based, optimisation-based. Solves a constrained optimisation problem at each time step based on a system model and forecasts.
Strengths	• Highly robust to uncertainty and unmodelled stochasticity. • Does not require an explicit, accurate system model. • Can learn complex, non-linear control policies. • Extremely fast inference time (a single forward pass) once trained.	• Explicitly and rigorously handles hard constraints, providing safety guarantees. • Proactive and anticipatory if forecasts are accurate. • Well-established, mature, and theoretically understood.
Weaknesses	• Can be highly sample-inefficient during the training phase. • Lacks hard safety guarantees (an active and important area of research). • The "black box" nature of neural network policies can make them difficult to interpret or verify.	• Performance is fundamentally shackled to the accuracy of the system model and external forecasts. • Computationally expensive at each time step, suffering from the "curse of dimensionality" with system size. • Can be brittle to unexpected forecast errors and unmodelled dynamics.
V2G Suitability	Excellent for dynamic, highly uncertain environments with complex, non-linear trade-offs and a large number of assets.	Good for problems with simple, well-defined dynamics and reliable forecasts, but struggles with the real-world stochasticity and scale of V2G.

Model Predictive Control (MPC) stands as the most powerful model-based alternative[29]. Its primary and most compelling strength lies in its native ability to handle hard constraints, which is critical for ensuring safe operation (such as never violating transformer limits or user energy requirements). However, its performance remains fundamentally constrained by the accuracy of its internal model

[29] 2.

and forecasts of future disturbances (prices, user behavior)[30]. In practice, creating accurate, tractable models for the entire V2G domain proves nearly impossible due to non-linear battery dynamics, extreme market volatility, and the profound unpredictability of human behavior. Furthermore, as EV fleet sizes grow, optimization problem dimensionality explodes, making online computation required at each time step intractable for large fleets[31].
Other methods, such as **meta-heuristic algorithms** (including genetic algorithms and particle swarm optimization), are sometimes proposed. However, these are typically employed for offline scheduling and lack the real-time, reactive responsiveness required for dynamic V2G control in rapidly changing environments[32].
Ultimately, DRL's singular advantage lies in its ability to learn and internalize the complex, non-linear trade-offs of multi-objective V2G problems directly from data, without requiring explicit models. This makes it uniquely suited to navigating the profound uncertainties of real-world operations. While other methods have their applications, DRL stands out as the most promising technology for deploying truly intelligent, autonomous, and scalable V2G management systems required to achieve the ambitious energy and climate objectives of the European Union.

[30] 10.
[31] 37.
[32] 16.

2.10 A Primer on Lithium-Ion Battery Chemistries and Degradation

The effectiveness, safety, and economic viability of any V2G strategy remain fundamentally constrained by the physical and chemical characteristics of vehicle batteries. Battery chemistry selection dictates EV operational envelopes, influencing energy density, power capabilities, lifespan, and safety profiles. Clear understanding of these trade-offs proves essential for developing robust, realistic, and responsible control algorithms.

2.10.1 Fundamental Concepts and Degradation Mechanisms

Battery degradation represents a complex, irreversible process that reduces battery capacity to store and deliver energy. It manifests primarily as capacity reduction (energy fade) and internal resistance increases (power fade). The mechanisms are broadly categorized into two types: calendar aging and cyclic aging[33].

- **Calendar Aging:** This refers to degradation occurring while batteries remain at rest, whether fully charged or empty. The primary, unavoidable mechanism involves slow, continuous growth of the **Solid Electrolyte Interphase (SEI)** layer on graphite anode surfaces. While thin, stable SEI layers are necessary for battery function, continued growth represents a parasitic reaction consuming active lithium ions and electrolyte, leading to irreversible capacity loss. SEI growth rates are strongly accelerated by two main factors:
  - **High Temperature:** Higher temperatures increase all chemical reaction rates, including parasitic ones. Storing batteries in hot environments proves highly detrimental.
  - **High State of Charge (SoC):** High SoC corresponds to low electrochemical potential at the anode, making graphite more reactive with electrolyte, thus promoting faster SEI growth[34]. This explains why leaving EVs stored at 100% SoC for extended periods is discouraged.
- **Cyclic Aging:** This refers to degradation occurring as a direct result of charging and discharging batteries. V2G operations directly contribute to this aging type. Key mechanisms include:
  - **Mechanical Stress:** During charging and discharging, lithium ions intercalate and de-intercalate from electrode materials, causing expansion and contraction. This repeated mechanical stress can cause micro-cracks in electrode particles, leading to electrical contact loss and active material reduction, thereby causing capacity fade. This effect becomes more pronounced with larger **Depths of Discharge (DoD)**.

[33] 4.
[34] 40.

- **SEI Layer Instability:** Volume changes from cycling can also crack protective SEI layers, exposing fresh anode material to electrolyte and triggering further, accelerated SEI growth.
- **Lithium Plating:** This represents a particularly damaging mechanism occurring under stressful conditions, such as very high charging rates (high C-rates) or low temperatures. Instead of smoothly intercalating into graphite anodes, lithium ions deposit on anode surfaces as metallic lithium. This causes rapid and severe capacity loss and can form sharp, needle-like structures called dendrites, which pose severe safety risks as they can grow through separators and cause internal short circuits[35].

For V2G applications, which inherently involve frequent and potentially deep cycling, understanding and mitigating cyclic aging remains paramount for ensuring long-term technology viability.

2.10.2 Key Automotive Chemistries

The EV market is dominated by several key lithium-ion battery families, primarily distinguished by cathode materials. Each chemistry presents different balances of performance characteristics.

- **Lithium Nickel Manganese Cobalt Oxide (NMC):** For years, this has been the most popular choice due to its excellent balance of energy density, power, and cycle life. The ratio of Nickel, Manganese, and Cobalt can be adjusted (such as NMC111, NMC532, NMC811) to prioritize either energy density (high Nickel) or safety and longevity (lower Nickel).

- **Lithium Nickel Cobalt Aluminum Oxide (NCA):** Offers very high specific energy (energy density by weight), enabling longer vehicle range. However, it typically comes at the cost of slightly lower cycle life and narrower safety margins compared to NMC, requiring more sophisticated thermal management.

- **Lithium Iron Phosphate (LFP):** Is rapidly gaining market share, particularly in standard-range vehicles, due to lower cost (no cobalt, an expensive and ethically contentious material) and superior safety. LFP batteries offer exceptional cycle life (often 3-4 times that of NMC/NCA) and are considered the safest common Li-ion chemistry due to high thermal stability. Their main drawback involves lower energy density.

- **Lithium Titanate Oxide (LTO):** This represents a more niche chemistry using titanate-based anodes instead of graphite. This provides exceptional safety (virtually no thermal runaway risk), extremely long cycle life ($>$10,000 cycles), and excellent low-temperature performance. However, very low energy density and high cost currently limit its use to specialized applications.

[35] 4.

2.10.3 Voltage Profiles and the Challenge of SoC Estimation

The relationship between battery open-circuit voltage and SoC represents a critical, non-linear function unique to each chemistry. The derivative of cell voltage with respect to DoD, $\frac{dV_{cell}}{d(DoD)}$, constitutes a key parameter for Battery Management Systems (BMS). Steep, monotonic slopes allow BMS to accurately infer SoC from simple voltage measurements. Conversely, flat slopes make this estimation extremely difficult.

