

March 2023

# Deep Reinforcement Learning Based Optimization Techniques for Energy and Socioeconomic Systems

Salman Sadiq Shuvo  
*University of South Florida*

Follow this and additional works at: <https://digitalcommons.usf.edu/etd>



Part of the [Artificial Intelligence and Robotics Commons](#), and the [Electrical and Computer Engineering Commons](#)

---

## Scholar Commons Citation

Shuvo, Salman Sadiq, "Deep Reinforcement Learning Based Optimization Techniques for Energy and Socioeconomic Systems" (2023). *USF Tampa Graduate Theses and Dissertations*.  
<https://digitalcommons.usf.edu/etd/9927>

This Dissertation is brought to you for free and open access by the USF Graduate Theses and Dissertations at Digital Commons @ University of South Florida. It has been accepted for inclusion in USF Tampa Graduate Theses and Dissertations by an authorized administrator of Digital Commons @ University of South Florida. For more information, please contact [digitalcommons@usf.edu](mailto:digitalcommons@usf.edu).

Deep Reinforcement Learning Based Optimization Techniques for Energy  
and Socioeconomic Systems

by

Salman Sadiq Shuvo

A dissertation submitted in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
Department of Electrical Engineering  
College of Engineering  
University of South Florida

Major Professor: Yasin Yilmaz, Ph.D.  
Ismail Uysal, Ph.D.  
Mahshid Rahnamay Naeini, Ph.D.  
Ankit Shah, Ph.D.  
Lu Lu, Ph.D.

Date of Approval:  
February 17, 2023

Keywords: Resource allocation, home energy management system, EV charging, sea level  
rise, hospital capacity

Copyright © 2023, Salman Sadiq Shuvo

## **Dedication**

To my dear wife, parents, and my son for their support, guidance, and companionship throughout my life. Thank you for always encouraging me and giving valuable advice whenever I needed it the most.

## Acknowledgments

First and foremost, I sincerely thank my advisor Dr. Yasin Yilmaz for his sincere, passionate, and efficient guidance and support in my research and other curricular needs. Dr. Yilmaz makes a role model through diligence, honesty, punctuality, discipline, and kindness. His prompt response and availability have helped me satisfy the research deadlines and paperwork requirements. This dissertation would not have been possible without him.

I want to thank the U.S. National Science Foundation for funding the sea level rise project.

I am grateful to my committee members - Dr. Ismayil Uysal, Dr. Mahshid Rahnamay Naeini, Dr. Ankit Shah, and Dr. Lu Lu for all of their support and invaluable feedback. I thank Dr. Christos Ferekides for his excellent leadership and support as my TA supervisor. I also want to thank Diana Hamilton and Gabriela Franco for their administrative help and the entire department for their warmth and welcoming manner.

I want to thank my lab mates Dr. Keval Doshi, Dr. Almuthanna Nassar, Shatha Abdalou, Ammar Haydari, Mahsa Mozaffari, Hamza Karim, Furkan Mumcu for the research discussions and fruitful collaborations. I am grateful to my wonderful friends Hasnat, Rasu, Rubel, Symum, Roxy, Ashu, Maliha, Shahbaz, Prayash, Naeem, Ayon, Shihab, Salekin, Parveen, Bappy, and Papul for all the sports, tours, and memorable moments.

Finally, I thank my wonderful wife, Sadia, who has stood by me through all my travails, my absences, my fits of pique and impatience. She gave me support and help, discussed ideas, and prevented several wrong turns. Along with her, I want to acknowledge my son, Zuber, who is my great source of love and relief. This dissertation remain incomplete without acknowledging the unconditional love and support that I got from my parents and sister. I am grateful to the Almighty God for enabling me to achieve my goal.

## Table of Contents

List of Tables . . . . .	iv
List of Figures . . . . .	vi
Abstract . . . . .	viii
Chapter 1: Introduction . . . . .	1
1.1 Deep Reinforcement Learning . . . . .	3
1.1.1 Value Based Methods . . . . .	5
1.1.2 Policy Based Methods . . . . .	5
1.1.3 Actor-Critic Methods . . . . .	6
1.2 Contributions and Outline of Dissertation . . . . .	6
Chapter 2: Home Energy Recommendation System (HERS): A Deep Reinforcement Learning Method Based on Residents' Feedback and Activity . . . . .	8
2.1 Introduction . . . . .	8
2.1.1 Contributions . . . . .	11
2.2 Model Development . . . . .	12
2.2.1 Environment . . . . .	12
2.2.1.1 Priority Devices (Type-1) . . . . .	13
2.2.1.2 Deferrable Devices (Type-2) . . . . .	13
2.2.1.3 Flexible Devices (Type-3) . . . . .	14
2.2.2 Action . . . . .	14
2.2.3 State . . . . .	15
2.2.4 Cost . . . . .	16
2.2.5 Next State . . . . .	17
2.3 Solution Approach . . . . .	17
2.4 Experimental Setup . . . . .	20
2.4.1 Activity Label . . . . .	20
2.4.2 Devices . . . . .	21
2.4.2.1 Central AC (Type-1) . . . . .	21
2.4.2.2 DW and WD (Type-2) . . . . .	22
2.4.2.3 EV Charging (Type-3) . . . . .	23
2.4.3 Discomfort Cost . . . . .	24
2.5 Results . . . . .	26
2.5.1 Benchmark Policies . . . . .	26
2.5.2 Scenarios . . . . .	27
2.5.2.1 Scenario 1: Unlimited Peak Demand . . . . .	27

2.5.2.2	Scenario 2: Limited Peak Demand . . . . .	28
2.5.3	Computational Statistics . . . . .	29
2.5.4	HERS Schedule Demonstration . . . . .	29
2.6	Discussion . . . . .	30
2.7	Conclusion . . . . .	31
Chapter 3: Modeling and Simulating Adaptation Strategies Against Sea-Level Rise Using Multi-Agent Deep Reinforcement Learning . . . . .		
3.1	Introduction . . . . .	33
3.2	Multi-Agent RL Framework . . . . .	35
3.2.1	Agent-based Modeling for Adaptation Strategies . . . . .	35
3.2.2	MDP Formulation . . . . .	36
3.2.3	Modeling Nature . . . . .	38
3.2.4	Modeling Stakeholders . . . . .	39
3.2.5	Optimal Policy Analysis . . . . .	41
3.2.6	Multi-Agent RL Algorithms . . . . .	46
3.3	Case Study . . . . .	49
3.3.1	Parameter Estimation . . . . .	49
3.3.2	Scenario Simulations . . . . .	55
3.4	Discussions and Conclusion . . . . .	57
Chapter 4: Multi-Objective Reinforcement Learning Based Healthcare Expansion Planning Considering Pandemic Events . . . . .		
4.1	Introduction . . . . .	59
4.2	Related Work . . . . .	62
4.3	Proposed Decision Model . . . . .	64
4.3.1	State . . . . .	66
4.3.1.1	Non-controllable States . . . . .	66
4.3.1.2	Controllable States . . . . .	67
4.3.2	Cost . . . . .	67
4.3.2.1	Expansion Cost . . . . .	67
4.3.2.2	DoS . . . . .	68
4.4	Solution Approach . . . . .	68
4.4.1	Multi-Objective Reinforcement Learning . . . . .	68
4.4.2	Deep MORL . . . . .	69
4.4.2.1	Critic Networks . . . . .	70
4.4.2.2	Actor Network . . . . .	71
4.5	Experimental Setup . . . . .	72
4.5.1	Data Generation . . . . .	73
4.5.2	Hospital Occupancy Forecasting . . . . .	74
4.5.3	Scenarios . . . . .	75
4.5.4	Objective Priority . . . . .	76
4.5.5	Neural Network Architecture and Computation Time . . . . .	76
4.6	Results . . . . .	77
4.6.1	Proposed Deep MORL-based Policy . . . . .	77

4.6.2	Benchmark Policies . . . . .	79
4.6.2.1	Target Occupancy Level Based Policy . . . . .	79
4.6.2.2	RNN Based Policy . . . . .	80
4.6.2.3	Single Objective RL Based Policy . . . . .	81
4.6.3	Comparative Analysis . . . . .	81
4.7	Discussions . . . . .	84
4.8	Conclusions . . . . .	85
Chapter 5: Demand-side and Utility-side Management Techniques for Increasing EV Charging Load . . . . .		
5.1	Introduction . . . . .	87
5.1.1	DSM for EV Charge Scheduling . . . . .	88
5.1.2	USM for XFR Maintenance . . . . .	89
5.1.3	Contributions . . . . .	90
5.2	EV Charge Scheduling for DSM . . . . .	91
5.2.1	Proposed Technique . . . . .	91
5.2.2	Consumer Incentive . . . . .	94
5.3	DRL Based XFR Replacement for USM . . . . .	95
5.3.1	Environment . . . . .	95
5.3.2	State . . . . .	97
5.3.3	Action . . . . .	97
5.3.4	Cost . . . . .	98
5.3.5	Next State . . . . .	98
5.4	Experiments . . . . .	99
5.4.1	Experimental Setup . . . . .	99
5.4.2	EV Charging (DSM) . . . . .	99
5.4.3	DRL Based Maintenance (USM) . . . . .	102
5.4.4	Comparative Analysis . . . . .	105
5.4.4.1	Fuse Blow . . . . .	105
5.4.4.2	XFR Failure . . . . .	106
5.4.4.3	Planned Maintenance . . . . .	106
5.4.4.4	Monetary Cost . . . . .	106
5.4.5	Key Insights . . . . .	106
5.5	Conclusion . . . . .	107
Chapter 6: Conclusions . . . . .		
		109
References . . . . .		
		113
Appendix A: Proof of Theorem 1 . . . . .		
		133
Appendix B: Copyright Permissions . . . . .		
		137

## List of Tables

Table 2.1	Device types . . . . .	14
Table 2.2	State input. . . . .	17
Table 2.3	ARAS activity dataset [18] . . . . .	21
Table 2.4	EV usage data generation. . . . .	23
Table 2.5	Monthly cost (\$) comparison for different policies. . . . .	29
Table 2.6	Computational statistics for the experiments. . . . .	29
Table 3.1	Model parameters. . . . .	37
Table 3.2	Relative sea level (mm) for different scenarios for St. Petersburg. . .	51
Table 3.3	Generalized Pareto parameters for Pinellas county. . . . .	53
Table 3.4	Action and cost parameters for Pinellas county. . . . .	54
Table 4.1	Computational details for the experiments. . . . .	77
Table 4.2	Parameter, cost, and DoS comparison among the policies for different objective priorities (non-pandemic scenario). . . . .	79
Table 4.3	Expansion cost, DoS, and selected sequential actions for each policy over a 30-year period for equal priority on the two objectives. . .	84
Table 5.1	Feeder data (Numbers in parentheses indicate the number of XFRs serving that many homes). . . . .	100
Table 5.2	Utility company’s equipment and labor cost for different maintenance types. . . . .	100
Table 5.3	Yearly cost (\$) for XFR-4 to customers and the utility for different charging techniques. . . . .	102
Table 5.4	Computational details for the experiments. . . . .	103
Table 5.5	Cumulative EV charging and maintenance cost for all 232 XFRs over a 30-year timeline. . . . .	104



Table 5.6	Comparison among XFR maintenance policies for uncoordinated charging (top) and proposed DSM (bottom) in terms of cumulative cost (left), power outage (middle), and XFR failure (left). . . . .	107
-----------	---	-----

## List of Figures

Figure 1.1	Reinforcement learning framework. . . . .	4
Figure 2.1	Proposed MDP model. . . . .	12
Figure 2.2	Action flow chart for active devices at each time $t$ . . . . .	15
Figure 2.3	Advantage Actor-Critic (A2C) network. . . . .	19
Figure 2.4	Daily cumulative cost in scenario 1 for devices: AC (top left), DW (top right), WD (bottom left), and EV (bottom right) for 1-month duration. . . . .	25
Figure 2.5	Daily cumulative cost comparison for all devices among the considered policies for scenario 1 (left) and scenario 2 (right). . . . .	28
Figure 2.6	Convergence of the proposed deep RL algorithm for HERS for scenario-1 total cost. . . . .	30
Figure 2.7	HERS scheduling results for a day under scenario-2 (peak demand limit 10 kW). . . . .	31
Figure 3.1	Proposed multi-agent MDP framework. . . . .	36
Figure 3.2	Expected total costs as a function of sea level for an example case with $A_G = 3$ . . . . .	44
Figure 3.3	Unified A2C structure for all three agents. . . . .	47
Figure 3.4	Separate A2C structure for each agent. . . . .	48
Figure 3.5	Average episodic total cost of all agents in the separate A2C policy for the high SLR scenario. . . . .	50
Figure 3.6	SLR projections by NOAA [137] (solid lines) and our fittings (dashed lines) for relative sea level change for St. Petersburg, FL. . . . .	52
Figure 3.7	Expected episodic cost under different policies for high SLR scenario. . . . .	55

Figure 3.8	100-year total cost for the intermediate-low, intermediate, high scenarios of SLR, and their average. . . . .	56
Figure 3.9	Yearly total cost under different policies for the high SLR scenario. . . . .	57
Figure 4.1	Proposed multi-objective MDP (MOMDP) model. . . . .	65
Figure 4.2	Proposed multi-objective A2C architecture. . . . .	70
Figure 4.3	Map of the 11 health regions of Florida for the case study. . . . .	73
Figure 4.4	Average accuracy for different regression models in predicting hospital admission for weekdays and weekends based on data from [1]. . . . .	75
Figure 4.5	Neural network architecture for the proposed multi-objective deep RL-based policy. . . . .	77
Figure 4.6	Episodic (30-year) total cost and DoS for healthcare authority under different objective priorities for the policies. . . . .	78
Figure 4.7	Episodic (30-year) total cost (solid lines) and DoS (dashed lines) for healthcare authority under different objective priorities for the policies for non-pandemic (left) and pandemic (right) scenarios. . . . .	82
Figure 4.8	Trade-off solutions for each policy for the non-pandemic (left) and pandemic (right) scenarios with equal (moderate) objective priority. . . . .	83
Figure 5.1	Proposed DSM flowchart (left) and DRL model for USM (right). . . . .	92
Figure 5.2	Comparison between the proposed utility-driven DSM and uncoordinated EV charging in terms of hourly load for XFR-4 for the first week of Year-1, Year-10, Year-20, and Year-30 (from left to right). . . . .	101
Figure 5.3	Convergence of the DRL based maintenance policy (with proposed DSM). . . . .	103

## Abstract

Optimization, which refers to making the best or most out of a system, is critical for an organization’s strategic planning. Optimization theories and techniques aim to find the optimal solution that maximizes/minimizes the values of an objective function within a set of constraints. Deep Reinforcement Learning (DRL) is a popular Machine Learning technique for optimization and resource allocation tasks. Unlike the supervised ML that trains on labeled data, DRL techniques require a simulated environment to capture the stochasticity of real-world complex systems. This uncertainty in future transitions makes the planning authorities doubt real-world implementation success. Furthermore, the DRL methods have limitations for different application environments; slow convergence, unstable learning, and being stuck in local optima are a few of them.

We address these challenges in our environmental, healthcare, and energy systems projects by carefully (1) modeling the system dynamics we achieved through research and collaboration with domain experts and (2) state-of-the-art DRL techniques for experimental analysis. Our experimental results and comparative analysis with the other optimization methods demonstrate the efficacy of DRL-based techniques. The success lies in appropriately modeling the critical decision-making features, reward function, and state transitions. In the process, we have developed novel DRL (Multi-agent and Multi-objective) algorithms.

## Chapter 1: Introduction

Recent advances in neural network-based Deep Reinforcement Learning (DRL) algorithms lead to widespread applications, including gaming [85], robotics [100], finance [72], energy systems [124], transportation [60], communications [90], environmental systems [119], and healthcare systems [122]. The success of DRL algorithms was first demonstrated in gaming; where the Deep Q Network (DQN) [85] algorithm outperformed other methods and human experts in many Atari games. Subsequently, these atari games provided experimental environments for new algorithms to demonstrate their performance and computational efficiency. Due to its adaptive learning capacity, DRL is very successful in robotics tasks [100]. The recent revolutionary chatbot 'ChatGPT' utilizes reinforcement learning from human feedback [100] to fine-tune the model.

DRL is popular in resource optimization tasks for an organization's strategic planning. Policymakers, planners, and management authorities can benefit from DRL-based optimization techniques. An extensive review of RL for Demand Side Management (DSM) of electricity in [146], showcases the suitability of RL for DSM techniques. Berlink et al. [28] were among the first to investigate RL-based DSM techniques for a smart home. The work [142] utilizes the inherent adaptability in deep RL algorithms by maintaining thermal comfort and optimal air quality while minimizing electricity usage. A large-scale home energy management system is proposed in [154] using a multi-agent deep RL framework. The applications of DRL-based methods in healthcare have provided adverse outcome predictions [156]. The works in [121, 120] utilized DRL to determine the optimal size of hospital capacity augmentation.

Resource optimization is vital in environmental, healthcare, energy systems, and other infrastructural planning domains. These complex systems often include multiple stakeholders and aim to satisfy multiple objectives. Deep Reinforcement Learning (DRL) is a popular Machine Learning technique for optimization and resource allocation tasks. This dissertation showcases projects on DRL-based optimization techniques for sea level rise (SLR) adaptation, healthcare expansion, smart home energy management, and electric vehicle (EV) charging management. These projects focus on appropriate system dynamics modeling and state-of-the-art DRL techniques for experimental analysis.

Our Home Energy Recommendation System (HERS) project proposes a DRL method for managing smart devices in a home to optimize electricity costs and residents' comfort. We incorporate human feedback in the objective function and human activity data in the DRL state definition to enhance energy optimization performance. The SLR project aims to solve a community-wide multi-stakeholder (government, residents, and businesses) problem. Simulating the local socioeconomic system around SLR, including the interactions between essential stakeholders and nature, can effectively facilitate evaluating different adaptation strategies and planning the best strategy for the local community. A well-developed and experimented approach helps the policymaker (the government) encourage other stakeholders (residents and businesses) to achieve collaborative success. The healthcare expansion planning project presents a multi-objective reinforcement learning (MORL) based solution for minimizing capacity expansion cost and minimizing the number of denial of service (DoS) for patients seeking hospital admission simultaneously for pandemic and non-pandemic scenarios. Our model provides a simple and intuitive way to set the balance between these two objectives by only determining their priority percentages, making it suitable for policymakers with different capabilities, preferences, and needs. Our fourth project aims to provide electric utility companies with scenario-based plans to cope with the varying EV penetration across locations and time. The contributions of this project are twofold. First, we propose a customer feedback-based EV charging scheduling to simultaneously minimize the peak load

for the distribution transformer (XFR) and satisfy the customer needs. Second, we present a DRL method for XFR maintenance, focusing on the XFR’s effective age and loading to periodically choose the best candidate XFR for replacement. The experimental results across all four projects demonstrate the efficacy of DRL-based techniques for challenging real-world optimization tasks.

## 1.1 Deep Reinforcement Learning

Reinforcement Learning (RL) is a machine learning (ML) based optimization technique. Optimization tasks often require making sequential decisions. Markov Decision Process (MDP), which is founded on Markovian or memorylessness property, is an effective tool for modeling optimal sequential decision-making problems. MDP provides a tractable mathematical formulation to model decision-making tasks in situations where outcomes are partly random and partly regulated by the agent actions [65, 25, 46]. RL provides a suitable theoretical framework for solving MDPs. As seen in Fig. 1.1, the RL agent interacts with the environment in state  $S_t$  by taking action  $A_t$  at each time and receiving a cost/reward  $R_t$  from the environment in return. The agent’s objective is to minimize/maximize an expected sum of costs/rewards over time by choosing optimal actions from an action set. At each time, as a result of the agent’s action, the system moves to a new state  $S_{t+1}$  according to a probability distribution. The optimal policy for deciding on actions maps system states to actions, i.e., determines which action to take in which state [136].

The RL agent aims to maximize the discounted cumulative reward in  $T$  time steps:

$$R_T = \sum_{t=0}^T \lambda^t R_t,$$

where  $\lambda$  is the discount factor, a critical parameter that represents the weight of future cost in current decision. In the MDP framework, the objective of agent is to maximize the expected total reward  $\mathbb{E}[R_T]$  in  $T$  time steps by following an optimal policy. Central to MDP is the

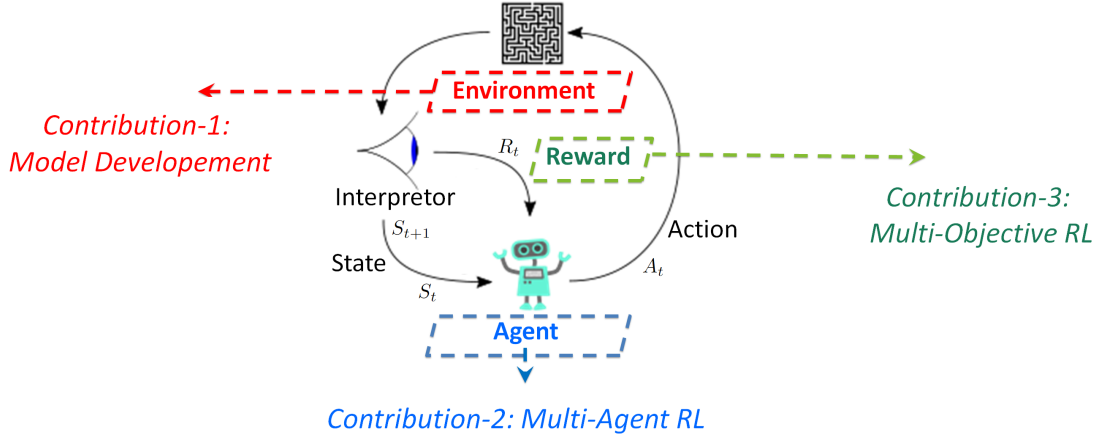


Figure 1.1: Reinforcement learning framework.

optimal value function

$$V(S_t) = \max_{\{A_t\}} \mathbb{E}[R_t | \{A_t\}],$$

which gives the maximum expected total reward possible at each state  $S_t$ , denoted as the optimal value of that state, by choosing the best action policy  $\{A_t\}$ . To find the optimal policy, the Bellman equation

$$V(S_t) = \max_{A_t} \mathbb{E}[R_t + \lambda V(S_{t+1}) | A_t] \quad (1.1)$$

provides a recursive approach by focusing on finding the optimal action  $A_t$  at each time step using the successor state value, instead of trying to find the entire policy at once.

RL provides a data-driven approach to solving in Eq. (1.1), which is of critical importance for the high-dimensional state and/or action spaces where exact dynamic programming solutions are not feasible [136]. Specifically, Deep Reinforcement learning (DRL) methods utilize deep neural networks to handle significantly high-dimensional problems, which are too complex for traditional tabular RL methods. As real world problems are high-dimensional inherently, the DRL methods have become widespread popular since its inception in 2013 by the landmark publication [84].



### 1.1.1 Value Based Methods

Value base methods aim to learn the value of a state from Eq. (1.1). To explore the environment sufficiently, the agent needs to take a mix of optimal (exploitation) and non-optimal (exploration) actions. In this regard, learning the value of the state-action pair  $Q(S_t, A_t)$  ( Q-value) is more appropriate.

$$Q(S_t, A_t) = \mathbb{E}[R_t + \lambda \max_{A_{t+1}} Q(S_t, A_{t+1} | S_t, A_t)].$$

The deep Q network (DQN) algorithm [85], which leverages a deep neural network to estimate the optimal action-value function is the most popular choice for deep RL. Many researches have made further improvements on the DQN framework to provide state-of-the-art solutions. Some of the popular extensions are Double DQN [85], Dueling-DQN [143], Prioritized Experience Replay [149], and Rainbow [62].

### 1.1.2 Policy Based Methods

Policy gradient based methods aims to learn the policy for a given state bypassing the need to learn the value function stated in Eq. (1.1). The REINFORCE [159] algorithm outputs the probability for each action through a softmax function. To that end, it finds the gradient of expected return  $J(\pi_\phi)$  of the policy  $\pi_\phi$  with respect to the weights  $\phi$  of the neural network through the following equation

$$\nabla_\phi J(\pi_\phi) = \mathbb{E}_{\pi_\phi}[\nabla_\phi \log(\pi_\phi(A_t|S_t))G_t], \tag{1.2}$$

where  $G_t$  is the expected estimated return from state  $S_t$ . As  $G_t$  is estimated from experience, it injects variance in learning the policy function and makes convergence challenging.

