# Application of Multi-Agent Reinforcement Learning for Battery Management in Renewable Mini-Grids

**Oluwatomisin I. Dada, Pierre Thodoroff, Neil D. Lawrence**

[1] Department of Computer Science, Cambridge University
tomisin.dada@gmail.com, pt440@cam.ac.uk, ndl21@cam.ac.uk

## Abstract

Electricity is an integral part of modern society, yet globally millions of people are without access. This lack of access, coupled with increasing concern over climate change represents a serious global challenge. Distributed energy storage will likely play a large part in the future of the grid, however, battery management remains an open problem. In this work, we re-frame the battery management problem in an Operations Research (OR) context as a multi-agent newsvendor problem. We benchmark seven Multi-Agent Reinforcement Learning (MARL) algorithms and compare their performance with five popular handcrafted heuristic strategies. We considered MARL algorithms due to their capacity to learn novel policies from data that may outperform handcrafted rule-based policies, especially as problem complexity increases. We find that all seven methods learn policies that achieve comparable results to each other and outperform a simple keep-fully-charged heuristic consistently. However, they do not consistently outperform all the heuristics considered in all the scenarios considered.

## Introduction

Access to electricity is an important sustainable development goal and index for measuring standard of living (IEA et al. 2011; Rao and Pachauri 2017; Klugman 2010; Bridge, Adhikari, and Fontenla 2016). With growing concern over climate change there is a need to rethink how electrical grids operate. Grids need to be redesigned to tackle what the World Energy Council has termed the *energy trilemma* (Song et al. 2017). The three aspects of the trilemma are *energy security*, *energy equity* and *environmental sustainability*. Energy security is concerned with the maintenance of a stable and dependable energy supply and the management of the supply and demand to maintain balance. Energy equity is concerned with increasing access to electricity and ensuring that electricity is affordable and environmental sustainability requires a significant increase in the use of low carbon energy sources and improved energy efficiency. It is challenging to address all three aspects of the energy trilemma simultaneously and yet it is important we do so not just to improve standards of living for individuals globally but also to reduce the environmental impact of electricity generation and help combat climate change. There is a need for '*smarter*' and more flexible power networks that will take advantage of the proliferation of power electronic devices to allow for increased usage of renewable energy sources and maximise the utility we receive from our power networks. Renewable mini-grids are currently considered a promising solution to the problem of rural electrification in many sub-Saharan countries (Russel; Muhoza and Johnson 2018; Eras-Almeida and Egido-Aguilera 2019) primarily addressing two aspects of the energy trilemma.

Renewable mini-grids are relatively small scale power networks where power is generated and used locally from a mix of generation sources (primarily a renewable source). They are typically restricted to lower power and voltage ratings than traditional power grids, serve a smaller geographical area and require significant energy storage/batteries. In the situation where batteries are located at the end-user side in the power network, the question of how much charge to store in these batteries arises. And as there are multiple users in the network this creates a situation in which we have to share a resource between multiple agents sensibly. In this work we reformulate the problem of resource distribution amongst multiple agents using an OR (Operations Research) framework, as a multi-agent newsvendor problem. We consider how multiple agents learn to share in a resource-constrained environment using reinforcement learning for multiple information structures. The term *information structure* here refers to the type of information that gets shared between agents as well as the employed communication protocol and network topology as seen in (Zhang et al. 2018; Pretorius et al. 2020).

There are two main reasons for the application of MARL to this problem: MARL has the potential to *self-discover* novel policies that outperform handcrafted rule-based policies (Janakiraman, Seshadri, and Shanthikumar 2007; Bijvank et al. 2014; Pretorius et al. 2020). This is particularly true as the complexity of the problem increases, while there is an optimal closed-form solution to the single-period single-agent newsvendor problem with increased lead time and periods the problem becomes increasingly complex with no exact solution (Zipkin 2000, 2008). The added dimension of multiple agents and variable prices further complicates the problem making the prospect of

learning policies through reinforcement learning appealing (Balaji et al. 2019). According to the `Autocurriculum hypothesis` proposed by Leibo et al (Leibo et al. 2019) "multi-agent systems sometimes display intrinsic dynamics arising from competition and cooperation that provide a naturally emergent curriculum, where the solution of one social task often begets new social tasks, continually generating novel challenges, and thereby promoting innovation".

In this work, we have created a mini-grid environment [1] where multiple agents share access to a pool of available electricity where the price is affected by the quantity demanded. The core research contributions of this work are:

- The formulation of the problem of resource distribution amongst multiple agents in an OR (Operations Research) framework, as a multi-agent newsvendor problem. In this framework batteries are treated as warehouses and electricity '*inventory*'. In this framework, the question then becomes how much '*inventory*' (electricity) to store given knowledge of the stochastic nature of demand for electricity and the economic cost of failing to meet demand.
- The implementation of baseline heuristics to establish performance benchmarks and the modification of the existing closed-form solution to the single-period, single-agent newsvendor problem for multiple agents and variable inventory cost.
- Implementation of MARL for seven different information structures and the evaluation of the resulting learned policies of the trained agents in comparison to the established baselines.

# Background

## Electrical Power Grids

**Basics** Traditionally, power grids consist of three sections, generation, transmission and distribution. Generation is the production of electricity from an energy source. This has generally been done using large turbines powered by coal, gas, or hydro. The rotation of these turbines in a magnetic field generates an alternating current (AC). Generation typically occurs at 23kV in the UK (Simmonds 2002) and large scale generation often occurs far away from load centres where the power is needed. After generation comes transmission where voltage is stepped-up using transformers from 132kV to 400kV (Simmonds 2002). As shown in equations 1 and 2, useful power ($P$) is directly proportional to the current ($I$) while power loss ($P_{loss}$) is proportional to the square of the current ($I^2$).

$$P = I \cdot V, \tag{1}$$

$$P_{loss} = I^2 \cdot R_{line} \tag{2}$$

As power in traditional grids is required to flow across large distances, there is an accumulation of significant resistance ($R_{line}$) in the power lines. It is necessary to reduce the current to mitigate power loss. Finally, after transmission comes distribution, where end-users (residential homes, offices) connect to the grid and voltage gets stepped down to 11kV - 230V. A voltage range where it is safe for consumers to connect.