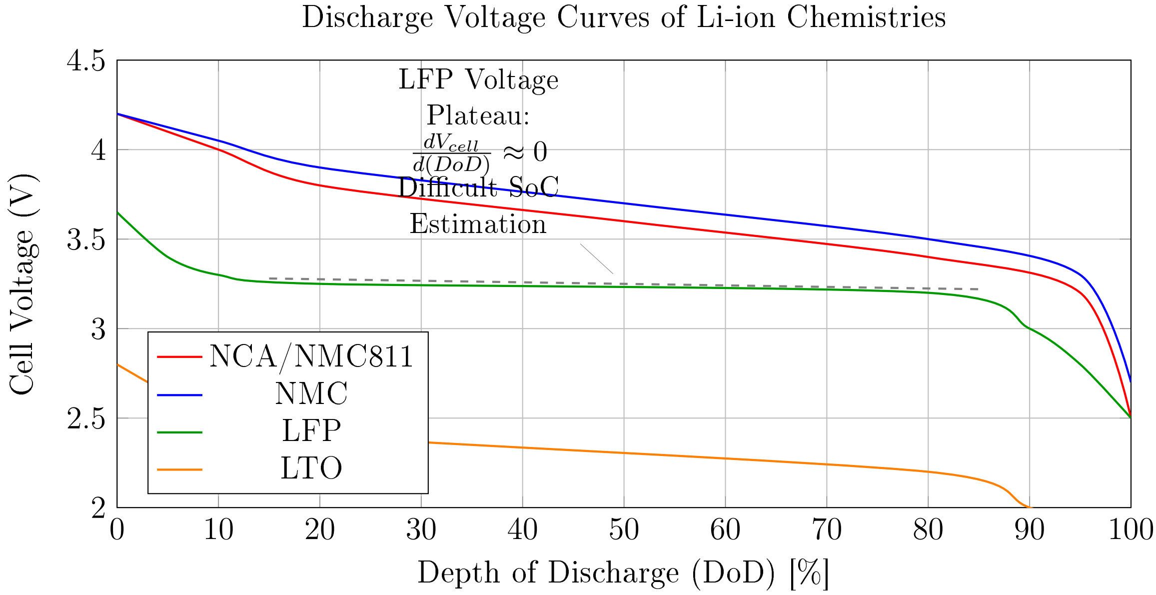


Figure 2.4: Typical discharge voltage curves for various lithium-ion chemistries. The extremely flat profile of LFP makes accurate SoC estimation challenging based on voltage alone, necessitating more complex estimation techniques like Coulomb counting and periodic recalibration[37].

As shown in Figure 2.4, LFP's remarkably flat voltage plateau makes it nearly impossible for BMS to determine precise SoC in the central operating range (from approximately 20% to 80%) using voltage alone. This necessitates more complex estimation techniques, such as Coulomb counting (integrating current over time), which can suffer from drift. To correct this drift, LFP-equipped vehicles require periodic full charges to 100% for BMS recalibration. This represents an important operational constraint that V2G control strategies must consider.

2.10.4 Comparative Analysis and Safety Considerations

The trade-offs between chemistries are summarized in Table 2.2. Safety remains paramount, with the primary risk being thermal runaway—a dangerous, self-sustaining exothermic reaction. This risk relates directly to cathode material chemical and thermal stability. Higher energy density generally means more energy packed into smaller mass, which can be released violently if cells are compromised. Consequently, critical temperatures for initiating thermal runaway are generally lower for higher energy density chemistries. LFP's stable phosphate-based structure makes it far more resistant to thermal runaway than nickel-based counterparts, a key reason for its growing popularity.

Table 2.2: Comparative analysis of key automotive battery chemistries, highlighting the trade-offs between performance and safety.

Metric	NCA	NMC	LFP	LTO	LCO
Energy Density (Wh/kg)	200 - 260 (Highest)	150 - 220 (High)	90 - 160 (Moderate)	60 - 110 (Low)	150-200 (High)
Cycle Life	1000 - 2000	1000 - 2500	2000 - 5000+	>10,000	500 - 1000
Safety	Good	Very Good	Excellent	Excellent	Poor
Thermal Runaway Temp (°C)	~150 - 180	~180 - 210	~220 - 270	> 250	~150

Chapter 3

An Enhanced V2G Simulation Framework for Robust Control

Developing, validating, and benchmarking advanced control algorithms for Vehicle-to-Grid (V2G) systems is a complex endeavor. Real-world experimentation is often impractical due to prohibitive costs, logistical challenges, and risks to grid stability and vehicle hardware. To bridge the gap between theory and practice, a realistic, flexible, and standardized simulation environment is a scientific necessity. This thesis builds upon the foundation of **EV2Gym**, a state-of-the-art, open-source simulator designed for V2G smart charging research[1]. This work, however, extends the original framework significantly, transforming it into a high-fidelity **digital twin** engineered not just for single-scenario optimization, but for the development and rigorous evaluation of **robust, generalist control agents**.

This enhanced framework offers a two-pronged approach to experimentation: it allows for deep-dive analysis of agents specialized for a single environment, while also introducing a novel methodology for training and testing agents designed to generalize across a multitude of diverse, unpredictable scenarios. This chapter provides an in-depth tour of this extended architecture, its data-driven models, and its unique evaluation capabilities, establishing the methodological bedrock for the rest of this work.

3.1 Core Simulator Architecture

The framework is built on the modular architecture of EV2Gym, which mirrors the key entities of a real-world V2G system. Its foundation on the OpenAI Gym (now Gymnasium) API is a cornerstone, providing a standardized agent-environment interface defined by the familiar language of states, actions, and rewards[2].

The architecture consists of several interacting components:

- **Charge Point Operator (CPO):** The central intelligence of the simulation, managing the charging infrastructure and serving as the primary interface for

[1] 28.
[2] 6.

the control algorithm (the DRL agent). The CPO aggregates system state information and dispatches control actions to individual chargers.

- **Chargers:** Digital representations of physical charging stations, configurable by type (AC/DC), maximum power, and efficiency. This allows for the simulation of heterogeneous charging infrastructures.
- **Power Transformers:** These components model the physical connection points to the grid, aggregating the electrical load from multiple chargers. Crucially, they enforce the physical power limits of the local distribution network and can model inflexible base loads (e.g., buildings) and local renewable generation (e.g., solar panels).
- **Electric Vehicles (EVs):** Dynamic and autonomous agents, each defined by its unique battery capacity, power limits, current and desired energy levels, and specific arrival and departure times.

The simulation process follows a reproducible three-phase structure: (1) **Initialization** from a comprehensive YAML configuration file, (2) a discrete-time **Simulation Loop** where the agent interacts with the environment, and (3) a final **Evaluation and Visualization** phase that generates standardized performance metrics.

3.2 Core Physical Models

The simulation's fidelity is anchored in its detailed, empirically validated models, which are essential for developing control strategies robust enough for real-world application.

3.2.1 EV Model and Charging/Discharging Dynamics

The framework implements a realistic two-stage charging/discharging model that captures the non-linear behavior of lithium-ion batteries, simulating both the **constant current (CC)** and **constant voltage (CV)** phases. Each EV is defined by a rich parameter set: maximum capacity (E_{max}), a minimum safety capacity (E_{min}), separate power limits for charging and discharging ($P_{ch}^{max}, P_{dis}^{max}$), and distinct efficiencies for each process (η_{ch}, η_{dis}).