### 1.1.3 Actor-Critic Methods

Actor-Critic methods reduce the variance of the REINFORCE [159] algorithm by Using two neural networks (actor and critic). The critic (value) network evaluates the value function to estimate the advantage function

$$\mathcal{A}(S_t; A_t) = \gamma V(S_{t+1}; \theta) - V(S_t; \theta). \quad (1.3)$$

where  $\theta$  is the weights of the critic network. The actor network utilizes this advantage function for finding the gradient of expected return  $J(\pi_\phi)$

$$\nabla_\phi J(\pi_\phi) = \mathbb{E}_{\pi_\phi} [\nabla_\phi \log(\pi_\phi(A_t|S_t)) \mathcal{A}(S_t; A_t)]$$

Many works have made significant improvement to the Actor-Critic algorithms. The notable works are DDPG [79], TRPO [113], PPO [114], A3C [83], and TD3 [50].

## 1.2 Contributions and Outline of Dissertation

The simulation environment development, multi-agent and multi-objective DRL algorithms to optimize the considered complex systems are the significant contributions of this dissertation, as marked in Fig. 1.1. The following chapters elaborate on these contributions

- Chapter 2: The Home energy management project incorporates *direct human feedback for discomfort* in the objective function through residents' manual overrides to the recommended device operations to learn residents' preferences; and uses *resident activities* in the state definition to learn device usage patterns.
- Chapter 3: The SLR project develops a Multi-agent RL framework for the SLR stakeholders (government, residents, businesses) and theoretically shows that the stakeholders should base their investment decision on the observed sea level instead of the incurred cost from nature.

- Chapter 4: The Health capacity expansion project develops a novel deep Multi-objective RL (MORL) algorithm based on the actor-critic framework. An extensive case study is performed for the state of Florida using real data to evaluate the proposed MORL approach.
- Chapter 5: The EV charging project provides the first comprehensive study of the problem of increasing stress on the distribution XFRs due to EV charging. Specifically, a combination of novel DSM and USM techniques is proposed for flattening the load curve and making timely maintenance of the distribution XFRs, respectively.

## Chapter 2: Home Energy Recommendation System (HERS): A Deep Reinforcement Learning Method Based on Residents' Feedback and Activity

### 2.1 Introduction

<sup>1</sup>Smart home systems can enhance human comfort and optimize electricity usage in an automated setup. While many devices have included sensor-based control for a long time, such as microwave ovens, air conditioning, etc., with the Internet of Things (IoT) revolution [106], many other smart appliances are entering our homes. Most of the devices will soon have such intelligence that will unlock the true potential of the smart home concept. Specifically, recent smart Home Energy Management (HEM) technologies can leverage state-of-the-art artificial intelligence (AI) techniques. As a result, residents can enjoy all the comfort smart devices offer according to their preferences in an automated way. In addition to personalized comfort, the HEM system can significantly reduce the electricity cost and flatten the demand curve by scheduling some devices to run during off-peak hours.

Utility companies employ Demand Response (DR) based techniques to encourage customers to shift their load to off-peak hours [129]. It serves two purposes: avoiding electricity purchases from expensive peaking power plants and keeping the system's maximum demand at check to avoid capacity expansion costs. They provide time-based pricing schemes for the customers, known as Time of Use (TOU) [107], such as real-time pricing, critical peak pricing, etc. Numerous researches have proposed appliance scheduling techniques for HEM systems [160] to capitalize TOU tariffs. Such Demand Side Management (DSM) techniques aim to modify the consumer's energy activities, e.g., shifting customers' electricity usage to-

---

<sup>1</sup>Portions of this chapter were published in IEEE Transactions on Smart Grid [126]. Copyright permissions from the publishers are included in Appendix B.

wards off-peak hours [132]. For instance, a hierarchical HEM system within a home microgrid is proposed in [80] that integrates photovoltaic (PV) energy into day-ahead load scheduling and aims to reduce the monthly peak demand and peak demand charges<sup>2</sup>. A state-space approximate dynamic programming (SS-ADP) approach is proposed in [158] to provide a fast real-time control strategy under uncertainty using the Bellman optimality condition. The work in [125] includes consumer input in their proposed EV charge scheduling technique. The uncertainties in electricity usage of smart building HEM as a nonlinear optimization problem is addressed in [117]. A microgrid where the users minimize cost by trading energy between each other before buying from the grid is presented in [154], where PV energy, home battery, and EV battery serve as intermittent sources.

The majority of the DSM techniques for HEM are based on a rule-based schedule for device usage, undermining consumers' comfort. Rule-based scheduling often suffers from the randomness inherent in human preference, weather, and other interventions, especially in realistic scenarios with multiple residents and multiple appliances. To this end, the works in [77, 93, 118] aim to dissolve the rigid scheduling of devices by including distributed energy generation and distributed energy storage devices in their HEM system. To realize the far-reaching potential of smart home technology, researchers have opted from rule based approaches to recent data-driven machine learning techniques for DSM.

Electricity consumption patterns are evolving with the fast-improving smart device technologies, which requires adaptability in HEM for scheduling devices. Reinforcement learning (RL) techniques are typically preferred for their data-driven online decision-making capability. Recent advances in neural network-based deep RL algorithms lead to widespread applications, including gaming [85], finance [72], energy systems [124], transportation [60], communications [90], environmental systems [119], and healthcare systems [122]. An extensive review of RL for DSM in [146], showcases the suitability of RL for DSM techniques. Berlink et al. [28] were among the first to investigate RL-based DSM techniques for a smart

---

<sup>2</sup>Not every utility charges for peak demand.

home. The work [142] utilizes the inherent adaptability in deep RL algorithms by maintaining thermal comfort and optimal air quality while minimizing electricity usage. A large-scale HEM is proposed in [154] using a multi-agent deep RL framework.

The authors, in their review of RL for demand response [146], emphasize the importance of incorporating human feedback in RL-based DSM techniques. Pilloni et al. [96] propose a smart HEM system in terms of the quality of experience, which depends on the information of consumers' discontent for changing home devices' operations. To replicate human feedback, they surveyed 427 people to generate residents' annoyance profiles for delayed scheduling of different appliances. Then, they incorporate a cost apart from the electricity price based on the annoyance levels from these profiles. In their following research [82], they used sensor-based activity recognition to predict future activities for appliance scheduling. The authors in [153] define human dissatisfaction by the difference between the maximum power rating and the delivered power rating of a device, an oversimplified way of representing human feedback for their RL-based HEM system. Murad et al. [71] calculate dissatisfaction if HEM turns off a device using an equation with different priority factors for different devices. Several other works, e.g., [157, 22], follow a similar approach to estimate discomfort cost rather than using actual feedback from residents. All these techniques lack adaptability to consumer preference, i.e., they may work well for certain types of users, but they are not general enough to ensure user convenience. Park et al. [94] provide theory and implementation for adaptive and occupant-centered lighting optimization in an office setup. They interpret switching on and off the lights by office employees as human feedback. This work has successfully incorporated human feedback for their RL algorithm; however, their scope is limited to lighting. Hence, the necessity for a human feedback-based HEM system still remains open.

The work in [78] proposes a deep sequential learning-based human activity recognition in smart homes. The benefits of labeled activity to analyze and assess the smart home residents' physical and psychological health has been reviewed in [38]. Chen et al. [35] analyze behavior patterns to predict energy consumption profile. Since the smart home

concept has the inherent capability of activity labeling, including the activity data as a feature for the DSM technique can greatly facilitate the RL agent’s learning capacity. The work [108] reviews sensor-based activity recognition techniques to implement in a smart home setup. Given the technology, our work includes human activity labels in the RL state definition for the first time to the best of our knowledge.

Although the smart home concept is originally introduced for the residents’ benefit, their comfort is often ignored in many existing methods. In this work, we propose a deep RL method that takes the residents’ feedback as a reward factor, apart from electricity prices and device status. We consider resident activities as part of the system state to better understand human comfort and feedback. Our work incorporates residents’ feedback every time they override the HEM system’s commands, a practical and novel way of extending the success of recommender systems (e.g., movie, book, shopping, video) to HEM. Recommender systems learn from customer usage patterns to recommend items/services [17]. A similar approach can be integrated into a HEM system by accommodating human input in a meaningful way.

### 2.1.1 Contributions

Our contributions lie in addressing two challenges in RL for HEM. Specifically,

- We propose a novel home energy recommender system (HERS) based on a Markov decision process (MDP) formulation and a deep RL solution to jointly minimize the electricity consumption cost and discomfort to the residents;
- HERS incorporates *direct human feedback for discomfort* in the objective function through residents’ manual overrides to the recommended device operations to learn residents’ preferences; and
- HERS uses *resident activities* in the state definition to learn device usage patterns.

We evaluate the performance of the proposed HERS method by comparing with a manually controlled, two rule-based[118, 96], and an RL-based approach [153].

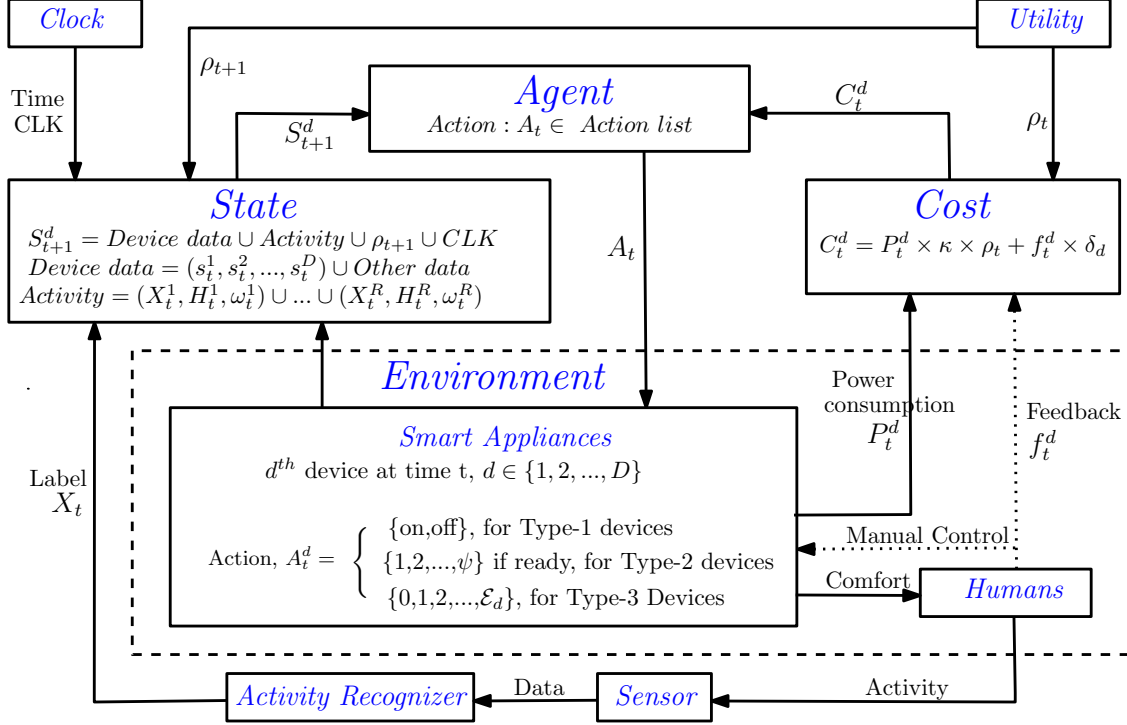


Figure 2.1: Proposed MDP model.

The remainder of the chapter is organized as follows. The MDP model is formulated in Section 2.2, and the deep RL algorithm for the optimal policy is given in Section 2.3. The experimental setup is presented in Section 2.4. Results are discussed in Section 2.5. Finally, after the key features of the chapter and the future research scope are discussed in Section 2.6, the chapter is concluded in Section 2.7.

## 2.2 Model Development

We propose an MDP framework shown in Fig. 2.1, where the smart home device manager is the MDP agent, called HERS.

### 2.2.1 Environment

The residents, activity recognizers, and devices form the MDP environment. The homes can be of different sizes, with multiple residents living in them. We assume access to the utility company's real-time pricing scheme,  $\rho_t$  (\$/kWh at time  $t$ ), and activity recognition



through multiple sensors placed throughout the home. Affordable and reliable activity recognition from sensor data has been studied by several works [78, 38, 35], which is out of the scope of this chapter. We assume the presence of an activity recognition set up, which provides the activity label  $X_t^1, X_t^2, \dots, X_t^R$  for all  $R$  residents at home.

HERS employs different methods to operate each of the  $d \in \{1, 2, \dots, D\}$  smart devices that we divide into three categories, as shown in Table 2.1. When switched on by a human or sensor, the device goes into the active status and will be considered for decision-making only during active status.

### 2.2.1.1 *Priority Devices (Type-1)*

These devices provide essential comfort to the residents, and they are not available for deferring. HERS can keep the active devices off intermittently without compromising the devices' functionality. Regular lights, TV, CCTV camera, alarm system, and air conditioner (AC) are examples of this type of appliance. Choosing the relevant data for the MDP state is a challenge for this task. For instance, if the resident is browsing the internet while the TV is on, turning it off may create discomfort. However, if the resident goes to sleep, keeping the TV on, turning it off may reduce electricity costs without compromising comfort. AC is the heaviest load for this device type, hence we focus on it in our experiments.

### 2.2.1.2 *Deferrable Devices (Type-2)*

These devices can be scheduled later to off-peak hours, reducing electricity cost and maintaining the peak demand lower than the threshold (if any). Dish Washer (DW) and Washer & Dryer (WD) fall in this category. These devices typically can evade human discomfort if it completes the task before the subsequent activation by the residents. So, the dynamic electricity price  $\rho_t$  and activation time are critical features for scheduling the deferrable devices.

Table 2.1: Device types

Devices	Priority (Type-1)	Deferrable (Type-2)	Flexible (Type-3)
Device Properties	Not deferrable, Rigid power consumption	Deferrable, Rigid power consumption	Deferrable, Flexible power consumption
Operational Objective	Minimize idle usage by turning on/off intermittently	Operate during low $\rho_t$	Charge highly at low $\rho_t$
Device Ready Status	Turned on by resident/sensor	Turned on by resident	Connected to charger
Action Selection	Every time step during active period	Once every activation	Every time step during active period
Actions	$A_t^d \in \{on, off\}$	$A_t^d \in \{0, 1, 2, \dots, \psi_d\}$	$A_t^d \in \{0, 1, 2, \dots, \mathcal{E}_d\}$
Device Examples	Regular lights, TV, AC	Sprinkler, DW, WD	EV, cell phone, laptop chargers

### 2.2.1.3 Flexible Devices (Type-3)

These devices are flexible in terms of time scheduling and power level. EV, cell phone, and laptop chargers are examples of these types of devices. These devices can consume different power levels  $\{0, 1, 2, \dots, \mathcal{E}_d\}$ , which changes their battery charge level  $\beta_t^d$ . Residents' activity patterns and  $\beta_t^d$  are important features in HERS for these devices.

### 2.2.2 Action

Our MDP model in Fig. 2.1 begins with the agent selecting actions  $A_t = (A_t^1 \cup A_t^2 \cup \dots \cup A_t^{D_{act}})$  about setting the operation mode for each of the smart devices in active status  $D_{act}$  ( $\leq D = m + n + o$ ). So, the total number of possible actions are

$$\underbrace{2^m}_{m \text{ Type-1}} \times \underbrace{(\psi_1 + 1) \times (\psi_2 + 1) \times \dots \times (\psi_n + 1)}_{n \text{ Type-2}} \times \underbrace{(\mathcal{E}_1 + 1) \times (\mathcal{E}_2 + 1) \times \dots \times (\mathcal{E}_o + 1)}_{o \text{ Type-3 Devices}} \quad (2.1)$$

where,  $m$  is the total number of Type-1 devices.  $\psi_1, \psi_2, \dots, \psi_n$  are scheduling time ranges for the  $n$  type-2 devices, and  $\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_o$  are charging power levels for the  $o$  type-3 devices. Fig. 2.2 shows the action flowchart for each type of devices at each time  $t$ .

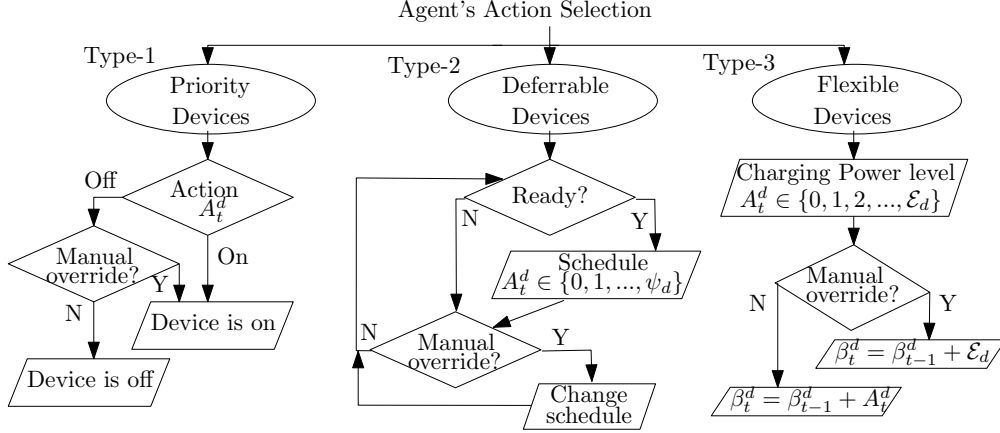


Figure 2.2: Action flow chart for active devices at each time  $t$ .

For the type-1 devices, there are two actions possible (on/off) for the device. For  $A_t^d = \text{off}$ , the agent changes its action if the residents' perform manual override. HERS schedules a Type-2 device when it is ready for a new run. No further decision is made until the current operation is finished, either scheduled or manually overridden. The device becomes ready again when the resident activates it for a new run. For Type-3 devices, HERS decides on a charge level  $A_t^d$  for each time  $t$ . If there is any manual override, then the charge level is set to full capacity  $\mathcal{E}_d$  to finish charging as soon as possible. Every manual override causes the discomfort cost through feedback  $f_t^d = 1$  to the RL agent for the corresponding device.

The actual number of possible actions will typically be smaller than Eq. 2.1 during a time step due to inactive devices. For example, when the residents are not at home, the AC will remain off and will not be considered for the agent's action. Similarly, the idle status of many devices can be determined to limit the number of actions. Furthermore, a deferrable device only remains active for one time step when HERS schedules its operation.

### 2.2.3 State

The MDP agent takes action based on the environment state. Appropriate design of the state is fundamental to the success of the MDP model. As the devices provide comfort to the residents, we hypothesize their activity data to be critical to define the states. An activity recognition system uses various home sensor data to label the residents' activity  $X_t$ . Apart

from the activity, the real-time electricity price  $\rho_t$  and clock time of the day (CLK) are other essential features that we include in the state definition, as shown in Table 2.2. The state at time  $t$  is defined as:

$$S_t^d = \text{Device data} \cup \text{Activity} \cup \rho_t \cup \text{CLK},$$

where  $\text{Activity} = (X_t^1, H_t^1, \omega_t^1) \cup \dots \cup (X_t^R, H_t^R, \omega_t^R)$  includes the current activity  $X_t^i$ , previous activities  $H_t^i$ , and duration of the current activity  $\omega_t^i$  for all  $R$  residents. Device data includes information like how long ago the device was activated, the number of dirty dishes or clothes for the Dishwasher and Washer Dryer, the charge level of the type-3 devices, that can be included in the state definition, as shown in Table 2.2. In practice, activity labels can be generated from activity recognition sensors as discussed in [78].

#### 2.2.4 Cost

The MDP agent tries to maximize a reward or minimize a cost by taking optimal actions for a given state. For instance, the RL-based Youtube video recommendation systems are rewarded when the user opens a recommended video[36]. Similarly, HERS receives cost (negative reward) whenever a resident is not happy with the selected action and changes the mode of a device. This human feedback  $f_t^d = 1$  is interpreted as discomfort and converted to a cost to the MDP agent through separate cost coefficients  $\delta_d$  for each device  $d$  for each manual override. The devices' operations are meant for human comfort, so HERS' objective is to minimize discomfort.

The total cost for the MDP agent is the sum of energy cost and human discomfort cost. The utility informs the agent of the electricity price for the current time step  $\rho_t$ , and future time step  $\rho_{t+1}$ . The energy usage at time  $t$  is obtained from the smart device's power consumption  $P_t^d$  and used to calculate the total cost for each active device for time step  $t$  as

$$C_t^d = P_t^d \times \kappa \times \rho_t + f_t^d \times \delta_d, \quad (2.2)$$

Table 2.2: State input.

Input	AC	DW	WD	EV
Activity	✓	✓	✓	✓
Clock time	✓	✓	✓	✓
Electricity price, $\rho_t$	✓	✓	✓	✓
Device status	✓	✓	✓	✓
Device activation duration	✓	✓	✓	✓
Battery charge level	×	×	×	✓
Travel upcoming	×	×	×	✓

where  $\kappa$  is the unit step time in hours. Cost coefficient  $\delta_d$  for each device is a critical modeling parameter that converts discomfort into monetary value.  $f_t^d$  represents the discomfort feedback of the residents, where 0 and 1 respectively indicates no override or override. The goal of the MDP agent is to minimize the following discounted cumulative cost for each device in  $T$  time steps:

$$C_T^d = \sum_{t=0}^T \lambda^t C_t^d, \tag{2.3}$$

where  $\lambda \in [0, 1]$  is the discount factor for future decisions.

### 2.2.5 Next State

At the end of a time step, the device state  $S_t^d$  changes according to the action  $A_t$ ; however, human activity data, electricity price data, etc., change stochastically. These features define the next state  $S_{t+1}^d$ , and the dynamic system moves to the next time step for the agent to act. These transitions satisfy the Markovian property of the MDP framework.

## 2.3 Solution Approach

HERS employs one separate MDP agent for each of the  $D$  devices to minimize the discounted total costs  $C_T^d$  in Eq. (2.3). To achieve the optimal policy  $\arg \min_{\{A_t^d\}} C_T^d$ , we need to solve the following Bellman equation. We drop the device index from here on for

---

**Algorithm 2.1** A2C algorithm for each device in HERS

---

*Input:* discount factor  $\lambda$ , discomfort cost coefficient  $\{\delta_d\}$

*Initialize:* Actor network with random weights  $\phi$  and critic network with random weights  $\theta$

**for** episode = 1, 2, ...,  $E$  **do**

**for**  $t = 1, 2, \dots, T$  **do**

    Collect activity data from Activity Recognizer, real-time electricity price  $\rho_t$  from Utility.

    Select action  $A_t^d$  using Actor Network (Fig. 2.3).

    Execute action  $A_t^d$  and observe human discomfort feedback  $f_t^d$ .

    Calculate cost  $C_t^d$  using Eq. (2.2).

    Store transitions  $(S_t^d, A_t^d, C_t^d, S_{t+1}^d)$ .

**end for**

  Update actor network  $\phi$  via Eq. (2.4).

  Update critic network  $\theta$  through back propagation.

**end for**

---

brevity. The agent’s value function at time step  $t$  is

$$V(S_t) = \min_{A_t} \left\{ \mathbb{E} [C_t + \gamma V(S_{t+1})] \right\}.$$

The above equation presents a solution dilemma in prioritizing between the immediate cost  $C_t$  and future expected cost  $\gamma V(S_{t+1})$ . Since the agent’s action changes the next state of the devices, the future discounted cost through the value function of the next state  $V(S_{t+1})$  depends on the action of the agent. Since in high-dimensional problems like the considered one here, it is not feasible to compute the expected future cost explicitly and find the value function for each possible state, deep neural networks are typically used in the modern practice of RL (known as deep RL) to learn the optimal policy of actions either directly (policy-based methods) or through the value function (value-based methods).

The Advantage Actor-Critic (A2C) algorithm, which is a hybrid (both value-based and policy-based) adaptation of policy gradient-based algorithm REINFORCE [133], is a popular choice for continuous state space, e.g., electricity price and battery charge level in our setup. We also considered using Deep Q Network (DQN), another popular deep RL algorithm, but

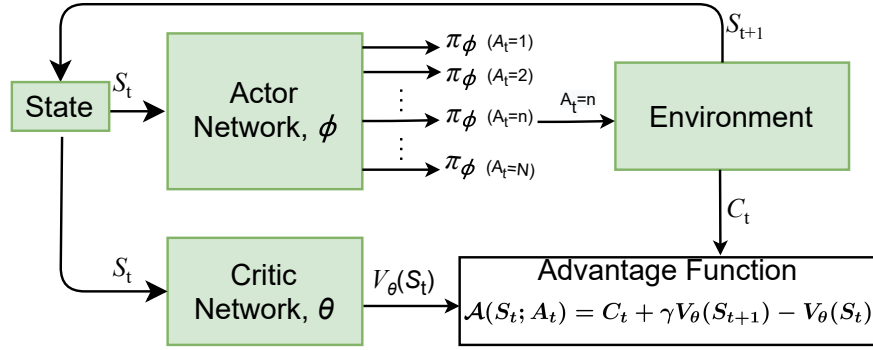


Figure 2.3: Advantage Actor-Critic (A2C) network.

