Reinforcement Learning for Three-Phase Voltage Regulation on Australian Distribution Networks

Eddie Atkinson

This report is submitted as partial fulfilment
of the requirements for the Honours Programme of the
Department of Computer Science and Software Engineering,
The University of Western Australia,
2022

Abstract

In the last decade, rooftop photo voltaic (PV) generation has grown exponentially in Australia. Installed capacity has increased 27-fold between 2010 and 2020, and in 2021 PV generation accounted for 7% of energy entering the national grid. However, growing pains are emerging. High levels of distributed generation increase stress on ageing electricity distribution networks that deliver electricity the "last mile" to homes and businesses.

Though distribution network monitoring is ad-hoc and limited in Australia, reports of voltage instability on networks with high PV penetration are emerging. Given the cost and disruption of upgrading network infrastructure, a new approach to managing voltage and PV power exports is required.

Reinforcement learning (RL) is a technique which has shown promise in this area. Previous work has demonstrated that distributed agents are able to maintain safe voltages whilst maximising power exports on IEEE test networks. However, the performance of this technique on Australian-style imbalanced three-phase distribution networks remains unexplored.

This project compares the performance of distributed RL agents with centralised linear optimisation models in regulating voltage on an imbalanced three-phase Australian network. Though agents appeared to become trapped in a local optima, they developed a voltage regulation strategy which outperformed a business as usual base case.

Existing works often use proprietary datasets, external voltage regulation infrastructure, and balanced, unrealistic benchmark networks. Therefore, a key contribution of this work is an open-source three-phase training environment and trained agents. The training environment includes a model of a real Australian distribution network, and the agents are trained to regulate voltage on this network.

Keywords: Multi-agent reinforcement learning, energy
CR Categories: G.3, I.2

Acknowledgements

I would like to thank the following people for their support throughout my Honours year:

- My supervisors, Mark Reynolds and Chris Townsend for their wisdom.
- Frederik Geth (formerly of the CSIRO, now GridQube) without whose patience, kindness, and deep expertise in energy modelling this project would not have been possible.
- Adam Green (formerly of Gridcognition) for his assistance in interpreting and improving the behaviour of my agents.
- Clare Nolan who provided me with much needed encouragement to keep going when nothing was working and my agents were curtailing everything.
- Sophie Giraudo for her unconditional support throughout this project, and unerring willingness to explain basic mathematics to me for the 8th time today.
- Other friends, family, and co-workers who were submitted to blistering and often unprovoked lectures on how the grid is broken and why it didn't need to be. I promise I'll have something else to talk about soon.

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1 Background and Motivation

Residential photo voltaic (PV) generation is a major success story of the renewable energy industry. It is cheap to deploy and maintain; low carbon; modular; and growing rapidly. Global installed capacity increased from 6 GW in 2009 to 79 GW in 2019 [6]. In Australia distributed PV is experiencing exponential growth - capacity grew 20% in 2018 bringing national PV penetration to 19% [7].

However, distributed generation is experiencing growing pains. Ageing distribution networks which deliver energy "the last mile" to consumers are increasingly strained. High levels of distributed generation creates two-way power flows which produce voltage and frequency instabilities and accelerate wear on network equipment [8].

Australian distribution networks are particularly susceptible to these challenges due to their world leading level of rooftop PV penetration and historically high network voltages [3, 7]. Voltage violations are often the first challenge to manifest at high penetration levels and are due to variable generation and power injections by inverters into distribution networks. These violations damage consumer and network equipment, and force inverters to disconnect if voltages exceed acceptable levels. This curtailment limits household generation and revenue.

Failure to resolve voltage-based curtailment will significantly reduce the realised value of distributed energy resources (DER). Though curtailment largely affects distributed PV at present, it will equally affect other DER such as battery energy storage systems (BESS) and electric vehicles (EVs). Voltage-based curtailment threatens the economic and environmental potential of these assets. The successful integration of DER into the grid is valued at $4 billion by the Australian Energy Market Operator (AEMO), and in 2021 rooftop PV was responsible for the abatement of 17.7 million tonnes of carbon emissions [9, 10]. Failure to realise the potential of these assets due to local voltage conditions will result in significant foregone value.

Voltage violations are typically addressed by a combination of inverter standards and infrastructure upgrades. However, inverter standards face compliance challenges (40% of inverters installed after 2016 are non-compliant), and infrastructure upgrades are expensive and disruptive [11, 12]. Prior research has instead treated voltage violations as an optimisation problem in which generation should be maximised whilst maintaining power quality [13]. These works propose a variety of centralised and decentralised approaches employing "classical" optimisation techniques such as linear optimisation and model predictive control [14, 15].

In recent years reinforcement learning (RL) has shown promise in real-time control scenarios such as strategy games [16]. Initial work indicates that RL may be an effective means of regulating voltage using decentralised agents [17, 18, 19]. However, there are gaps in this work. Existing work explores the effectiveness of using reinforcement learning on IEEE benchmark datasets. Though convenient, these datasets do not resemble the voltage stressed, imbalanced three-phase networks found across Australia.

1.2 Research Aims

The aim of this project is to explore the viability of using distributed RL agents to control voltage on Australian distribution networks. Existing work focuses on balanced three-phase or single-phase networks which do not suffer from phase imbalance. Phase imbalance is a common feature of Australian distribution networks, is often exacerbated by high penetrations of rooftop PV, and worsens voltage instability. It is therefore of interest to examine the performance of distributed reinforcement learning agents in regulating voltage under these conditions.

The RL approach will be compared with centralised optimisation and an extreme "unrestricted export" base case. If successful, this approach has the potential to increase the hosting capacity of distribution networks whilst minimising curtailment of household DER.

1.3 Contribution

Existing works applying RL techniques to voltage management share the following properties:

- **Are multi-agent**: PV inverters are owned by different households, therefore works such as [17, 18, 20, 21] model them as individual agents who collaborate to regulate voltage.
- **Use IEEE benchmark networks**: The IEEE maintains a set of network models that are used for comparing the performance of network solvers. The IEEE 13, 33, and 133 bus networks used in [17, 18, 20, 21] are based on North American electricity networks, and as such bear little resemblance to Australian networks.
- **Utilise existing voltage regulation infrastructure**: Work by [17, 20] utilises network voltage regulation infrastructure such as capacitor banks and transformer on-load tap changers in addition to PV inverters to control voltage.
- **Use balanced three-phase or single-phase networks**: Balanced three-phase and single-phase networks lack phase imbalance which contributes to increased voltage volatility. Networks of this type are used in work by [17, 18, 21].

- **Employ centralised training and decentralised execution**: It is common for multi-agent deep reinforcement learning (MADRL) to employ centralised training which uses shared state to stabilise agents' policies. This approach is used by [17, 18, 20, 21].

In light of these properties the contribution of this work is an implementation of a multi-agent reinforcement learning voltage regulation strategy which:

- Uses freely available network models, load data and PV data
- Only leverages the voltage regulation capabilities of domestic PV inverters
- Operates on an imbalanced three-phase network
- Is trained on a model of a real Australian distribution network

1.4 Dissertation Outline

This dissertation first explores the structure of modern electricity grids, voltage conditions in Australia, and previous voltage management approaches through a literature review in chapter 2. The proposed methodology is described in chapter 3, and results are presented in chapter 4. Conclusions and future work are discussed in chapter 5.

CHAPTER 2

Literature Review

2.1 The Electricity Grid

The modern electricity grid is both a physical and economic concept. The physical grid is the infrastructure necessary for the generation, transmission, and distribution of electricity. The economic grid consists of markets which model the movement of electricity as financial flows.

2.1.1 The Physical Grid

The infrastructure of the grid comprises three systems – generators, the transmission network, and the distribution network. Figure 2.1 shows the relationship between these systems.

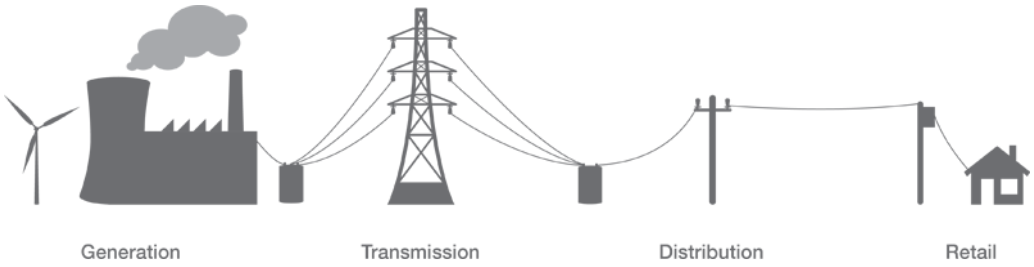

Figure 2.1: Diagram of the structure of the electricity grid [1]

Traditional energy grids are highly centralised. Large generators are located near fuel sources such as coal mines, large bodies of water, and natural gas wells, typically far from residential and industrial customers. Therefore, electricity must be transmitted over large distances from generators to load centres by a network of high voltage (HV) lines – the transmission network [22]. As it handles bulk electricity transfer, robustness is essential for the transmission network [23]. The network must support two-way electricity flow in case a branch of the network is unavailable due to line, transformer or generator failure.

The distribution network starts at the edge of populated areas where HV electricity from the transmission network is stepped down to medium voltage (MV). MV feeders distribute electricity throughout a neighbourhood before being stepped down to low voltage (LV). LV feeders then bus electricity to individual homes and businesses in

a small area, such as a single street. Figure 2.2 shows the topology of a typical distribution network in Perth, Western Australia.

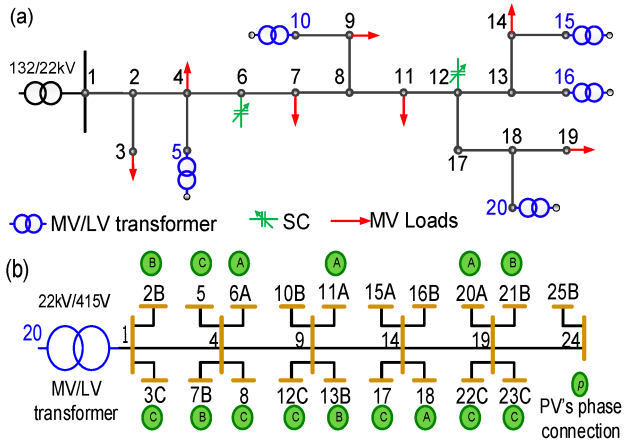

Figure 2.2: Topology of a 145-bus distribution network participating in the Perth Solar City project: (**a**) the MV feeder, (**b**) the LV feeder on bus 20 of the MV feeder [2]

Unlike the transmission network, the distribution network is not designed for bi-directional power flows [8].

In Australia the electricity network operates on three-phase alternating current (AC) power. In a three-phase power system three wires and a neutral return wire are used to transmit electrical energy. The AC waveform in each wire is 120° out of phase with the other wires. Three-phase power is a natural consequence of the use of spinning generators for the production of electricity. As a magnet's electrical field moves past a generator's coil it induces a current in that coil proportional to the strength of the magnetic field felt by the coil. The regular rotation of a magnet in a generator therefore causes an oscillation in the current output of each coil. This regular oscillation is an AC waveform.

Generators typically have three coils equidistant apart, producing three phases of electricity, each with its peak current output 120° apart. This three-phase electricity is transmitted through the grid and individual properties on the distribution network are connected to one or all of the phases. Figure 2.3 shows the relationship between the coils of a spinning generator and houses on the distribution network.

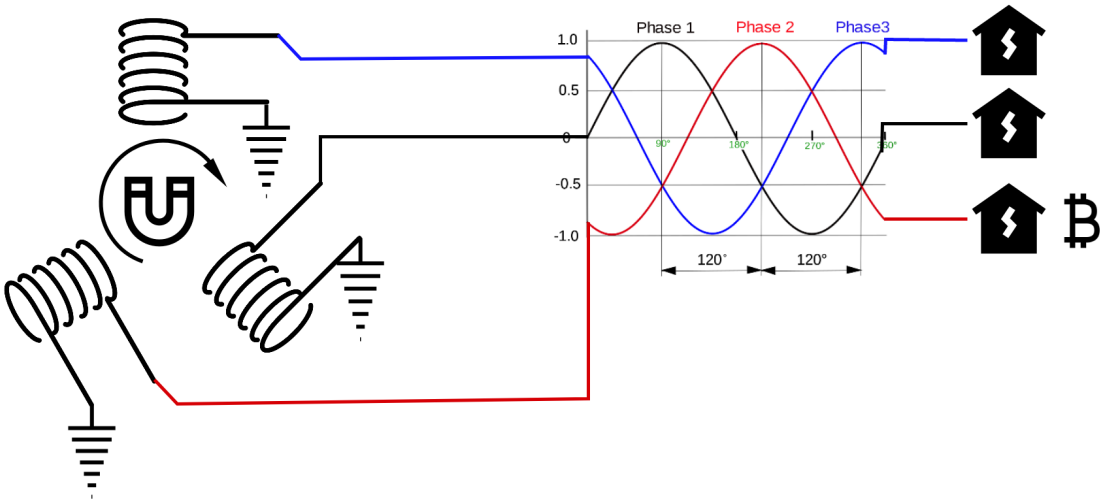

Figure 2.3: Diagram of the relationship between the magnet and coils of a spinning generator and households on the distribution network. One house on phase 2 has significant load compared to the other phases, resulting in phase imbalance, and increased voltages on the other phases.

One challenge that can emerge with three-phase power is phase imbalance which occurs when there is more load or generation on one phase of a three-phase network than others [12]. High load on one phase can induce voltage rises on adjacent phases, whilst low load due to high generation can reduce voltage on adjacent phases [12]. The installation of PV on a distribution network may exacerbate or ameliorate phase imbalance by increasing or reducing the load difference between phases.

2.1.2 The Economic Grid

Besides its physical dimension the modern grid also consists of a financial abstraction – the economic grid. The economic grid models and assigns value to the flow of electrical energy between participants in various markets.

Modern deregulated grids operate wholesale electricity and ancillary services markets [24]. Wholesale markets exchange electrical energy, whilst ancillary services markets maintain power quality by providing services such as frequency regulation and system restarts [24].

Australia's main electricity market is the National Electricity Market (NEM). The NEM serves 10 million customers and covers all states but Western Australia and the Northern Territory [25]. The Australian Energy Market Operator (AEMO) operates

all wholesale and ancillary services markets in the NEM [25].

Traditionally residential customers do not directly participate in wholesale electricity and ancillary services markets. Instead, retailers manage the volatility of wholesale markets on behalf of customers, purchasing financial products such as swaps, hedges, options, and caps, and charging customers a predictable retail tariff [25]. For customers with distributed generation retailers pay a flat feed in tariff (FiT) for their generated energy.

However, distributed energy resources (DER) are changing this dynamic. Retailers are increasingly offering products which expose customers to the wholesale spot price for both generation and consumption [26]. Customers are also able to participate in energy markets via virtual power plants (VPPs). VPPs aggregate DER such as PV and BESS from a large number of customers into a single operating unit to provide energy and market services [26]. AEMO estimates that successfully integrating DER into the grid via structures such as VPPs will yield $4 billion worth of value [9]. However, failure to resolve voltage challenges on local distribution networks threatens the ability of DER to freely participate in the electricity grid.

2.2 Distribution Network Voltage Management

To understand and resolve the voltage challenges posed by distributed generation, a mathematical formulation based on physical principles is required.

Electrical power is described by three distinct but related components – apparent, active and reactive power. Apparent power is the total power flowing through a circuit, active power performs work, and reactive power is waste. A pint of beer is often used to describe these components. A customer pays for the entire pint (apparent power), has their thirst quenched by the liquid component (active power), and derives no value from the foam (reactive power). The mathematical and intuitive relationship between these components of power is depicted in Figure 2.4.

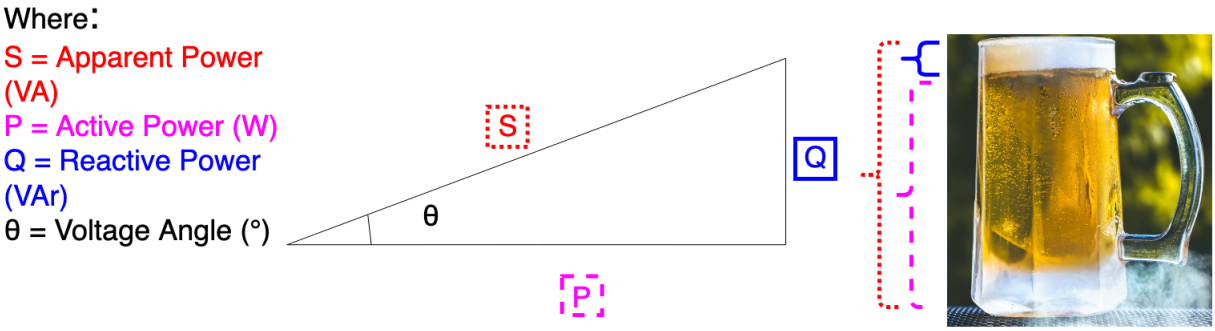

Figure 2.4: The power triangle describing the mathematical relationship between apparent, active, and reactive power with an analogy drawn to beer.

The ratio of a load or generator's real power to its apparent power is its power factor. The power factor determines the size and direction of distribution network voltage changes due to the operation of a load or generator. The effect of reactive and active power injections on the distribution network are described by Figure 2.5 and Equation 2.1.