3.2.2 Battery Degradation Model

To address the critical issue of battery health in V2G operations, the simulator incorporates a semi-empirical battery degradation model. It quantifies capacity loss (Q_{lost}) as the sum of two primary aging mechanisms[3]:

- **Calendar Aging (d_{cal}):** Time-dependent capacity loss, influenced by the battery's average State of Charge (SoC) and temperature.

[3] 28.

- **Cyclic Aging (d_{cyc}):** Wear resulting from charge/discharge cycles, dependent on energy throughput, depth-of-cycle, and C-rate.

This integrated model allows for the direct quantification of how different control strategies impact the battery's long-term State of Health (SoH), enabling the training of agents that balance profitability with battery preservation.

3.2.3 EV Behavior and Grid Models

To ensure realism, the simulation is driven by authentic, open-source datasets. EV arrival/departure patterns and energy requirements are modeled using probability distributions derived from a large real-world dataset from **ElaadNL**. Grid conditions are similarly grounded in reality, using inflexible load data from the **Pecan Street** project and solar generation profiles from the **Renewables.ninja** platform[4].

3.3 A Unified Experimentation and Evaluation Workflow

A key contribution of this thesis is the development of a unified and powerful experimentation workflow, orchestrated by the main script `run_experiments.py`. This script replaces the previous fragmented approach, providing a single, interactive interface to manage the entire lifecycle of training, benchmarking, and evaluation for V2G control agents. This workflow is designed to be both flexible for research and rigorous for evaluation, supporting the dual goals of developing specialized and generalized agents.

3.3.1 Orchestration via `run_experiments.py`

The `run_experiments.py` script acts as the central hub for all experimentation. It guides the user through an interactive command-line process, ensuring consistency and reproducibility. The key steps are:

1. **Algorithm Selection:** The user can select from a predefined list of algorithms to benchmark. This includes Deep Reinforcement Learning agents (e.g., SAC, DDPG+PER, TQC), classical optimization methods (Model Predictive Control), and rule-based heuristics (e.g., Charge As Fast As Possible).

2. **Scenario Selection:** The script automatically detects all available `.yaml` configuration files, allowing the user to choose one or more scenarios for the experiment. This choice determines the mode of operation (single-domain vs. multi-scenario).

3. **Reward Function Selection:** The framework's flexibility is enhanced by allowing the user to dynamically select the reward function for the RL agents from the `reward.py` module.

[4] 28.

4. **Training and Benchmarking:** Based on the user's selections, the script proceeds to the optional training phase and then to a comprehensive benchmark, saving all results in a timestamped directory.

3.3.2 Dual-Mode Training: Specialists and Generalists

The new workflow elegantly unifies the training of both "specialist" and "generalist" agents, a concept previously handled by separate scripts. The behavior is determined implicitly by the number of selected scenarios:

- **Single-Domain Specialization:** If the user selects a single scenario, the script trains an RL agent exclusively on that environment. This produces a specialist agent, optimized to extract maximum performance from a specific, known set of conditions (e.g., a particular charging station topology and price profile).

- **Multi-Scenario Generalization:** If multiple scenarios are selected, the script automatically utilizes the `MultiScenarioEnv` wrapper. This custom Gymnasium environment dynamically switches between the different selected configurations at the start of each training episode. This process forces the agent to learn a robust and generalizable policy that performs well across a wide range of conditions, preventing overfitting to any single scenario. To handle the technical challenge of varying observation and action space sizes across scenarios, a `CompatibilityWrapper` is used to pad and slice the state-action vectors, enabling a single neural network policy to control heterogeneous environments.

3.3.3 Reproducible Benchmarking and Evaluation

To ensure a fair and scientifically valid comparison, the `run_benchmark` function implements a rigorous evaluation protocol. For each scenario, it first generates a "replay" file containing the exact sequence of stochastic events (e.g., EV arrivals, energy demands). This exact same sequence is then used to evaluate every algorithm, eliminating randomness as a factor in performance differences. The script runs multiple simulations for statistical robustness, aggregates the mean results, and automatically generates a suite of comparative plots, including overall performance metrics and detailed battery degradation analyses.

3.4 Evaluation Metrics

To ensure a fair and comprehensive comparison, all algorithms are evaluated against an identical set of pre-generated scenarios through a "replay" mechanism. The **mean** and **standard deviation** of performance are calculated across multiple simulation runs. The key metrics include:

- **Total Profit (\$):** The net economic outcome, calculated as revenue from energy sales minus the cost of energy purchases.

$$\Pi_{\text{total}} = \sum_{t=0}^{T_{\text{sim}}} \sum_{i=1}^{N} \left(C_{\text{sell}}(t) P_{\text{dis},i}(t) - C_{\text{buy}}(t) P_{\text{ch},i}(t) \right) \Delta t$$

- **Tracking Error (RMSE, kW):** For grid-balancing scenarios, this measures the root-mean-square error between the fleet's aggregated power and a target setpoint.

$$E_{\text{track}} = \sqrt{\frac{1}{T_{\text{sim}}} \sum_{t=0}^{T_{\text{sim}}-1} \left(P_{\text{setpoint}}(t) - P_{\text{total}}(t) \right)^2}$$

- **User Satisfaction (Average):** The fraction of energy delivered compared to what was requested by the user, averaged across all EV sessions. A score of 1 indicates perfect service.

$$US_{\text{avg}} = \frac{1}{N_{\text{EVs}}} \sum_{k=1}^{N_{\text{EVs}}} \min \left(1, \frac{E_k(t_k^{\text{dep}})}{E_k^{\text{des}}} \right)$$

- **Transformer Overload (kWh):** The total energy that exceeded the transformer's rated power limit. An ideal controller should achieve a value of 0.

$$O_{\text{tr}} = \sum_{t=0}^{T_{\text{sim}}} \sum_{j=1}^{N_T} \max(0, P_j^{\text{tr}}(t) - P_j^{\text{tr,max}}) \cdot \Delta t$$

- **Battery Degradation (\$):** The estimated monetary cost of battery aging due to both cyclic and calendar effects.

$$D_{\text{batt}} = \sum_{k=1}^{N_{\text{EVs}}} (\text{CyclicCost}_k + \text{CalendarCost}_k)$$

3.5 Simulator Implementation Details

During the analysis and implementation of new metrics, fundamental details about the `EV2Gym` simulator's architecture emerged, which warrant documentation. The configuration of Electric Vehicles (EVs) and the calculation of their degradation follow a specific logic dependent on a key parameter in the `.yaml` configuration files.

Vehicle Definition Modes

The simulator operates in two distinct modes, controlled by the boolean flag `heterogeneous_ev_spec`

- **Heterogeneous Mode (`True`):** In this mode, the simulator ignores the default vehicle specifications in the `.yaml` file. Instead, it loads a list of vehicle profiles from an external JSON file, specified by the `ev_specs_file` parameter (e.g., `ev_specs_v2g_enabled2024.json`). This allows for the creation of a realistic fleet with diverse battery capacities, charging powers, and efficiencies. For instance, the fleet may include a **Peugeot 208** with a 46.3 kWh battery and a 7.4 kW charge rate, alongside a **Volkswagen ID.4** with a 77 kWh battery and an 11 kW charge rate. A vehicle is randomly selected from this list for each new arrival event.