A2C performed better in the proposed state space, as expected. A2C uses the advantage functions for policy update, which reduces the REINFORCE algorithm’s variance as shown in Fig. 2.3.

The actor network, also known as the policy network, outputs probability for each action value  $\pi_\phi(A_t)$  through a softmax function. Then the agent samples an action  $A_t$  based on the policy  $\pi_\phi$  and the environment moves to the next state  $S_{t+1}$  and provides the immediate cost  $C_t$ . The actor network aims to find the gradient of expected return  $J(\pi_\phi)$  of the policy  $\pi_\phi$  with respect to the weights  $\phi$  of the neural network through the following equation:

$$\nabla_\phi J(\pi_\phi) = \mathbb{E}_{\pi_\phi} [\nabla_\phi \log(\pi_\phi(A_t|S_t)) \mathcal{A}(S_t; A_t)], \quad (2.4)$$

where the advantage function  $\mathcal{A}$  is given by

$$\mathcal{A}(S_t; A_t) = C_t + \gamma V_\theta(S_{t+1}) - V_\theta(S_t). \quad (2.5)$$

The critic-network learns the value function  $V_\theta(S_t)$  for each state. It uses the advantage function  $\mathcal{A}$  as the critic loss to update its network parameters  $\theta$  through back propagation. A pseudo code for the proposed A2C algorithm is given in Algorithm 2.1.

## 2.4 Experimental Setup

The ideal experimental setup would be implementing the HERS algorithm in an existing smart home. However, a fully equipped smart home capable of taking human feedback is yet to be available. We will hypothetically generate human feedback and interactions with the devices based on the residents’ activity data. HERS select different features for operating different devices, as shown in Table 2.2. We include clock time in minutes and real-time electricity price  $\rho_t$  as the common states for all the devices. The New York Independent System Operator (NYISO) provides real-time electricity prices; we use Long Island, NY prices for March 13 and 19, 2021 as the electricity price respectively for weekends and weekdays in our simulation [91]. We find that  $\kappa= 0.25$  hour (15 minutes) is suitable for the experimental setup.

### 2.4.1 Activity Label

For residents’ indoor activity data, we use the ARAS dataset [18]. The attributes of the dataset for the two homes are shown in Table 2.3. The dataset contains 27 types of activities labeled by sensors and validated by the residents. This dataset is comparatively newer and has more activity types than other datasets in the literature. We choose House B for the experiments. HERS takes the current activity label, duration, and the last activity label for each resident (6 inputs in total for the two residents in the house). Apart from providing the dataset, [18] also gives a guideline about the sensors required for activity recognition. To collect the activity data, they used a total of 20 binary sensors of 7 types: (1) force sensor, (2) photocell, (3) contact sensors, (4) proximity sensors, (5) sonar distance sensors, (6) temperature sensors, and (7) infrared sensors.



Table 2.3: ARAS activity dataset [18]

	House A	House B
Size	538ft <sup>2</sup>	969ft <sup>2</sup>
Layout	One bedroom, one living room one kitchen, one bathroom	Two bedrooms, one living room one kitchen, one bathroom
Residents	2 males at their twenties	Married couple at their thirties
Duration	30 days	30 days
Published in	2013	2013
Labelled Activities: (1) Going out, (2-4) Cooking (breakfast, lunch, dinner), (5-7) Having breakfast, lunch, dinner, (8) Washing dishes, (9) Having snack, (10) Sleeping, (11) Watching TV, (12) Studying, (13) Bath, (14) Toileting, (15) Napping, (16) Using Internet, (17) Reading book, (18) Laundry, (19) Shaving, (20) Brushing teeth, (21) Phone conversation, (22) Listening music, (23) Cleaning, (24) Conversation, (25) Having guest, (26) Changing clothes, (27) Other		

#### 2.4.2 Devices

HERS can provide optimal control for all the smart devices in a home. However, we limit our case study to high power loads that renders significant energy cost. Specifically, we choose the following four devices for our experiments.

##### 2.4.2.1 Central AC (Type-1)

We estimate a 12000 BTU (3.5 kWh) AC capacity for the 90 m<sup>2</sup> (968.75 ft<sup>2</sup>) area of the home, located in a mild temperature zone. In reality, the average AC load is typically half of the capacity[32], so we model the AC load with the following normal distribution:

$$P_{AC} \sim \mathcal{N}(\mu = 1.8 \text{ kW}, \sigma = 0.5 \text{ kW}).$$

The AC will be in the idle status ( $s_t^{AC} = 0$ ) if none of the residents are at home or active ( $s_t^{AC} = 1$ ) otherwise. The agent may keep the AC off intermittently under active status; however, the resident will manually turn the AC on if it causes discomfort. We generate this feedback  $f_t^{AC}$  if AC goes off within  $T_{AC}$  minutes of being turned on. In that case, the residents turn on the AC manually, which penalizes the HERS agent by \$  $\delta_{AC}$ . We model

$T_{AC}$  with a uniform distribution between 45 to 90 minutes and intermittent off duration as 15 minutes. We include the on duration as a state for AC.

#### 2.4.2.2 *DW and WD (Type-2)*

The activity pattern of house B indicates the lack or no usage of a dishwasher (DW). We generate the dishwashing events to be activated, i.e.,  $s_t^{DW} = 1$ ,  $T_{DW}$  minutes after any resident finishes dinner. We model the delay time  $T_{DW}$  with the Poisson distribution with 60 minutes mean value. There will be no dishwashing events for the days when none of the residents have cooked, as there will not be a significant load for the dishwasher. The analysis in [27] estimates 152-minute automatic dishwashing for a comparable load to the considered household. Hence, we model the dishwashing event as a 2.5-hour continuous operation with 1.1 kW power. The Bosch 500 series smart dishwashers are among the most popular models of the year 2020 and serve as the DW model in our experiments [30]. The agent needs to complete dishwashing before the subsequent switching by the residents; otherwise, it receives the discomfort cost  $\delta_{DW}$ , and the DW is turned on manually to clean the previous dishes.

House B has a regular heavy load washer & dryer (WD), so following its laundry schedule would not be practical. The future smart homes will utilize the high-tech WD combos like the LG WM3900HBA, a single compartment light-duty device that takes around 1 hour for washing and 1.5 hours for drying for an average cloth load. We estimate that the residents produce this cloth load every three baths, hence fill and switch the WD in active mode on average 30 minutes (Poisson mean) after their second or third bath (with equal probability) from the previous laundry. Then the RL agent has to turn the WD on for a 1-hour continuous washing cycle, followed by 1.5-hour drying cycles with 1.2 kW power to complete the laundry. If the agent does not complete the process before the next switching by the residents, the resident provides negative feedback  $\delta_{WD}$  and turns on the WD immediately to clean the previous cloths.

Table 2.4: EV usage data generation.

	Weekdays			Weekends	
Duration, $t_a$ (hrs)	< 8	8-16	> 16	< 10	> 10
Purpose	Leisure	Office	Travel	Leisure	Travel
Miles driven, $M$	$f(t_a)$	$40+f(t_a-10)$	n/a	$f(t_a)$	n/a
Minimum Battery Before Trip	40%	40%	70%	40%	70%
Battery Status After Trip	$\beta - \frac{M}{220}$	$\beta - \frac{M}{220}$	20%	$\beta - \frac{M}{220}$	20%

### 2.4.2.3 EV Charging (Type-3)

The residents' activity pattern shows that they mostly go out of the home together. Considering an EV in the house, we assume that the second resident drives it. The EV driver's work pattern seems to consist of long hours with some off days throughout the week. We set his one-way drive to work as 20 miles; 69th percentile driving distance from the data collected by The American Time Use Survey (ATUS) [89], which includes over 13,000 respondents. The activity data provides us with the duration the resident is away from home. Based on the duration, we label such away time as leisure, office time, and travel as in Table 2.4. We assume the EV is always connected to the charger when the resident is at home.

For weekdays, if the resident stays away for less than 8 hours, it is labeled as a leisure activity, which includes going shopping, visiting friends, short trips, theater, etc. Residents spend more time in leisure activities during the weekend, extending the leisure activity labeling time to 10 hours for the weekend. Driving distance in miles during leisure trip for  $t_a$  time duration is approximated as;

$$M = f(t_a) = t_{\text{driving}} \times v_{\text{avg}} = \alpha \times t_a \times v_{\text{avg}}.$$

where,  $\alpha = t_{\text{driving}}/t_a$  is the ratio of time spent for driving and the total time spent away. We model it with a normal distribution

$$\alpha \sim \mathcal{N}(\mu = 0.33, \sigma = 0.1).$$

Average speed  $v_{\text{avg}}$  is taken as 30 mph. The instances in which the resident spends 8-16 hours out of home is labeled as office and leisure activity during weekdays. Round trip to the office is taken as 40 miles, additional time after 10 hours is considered a leisure activity, and driving distance is calculated as  $M = 40 + f(t_a - 10)$ .

2021 Tesla Model 3 Standard Range is one of the most popular latest EV models with a 450 hp (336 kW) engine 50 kWh battery. The level-2 charging of 7.68 kW (240 V 32 A) capacity would require 6.5 hours to charge the completely depleted EV battery fully. Battery status after a trip is the initial battery status when going out of home  $\beta$  minus  $\frac{M}{220}$  as the Tesla 3 model has a standard driving range of 220 miles. The resident does not use the EV if  $\beta$  is less than 40% before starting a trip. The resident takes some other transportation mode and assigns a discomfort cost  $\delta_{EV1}$  to the RL agent. If the resident stays more than 16 and 10 hours out of home, respectively, on weekdays and weekends, we label this activity as travel that may require outside charging. We do not calculate driving distance for traveling; however, we set battery status after the travel to be 5-20 %, as home charging is the cheapest and the resident would try outside (paid) charging as little as necessary to reach home. The resident requires a higher initial charge for traveling. We set a higher discomfort cost  $\delta_{EV2}$  if  $\beta < 70\%$  before travel. As the initial charge level is higher for travel, we include the next trip type as an input state for the EV.

### 2.4.3 Discomfort Cost

Discomfort costs  $\delta_d$  for each device are critical parameters in the HERS setup. So, we model it as user-defined numbers that the residents can set initially and update while the

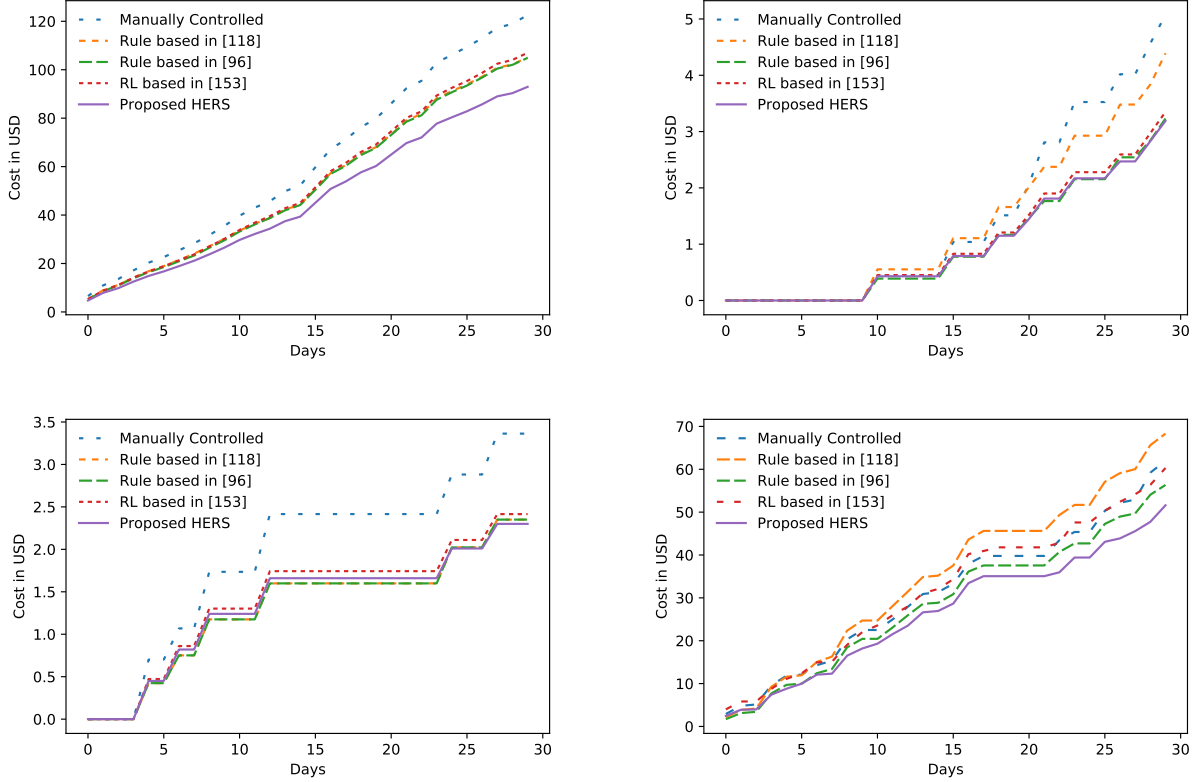


Figure 2.4: Daily cumulative cost in scenario 1 for devices: AC (top left), DW (top right), WD (bottom left), and EV (bottom right) for 1-month duration.

HERS is at service. The discomfort costs also represent the comfort and device priority mindset of the residents as low discomfort cost will emphasize electricity cost, and high discomfort cost will prioritize human feedback. In case of an update to the discomfort costs  $\delta_d$ , thanks to its adaptive nature, the RL agent will update the policy in an online fashion. For the experimental purpose, we performed a survey among twenty participants with different backgrounds (e.g., student, homemaker, engineer, etc.) to set the discomfort cost for each of the four devices. Survey results suggest EV charging failure creates the maximum discomfort. Other discomfort costs in decreasing order are for WD, DW, and AC. We select discomfort cost coefficients as  $\delta_{AC} = 20$ ,  $\delta_{DW} = 40$ ,  $\delta_{WD} = 50$ ,  $\delta_{EV1} = 100$ ,  $\delta_{EV2} = 300$  in USD.

## 2.5 Results

### 2.5.1 Benchmark Policies

(1) Manually controlled policy: In this policy, the residents operate the devices themselves, so a device turns on immediately upon its activation without any scheduling consideration. We assume the residents turn off the AC when both of them are out of home and turn it on upon returning. This policy ignores the benefit of smart scheduling, and we will refer to it as the baseline policy to evaluate the other policies' success.

(2) Rule based HEM in [118]: Shirazi et al. [118] present a home energy management with DERs and appliance scheduling (HEMDAS). The energy management problem in a house is modeled as a mixed-integer nonlinear programming (MINLP) that includes constrained optimization for managing DERs and appliance usage. More precisely, the devices are scheduled based on real-time pricing of electricity during a time window. They define separate earliest starting times (EST) and latest finish times (LFT) for DW, WD, and EV to ensure user convenience. Each device is scheduled based on the real-time electricity price during its operating time window. The AC maintains the desired temperature decided by the customer, which our smart home agent ensures by keeping the AC on for 90 minutes before every 15-minute interruption.

(3) Rule-based HEM in [96]: Pilloni et al. [96] survey 427 people about their degree of annoyance if a device performs under-capacity or is scheduled for later periods. The survey responses are used for generating different types of resident profiles. During training, the smart home residents' usage pattern is matched to one of those profiles. Once the resident's appliance usage profile is assigned, the algorithm minimizes the cost for each device,

$$C_t^d = \frac{P_t^d \times \kappa \times \rho_t}{\sigma(\Delta X)}$$

where the numerator represents the electricity cost and  $\sigma(\Delta X) \in (0, 1]$  is the relative satisfaction level of the home residents for the device. This rule-based method accommodates

user preference and provides a good analogy to our discomfort feedback-based RL approach. The resident feedback pattern in our setup for the AC, DW, and WD matches most of the resident profiles in the survey. Since [96] does not provide an EV charging profile, we assume that this policy schedules EV only if its battery is more than 50% charged, otherwise charges at full capacity.

(4) RL-based HEM in [153]: In [153], Xu et al. utilize hour-ahead electricity price as a state to minimize electricity cost. We tailor their approach to fit this comparative analysis with the following modifications: (i) Agent makes decisions every 15 minutes instead of hourly decisions. (ii) There is no PV generation in our setup, so the MDP state consists of electricity price of the next 24 hours, with 4.67 % prediction error following the case-1 (best prediction) in that chapter. (iii) We consider the AC as a priority device that maintains the user set the temperature on its own. Hence, the possible actions for the AC remain turn on or off instead of different power ratings, (iv) We include EV battery depletion, which is overlooked in [153].

## 2.5.2 Scenarios

### 2.5.2.1 Scenario 1: Unlimited Peak Demand

There is no restriction for keeping the electricity usage within a limit in this scenario. Fig. 2.4 shows the daily cumulative cost comparison among policies for different devices, and Table 2.5 summarizes the results. The manually controlled policy has the maximum monthly total cost of \$193. Among the rule-based approaches, the Piloni et al. method [96] costs \$166 and performs better than the Shirazi et al. method [118] with \$180 monthly cost. The RL-based approach in [153] attains \$172 monthly, and the proposed deep RL-based HERS policy achieves the lowest cost with \$149 and minimizes the cost by 23 % from the baseline manual control policy. The manually controlled policy starts operation immediately, thus does not take advantage of the lower electricity rate at off-peak hours, unlike the rule-based ones. However, the rule-based policy follows a conservative approach

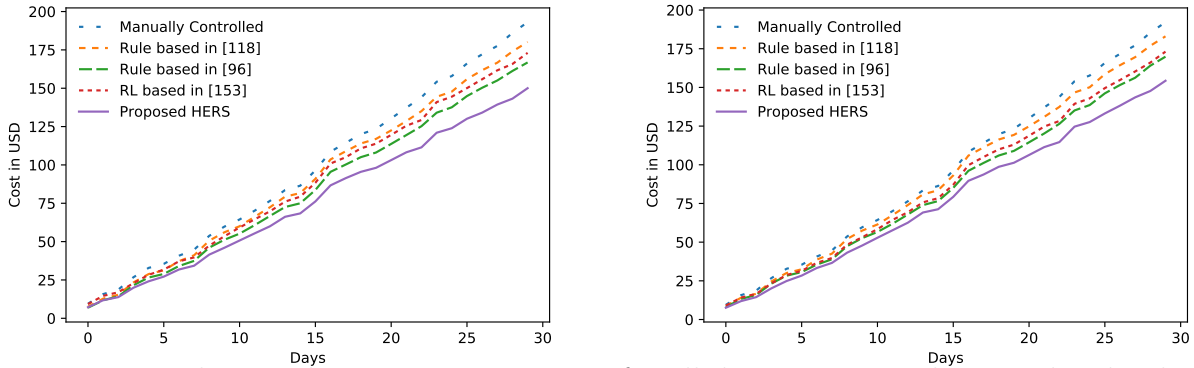


Figure 2.5: Daily cumulative cost comparison for all devices among the considered policies for scenario 1 (left) and scenario 2 (right).

for optimization by searching low tariffs in a smaller time window to avoid creating resident discomfort. Especially, the EV charging time window in method [118] overlaps with the peak hours. So, these policies minimize the cost for all appliances on a smaller scale. The RL-based approach in [153] achieves comparable results with the rule-based policies. The success of this policy is limited due to only including electricity price in its state definition and overlooking many critical features that the HERS policy capitalizes on (see Table 2.2). The HERS policy focuses on human feedback in its cost and runs the devices optimally. For instance, HERS keeps the AC off for shorter intervals during midnight without causing any resident discomfort. The proposed deep RL-based policy is expected to decrease the cost further for a system with more devices.

### 2.5.2.2 Scenario 2: Limited Peak Demand

To avoid overloading a distribution system, the utility company often restricts users to keep energy usage under a threshold. Under this scenario, we limit the peak electricity usage to 10 kW to obey such restrictions. The EV charging can take up to 7.68 kW of electricity, even greater than the sum of other loads. So, all the devices other than the EV receive their unrestricted electricity. Hence, the other devices' electricity cost is the same for both scenarios. The EV charging gets the least priority and can consume up to the remaining electricity. Fig. 2.5 compares the total cost among different policies for both of the scenarios.



Table 2.5: Monthly cost (\$) comparison for different policies.

Policy Device	AC	DW	WD	EV		Total	
				S1	S2	S1	S2
Manual Control	122.9	5.06	3.48	62.01	62.16	193.3	193.4
Rule-based in [118]	104.6	4.39	2.35	68.32	70.55	179.6	181.8
Rule-based in [96]	104.6	3.23	2.35	56.23	60.97	166.4	171.1
RL-based in [153]	106.7	3.35	2.41	59.76	60.37	172.2	172.8
Proposed HERS	92.0	3.19	2.3	51.6	55.88	149.1	153.4

Table 2.6: Computational statistics for the experiments.

Hardware	Software	Task	Computation time
Intel(R) Core i7,3.60 GHz, 16 GB RAM	Python 3.7 Pytorch 1.8.1	Training	128 min
		Online Scheduling	4 sec

In Scenario 2, all the policies attain similar results as in Scenario 1, however with a small increase in cost due to the restrictions. With more devices or lower peak limiting, the results may vary more compared with Scenario 1.

### 2.5.3 Computational Statistics

Fig. 2.6 shows that the proposed deep RL algorithm learns the optimal policy within 600 episodes. Table 2.6 shows that training convergence takes 128 minutes and online decision making requires only 4 seconds in our computer (Intel(R) Core i7,3.60 GHz, 16 GB RAM), exhibiting the real-world applicability. The high cost in the early episodes indicates discomfort among residents; however, it ceases very fast. The trained HERS will take feedback from the consumer to optimize the electricity cost of the house. We examine the above policies for two scenarios.

### 2.5.4 HERS Schedule Demonstration

Fig. 2.7 shows the implemented schedule by HERS for a particular day. The black curve shows the electricity price. From the residents' feedback, HERS learns that switching the AC off for 15 minutes after 1 hour of continuous operation is its optimal schedule that minimizes

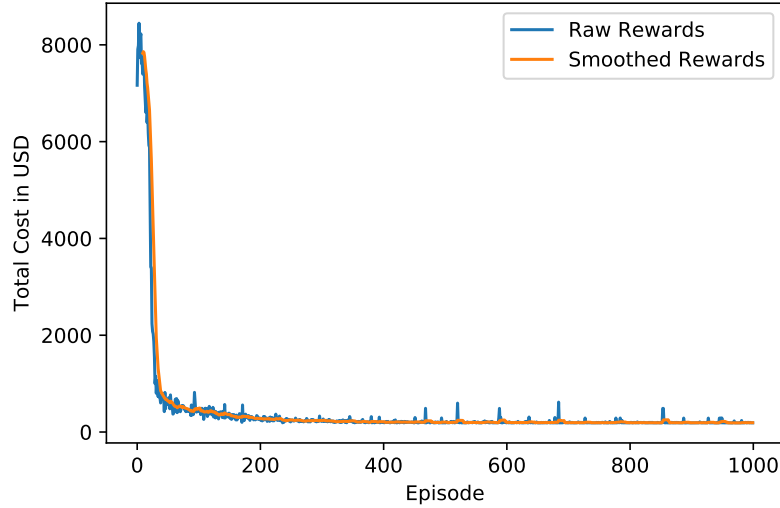


Figure 2.6: Convergence of the proposed deep RL algorithm for HERS for scenario-1 total cost.

electricity cost and does not create any discomfort. Hence, HERS follows this pattern and keeps the AC off when no one is home (9:30 am-5:45 pm). The EV is charged at maximum capacity (7.68 kW) during the low tariff early hours (12 am-3 am) and its remaining charge at 75% capacity (5.76 kW) during a slightly higher tariff (3 am-4 am). The EV returns home at 5:45 pm; however, it waits for lower electricity tariffs at 11 pm-12 am. The DW and WD require 2.5 hours of continuous power that the HERS schedules for the low-demand low-tariff hours during mid-day (12:30 pm-3 pm). Notably, HERS chooses this schedule instead of 12:00 pm-2:30 pm as the electricity price is lower during 2 pm-3 pm compared to 12 pm-1 pm. This sample schedule shows that HERS learns to minimize electricity cost and resident discomfort by utilizing the human feedback and activity labels in the proposed deep RL setup.