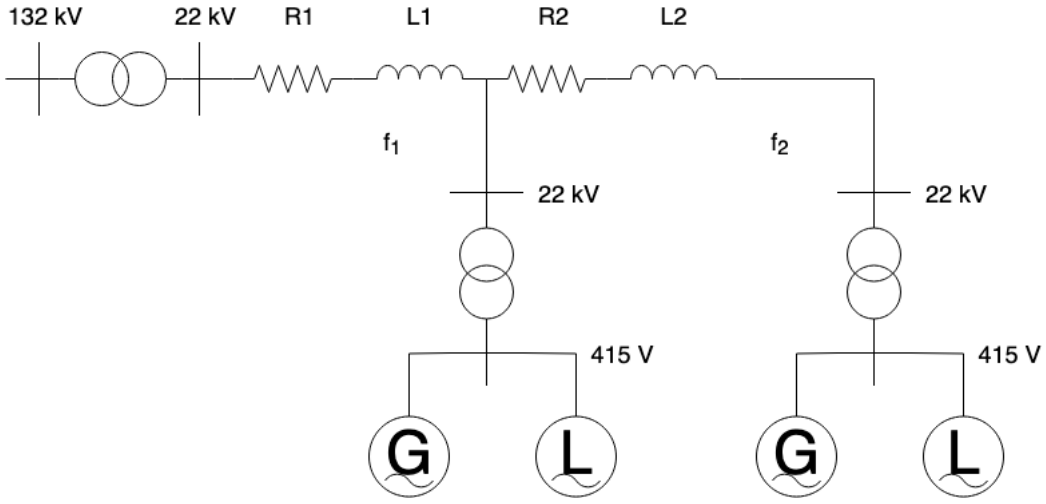

Figure 2.5: A simple distribution network with two LV feeders, f_1 and f_2 Each hosts generators (G) and loads (L).

$$\frac{\Delta \hat{V}}{\hat{V}_{f1}} = \frac{R_{th} \times P_{f1} + X_{th} \times Q_{f1}}{\hat{V}_{f2}^2} \tag{2.1}$$

where:
R_{th} = Approximate conductor resistance between the HV-MV transformer and f_1
P_{f1} = Active power absorbed at f_1
X_{th} = Approximate conductor reactance between the HV-MV transformer and f_1
Q_{f1} = Reactive power absorbed at f_1

$\hat{V}_{f2}$ = Nominal line voltage at f_2

Figure 2.5 shows a simplified MV feeder with two LV distribution transformers, each hosting loads and generators, and Equation 2.1 describes the voltage drop across the first branch in the feeder, f_1. Assuming low levels of reactive power absorption, if load on f_1 exceeds generation there is a voltage drop across f_1. Conversely, if generation exceeds load, then there is a voltage rise across f_1.

However, Equation 2.1 implies that voltage rises due to distributed generation can be managed by modifying the inverter's power factor and increasing reactive power absorption. Loads with a power factor between 0 and 1 are said to have a *leading* power factor and absorb reactive power, whilst those with a power factor between -1 and 0 have a *lagging* power factor and inject reactive power.

Controlling the power factor of an inverter to manage voltage is referred to as volt-var control. Inverters in Australia are required to be capable of injecting or absorbing reactive power with a power factor of 0.8. But under normal circumstances, inverters produce power with a unity power factor of 1 (pure active power injection) [4]. Injecting or absorbing reactive power to manage voltage reduces the export of useful active power. It is thus considered a "loss" and is avoided unless absolutely necessary.

2.3 Voltage Conditions in Australia

Australia is an ideal case study for voltage regulation approaches in high DER distribution networks. One-fifth of Australian homes have PV installed, with capacity growing 40% in 2019 [7, 27]. Though world-leading, Australia's level of PV penetration is unlikely to remain unique. Two-thirds of the world's population live in countries where PV or wind is the cheapest form of generation, and over half of renewable capacity installed in 2019 was PV [28].

The AS61000.3.100 standard specifies that the acceptable voltage range for domestic electricity in Australia is between 216V and 253V [29]. Despite high levels of PV penetration, the extent of voltage violations in Australian distribution networks is poorly understood. A 2019 AEMO report on distribution network conditions notes that there "...is little direct monitoring of loads and voltages on LV transformers and circuits, and on individual phases of those circuits" [30]. Work by [12] studied data gathered from PV inverters across the NEM, finding a large number of distribution networks operate near the 253V limit for some of the year, and in certain cases, at all times.

Follow up work by [3] examines PV generation data from South Australia to understand the extent and costs of PV curtailment to households. This work found an average daily curtailment of 1.1% on clear spring days where generation is high and load is low [3]. However, considerable variation in curtailment exists with some sites experiencing 46%–95% curtailment per day [3]. Figure 2.6 shows curtailment of a PV inverter in South Australia on a clear day, likely due to voltage violations on the distribution network [3].

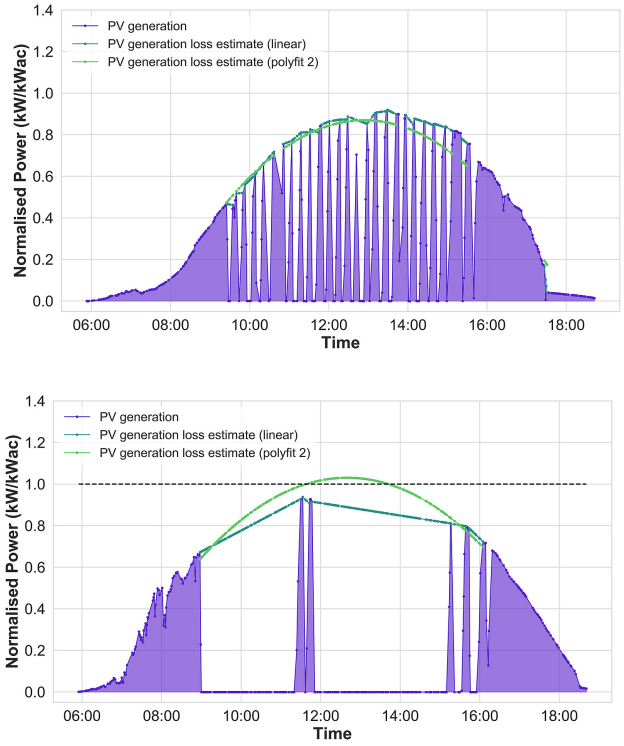

Figure 2.6: Curtailment of a PV inverter in South Australia, likely due to over voltage [3]

At present, curtailment largely affects PV, but poses a risk to all DER. Voltage-based curtailment threatens the future participation of DER in VPPs, limiting the availability of cheap, clean energy to the grid, and reducing the significant economic benefits of a successful DER-grid integration [9].

Given the continued growth of distributed generation in Australia, its economic potential and its environmental value, finding a solution to voltage violations is essential.

2.4 Voltage Regulation Approaches

Voltage regulation on distribution networks is a well-studied problem which has been approached from many angles. These approaches fall into three broad categories – infrastructure, centralised, and decentralised. Figure 2.7 shows the relationship between the categories, and their sub-categories, which are surveyed in this literature review.

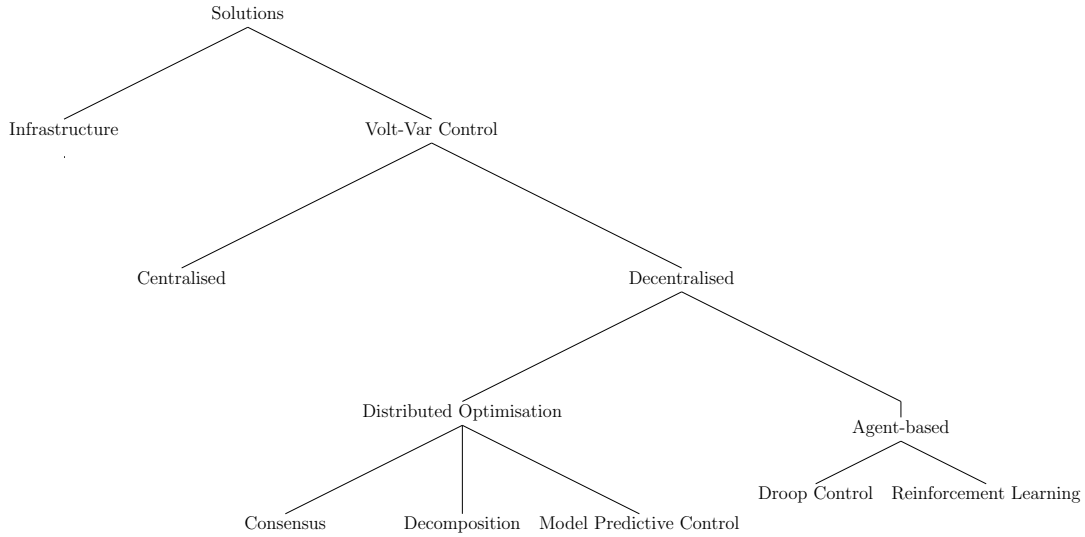

Figure 2.7: Taxonomy of voltage regulation approaches

2.4.1 Infrastructure

Traditionally distribution network voltage is managed by investment in infrastructure such as on-load tap changers (OLTCs), no-load tap changers (NLTCs), and switched capacitors (SCs) [12]. Though effective, these regulation devices are slow acting, expensive, and may be damaged when used to moderate voltage fluctuations from stochastic generators such as PV.

Distribution network operators have historically resisted the use of reactive and active power output for managing voltage as it may interfere with operation of network equipment [13].

However, given the disadvantages of employing infrastructure to address voltage violations, alternate approaches based on volt-var control have been proposed. This change in policy is reflected in recent revisions to inverter standards in Australia which require new inverters to have volt-var capabilities [31, 4].

2.4.2 Centralised

Work by [13] classifies voltage regulation methods as centralised or decentralised according to the way in which decisions regarding reactive and active power set points are determined. Centralised methods treat voltage management as a dispatch problem in which a central controller monitors network conditions and directly controls inverters' power output [13]. Decentralised approaches meanwhile rely on inverters locally sensing network conditions and autonomously prosecuting a voltage regulation strategy [13].

Centralised voltage regulation strategies determine active and reacting power set points by solving the optimal power flow as defined in [21]. Given the optimal power flow problem is a system of linear equations, linear, mixed-integer and nonlinear programming are commonly used for centralised voltage regulation. Work by

[32] uses mixed integer non-linear programming to manage voltage using a SCADA system to control DER on a simulated Finnish distribution network. This approach was able to maintain acceptable voltage ranges whilst minimising DER curtailment.

Though undoubtedly effective, centralised voltage regulation is impractical to implement in Australia. Centralised methods require an accurate model of the entire distribution network, up-to-date consumption and generation data, and some means of controlling assets to respond to voltage conditions. This is simply impossible in Australia, where distribution network monitoring, let alone control, is virtually non-existent [30]. Direct control of customer assets also raises questions of fairness – optimality is great as long it's not your house's PV getting curtailed. Finally, centralised approaches require significant computational resources to solve the optimal power flow problem even for small networks [21]. Given the real-time nature of voltage management the computational requirements of centralised optimisation are neither practical nor cost effective.

2.4.3 Decentralised

Decentralised voltage regulation approaches select active and reactive power set points on the basis of local and shared observations. They are an attractive alternative to centralised approaches as they require little investment in additional infrastructure to implement.

Decentralised approaches fall into two broad categories – distributed optimisation, and agent-based approaches.

Distributed Optimisation

Distributed optimisation approaches are derived from centralised formulations of the non-convex optimal power flow problem [33]. In the centralised setting convex relaxations are applied, offering guaranteed convergence to global optima under certain conditions (for example, a radial network topology) [33]. Distributed optimisation approaches divide the centralised optimal power flow problem and iteratively solve it using local communication. The most common distributed optimisation techniques are:

- **Consensus**: Consensus methods share a variable between neighbours to compute a common state. This state is used to solve the distributed optimisation problem iteratively until a global optimum of decision variables is agreed upon which minimises the problem's cost function [33].

  Various implementations of consensus algorithms for voltage regulation exist. The work of [34] proposes a gradient ascent method which optimises reactive and active power output. Meanwhile [35] employs a dual decomposition and feedback mechanism that alternates between sensing network conditions and altering reactive power injections. In [36] distributed evolutionary algorithms achieve fast-convergence to near-optimal solutions for real-time control.

- **Distributed Model Predictive Control**: Model Predictive Control (MPC) is a standard control method for large industrial plants which is capable of handling multi-variable control problems, is easy to tune, and explicitly specifies constraints [15]. MPC operates in discrete time steps by minimising a cost function associated with deviation from an ideal system state [15]. The cost function is evaluated with respect to an N step time horizon to minimise the sum of errors and produce an N step control sequence [15]. Only the first step of this control sequence is used on each iteration and is recomputed at each time step [15].

  The work of [37] proposes a network of hybrid distribution transformers (HDT) which regulate voltage using robust MPC based on local voltage measurements. Though successful in regulating over and under voltage events, this approach requires the replacement of some or all of the MV-LV transformers on a distribution network in order to operate, a non-trivial expense [37]. Communication between adjacent distributed generators is used by [38] to solve a local optimal control problem to derive reactive and active power set points. A similar approach is proposed by [39] in which BESS share measurements with their neighbours to regulate voltage via MPC. Using consensus to distribute computation is an effective means of scaling MPC, whose computational requirements grow rapidly with an increasing number of devices [37].

- **Decomposition**: Decomposition techniques reduce a large optimisation problem to a set of sub-problems which can be iteratively solved in a distributed manner [15]. These methods divide the optimisation problem into areas according to criteria such as proximity, information availability, or bus controllability [15].

A significant limitation of distributed optimisation approaches is their requirement for robust, real-time communication between neighbours. This necessitates infrastructure expenditure, and is associated with privacy and cybersecurity risks. Furthermore, the speed of convergence for the optimisation problem, and therefore the response time of agents, is dependent on the speed and reliability of communication links [33].

Agent-based

Agent-based modelling is a technique which models complex systems through the interactions of discrete autonomous entities and their environment [40]. These entities are agents which are situated in some environment and act independently to achieve their design goals [40]. In addition to the ability to sense and affect change in their surroundings, intelligent agents have the following characteristics:

- **Reactivity**: Agents detect and respond to changes in their environment in a timely manner in a way that best achieves their goals.

- **Pro-activeness**: Agents "take the initiative" and flexibly update their behaviour to achieve their goals.

- **Social ability**: Agents are able to communicate with other agents, not only to exchange information, but also for the purposes of cooperation and negotiation.

Multi-agent systems (MAS) are systems comprised of two or more intelligent agents acting to achieve their objectives [40].

The most common agent-based voltage regulation strategies are Droop control and Reinforcement Learning.

In traditional grids, frequency and voltage are maintained by the inertia of spinning steam generators. Under high load, the rotation, and therefore frequency, of these generators droops to maintain system voltage, triggering the release of more steam and an increase in speed. Though inverters have no rotational inertia, Q-V droop control is a widely studied technique for inverter-based voltage regulation [41]. Under Q-V droop control, an increase in voltage due to high generation is compensated for by an increase in reactive power absorption. Inverters implementing droop control moderate system voltage by varying their reactive power output in real-time in response to local voltage readings [41].

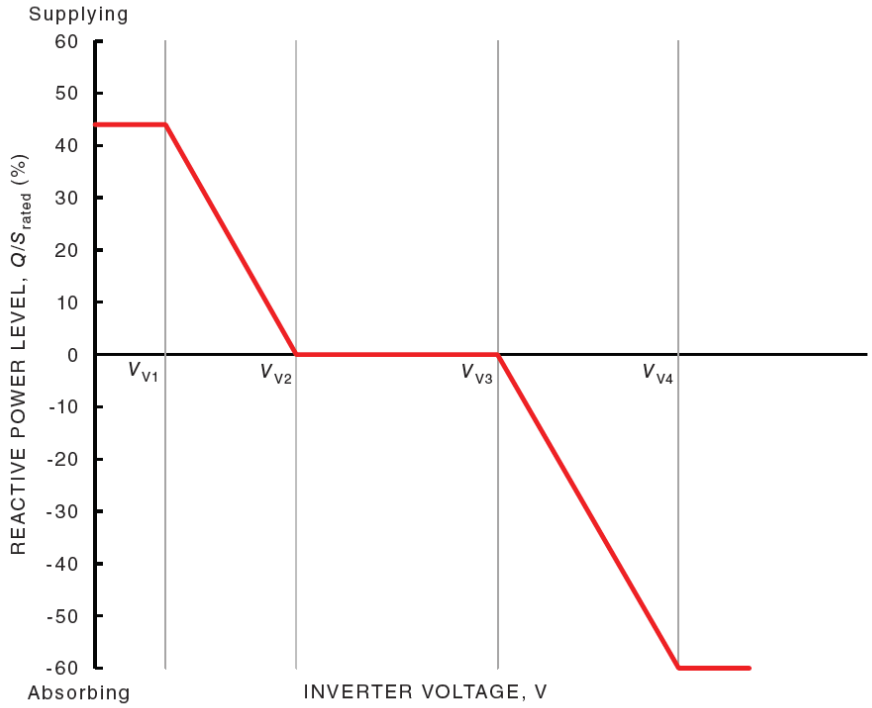

Figure 2.8: Example curve for volt-var control from the AS/NZ 4777.2 standard [4]

An example droop curve from the Australian inverter standard AS/NZ 4777.2 can be seen in Figure 2.8. As voltage increases, a greater proportion of apparent power is devoted to absorbing reactive power to reduce voltage.