In contrast, renewable mini-grids only consists of generation and distribution. Given their smaller geographical scope it is not necessary nor practical to have the high voltages found in transmission networks. Due to the tighter economic constraints of renewable mini-grids, operators must extract the maximum utility they can from constructed infrastructure. This leads to one of the reasons for increased energy storage in the network at the consumer end as it allows for increased demand-side management. With demand-side management grid operators can reduce peak demand and shift demand to periods with reduced grid activity. Mini-grid operators gain two benefits by reducing the size of the demand peak:

- A reduction in power loss due to the reduced current magnitude which reduces operating cost.
- A reduction in the required peak capacity of the network which diminishes construction cost.

Additionally, when balancing demand and supply, traditional power networks have the advantage of having many large spinning turbines within the network responsible for generating electricity. These spinning turbines provide inertia to the system and prevent rapid changes in the network frequency. Most renewable energy sources do not exhibit this inertia and so any mismatch in demand and supply in a renewable mini grid results in faster variations of frequency of greater magnitude (Castro, Fuerte-Esquivel, and Tovar-Hernandez 2012). For a renewable mini-grid, there is significant utility in a battery management system that forecasts its demand and responds to price signals, validating the problem framework chosen for this work.

**The Electricity Market** The electricity market consists of the demand and supply curves for electricity. A demand curve shows the quantity of electricity users would consume at different price levels. The supply curve shows the quantity of electricity suppliers are willing to sell at different price levels. In the short term, electricity demand is usually considered price inelastic where the price of electricity has little impact on the quantity of electricity demanded. The supply curve of electricity resembles a series of steps. Suppliers only agree to supply electricity when the price reaches the marginal cost of electricity production using their corresponding technology. Each step represents a different technology with a higher marginal cost. Figure 1 illustrates a demand and supply curve for the electricity market, where demand and supply curves intersect at market equilibrium.

## Newsvendor Problem

The classic newsvendor problem is a three-node supply chain that consists of a supplier, buyer and customer (Arrow, Harris, and Marschak 1951; Silver et al. 1998). There are many possible extensions to this problem involving variations in customer demand (Qin et al. 2011; Burke, Carrillo, and Vakharia 2007), buying risk profiles (Qin et al. 2011;

[1] Code available at https://github.com/dadatomisin/cam2021

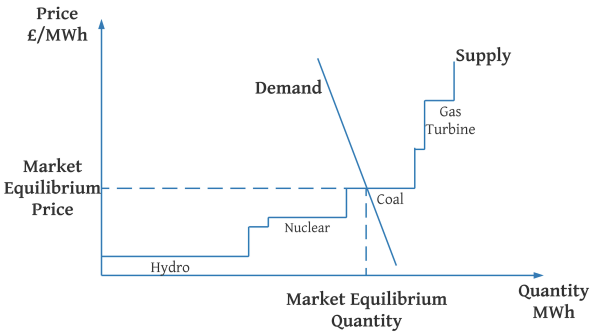


Figure 1: Illustration of Demand and Supply Curve for Electricity showing various generating technologies.

Burke, Carrillo, and Vakharia 2007), extended lead times and supplier pricing profiles (Qin et al. 2011). In this study, we examine the simple single period price-taking baseline problem. This form of the problem has an intuitive closed-form solution and serves as a building block for the multi-agent environment built in this project.

In the single-period, price-taking newsvendor problem, stock order quantity is set and arrives before the selling period starts. During the selling period, there is a realisation from the *stationary* stochastic demand. Thus the ordering is done with uncertainty about the demand realisation and it is not possible to order additional units to cover an unexpected part of demand (Petruzzi and Dada 1999; Arikan 2018). At the end of the selling period, earnings from unmet demand are lost and surplus inventory is rendered obsolete. As a price taker, the vendor is seen as a small part of a perfectly competitive market and has no power to affect the selling price of its goods to customers. For a given cost price $c$, selling price $p$, stock order quantity $y$ and demand realisation $x$ the vendors profit $\Upsilon(y)$ is given by equation 3 (Arikan 2018),

$$\Upsilon(y) = p \cdot \min(x, y) - cy. \tag{3}$$

The vendor's goal is to find the optimal order quantity which maximises its expected profit for demand with a probability density function $f(x)$,

$$\begin{aligned}\mathbb{E}[\Upsilon] &= \int_0^y |px - cy| f(x)\mathrm{d}x + \int_y^\infty |py - cy| f(x)\mathrm{d}x \\ &= (p-c)y - p\int_0^y F(x)\mathrm{d}x,\end{aligned} \tag{4}$$

Setting the derivative of the expected profit equal to zero and solving for $y$ yields the optimal order quantity shown in equation 5,

$$y^* = F^{-1}\left(\frac{p-c}{p}\right), \tag{5}$$

where $F^{-1}(\cdot)$ is the inverse cumulative distribution function of demand. Taking the second derivative $\Upsilon(y)''$ of the expected profit function shows that $y^*$ is a unique optimiser as this derivative is strictly concave, given its negative second derivative shown in equation 6 (Arikan 2018),

$$\Upsilon(y)'' = -pf(y) < 0. \tag{6}$$

This result is the building block for the order-up-to policy used in the frequently seen problem extension of multi-period settings. Due to the concavity of the objective function, this policy considers if there is existing stock when setting order quantity at the start of the selling period. If this stock is below the optimal $y^*$, the optimal action is to order additional stock to reach this level, otherwise do nothing. The numerator $(p-c)$ in equation 5 represents the opportunity cost of missing out on the sale of one unit of stock due to ordering too few and is commonly called the underage cost $C_u$. The denominator $p$, however, is the cost of ordering a single additional unit that is not sold and is called the overage cost $C_o$. So the optimal order quantity balances the expected overage and underage cost under a given probability distribution. It is common to add a penalty for loss of goodwill caused by missing out on a potential sale to the underage cost and subtract the salvage value gotten from disposing of excess inventory from the overage cost. Thus this fraction can be rewritten as shown in equation 7 and is commonly called the critical ratio (CR),

$$CR = \left(\frac{p-c+k}{p-c+k+h}\right), \tag{7}$$

where $p$ is the selling price, $c$ is the cost price $k$ is the penalty for loss of goodwill and $h$ is the holding cost. The critical ratio reflects what fraction of uncertainty to cover in the order quantity.