## 2.6 Discussion

This work focuses on key features derived from residents' activity for operating smart devices. The reward of the RL agent accommodates direct human feedback, thus providing a setup similar to the popular recommendation systems (e.g., video, book, music, etc.). We

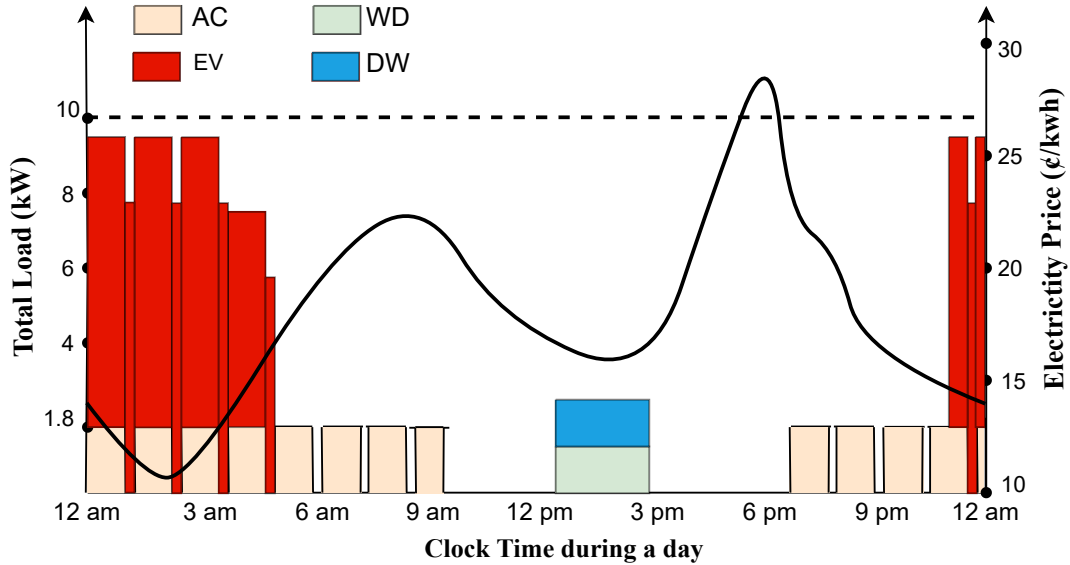


Figure 2.7: HERS scheduling results for a day under scenario-2 (peak demand limit 10 kW).

understand that any other approach incorporating more customized features for different appliances may achieve further improved results. So, the RL-based recommendation approach for device-specific policy-making has a high potential. This work demonstrates the benefit of including human activity-based states and human feedback-based rewards for adaptive HEM. Our model provides usage control of devices that do not include PV sources, energy storage, microgrid, and data sharing with other homes or a multi-agent setup. However, our core architecture can accommodate these features in the future to open up further research opportunities in this domain.

## 2.7 Conclusion

This work presents a deep Reinforcement Learning (RL) based recommendation system for smart home energy management (HEM). Residents' manual override for a device is interpreted as a negative reward to the RL agent that operates the device. So, the goal of the RL agent is to capitalize low-tariff electricity without creating human discomfort. To the best of our knowledge, this is the first work that takes direct human feedback for device management in a general smart home setup. Intuitively, this method works similarly

to the popular recommendation applications that suggest a video, book, music, etc., based on a user's usage pattern, so we call it Home Energy Recommendation System (HERS). Furthermore, the RL agent considers the human activities for state definition, another novelty the existing literature lacks. The experimental results show that the human activity pattern plays a vital role in device operation, in comparison with the RL approach of Xu et al. [153] that only considers electricity price for state definition. Our comparative analysis shows that HERS minimizes the electricity cost significantly with respect to the manually controlled policy, rule-based policies in [118, 96], and the RL-based policy presented in [153].

## Chapter 3: Modeling and Simulating Adaptation Strategies Against Sea-Level Rise Using Multi-Agent Deep Reinforcement Learning

### 3.1 Introduction

<sup>3</sup>Sea-level rise (SLR) is one of the most catastrophic outcomes of the global increase in greenhouse gas (GHG) emissions and climate change. While many policy makers have committed to reducing GHG emissions since the Paris agreement in 2015, coastal communities will require adaptation strategies to deal with SLR problems before harnessing the benefit of worldwide GHG emission reduction [24]. Due to climate change and SLR, storm surge, recurrent hurricanes, and permanent inundation pose significant challenges to most coastal cities, many of which are among the world’s largest cities [147]. Underdeveloped areas will also face many social and financial crises apart from the property loss [70, 112].

Recently, in the literature, a quantification of present and future flood damages in 136 major coastal cities is presented in [56]. Population growth is also considered in [59] to assess the potential magnitude of future impacts in the continental US. The study in [52] proposes a coherent statistical model for coastal flood frequency analysis and validates a mixture model for 68 tidal stations along the contiguous United States coast with long-term observed data. [57] demonstrates a methodology to assess the economic impacts of climate change at city scale (Copenhagen, Denmark) and the benefits of SLR adaptation. [51] uses HAZUS-MH [110] coastal flood hazard modeling and loss estimation tools to determine flood extent and depth and the corresponding monetary losses to infrastructure in Miami-Dade County. A

---

<sup>3</sup>Portions of this chapter were published in IEEE Transactions on Computational Social Systems [128]. Copyright permissions from the publishers are included in Appendix B.

case study comparing the cost-effectiveness of nature-based and coastal adaptation for the Gulf Coast of the United States is presented in [103].

Several countries are already making significant investments toward reducing the catastrophic and long-term impacts of SLR. However, the progress in risk reduction is far behind the coastal development and population growth globally [95]. The need for appropriate planning and execution for hurricane and flood protection is becoming more prominent as SLR risks grow. The major challenge is the daunting cost of undertaking mega projects, and building megastructures [66]. The US government spends billions of dollars to fund agencies like the US Army Corps of Engineers and the US Department of Transportation for hazard mitigation. In 2020, the Federal Emergency Management Agency (FEMA) announced to grant up to \$660 million in grant funding, including a record-breaking \$500 million for the Building Resilient Infrastructure and Communities (BRIC) pre-disaster mitigation grant program and \$160 million for the Flood Mitigation Assistance program [13].

The success of such investments depends on understanding different risk drivers from a financial viewpoint, including SLR and the current state of infrastructure [55]. A disaster cost and investment benefit analysis for a region can provide a guideline to the government about budgeting its funds towards different mitigation programs. Furthermore, since SLR is not uniform across the globe [63], the adaptation planning for different regions may differ significantly. Risk assessment and investment planning for different regions might require separate analyses and consider different sea-level projections [31]. Governments, in particular administrators and officials in the corresponding agencies, need substantial information for better strategic vision and adaptation planning [24, 49, 104]. The challenging task of adaptation planning demands considering various stakeholder dynamics and SLR scenarios [139]. Explicitly modeling the stakeholders' reactions to SLR scenarios can help create strategies suitable for local impacts and resilience management, and requires planning mechanisms such as agent-based modeling and sequential decision making [26, 42, 98].

To this end, we here study the interactions between SLR stakeholders under different SLR scenarios using multi-agent deep reinforcement learning (RL). Specifically, we use a probabilistic model for nature’s response to the collaborative policies of three local agents (government, residents, businesses). The proposed multi-agent RL framework serves two purposes. It provides a general scenario planning tool to investigate the cost-benefit analysis of natural events (e.g., flooding, hurricane) and agents’ investments (e.g., infrastructure improvement), and also shows how much the total cost due to SLR can be reduced over time by optimized adaptation strategies. We demonstrate the proposed scenario planning tool using available economic data and sea-level projections for Pinellas County, Florida, in a case study. Although we here focus on the SLR problem, the proposed scenario planning framework can be adapted to other natural and socioeconomic systems. A preliminary version of this work was presented in [119]. This submission greatly enhances both the intellectual merit and the broader impacts of the work. The major improvements include a more realistic multi-agent setup, more effective state-of-the-art deep RL methods, and a case study with extensive experimental results based on real economic data from the Tampa Bay area.

The remainder of the chapter is organized as follows. The proposed multi-agent RL framework is presented and analyzed in Sec. 3.2. The case study is given in Sec. 3.3, and the concluding discussions and remarks are provided in Sec. 3.4.

## **3.2 Multi-Agent RL Framework**

### **3.2.1 Agent-based Modeling for Adaptation Strategies**

The long-term effects of adaptation strategies can be effectively simulated using agent-based modeling, where an agent represents each stakeholder, and its actions are modeled through realistic policies. In the considered sequential decision-making setup, at the beginning of a year, residents and businesses decide on their additional tax contributions towards SLR adaptation; then the government decides on its own investment amount against SLR

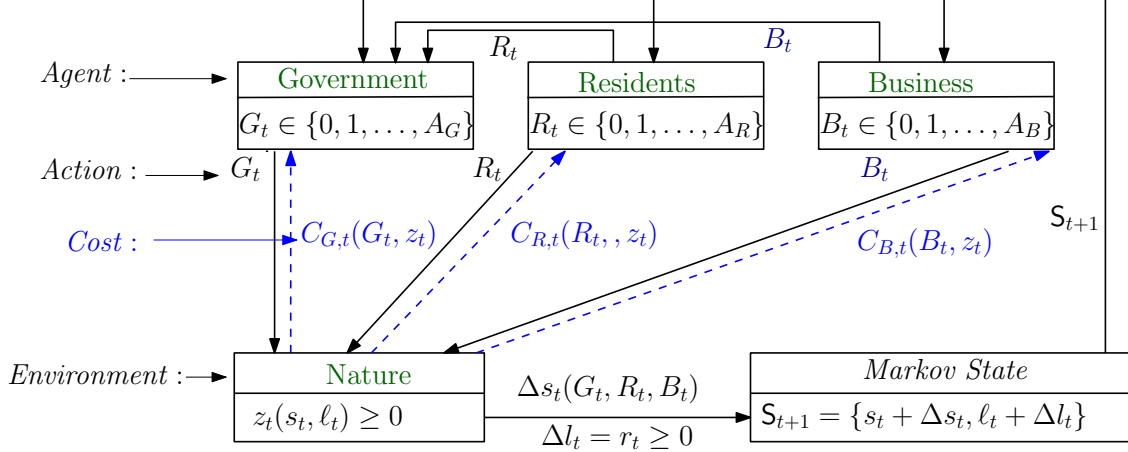


Figure 3.1: Proposed multi-agent MDP framework.

and implements an SLR adaptation strategy based on the total investment amount from all stakeholders. Finally, at the end of the year, the total cost from natural events is observed, which can serve as a feedback about the recent SLR adaptation strategies and inform the agents' future actions. A straightforward but realistic policy is the cost-based policy in which an agent decides whether to invest and the investment amount depending on the cost it experiences from the natural events. While different threshold levels on the natural cost can be used to model different agent prototypes (i.e., lower threshold for more reactive agents), it is hard to select such thresholds to link them to realistic prototypes. We here show that a multi-agent RL framework can be used to model realistic stakeholder policies in an easily controllable way. Our proposed RL framework provides an intuitive parameterization (cooperation indices between zero and one) to simulate different stakeholder prototypes conveniently. Moreover, the proposed RL framework illustrates how much cost can be saved by proactive and fully cooperative stakeholders with optimized decision policies.

### 3.2.2 MDP Formulation

We next explain the proposed cost models for the environment and the agents under the MDP framework. We propose a multi-agent MDP framework to model the behaviors of local SLR stakeholders (government, residents, and businesses), and their interactions



among themselves and with nature. As shown in Fig. 3.1, government, which is the major decision-maker in dealing with the SLR problem, at each time step  $t$  takes an action (i.e., investment decision)  $G_t$  and receives feedback from nature through the cost  $z_{G,t}$ . The other two agents, residents and businesses, similarly take actions  $R_t$ ,  $B_t$  and receives feedback from nature through the costs  $z_{R,t}$  and  $z_{B,t}$ , respectively. Then, this natural and socioeconomic system moves to a new state  $\mathcal{S}_{t+1}$  based on the current state  $\mathcal{S}_t$ , agents' action  $G_t$ ,  $R_t$ ,  $B_t$ , and SLR  $r_t$ . The system state consists of the city's infrastructure state  $s_t$  and the sea level  $\ell_t$ , i.e.,  $\mathcal{S}_t = \{s_t, \ell_t\}$ . The agents' decisions determine the infrastructure state  $s_t = s_0 + \sum_{m=1}^t (\Delta s_{G,m} + \Delta s_{R,m} + \Delta s_{B,m}) = s_{t-1} + \Delta s_t(G_t, R_t, B_t)$ . Likewise, the sea level at time  $t$  is given by the cumulative SLR values:  $\ell_t = \ell_0 + \sum_{m=1}^t r_m = \ell_{t-1} + r_t$ , where  $r_m$  is the SLR value at time  $m$ . Here,  $s_0$  and  $\ell_0$  are respectively the initial infrastructure state and the initial sea level of the region relative to a reference year. In terms of simulations, these are two user-defined numbers representing the existing states at the beginning of the simulations. The system state satisfies the Markov property:  $(\mathcal{S}_{t+1} | \mathcal{S}_t, \dots, \mathcal{S}_0) = (\mathcal{S}_{t+1} | \mathcal{S}_t)$ . We assume that the government observes the other agents' actions  $R_t, B_t$ , hence has the complete knowledge of multi-agent MDP. However, the residents and businesses do not necessarily know the other agents' actions, hence the MDP is partially observable to them. The parameters of the proposed multi-agent MDP framework are summarized in Table 3.1.

Table 3.1: Model parameters.

Initial sea level	$\ell_0 \geq 0$
SLR at time $t$	$r_t \geq 0$
Sea level at time $t$	$\ell_t = \ell_0 + \sum_{m=1}^t r_m$
Initial infrastructure state	$s_0 \in \{1, 2, \dots\}$
Infrastructure improvement at time $t$ ,	$\Delta s_t(G_t, R_t, B_t)$
Infrastructure state at time $t$ ,	$s_t = s_0 + \sum_{m=1}^t \Delta s_m$
Government's decision at time $t$	$G_t \in \{0, 1, \dots, A_G\}$
Residents' decision at time $t$	$R_t \in \{0, 1, \dots, A_R\}$
Businesses' decision at time $t$	$B_t \in \{0, 1, \dots, A_B\}$

### 3.2.3 Modeling Nature

We model nature's cost  $z_t = z_{G,t} + z_{R,t} + z_{B,t}$  using the generalized Pareto distribution, which is commonly used to model catastrophic losses, e.g., [34, 140, 40]. It is known that the storm- and flooding-related costs for the stakeholders have been increasing with SLR [7]. Thus, we model the scale parameter of generalized Pareto distributed  $z_t$  directly proportional to the most recent sea level  $\ell_t$  and inversely proportional to the most recent infrastructure state  $s_t$ . The cost from nature is distributed among the stakeholders through the multiplying factors ( $m_G + m_R + m_B = 1$ ) for government ( $z_{G,t} = m_G \times z_t$ ), residents ( $z_{R,t} = m_R \times z_t$ ), and businesses ( $z_{B,t} = m_B \times z_t$ ). These factors vary with regions; however, generally  $m_G > m_R, m_B$  since typically government is faced with most of the cost from nature.

The probabilistic model for the cost from nature is given by

$$z_t \sim \text{GeneralizedPareto}(\xi, \sigma_t, \mu)$$

$$\mu \geq 0, \xi < 0, \sigma_t = \frac{\eta(\ell_t)^p}{(s_t)^q} \quad (3.1)$$

where  $\mu, \sigma_t, \xi$  are the location, scale, and shape parameters of generalized Pareto, respectively; and  $\eta > 0, p \in (0, 1), q > 0$  are our additional model parameters. The parameters  $\xi, \mu, \eta, p, q$  help to regulate the impact of the most recent sea level  $\ell_t$  over the nature's cost  $z_t$  relative to the most recent infrastructure state  $s_t$ . Choosing an appropriate set of parameters depends on the region considered for simulations. Our preference for modeling the scale parameter and not the location parameter is due to the fact that the scale parameter can control both the mean and the variance, whereas the location parameter appears only in the mean. For certain values of shape parameter  $\xi$ , the expected value and range of the cost are as follows:

$$\mathbb{E}[z_t] = \mu + \frac{\eta \ell_t^p}{(1 - \xi) s_t^q} \quad \text{for } \xi < 1$$

$$\mu \leq z_t \leq \mu - \frac{\eta \ell_t^p}{\xi s_t^q} \quad \text{for } \xi < 0.$$

The location parameter is set to be positive,  $\mu > 0$ , to generate positive disaster cost,  $z_t > 0$ . We choose  $\mu = \$30$  million from historical data provided in Table A-1 in [5], which indicates that a year with no serious natural disaster might produce this cost, typically to cover maintenance. To get an upper bound on  $z_t$ , we need the shape parameter to be negative,  $\xi < 0$ . We select  $\xi = -0.1$ , which limits the upper bound to roughly 10 times the expected cost. The other parameters  $\eta, \rho, q$  are set according to the cost projections in Pinellas County presented in [4], and in Sec. 3.3.

### 3.2.4 Modeling Stakeholders

In our model, the government is the biggest stakeholder and implementer of the investment decisions for other agents too. At each time step, e.g., a year, the government decides the degree of its investment  $G_t \in \{0, 1, 2, \dots, A_G\}$  for infrastructure development, where  $A_G$  is a finite positive integer.  $G_t = 0$  means no investment at step  $t$ . Hence, there are  $A_G + 1$  possible actions for the government at each time step. The numerical value of  $G_t = m$  can be interpreted as spending  $m$  unit money towards infrastructure development or the  $m + 1$  th action among  $A_G + 1$  different actions with increasing cost and effectiveness. Possible government actions include but are not limited to building seawalls, raising roads, widening beaches, building traditional or horizontal levees, placing storm-water pumps, improving sewage systems, relocating seaside properties, etc. [116, 2, 152]. The range of  $G_t$  is designed to cover the real world costs from the cheapest investment like cleaning the pipes to the most expensive investment like buying lands and property to relocate the seaside inhabitants and businesses. The total cost  $C_{G,t}$  to the agent at each time  $t$  consists of the investment cost and cost from nature. We assume most of the business and residential properties are insured by the government. So  $f \in (0, 1)$  fraction of their insurance payments,  $f \times I_{R,t}$  and  $f \times I_{B,t}$  respectively for residents and businesses go to the government, hence negatively contribute to  $C_{G,t}$ . We explain modeling  $I_{R,t}$  and  $I_{B,t}$  later in this section while presenting the models for residents and businesses. Since the government's investment decision has an integer value,

we model the total cost as  $C_{G,t} = \alpha_G G_t + z_{G,t} - f(l_{R,t} + l_{B,t})$  using parameter  $\alpha_G$  to map the decision to monetary value. The discounted cumulative cost for the government in  $T$  time steps is given by  $C_{G,T} = \sum_{t=0}^T \lambda_G^t [\alpha_G G_t + z_{G,t} - f(l_{R,t} + l_{B,t})]$ , where the discount factor  $\lambda_G \in (0, 1)$  discounts the weight of future costs following the common practice in MDP. In our model, this discount factor also serves as a measure of the government's cooperation towards long-term welfare, hence, termed as the *government's cooperation index* in this chapter. Higher  $\lambda_G$  corresponds to a more cooperative government which better recognizes the future SLR costs from nature, compared to a more short-sighted government represented by lower  $\lambda_G$ . The government's objective is to minimize the expected cumulative cost  $\mathbb{E}[C_{G,T}]$  by taking investment actions  $\{G_t\}$  over time.

Residents' community decides its own action based on its learning of the environment and hence modeled as an agent in our multi-agent setup. The community organization decides the degree of its investment  $R_t \in \{0, 1, 2, \dots, A_R\}$ , i.e., how much additional tax they are going to pay to the authority to build infrastructure for them. The unit cost of residents' investment  $\alpha_R$  is limited to some fraction of the government investment unit  $\alpha_G$  since the government is expected to cover the majority of infrastructure investment costs. The change in infrastructure state by residents' investment decision  $R_t$  is set relative to the government:  $\Delta s_{R,t} = R_t \times \alpha_R / \alpha_G$  where  $\Delta s_{G,t} = G_t$ . It is assumed that these coastal residents typically insure their vulnerable properties, so their cost from nature is mainly due to the raise of insurance premiums. Insurance premiums go up if the insurance had to pay more for recent catastrophic events. Hence, the insurance premium can be approximated based on the historical cost from nature as  $l_{R,t} = l_{R,0} \times \rho_R^t + l_R \times \sum_{m=1}^{t-1} \rho_R^{t-m} z_{G,m}$ , where  $\rho_R \in (0, 1)$  is the insurance company's memory factor for past events, and  $l_R$  is the coefficient that maps the total recent natural cost of the insurance company (i.e., the government) to insurance premium. Pre-existing insurance premium for the region,  $l_{R,0}$ , can serve as the initial value for the simulation. Apart from the insurance cost, the residents also endure a fraction of the cost from nature, represented by  $z_{R,t} = m_R \times z_t$ . The discounted cumulative

cost for the residents in  $T$  time steps is given by  $C_{R,T} = \sum_{t=0}^T \lambda_R^t (\alpha_R R_t + z_{R,t} + I_{R,t})$ , where the discount factor  $\lambda_R \in (0, 1)$  can be interpreted as the *residents' cooperation index*, as in the government's model. The residents' objective is to minimize the expected cumulative cost  $\mathbb{E}[C_{R,T}]$  by taking investment actions  $\{R_t\}$  over time.

Businesses are another major stakeholder of SLR impacts. Businesses get monetary loss through inundation, loss of customers, property damages, and increasing insurance premiums. Similar to the residents' model, we consider a business association to implement their collective actions. The business association takes action  $B_t \in \{0, 1, 2, \dots, A_B\}$ , i.e., decides on their degree of monetary contribution towards infrastructure development. The unit cost of businesses investment  $\alpha_B$  typically ranges between  $\alpha_R$  and  $\alpha_G$ . The change in infrastructure state by businesses' investment decision  $B_t$  is  $\Delta s_{B,t} = B_t \times \alpha_B / \alpha_G$ . Similar to the residents, businesses have insurance and non-insurance costs. Insurance premiums go up if the insurance had to pay more for recent catastrophic events. Hence, the insurance premium for businesses is modeled as  $I_{B,t} = I_{B,0} \times \rho_B^t + I_B \times \sum_{m=1}^{t-1} \rho_B^{t-m} z_{G,m}$ , where  $\rho_B \in (0, 1)$  is the insurance company's memory factor for past events,  $I_B$  and  $I_{B,0}$  are the insurance coefficient and the initial insurance premium for businesses, respectively. The discounted cumulative cost for the business agent in  $T$  time steps is given by  $C_{B,T} = \sum_{t=0}^T \lambda_B^t (\alpha_B B_t + z_{B,t} + I_{B,t})$ , where discount factor  $\lambda_B$  can represent businesses' awareness and cooperation against SLR and is called *businesses' cooperation index*. The businesses' objective is to minimize the expected cumulative cost  $\mathbb{E}[C_{B,T}]$  by taking investment actions  $\{B_t\}$  over time.