Though effective in some circumstances, droop control techniques cannot be universally applied due to their sensitivity to the ratio of resistance and inductance in lines [42]. Furthermore, voltage regulation via droop control comes at the expense

of inverter active power curtailment which grows with distance from the distribution transformer [43]. Inverters further down the feeder must curtail their output significantly more to maintain safe voltages than those nearer the transformer [43]. A method for fairly distributing curtailment from droop control along a feeder is proposed by [43]. Though successful in evenly sharing curtailment, this approach results in the loss of almost one-third of generation on a test feeder, severely limiting generators' ability to contribute to the grid [43].

RL offers a more nuanced approach to voltage regulation in which agents learn the conditions of their network and modify their volt-var control strategy accordingly. As it is the focus of this work, an entire section of this literature review is dedicated to RL and its application to voltage regulation.

2.5 Reinforcement Learning

Reinforcement learning is a branch of machine learning in which an agent learns to maximise a reward signal via experimentation [44]. RL is attractive as it offers a means of programming agents to achieve goals without specifying *how* to achieve those goals [45].

In the context of energy research RL is a promising and understudied technique. The review of RL applictions in energy literature undertaken by [46] found few publications utilise state of the art methods. However, many of those that did employ RL methods reported a 10-20% improvement in performance over traditional approaches [46]. Successfully applied in control scenarios including building energy management and robotics, RL has potential as a voltage management technique [47, 48].

2.5.1 Reinforcement Learning Problem Definition

In an RL problem agents sense information about their state, have a set of actions available, and attempt to learn the mapping between states, actions, and rewards [44].

Deriving a pattern of behaviour which maximises reward in an RL problem may be framed as a search and optimisation problem, or as a statistical modelling problem [45]. In the former case, optimal agent behaviours for a particular environment are found by directly searching the space of possible behaviours [45]. Genetic programming and evolutionary algorithms are typically used for this task [45]. Genetic programming generates behaviours according to a heuristic, ranks them against a scoring criteria, combines the best candidates, and repeats the process until a stopping criteria is met. In the latter case, estimates for the utility of actions are generated using statistical methods and dynamic programming [45].

2.5.2 Statistical Approaches to Reinforcement Learning

Statistical approaches to RL are concerned with finding an approximation for the utility of states and/or state-action pairs. Markov decision processes (MDPs) are a convenient means of mathematically representing an RL problem which involves delayed reward - a long sequence of actions with insignificant rewards eventually results in an action which yields significant reward [45]. MDPs have the same set of actions and states as the original RL problem, and introduce two new concepts [45]:

- A reward function: $R : S \times A \rightarrow \Re$.
- A state transition function: $T : S \times A \rightarrow \prod S$ where $\prod S$ is a probability distribution over the set of states.

The reward function specifies the expected immediate reward associated with taking a given action in a given state, whilst the state transition function specifies the probability of transitioning to a state from another state given a particular action is selected [45]. Collectively these functions are referred to as the transition model.

Using the state transition and reward functions, the utility of a state can be represented using a policy or value function. A policy π is a mapping from states to actions [44]. When an agent is in a state s, a policy will direct it to select the action a which is associated with the highest expected reward. A value function $v_\pi(s)$ evaluates states according to the discounted future rewards that an agent expects to accrue starting from a given state [44]. When an agent is in state s, a value function will specify the utility of all the states reachable from the current state, allowing the agent to select an action a which maximises its expected reward by placing it in a high value state s'. A policy may be used to derive a value function - the value of a state is the reward associated with prosecuting a given policy starting from that state [44]. Similarly, a policy can be derived from a value function - the ideal policy is that which selects the actions which lead to the most rewarding states [44].

Statistical RL methods attempt to estimate the utility of actions by value function or policy approximation. Methods which perform value function or policy estimation directly are termed *model-free*, whilst those which estimate the transition model and use it construct a policy or value function are *model-based*.

For toy examples such as the Gridworld problem proposed by [44] complete tables of states and actions can be stored in memory and enumerated. However, most real-world problems consist of too many states and actions to be represented in this way. Instead, policies and value functions are encoded as parameterised functions which are learned [49]. Methods for learning these functions include:

- **Value function methods**: Value function methods attempt to approximate the ideal Q function $Q_\pi*$ – the mapping between states and actions that maximises long-term reward. The Q-learning algorithm proposed by [50] is a popular means of approximating $Q_\pi*$. Q-learning agents maintain an estimate of

the Q function $\hat{Q}(s,a)$, which is updated on each state transition according to:

$$\hat{Q}(s,a) = Q(s,a) + \alpha(r + \gamma max(Q(s',a')) - Q(s,a))$$

Where α is a learning rate parameter, r is the immediate reward for taking action a, and γ is the discount factor for future rewards. Q-learning incrementally improves the Q function by taking an action, evaluating its utility in terms of immediate and discounted future rewards, and adjusting its function estimate accordingly.

- **Policy based methods**: Policy based methods search the policy space directly using parameterised functions to estimate the optimal policy. Policy gradient approaches iteratively improve a policy by updating parameters along the gradient of long-term reward. The policy gradient at time t is given by:

$$\hat{g} = \hat{\mathbb{E}}_t \left[\Delta_\theta log \pi_\theta(a_t|s_t) \ \hat{A}_t \right]$$

Where $\hat{A}_t$ is the estimated advantage of an action - whether the final reward from taking an action was better or worse than expected, and $log\pi_\theta(a_t|s_t)$ is the probability of selecting an action a_t under the current policy. The gradient function is used to update parameter policy weights according to whether a state resulted in a better or worse total reward than anticipated. Popular gradient-based methods include REINFORCE, G(PO)MDP, and actor-critic algorithms such as soft actor-critic, TRPO, and PPO [51].

2.5.3 Deep Reinforcement Learning

In 2015 [16] developed an effective method for parameterising a value function using a deep neural network (DNN), sparking significant interest in the use of DNNs for RL in complex environments such as games [49]. A great deal of recent work in RL has involved integrating DNNs into existing approaches such as Q-learning, actor-critic, and policy gradient methods. This work has achieved human-level performance at tasks such as Go, classic Atari video games, and real-time strategy games [51].

2.5.4 Reinforcement Learning Voltage Regulation Approaches

Though deep RL is becoming increasingly mainstream, it remains uncommon as a voltage regulation technique. Existing works applying RL techniques to voltage regulation share several properties:

- **Multi-agent RL used**: As PV inverters are owned by different households which have shared, but competing interests, modelling voltage regulation as a multi-agent problem is a logical choice. Training multi-agent deep RL (MADRL) agents is often performed using offline simulations in which some information is shared amongst agents for training purposes and decentralised execution where no information is shared.

- **Use IEEE benchmark networks**: The IEEE maintains a set of network models that are used for comparing the performance of network solvers. These network models are freely accessible, well understood, and easy to use.
- **Utilise existing voltage regulation infrastructure**: In some jurisdictions external voltage regulation infrastructure such as switched capacitors and on-load transformer tap changers are available for voltage regulation. Combining these centrally controlled assets with decentralised RL agents enables multi-level voltage control.
- **Use balanced three-phase or single-phase networks**: Simulations of balanced three-phase or single-phase networks are straightforward to construct and pose a simpler voltage regulation task for agents than imbalanced three-phase networks.

A summary of the surveyed works and whether they have these properties may be seen in Table 2.1.

Work	MADRL	IEEE Benchmark Networks	Regulation Infrastructure	Balanced Three-phase or Single-phase
[17]	✓	✓	✓	✓
[18]	✓	✓		✓
[20]	✓	✓	✓	
[21]	✓	✓		✓

Table 2.1: Comparison of surveyed RL voltage regulation works

In [17] a two stage control process which coordinates slow acting electro-mechanical regulation devices with fast acting inverters is developed. OLTCs and capacitor banks are dispatched a day in advance by a centralised controller, whilst decentralised RL agents control the reactive power output of inverters in real-time using local information. The combination of agents trained using a multi-agent deep deterministic policy gradient (MADDPG) algorithm and centralised control performs similarly to centralised optimisation on the IEEE 33 bus network without requiring real-time communication [17].

Work by [18] presents a fully decentralised solution in which MADDPG agents learn offline and implement an online control strategy using local information. Performance of this method in regulating voltage is comparable to centralised optimisation on the IEEE 33 bus feeder [18].

The work of [20] uses a deep Q network (DQN) to train agents to regulate voltage on imbalanced IEEE 13 and 123 bus feeders. In this work the action space for agents includes inverter active and reactive power set points, and the state of switched capacitors and voltage regulators on the network. Consequently, the agents are able to successfully regulate the voltage whilst achieving minimal power loss [20].

Finally, work by [21] compares the performance of 10 MADRL algorithms to traditional control methods on single phase IEEE 33, 141 and 322 bus networks. Of the examined algorithms MADDPG had the best performance across the three networks, though did not outperform centralised optimisation techniques [21].

These works have several limitations in an Australian context:

- Balanced single-phase or three-phase networks rarely occur in Australia. Feeders which suffer from voltage violations are often imbalanced, therefore voltage regulation techniques should be tested under these conditions [12].
- The IEEE test cases used are not intended for benchmarking voltage regulation strategies, nor are they representative of Australian distribution network topologies.
- Australian networks often do not have the voltage regulation infrastructure used in these works such as capacitor banks or OLTCs.

This work intends to address these gaps in the literature by training MADRL agents on an imbalanced three-phase Australian distribution network without the use of external voltage regulation equipment.

CHAPTER 3

Methodology

3.1 Overview

This chapter describes the design of experiments used to assess the performance of RL agents in regulating voltage on an imbalanced three-phase Australian distribution network. The design of an environment is problem-specific and determines whether agents are able to learn desirable behaviours. Therefore, an in-depth discussion of the following inputs to the environment is provided:

- Network models
- Load and PV data
- Reinforcement Learning Problem Formulation
- Simulation environment

Specific details of experimental setups, including the construction of comparison models and performance metrics are also discussed.

3.2 Environment

In RL the environment refers to the world which agents observe, act, and receive rewards from. In the context of voltage regulation this environment is a distribution network comprised of loads, generators, and buses. The load, generation, voltage, power and phase angle on each bus comprise the state of this environment. Agents are individual PV inverters which observe the state of the network through local and shared measurement. Agents act to vary their output power and power factor to control voltage, and receive rewards if they are successful. Figure 3.1 shows the relationship between the components of a voltage regulation RL environment.

Given their importance in shaping agents' learning, the acquisition and treatment of each input to the environment is described in detail below.

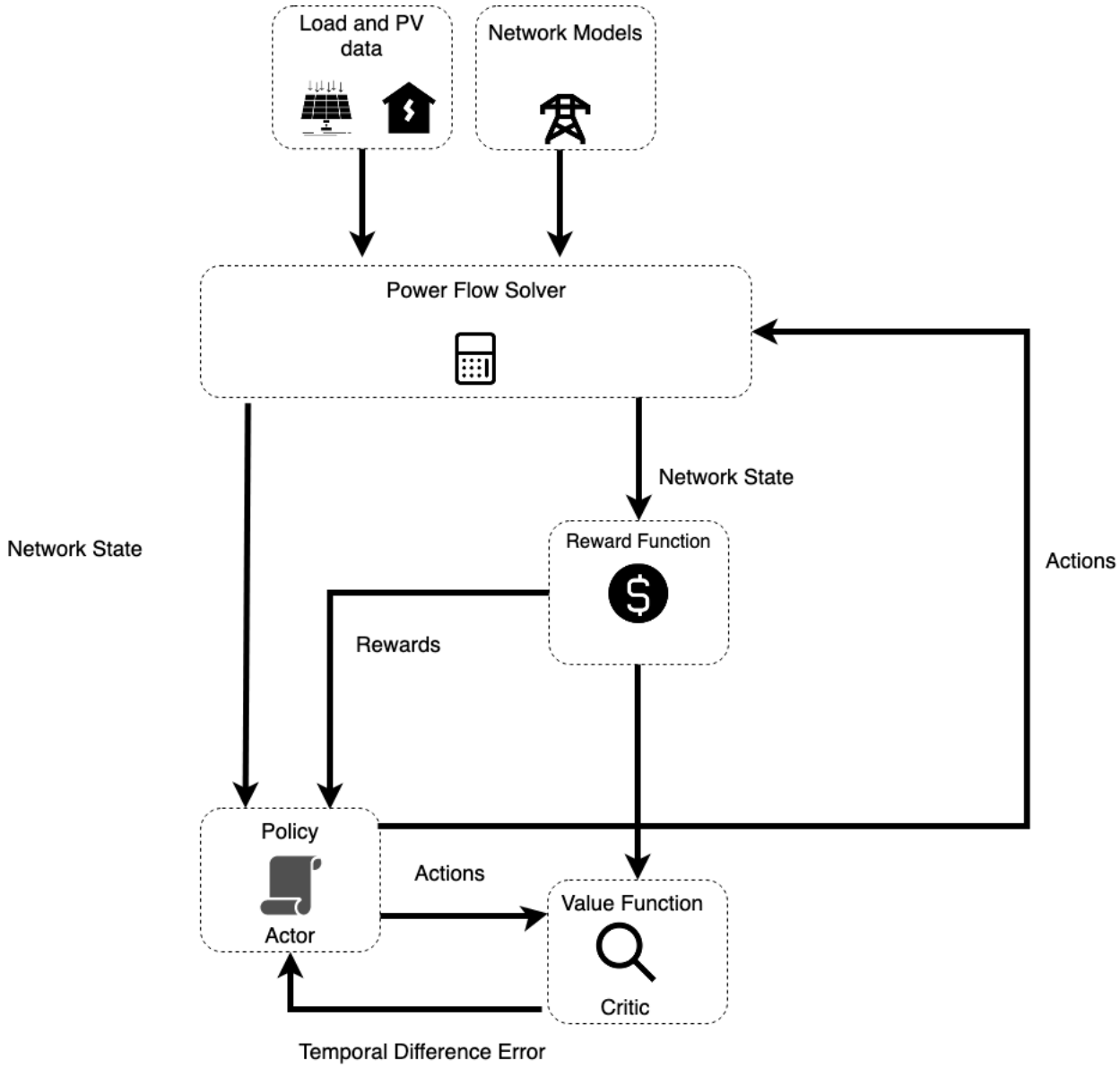

Figure 3.1: Overview of the RL environment used for training and testing agents

3.2.1 Network Models

Network models describe the physical configuration of assets in a distribution network. The topology of a network, and its distribution of assets play a leading role in determining whether a network is prone to voltage instability. Accurate, real-world network models are an important means of validating the viability of voltage management techniques via simulation. However, finding such network models is challenging.

Privacy concerns and difficulties in extracting data mean that network operators rarely publish data. Any data provided is incomplete, poor quality and often in proprietary formats. As a consequence, benchmark datasets such as the IEEE test cases are used to compare the performance of new techniques [52, 17, 21, 53]. Though

well understood and accessible, the most commonly used IEEE datasets are intended for testing network solvers, not for comparing voltage regulation techniques [54]. More realistic network models should be used to evaluate the performance of RL techniques.

Despite challenges in acquiring real network models, several open source datasets are available. For this project two such datasets were considered: the Australian Low Voltage Feeder Taxonomy (LVFT) and the Electricity North West Low Voltage Solutions (ENWL) dataset [5, 55].

The LVFT clustered 100,000 low voltage networks from 9 Australian distribution network operators into a set of 23 representative networks [5]. Meanwhile the ENWL project created 25 network models from Electricity North West's network [55].

Though networks in both of these data sets showed signs of voltage instability, it was more pronounced in the LVFT data, and directly exacerbated by the presence of distributed generation. Given this project's focus on an Australian context, it was therefore an obvious choice.

In addition to creating network models the original LVFT project analysed voltage network voltage conditions at varying degrees of PV penetration. At 50% penetration, Network J, an urban radial distribution feeder from South Australia experienced considerable voltage instability. Figure 3.2 shows the topology of network J, and Figure 3.3 shows its distribution of voltages across buses and phases at 50% PV penetration.