- **Homogeneous Mode (`False`):** In this mode, the external JSON file is ignored. All vehicles created in the simulation are identical, and their characteristics are defined exclusively by the `ev:` block within the `.yaml` configuration file. The `battery_capacity` parameter in this block becomes the single source of truth for the entire fleet.

Empirical Calibration of the Degradation Model

A significant enhancement in this work is the move towards a more physically representative and flexible battery degradation model. While the underlying semi-empirical model for calendar and cyclic aging remains, the methodology for parameterizing it has been fundamentally improved, addressing previous inconsistencies.

This is achieved through the `Fit_battery.py` script, a new utility for empirical model calibration. The script implements the following workflow:

1. **Data Loading:** It loads time-series data from real-world battery aging experiments. The expected data includes measurements of capacity loss over time, along with contextual variables like state of charge (SoC), temperature, and energy throughput.

2. **Model Fitting:** Using the `curve_fit` function from the SciPy library, the script fits the parameters of the `Qlost_model` (which combines calendar and cyclic aging) to the empirical data. This optimization process finds the physical constants (e.g., ϵ_0, ζ_0) that best explain the observed degradation.

3. **Parameter Export:** The script outputs the calibrated parameters. These values can then be used directly in the simulator's configuration, ensuring that the degradation model for a specific EV fleet is grounded in experimental evidence for that battery type.

This calibration workflow, integrated optionally into the main `run_experiments.py` script, elevates the simulation's fidelity. It allows the framework to move beyond a single, fixed degradation model (previously calibrated for a 78 kWh battery) and enables the creation of high-fidelity digital twins for a wide variety of EV batteries, provided that the necessary experimental data is available.

3.6 Reinforcement Learning Formulation

The control problem is formalized as a Markov Decision Process (MDP), defined by the tuple (S, A, P, R, γ).

3.6.1 State Space (S)

The state $s_t \in S$ is a feature vector providing a snapshot of the environment at time t. A representative state, as defined in modules like `V2G_profit_max_loads.py`, includes:

$$s_t = [t, P_{\text{total}}(t-1), \mathbf{c}(t,H), \mathbf{L}_1(t,H), \mathbf{PV}_1(t,H), \ldots, \mathbf{s}_1^{\text{EV}}(t), \ldots, \mathbf{s}_N^{\text{EV}}(t)]^T$$

where the components are:

- t: The current time step.
- $P_{\text{total}}(t-1)$: The aggregated power from the previous time step.
- $\mathbf{c}(t,H)$: A vector of **predicted future** electricity prices over a horizon H.
- $\mathbf{L}_j(t,H), \mathbf{PV}_j(t,H)$: Forecasts for inflexible loads and solar generation.
- $\mathbf{s}_i^{\text{EV}}(t) = [\text{SoC}_i(t), t_i^{\text{dep}} - t]$: Key information for each EV i, including its State of Charge and remaining time until departure.

3.6.2 Action Space (A)

The action $a_t \in A$ is a continuous vector in $\mathbb{R}^N$, where N is the number of chargers. For each charger i, the command $a_i(t) \in [-1, 1]$ is a normalized value that is translated into a power command:

- If $a_i(t) > 0$, the EV is charging: $P_i(t) = a_i(t) \cdot P_{\text{charge},i}^{\text{max}}$.
- If $a_i(t) < 0$, the EV is discharging (V2G): $P_i(t) = a_i(t) \cdot P_{\text{discharge},i}^{\text{max}}$.

3.6.3 Reward Function

The reward function $R(t)$ encodes the objectives of the control agent. The framework allows for the selection of different reward functions from the `reward.py` module to suit various goals. Key examples include:

- **Profit Maximization with Penalties** (`ProfitMax_TrPenalty_UserIncentives`): This function creates a balance between economic gain and physical constraints.

$$R(t) = \underbrace{\text{Profit}(t)}_{\text{Economic Gain}} - \underbrace{\lambda_1 \cdot \text{Overload}(t)}_{\text{Grid Penalty}} - \underbrace{\lambda_2 \cdot \text{Unsatisfaction}(t)}_{\text{User Penalty}}$$

  The agent is rewarded for profit but penalized for overloading transformers and for failing to meet the charging needs of departing drivers.

- **Squared Tracking Error** (`SquaredTrackingErrorReward`): Used for grid service applications where precision is paramount.

$$R(t) = -\left(P_{\text{setpoint}}(t) - \sum_{i=1}^{N} P_i(t)\right)^2$$

The reward is the negative squared error from the power setpoint, incentivizing the agent to minimize this error at all times.

By using this enhanced framework, this thesis moves beyond single-scenario optimization to develop and validate an intelligent V2G control agent that is not only high-performing but also robust, adaptable, and ready for the complexities of real-world deployment.

3.7 Reinforcement Learning Algorithms

This work benchmarks several state-of-the-art Deep Reinforcement Learning algorithms. The following sections provide a detailed mathematical description of the selected off-policy, actor-critic algorithms.

Soft Actor-Critic (SAC)

SAC is an off-policy actor-critic algorithm designed for continuous action spaces that optimizes a stochastic policy. Its core feature is entropy maximization, which encourages exploration and improves robustness. The agent aims to maximize not only the expected sum of rewards but also the entropy of its policy.

The objective function is:

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim\rho_\pi} \left[r(s_t, a_t) + \alpha\mathcal{H}(\pi(\cdot|s_t))\right]$$

where $\mathcal{H}$ is the entropy of the policy π and α is the temperature parameter, which controls the trade-off between reward and entropy.

SAC uses a soft Q-function, trained to minimize the soft Bellman residual:

$$L(\theta_Q) = \mathbb{E}_{(s_t,a_t,r_t,s_{t+1})\sim D} \left[\left(Q(s_t, a_t) - \left(r_t + \gamma V_{\bar{\psi}}(s_{t+1})\right)\right)^2\right]$$

where D is the replay buffer and the soft state value function V is defined as:

$$V_{\text{soft}}(s_t) = \mathbb{E}_{a_t\sim\pi}[Q_{\text{soft}}(s_t, a_t) - \alpha \log \pi(a_t|s_t)]$$

To mitigate positive bias, SAC employs two Q-networks (Clipped Double-Q) and takes the minimum of the two target Q-values during the Bellman update.

Deep Deterministic Policy Gradient + PER (DDPG+PER)

DDPG is an off-policy algorithm that concurrently learns a deterministic policy $\mu(s|\theta^\mu)$ and a Q-function $Q(s, a|\theta^Q)$. It is the deep-learning extension of the DPG algorithm for continuous action spaces.