### 3.2.5 Optimal Policy Analysis

In our proposed MDP structure, each agent tries to minimize its expected total cost  $\mathbb{E}[C_T]$  in  $T$  time steps by following an optimal investment policy. At each time step, first, residents and businesses take their actions  $R_t$  and  $B_t$  respectively; then, the government collects their investments, makes its decision  $G_t$ , and implements the monetary investment towards developing infrastructure. So, the next state transition is fully observable to the

government and at the same time partially observable to the other two agents. We begin with the optimal policy analysis for the government. Its optimal value function, which gives the minimum expected total cost possible at each state  $(s_t, \ell_t)$ , characterizes the best action policy  $\{G_t\}$ , and is written as

$$V_G(s_t, \ell_t, O_{G,t}) = \min_{\{G_t\}} \mathbb{E}[C_{G,t}^T | \{G_t\}],$$

where  $C_{G,t}^T = \sum_{\tau=0}^T \lambda_G^\tau C_{G,t+\tau}$  is the cumulative cost starting from time  $t$ . We know from the main body of the chapter

$$C_{G,t} = \alpha_G G_t + z_{G,t} - f(l_{R,t} + l_{B,t}) \quad (3.2)$$

and,  $C_{G,T} = \sum_{t=0}^T \lambda_G^t [\alpha_G G_t + z_{G,t} - f(l_{R,t} + l_{B,t})]$ . Here, the government's observation  $O_{G,t}$  represents its knowledge about other agents' actions at time  $t$ . To find the optimal policy, the Bellman equation

$$V_G(s_t, \ell_t, O_{G,t}) = \min_{G_t} \mathbb{E}[C_{G,t} + \lambda_G V_G(s_{t+1}, \ell_{t+1}) | G_t]$$

provides a recursive approach by focusing on finding the optimal action  $G_t$  at each time step using the successor state value instead of trying to find the entire policy  $\{G_t\}$  at once. Using the cost expression given by (3.2) and considering possible  $A_G + 1$  actions for  $G_t$  this iterative

equation can be rewritten as

$$\begin{aligned}
V_G(\mathbf{s}_t, \ell_t, O_{G,t}) = \min \left\{ \underbrace{\mathbb{E}[z_{G,t} - f(I_{R,t} + I_{B,t}) + \lambda_G V(\hat{\mathbf{s}}_t, \ell_t + r_t)]}_{F_0(\hat{\mathbf{s}}_t, \ell_t)}, \right. \\
\underbrace{\mathbb{E}[\alpha_G + z_{G,t} - f(I_{R,t} + I_{B,t}) + \lambda_G V(\hat{\mathbf{s}}_t + \mathbf{1}, \ell_t + r_t)]}_{F_1(\hat{\mathbf{s}}_t, \ell_t)}, \\
\underbrace{\mathbb{E}[2\alpha_G + z_{G,t} - f(I_{R,t} + I_{B,t}) + \lambda_G V(\hat{\mathbf{s}}_t + 2, \ell_t + r_t)]}_{F_2(\hat{\mathbf{s}}_t, \ell_t)}, \dots, \\
\left. \underbrace{\mathbb{E}[A_G \alpha_G + z_{G,t} - f(I_{R,t} + I_{B,t}) + \lambda_G V(\hat{\mathbf{s}}_t + A_G, \ell_t + r_t)]}_{F_{A_G}(\hat{\mathbf{s}}_t, \ell_t)} \right\} \quad (3.3)
\end{aligned}$$

where  $\hat{\mathbf{s}}_t = \mathbf{s}_t + \mathbf{R}_t \times \alpha_R / \alpha_G + \mathbf{B}_t \times \alpha_B / \alpha_G$  is the deterministic next infrastructure state before government investment, termed as augmented infrastructure state. The knowledge of other agents' current actions provides the basis of our optimal policy analysis for the government and Theorem 1.

At each time step  $t$ , action  $G_t$  shapes the instant cost  $C_{G,t}$  and moves the system to the next state, which determines the discounted future cost  $\lambda_G V(\mathbf{s}_{t+1}, \ell_{t+1})$ . The optimum policy chooses among the investment actions  $G_t \in \{0, 1, 2, \dots, A_G\}$  that has the minimum expected total cost,  $\min_m \{F_m(\hat{\mathbf{s}}_t, \ell_t)\}$ , as shown in (3.3). Since the functions  $\{F_0(\hat{\mathbf{s}}_t, \ell_t), \dots, F_{A_G}(\hat{\mathbf{s}}_t, \ell_t)\}$  determine the optimal policy, we next analyze them to characterize the optimal government policy.

**Theorem 1.** *For  $m = 0, 1, \dots, A_G$ ,  $F_m(\hat{\mathbf{s}}_t, \ell_t)$  is nondecreasing and concave in  $\ell_t$  for each  $\hat{\mathbf{s}}_t$ ; and the derivative of  $F_m(\hat{\mathbf{s}}_t, \ell_t)$  with respect to  $\ell_t$  is lower than that of  $F_{m-1}(\hat{\mathbf{s}}_t, \ell_t)$ .*

Proof is provided in the Appendix. For a specific infrastructure state  $\hat{\mathbf{s}}_t$ , expected costs  $F_0(\hat{\mathbf{s}}_t, \ell_t), \dots, F_3(\hat{\mathbf{s}}_t, \ell_t)$  are illustrated in Figure 3.2 according to Theorem 1 where the expected total costs intersect each other only once for any given infrastructure state. The optimum policy picks the minimum of the  $A_G + 1 = 4$  curves at each time, which is shown with the

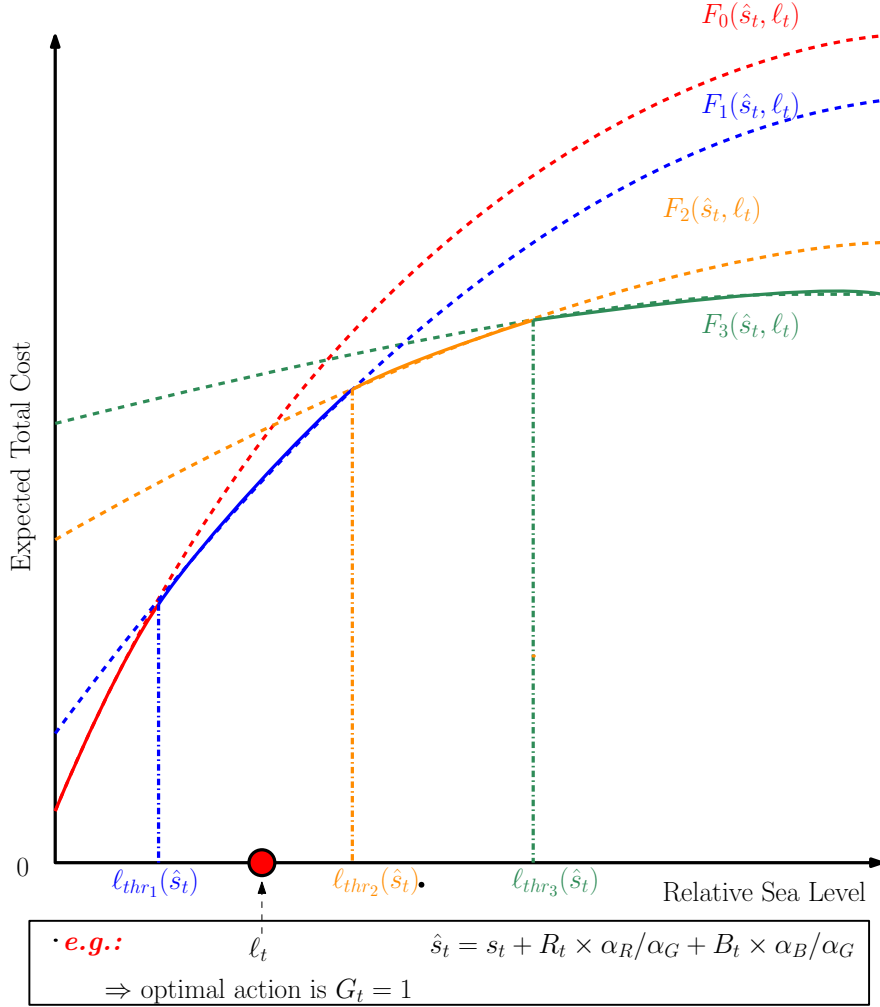


Figure 3.2: Expected total costs as a function of sea level for an example case with  $A_G = 3$ .

solid curve in Figure 3.2. As a result of Theorem 1, we next give the outline of optimum policy in Corollary 3.2.5.

The optimum policy, at each augmented infrastructure state  $\hat{s}_t$ , compares the sea level  $l_t$  with at most  $A_G$  thresholds where each threshold signifies a change of optimal action.

To prove Corollary 3.2.5, note that  $F_{m-1}(\hat{s}_t, l_t = 0) < F_m(\hat{s}_t, l_t = 0)$  for  $m \in \{1, 2, \dots, A_G\}$  because  $l_t = 0$  corresponds to the fictional case of zero sea level where there is no risk. That is,  $F_{m-1}(\hat{s}_t, l_t)$  starts at a lower point than  $F_m(\hat{s}_t, l_t)$ , but increases faster than  $F_m(\hat{s}_t, l_t)$  since its derivative is higher (Theorem 1). Also from Theorem 1, it is known that both of them are concave and bounded, hence  $F_{m-1}(\hat{s}_t, l_t)$  and  $F_m(\hat{s}_t, l_t)$  intersect exactly at one point for



$m \in \{1, 2, \dots, A_G\}$ . While for  $\ell_t$  less than the intersection point the action  $A_G = m$  is less effective than the action  $A_G = m - 1$  in terms of immediate cost and expected future cost, it becomes more effective when  $\ell_t$  exceeds the intersection point.

Figure 3.2 gives an example case with  $A_G = 3$  thresholds  $\ell_{thr_1}(\hat{s}_t), \ell_{thr_2}(\hat{s}_t), \ell_{thr_3}(\hat{s}_t)$ , which depend on  $\hat{s}_t$  and indicate change points of optimal action. However, depending on the slopes of  $\{F_m(\hat{s}_t, \ell_t)\}$  curves at each augmented infrastructure state  $\hat{s}_t$ , there may be less than  $A_G$  change points. To summarize, for a given state  $(\hat{s}_t, \ell_t)$ , the optimum policy chooses  $G_t$  based on the relative value of the current sea level  $\ell_t$ , with respect to the augmented infrastructure state  $\hat{s}_t$ .

The thresholds also depend on the cooperation index  $\lambda_G$ . Higher cooperation indices set the thresholds lower and vice versa. Intuitively, as  $\lambda_G$  grows, the government becomes more cautious about (i.e., sees more objectively without severely discounting) the expected future natural costs and sets a lower threshold for investment actions. On the contrary, small  $\lambda_G$  implies underestimated future costs and thus overemphasized immediate investment costs, which results in a high threshold for investment.

Optimal value functions for residents' and businesses' are similar to the government's. The Bellman equation for residents' is

$$V_R(s_t, \ell_t, O_{R,t}) = \min_{R_t} \mathbb{E}[C_{R,t} + \lambda_R V_R(s_{t+1}, \ell_{t+1}) | R_t].$$

Ideally, each agent would like to see other agents' actions. Nevertheless, in the considered problem, residents and businesses do not have the information of others' actions. Since the states do not change drastically, it is reasonable to approximate the agents' previous optimal action as their recent action. So, the residents approximate  $G_t \approx G_{t-1}$  and  $B_t \approx B_{t-1}$  in its observation  $O_{R,t}$ , where  $G_{t-1}$ , and  $B_{t-1}$  are the actions taken in previous time step by the corresponding agents. Similarly, the business approximate  $G_t \approx G_{t-1}$  and  $R_t \approx R_{t-1}$  in its

observation  $O_{B,t}$  for the value function

$$V_B(s_t, \ell_t, O_{B,t}) = \min_{B_t} \mathbb{E}[C_{B,t} + \lambda_B V_B(s_{t+1}, \ell_{t+1}) | B_t].$$

Due to such partial knowledge and approximations, functional analysis as in Theorem 1 is not tractable for residents and businesses.

### 3.2.6 Multi-Agent RL Algorithms

The continuous sea level values, which cause an infinite number of possible states, necessitate a deep RL algorithm instead of a traditional RL algorithm. We consider two deep RL approaches for comparison. The deep Q-network (DQN) algorithm [85], which is a popular choice for deep RL, addresses well the infinite-dimensional state space problem. It leverages a deep neural network to estimate the optimal action-value function for each of the three agents. The Advantage Actor-Critic (A2C) algorithm, a policy gradient-based algorithm [133], is a popular choice for multi-agent deep RL. A2C uses two neural networks:

(1) The actor network, also known as the policy network, outputs the probability for each action through a softmax function. It is updated using the gradient of expected return of the policy  $\pi_\theta$  with respect to the weights  $\theta$  of the neural network, e.g., for the government  $\mathbb{E}[\nabla_{\theta_G} \log \pi_{\theta_G}(G_t | \mathcal{S}_t) D_G(\mathcal{S}_t, G_t)]$  where  $\pi_{\theta_G}(G_t | \mathcal{S}_t)$  denotes the probability for action  $G_t$  at state  $\mathcal{S}_t$ , and the advantage function is given by  $D_G(\mathcal{S}_t, G_t) = C_{G,t} + \lambda_G V_{\psi_G}(\mathcal{S}_{t+1}) - V_{\psi_G}(\mathcal{S}_t)$ , where  $V_{\psi_G}$  is the output of the critic network (see below) with  $\psi$  denoting the network weights.

(2) The critic network, which is also known as the value network, is used to learn the value function for each state, e.g.,  $V_{\psi_G}(\mathcal{S}_t)$  for the government. It is updated using the gradient of the squared advantage function,  $\mathbb{E}[\nabla_{\psi_G} D_G^2]$ .

The cost from natural events, which is modeled with a generalized Pareto distribution, can have a high variance depending on the parameter settings, i.e., regular flooding costs

in a typical year vs. major hurricane costs in another year. Our experiments in the case study, explained next, corroborate the previous findings that A2C in general deals with high variance more successfully than DQN [133].

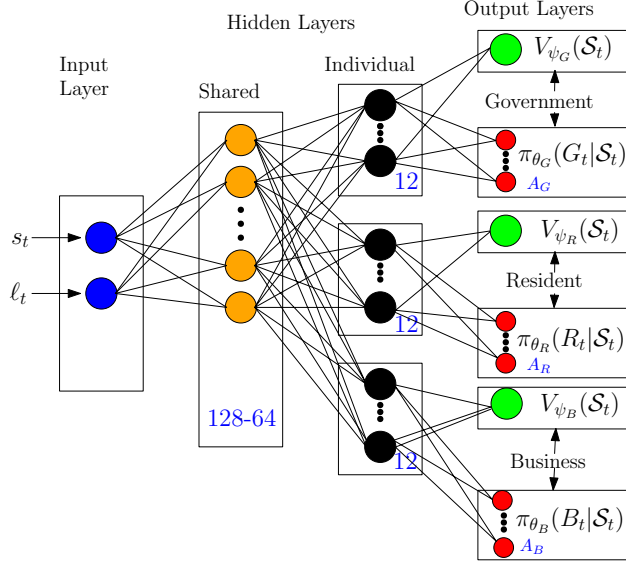


Figure 3.3: Unified A2C structure for all three agents.

For the multi-agent implementation of A2C, we consider three different structures. In the first one, a single deep neural network structure is used for all agents based on the similarities between their state and cost definitions. As shown in Fig. 3.3, the input layer, which represents the common system state  $\mathcal{S}_t = \{s_t, l_t\}$ , is the same for all agents. Numbers inside the box give the neuron numbers in each layer. Since the cost functions for the agents have similarities, they also share some hidden layers. From there on, the agents have their individual hidden layers to output their state values  $V_{\psi_G}, V_{\psi_R}, V_{\psi_B}$  (critic network) and action probabilities  $\pi_{\theta_G}, \pi_{\theta_R}, \pi_{\theta_B}$  (actor network). In this unified A2C structure, the interaction between agents is not explicitly implemented through observations of other agents' actions  $O_{G,t}, O_{R,t}, O_{B,t}$  at the input.

We next consider using a separate neural network for each agent with  $s_t, l_t$ , and  $O_t$  at the input, as shown in Fig. 3.4. Numbers inside the box give the neuron numbers in each layer. In this structure, the agents explicitly use the other agents' actions  $O_t$  in their input states. Government observes the residents' and businesses' actions before taking its own action, i.e.,

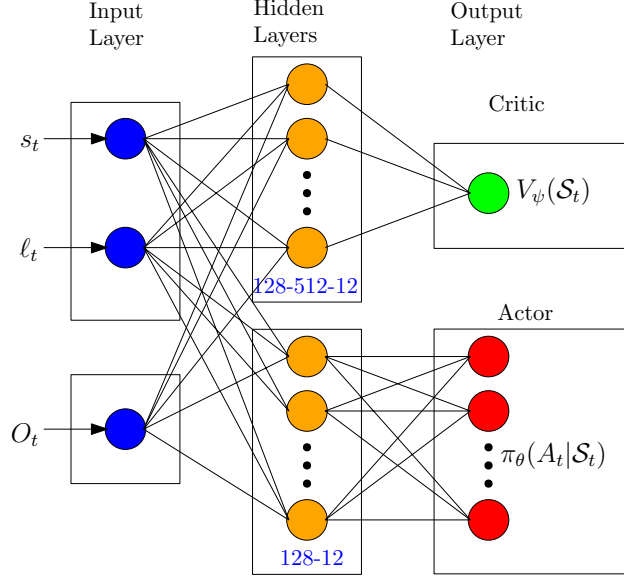


Figure 3.4: Separate A2C structure for each agent.

$O_{G,t} = \{R_t, B_t\}$ . However, residents and businesses only know the previous actions of other agents, i.e.,  $O_{R,t} = \{G_{t-1}, B_{t-1}\}$  and  $O_{B,t} = \{G_{t-1}, R_{t-1}\}$ .

As a third (hybrid) structure, we also consider using a single critic network common to all agents. Specifically, each agent has its own actor network, as in Fig. 3.4, but shares a common critic network. According to our experimental results in the case study, among the three multi-agent A2C structures, the separate A2C structure (Fig. 3.4) performs the best.

Algorithm 3.1 summarizes the separate A2C algorithm (Fig. 3.4). Each episode consists of Monte-Carlo simulations in which several states are visited according to the current policy defined by the current actor network. Line 1 initializes the disaster cost and investment cost parameters. Line 2 sets up the discount factors and insurance parameters. An episode starts with the initial relative sea level and infrastructure state. Line 7 shows the action selection procedure for the A2C agents. Then, the simulator calculates the costs. Actor and critic networks are updated at the end of an episode. The convergence of the separate A2C algorithm used in the experiments is shown in Fig. 3.5.

---

**Algorithm 3.1** Multi-agent A2C algorithm (Fig. 3.4)

---

- 1: *Input:*  $\mu, \epsilon, \eta, \rho, \mathbf{q}, \mathbf{m}_G, \mathbf{m}_R, \mathbf{m}_B, \alpha_G, \alpha_R, \alpha_B,$
  - 2: *Input:*  $\lambda_G, \lambda_R, \lambda_B, \rho_R, \rho_B, l_R, l_B$
  - 3: *Initialize* policy network with random weights  $\theta_G, \theta_R, \theta_B$  and critic network with random weights  $\psi_G, \psi_R, \psi_B$ .
  - 4: **for** episode = 1, 2, ... **do**
  - 5:   *Initialize* state  $\mathcal{S}_0 = (s_0, l_0)$
  - 6:   **for**  $t = 1, 2, \dots, T$  **do**
  - 7:     Sample action  $G_t, R_t, B_t$  from probability distribution generated by actor networks  $\theta_G, \theta_R, \theta_B$ .
  - 8:     Execute action  $G_t, R_t, B_t$  and observe costs  $C_{G,t}, C_{R,t}, C_{B,t}$
  - 9:   **end for**
  - 10:   Update actor network  $\theta_G$  (and similarly  $\theta_R, \theta_B$ ) by back propagating  $\mathbb{E}[\nabla_{\theta_G} \log \pi_{\theta_G}(G_t | \mathcal{S}_t) D_G(\mathcal{S}_t, G_t)]$ .
  - 11:   Update critic network  $\psi_G$  (and similarly  $\psi_R, \psi_B$ ) by back propagating  $\mathbb{E}[\nabla_{\psi_G} D_G^2]$ .
  - 12: **end for**
- 

### 3.3 Case Study

We here present a case study for Pinellas County, Florida, USA, using our multi-agent RL framework. Pinellas County, home to nearly one million residents, is the most visited destination on the US Gulf Coast. About 15 million tourists yearly spent over \$20 Billion over the past five years [10]. The top two cities in the county, St. Petersburg and Clearwater, are ranked among the cities with a high risk of flooding [56].

#### 3.3.1 Parameter Estimation

Hurricanes, floods, and stagnant water are some of the many SLR-related natural events that cause costs in different ways, such as loss of properties, jobs, taxes, and tourism incomes due to submerged areas. To model this cost  $\mathbf{z}_t$ , we begin with modeling the SLR amount  $r_t$ . For the Tide Gauge #8726520 located in St. Petersburg, FL, we utilize the online sea level change calculator developed by the US Army Corps of Engineers (USACE) in [3], which uses the NOAA projections for Pinellas County [137]. Among the seven SLR projections, with relative sea level (RSL) zero for the year 2000, the Tampa Bay Climate Science Advisory

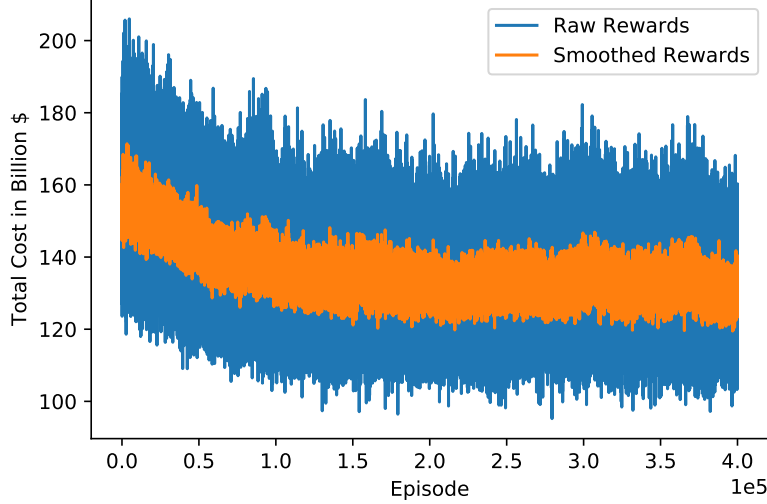


Figure 3.5: Average episodic total cost of all agents in the separate A2C policy for the high SLR scenario.

Panel ruled out the very low, low, and extreme scenarios for planning purposes [33]. We also disregard the intermediate-high scenario and limit our simulations for intermediate-low, intermediate, and high SLR scenarios.

Our simulations target the hundred years 2020–2120, hence we adjust the initial sea level value for 2020 to  $\hat{\ell}_0 = 100\text{mm}$  for all scenarios. For the following years, we follow SLR projections in [137] till 2120. RSL for these scenarios from [3] and our adjusted values are shown in Table 3.2. The randomness in SLR at each time step is modeled using the Gamma distribution, which is commonly used for modeling positive variables, including environmental applications, e.g., daily rainfall [16]. We use these adjusted projections  $\{\hat{\ell}_t\}$  as the mean RSL values for the Gamma distribution, i.e.,  $r_t \sim \text{Gamma}(\alpha, \beta)$ . Specifically, we set the scale parameter to  $\beta = 0.5$  and vary the shape parameter  $\alpha$  in a range to match the mean RSL, given by  $\sum_{n=1}^t \mathbb{E}[r_n]$ , with the adjusted NOAA projection curves. The successful curve fitting for mean RSL values shown in Fig. 3.6 is achieved by the following time series

Table 3.2: Relative sea level (mm) for different scenarios for St. Petersburg.

Year	NOAA 2017 [3]			Simulation adjusted RSL, $\hat{\ell}_t$		
	Int-low	Intermediate	High	Int-low	Intermediate	High
2000	0	0	0	n/a	n/a	n/a
2010	50	70	110	n/a	n/a	n/a
2020	110	150	220	100	100	100
2030	170	240	380	160	190	260
2040	220	330	540	210	280	420
2050	290	440	780	280	390	660
2060	350	570	1060	340	520	940
2070	410	710	1390	400	660	1270
2080	470	860	1740	460	810	1620
2090	520	1030	2150	510	980	2030
2100	580	1190	2590	570	1140	2470
2120	670	1430	3460	660	1380	3340
2150	840	1980	5230	n/a	n/a	n/a
2200	1080	2970	8940	n/a	n/a	n/a

equations,

$$\text{Int-low: } \alpha_t = 11.102 + 0.012 \times t$$

$$\text{Intermediate: } \alpha_t = 15.8 + 0.211 \times t \tag{3.4}$$

$$\text{High: } \alpha_t = 24 + 0.85 \times t,$$

where the subscript  $t$  represents calendar year  $2020 + t$ .

The generalized Pareto distribution used to model the cost from nature  $z_t$ , has three parameters, location, shape, and scale. The location parameter determines the range of  $z_t$ , in particular the lower limit. We set it as \$30 million according to the data provided in Table A-1 in [5], which indicates that a year with no serious natural disaster might produce this cost, typically for maintenance. To get an upper limit on  $z_t$ , we need the shape parameter to be negative. We set it as  $-0.1$  for the upper limit to be roughly ten times the expected cost. The scale parameter establishes the relation between sea level, infrastructure state, and disaster cost in our model. A recent report by the Tampa Bay Regional Planning

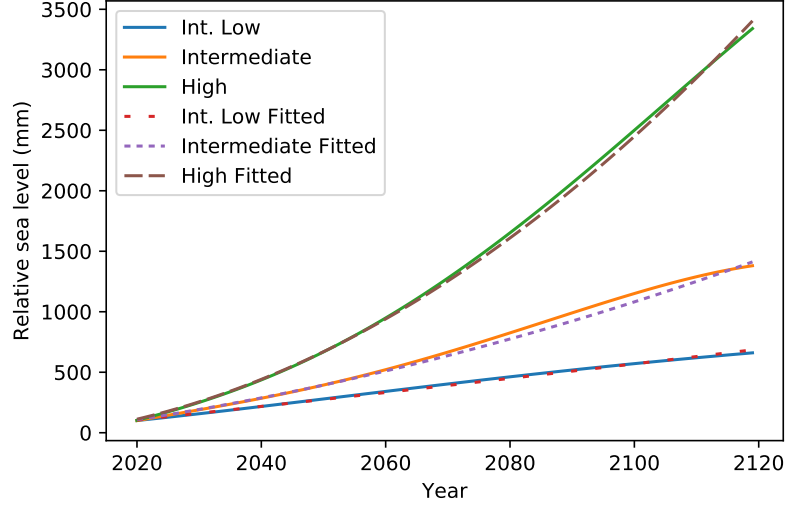


Figure 3.6: SLR projections by NOAA [137] (solid lines) and our fittings (dashed lines) for relative sea level change for St. Petersburg, FL.