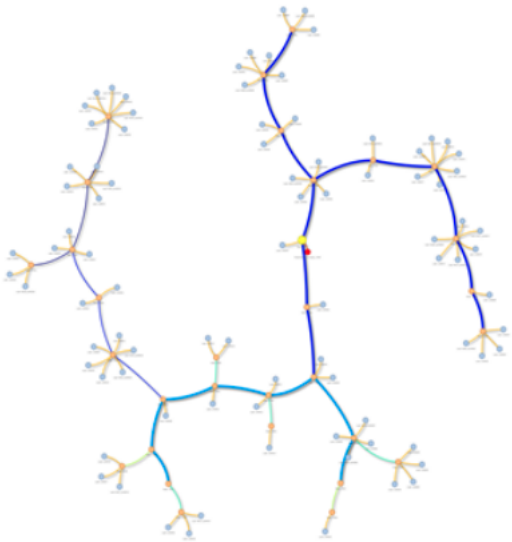

Figure 3.2: Topology of network J from the Australian LVFT study [5]

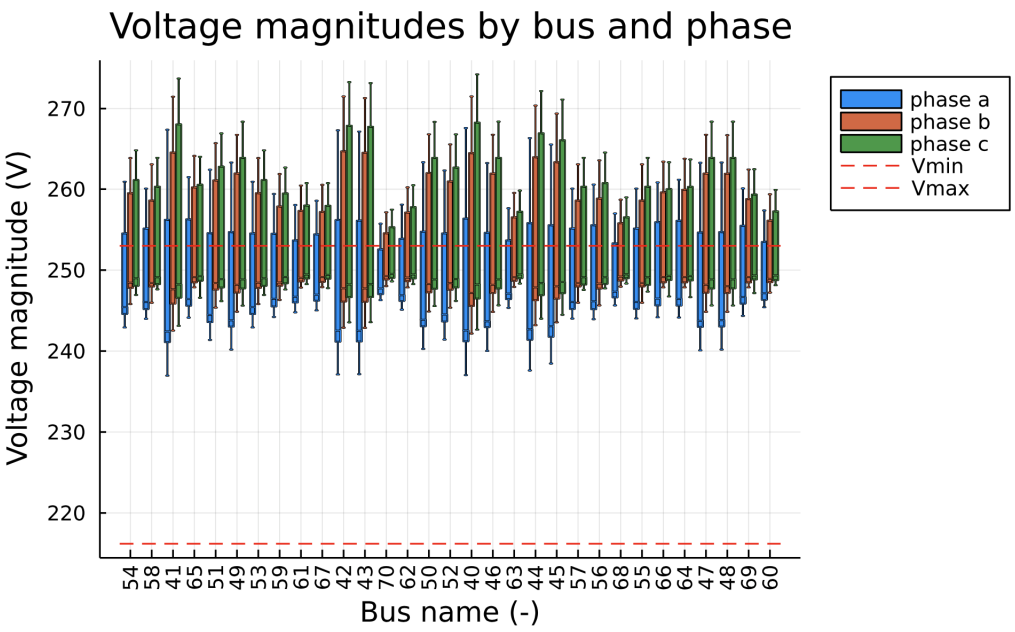

Figure 3.3: At 50% PV penetration network J experiences considerable voltage instability [5]. The vertical red dotted lines demonstrate acceptable voltage bounds.

Network J is a relatively small network of 87 loads and 31 buses experiencing voltage challenges at 50% penetration. It is therefore an ideal real-world test case for RL-based voltage regulation. Its instability provides sufficient intervals to train agents, whilst its small size reduces training time and aids interpretability of agent behaviour.

Though all networks produced by the LVFT are three-phase some networks, including Network J, have all their loads on the same phase. As a primary goal of this work is to benchmark the performance of RL agents in regulating voltage in the presence of phase imbalance, distributing loads across the phases was necessary.

This was achieved by randomly assigning loads to phases such that the phases contained the same number of loads, but not necessarily the same size of load. Given the voltage instability exhibited by Network J at 50% PV penetration, this level of penetration was used for this project. After loads were assigned to phases, PV systems were randomly assigned to 50% of the loads.

3.2.2 Load and PV Data

Load and PV data determine the scale of the voltage challenge faced by agents and their capacity to influence network voltage conditions. As such, using high resolution load and PV data which includes reactive and active power for simulations is essential. Given the role weather plays in determining energy usage patterns, data which includes different seasons is highly desirable. However, acquiring this data, specifically load data, is not straightforward.

Household load data is a sensitive form of personal information, and is therefore covered by strict privacy regulation in many jurisdictions [56]. Work by [57] provides a comprehensive review of publicly available load datasets. A subset of the surveyed datasets was considered for this project and is summarised in Table 3.1.

Name	Size	Households	Resolution	Reactive Power
Smart Grid Smart City	4 years	17,000	30 min	No
Pecan Street Free License	1 year	75	1s	No
ECO	6 months	9	1s	Yes
UK DALE	39-284 days	5	6s	Yes
ADRES	14 days	30	1s	Yes

Table 3.1: Comparison of available load data sets

Of the available datasets the Pecan Street dataset was selected. It is one of the largest and most comprehensive dataset available, covering an entire year and including PV data in addition to load data.

Though the Pecan Street dataset includes 75 households, they are spread across three climatic regions – California, New York, and Austin. As weather-dependent load differs greatly between these regions throughout the year, only data from Austin was used. Of the three regions Austin is most climatically similar to many major Australian cities and receives a similar level of solar irradiance [58].

Though high quality, this dataset could not be used in its raw form. The data cleaning process applied to the load and PV data is depicted in Figure 3.4.

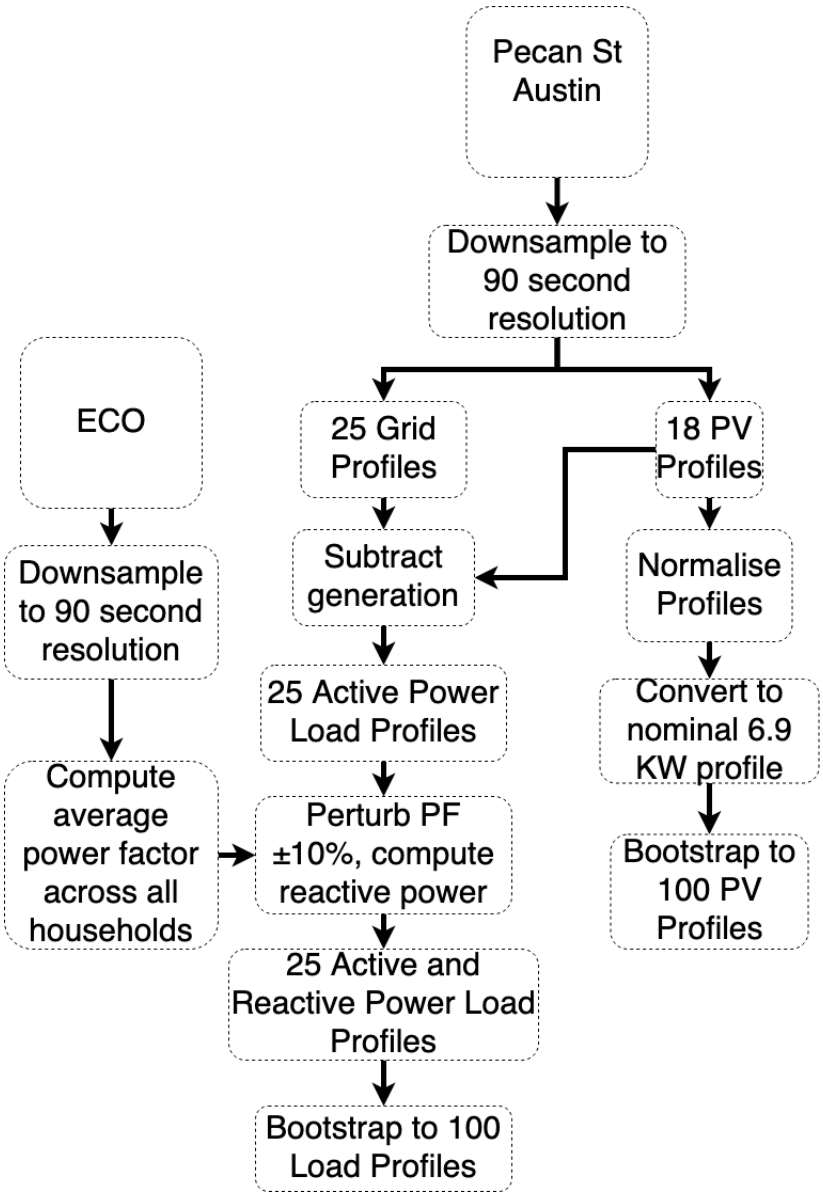

Figure 3.4: An overview of the data cleaning process applied to the Pecan Street data set

Though high resolution input data is important, a balance must be struck between accuracy and computational resources. The raw data for 25 households in the original Pecan Street dataset is over 40 GB. Given that 87 load profiles are required for Network J, and training deep neural networks is memory intensive, downsampling to 90 seconds was necessary.

Household active power load profiles are constructed by substracting grid draw from solar generation. To synthesise reactive power load profiles an average power factor of 0.99 from the ECO dataset was computed. For each interval in the Pecan Street dataset, the 0.99 power factor is perturbed by $\pm 10\%$ and used to compute reactive power.

Only 18 of the households in the Pecan Street data have PV profiles. The size of these systems varies considerably and is often small by Australian standards. In the

interest of consistency the solar generation profiles were normalised to the range [0, 1] and converted to a nominal system size of 6.9 kW. This was the average size of an Australian solar system in 2019 [59].

To produce a final dataset of 87 load and 42 generation profiles a block bootstrap was performed [60]. Bootstrapped profiles were constructed day by day by randomly selecting a given day's data from the existing dataset with replacement.

3.2.3 Reinforcement Learning Problem Formulation

As each rooftop PV installation in a distribution network is a separate system with different owners, the problem of distributed voltage regulation naturally tends towards a multi-agent formulation. Agents must collaborate in real-time with limited local and shared information to moderate voltage whilst limiting active power loss. Multi-agent RL (MARL) problems requiring cooperation with incomplete information are typically formulated as a Decentralised Partially Observable Markov Decision Process (Dec-POMDP) [61].

Dec-POMDPs are mathematically described by a 10-tuple ($\mathcal{I}$, $\mathcal{S}$, $\mathcal{A}$, $\mathcal{R}$, $\mathcal{O}$, $\mathcal{T}$, r, Ω, ρ, γ), where ρ is a probability distribution function for the initial state, and γ is the discount for future rewards. The components of this 10-tuple are defined below:

- $\mathcal{I}$: The set of agents controlling PV inverters. Agents are located at buses within the network. A function $g : \mathcal{I} \to V$ maps agents to the node on the network where they are located
- $\mathcal{R}$: The regions of the network. The network is divided into three regions, one for each phase in the set $\{A, B, C\}$. The function $f : \mathcal{I} \to R$ maps an agent to its region.
- $\mathcal{S}$: The state set is defined as $\mathcal{S} = \mathcal{L} \times \mathcal{P} \times \mathcal{Q} \times \mathcal{V}$ where:
  - $\mathcal{L} = \{(p^L, q^L) : p^L, q^L \in (0, \infty)^{|\mathcal{V}|}\}$ the set of active and reactive power for all loads on the network
  - $\mathcal{P} = \{p^{PV} : p^{PV} \in (0, \infty)^{|\mathcal{I}|}\}$ the set of active power outputs for all PV inverters in the previous step
  - $\mathcal{Q} = \{q^{PV} : q^{PV} \in (0, \infty)^{|\mathcal{I}|}\}$ the set of reactive power outputs for all PV inverters in the previous step
  - $\mathcal{V} = \{(v_a, v_b, v_c) : v_{a,b,c} \in (0, \infty)^{|\mathcal{V}|}\}$ the set of voltages by phase for each node in the network

  The function $h : \mathbb{P}(V) \to \mathbb{P}(S)$ maps a subset of nodes to the state measurements that relate to them, where $\mathbb{P}(\cdot)$ denotes the power set
- $\mathcal{O}$: The observation set is defined as $\mathcal{O} = \times_{i \in \mathcal{I}} \mathcal{O}_i$ where $\mathcal{O}_i = (h \circ f \circ g)(i)$ denotes the measurements of voltage, load active and reactive power, and PV active and reactive power on an agent's phase. In other words, every agent observes the voltage conditions and actions of other agents on their phase.

- $\mathcal{A}$: The action set. Each agent $i \in \mathcal{I}$ has two continuous actions:
  - $\mathcal{A}_{ij} = \{0 \leq a_{ij} \leq 1.0\}$ the ratio of apparent power utilised by the agent to the total amount of apparent power available from generation. This is effectively a curtailment factor, where a value of 0 indicates that all available apparent power is curtailed and 1.0 indicates that all available apparent power is utilised.
  - $\mathcal{A}_{ik} = \{-1.0 \leq a_{ik} \leq 1.0\}$ the ratio of maximum reactive power deployed. Australian regulation specifies a minimum a power factor of 0.8 [4]. Therefore the maximum amount of reactive power available for deployment is given by $q_{max} = 0.2 \times s$. The sign indicates whether the reactive power is injected (negative) or absorbed (positive).
- $\mathcal{T}$: The state transition probability function is defined as $\mathcal{T} = \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$. Specifically, $\mathcal{T}(s_t + 1|s_t, a_t) = Pr(s_t + 1|\delta(s_t, a_t))$, where:
  - $a_t \in \mathcal{A}$
  - $s_t, s_{t+1} \in \mathcal{S}$
  - $\delta(s_t, a) \rightarrow s_{t+\mathcal{T}}$ represents the result of the power flow
  - $Pr(s_{t+1}|s_{t+\mathcal{T}})$ denotes the change in load
- Ω: The observation probability function is given by $\Omega = \mathcal{S} \times \mathcal{A} \times \mathcal{O} \rightarrow [0, 1]$. In regards to electricity networks the observation probability function refers to the measurement errors of sensors such as electric meters. Specifically the observation probability function is defined as $\Omega(o_{t+1}|s_{t+1}, a_t) = s_{t+1} + \mathcal{N}(0, \Sigma)$, where:
  - $\mathcal{N}(0, \Sigma)$ is an isotropic multi-variable normal distribution with variance dependent on the measurement error of the sensors
- r: The reward function is defined as $r = -loss$ where the loss function is given by:

$$loss = \frac{1}{|V|}\frac{1}{|R|}\sum_{i \in \mathcal{V}}\sum_{j \in \mathcal{R}} l_v v_{ij} - \alpha \cdot l_p(s_{ij}^{PV}) \tag{3.1}$$

  where:
  - $l_v(\cdot)$ is a voltage barrier function
  - $l_p(s_i^{PV}) = smax_i - p_i$ the difference between available apparent power and active power exported for inverter i
  - α is a scaling factor used to scale the punishment for wasting active power. In this work this scaling factor is set to 0.2

  In other words, agents are rewarded for regulating voltage across all phases and nodes and are penalised for any power lost in doing so.
- Objective function: For this problem the objective function is $max_\pi \mathbb{E}_\pi \sum_{t=0}^{\infty} \gamma^t r_t$, where:

- The overall policy, π, is given by $\pi = \times_{i \in \mathcal{I}} \pi_i$. That is, the overall policy is derived from the Cartesian product of the individual agents' policies
- Individual agents' policies are defined as the probability function $\pi_i : \bar{\mathcal{O}}_i \times \mathcal{A}_i \to [0, 1]$
- Agent observations are of length h and are given by $\bar{\mathcal{O}}_i = (\bar{\mathcal{O}}_i^{\mathcal{T}})_{\mathcal{T}^h}$

Put simply, the objective is to find an optimal joint policy, π, that maximises discounted cumulative future rewards.

The aim of voltage regulation in an Australian context is to maintain a voltage range of 230V±10%, 0.9-1.1 in the per unit (pu) system. To provide agents with a strong signal to maintain this range, a voltage barrier function is used. For this work the "bowl" voltage barrier function proposed by [21] is used. This function strongly punishes agents when voltage limits are breached without unduly punishing them when conditions are acceptable. The bowl voltage barrier function can be seen in Figure 3.5

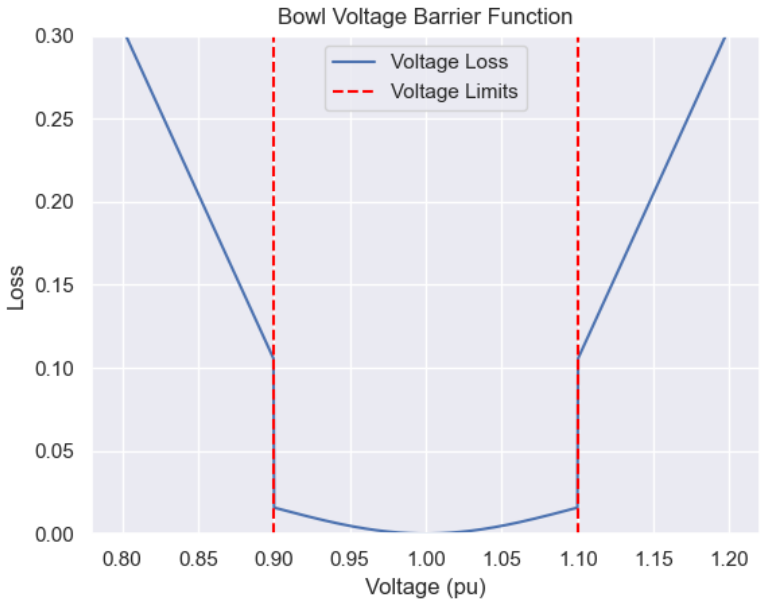

Figure 3.5: The bowl-shaped voltage barrier function used to penalise agents for voltage violations. When voltages fall outside the range 0.9-1.1 pu, the loss function increases steeply.

3.2.4 Simulation Environment

Simulation environments tie together all the elements of an RL problem to facilitate the training and testing of agents. Constructing such a simulation environment for voltage regulation is a major undertaking, requiring significant domain knowledge. As such, it is prudent to extend the work of others where possible to avoid duplicated

effort and mistakes. This project extends the single-phase environment constructed by [21] to support three-phase networks.

This simulation environment uses Pandapower to estimate voltage conditions on the distribution network [62]. Pandapower is a thoroughly tested network solver capable of performing single and three-phase power flow simulations [62].

The RL algorithm used in this project was MADDPG [63]. MADDPG is an actor-critic algorithm with several attractive properties for distributed voltage regulation. Actor-critic algorithms attempt to learn an optimal policy by combining value-function and policy gradient methods in a regime of centralised training and decentralised execution.