- **Critic Update:** The critic is updated by minimizing the mean-squared Bellman error, similar to Q-learning. Target networks (Q' and μ') are used to stabilize training.

$$L(\theta^Q) = \mathbb{E}_{(s_t,a_t,r_t,s_{t+1})\sim D} \left[(y_t - Q(s_t, a_t|\theta^Q))^2\right]$$

where the target y_t is given by:

$$y_t = r_t + \gamma Q'(s_{t+1}, \mu'(s_{t+1}|\theta^{\mu'})|\theta^{Q'})$$

- **Actor Update:** The actor is updated using the deterministic policy gradient theorem:
$$\nabla_{\theta^\mu} J \approx \mathbb{E}_{s_t \sim D}[\nabla_a Q(s, a|\theta^Q)|_{s=s_t, a=\mu(s_t)} \nabla_{\theta^\mu} \mu(s_t|\theta^\mu)]$$

- **Prioritized Experience Replay (PER):** This work enhances DDPG with PER. Instead of uniform sampling from the replay buffer D, PER samples transitions based on their TD-error, prioritizing those where the model has the most to learn. The probability of sampling transition i is:

$$P(i) = \frac{p_i^\beta}{\sum_k p_k^\beta}$$

where $p_i = |\delta_i| + \epsilon$ is the priority based on the TD-error δ_i, and β controls the degree of prioritization. To correct for the bias introduced by non-uniform sampling, PER uses importance-sampling (IS) weights.

Truncated Quantile Critics (TQC)

TQC enhances the stability of SAC by modeling the entire distribution of returns instead of just its mean. This is achieved through quantile regression and a novel truncation mechanism to combat Q-value overestimation.

- **Distributional Learning:** TQC employs a set of N critic networks, $\{Q_{\phi_i}(s, a)\}_{i=1}^N$, each trained to estimate a specific quantile τ_i of the return distribution. The target quantiles are implicitly defined as $\tau_i = \frac{i-0.5}{N}$. The critics are trained by minimizing the quantile Huber loss, L_{QH}.

- **Distributional Target Calculation:** A distributional target is constructed for the Bellman update. First, an action is sampled from the target policy for the next state: $\tilde{a}_{t+1} \sim \pi_{\theta'}(\cdot|s_{t+1})$. Then, a set of N Q-value estimates for the next state is obtained from the N target critic networks: $\{Q_{\phi'_j}(s_{t+1}, \tilde{a}_{t+1})\}_{j=1}^N$.

- **Truncation:** This is the key idea of TQC. To combat overestimation, the algorithm discards the k largest Q-value estimates from the set of N target values. This truncation removes the most optimistic estimates, which are a primary source of bias, leading to more conservative and stable updates.

- **Critic Update:** The target value for updating the i-th critic is formed using the Bellman equation with the truncated set of next-state Q-values. The overall critic loss is the sum of the quantile losses across all critics:

$$L(\phi) = \sum_{i=1}^{N} \mathbb{E}_{(s,a,r,s')\sim D} \left[L_{QH} \left(r + \gamma Q_{\text{trunc}}(s', \tilde{a}') - Q_{\phi_i}(s, a) \right) \right]$$

where Q_{trunc} represents the value derived from the truncated set of target quantiles.

3.7.1 A History-Based Adaptive Reward for Profit Maximization

To effectively steer the learning agent towards a policy that is both highly profitable and reliable, we have designed and implemented a novel, history-based adaptive reward function, named `FastProfitAdaptiveReward`. This function departs from traditional static-weight penalties and instead introduces a dynamic feedback mechanism where the severity of penalties is directly influenced by the agent's recent performance. The core philosophy is to aggressively prioritize economic profit while using adaptive penalties as guardrails that become stricter only when the agent begins to consistently violate operational constraints.

The total reward at each timestep t, R_t, is calculated as the net economic profit minus any active penalties for user dissatisfaction or transformer overload.

$$R_t = \Pi_t - P_t^{\text{sat}} - P_t^{\text{tr}} \tag{3.1}$$

Economic Profit

The foundation of the reward signal is the direct, instantaneous economic profit, Π_t. This component provides a clear and strong incentive for the agent to learn market dynamics, encouraging it to charge during low-price periods and discharge (V2G) during high-price periods.

$$\Pi_t = \sum_{i=1}^{N} \left(C_t^{\text{sell}} \cdot P_{i,t}^{\text{dis}} - C_t^{\text{buy}} \cdot P_{i,t}^{\text{ch}} \right) \Delta t \tag{3.2}$$

where N is the number of connected EVs, C_t^{sell} and C_t^{buy} are the electricity prices, and $P_{i,t}^{\text{dis}}$ and $P_{i,t}^{\text{ch}}$ are the discharging and charging powers for EV i.

Adaptive User Satisfaction Penalty

The penalty for failing to meet user charging demands, P_t^{sat}, is not a fixed value. Instead, it adapts based on the system's recent history of performance. The environment maintains a short-term memory of the average user satisfaction over the last 100 timesteps. From this history, we calculate an average satisfaction score, $\bar{S}_{hist}$.

A *satisfaction severity multiplier*, λ_t^{sat}, is then calculated. This multiplier grows quadratically as the historical average satisfaction drops, meaning that if the system has been performing poorly, the consequences for a new failure become much more severe.

$$\lambda_t^{\text{sat}} = \lambda_{\text{base}}^{\text{sat}} \cdot (1 - \bar{S}_{hist})^2 \tag{3.3}$$

where $\lambda_{\text{base}}^{\text{sat}}$ is a base scaling factor (e.g., 20.0). A penalty is only applied if any departing EV's satisfaction, S_k, is below a critical threshold (e.g., 95%). The magnitude of the penalty is the product of the adaptive multiplier and the current satisfaction deficit.

$$P_t^{\text{sat}} = \lambda_t^{\text{sat}} \cdot (1 - \min(S_k)) \quad \forall k \in \text{EVs departing at } t \tag{3.4}$$

This creates a powerful feedback loop: a single, isolated failure in an otherwise well-performing system results in a mild penalty. However, persistent failures lead to a rapidly escalating penalty, forcing the agent to correct its behavior.

Adaptive Transformer Overload Penalty

Similarly, the transformer overload penalty, P_t^{tr}, adapts based on the recent frequency of overloads. The environment tracks how often an overload has occurred in the last 100 timesteps, yielding an overload frequency, F_{hist}^{tr}.

This frequency is used to compute a linear *overload severity multiplier*, λ_t^{tr}. The more frequently overloads have happened, the higher the penalty for a new one.

$$\lambda_t^{\text{tr}} = \lambda_{\text{base}}^{\text{tr}} \cdot F_{hist}^{\text{tr}} \tag{3.5}$$

where $\lambda_{\text{base}}^{\text{tr}}$ is a base scaler (e.g., 50.0). If the total power drawn, $P_j^{\text{total}}(t)$, exceeds the transformer's limit, P_j^{max}, a penalty is applied. This penalty consists of a small, fixed base amount plus the adaptive component, which scales with the magnitude of the current overload.

$$P_t^{\text{tr}} = P_{\text{base}} + \lambda_t^{\text{tr}} \cdot \sum_{j=1}^{N_T} \max(0, P_j^{\text{total}}(t) - P_j^{\text{max}}) \tag{3.6}$$

This mechanism teaches the agent that while a rare, minor overload might be acceptable in pursuit of high profit, habitual overloading is an unsustainable and heavily penalized strategy.

Rationale and Significance

This history-based adaptive reward function represents a significant advancement over static or purely state-based approaches. By making the penalty weights a function of the system's recent performance history, we provide a more nuanced and stable learning signal. The agent is not punished excessively for isolated, exploratory actions that might lead to a minor constraint violation. Instead, it is strongly discouraged from developing policies that lead to chronic system failures.