## Reinforcement Learning

Reinforcement Learning (RL) involves an agent sensing and acting within a given environment to achieve a goal by maximising a given reward signal (Sutton and Barto 2018). Through the interaction of the agent with the environment, RL attempts to create a mapping between situations/states and actions. An agent learns a *policy* $\pi$ which determines its action $a$ from a set of possible *actions* $a \in \mathbf{A}$ given a state $s$ from a set of possible *states* $s \in \mathbf{S}$. The reward signal $r(s)$ and *value functions* $\mathbf{V}_\pi(s)$, $\mathbf{Q}_\pi(s,a)$ serve as the primary and secondary signals that inform the agent about the appropriateness of an action and are used to modify the agent's policy (Sutton and Barto 2018). The reward signal is the direct feedback the agent receives from the environment while the value function estimates how good a specific state $s$ or action $a$ is by calculating the long term cumulative rewards associated with that state or action.

Markov decision processes (MDPs) (and partially observable Markov decision processes (POMDPs)) are the mathematical framework which allow us to formalise sequential decision making in RL. MDPs are used for modelling decision making in situations where outcomes are partly random and partly under the control of a decision maker. MDPs include a *transition function* ($T(s_t, a_t, s_{t+1}) : \mathbf{S} \times \mathbf{A} \times \mathbf{S} \rightarrow [0,1]$) which shows that the next state $s_{t+1}$ is dependent on the current state $s_t$ and agent action $a_t$. State transitions functions in MDPs must satisfy the Markov property that for a given state $s$ and action $a$ the state transition is conditionally independent of all previous states and actions (Sutton and Barto 2018).

In an episodic setting the goal of RL is to find the optimal policy which maximises the long term reward or expected return $G_t$ which can be expressed as $G_t = r_{t+1} + r_{t+2} + ... + r_T$ for final time step $T$. For continuous problems where $T = \infty$ a discount factor $\gamma$ which varies from $0 - 1$ is included which reflects the importance of immediate rewards and discounts future rewards and so the expected return can be rewritten as $G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$ (Sutton and Barto 2018). The state-value function $\mathbf{V}_\pi(s)$ estimates the expected return from following a policy $\pi$ starting from a given state $s$ and the action-value function $\mathbf{Q}_\pi(s, a)$ estimates the expected reward from following a policy $\pi$ starting from a given state $s$ after taking an action $a$. These functions have a recursive property with successor states which gives the *Bellman equation* for the value function (Sutton and Barto 2018). Using this property the value of states and actions can be solved recursively. However, this can quickly become untenable because if states are unique then there are exponentially many state-action pairs to consider, a phenomenon called the curse of dimensionality. In deep RL, we approximate functions using deep neural networks (NNs) and compare the relative closeness of states using a learned metric and estimate values based on this (Sutton and Barto 2018). The optimal policy is the policy that maximises the state-value and action-value functions.

**Multi Agent Reinforcement Learning (MARL)** MARL considers the problem of multiple agents acting within an environment. Multi-agent systems improve robustness to single-agent failure and many problems naturally lend themselves to multi-agent frameworks (Tian 2012). However, major challenges to MARL include the *credit assignment problem* (Sutton 1984) where there is a decreased correlation between agent action and reward signal and the *non-stationarity* of the environment as from the perspective of a single agent the environment is constantly changing which results in learning stability issues (Foerster et al. 2017). *Centralised learning, decentralised execution* which is a standard paradigm for multi-agent planning (Foerster et al. 2016; Kraemer and Banerjee 2016; Oliehoek, Spaan, and Vlassis 2008) is one approach to mitigating this problem. In addition to this various communication protocols have been developed to help agents collaborate better in these environments. MARL work can be classified into four categories based on the communication method employed (Chu, Chinchali, and Katti 2020). The first category is non-communicative and focuses on stabilizing training with advanced value estimation methods (Lowe et al. 2017). These include methods such as multi-agent deep deterministic policy gradients (MADDPG Lowe et al. 2017), counterfactual multi-agent (COMA Foerster et al. 2018) policy gradients and parameter sharing trust region policy optimisation (PS-TRPO Gupta, Egorov, and Kochenderfer 2017). The second category considers direct information sharing where low dimension policy fingerprints (Foerster et al. 2017) or average neighbourhood policies are shared (Yang et al. 2018). The communication policies in this category are designed explicitly before training and could be redundant or inefficient. The third category consists of learned communication policies where the communication channel is differentiable in training and agents learn messages to communicate in training (Foerster et al. 2016; Chu et al. 2019). The fourth category focuses on communication *attention* to selectively send messages (Jiang and Lu 2018; Singh, Jain, and Sukhbaatar 2018). The first category simply treats other agents as part of the environment and tries to counter their effects on the environment with more advanced value estimation methods. The second category allows agents to consider the policies/actions of other agents in the environment, however, the information shared is explicitly determined beforehand by the designer. The third category allows agents to learn a communication policy during training that is beneficial to the task. In this work, we employ MARL techniques from the first three categories.

MARL also allows agents to act not just to maximise their rewards but also a shared global reward and neighbourhood reward functions play a key role in helping agents learn to collaborate (Pretorius et al. 2020),

$$\mathbb{E}\left[\sum_{t=1}^{T} \gamma^{t-1} \left( r_{i,t} + \sum_{j \in N_i} \alpha r_{j,t} \right)\right]. \tag{8}$$

To encourage collaboration in the neighbourhood reward function, agents also include a discounted sum of the rewards of neighbouring agents. $\alpha$ is the neighbourhood discount factor which varies between 0 and 1 and determines the importance agents attach to neighbouring agents rewards. Neighbouring agents are agents with a direct communication channel connecting them.