Council [4] gives “the cost of doing nothing” due to SLR impacts under the NOAA’s high SLR projections for the Tampa Bay region, including Pinellas County. This report uses the widely accepted REMI PI+ economic modeling tool for their estimations. The following equations are obtained using the data provided in [4], where the cost unit is in million USD and subscripts represent the calendar year:

$$\mathbb{E}[z_{2060}] = \mu + \frac{\eta \ell_{2060}^p}{(1 - \xi) s_{2020}^q} = 5057,$$

$$\mathbb{E} \left[ \sum_{y=2020}^{2060} z_y \right] = 89000.$$

The report in [4] discusses the cost of doing nothing; hence we keep the infrastructure state  $s_{2020} = 20$  constant between the years 2020 and 2060 in the above equations. We further set the relative sea level  $\ell_{2020} = 100$ , and following the high SLR scenario we obtain  $\eta = 100$ ,  $p = 0.92$ ,  $q = 0.8$  as a set of values suitable for our simulations. Table 3.3 summarizes the values of the generalized Pareto parameters for the Pinellas County case study.



Table 3.3: Generalized Pareto parameters for Pinellas county.

$\xi$	$\mu$	$\eta$	p	q
-0.1	\$ 30M	\$ 100M	0.92	0.8

To determine the distribution of cost from nature, we use the economic data [10] and the cost projections [4] for tourism in Pinellas County. The tourism industry contributed \$9.25 billion annual spending impact to the Pinellas County local economy [10] in 2019. We consider only the tourism business in our model as they are the main business stakeholder of SLR impacts. The “cost of doing nothing” report gives the tourism loss in 2060 as \$898 million [4]. The principal cost for the business is the loss of net profit, which is estimated as the 5% of total tourism income. With the current sea level, we estimate the upper bound of business loss for the year 2019 as 10% of the net profit. This loss grows with 3% yearly growth for tourism business, in line with US GDP growth. With the high SLR scenario, business damage loss will increase to 100% of net profit in 2060, up from 10% in 2019, which is equivalent to saying that the tourism sector will lose all profit if no infrastructure is developed in the next 40 years. This cost model for tourism business in Pinellas County corresponds to the 22% of total cost from nature, i.e.,  $m_B = 0.22$  in (3.1). Together with the insurance cost explained below, it also gives similar costs in our simulations to the cost-of-doing-nothing estimation in [4].

Since the government is the major stakeholder with infrastructures, including buildings, roads, parks, etc. under its liability, we set the government’s portion within the cost from nature as 75%, i.e.,  $m_G = 0.75$  in (3.1). The majority of residents in coastal regions, in particular Pinellas County, have flood insurance, as explained next, hence most of the property inundation cost, which is estimated to be more than \$16 billion in the worst case scenario [4], is covered by the government. The direct cost to residents from nature is set as 3% of total cost from nature, i.e.,  $m_R = 0.03$ , to account for the uninsured and uninsurable properties.

For the insurance cost, we use a topology-based data set provided in [103] for exposed assets by ground heights for all the Gulf Coast. Pinellas County falls under a high-risk

Table 3.4: Action and cost parameters for Pinellas county.

Agent	Action	Action multiplier	Portion of natural cost	Insurance factor	Insurance memory
Govt.	0,1,2,3,4	\$140M	$m_G = 0.75$	n/a	n/a
Resident	0,1,2,3,4	\$20M	$m_R = 0.03$	$l_R = 0.04$	$\rho_R = 0.9$
Business	0,1,2,3,4	\$50M	$m_B = 0.22$	$l_B = 0.006$	$\rho_B = 0.9$

flood zone, with many of its 407,720 residential properties considered as exposed assets by ground heights [8]. The homeowners typically have the National Flood Insurance Program (NFIP) provided by the government. The residents of St. Petersburg paid an average insurance premium of \$950 and around \$33 million in total annually [6], which is the highest in Florida. Scaling this total insurance premium payment in St. Petersburg to the entire Pinellas County according to the almost 1/4 ratio of households [9], we set the base insurance premium by residents as  $l_{R,0} = \$132$  million. The insurance cost and action parameters for each agent are given in Table 3.4. Although most of the cost from nature is covered by the insurance, increasing costs due to SLR is reflected to the residents through a higher premium rate in the future. We empirically determined the insurance factor as  $l_R = 0.04$  and the insurance memory factor as  $\rho_R = 0.9$  to match the insurance data stated above. Similarly, the structural properties of businesses are mostly covered by insurance, and the premiums increase with accumulating cost from nature,  $\rho_B = 0.9$ . Since the commercial land use and the number of commercial insurance policies are less than the residential ones [4], we set the initial insurance premium for business as  $l_{B,0} = \$20$  million and the insurance coefficient as  $l_B = 0.006$ .

We assume the infrastructure improvement is proportionate to the total investment amount. Infrastructure development may include activities like beach and wetland restoration, home elevation, dykes, local levees, sandbags, etc. We estimate the cost of these actions to range between a couple of millions to billions of USD based on [103, supplementary]. The investment ranges for agents are determined such that the maximum continuous investment from all agents in 40 years prevents any cost from nature until 2060, e.g., 10-foot home el-

evation or 20-foot dyke all over the coastline. The investment cost parameters are given in Table 3.4 along with the insurance and natural cost parameters.

### 3.3.2 Scenario Simulations

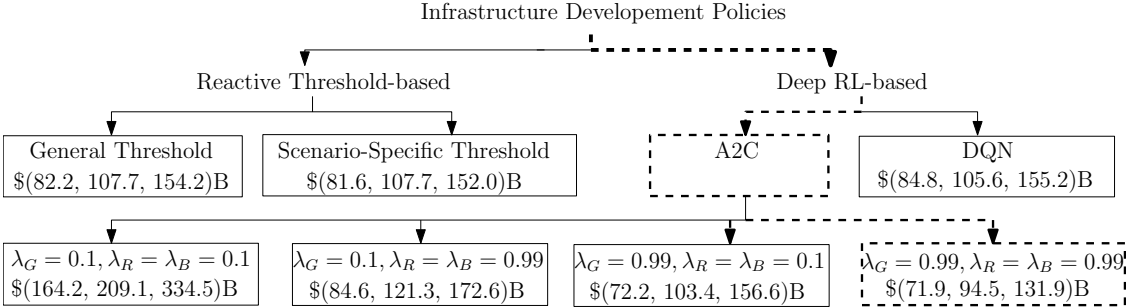


Figure 3.7: Expected episodic cost under different policies for high SLR scenario.

To benchmark the performance of the “proactive” deep RL framework, we also consider a straightforward “reactive” policy that makes an infrastructure improvement only after its need is proven by high natural cost. In a reactive community, the government and other stakeholders generally become active in infrastructure development after a major natural disaster. This trend can be portrayed through a threshold-based policy, where an agent invests in infrastructure if the cost from nature exceeds a predefined threshold. Although simple, this policy is not tractable for generating simulations that represent realistic stakeholder behaviors because of the difficulty in selecting the thresholds for natural cost. Whereas, the cooperation indices in our simulation tool can be intuitively varied between zero and one to jointly simulate the adaptation strategies of different stakeholder prototypes.

Fig. 3.7 presents the total cost for all stakeholders over the 100-year period 2020-2120 when the threshold-based and deep RL policies are deployed. The values within the parenthesis indicate the cost for Intermediate Low, Intermediate, and High SLR, respectively. The three values in parentheses are the total cost over 100 years in billion US dollars considering the intermediate-low, intermediate, and high SLR projections of NOAA, respectively. The threshold-based policy is used in two forms: the general threshold policy represents the case where the agents are agnostic to SLR projection scenario in the simulations, and the

scenario-specific threshold policy corresponds to the case where the best threshold is used for each SLR scenario. In both threshold-based policies, all three agents take the maximum investment action shown in Table 3.4 once the cost from nature in a year exceeds the same threshold. This common threshold is optimized for each scenario in the scenario-specific policy, and for the average of three scenarios in the general threshold-based policy to demonstrate the best performance such threshold-based policies can attain (Fig. 3.8). The vertical dashed line represents the best general threshold for the average SLR scenario. As shown in Fig. 3.7, the proposed deep RL policy based on the A2C algorithm can intuitively simulate a variety of stakeholder prototypes by varying their cooperation indices between zero and one. While the fully non-cooperative case with all three cooperation indices equal to 0.1 results in huge costs, double the costs of the best threshold-based policy, the fully-cooperative case with cooperation indices equal to 0.99 reduces the total cost by 13% with respect to the best threshold-based policy. The costs presented for the DQN-based policy are for the fully cooperative case ( $\lambda_G = \lambda_R = \lambda_B = 0.99$ ). They are significantly worse than their counterparts in the A2C policy due to the high variance in the cost from nature. Finally, Fig. 3.9 shows the cumulative yearly cost for the high SLR scenario for each policy.

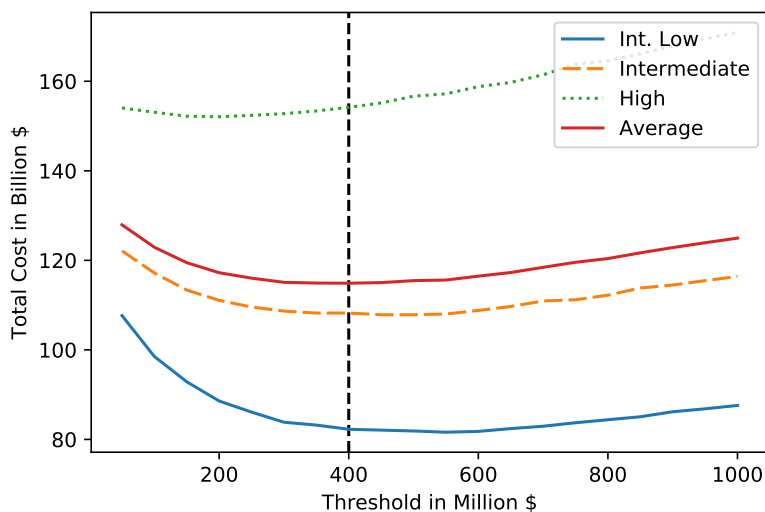


Figure 3.8: 100-year total cost for the intermediate-low, intermediate, high scenarios of SLR, and their average.

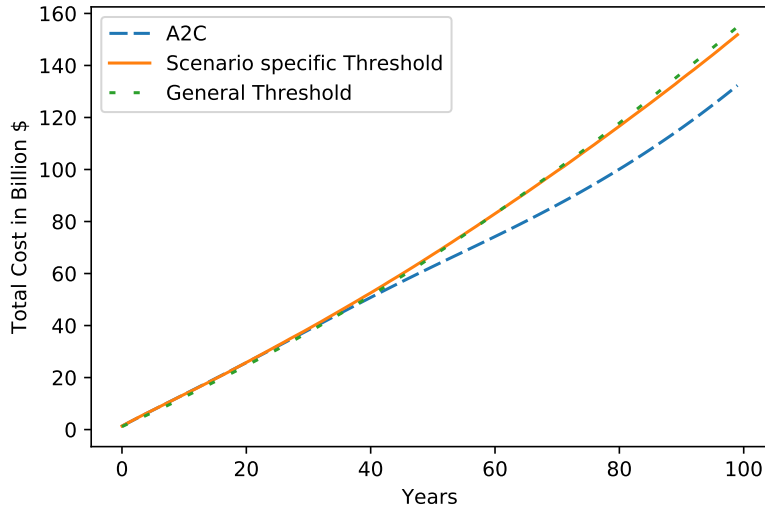


Figure 3.9: Yearly total cost under different policies for the high SLR scenario.

### 3.4 Discussions and Conclusion

In this work, we presented how a socioeconomic system around the sea-level rise (SLR) problem can be modeled as a Markov Decision Process (MDP) and simulated using Deep Reinforcement Learning (RL) algorithms. In addition to providing a general scenario planning tool to investigate the cost-benefit analysis of natural events and stakeholders’ investments, the proposed framework also illustrates, through a case study for the Tampa Bay region based on real data, how optimizing the adaptation strategies can effectively minimize the total cost due to SLR. Being the first in the literature, the proposed MDP model relies on some simplifying assumptions.

For example, we assumed a uniform cost-benefit economic model for the adaptation actions to represent the natural disaster cost with a tractable model with respect to the taken adaptation actions so far (i.e., setting the scale parameter of the generalized Pareto distribution as a function of the sea level and infrastructure state). In the uniform model, the action level (Table 3.4) also determines the development level in the infrastructure. A detailed cost-benefit model for different actions can easily replace the considered uniform model. Specifically, for a set of actions such as beach restoration, raising roads, building

seawalls, and relocating coastal properties, the cost levels and development levels can be non-uniformly set after a detailed study. Another feasible improvement over the proposed framework would be to represent residents and businesses with multiple independent agents and consider local infrastructure improvement actions for each subregion defined by a resident/business agent. Such an extension will increase the number of agents and the number of actions to decide for the government while the structure of the overall model remains the same. Note that the proposed multi-agent MDP model is not restricted to the RL policies; any action policy can be followed by the agents. The sequential and agent-based structure allows for a turn-taking game mode, where each agent decides on its action sequentially in a round, and at the end of the round nature imposes its cost on the agents. We developed a board game in which players can cooperate on adaptation strategies to mitigate SLR-related damages from nature [155]. A digital and improved version of the game is planned as a future work.

## Chapter 4: Multi-Objective Reinforcement Learning Based Healthcare Expansion Planning Considering Pandemic Events

### 4.1 Introduction

<sup>4</sup>Healthcare is a universal need that includes health promotion, prevention, treatment, rehabilitation, and palliative care. The distribution, management, operation, augmentation, and demand of healthcare facilities have become a delicate and crucial reality for our existence in this world. Pandemic events such as COVID-19 have highlighted the lack of a resilient and sustainable augmentation plan for healthcare facilities, even in developed countries. High population growth and the nationwide increase of median age in the US [151] indicate the need for widespread healthcare facilities.

The dynamics of hospital bed demand results from a wide range of stochastic variables, making it very challenging to model the future demand and augmentation scenarios. Emergency department crowding, natural disasters, and humanitarian crises are often not adequately addressed in the current annual development plans. Augmentation plans based on the yearly demand statistics can often be misleading [41]. The number of beds is often increased by observing the local population's needs, known as the Certificate of Needs (CON). In many states in the US, the hospital bed capacity is regulated based on the CON [54]. This method aims to maintain a target occupancy level of hospital capacity to minimize expenditures. The large number of casualties caused by the COVID-19 pandemic proved that this method has limitations in forecasting future bed demands. Lack of treatment often causes irreparable damage to the patients and families, physically and psychologically. Therefore,

---

<sup>4</sup>Portions of this chapter were published in IEEE Journal of Biomedical and Health Informatics [123]. Copyright permissions from the publishers are included in Appendix B.

hospital bed demand forecasting and facility augmentation planning need meticulous attention from the planners, administrators, and research community to ensure sustainable and accessible healthcare for all.

This study aims to address this research gap in hospital capacity expansion planning, especially under pandemic events like COVID-19. We include critical features in forecasting hospital bed demand for making augmentation plans. Firstly, different age-based population groups (e.g., infants, older people) require different levels of hospitalization, hence the age distribution is a primary factor in hospital occupancy forecasting. Secondly, Disease Burden (DB), which represents the hospital dependency of the residents in a region for critical diseases, is another major factor for hospitalization requirement. Moreover, the Social Vulnerability Index (SVI) of a region represents the vulnerability of its residents to diseases. Finally, a pandemic event is another factor that shapes the hospitalization need.

Beside the human health factors, the economics of maintaining hospitals should also be considered in the augmentation plan as the demand and supply in this sector is non-trivial. Maintaining an enormous capacity to meet uneven demand is not economically sustainable. Furthermore, cost of goods and services vary region to region because of transportation costs, tariff/taxes, or other reasons. Different administrative regions control the prices of goods/services in different ways, which can be summarized by the regional price parity index (PPI). PPI measures the cost of goods and services compared to the national average, making it a good regional cost indicator for establishing, expanding, and operating a hospital. Beside serving the health needs of community, for-profit and non-profit hospitals are significant revenue and employment providers locally and nationally. An oversight to these critical health and economic factors while devising an augmentation plan can significantly harm a region's health and economy. A robust, dynamic, and detailed hospital augmentation plan can benefit both the government and private parties, underscoring the scope of this work for the policymakers.



For a sustainable solution to this highly stochastic problem [115, 58], we utilize the important factors discussed above in a systematic artificial intelligence (AI) framework, deep reinforcement learning (RL). We propose a Multi-Objective Reinforcement Learning (MORL) method based on deep neural networks to satisfy two objectives: minimize augmentation cost and Denial of Service (DoS) to the patients. In our preliminary work [122], we proposed an RL approach that converts the patients’ discomfort caused by DoS into monetary cost through fixed coefficients. However, defining fixed coefficients for different places and periods is not feasible, causing a practical challenge for the applicability of the work [122]. To this end, in this work, motivated by the Pareto-optimal Q-learning (PQL) method [144] we propose multi-objective actor-critic method to avoid the forced conversion of DoS discomfort to monetary cost. Since we need to deal with high-dimensional state and action spaces for hospital augmentation planning, we utilize deep neural network based approximations for the MORL task, similar to [105, 148].

In the proposed method, the healthcare authority is the MORL agent that selects a region for hospital augmentation at each decision time (e.g., annually) by considering several important factors, the age-partitioned population, DB, SVI, PPI, and the existing hospital capacity for all regions. As a result of its augmentation actions, the agent observes the DoS and expansion costs. We modify Advantage Actor-Critic (A2C) [133], a popular deep RL algorithm, to address the considered MORL problem. The contributions of this work can be summarized as follows.

- A novel Markov decision process (MDP) formulation based on important health and economic factors, such as DB, SVI, and PPI, is proposed to learn the optimal hospital capacity expansion policy which minimizes the expansion cost and DoS.
- A novel deep MORL algorithm is developed based on the actor-critic framework.
- An extensive case study is performed for the state of Florida using real data to evaluate the proposed MORL approach.

The rest of the chapter is arranged as follows. Section 4.2 presents literature review for hospital capacity expansion planning. We present the MDP formulation in Section 4.3, and the proposed deep MORL algorithm in Section 4.4. The experimental setup, results, and discussions are given in Section 4.5, Section 4.6, and Section 4.7, respectively, and the chapter is concluded in Section 4.8.

## 4.2 Related Work

Proactive planning to address hospital bed occupancy problems and future expansion decisions under population changes and emergencies have been a critical problem for hospitals and care providers. The challenges in hospital bed management and expansion decision have been approached by several researchers based on the different understanding of the problem [101, 115]. The hospital bed occupancy and expansion decision literature can be divided into two major areas: (1) bed occupancy management and allocations within a hospital and (2) capacity planning and allocation of the hospital beds within a region. In the first type of study, researchers typically proposed a mathematical framework addressing systemic issues such as overcrowding within the hospital settings focusing on optimum use of healthcare resources that maximize bed usage and reduce boarding time. These studies include forecasting hospital bed occupancy and resources, healthcare personnel and critical resource allocation, and patient allocation and ambulance diversion [14, 21]. Prior studies in this area are widely varied by hospital division (e.g., psychiatric, emergency medicine, and maternity ward), care delivery setting (e.g., trauma hospital, children hospital, and specialty care), forecasting horizon (e.g., one hour to seven days), hospital resources (e.g., ICU bed, ventilation equipment, and physicians), patient case-mix setting (e.g., children, elderly, and pregnancy) and data-source (e.g., EMR, EHR, and clinical data) [23, 75, 99, 161]. However, the majority of these forecasting and resource allocation-based studies focused on supporting optimal use of crucial healthcare resources within the hospital setting rather than long-term bed expansion planning. Given the importance of long-term hospital bed

capacity and geographical allocation, we focus our study on models intended to address hospital bed expansion within a region considering the increased demand, shifts in population demographics, and emergencies such as COVID-19.

There are a few existing studies that considered capacity and expansion decisions for the medium or long-term planning horizon. These studies implemented various forecasting methods, including the simple ratio method, formula method, Michigan’s bed need model, and usage projection model to predict estimates at different regional settings (e.g., county, city) [138, 115]. Implementing these methods into long-term hospital expansion decision-making might lead to several critical limitations [101]. First, they do not consider the importance of complex interactions between the hospital bed need and population demographics changes, which might play a fundamental role in determining decisions [81]. Second, most of the studies faced challenges in forecasting accuracy, model fitting (e.g., over and underestimating), and incorporating geographical and hospital administration variations. It is suggested that an alternative robust decision support model incorporating uncertainty might have the potential to reliably predict hospital capacity planning and bed extension decisions for the regions with rapidly changing demographics and patient case-mix population [58, 67].

A promising alternative approach for hospital expansion decision consists of modeling with data-driven approaches and constrained optimization in the decision-making framework [138, 73]. A few studies showed potential with improved prediction performance for forecasting bed occupancy in various hospital settings and geographical regions through the data-driven forecasting approach [61, 74]. These studies used various statistical and machine learning (ML) methods, such as linear regression models. However, most of these studies are limited to forecasting, considering only the patient volume. A few studies utilized several neural network-based algorithms in forecasting intensive and critical care bed usage, surgical room prediction, and overall bed capacity estimations [74, 111]. However, only a handful of studies implemented ML-based methods to investigate hospital expansion planning at the regional and state level.

With the recent theoretical and technical achievements in RL approaches, the application of deep RL methods can potentially integrate prediction models with optimizing conflicting multiple objectives. Therefore, RL-based methods have been widely used in various applications areas, including robotics, virtual reality, finance, communications, and transportation [119, 60]. The applications of RL-based methods in the healthcare domain are mainly focused on adverse outcome predictions, rather than healthcare policy-related decision making [156]. Based on the RL-based studies in non-healthcare settings, RL-based algorithms have the potential to improve the hospital augmentation design with capabilities of incorporating multiple decision criteria and critical covariates under the same framework [53]. The works in [121, 120] utilized RL to determine the optimal size of hospital capacity augmentation; however, these methods neglect the fact that the capacity expansion usually happens in bulk numbers (e.g., 120-bed extension unit) [68, 47]. Furthermore, they did not consider interactions between covariates (e.g., patient case-mix and changes in demographics) and appropriate health administration division, which may significantly influence the hospital bed augmentation decisions [39]. Also, these studies assumed a single isolated objective and demand targets that are not necessarily main factors for hospital bed expansion decisions [81, 44]. Unlike previous approaches, our study deals with multiple decision criteria for hospital expansion decision making. In particular, our method aims to simultaneously minimize the capacity expansion cost and the number of service denials.

### 4.3 Proposed Decision Model

The proposed MORL method is based on a Markov decision process (MDP) specifically designed for the considered healthcare expansion planning problem. The proposed MDP formulation follows the Markov property: transition to the next state depends only on the agent’s current state and action. Fig. 4.1 shows our multi-objective MDP (MOMDP) model, where the healthcare administration is the agent which manages the healthcare facilities in  $R$  regions. The system state at time  $n$ ,  $\mathcal{S}_n$ , is defined by the non-controllable state variables

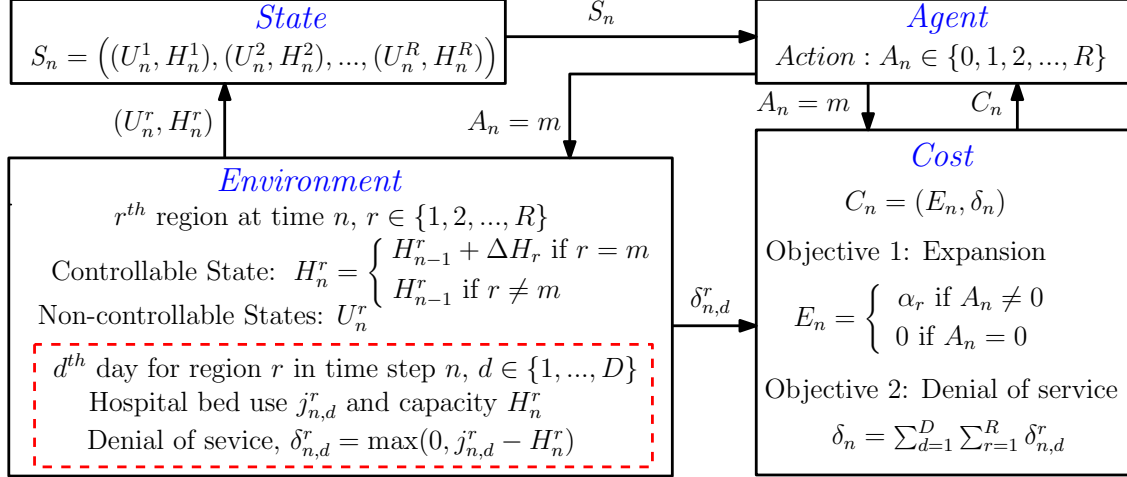


Figure 4.1: Proposed multi-objective MDP (MOMDP) model.