The intuition is to mimic a more realistic management objective: maintain high performance on average, and react strongly only when performance trends begin to degrade. This method is also computationally efficient, avoiding complex state-dependent calculations in favor of simple updates to historical data queues. Ultimately, this reward structure guides the agent to discover policies that are not only profitable but also robust and reliable over time, striking a more intelligent balance between economic ambition and operational safety.

3.8 Model Predictive Control (MPC)

The MPC, implemented in `mpc.py` and `eMPC.py`, solves an optimization problem at every time step over a prediction horizon H.

3.8.1 System Model

The system is modeled in linear state-space form. The state $\mathbf{x}_k \in \mathbb{R}^N$ is the vector of SoCs of all EVs at time k. The input $\mathbf{u}_k \in \mathbb{R}^{2N}$ is the vector of charging and discharging powers.

$$\mathbf{x}_{k+1} = A_k \mathbf{x}_k + B_k \mathbf{u}_k$$

The matrices A_k (`Amono`) and B_k (`Bmono`) are time-varying because they depend on which EVs are connected. A_k is typically a diagonal identity-like matrix modeling the persistence of EVs. B_k maps power to SoC change, including efficiencies and Δt.

3.8.2 Optimization Problem

At time t, the MPC solves:

$$\min_{\{\mathbf{u}_k\}_{k=t}^{t+H-1}} \sum_{k=t}^{t+H-1} \mathbf{f}_k^T \mathbf{u}_k$$

subject to:

$$\begin{aligned}
&\mathbf{x}_{k+1} = A_k \mathbf{x}_k + B_k \mathbf{u}_k, \quad \forall k \in [t, t+H-1] && \text{(Dynamics)} \\
&\mathbf{x}_k^{\min} \leq \mathbf{x}_k \leq \mathbf{x}_k^{\max} && \text{(SoC limits)} \\
&\mathbf{0} \leq \mathbf{u}_k^{\text{ch}} \leq \mathbf{u}_k^{\text{ch,max}} \cdot \mathbf{z}_k && \text{(Charge limits)} \\
&\mathbf{0} \leq \mathbf{u}_k^{\text{dis}} \leq \mathbf{u}_k^{\text{dis,max}} \cdot (1 - \mathbf{z}_k) && \text{(Discharge limits)} \\
&\sum_{i \in \text{CS}_j} (u_i^{\text{ch}} - u_i^{\text{dis}}) + L_j(k) - PV_j(k) \leq P_j^{\text{tr,max}}(k) && \text{(Transformer limits)}
\end{aligned}$$

where $\mathbf{z}_k$ is a vector of binary variables to prevent simultaneous charge and discharge. The cost vector $\mathbf{f}_k$ contains the energy prices. The code formulates this problem compactly as $\mathbf{AU} \leq \mathbf{bU}$, where $\mathbf{U}$ is the vector of all actions over the horizon.

3.9 Offline Optimization with Gurobi

Gurobi is used to find the optimal offline (a posteriori) solution, providing a performance benchmark. The files `profit_max.py` and `tracking_error.py` define the optimization problem over the entire simulation horizon T_{sim}.

3.9.1 Decision Variables

- $E_{p,i,t}$: Energy in the EV at port p of station i at time t.
- $I_{p,i,t}^{\text{ch}}, I_{p,i,t}^{\text{dis}}$: Charging/discharging currents.

- $\omega^{\text{ch}}_{p,i,t}, \omega^{\text{dis}}_{p,i,t}$: Binary variables for operating modes.

3.9.2 Objective Function (Example: Profit Maximization)

$$\max \sum_{t=0}^{T_{\text{sim}}} \sum_{i=1}^{N_{CS}} \sum_{p=1}^{N_p} \left(C_{\text{sell}}(t) P^{\text{dis}}_{p,i,t} - C_{\text{buy}}(t) P^{\text{ch}}_{p,i,t} \right) \Delta t - \lambda \sum_{k \in \text{EVs departed}} (E^{\text{des}}_k - E_k(t^{\text{dep}}_k))^2$$

where $P = V \cdot I \cdot \eta$.

3.9.3 Main Constraints

- **Energy Balance:**

$$E_{p,i,t} = E_{p,i,t-1} + (\eta_{\text{ch}} V_i I^{\text{ch}}_{p,i,t} - \frac{1}{\eta_{\text{dis}}} V_i I^{\text{dis}}_{p,i,t}) \Delta t$$

- **Activation of Current:**

$$I^{\text{ch}}_{p,i,t} \leq M \cdot \omega^{\text{ch}}_{p,i,t} \quad , \quad I^{\text{dis}}_{p,i,t} \leq M \cdot \omega^{\text{dis}}_{p,i,t}$$

- **Mutual Exclusion:**

$$\omega^{\text{ch}}_{p,i,t} + \omega^{\text{dis}}_{p,i,t} \leq 1$$

- **Current and SoC Limits:**

$$I^{\min} \leq I_{p,i,t} \leq I^{\max} \quad , \quad E^{\min} \leq E_{p,i,t} \leq E^{\max}$$

- **SoC at Departure:**

$$E_{p,i}(t^{\text{dep}}) \geq E^{\text{des}}_{p,i}$$

3.10 Online MPC Formulation (PuLP Implementation)

The Model Predictive Control (MPC) implemented with PuLP solves a profit maximization problem at each time step t over a finite prediction horizon H. This formulation is designed for online, real-time control, where decisions are made based on the current system state and future predictions.

3.10.1 Mathematical Formulation

At each time step t, the MPC controller solves the following optimization problem.

Objective Function: Net Operational Profit

The objective is to maximize the total net operational profit over the control horizon H. This provides a comprehensive economic model that goes beyond simple energy arbitrage.

$$\max_{P^{\mathrm{ch}},P^{\mathrm{dis}},z} \sum_{k=t}^{t+H-1} \sum_{i\in\mathrm{CS}} (\mathrm{Revenues}_{i,k} - \mathrm{Costs}_{i,k}) \tag{3.7}$$

The revenue and cost components are defined for each station i at time step k as:

- **Revenues** consist of:
  - Grid Sales Revenue (V2G): $c_k^{\mathrm{sell}} \cdot P_{i,k}^{\mathrm{dis}} \cdot \Delta t$
  - User Charging Revenue: $c^{\mathrm{user}} \cdot P_{i,k}^{\mathrm{ch}} \cdot \Delta t$
- **Costs** consist of:
  - Grid Purchase Cost: $c_k^{\mathrm{buy}} \cdot P_{i,k}^{\mathrm{ch}} \cdot \Delta t$
  - Battery Degradation Cost: $c^{\mathrm{deg}} \cdot (P_{i,k}^{\mathrm{ch}} + P_{i,k}^{\mathrm{dis}}) \cdot \Delta t$

where c_k^{sell} and c_k^{buy} are the time-varying electricity prices, c^{user} is the fixed price for the end-user, c^{deg} is the estimated cost of battery degradation per kWh cycled, and Δt is the time step duration.

System Constraints

The optimization is subject to the following constraints for each station i and time step $k \in [t, t+H-1]$.