**Differentiable Communication Protocols** Differentiable communication protocols can play a vital role in successfully learning to complete cooperative tasks in multi-agent environments. There are three differentiable communication protocols considered in this paper, differentiable inter agent learning (DIAL), CommNet and NeurComm. In CommNet agents generate real value messages which are averaged at the receiving side (Sukhbaatar, Szlam, and Fergus 2016). In DIAL agents generate not just messages but action-value estimates which are not averaged but encoded and summed (Foerster et al. 2016). Chu, Chinchali, and Katti (2020) argue that the aggregation of input signals in both works leads to a loss of information and so developed NeurComm. NeurComm encodes and concatenates messages instead of summing them and includes policy fingerprints to help further reduce non-stationarity.

## Related Work

There is existing work that considers the battery management problem in an OR framework. Marchi, Zanoni, and Pasetti (2019) conducted a numerical study to demonstrate the benefits of the application of traditional inventory management techniques to this problem. Saran, Goentzel, and Siegert (2010) focused on the evaluation and formulation of operation policies for a wind plant using concepts from the newsvendor problem while Schneider et al. (2016) focused on the optimal sizing of batteries under uncertain supply conditions. However, these papers do not consider the

implication of multiple agents interacting within this environment or apply reinforcement learning techniques to this problem. Existing work on the newsvendor problem in the field of reinforcement learning is primarily focused on the single-period problem and are often trying to learn one of the inputs, such as demand in work by Oroojlooyjadid, Snyder, and Takác (2016). Gijsbrechts et al. (2018) tackled the dual sourcing problem using RL and Balaji et al. (2019) tackled the multi period newsvendor problem with lead time. However, these papers still consider a single agent framework of the problem and do not apply multi-agent reinforcement learning techniques. Pretorius et al. (2020) conducted a game-theoretic analysis of MARL for the distribution of common-pool resources. In this work, MARL was applied to a water management systems where at each time step agents have the binary choices of consuming or abstaining from consuming the available resource.

# Experiment Setup

## Environment

We consider a multi-agent battery management problem with stationary stochastic demand. We formalise our problem as a Markov decision process (MDP). The demand $D$ of each agent $i$ is assumed to be stationary and Poisson distributed with mean $\mu_i$. Electricity is purchased at a cost dependent on the point on the supply curve the aggregate demand reaches. We assume that the electricity supply in the mini-grid is a mix of two sources. The first source is a renewable resource such as solar or wind which makes up the majority of the electricity supply and has a low/zero marginal cost of energy production (Blazquez et al. 2018) while the second energy source is a fuel-burning source such as a natural gas-powered turbine which has a higher marginal cost of production. We also assume that government regulation ensures that despite the monopolistic nature of supply, generators charge prices equal to their marginal cost of production. These assumptions result in a supply curve which is a simple step function. Here price increases to the higher marginal cost when the quantity demanded exceeds the available quantity of renewable energy.

As in the newsvendor framework, for each unit of demand met from available inventory the end users receives a fixed utility $p$. Demand not satisfied by available inventory incurs a penalty $k$ which is the cost of electricity at a premium, as demand and supply must always be balanced in the power grid. Any units leftover incur a holding cost $h$, which reflects the natural self-discharge rate and inefficiencies of the energy storage device and is represented as a fraction of the current cost of electricity given by the electricity supply curve.

**State** The state $S$ of the problem is given by an $N \times 7$ dimensional matrix where $N$ represents the number of agents in the system,

$$S = \begin{bmatrix} p & c & h & k & r & \mu_1 & b_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \mu_i & b_i \\ p & c & h & k & r & \mu_N & b_N \end{bmatrix}, \tag{9}$$

where $c$ represents the higher marginal cost associated with natural gas electricity generation, $r$ is the available quantity of renewable resource and $b_i$ represents the current state of charge of agent $i$'s battery.

Each agent $i$ in the environment receives the corresponding row $i$ as an observation. And while the values $p$, $c$, $h$ and $k$ were kept constant between agents in this dissertation it would be possible to modify it to reflect heterogeneity in the battery efficiencies of agents or the utility they receive from electricity.

**Action Space** In this work we consider a discrete action space. At the start of each period, each agent $i$ decides how much charge to store in their battery to satisfy demand while minimising cost. It is assumed that the battery sizes of each agent are equivalent and the actions space is discretised into 101 possible actions where agents can demand between 0 to 100 units of electricity as long as they have the available spare capacity to store the charge otherwise demand is limited to this available spare capacity.

**Reward** Each agent first incurs the purchasing cost corresponding to the procured units given the action $a_i$ and the price of electricity given the collective action $\sum_{i=1}^{N} a_i$ and supply curve function $\mathbf{S}(\cdot)$. For each agent a realization $d_i$ of the demand $D_i$ (Poisson distributed with mean $\mu_i$) is observed, and demand is satisfied as much as is possible given available energy storage. Underestimated demand incurs a penalty of $k$ per unit, while leftover units incur a holding cost $h$ resulting in reward $r_i$,

$$r_i = p \cdot \min(b_i, d_i) - c_s a_i - h c_s (b_i - d_i)^+ - k(d_i - b_i)^+, \tag{10}$$

where $c_s = \mathbf{S}(\sum_{i=1}^{N} a_i)$ and $(x)^+ = \max(x, 0)$.

**Transition** The state of the system $S$ is updated to $S_+$ by updating the battery levels of the agents,

$$S_+ = \begin{bmatrix} p & c & h & k & r & \mu_1 & (b_1 - d_1)^+ \\ \vdots & \vdots & \vdots & \vdots & \vdots & \mu_i & (b_i - d_i)^+ \\ p & c & h & k & r & \mu_N & (b_N - d_N)^+ \end{bmatrix}. \tag{11}$$

**Software Packages** We use `OpenAI Gym` (Brockman et al. 2016), `Ray RLlib` (Liang et al. 2017) to build the simulation environment . We consider seven RL algorithms originally implemented by Chu et al. (2019) and extended by Pretorius et al. (2020), however, with the unified API offered by `RayRLlib` other RL algorithms can easily be applied to this environment.