During centralised training MADDPG maintains an actor and critic for each agent. The actor receives local measurements and selects an action according to the current policy. The critic uses these actions, the agent's observations, and the actions of the other agents to improve its estimation of the Q function. This Q function in turn produces a temporal difference error which is used to guide weight updates for each actor's policy network. Because the critic sees the actions taken by all agents the environment remains stationary, even as individual agent policies change [63]. In practice a shared replay buffer containing the experiences of all agents is used to improve the critic's Q function during training.

During decentralised execution each agent uses its actor's policy to select actions, and the critic is discarded. Not only does this execution model resemble real-world environments where agents must operate using local measurements, but also has been successfully applied in other works [18, 53, 21].

Table 3.2 shows the architecture of the actor and critic neural networks and Table 3.3 shows the training hyperparameters used in this project.

Network	Architecture
Actor	$Linear(42, 164) \rightarrow LayerNorm() \rightarrow ReLU() \rightarrow GRU(42, 64) \rightarrow Linear(64, 64) \rightarrow Linear(42, 2)$
Critic	$Linear(64, 164) \rightarrow LayerNorm() \rightarrow ReLU() \rightarrow Linear(42, 64) \rightarrow ReLU() \rightarrow Linear(42, 1)$

Table 3.2: The architecture of the actor and critic neural networks. There are 42 agents controlling PV systems in network J, producing 2 continuous actions each

Parameter	Value
Batch Size	32
Replay Buffer Size	5,000
Episode Length	240 intervals (6 hours)
Number of Training Episodes	1000
Optimiser	RMSProp
Policy Learning Rate	1.0e-4
Value Function Learning Rate	1.0e-4
γ	0.99
Parameter Initialisation Distribution	$\mathcal{N}(0, 0.1)$
Behaviour Value Function / Policy Update Frequency	60 intervals
Target Value Function / Policy Update Frequency	120 intervals

Table 3.3: Training hyperparameters for MADDPG

3.3 Experiments

The aim of this work is to assess the performance of RL agents in regulating voltage on three-phase Australian distribution networks. To do so, performance metrics and comparison models must be devised.

3.3.1 Voltage Regulation Strategies

For this work three voltage regulation strategies were compared:

- **Unconditional Export**: An extreme base case in which PV inverters export all of their available apparent power as active power without regard for voltage conditions.
- **Optimal Power Flow (OPF)**: The problem of maximising active power export whilst minimising voltage violations is expressed as linear constraints and solved using an optimal power flow solver.
- **Decentralised RL**: Agents trained using the MADDPG algorithm outlined above are deployed to regulate voltage and maximise active power exports.

The unconditional export strategy provides a lower bound on agent performance. Unconditionally exporting active power will minimise active power loss, at the expense of voltage violations. In the worst case we expect agents to incur fewer voltage violations by significantly curtailing generation.

OPF, meanwhile, provides the upper bound on agent performance. Providing the power flow problem is solvable and numerically well-conditioned, the OPF strategy is guaranteed to find an optimal, or near optimal, solution. In the best case we

expect the RL agents to learn a control strategy similar to OPF which minimises voltage violations whilst maximising active power exports.

The unconditional export strategy is trivial to implement using the Pandapower timeseries module. The OPF strategy is more challenging. Pandapower is capable of performing three-phase power flows, but is unable to solve optimal power flows.

Instead, *PowerModelsDistribution.jl*, an open-source optimal power flow solver was used [64]. Network J's model was loaded and the following constraints were applied:

- Per unit voltages must lie in the range [0.9, 1.1]
- Reactive power cannot account for more than 20% of available apparent power

Initially, the model was run without solar to ensure that the problem had a feasible solution that does not rely on volt-var control. This provides a mathematical guarantee that a solution to the voltage regulation problem exists even when solar is added to the model. In the worst case, the power flow solver and RL agents could completely curtail solar to alleviate voltage challenges. The model was then re-run with solar to derive the optimal reactive and active power set points for PV. These optimal set points were then fed back into Pandapower's power flow solver to eliminate the potential for error due to differences between Pandapower's and *PowerModelsDistribution.jl*'s power flow solvers.

3.3.2 Evaluation Metrics

In the context of residential PV there are two clear goals for a voltage regulation strategy – maintain voltage limits and minimise active power loss. This work measures success in achieving these goals using the following metrics:

- **Uncontrollable Rate**: The number of intervals in which voltages fell outside the acceptable range on any bus and phase
- **Power Loss**: The amount of apparent power not converted to active power by PV inverters

We aim to train agents which minimise the uncontrollable rate and power loss.

CHAPTER 4

Results and Discussion

To appraise the effectiveness of a voltage regulation strategy, understanding the conditions of the network is essential. As such, we will first consider the network conditions faced by and produced by each voltage regulation strategy, and then compare their performance in terms of uncontrollable rate and power loss. Finally, the performance of the RL agents, potential shortcomings in their design, and areas for future work are discussed.

4.1 Baseline Network Conditions

The difficulty of the task faced by agents is the sum of the baseline network conditions and the challenges introduced by PV generation. Therefore, we first consider the voltage and phase conditions of the network without generation.

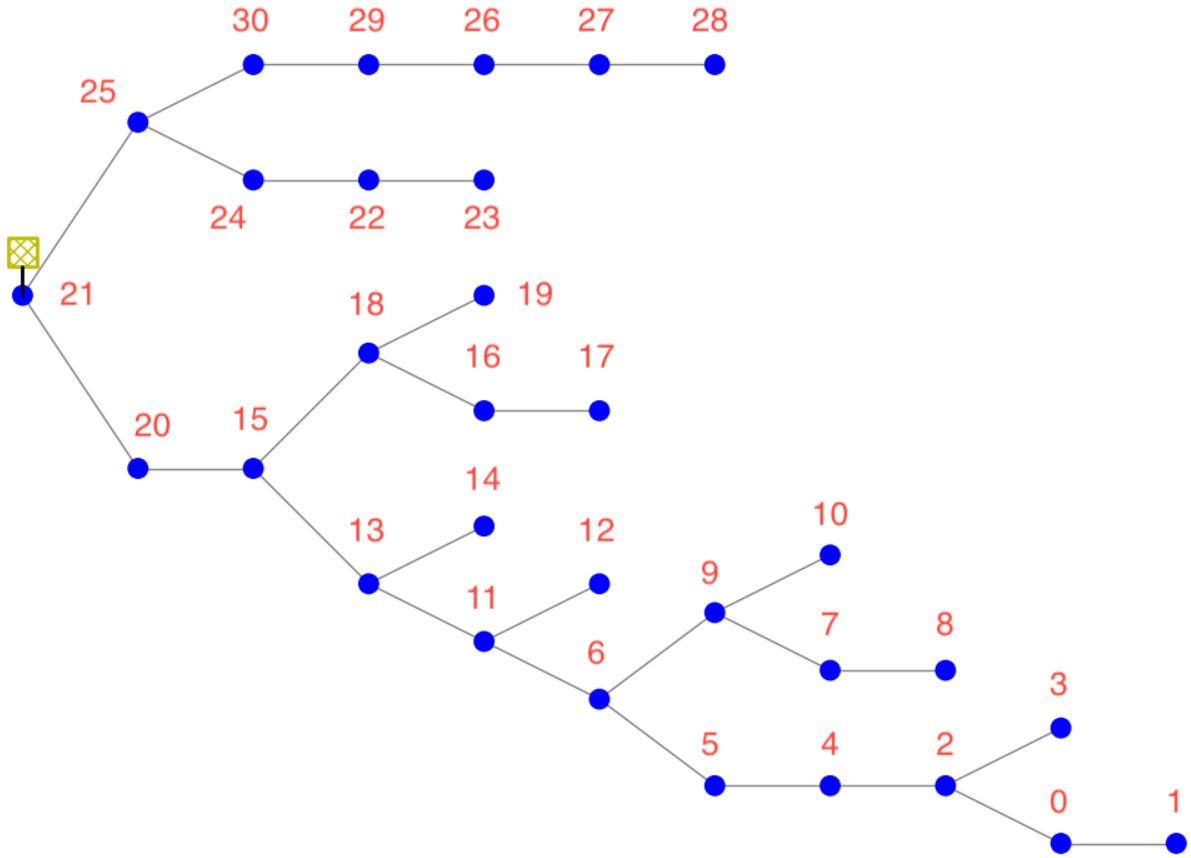

Figure 4.1: Labelled map of the network buses.

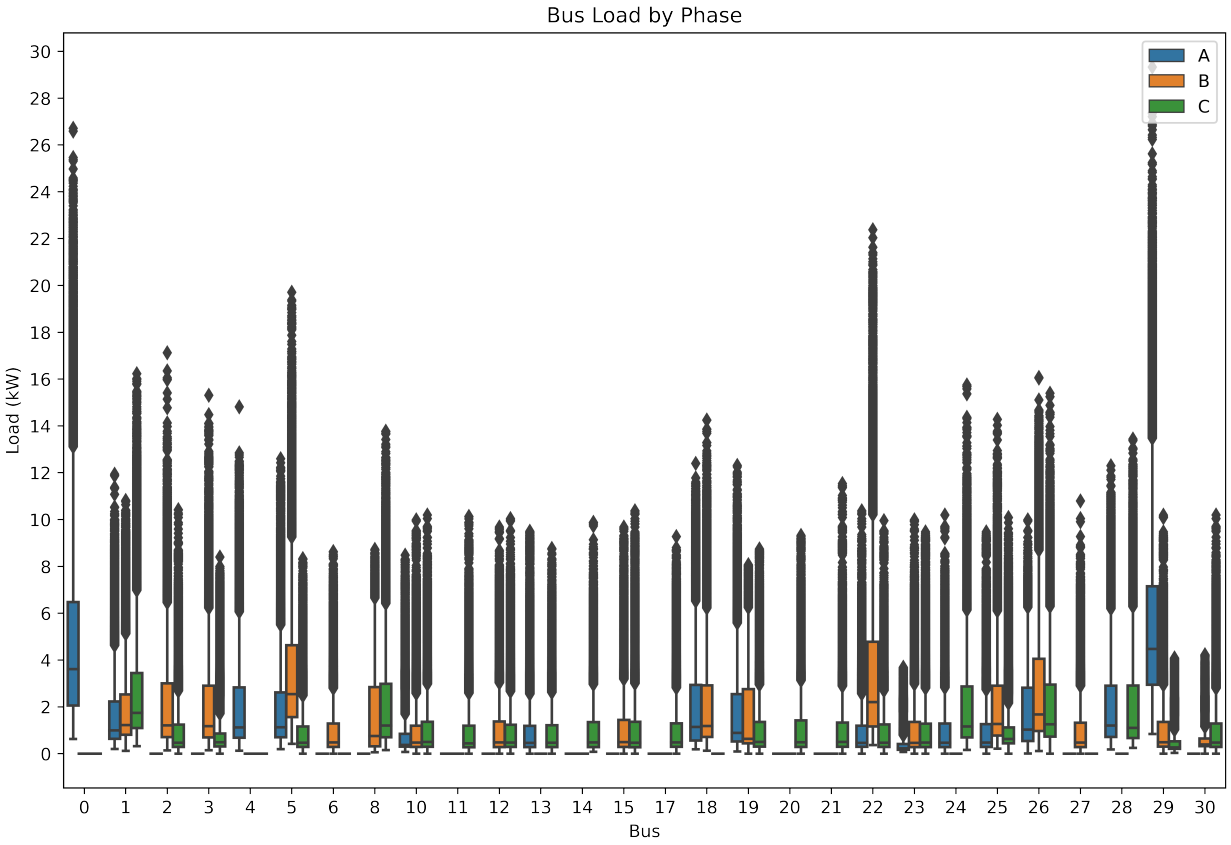

Figure 4.2: Load by bus and phase for Network J. Even without solar generation this network has significant phase imbalance.

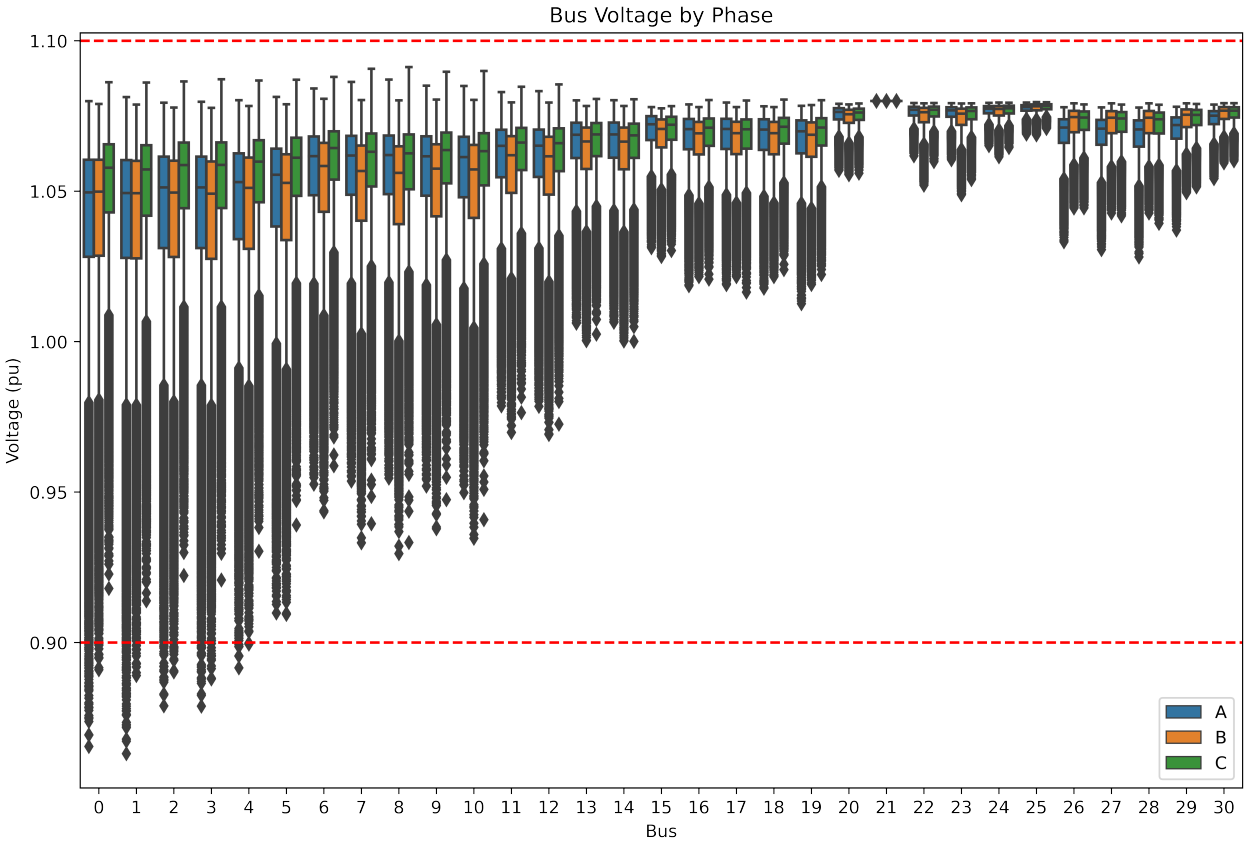

Figure 4.3: Bus voltages by phase in the baseline network. Despite the significant phase imbalance almost no voltage violations occur.

A graph representation of Network J with labelled buses is depicted in Figure 4.1. Though the shape of the graph does not exactly match the original network shape in Figure 3.2, the connections between buses and their relative distance from the source bus, bus 21, are correct. Figure 4.2 shows the load on each bus of the network by phase. Randomly assigning loads to phases has produced significant phase imbalance, with some buses having no load on some phases at all. Buses 7, 9, and 16 have no load whatsoever.

Finally, Figure 4.3 shows bus voltages by phase. The range of voltages grows as distance from the source bus increases, but rarely breaches the statutory voltage limits. Phase imbalance appears to widen the voltage range on some buses, such as buses 0-6 which have light load on phase C. The whiskers of phase C's boxplot for these buses extends well above the level of the source bus, the nominal baseline voltage for the network. This suggests that voltage rise due to phase imbalance is occurring.

An interesting visual feature of these charts is the presence of a significant number of outliers. This is to be expected both for voltages and loads which are typically stable but can fluctuate wildly in extreme conditions. Voltage challenges often arise during extreme load or generation events, and thus this work largely focuses on the ability of voltage regulation strategies to minimise voltage violations that arise due to these conditions. In other words, the ability of a strategy to eliminate outlier

conditions is a rough approximation for its effectiveness.

4.2 Unrestricted Export Network Conditions

The lack of voltage rise, except where it is due to phase imbalance, is logical for the baseline network. There are no power sources in this network, only sinks. Adding solar generators will introduce active power injection, more phase imbalance, and more voltage challenges.