Energy Balance Dynamics. The state of energy of the EV battery evolves according to:

$$E_{i,k} = E_{i,k-1} + \left(\eta^{\mathrm{ch}} P_{i,k}^{\mathrm{ch}} - \frac{1}{\eta^{\mathrm{dis}}} P_{i,k}^{\mathrm{dis}}\right) \cdot \Delta t \tag{3.8}$$

where the initial state $E_{i,t-1}$ is the currently measured energy level of the EV.

Power Limits and Mutual Exclusion. Charging and discharging powers are bounded by the EV's capabilities and controlled by a binary variable $z_{i,k}$ to prevent simultaneous operation.

$$0 \le P_{i,k}^{\mathrm{ch}} \le P_i^{\mathrm{ch,max}} \cdot z_{i,k} \tag{3.9}$$

$$0 \le P_{i,k}^{\mathrm{dis}} \le P_i^{\mathrm{dis,max}} \cdot (1 - z_{i,k}) \tag{3.10}$$

State of Energy (SoE) Limits. The battery energy level must remain within its physical operational window.

$$E_i^{\mathrm{min}} \le E_{i,k} \le E_i^{\mathrm{max}} \tag{3.11}$$

User Satisfaction (Hard Constraint). The desired energy level must be met at the time of departure. This is modeled as a hard constraint, reflecting a non-negotiable service requirement.

$$E_{i,k_{\text{dep}}} \geq E_i^{\text{des}} \tag{3.12}$$

where k_{dep} is the predicted departure step of the EV within the horizon.

Transformer Power Limit. The total net power drawn from (or injected into) the grid by all charging stations must not exceed the transformer's maximum capacity.

$$\sum_{i \in \text{CS}} (P_{i,k}^{\text{ch}} - P_{i,k}^{\text{dis}}) \leq P^{\text{tr,max}} \tag{3.13}$$

3.11 Conceptual Comparison: PuLP MPC vs. Gurobi Offline Optimizer

While both the PuLP MPC and the Gurobi offline optimizer are used to solve the EV charging problem, they operate on fundamentally different principles and serve distinct purposes. This section provides a discursive comparison of their core concepts.

3.11.1 Core Philosophy: Controller vs. Judge

The most significant difference lies in their philosophy. The **PuLP MPC** is designed as a **controller**. It operates online, making decisions in real-time with incomplete information about the future (e.g., EV arrivals, price fluctuations beyond the prediction horizon). Its goal is to find a practical and robust strategy for the immediate future.

Conversely, the **Gurobi formulation** acts as a **judge**. It is an offline tool that solves the problem over the entire simulation period with perfect hindsight (a-posteriori). Its purpose is not to control the system in real-time, but to establish a theoretical performance benchmark—the "perfect score"—against which the performance of a practical controller like the MPC can be measured.

3.11.2 Objective Function: Operational Profit vs. Energy Arbitrage

The objectives, while both related to profit, reflect their different roles. The PuLP MPC maximizes a detailed **Net Operational Profit**, incorporating a realistic business model that includes revenue from end-users and operational costs like battery degradation. This makes its decisions economically grounded from a business perspective.

The Gurobi optimizer, on the other hand, typically maximizes profit from a simpler **energy arbitrage** model, focusing on the difference between buying and selling electricity. While it includes a penalty for not meeting user demand, it does not explicitly account for the same level of operational economic detail as the MPC.

3.11.3 Handling of User Satisfaction: Hard vs. Soft Constraints

This distinction is critical from an operational standpoint. The PuLP MPC treats user satisfaction as a **Hard Constraint**. The EV *must* reach its desired energy level by its departure time. If the model determines this is impossible, the optimization problem becomes infeasible, signaling a failure to meet a mandatory service level agreement.

The Gurobi formulation treats user satisfaction primarily as a **Soft Constraint** via a penalty term in its objective function. This allows the optimizer to make a trade-off: it can choose to not fully charge a vehicle if the economic benefit of doing so (e.g., selling a large amount of energy to the grid at a high price) outweighs the penalty for customer dissatisfaction. This is useful for theoretical analysis but less practical for guaranteeing service.

Bibliography

[1] Farag Alfaverh, Mouloud Denaï, and Yifei Sun. "Optimal vehicle-to-grid control for supplementary frequency regulation using deep reinforcement learning". In: *Applied Energy* 325 (2022), p. 119881. DOI: 10.1016/j.apenergy.2022.119881.

[2] Mohamad Alsabbagh and Wan-Chi Siu. "Reinforcement learning for vehicle-to-grid: a review". In: *Journal of Energy Storage* 53 (2022), p. 105149. DOI: 10.1016/j.est.2022.105149.

[3] Alberto Bemporad. "Explicit Model Predictive Control". In: *Springer Handbook of Automation* (2013), pp. 883–898.

[4] Christoph R Birkl et al. "Degradation diagnostics for lithium ion cells". In: *Journal of Power Sources* 341 (2017), pp. 373–386.

[5] Francesco Borrelli, Alberto Bemporad, and Manfred Morari. *Predictive Control for Linear and Hybrid Systems*. Cambridge University Press, 2017.

[6] Greg Brockman et al. "OpenAI Gym". In: *arXiv preprint arXiv:1606.01540* (2016). DOI: 10.48550/arXiv.1606.01540.

[7] Eduardo F Camacho and Juan Bordons. *Model Predictive Control*. Springer Science & Business Media, 2013.

[8] European Commission. *Fit for 55: Delivering the EU's 2030 Climate Target*. Accessed: 2025-09-21. 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX%3A52021DC0550.

[9] C. R. Cutler and B. L. Ramaker. "Dynamic Matrix Control – A Computer Control Algorithm". In: *Proceedings of the Joint Automatic Control Conference*. Paper No. WP5-B. American Control Conference. San Francisco, USA, 1980.

[10] Gabriele Faggio. "Design and Testing of Online and Offline Optimization Algorithms for Vehicle-to-Grid (V2G) Industrial Applications". MA thesis. Politecnico di Milano, 2023.

[11] Scott Fujimoto, Herke van Hoof, and David Meger. "Addressing Function Approximation Error in Actor-Critic Methods". In: *Proceedings of the 35th International Conference on Machine Learning (ICML)*. 2018. URL: https://arxiv.org/abs/1802.09477.

[12] Tuomas Haarnoja et al. "Soft Actor-Critic Algorithms and Applications". In: *arXiv preprint arXiv:1812.05905* (2019). URL: https://arxiv.org/abs/1812.05905.

[13] Ghaith Al-Hmoud and Hamed Al-Raweshidy. "A Review of Smart Grid Evolution and Reinforcement Learning: Applications, Challenges and Future Directions". In: *Energies* 18.7 (2024), p. 1837. DOI: 10.3390/en18071837.

[14] Sinan Ibrahim et al. "Comprehensive Overview of Reward Engineering and Shaping in Advancing Reinforcement Learning Applications". In: *IEEE Access* PP.99 (2024). Accessed: 2025-09-21, pp. 1–1. DOI: 10.1109/ACCESS.2024.3504735. URL: https://arxiv.org/abs/2408.10215.

[15] Sheroz Ullah Khan et al. "A Review of Bidirectional Charging Grid Support Applications and Control". In: *Energies* 17.6 (2024), p. 1320. DOI: 10.3390/en17061320.