## Methods

In this work, we implemented 5 heuristic methods and MARL methods with 7 different information structures. The heuristic methods serve as benchmarks on performance for comparison with the MARL methods implemented. The heuristic methods implemented fall broadly into two categories *compromising* and *uncompromising* strategies.

**Compromising Strategy** Compromising strategies deliberately restrict the aggregate demand of the agents in the environment to below the available lower-cost renewable resource supply prioritising minimising cost. Two of the heuristics implemented in this work employ compromising strategies.

- **Equal Division of Available Renewable Resource:** This policy distributes the available low-cost renewable resource equally amongst all agents.
- **Proportional Division of Available Renewable Resource:** This policy distributes the available low-cost renewable resource proportionally amongst all agents to reduce inequality.

**Uncompromising Strategy** Uncompromising strategies make no deliberate effort to restrict the aggregate demand of the agents in the environment but focus on satisfying realisation of demand. Three of the heuristics implemented in this work employ uncompromising strategies.

- **Keep-Fully-Charged:** Each agents take actions to keep respective batteries fully charged.
- **Mean Action:** In this policy each agent requests its mean demand $\mu_i$ at each time step.
- **Converged Aggregate Critical Ratio:** This policy builds off the Critical Ratio closed-form solution to the single period, single-agent newsvendor problem. In this approach we generate individual demand curves $D_i(p) = (F^{-1}(CR, \mu_i) - b_i)^+$ for each agent which account for current battery levels, agent's mean demand and associated economic costs. These individual demand curves reflect the optimal quantity of inventory each agent would stock at each price point. Individual demand curves are then summed to generate an aggregate demand curve $D_{agg}(p) = D_1(p) + \cdots + D_i(p) + \cdots + D_n(p)$ and using the bisection method the intersection of the aggregate demand and supply curves is determined. Agents then place quantity orders that correspond to this point of intersection. This process is repeated at each time step. At this point, there is an economic equilibrium between supply and aggregate demand. However, it is not necessarily optimal. It is worth noting that the $\sum_{i=1}^{N}\left(F^{-1}(CR, \mu_i)\right) \neq F^{-1}(CR, \sum_{i=1}^{N}(\mu_i))$ and so aggregate demand $D_{agg}$ can not be found by summing the individual means and taking the inverse cumulative function of the resulting value even when the critical ratio (CR) is the same for all agents.

**Multi-Agent Reinforcement Learning Methods** We consider MARL with seven different information structures which were recently used by Chu et al. (2019) and Pretorius et al. (2020) who studied the applications of MARL to adaptive traffic signal control and common-pool resource management respectively. The information structure considered are as follows:

- **Independent A2C (IA2C):** This framework is non-communicative wherein all agents are disconnected from each other and A2C is implemented in fully independent manner (Pretorius et al. 2020; Tan 1993).

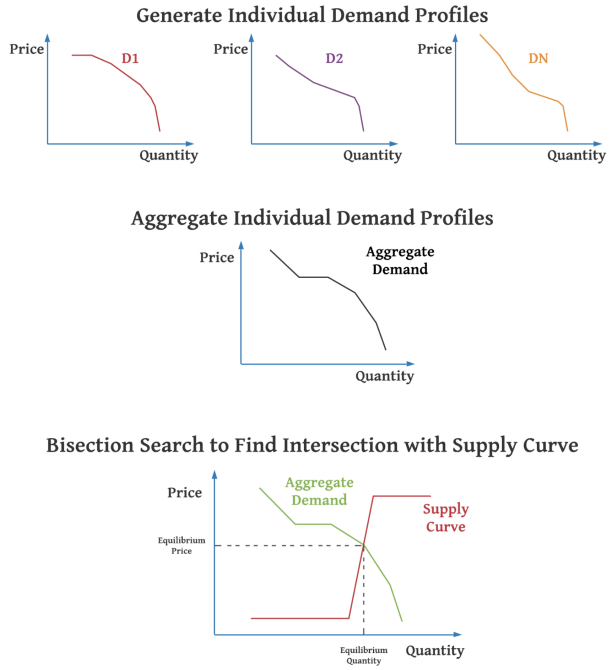


Figure 2: Illustration of Operation of Converged Aggregate Critical Ratio. This figure shows the individual demand curves formulated using the Critical Ratios that are aggregated to form the aggregate demand curve. The intersection of these two curves is found using a binary search.

- **Networked A2C (NA2C):** This is a networked A2C implementation of the MADDPG algorithm (Lowe et al. 2017) without explicit communication. Agents receive the observations from other agents they are connected to.
- **NA2C with Policy Fingerprints (NA2C FPrint ):** This method extends the previous version by including direct information sharing of low dimension policy fingerprints (Foerster et al. 2017).
- **NA2C with Consensus Network (NA2C ConseNet):** This method also involves direct information sharing with the consensus update (Zhang et al. 2018).
- **NA2C with CommNet (NA2C CommNet):** The method implements the differentiable communication protocol CommNet (Sukhbaatar, Szlam, and Fergus 2016).
- **NA2C with DIAL (NA2C DIAL):** The method implements the differentiable communication protocol DIAL (Foerster et al. 2016).
- **NA2C with NeurComm (NA2C NeurComm):** The method implements the differentiable communication protocol NeurComm (Chu, Chinchali, and Katti 2020).

All the algorithms implemented are decentralised in that none of them employs a centralised critic or global policy network. However, the value estimates from the critic network for each algorithm are conditioned on shared neighbourhood actions.

## Evaluation Metrics

We consider two evaluation metrics for assessing the performance of a policy, the weighted average individual reward (average individual satisfaction) and an inequality measure.