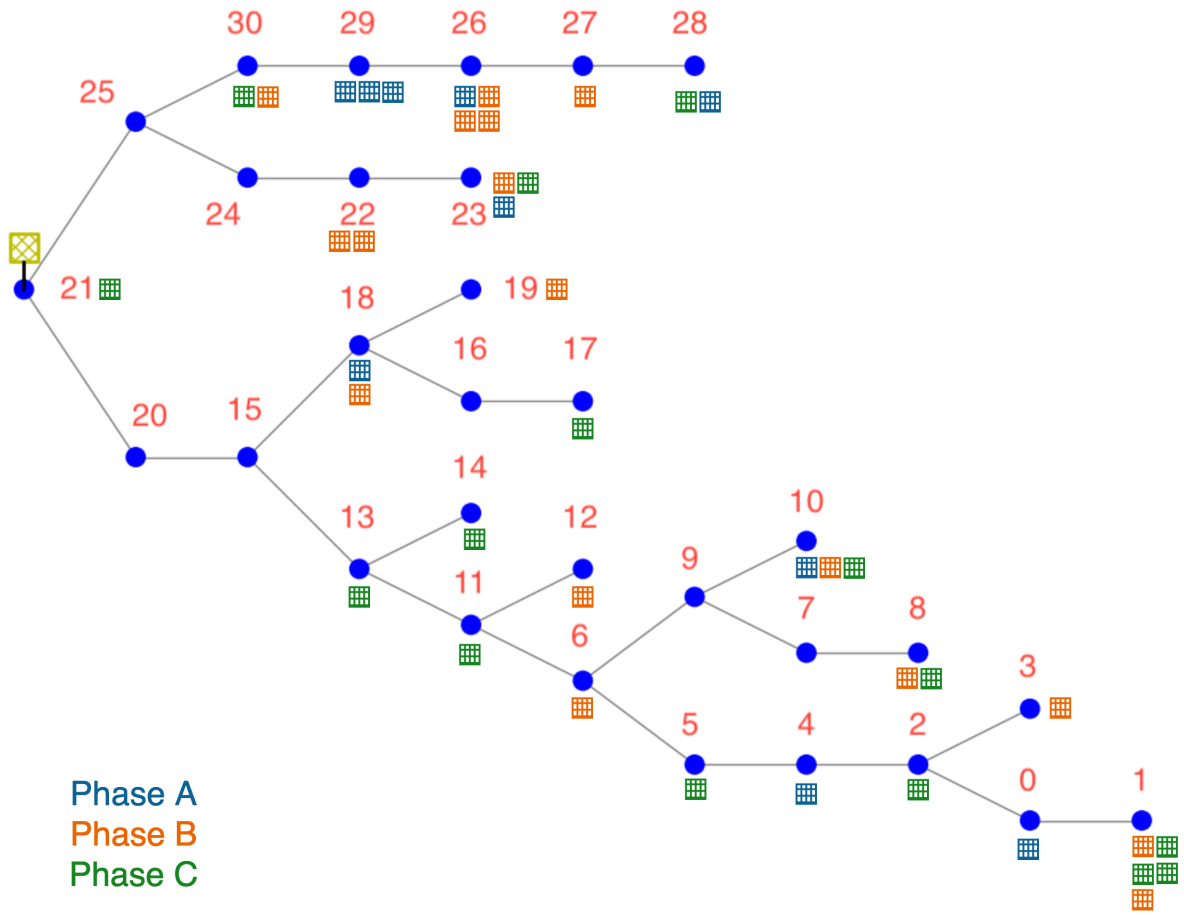

Figure 4.4: Topology of Network J with solar generation randomly assigned to loads.

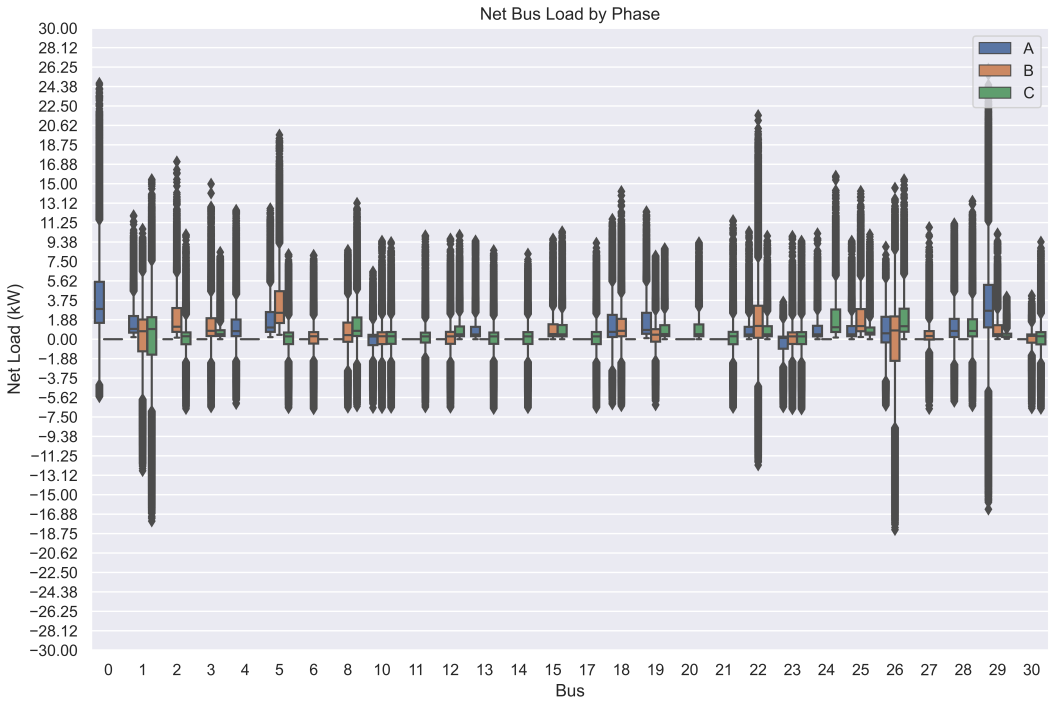

Figure 4.5: Load by bus and phase net of solar generation.

The distribution of solar generation across the buses of Network J is shown in Figure 4.4. Once again, the random assignment of generators to loads has produced an unusual distribution of generation across the network.

The effect of this unusual distribution, and of the addition of solar generation on net bus loads per phase can be seen in Figure 4.5. Solar generation narrows the average level of phase imbalance, but widens the total range of load on buses. For example, the inter-quartile range for load on phase B of bus 22 has fallen from approximately 6 kW to 3 kW, but its total range has increased from 24 kW to 34 kW. In other words, the addition of solar generation has caused average load to fall, but outlier conditions have become more extreme.

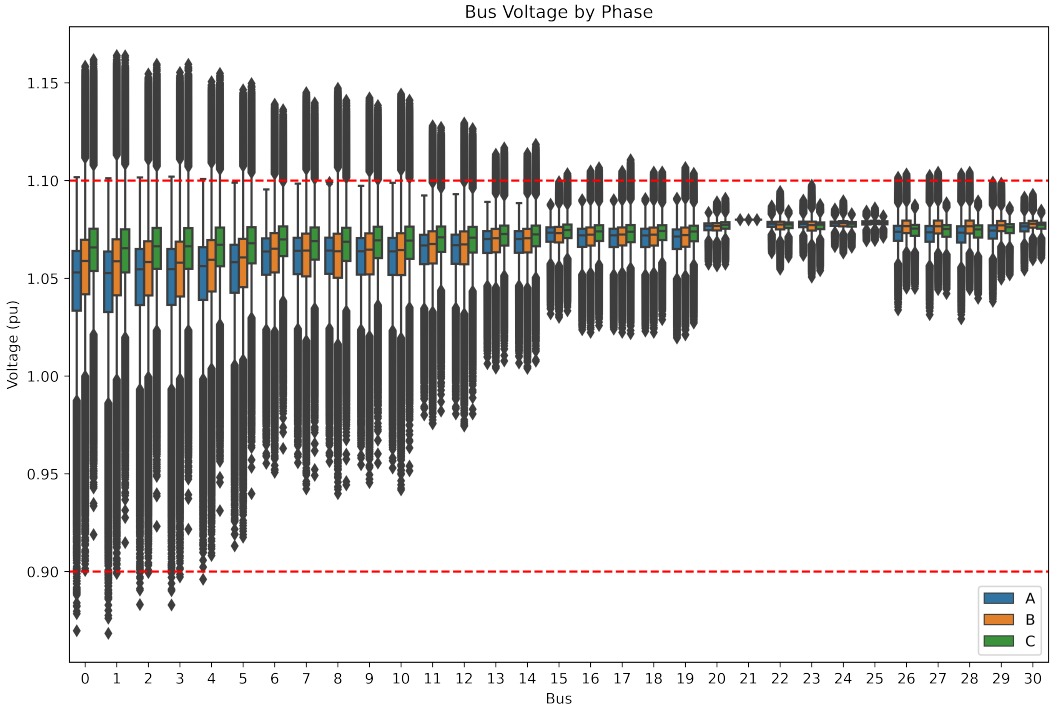

Figure 4.6: Voltage by bus and phase with unrestricted solar export.

Voltage conditions have also become more extreme. Figure 4.6 shows the impact of solar generation on bus voltages, whose range and maximum values have increased across the network. The whiskers for almost all buses extend above the statutory voltage limit, even for buses which formerly lay comfortably within voltage limits.

The addition of generation also exacerbates the phase imbalance on buses 0-6. The load gap between phase C and phases A and B has widened, which combined with active power injection from solar generation induces further voltage rise on phase C and more over-voltage events.

4.3 Optimal Power Flow Network Conditions

OPF, in stark contrast to unrestricted export, is a targeted and highly effective voltage control strategy. If the power flow problem is numerically well-conditioned and an optimal solution exists, OPF is guaranteed to find it. For Network J the OPF solver found an optimal, or near optimal, solution in 98.6% of iterations.

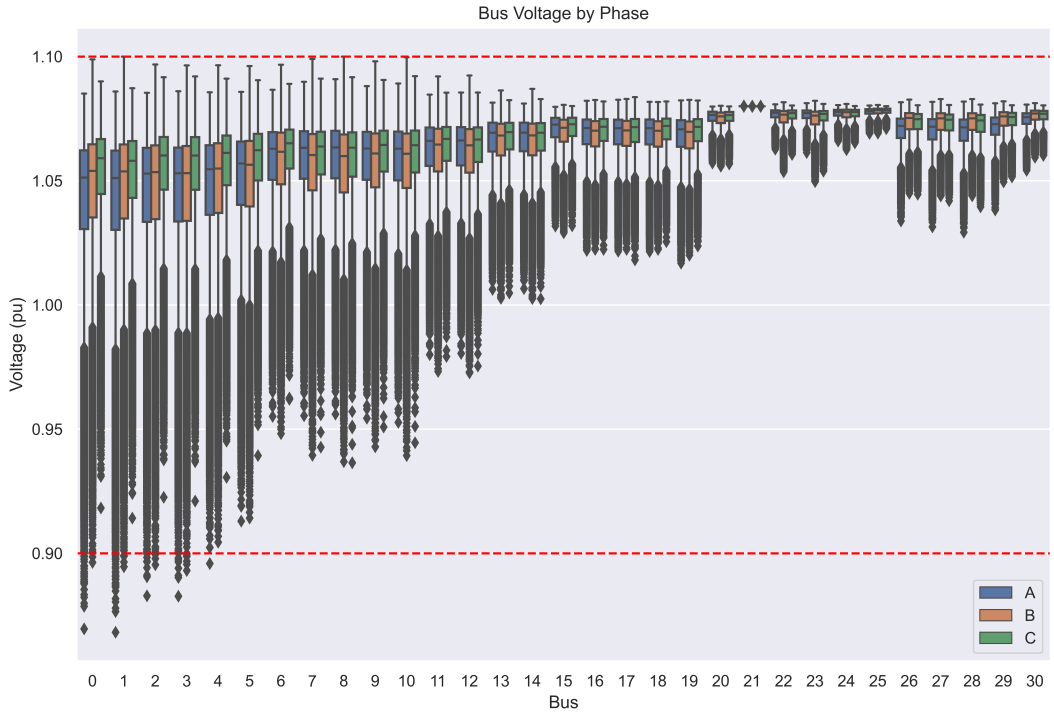

Figure 4.7: Voltage by bus and phase with Optimal Power Flow.

From the voltage by phase graph in Figure 4.7 alone, it is clear that OPF has performed well on this network. There are almost no voltage violations and phase voltages are reasonably balanced. From Figure 4.8 it is clear that voltage is controllable without significant curtailment. Net bus load is not drastically different from the unrestricted export case – the whiskers have simply been trimmed to prevent voltage excursions.

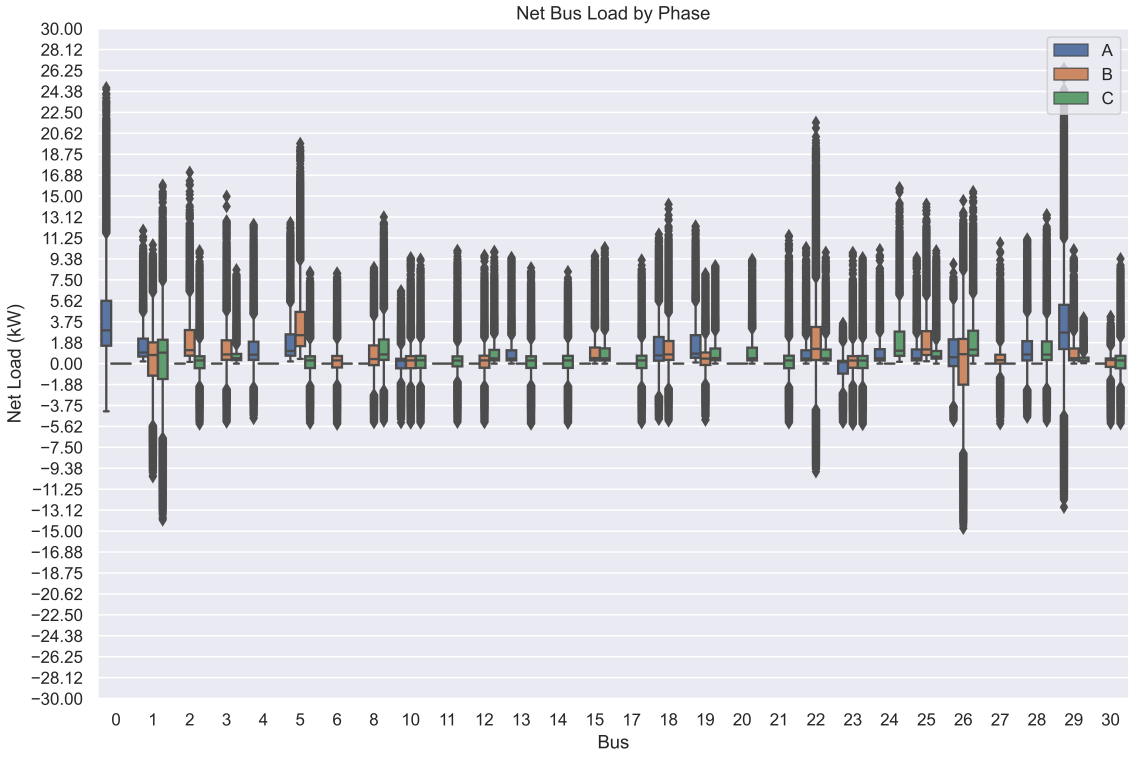

Figure 4.8: Load by bus and phase net of solar generation for the OPF strategy.

4.4 Reinforcement Learning Agents Network Conditions

Clearly, an optimal strategy which maintains voltage limits whilst minimising the curtailment of active power exists. To what extent have the RL agents learned that strategy?

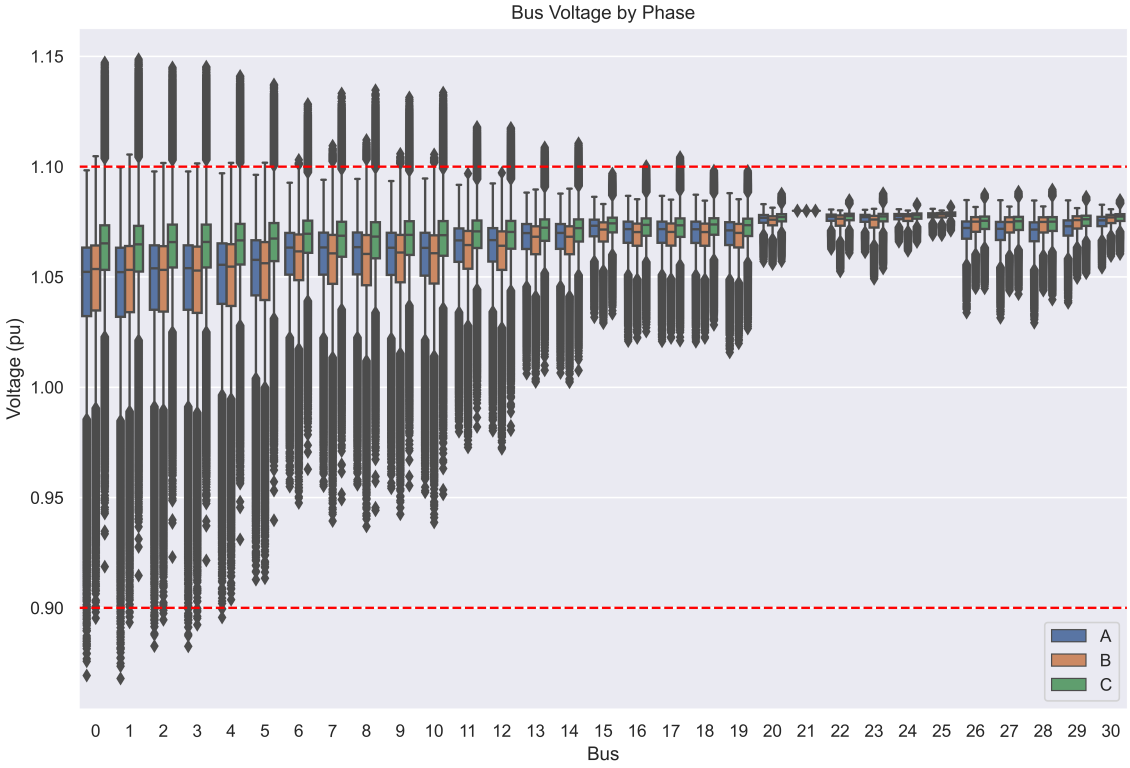

Figure 4.9: Voltage by bus and phase with RL agents controlling PV inverters.