[16] Vinay Kumar, Surender Singh, and Dinesh Kumar. "Integration of electric vehicle into smart grid: a meta heuristic algorithm for energy management between V2G and G2V". In: *Frontiers in Energy Research* 12 (2024), p. 1357863. DOI: 10.3389/fenrg.2024.1357863.

[17] Igor Kuznetsov et al. "Controlling Overestimation Bias with Truncated Mixture of Continuous Distributional Quantile Critics". In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*. 2020. URL: https://proceedings.mlr.press/v119/kuznetsov20a/kuznetsov20a.pdf.

[18] Haijie Li et al. "Investigating Dynamic Behavior in SAG Mill Pebble Recycling Circuits: A Simulation Approach". In: *Minerals* 14.7 (2024). Accessed: 2025-09-21, p. 716. DOI: 10.3390/min14070716. URL: https://www.mdpi.com/2075-163X/14/7/716.

[19] Timothy P Lillicrap et al. "Continuous Control with Deep Reinforcement Learning". In: *arXiv preprint arXiv:1509.02971* (2015).

[20] Zhaomiao Liu et al. "Optimal scheduling for charging and discharging of electric vehicles based on deep reinforcement learning". In: *Frontiers in Energy Research* 11 (2023), p. 1273820. DOI: 10.3389/fenrg.2023.1273820.

[21] Logeshwaran et al. "A Comparative Study of Deep Reinforcement Learning Algorithms". In: *arXiv preprint arXiv:2407.14151* (2022). URL: https://arxiv.org/abs/2407.14151.

[22] Horia Mania, Aurelia Guy, and Benjamin Recht. "Simple Random Search Provides a Competitive Approach to Reinforcement Learning". In: *arXiv preprint arXiv:1803.07055* (2018). URL: https://arxiv.org/abs/1803.07055.

[23] D. Q. Mayne et al. "Constrained model predictive control: Stability and optimality". In: *Automatica* 36.6 (2000), pp. 789–814. DOI: 10.1016/S0005-1098(99)00214-9.

[24] Carlos A Minchala-Ávila, Paúl Arévalo, and Diego Ochoa-Correa. "A Systematic Review of Model Predictive Control for Robust and Efficient Energy Management in Electric Vehicle Integration and V2G Applications". In: *Modelling* 6.1 (2025), p. 20. DOI: 10.3390/modelling6010020.

[25] Volodymyr Mnih et al. "Asynchronous Methods for Deep Reinforcement Learning". In: *Proceedings of the 33rd International Conference on Machine Learning (ICML)*. 2016, pp. 1928–1937. URL: https://arxiv.org/abs/1602.01783.

[26] Andrew Y Ng, Daishi Harada, and Stuart Russell. “Policy invariance under reward transformations: Theory and application to reward shaping”. In: *ICML* 99 (1999), pp. 278–287.

[27] Stylianos Orfanoudakis et al. “A Deep Reinforcement Learning-Based Charging Strategy for Electric Vehicles with V2G Services in a Volatile Market”. In: *arXiv preprint arXiv:2209.09772* (2022). DOI: 10.48550/arXiv.2209.09772.

[28] Stylianos Orfanoudakis et al. “EV2Gym: A Flexible V2G Simulator for EV Smart Charging Research and Benchmarking”. In: *arXiv preprint arXiv:2404.01849* (2024). DOI: 10.48550/arXiv.2404.01849.

[29] Alessandra Parisio, Evangelos Rikos, and Luigi Glielmo. “A Model Predictive Control Approach to Microgrid Operation Optimization”. In: *IEEE Transactions on Control Systems Technology* 22.5 (2014), pp. 1813–1827. DOI: 10.1109/TCST.2013.2295737.

[30] D. Qiu. “Reinforcement learning for electric vehicle applications in energy management: A review”. In: *Renewable and Sustainable Energy Reviews* 163 (2023), p. 112443. DOI: 10.1016/j.rser.2022.112443.

[31] J. Richalet et al. “Model predictive heuristic control: Applications to industrial processes”. In: *Automatica* 14.5 (1978), pp. 413–428. DOI: 10.1016/0005-1098(78)90001-8.

[32] Arman Sadeghi. “Cost and power loss aware coalitions under uncertainty in transactive energy systems”. In: *Université d’Ottawa / University of Ottawa* (2021). PhD Thesis.

[33] Gabriel Antonio Salvatti et al. “Electric Vehicles Energy Management with V2G/G2V Multifactor Optimization of Smart Grids”. In: *Energies* 13.5 (2020), p. 1191. DOI: 10.3390/en13051191.

[34] Tom Schaul et al. “Prioritized experience replay”. In: *arXiv preprint arXiv:1511.05952* (2015). DOI: 10.48550/arXiv.1511.05952.

[35] John Schulman et al. “Proximal Policy Optimization Algorithms”. In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 2017, pp. 3371–3380. URL: https://arxiv.org/abs/1707.06347.

[36] John Schulman et al. “Trust Region Policy Optimization”. In: *Proceedings of the 32nd International Conference on Machine Learning* 37 (2015), pp. 1889–1897. URL: https://proceedings.mlr.press/v37/schulman15.html.

[37] Johannes Schwenk et al. “A computationally efficient MPC for residential EV charging with V2G and 3-phase power flow”. In: *Energies* 15.16 (2022), p. 5923. DOI: 10.3390/en15165923.

[38] Mohamed M Shibl, Loay S Ismail, and Ahmed M Massoud. “Electric vehicles charging management using deep reinforcement learning considering vehicle-to-grid operation and battery degradation”. In: *Energy Reports* 10 (2023), pp. 494–509. DOI: 10.1016/j.egyr.2023.07.039.

[39] Ahmad Tavakoli et al. "Impacts of grid integration of solar PV and electric vehicle on grid stability, power quality and energy economics: a review". In: *IET Energy Systems Integration* 2.3 (2020), pp. 233–245. DOI: 10.1049/iet-esi.2019.0047.

[40] Jürgen Vetter et al. "Ageing mechanisms in lithium-ion batteries". In: *Journal of Power Sources* 147.1-2 (2005), pp. 269–281.

[41] Zhong Wan et al. "A dynamic reward-based deep reinforcement learning for V2G control with EV battery degradation". In: *Applied Energy* 309 (2022), p. 118462. DOI: 10.1016/j.apenergy.2021.118462.

[42] Mingyu Wang et al. "Multi-Agent Reinforcement Learning is a Sequence Modeling Problem". In: *arXiv preprint arXiv:2205.14953* (2022). URL: https://arxiv.org/abs/2205.14953.

[43] C. David White and Willett Kempton. "Vehicle-to-grid technology for frequency regulation and peak shaving". In: *Journal of Power Sources* 196.3 (2011), pp. 3972–3980. DOI: 10.1016/j.jpowsour.2010.11.010.

[44] H. Xie. "Reinforcement learning for vehicle-to-grid: A review". In: *ScienceDirect* (2025). DOI: 10.1016/j.sciad.2025.100008.

[45] Y. Zhang. "Transfer deep reinforcement learning-based large-scale V2G continuous charging coordination with renewable energy sources". In: *arXiv preprint arXiv:2210.07013* (2023).

[46] Hongbin Zou. "Deep reinforcement learning-based control strategies for electric vehicle charging and V2G services". In: *Advances in Applied Energy* (2025), p. 100214. DOI: 10.1016/j.adapen.2025.100214.