**Average Individual Satisfaction** The weighted average individual reward $R_{WAI}$ reflects the average individual utility gained by the agent from its actions and is weighted by the agent's mean demand $\mu_i$. The $R_{WAI}$ is weighted by individual means as larger means result in larger reward values for the same percentage of demand satisfaction. Individual weighting considers average reward per unit mean demand which we believe is the right metric when considering equity in the reward distribution. Equation 12 gives the formula for weighted average individual reward $R_{WAI}$,

$$R_{WAI}(i) = \frac{1}{T\mu_i} \sum_{t=0}^{T} r_i(t). \tag{12}$$

**Inequality** Given potential inequalities in the distribution of rewards, we consider an inequality measure between weighted average individual rewards using the *Gini Coefficient* (Dorfman 1979)

$$G = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} |x_i - x_j|}{2N \sum_{i=1}^{N} x_i}, \tag{13}$$

which is frequently used to measure income inequality by economists. The coefficient range from 0 representing complete equality to 1 representing complete inequality where all resources are held by one individual assuming non-negative values. In situations with negative values inequality may exceed 1. The coefficient is defined mathematically as shown in Equation 13 where $x_i$, in this case, is the individual satisfaction score and there are $N$ individuals/agents.

## Results

We consider two different levels of available renewable resource, moderate (60% of aggregate mean demand) and high (90% of aggregate mean demand). We average the results of 10 episodes using each method with randomly generated mean demands and economic cost factors ($p$, $c$, $h$, $k$) held constant and each episode lasting 10,000 steps.

**Moderate Renewable Resource Availability** In the moderate renewable resource availability setting, performance can be grouped into four levels based on average satisfaction. CACR is the best performing method with the highest average satisfaction (**0.45362**), lowest variance, and forms the first level as shown in figure 3. Mean action is the second-best performing method and forms the second level, collectively all 7 MARL methods form the third level with similar performance and keep-fully-charged, equal and proportional distribution form the last level. While keep-fully-charged, equal and proportional distribution all have similar average satisfaction, keep-fully-charged has a much larger variance in performance. A keep-fully-charged approach has better performance with higher mean demands and so variations in performance are due to variation in mean demand.

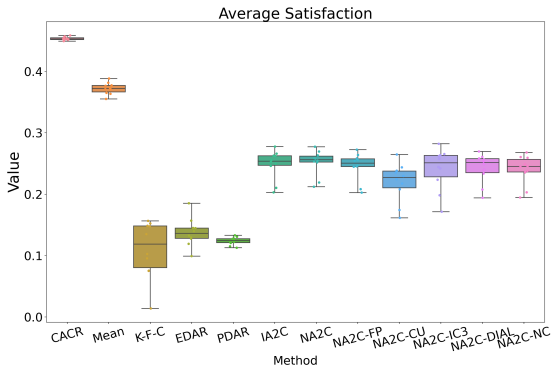


Figure 3: Box plots for average satisfaction of heuristic and MARL methods with a renewable resource availability of 60%. From this plot CACR is the best performing method with a tight distribution, mean action is the second-best performing method followed by the MARL methods.

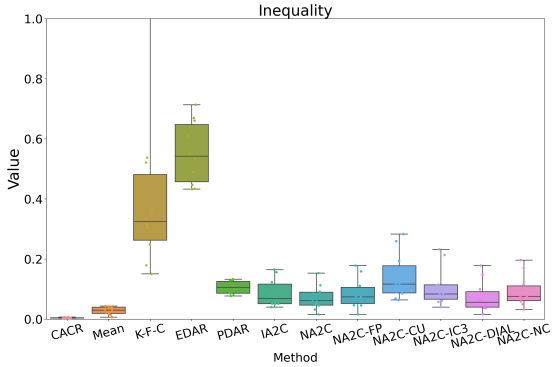


Figure 4: Box plots for inequality of heuristic and MARL methods with a renewable resource availability of 60%. CACR is the best performing method with a tight distribution and low inequality. All methods apart from keep-fully-charged and equal distribution resulted in low inequality.

All 7 MARL methods have similar average satisfaction scores, with only approximately a **0.03** difference in average satisfaction between the highest (**NA2C**) and lowest-performing (**NA2C - ConseNet**) MARL methods. These results do not indicate a clear superiority in any of the MARL methods considered. All methods resulted in low inequality ($<$ **0.15000**) with the exceptions of keep-fully-charged and equal distribution. There is significant variance in inequality scores for keep-fully-charged with the final point of the strip plot omitted in figure 4 to maintain a scale that allows for visual comparison of all methods. Inequality exceeded 1.0 due to negative individual satisfaction scores in some agents. Of the seven MARL methods considered NA2C - ConseNet and NA2C - CommNet which employ consensus in their information structures resulted in the highest mean inequality scores ($>$ **0.10000**) with the largest variance.

**High Renewable Resource Availability** In the high renewable resource availability setting we see a reorganisation of the performance levels different methods fall into based on average satisfaction as can be seen from figure 5. Equal and proportional distributions achieve the best performance with average satisfaction ($>$ **0.65000**). Proportional distribution achieves the highest mean scores and has significantly lower variance than the equal distribution method. CACR forms the next performance level, followed by the seven MARL methods which outperform mean action and achieve performance approaching the level achieved by CACR. Keep-fully-charged and mean action form the last performance level, although with mean action having higher mean values and lower variance.

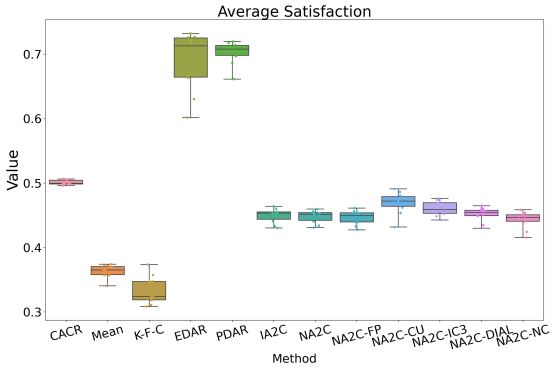


Figure 5: Box and strip plots for average satisfaction of heuristic and MARL methods with a renewable resource availability of 90%. Equal and proportional distributions are the best performing methods although equal distribution has a wider performance spread, CACR is the second-best performing method followed by the MARL methods