$U_n^r$  and the controllable state (hospital bed capacity)  $H_n^r$  of each region  $r$ . Every time step  $n$ , the agent takes action  $A_n = m$ , which means selecting the  $m^{\text{th}}$  region for a capacity expansion of  $\Delta H_m$  beds at the cost of  $E_n = \alpha_m$ . The agent can decide to decline expansion ( $A_n = 0$ ), resulting in a total of  $R+1$  decision options. The number of patients requiring hospitalization for a day can be greater than the region's capacity. In those days, some patients face with denial of service (DoS).  $\delta_{n,d}^r$  denotes the number of DoS for region  $r$  in day  $d \in \{1, 2, \dots, D\}$  of time step  $n$ , where  $D$  is the total number of days in a time step. The agent has two conflicting objectives:

- Minimize the cumulative capacity expansion cost,  $\sum_n E_n$ ,
- Minimize the total DoS,  $\sum_n \delta_n = \sum_n \sum_d \sum_r \delta_{n,d}^r$ .

While the expansion actions ( $A_n \neq 0$ ) incur monetary cost, they also reduce the future number of DoS. The agent aims to simultaneously minimize the monetary cost and DoS (i.e., find an optimum balance between them) over a finite time horizon by taking optimal actions. Before explaining our MORL solution to this problem, we next elaborate the state variables and the cost function.

### 4.3.1 State

#### 4.3.1.1 Non-controllable States

Selecting appropriate variables for the state definition is a critical task in MDP formulation. The agent gathers essential information from the environment by observing several state variables to inform its actions. The variables which are not directly affected by the agent’s actions are called the non-controllable states in our model. The required hospital capacity has a strong correlation with the following factors, which we choose as non-controllable states for our model.

- The work [122] shows that the number of hospital admissions is better captured by age-grouped population data, consistent with the general understanding that some age groups require more medical attention (especially children and older people). The age-partitioned population of the  $r^{th}$  region at time  $n$ ,  $p_n^r = [p_n^{r1}, \dots, p_n^{rG}]$ , is a vector of  $G$  age groups. We separate the population among 4 age groups: 0-18, 18-44, 44-65, and 65+ years for our case study.
- Disease Burden,  $DB_n^r$ , represents the age adjusted death rate per 100,000, which ranges between 450 and 1600 for the regions in our experiment.
- Social Vulnerability Index,  $SVI_n^r$ , represents the vulnerability of the population towards diseases and ranges between 0 and 1. SVI is a surrogate measure of the potential negative effects on communities caused by external stresses on human health [48]. Another relevant measure Health Deprivation Index (HDI) is available primarily at the census block group level, which can be used for fine-grained modeling at the neighborhood level. Since larger regions for healthcare administration are considered in this work, we prefer SVI, which is available at the county or census tract level.
- Price Parity Index,  $PPI_n^r$ , represents the cost of living in a region normalized by the national average.

- During pandemics the healthcare system allocates part of its capacity to deal with those pandemic-affected patients, which significantly changes the environment. Therefore, we include it in the non-controllable states as a single binary variable  $\text{Pand}_n \in \{0, 1\}$ . This pandemic flag may also cover other humanitarian crises due to natural disasters or other catastrophic events.

The agent only observes these states from the environment, but cannot control them. In the experiments in Section 4.5, we explain how to reliably estimate these state variables using real-world data. We include variance in the estimated values for these non-controllable states to simulate a realistic environment in the case study.

#### 4.3.1.2 Controllable States

The agent’s action controls the hospital bed capacity for each region, which is the only controllable state in this setup. The current hospital capacity for the  $r^{\text{th}}$  region at time  $n$  is given by

$$H_n^r = H_{n-1}^r + \Delta H_n^r = H_0^r + \sum_{\tau=1}^n \Delta H_\tau^r,$$

where  $\Delta H_n^r = \Delta H_r$  if the region is selected for capacity expansion ( $A_n = r$ ), otherwise  $\Delta H_n^r = 0$ . The expansion size  $\Delta H_r$  may vary among the regions.  $H_0^r$  is the initial hospital capacity for the region at the beginning of the study.

#### 4.3.2 Cost

Since we have two objectives, the cost in this MOMDP setup is a vector  $C_n = (E_n, \delta_n)$ .

##### 4.3.2.1 Expansion Cost

The agent can implement capacity expansion by building a new hospital or augmenting an existing facility. The different capacity expansion size  $\Delta H_r$  for each region incurs the expansion cost  $\alpha_r$ . We assume the healthcare authority has the proper understanding to

determine these parameters in practice, as demonstrated in our case study. Hence, the capacity expansion cost  $E_n = \alpha_r$  varies for different actions  $A_n = r$ , and  $E_n = 0$  for the no capacity expansion decision.

#### 4.3.2.2 DoS

The per capita (per 1000 people) hospital bed capacity varies widely among countries, e.g., Japan has 13 hospital beds per capita while Mali has only 0.1 [150]. The US has a moderate per capita of 2.5, where South Dakota leads the chart with 4.8 in comparison with Oregon’s 1.6 per capita hospital bed [20]. For any region in the world, the actual hospital admission on a given day can be more than the available capacity, especially during pandemic times such as COVID-19. Since it is not financially feasible to maintain capacity capable of providing healthcare for all scenarios, the healthcare authority tries to maintain a reasonable capacity. However, the patients living in lower per capita capacity regions are prone to more frequent DoS. The DoS for the  $r^{th}$  region is

$$\delta_{n,d}^r = \max(0, j_{n,d}^r - H_n^r),$$

where  $j_{n,d}^r$  is the hospital admission requirement for day  $d \in \{1, 2, \dots, D\}$  at time step  $n$ . So, the total number of DoS for the  $n^{th}$  time step is

$$\delta_n = \sum_{r=1}^R \sum_{d=1}^D \delta_{n,d}^r.$$

## 4.4 Solution Approach

### 4.4.1 Multi-Objective Reinforcement Learning

In a Reinforcement Learning (RL) setup, the agent takes an action that changes the environment, and the environment responds by providing an immediate reward/cost. In the standard setting, the goal of the RL agent is to maximize the discounted cumulative



reward  $R_N = \sum_{n=0}^N \gamma^n R_n$  by taking optimal actions over a time horizon of  $N$  steps. The discount factor  $\gamma \in (0, 1)$  determines the weight of future rewards/costs relative to the immediate one for the RL agent. The traditional way to obtain a scalar reward from the multiple costs present in the original objectives (i.e., expansion cost and DoS in our case) is to combine them using a conversion parameter,  $R_n = -E_n - \beta\delta_n$ . There is a significant challenge in setting the conversion parameter  $\beta$  to a realistic value since it is in general not obvious what the conversion rate should be. Specifically, DoS is a health-related discomfort cost for the patients, which is not easy to convert into a monetary cost like the expansion cost. Although one can find studies that try to assign economic value to such an important discomfort cost, there is no unique and optimum way of doing this. Avoiding such a forced conversion, we treat each objective in a natural way through a deep MORL algorithm.

To this end, instead of a single value function used for the scalarized cost in the traditional RL approach, we define two value functions for the expansion cost and DoS, which are given by the Bellman equations [133]

$$\begin{aligned} V_E(S_n) &= \max_{A_n} \{ \mathbb{E} [-E_n + \gamma_E V_E(S_{n+1}) | S_n, A_n] \}, \\ V_\delta(S_n) &= \max_{A_n} \{ \mathbb{E} [-\delta_n + \gamma_\delta V_\delta(S_{n+1}) | S_n, A_n] \}. \end{aligned} \tag{4.1}$$

The value functions  $V_E(S_n)$  and  $V_\delta(S_n)$  represent the maximum expected reward at a certain state achievable by taking the optimum actions in the current time step and in the future.

#### 4.4.2 Deep MORL

Our MOMDP model consists of  $8R$  states and  $R + 1$  actions for  $R$  regions. This high-dimensional state-action space requires neural network (NN) based approximations to learn the value functions in Eq. (4.1). The NN-based RL approaches are called deep RL, and Advantage Actor-Critic (A2C) is a popular deep RL technique. A2C is known to be more successful for high-dimensional action space than its most prominent alternative Deep Q

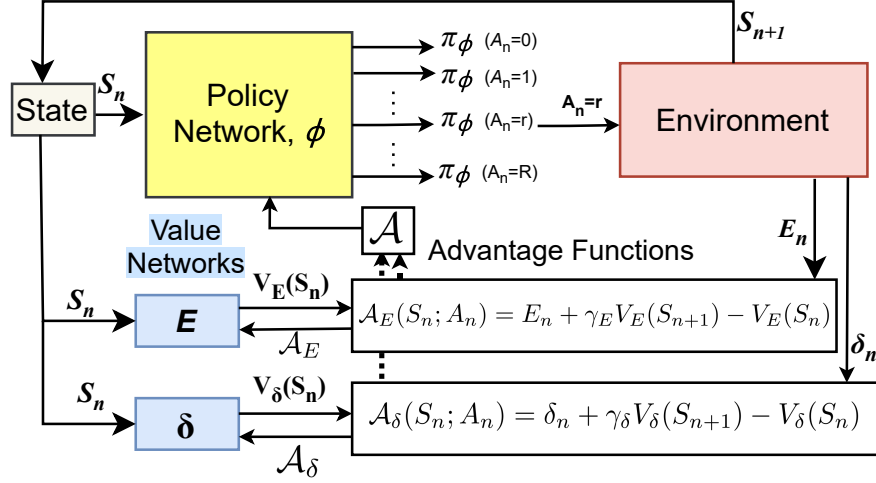


Figure 4.2: Proposed multi-objective A2C architecture.

Network (DQN) [133] and thus is suitable for our problem. A2C uses a function called the advantage function for policy update to address the high variance problem of its predecessor, the REINFORCE algorithm [133]. We propose a multi-objective A2C algorithm for the considered MORL problem, following the Pareto optimality approach [148]. Fig. 4.2 shows the proposed multi-objective A2C architecture. The pseudo code is also given in Algorithm 4.1. A2C uses two different type of networks, the actor network and the critic network. In our multi-objective A2C architecture, while there is a single actor network for action decisions, we utilize two critic networks for the two objectives, as explained next.

#### 4.4.2.1 Critic Networks

The two *value networks* for the two objectives aim to learn the value functions  $V_E(S_n)$ , and  $V_\delta(S_n)$  for a given state  $S_n$ . Based on the agent's action  $A_n$ , the target values are estimated from the immediate cost and the value function for the next state  $S_{n+1}$ . Then, the advantage functions for the state action pair  $(S_n, A_n)$  are calculated as the difference between

the target values and the predicted values:

$$\begin{aligned}\mathcal{A}_E(S_n; A_n) &= -E_n + \gamma_E V_E(S_{n+1}) - V_E(S_n), \\ \mathcal{A}_\delta(S_n; A_n) &= -\delta_n + \gamma_\delta V_\delta(S_{n+1}) - V_\delta(S_n).\end{aligned}\tag{4.2}$$

The value networks use the advantage functions as the temporal difference error for gradient descent update through backpropagation. They are called *critic networks* as they guide the policy network about the quality of its actions through the advantage functions.

#### 4.4.2.2 Actor Network

The *policy network* outputs probability  $\pi_\phi(A_n = r)$  for each action through a softmax function, i.e.,  $\sum_{r=0}^R \pi_\phi(A_n = r) = 1$ . It aims to maximize the expected return  $J(\pi_\phi)$  by performing gradient ascent with respect to the weights  $\phi$  of the NN through the following equation:

$$\nabla_\phi J(\pi_\phi) = \mathbb{E}_{\pi_\phi}[\nabla_\phi \log(\pi_\phi(A_n|S_n))(S_n; A_n)].\tag{4.3}$$

While updating the critic networks by their corresponding advantage functions is straightforward in this multi-objective setup, we define the following advantage function for the actor network

$$\mathcal{A} = \begin{cases} w_E \mathcal{A}_E + w_\delta \mathcal{A}_\delta, & \text{if } |\mathcal{A}_E + \mathcal{A}_\delta| = |\mathcal{A}_E| + |\mathcal{A}_\delta| \\ 0, & \text{otherwise,} \end{cases}\tag{4.4}$$

where  $w_E + w_\delta = 1$ . The coefficients  $w_E$  and  $w_\delta$  reflect the priority of the policymaker for the two objectives mentioned above. Such a flexibility is missing in [148]. Notably, the actor network is updated only when both advantage functions have the same sign (both positive or both negative), as seen in Eq. (4.4). This intuition is in line with the Pareto optimality discussed in [148], which prescribes to update only when the gradient ascent directions (advantage functions) corresponding to all objectives are the same. Updating in the same

---

**Algorithm 4.1** Multi-objective A2C algorithm (Fig. 4.2)

---

- 1: *Input*: discount factors  $\gamma_E$  and  $\gamma_\delta$ , objective weights  $w_E$  and  $w_\delta$ , learning rate  $\alpha$ .
  - 2: *Initialize* policy network with random weights  $\phi$  and the value networks with random weights  $E$  and  $\delta$ .
  - 3: **for** episode = 1, 2, ... **do**
  - 4:   *Initialize* the MOMDP, obtain the initial state  $S_0$ ;
  - 5:   **for**  $n = 1, 2, \dots, N$  **do**
  - 6:     Sample action  $A_n$ , from probability distribution generated by the actor network  $\phi$ .
  - 7:     Execute action  $A_n$ , and observe reward vector  $R_n = [-E_n, -\delta_n]$  and next state  $S_{n+1}$ .
  - 8:   **end for**
  - 9:   Calculate the advantage functions for the value networks from Eq. (4.2).
  - 10:   Update policy network  $\phi$  using the advantage function  $\mathcal{A}$  (Eq. (4.4)) in gradient ascent (Eq. (4.3)).
  - 11:   Update value networks  $E$  and  $\delta$  using their advantage functions  $\mathcal{A}_E$ , and  $\mathcal{A}_\delta$ .
  - 12: **end for**
- 

gradient ascent direction will discover new undominated points on the Pareto front. On the contrary, different gradient ascent directions for different objectives indicate the discovery of dominated points since they do not increase all objectives concurrently. Hence, we do not update the actor network when the advantage functions have different directions.

## 4.5 Experimental Setup

Having warm tropical weather, Florida is an attractive retirement home for an increasing number of older people in the US. Older people are more prone to medical care and longer stays in hospital. The high population growth in both infant and older age groups requires robust planning and expansion of healthcare facilities in Florida. So, we assess our MORL policy for Florida, where the Agency for Healthcare Administration (AHCA) can represent the MORL agent. AHCA grouped the 67 counties of the state in  $R = 11$  regions or health districts, as shown in Fig. 4.3. Diamond shape with number  $n$  next to the region label indicates capacity expansion decisions for that region by the proposed MORL in year  $n$  for both scenarios, whereas round shape indicates expansion for the pandemic scenario only.

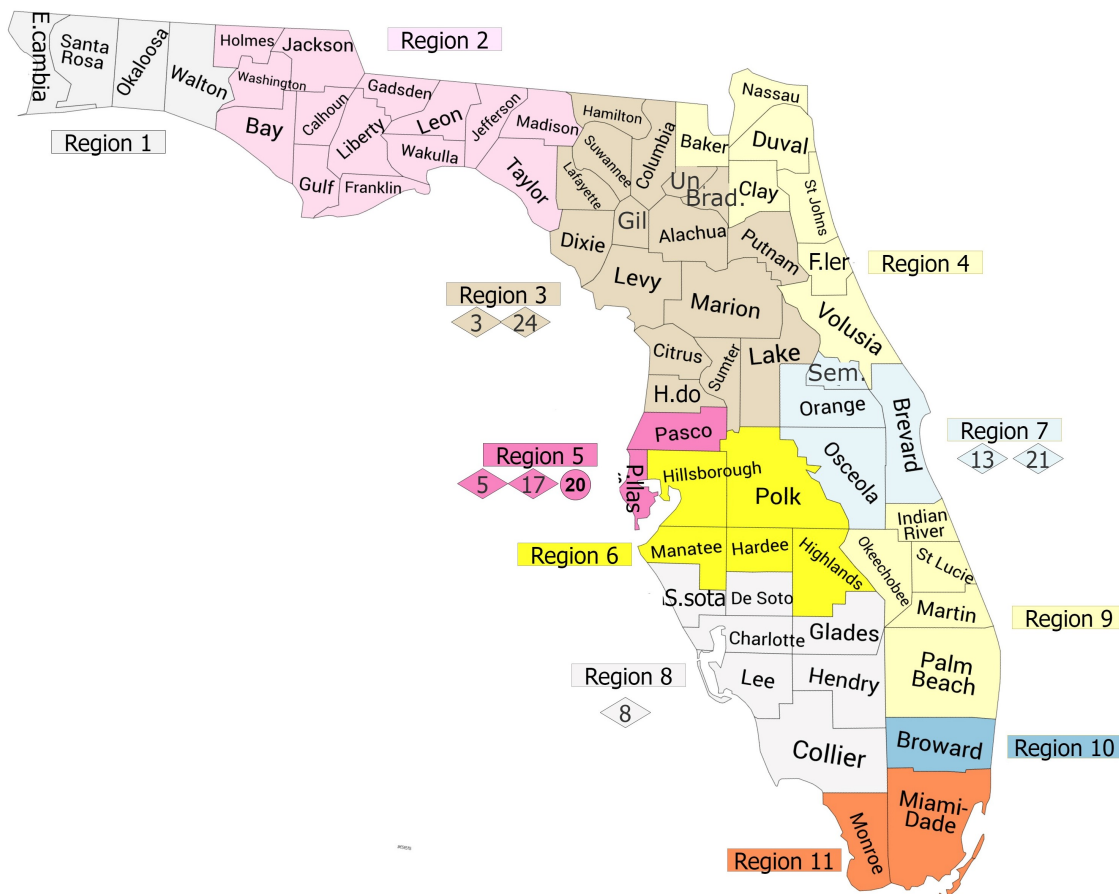


Figure 4.3: Map of the 11 health regions of Florida for the case study.

#### 4.5.1 Data Generation

The Bureau of Economics and Business Research provides Florida’s county-wise population history and projections up to the year 2045 [102]. We extracted historical hospital admission, Disease Burden (DB), and Social Vulnerability Index (SVI) data between 2010-2019 from State Inpatient Databases (SID) [1].

Although the US Bureau of Economic Analysis (BEA) publishes the state-wise Price Parity Index (PPI) [11], county-wise PPI data is yet to be published. Since PPI has a strong correlation with the household income of a region [11], we generate the PPI data for the  $r^{th}$  region in our case study as follows

$$PPI_n^r = \frac{HI_n^r}{Med(HI_n)}.$$

$HI_n^r$  is the household income for the region, and  $Med(HI_n)$  is the median of household incomes for all 67 Florida counties. We found that Region 6 is the costliest in Florida, which closely matches the BEA’s [11] map of real personal income and regional price parity map of the major metropolitan areas in the US. Hence, establishing and extending hospital facilities in Region 6 will be the costliest in Florida. We devise yearly decisions in each policy to expand the capacity of a region with  $\Delta b = 120$  hospital beds. We set the cost of adding 120 hospital beds with a normal distribution of mean  $\mu = 50$  and variance  $\sigma^2 = 3$  M USD for Florida [68, 47]. So, the expansion cost at the  $n^{th}$  time step depends on the PPI of the selected region as in

$$E_n = \alpha_r = PPI_n^r \times \mathcal{N}(\mu, \sigma^2).$$

Instead of projecting PPI values, we follow the PPI data of the year 2019 for the simulation period, i.e.,  $PPI_n^r = PPI_{2019}^r, \forall n$ .

#### 4.5.2 Hospital Occupancy Forecasting

Historical hospital admission for the regions was obtained from the Florida State’s health-care website [1]. Although the hospital admission requirement for an area depends on multiple factors, we hypothesize that the elements in our MOMDP state space  $U_n^r$  (except PPI) be sufficient to predict future hospital bed requirements. In this regard, Harrison et al. [58] shows the suitability of Poisson distribution in predicting hospital admission. Data shows higher hospital admission on weekdays than weekends on average [1]. So we fit the historical hospital admission data using the features (age-partitioned population, disease burden, social vulnerability index) within separate models for weekdays and weekends. Prediction accuracy above 90% with different regression algorithms shows the appropriateness of input features. Fig. 4.4 shows the prediction accuracies for different regression models, where data from 2010-2016 forms the training set, and 2017 data is used as the test set. We choose the best regressor (decision tree with Mean Absolute Error) to predict the Poisson distribution mean for weekdays and weekends in each region. Based on our observation of the hospital

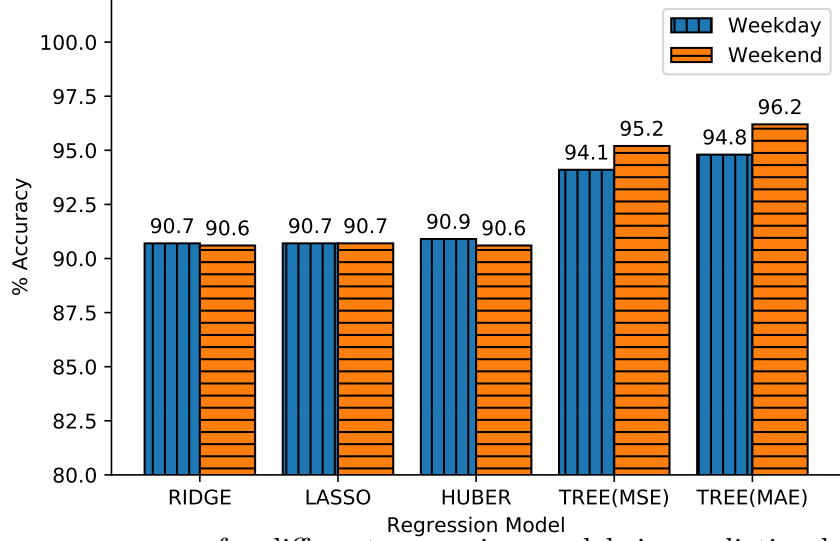


Figure 4.4: Average accuracy for different regression models in predicting hospital admission for weekdays and weekends based on data from [1].

admission data, to better account for the day-to-day variation, we add 20% Gaussian variance around the predicted value from the regression model to obtain  $\lambda_n^r$ . Finally, the number of beds required for a particular day in the  $r^{th}$  region is modeled as

$$j_n^r \sim \text{Poisson}(\lambda_n^r). \quad (4.5)$$

As the average length of stay per admission is 4.7 days throughout the US [20], we set the number of hospital bed requirements as 4.7 times the random number  $j_n^r$  generated in Eq. (4.5).

### 4.5.3 Scenarios

During the COVID-19 outbreak, the healthcare system allocated part of its capacity to deal with the pandemic-affected patients, decreasing the regular healthcare capacity. Hence, the healthcare authority needs to include pandemic scenarios in its policymaking. We define two scenarios with no pandemic and a 3-year long pandemic (between 20<sup>th</sup> – 22<sup>nd</sup> year) event within the 30 year decision time horizon (year 2021-2050). During the peak period of the COVID-19 pandemic, the average hospitalization in Florida was 12250, which is around 20%

of the hospital beds in Florida [12, 61], which we integrate into this case study. Specifically, in the pandemic years, 80% of beds will be available for regular healthcare, keeping the rest 20% reserved to handle the pandemic. The pandemic scenario may also cover other humanitarian crises due to natural disasters or other catastrophic events.