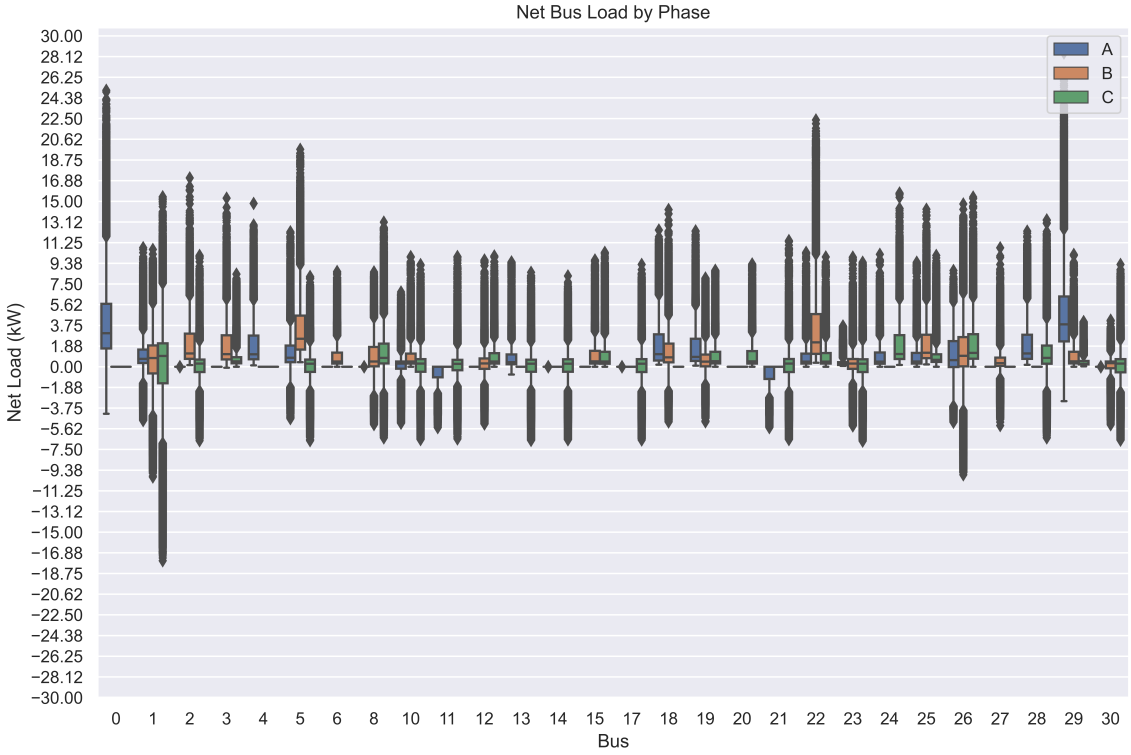

Figure 4.10: Load by bus and phase net of solar generation with RL agents controlling PV inverters.

The plot of voltage by phase for the RL agents in Figure 4.9 shows that the agents have achieved some degree of voltage regulation. There are almost no voltage violations on phase B, and the number of buses experiencing voltage challenges on phase C is reduced.

Comparing net bus load for the RL agents in Figure 4.10 with that of OPF in Figure 4.8 we see that that generation on buses 21-29 has been significantly curtailed. The box plots for net load on phase A of bus 29 and phase B of bus 26 have shifted upwards considerably relative to the OPF strategy. Under the RL agents' strategy large negative net load is less frequent than under the OPF control strategy for these buses.

Phases A and B on buses 3, 4 and 6 have been similarly curtailed relative to the OPF model. Given that phase C is lightly loaded in this part of the network, increasing load on these phases by curtailing generation undoubtedly increases phase imbalance and induces further voltage rise on phase C. It is likely that the near-elimination of voltage violations on phases A and B in this part of the network is achieved at the expense of uncontrollable voltage rise on phase C.

4.5 Performance Comparison

Examining the network conditions for each voltage regulation strategy in isolation we can qualitatively conclude that OPF and the RL agents are able to regulate

voltage. To quantitatively compare *how* well they regulated voltage we compare their performance over time using uncontrollable rate and power loss as metrics.

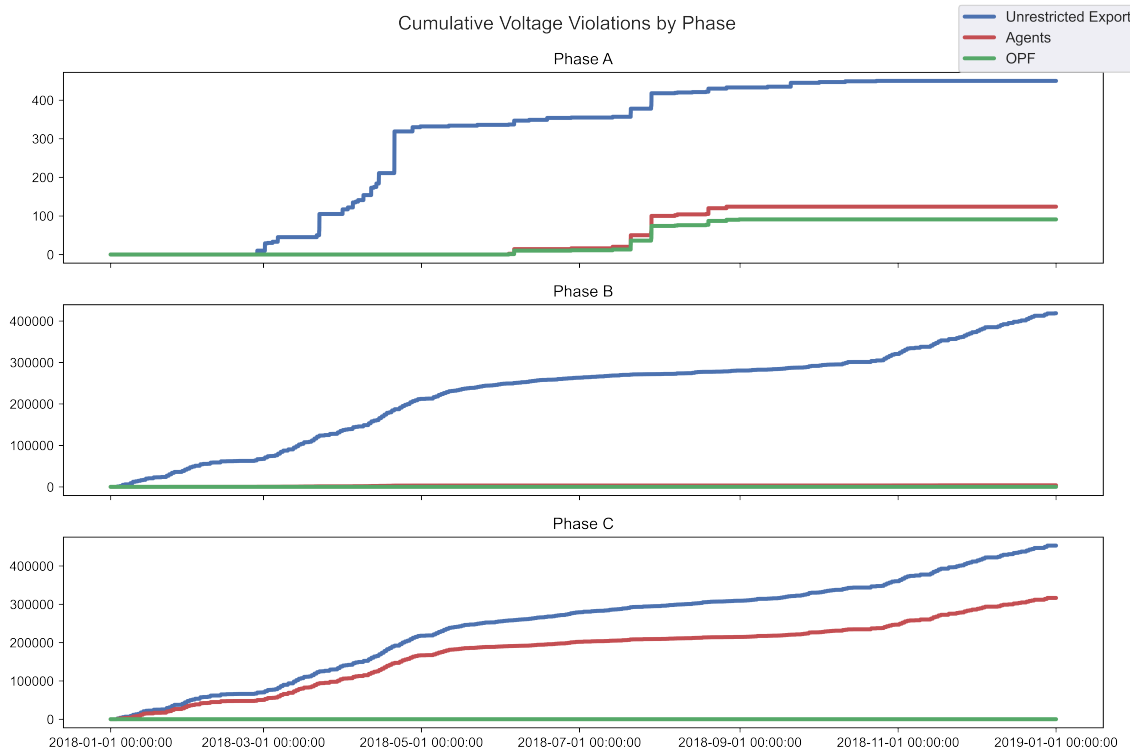

Figure 4.11: Cumulative voltage violations by phase for each voltage regulation strategy.

Cumulative voltage violations over the course of the entire dataset are shown in Figure 4.11. The greater the number of voltage violations, the greater the uncontrollable rate, and the poorer the voltage regulation strategy. Unsurprisingly, the unrestricted export strategy produces the most voltage violations across each phase. On phases A and B the RL agents achieve near identical performance to the OPF strategy, with almost no voltage violations on either phase. But, this came at the expense of significant voltage violations on phase C, where the performance of the OPF and RL agents diverged sharply. Despite this difference in performance, the RL agents achieved a 25% reduction in voltage violations relative to the unrestricted export strategy.

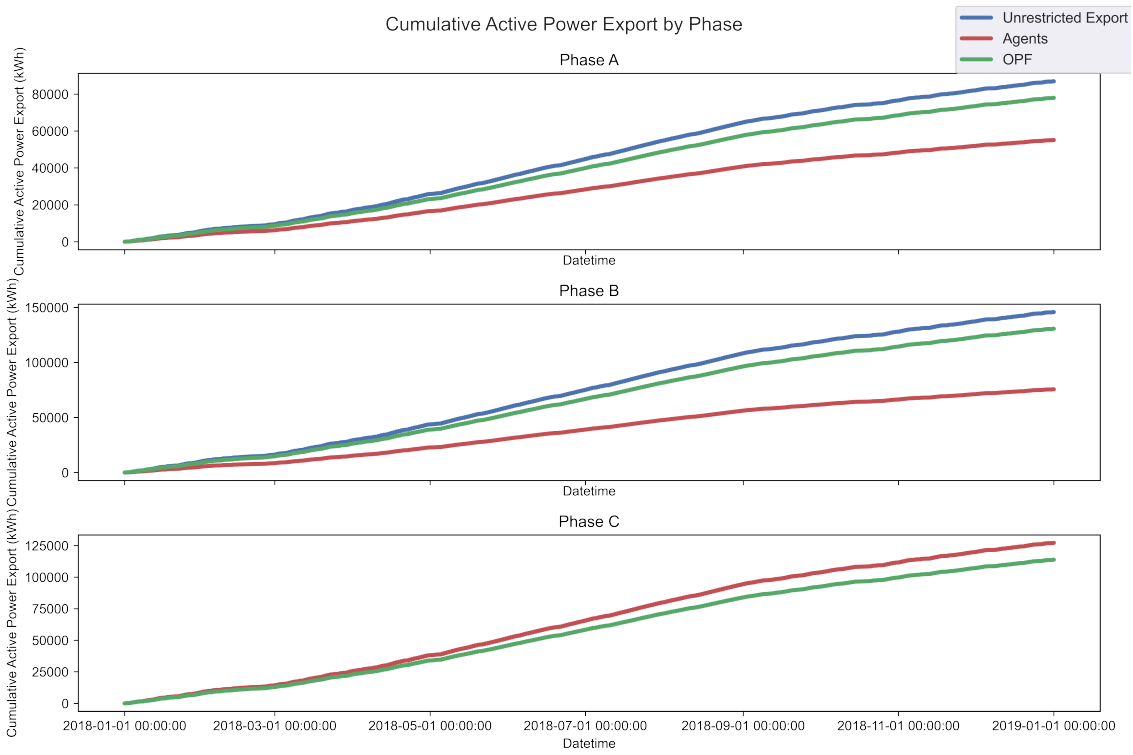

Figure 4.12: Cumulative active power export by phase for each voltage regulation strategy.

However, maintaining statutory limits is only one half of an effective voltage regulation strategy. Voltage limits must be maintained whilst minimising active power loss. Figure 4.12 shows the cumulative active power export for each approach over time. The greater the active power export, the smaller the active power loss for the voltage regulation strategy.

The unrestricted export strategy trivially exports the most active power on all phases. For phases A and B OPF significantly outperforms the RL agents, exporting nearly twice as much active power. Meanwhile, on phase C OPF curtails approximately 10% of active power, whilst the RL agents export all available power as active power.

Though the RL agents successfully regulated voltage on phases A and B, they did so by significantly curtailing generation. On phase C they achieved some success without any curtailment, but drastically under-performed compared to the OPF strategy which eliminated voltage violations with minimal active power loss.

Examining intervals where the RL agents failed to regulate voltage illustrates the shortcomings of their learned policy. Figure 4.13 shows the active power exports of each strategy by phase from midday to sunset on the 30th of January. This interval reflects the trend seen in the aggregate performance data – OPF slightly curtails generation on all phases, whilst the RL agents strongly curtail phases A and B whilst exporting all power on phase C. Looking at the voltage by phase and regulation strategy for this interval in Figure 4.14, the overall voltage trend is also reflected. OPF successfully controls voltages on all phases, and the RL agents

control voltages on phases A and B, but suffer voltage violations on phase C.

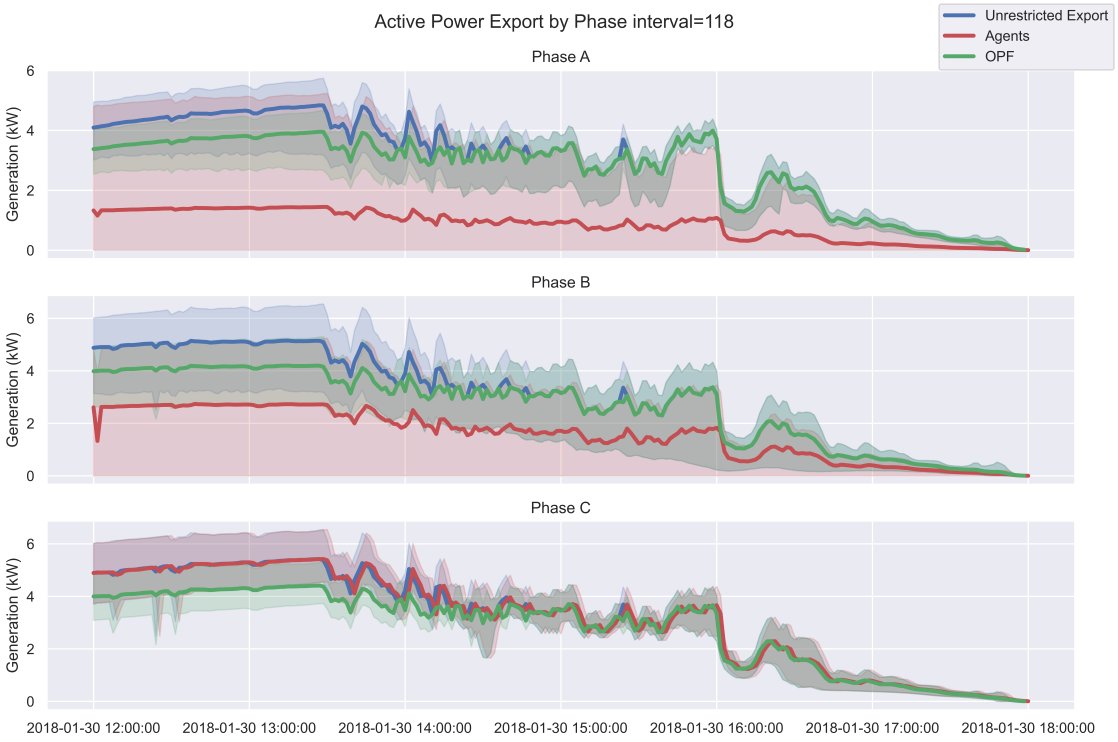

Figure 4.13: Average active power export by phase for each voltage regulation strategy from midday to sunset on the 30th of January 2018. The shaded area above and below the line spans the maximum and minimum active power export respectively.

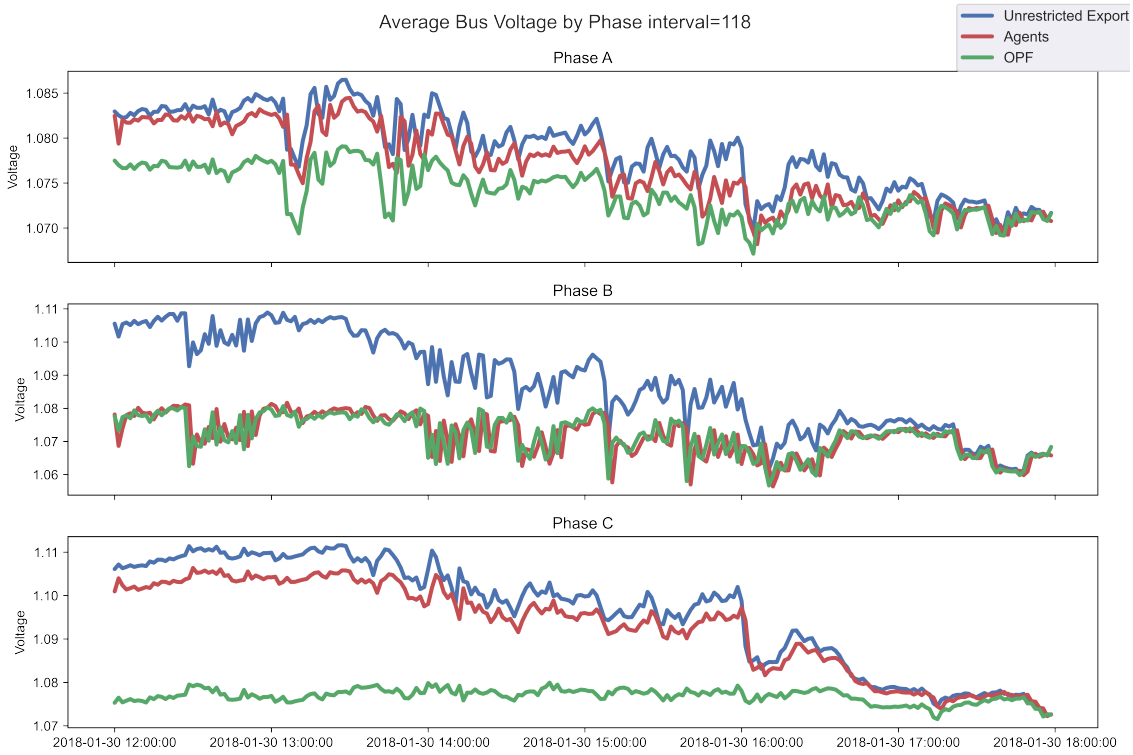

Figure 4.14: Average voltage by phase for each voltage regulation strategy from midday to sunset on the 30th of January 2018.