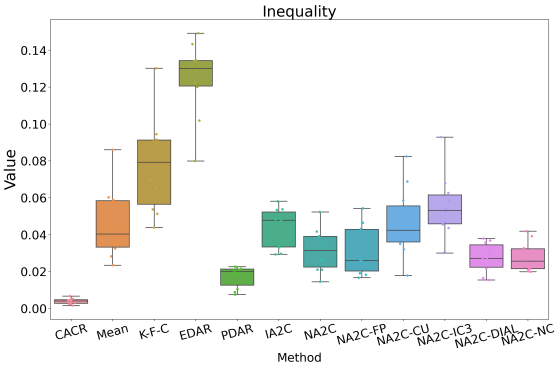


Figure 6: Box and strip plots for inequality of heuristic and MARL methods with a renewable resource availability of 90%. CACR is the best performing method with a very tight distribution and low inequality. All methods resulted in low inequality ($<$ 0.15000).

Again all 7 MARL methods have similar average utility and average satisfaction scores, with only approximately a **0.026** difference in average satisfaction between the highest (**NA2C - ConseNet**) and lowest-performing (**NA2C - NC**) MARL methods. These results show a reversal in performance by NA2C - ConseNet but still do not indicate a clear superiority in any of the MARL methods considered. With regards to inequality, generally speaking, all methods resulted in low inequality ($<$ **0.15000**) with keep-fully-charged (**0.07794**) and equal distribution (**0.12458**) again having the highest mean inequality scores. Of the seven MARL methods considered NA2C - ConseNet and NA2C - CommNet resulted in the highest mean inequality scores ($>$ **0.04500**) with the largest variance.

## Discussion

At moderate levels, the MARL methods achieved average satisfaction scores below mean action, however, at high levels of renewable resource availability they achieved higher scores than mean action. The policies learnt by the MARL methods appear to fall in between compromising and uncompromising benefiting significantly from the increased availability but not reaching the scores attained by equal and proportional distribution. For all the information structures implemented MARL achieved low inequality scores which were lowered further when availability was increased. Although, there were differences observed in the scores achieved by the different MARL methods the margin was not large enough to indicate the clear superiority of any one method over the others for this problem. And while the CACR heuristic has no training time it has a significantly longer execution time than the MARL methods as it performs a binary search at each time step. In conclusion, no single method heuristic or MARL produces the best performance in all considered problem settings. The performance of the MARL methods are encouraging as they learned policies that differed from the handcrafted heuristics, consistently outperformed a simple "keep-fully-charged" policy, had low inequality and responded positively to an increase in renewable resource availability. While the MARL methods did not outperform all the heuristics considered there is still potential room for improvement with more extensive hyperparameter tuning and longer training times.

We also identified several areas for potential future research that are not fully considered in this work. We consider curriculum learning a potential path to MARL consistently outperforming the established hand-crafted heuristics. In this curriculum, this complex task could be broken down into three sub-tasks of learning optimal uncompromising and compromising strategies and then learning when to apply which. We also believe it would be interesting to consider the performance of MARL methods with a large number of agents and diverse neighbourhood connectivity structures as this is more reflective of potential real-world deployment. Potential real-world deployments also raise questions with regards to privacy, the information shared between agents could be considered sensitive and may compromise the privacy of individual agents. Finally, the effects on the performance of deceptive/non-co-operative agents should be considered to ensure that any deployed MARL system is robust.

## References


Arikan, E. 2018. *Single Period Inventory Control and Pricing: An Empirical and Analytical Study of a Generalized Model*. Peter Lang International Academic Publishers.

Arrow, K. J.; Harris, T.; and Marschak, J. 1951. Optimal inventory policy. *Econometrica: Journal of the Econometric Society*, 250–272.

Balaji, B.; Bell-Masterson, J.; Bilgin, E.; Damianou, A. C.; Garcia, P. M.; Jain, A.; Luo, R.; Maggiar, A.; Narayanaswamy, B.; and Ye, C. 2019. ORL: Reinforcement Learning Benchmarks for Online Stochastic Optimization Problems. *CoRR*, abs/1911.10641.

Bijvank, M.; Huh, W. T.; Janakiraman, G.; and Kang, W. 2014. Robustness of order-up-to policies in lost-sales inventory systems. *Operations Research*, 62(5): 1040–1047.

Blazquez, J.; Fuentes-Bracamontes, R.; Bollino, C. A.; and Nezamuddin, N. 2018. The renewable energy policy Paradox. *Renewable and Sustainable Energy Reviews*, 82: 1–5.

Bridge, B. A.; Adhikari, D.; and Fontenla, M. 2016. Electricity, income, and quality of life. *The Social Science Journal*, 53(1): 33–39.

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. OpenAI Gym. .

Burke, G. J.; Carrillo, J. E.; and Vakharia, A. J. 2007. Single versus multiple supplier sourcing strategies. *European journal of operational research*, 182(1): 95–112.

Castro, L. M.; Fuerte-Esquivel, C. R.; and Tovar-Hernandez, J. H. 2012. Solution of power flow with automatic load-frequency control devices including wind farms. *IEEE Transactions on power systems*, 27(4): 2186–2195.

Chu, T.; Chinchali, S.; and Katti, S. 2020. Multi-agent reinforcement learning for networked system control. *arXiv preprint arXiv:2004.01339*.

Chu, T.; Wang, J.; Codecà, L.; and Li, Z. 2019. Multi-Agent Deep Reinforcement Learning for Large-Scale Traffic Signal Control. *IEEE Transactions on Intelligent Transportation Systems*.

Dorfman, R. 1979. A formula for the Gini coefficient. *The review of economics and statistics*, 146–149.

Eras-Almeida, A.; and Egido-Aguilera, M. 2019. Hybrid renewable mini-grids on non-interconnected small islands: Review of case studies. *Renewable and Sustainable Energy Reviews*, 116: 109417.