From these results it appears as though the RL agents' policies became trapped in a local optima. Strongly curtailing phases A and B eliminates voltage violations on those phases but creates a large phase imbalance between phases A and B and phase C. This phase imbalance worsens the voltage violations on phase C which cannot be resolved unilaterally by curtailing generation on phase C. Therefore, agents on phase C accept voltage violations as a fact of life and attempt to maximise their reward by eliminating active power loss.

The optimum strategy employed by OPF is obvious when Figures 4.13 and 4.14 are considered, but is counter-intuitive for agents to learn. To limit phase imbalance and minimise voltage violations agents must *increase* their active power export on phases A and B, whilst slightly curtailing export on phase C.

Clearly, the RL agents have failed to learn about phase imbalance and its implications for voltage regulation.

There are several avenues for improving RL agent performance that warrant further exploration:

- **Increased training**: During training the loss function consistently levelled off after approximately 1000 episodes. Thus this value was chosen as the cut-off point for training. However, it is possible that with further training episodes the agents may have learned a more nuanced policy. Though it is unclear how much more training would be required.

- **More complicated critic network architecture**: The critic network for this project was a feed forward neural network with two hidden layers. Given the complexity of the relationship between phase, load and voltage this network may not be able to accurately approximate the Q function. Experimenting with more advanced critic network architectures may yield better results as agents' policies are guided by better critics.
- **Different reward function**: The averaging effect of the reward function in this work likely played a role in the mediocre performance of agents. In effect, agent policies were "too greedy" in their pursuit of a global optima. Agents were guided as much by the state of distant buses as those that were nearby, but their actions only directly affected local buses. With the benefit of hindsight it would have been better to scale agent rewards on the basis of bus distance. Agents should be punished more strongly if their actions contribute to local voltage instability.

CHAPTER 5

Conclusion & Future Work

The results of this work are promising. Though far from perfect, the RL agents trained in this project were able to learn a voltage regulation strategy in a three-phase network which outperformed the base case. This suggests that this problem is both tractable and deserving of continued attention. If efforts to create a decarbonised, cost-effective and stable electricity grid are to succeed the fair participation of distributed assets such as PV, batteries, and electric vehicles must be guaranteed. Without swift action to address the emerging challenge of voltage instability in low voltage distribution networks, the economic and environmental potential of these assets will be significantly curtailed. Suffice to say, the solution to this problem is unclear; the necessity of finding one is not.

A consideration of the contribution of this work is incomplete without describing the challenges faced in its inception. Voltage regulation using RL is a new and exciting field that lies at the intersection of several challenging research areas. As a result, standard tools and approaches are yet to emerge and become widely disseminated. A large component of this work was discovering, appraising, and 'gluing together' the appropriate mix of legacy and modern tools to create a robust simulation environment. This effort was made all the more difficult for the lack of published implementations from other works, the use of proprietary datasets, and/or the vague description of methodologies (often by the same works that did not publish their code). As a community, this is an area where we must improve.

In addition to this lack of standard "tech stacks" for projects of this nature, data quality was also an issue. Acquiring good quality load and generation data was easier than anticipated, the same cannot be said for network models. Despite the best efforts of the CSIRO team who carried out the Low Voltage Feeder Taxonomy, the network models produced were challenging to use. Many networks had negative loads and bizarre network configurations which made constructing OPF models difficult.

In light of these difficulties the contributions of this project are as follows:

- Open source implementation of RL-driven voltage regulation using datasets freely available to researchers
- Use of real Australian three-phase distribution network models
- Development of agents that perform voltage regulation using standard household inverters without leveraging existing network infrastructure

Given these challenges and the contributions of this work, several avenues for future work present themselves:

- **Comprehensively review LV network modelling tools**: Much of the available literature in this area uses tools which are no longer state of the art and are unwieldy compared to modern tools. A review of available tools, and a consideration of their trade-offs and capabilities would reduce the barrier to entry for newcomers in this area.
- **Making more network models available**: Agents are only as good as the data used to train them. Acquiring, cleaning and open sourcing network models from operators would provide a greater variety of training environments for agents.
- **Add batteries into the mix**: Stationary home batteries and EV significantly alter household electricity usage patterns, and provide an additional degree of freedom for agents. Given the proliferation of these assets, exploring their effect on the behaviour and performance of agents is an interesting extension of this work.
- **Explore different reward functions and observations**: Different reward and observation structures will undoubtedly change the behaviour of agents, and may guide them to a policy which takes account of phase dynamics.

This work, much like the agents it produced, has reached a local optima. Unlike the agents, I am foolish enough to believe that a global optima is somewhere over the next hill.

Bibliography

[1] Department for Energy and Mining, "SA's electricity supply and market," https://www.sa.gov.au/topics/energy-and-environment/energy-supply/sas-electricity-supply-and-market, 2021, [Online; accessed 12-September-2021].

[2] X. Su, J. Liu, S. Tian, P. Ling, Y. Fu, S. Wei, and C. SiMa, "A multi-stage coordinated volt-var optimization for integrated and unbalanced radial distribution networks," *Energies*, vol. 13, no. 18, p. 4877, 2020.

[3] N. Stringer, N. Haghdadi, A. Bruce, and I. MacGill, "Fair consumer outcomes in the balance: Data driven analysis of distributed PV curtailment," *Renewable Energy*, vol. 173, pp. 972–986, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960148121005322

[4] "Grid Connection of Energy Systems via Inverters," Standards Australia/Standards New Zealand, Standard, Dec. 2020.

[5] F. Geth, T. Brinsmead, S. West, P. Goldthorpe, B. Spak, G. Cross, and J. Braslavsky, "National low-voltage feeder taxonomy study," *CSIRO Research Publications Repository*, 2021.

[6] IEA, "Renewables 2020," 2020. [Online]. Available: https://www.iea.org/reports/renewables-2020

[7] Australian Energy Council, "Solar report. 2019," 2021. [Online]. Available: https://www.energycouncil.com.au/media/15358/australian-energy-council-solar-report_$january - 2019.pdf$

[8] M. Viyathukattuva Mohamed Ali, M. Babar, P. Nguyen, and J. Cobben, "Overlaying control mechanism for solar PV inverters in the LV distribution network," *Electric Power Systems Research*, vol. 145, pp. 264–274, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378779616305156

[9] AEMO, "Integrated system plan for the national electricity market," 2018.

[10] Dept. Industry, Science, Energy and Resources. (2021) Australia achieves 3 million rooftop solar PV installations. [Online]. Available: https://www.energy.gov.au/news-media/news/australia-achieves-3-million-rooftop-solar-pv-installations

[11] AEMO, "AS/NZS 4777.2 inverter requirements standard," 2020. [Online]. Available: https://aemo.com.au/en/initiatives/major-programs/nem-distributed-energy-resources-der-program/standards-and-connections/as-nzs-4777-2-inverter-requirements-standard

[12] S. Heslop, N. Stringer, B. Yaldiz, A. Bruce, P. Heywood, I. MacGill, and R. Passey, "Voltage analysis of the LV distribution network in the Australian National Electricity Market," 2020.

[13] P. N. Vovos, A. E. Kiprakis, A. R. Wallace, and G. P. Harrison, "Centralized and distributed voltage control: Impact on distributed generation penetration," *IEEE Transactions on power systems*, vol. 22, no. 1, pp. 476–483, 2007.

[14] G. C. Kryonidis, C. S. Demoulias, and G. K. Papagiannis, "A nearly decentralized voltage regulation algorithm for loss minimization in radial mv networks with high dg penetration," *IEEE Transactions on Sustainable Energy*, vol. 7, no. 4, pp. 1430–1439, 2016.

[15] M. Yazdanian and A. Mehrizi-Sani, "Distributed control techniques in microgrids," *IEEE Transactions on Smart Grid*, vol. 5, no. 6, pp. 2901–2909, 2014.

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

[17] X. Sun and J. Qiu, "Two-stage volt/var control in active distribution networks with multi-agent deep reinforcement learning method," *IEEE Transactions on Smart Grid*, 2021.

[18] D. Cao, W. Hu, J. Zhao, Q. Huang, Z. Chen, and F. Blaabjerg, "A multi-agent deep reinforcement learning based voltage regulation using coordinated pv inverters," *IEEE Transactions on Power Systems*, vol. 35, no. 5, pp. 4120–4123, 2020.

[19] K. Beyer, R. Beckmann, S. Geißendörfer, K. v. Maydell, and C. Agert, "Adaptive online-learning volt-var control for smart inverters using deep reinforcement learning," *Energies*, vol. 14, no. 7, p. 1991, 2021.

[20] Y. Zhang, X. Wang, J. Wang, and Y. Zhang, "Deep reinforcement learning based volt-var optimization in smart distribution systems," *IEEE Transactions on Smart Grid*, vol. 12, no. 1, pp. 361–371, 2020.

[21] J. Wang, W. Xu, Y. Gu, W. Song, and T. Green, "Multi-agent reinforcement learning for active voltage control on power distribution networks," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[22] Federal Energy Regulatory Commission *et al.*, "Energy primer: A handbook of energy market basics," *Federal Energy Regulatory Commission: Washington, DC, USA*, 2020.

[23] R. P. Sedano and M. H. Brown, "Electricity transmission: a primer," *National Council on Electricity Policy, June, Washington, DC*, 2004.

[24] N. Naval and J. M. Yusta, "Virtual power plant models and electricity markets - a review," *Renewable and Sustainable Energy Reviews*, vol. 149, p. 111393, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S136403212100678X

[25] AEMO, "The national electricity market: Fact sheet," https://www.aemo.com.au/-/media/Files/Electricity/NEM/National-Electricity-Market-Fact-Sheet.pdf, 2020, accessed: 2021-10-10.

[26] F. P. Sioshansi, *Variable Generation, Flexible Demand*. Academic Press, 2020.

[27] Australian Institute of PV, "Australian PV market since April 2001," *Australian PV installations since April*, 2001, https://pv-map.apvi.org.au/analyses.

[28] T. Ajadi, V. Cuming, R. Boyle, D. Strahan, M. Kimmel, M. Logan, A. McCrone *et al.*, "Global trends in renewable energy investment 2020," 2020.

[29] "Electromagnetic compatibility (EMC)," Standards Australia/Standards New Zealand, Standard, Mar. 2011.

[30] Australian Energy Market Operator, "Integrating Distributed Energy Resources For The Grid Of The Future," 2019.

[31] AEMO, "National Energy Retail Amendment (Technical Standards For Distributed Energy Resources) Rule 2021," 2021.

[32] A. Kulmala, S. Repo, and P. Järventausta, "Coordinated voltage control in distribution networks including several distributed energy resources," *IEEE Transactions on Smart Grid*, vol. 5, no. 4, pp. 2010–2020, 2014.

[33] K. E. Antoniadou-Plytaria, I. N. Kouveliotis-Lysikatos, P. S. Georgilakis, and N. D. Hatziargyriou, "Distributed and decentralized voltage control of smart distribution networks: Models, methods, and future research," *IEEE Transactions on smart grid*, vol. 8, no. 6, pp. 2999–3008, 2017.

[34] B. Zhang, A. Y. Lam, A. D. Domínguez-García, and D. Tse, "An optimal and distributed method for voltage regulation in power distribution systems," *IEEE Transactions on Power Systems*, vol. 30, no. 4, pp. 1714–1726, 2014.

[35] S. Bolognani, R. Carli, G. Cavraro, and S. Zampieri, "Distributed reactive power feedback control for voltage regulation and loss minimization," *IEEE Transactions on Automatic Control*, vol. 60, no. 4, pp. 966–981, 2014.

[36] K. Utkarsh, A. Trivedi, D. Srinivasan, and T. Reindl, "A consensus-based distributed computational intelligence technique for real-time optimal control in smart distribution grids," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 1, no. 1, pp. 51–60, 2016.

[37] P. Kou, D. Liang, R. Gao, Y. Liu, and L. Gao, "Decentralized model predictive control of hybrid distribution transformers for voltage regulation in active distribution networks," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 4, pp. 2189–2200, 2019.

[38] Y. Guo, Q. Wu, H. Gao, and F. Shen, "Distributed voltage regulation of smart distribution networks: Consensus-based information synchronization and distributed model predictive control scheme," *International Journal of Electrical Power Energy Systems*, vol. 111, pp. 58–65, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0142061518330977

[39] K. Meng, Z. Y. Dong, Z. Xu, and S. R. Weller, "Cooperation-driven distributed model predictive control for energy storage systems," *IEEE Transactions on Smart Grid*, vol. 6, no. 6, pp. 2583–2585, 2015.

[40] S. D. McArthur, E. M. Davidson, V. M. Catterson, A. L. Dimeas, N. D. Hatziargyriou, F. Ponci, and T. Funabashi, "Multi-agent systems for power engineering applications—part i: Concepts, approaches, and technical challenges," *IEEE Transactions on Power systems*, vol. 22, no. 4, pp. 1743–1752, 2007.

[41] X. Yu, A. M. Khambadkone, H. Wang, and S. T. S. Terence, "Control of parallel-connected power converters for low-voltage microgrid—part i: A hybrid control architecture," *IEEE Transactions on Power Electronics*, vol. 25, no. 12, pp. 2962–2970, 2010.

[42] J. Rocabert, A. Luna, F. Blaabjerg, and P. Rodriguez, "Control of power converters in ac microgrids," *IEEE transactions on power electronics*, vol. 27, no. 11, pp. 4734–4749, 2012.

[43] M. V. M. Ali, P. Nguyen, W. Kling, A. Chrysochos, T. Papadopoulos, and G. Papagiannis, "Fair power curtailment of distributed renewable energy sources to mitigate overvoltages in low-voltage networks," in *2015 IEEE Eindhoven PowerTech*. IEEE, 2015, pp. 1–5.

[44] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[45] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.

[46] A. Perera and P. Kamalaruban, "Applications of reinforcement learning in energy systems," *Renewable and Sustainable Energy Reviews*, vol. 137, p. 110618, 2021.

[47] K. Mason and S. Grijalva, "A review of reinforcement learning for autonomous building energy management," *Computers Electrical Engineering*, vol. 78, pp. 300–312, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045790618333421

[48] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke *et al.*, "Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation," *arXiv preprint arXiv:1806.10293*, 2018.

[49] M. Botvinick, S. Ritter, J. X. Wang, Z. Kurth-Nelson, C. Blundell, and D. Hassabis, "Reinforcement learning, fast and slow," *Trends in cognitive sciences*, vol. 23, no. 5, pp. 408–422, 2019.

[50] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[51] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of Reinforcement Learning and Control*, pp. 321–384, 2021.

[52] T.-H. Fan, X. Y. Lee, and Y. Wang, "Powergym: A reinforcement learning environment for volt-var control in power distribution systems," *arXiv preprint arXiv:2109.03970*, 2021.

[53] X. Wang, C. Wang, T. Xu, L. Guo, S. Fan, and Z. Wei, "Decentralised voltage control with built-in incentives for participants in distribution networks," *IET Generation, Transmission & Distribution*, vol. 12, no. 3, pp. 790–797, 2018.

[54] K. P. Schneider, B. Mather, B. Pal, C.-W. Ten, G. J. Shirek, H. Zhu, J. C. Fuller, J. L. R. Pereira, L. F. Ochoa, L. R. de Araujo *et al.*, "Analytic considerations and design basis for the ieee distribution test feeders," *IEEE Transactions on power systems*, vol. 33, no. 3, pp. 3181–3188, 2017.

[55] Electricity North West, "Low voltage network solutions report," 2014. [Online]. Available: https://www.enwl.co.uk/go-net-zero/innovation/smaller-projects/low-carbon-networks-fund/low-voltage-network-solutions/low-voltage-network-solutions-closedown-report/

[56] M. R. Asghar, G. Dán, D. Miorandi, and I. Chlamtac, "Smart meter data privacy: A survey," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2820–2835, 2017.

[57] S. Kapoor, B. Sturmberg, and M. Shaw, "A review of publicly available energy data sets," *Wattwatchers' My Energy Marketplace (MEM)(The Australian National University, Canberra, Australia, 2020)*, 2020.

[58] J. Estima, N. Fichaux, L. Menard, and H. Ghedira, "The global solar and wind atlas: a unique global spatial data infrastructure for all renewable energy," in *Proceedings of the 1st ACM SIGSPATIAL International Workshop on mapinteraction*, ser. MapInteract '13. ACM, 2013, pp. 36–39.

[59] Clean Energy Regulator, "Australians install two million solar pv systems," *Canberra, Australian Government*, 2018.

[60] G. Hongyi Li and Maddala, "Bootstrapping time series models," *Econometric reviews*, vol. 15, no. 2, pp. 115–158, 1996.

[61] F. A. Oliehoek and C. Amato, *A concise introduction to decentralized POMDPs*. Springer, 2016.

[62] L. Thurner, A. Scheidler, F. Schäfer, J.-H. Menke, J. Dollichon, F. Meier, S. Meinecke, and M. Braun. (2017) pandapower - an Open Source Python Tool for Convenient Modeling, Analysis and Optimization of Electric Power Systems. Preprint. [Online]. Available: https://arxiv.org/abs/1709.06743

[63] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *Advances in neural information processing systems*, vol. 30, 2017.

[64] D. M. Fobes, S. Claeys, F. Geth, and C. Coffrin, "Powermodelsdistribution.jl: An open-source framework for exploring distribution power flow formulations," *Electric Power Systems Research*, vol. 189, p. 106664, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378779620304673