Foerster, J.; Farquhar, G.; Afouras, T.; Nardelli, N.; and Whiteson, S. 2018. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Foerster, J.; Nardelli, N.; Farquhar, G.; Afouras, T.; Torr, P. H.; Kohli, P.; and Whiteson, S. 2017. Stabilising experience replay for deep multi-agent reinforcement learning. In *International conference on machine learning*, 1146–1155. PMLR.

Foerster, J. N.; Assael, Y. M.; De Freitas, N.; and Whiteson, S. 2016. Learning to communicate with deep multi-agent reinforcement learning. *arXiv preprint arXiv:1605.06676*.

Gijsbrechts, J.; Boute, R. N.; Van Mieghem, J. A.; and Zhang, D. 2018. Can deep reinforcement learning improve inventory management? performance and implementation of dual sourcing-mode problems.

Gupta, J. K.; Egorov, M.; and Kochenderfer, M. 2017. Cooperative multi-agent control using deep reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems*, 66–83. Springer.

IEA, I.; et al. 2011. World energy outlook 2011. *Int. Energy Agency*, 666.

Janakiraman, G.; Seshadri, S.; and Shanthikumar, J. G. 2007. A comparison of the optimal costs of two canonical inventory systems. *Operations Research*, 55(5): 866–875.

Jiang, J.; and Lu, Z. 2018. Learning attentional communication for multi-agent cooperation. *arXiv preprint arXiv:1805.07733*.

Klugman, J. 2010. Human development report 2010–20th anniversary edition. The real wealth of nations: pathways to human development.

Kraemer, L.; and Banerjee, B. 2016. Multi-agent reinforcement learning as a rehearsal for decentralized planning. *Neurocomputing*, 190: 82–94.

Leibo, J. Z.; Hughes, E.; Lanctot, M.; and Graepel, T. 2019. Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. *arXiv preprint arXiv:1903.00742*.

Liang, E.; Liaw, R.; Nishihara, R.; Moritz, P.; Fox, R.; Gonzalez, J.; Goldberg, K.; and Stoica, I. 2017. Ray RLLib: A Composable and Scalable Reinforcement Learning Library. *CoRR*, abs/1712.09381.

Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; and Mordatch, I. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *arXiv preprint arXiv:1706.02275*.

Marchi, B.; Zanoni, S.; and Pasetti, M. 2019. Multi-Period Newsvendor Problem for the Management of Battery Energy Storage Systems in Support of Distributed Generation. *Energies*, 12(23): 4598.

Muhoza, C.; and Johnson, O. W. 2018. Exploring household energy transitions in rural Zambia from the user perspective. *Energy Policy*, 121: 25–34.

Oliehoek, F. A.; Spaan, M. T.; and Vlassis, N. 2008. Optimal and approximate Q-value functions for decentralized POMDPs. *Journal of Artificial Intelligence Research*, 32: 289–353.

Oroojlooyjadid, A.; Snyder, L. V.; and Takác, M. 2016. Applying Deep Learning to the Newsvendor Problem. *CoRR*, abs/1607.02177.

Petruzzi, N. C.; and Dada, M. 1999. Pricing and the newsvendor problem: A review with extensions. *Operations research*, 47(2): 183–194.

Pretorius, A.; Cameron, S.; van Biljon, E.; Makkink, T.; Mawjee, S.; Plessis, J. d.; Shock, J.; Laterre, A.; and Beguir, K. 2020. A game-theoretic analysis of networked system control for common-pool resource management using multi-agent reinforcement learning. *arXiv preprint arXiv:2010.07777*.

Qin, Y.; Wang, R.; Vakharia, A. J.; Chen, Y.; and Seref, M. M. 2011. The newsvendor problem: Review and directions for future research. *European Journal of Operational Research*, 213(2): 361–374.

Rao, N. D.; and Pachauri, S. 2017. Energy access and living standards: some observations on recent trends. *Environmental Research Letters*, 12(2): 025011.

Russel, W. ???? Renewable energy mini-grids: An alternative approach to energy access in southern Africa.

Saran, P.; Goentzel, J.; and Siegert, C. W. 2010. Economic analysis of wind plant and battery storage operation using supply chain management techniques. In *IEEE PES General Meeting*, 1–8. IEEE.

Schneider, M.; Biel, K.; Pfaller, S.; Schaede, H.; Rinderknecht, S.; and Glock, C. H. 2016. Using inventory models for sizing energy storage systems: An interdisciplinary approach. *Journal of Energy Storage*, 8: 339–348.

Silver, E. A.; Pyke, D. F.; Peterson, R.; et al. 1998. *Inventory management and production planning and scheduling*, volume 3. Wiley New York.

Simmonds, G. 2002. *Regulation of the UK electricity industry*, volume 73. University of Bath School of Management.

Singh, A.; Jain, T.; and Sukhbaatar, S. 2018. Learning when to communicate at scale in multiagent cooperative and competitive tasks. *arXiv preprint arXiv:1812.09755*.

Song, L.; Fu, Y.; Zhou, P.; and Lai, K. K. 2017. Measuring national energy performance via energy trilemma index: A stochastic multicriteria acceptability analysis. *Energy Economics*, 66: 313–319.

Sukhbaatar, S.; Szlam, A.; and Fergus, R. 2016. Learning multiagent communication with backpropagation. *arXiv preprint arXiv:1605.07736*.

Sutton, R. S. 1984. *Temporal credit assignment in reinforcement learning*. Ph.D. thesis, University of Massachusetts Amherst.

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

Tan, M. 1993. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, 330–337.

Tian, Y.-P. 2012. *Frequency-domain analysis and design of distributed control systems*. John Wiley & Sons.

Yang, Y.; Luo, R.; Li, M.; Zhou, M.; Zhang, W.; and Wang, J. 2018. Mean field multi-agent reinforcement learning. In *International Conference on Machine Learning*, 5571–5580. PMLR.

Zhang, K.; Yang, Z.; Liu, H.; Zhang, T.; and Basar, T. 2018. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning*, 5872–5881. PMLR.

Zipkin, P. 2008. Old and new methods for lost-sales inventory systems. *Operations Research*, 56(5): 1256–1263.

Zipkin, P. H. 2000. *Foundations of inventory management*.