#### 4.5.4 Objective Priority

In the current practice, following the certificate of need (CON) process, the healthcare authority sets a threshold on the occupancy level to make the expansion decision, which implicitly represents their priority levels for the DoS and expansion cost objectives [101]. We reflect the healthcare authority’s priority levels for the two objectives in its healthcare expansion policy by explicitly considering varying weights for the actor network’s advantage function:

$$(w_E, w_\delta) = \left( (0.1, 0.9), (0.2, 0.8), \dots, (0.9, 0.1) \right). \quad (4.6)$$

These weight pairs respectively represent a range of policies from service-centric to cost-centric.

#### 4.5.5 Neural Network Architecture and Computation Time

We have one policy (actor) and two value (critic) networks in the A2C architecture for our MORL-based policy presented in Section 4.4. Fig. 4.5 shows the NN architecture of our method. We use a learning rate of 0.0003 and a discount factor of 0.99 for all three networks. Although the input state is the same for all 3 deep NNs, they have separate hidden layers to output the policy and value estimates. The hidden layers have 48, 120, and 48 neurons for all 3 deep NNs. The two value networks output one value for each value function estimate. However, the policy-network outputs  $R + 1 = 12$  action probabilities for the given state. For each combination of weights  $(w_E, w_\delta)$  in Eq. (4.6), both objectives



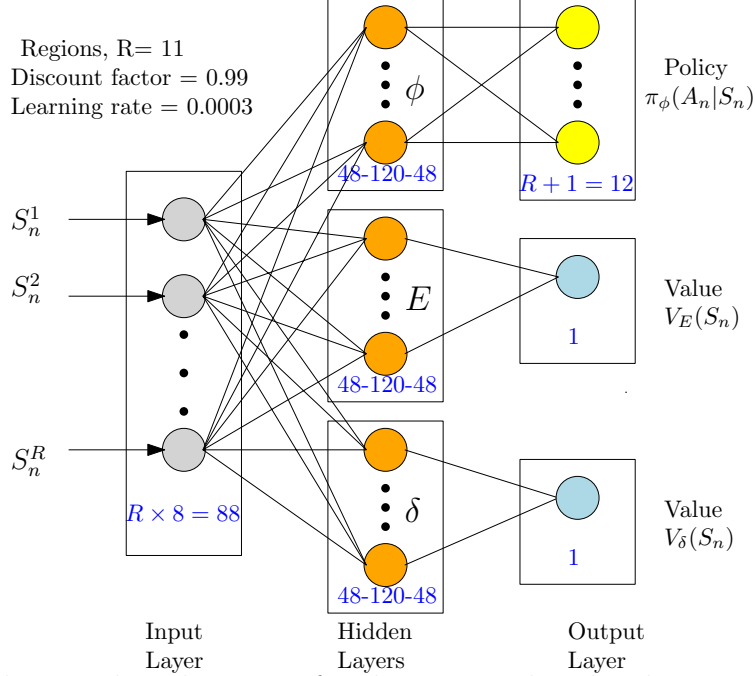


Figure 4.5: Neural network architecture for the proposed multi-objective deep RL-based policy.

Table 4.1: Computational details for the experiments.

Hardware	Software	Task	Computation time
Intel <sup>®</sup> Core i7 3.60GHz 16 GB RAM	Python 3.7	Data Preparation	5 min
	Pytorch 1.8.1	MORL Convergence	510 min
	sklearn 0.23.2	MORL Decision	0.33 sec

converge within 3,000 episodes, i.e., the agent learns the optimal policy after 3,000 runs. Table 4.1 shows the computation time for the proposed method. It takes 5 minutes for data processing, including hospital occupancy forecasting by using an Intel<sup>®</sup> Core i7, 3.60 GHz, 16 GB RAM computer. The MORL algorithm needs 510 minutes to perform the 3,000 episodes for convergence. Notably, the computational time for each decision is 0.33 seconds; negligible compared to our approach’s policy-making steps (i.e., 1 year).

## 4.6 Results

### 4.6.1 Proposed Deep MORL-based Policy

Our method provides a set of trade-off solutions for the healthcare authority. Depending on the objective weight range from Eq. (4.6), Fig. 4.6 represent objective priorities are

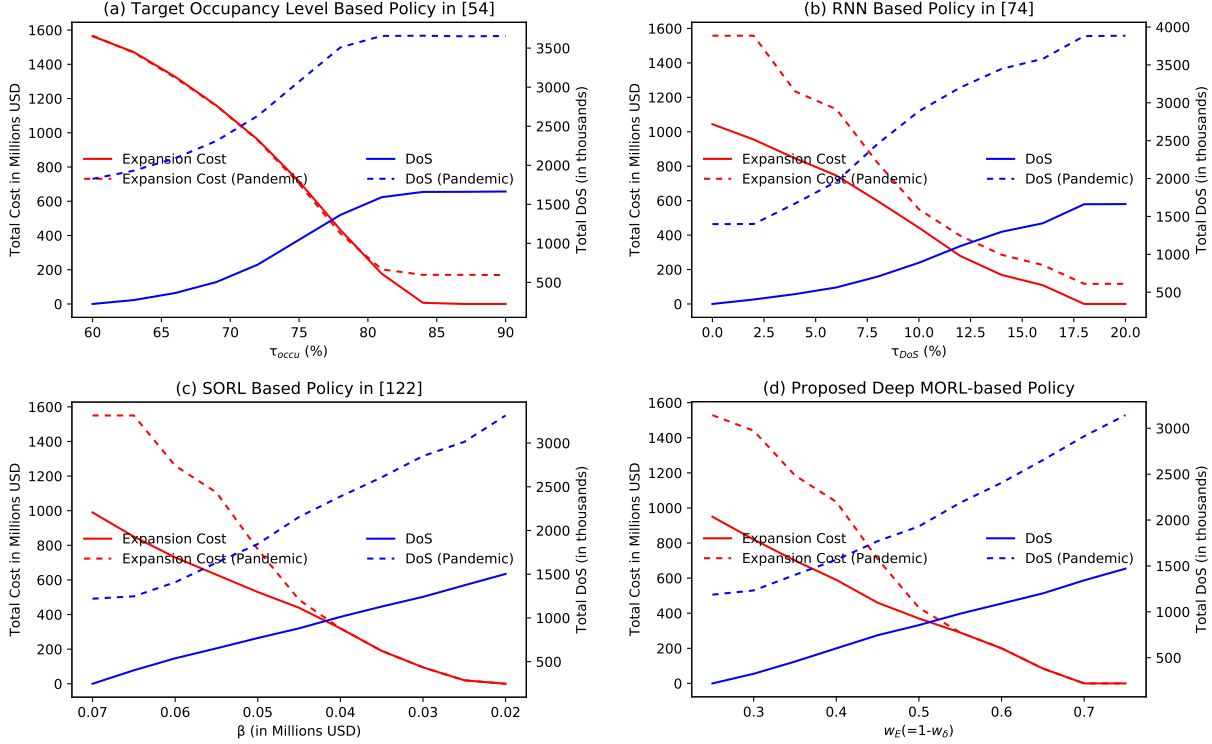


Figure 4.6: Episodic (30-year) total cost and DoS for healthcare authority under different objective priorities for the policies.

represented by (a) occupancy threshold levels for the target occupancy level based policy in [54], (b) DoS threshold levels for the RNN based method in [74], (c)  $\beta$  values for each DoS for the SORL based policy in [122], and (d) objective weights  $w_E (=1-w_\delta)$  for the proposed MORL based policy. Dashed lines represent the pandemic scenario. Fig. 4.6 (d) shows the cumulative expansion cost (left y-axis) and DoS (right y-axis) for the 30-year timeline obtained from the proposed deep MORL-based optimal policy. For the most service-centric healthcare authority ( $w_E = 0.2 < w_\delta = 0.8$ ), the expansion cost is 990 and 1550 Million USD, respectively, for the non-pandemic and pandemic scenarios over a 30-year period. The cumulative DoS for the regions is 220 and 1187 thousand, respectively, for the two scenarios. Although pandemic occurrence makes both costs worse, the obtained DoS is a lot more acceptable than the actual situation during the COVID-19 pandemic. With more emphasis on cost minimization ( $w_E > 0.2$ ), the DoS number goes up, and the expansion cost goes down, as shown in Fig. 4.6(d). For the most cost-centric policy in our setup ( $w_E = 0.8, w_\delta = 0.2$ ),

Table 4.2: Parameter, cost, and DoS comparison among the policies for different objective priorities (non-pandemic scenario).

Proposed Deep MORL			Target Occupancy [54]			RNN based [74]			SORL[122]		
$w_E$	Exp. cost	DoS(K)	$\tau_{occu}$	Exp. cost	DoS(K)	$\tau_{DoS}$	Exp. cost	DoS(K)	$\beta$	Exp. cost	DoS(K)
0.25	\$950 M	220	60%	\$1564 M	224	0%	\$1044 M	345	0.07	\$990 M	247
0.35	\$700 M	458	66%	\$1327 M	364	4%	\$846 M	475	0.06	\$730 M	538
0.45	\$460 M	745	72%	\$960 M	728	8%	\$598 M	706	0.05	\$530 M	770
0.55	\$290 M	980	78%	\$433 M	1364	12%	\$279 M	1108	0.04	\$320 M	1012
0.65	\$85 M	1201	84%	\$6 M	1659	16%	\$109 M	1410	0.03	\$95 M	1241
0.75	\$0 M	1470	90%	\$0 M	1664	20%	\$0 M	1663	0.02	\$0 M	1503

the healthcare authority makes no investment actions and endures 1470 and 3143 thousand DoS, respectively, for the two scenarios. The source codes are available at GitHub <sup>5</sup>.

#### 4.6.2 Benchmark Policies

We compare our deep MORL-based policy with a myopic policy from [54], a Recurrent Neural Network (RNN)-based policy from [74], and a single objective RL-based policy from [122] for a 30-year scheme. We selected decision thresholds to incur investment costs ranging from maximum to minimum for every policy to make a head-to-head comparison with our proposed MORL method.

##### 4.6.2.1 Target Occupancy Level Based Policy

Historically, many states regulated the number of hospital beds by the certificate of need (CON) process, under which hospitals could only expand under state review and approval. The CON process follows a target occupancy level of hospital beds as the decisive factor [54]. We select this method as a baseline policy where the region with the maximum percentile occupancy on the previous year is selected for augmentation. No augmentation action is selected if target occupancy for each region is lower than a threshold  $\tau_{occu}$ . We sweep the threshold  $\tau_{occu}$  over a range to represent priority over the objectives, as shown in Fig. 4.6(a). For lower thresholds ( $\tau_{occu} = 60\%$ ), the expansion cost is high, but DoS is low (service-centric) and vice versa (cost-centric) for higher thresholds ( $\tau_{occu} = 90\%$ ). The DoS is higher

<sup>5</sup><https://github.com/Secure-and-Intelligent-Systems-Lab/MORL-Based-Healthcare-Expansion-Planning>

for the pandemic scenario (dashed lines) than no-pandemic scenario over the entire range. However, the expansion cost is higher only for  $\tau_{occu} > 80\%$ .

#### 4.6.2.2 RNN Based Policy

Kutafina et al. [74] provide an RNN-based hospital occupancy forecasting method. They achieved an accuracy rate of 93.76 % on eight validation sets from a German hospital’s 13-year (2002-2015) hospital records data set. They included the day of the week, day of the year, public holidays, and school holidays as the features for the RNN. We include their method for the comparative analysis with the following adaptations:

- We include age-based population vector, DB, SVI, and pandemic flag on top of the features used in [74].
- We select a region for expansion that is predicted to have the most DoS based on the RNN forecast for the next step.
- We select no expansion if the DoS of the selected region is less than a threshold  $\tau_{DoS}$ ; otherwise, that region gets the capacity expansion.

The  $\tau_{DoS}$  indicates how much percentile DoS the healthcare allows, i.e., it does not make any expansion if all of the regions’ DoS is below that threshold. We do a sweep search for  $\tau_{DoS}$  between 0-20% that characterizes a range between service-centric to cost-centric healthcare authority as shown in Fig. 4.6(b). This method has a one-step look ahead benefit compared to the Target Occupancy Level method of [54]. Hence, it incurs higher expansion cost and less increase in DoS for the pandemic scenario throughout the threshold range compared to the Target Occupancy Level method.

### 4.6.2.3 Single Objective RL Based Policy

In our preliminary work [122], we converted the DoS into monetary cost by assigning DoS cost for each region each day as in

$$c_{n,d}^r = \begin{cases} \beta(j_{n,d}^r - H_n^r), & \text{if } j_{n,d}^r - H_n^r > 0 \\ 0, & \text{otherwise,} \end{cases}$$

where we selected  $\beta = \$0.04\text{M}$ , which represents the monetary cost equivalent of the discomfort of an unattended patient, based on the study [45]. This cost is summed up over all regions as the DoS cost for the time step  $n$  as

$$E_n^{DoS} = \sum_{r=1}^R \sum_{d=1}^D c_{n,d}^r. \quad (4.7)$$

The sum of the expansion cost and the DoS cost is used as the negative reward,  $R_n = -(E_n + E_n^{DoS})$  for the single objective RL approach in [122]. This method does not provide a handle over preference between the two objectives. Furthermore, the monetary equivalence for DoS is an abstract idea, and setting a universal value for  $\beta$  is impossible. In fact, this value can represent the mindset of the healthcare authority about how much it cares about the population. So, we use a range of values  $\beta = (0.07, 0.065, \dots, 0.02)$  that represents from service-centric to cost-centric policies with the decreasing value of  $\beta$  as seen in Fig. 4.6(c). The pandemic scenario incurs higher cost and DoS; however, the expansion actions are similar when the agent puts less value on  $\beta < 0.045$ .

### 4.6.3 Comparative Analysis

We conduct a comparative analysis among the policies mentioned above in terms of the total expansion cost and total DoS for the 30-year timeline (Fig. 4.7). The x-axis represents the objective priority that we generalized as cost-centric, moderate, and service-

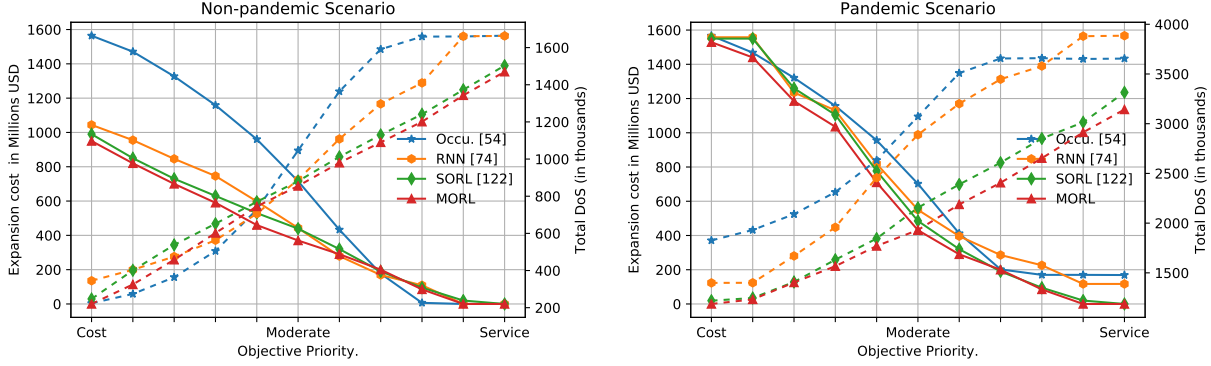


Figure 4.7: Episodic (30-year) total cost (solid lines) and DoS (dashed lines) for healthcare authority under different objective priorities for the policies for non-pandemic (left) and pandemic (right) scenarios.

centric (from left to right) based on the decision process range discussed and shown in Fig. 4.6. In particular, Fig. 4.7 puts all the policies' outcomes in a single frame to better understand the benefit of our proposed MORL-based policy. The target occupancy level based policy in [54] performs worst among the policies as it is always one step behind, i.e., its decision is based on the previous year's experience. The RNN forecasting based policy in [74] performs better mainly due to the inclusion of the observable state's data in the forecasting method. However, this policy in [74] lacks the RL mechanism to minimize the cumulative costs from all future states. The single objective RL-based policy in [122] and our proposed MORL perform better than the other two policies. However, our MORL based policy outperforms the SORL in [122] by utilizing more data (information) in its state definition. Especially with the PPI data, our method picks a less expensive region for expansion when there is a tie between two regions of different PPI levels. This better performance of MORL is more emphasized in the pandemic scenario, as shown in Fig. 4.7.

Fig. 4.8 provides a synonymous view of the Pareto front of trade-off solutions discussed in [144] for the initial state (i.e., year 2021). The x-axis represents expansion cost, and the y-axis represents the number of DoS achieved from the initial state (i.e.,  $n = 0$ ) for all the policies considering equal (moderate) priority between the two objectives. The ideal Pareto front would be a curve that no other policy can go under, i.e., no other points can decrease

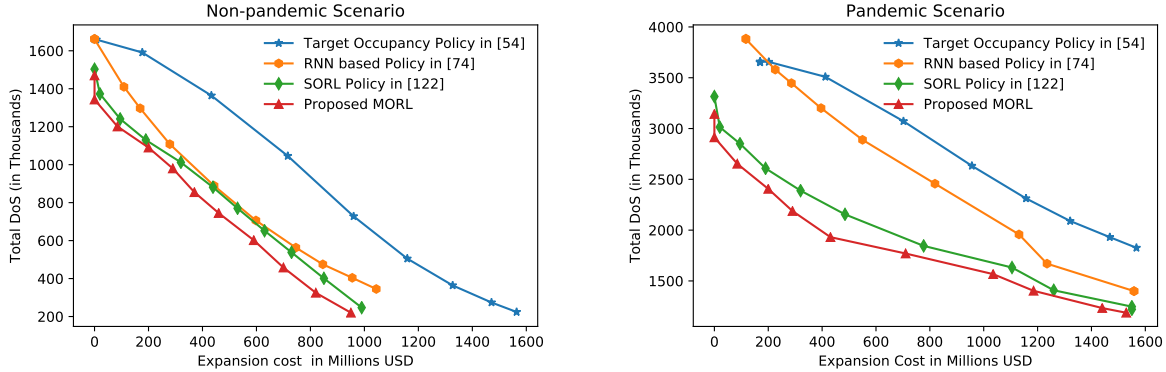


Figure 4.8: Trade-off solutions for each policy for the non-pandemic (left) and pandemic (right) scenarios with equal (moderate) objective priority.

expansion cost without increasing DoS and vice versa. Many works [144, 105, 148] focus on obtaining the ideal Pareto front for multi-objective optimization tasks, yet we can only approximate for our high dimensional state and action space problem. In Fig. 4.8, for the deep MORL based policy, each point on the curve is a non-dominated solution among the compared policies, which means none of the competing policies achieves better optimization for the corresponding objective priority level. The points in the curves refer to episodic (30-year) total cost and DoS for the healthcare authority at the initial state (i.e.,  $n = 0$ ). Furthermore, our method provides an easy-to-use and natural way to control the trade-off between the objectives through setting simple weights between zero and one ( $w_E$  and  $w_\delta$ ), whereas the SORL policy requires further studies to strike a desired balance between the objectives.

The comparative analysis is summarized in Table 4.3 for a healthcare authority that puts equal (moderate) priority for both objectives. Different actions in the pandemic scenario are shown in the parentheses. Our MORL based policy provides the lowest cost and DoS among the other policies. To compare the policies in more detail, their selected actions for the 30-year period are also shown in Table 4.3. The different actions in the pandemic scenario are given in parentheses. The target occupancy level based policy in [54] takes the same actions under both scenarios. The RNN based policy in [74] takes similar actions like in [54]; however, with its prediction capability it takes those actions early to prevent higher

Table 4.3: Expansion cost, DoS, and selected sequential actions for each policy over a 30-year period for equal priority on the two objectives.

Policy	Non-pandemic		Pandemic		Selected Sequential Actions (in year 1 to 30)
	Exp. cost	DoS(K)	Exp. cost	DoS(K)	
Target Occupancy [54]	\$716 M	1046	\$716 M	3072	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0, 5, 5, 7, 5, 7, 3, 2, 5, 2, 7, 5, 3
RNN Based [74]	\$443 M	889	\$550 M	2889	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0, 7, 5, 5, 0(7), 0(6), 7, 3, 2, 5, 2, 0, 0, 0
SORL Based [122]	\$439 M	881	\$485 M	2155	0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 7, 0, 0, 3, 5, 0, 2, 5, 7, 5, 0(7), 3, 0, 0, 0, 0, 0, 0
Proposed MORL	\$370 M	855	\$425 M	1931	0, 0, 0, 0, 5, 0, 0, 8, 0, 0, 0, 0, 7, 0, 0, 3, 5, 0, 2, 0(5), 7, 0, 0, 3, 0, 0, 0, 0, 0, 0

costs. SORL in [122] and the proposed MORL policy perform better as they are even more proactive in making preventive expansions. Both policies expand in the early years to keep the system in a balanced state in the future. Region 5 gets the most expansions, suggesting this region expects higher DoS in the future. However, the proposed MORL also selects Region 8 for expansion in the early years, which is one of the significant differences from the other policies. Since it is a larger region with a high population, as a result of that expansion, the overall DoS goes down significantly. The expansion decisions of the proposed MORL policy are also shown in Fig. 4.3.

4.7 Discussions

The key insights from the conducted study are

- The historical patient data based methods (e.g., target occupancy policy [54]) focus on the trend and lacks the root cause analysis (RL states) for future patient estimation.
- The RNN based policy [74] predicts the future hospital needs well; however, it lacks prescriptive analysis. It requires a dynamic DoS threshold for making decisions to adapt to different situations.
- Deep RL-based prescriptive analysis is suitable for the task as evident in experimental results for SORL [122] and our deep MORL method. As the SORL policy converts the DoS into a monetary cost with the help of a coefficient ( $\beta$  here), it requires literature to support spatial and temporal generalization for a suitable value of  $\beta$ . Our deep



MORL approach addresses this issue and provides an easy handle to the authority to set the relative weights for the two objectives.

While significantly improving the state-of-the-art, the proposed method also has certain limitations. For instance, it does not consider selecting multiple regions concurrently for expansion. Also, our method does not provide a mechanism for selecting different capacity expansion sizes for different regions in this current form. This research can be further utilized for allocating human resources such as physicians and healthcare personnel for a region. Private organizations often provide significant healthcare, hence including them in policymaking can provide the basis for a multi-agent RL setup. In that context, the reward function for the private organizations may include their financial benefit, and the central RL agent (healthcare authority) may consider the private facilities as buffer capacity to accommodate emergencies. Addressing these limitations and new scopes can provide several future research directions.

#### **4.8 Conclusions**

We proposed a multi-objective reinforcement learning (MORL) framework to develop a healthcare expansion plan and demonstrated its efficacy in a case study for Florida. The MORL method enables the user to conveniently set different weights for its two objectives, minimization of expansion cost and number of Denial of Service (DoS), in a natural way by only setting their priority percentages. Our data-driven approach is suitable for coping with the dynamic behavior of the region’s healthcare needs over a long period, especially to deal with emergency scenarios like pandemic events. We significantly improved our preliminary work in [122], which follows a single objective RL approach through converting DoS into monetary cost, by developing a multi-objective solution to enable intuitive objective priority setting; including Disease Burden (DB) and Social Vulnerability Index (SVI), apart from the age-partitioned population, for hospital occupancy prediction and making expansion

decisions; utilizing Price Parity Index (PPI) to accommodate different expansion costs for different regions.

## Chapter 5: Demand-side and Utility-side Management Techniques for Increasing EV Charging Load

### 5.1 Introduction

<sup>6</sup>Technological development throughout the previous decades paved the way for electric vehicles (EVs) to replace gasoline-based vehicles at an increasing rate. Specifically, the battery capacity and cost, which are the major impediments to EV adaptation, have been significantly improved.

As a result, today, governments, manufacturers, and customers are more convinced about EVs' environmental, commercial, and economic benefits, escalating EV popularity and adoption. According to Bloomberg New Energy Finance, which provides a comprehensive analysis of predictions from different entities like oil manufacturing companies and independent research groups [29], there are already 13 million EVs on the road globally, with 2.7 million sales in 2021. Following the planned expansion of charging infrastructure, EV growth predictions are mostly optimistic. For instance, the International Energy Agency predicts the total number of EVs will go over 250 million by 2030 from the estimated 5 million on the streets globally in 2018 [15].

While expanding the charging infrastructure is critical for large-scale EV adoption, a significant portion of daily EV charging occurs at homes and creates stress on distribution transformers (XFRs). Since EV charging requires significantly higher power than the other loads in a household, a combination of effective demand-side management (DSM) techniques for EV charge scheduling and utility-side management (USM) policies to cope with

---

<sup>6</sup>Portions of this chapter were published in IEEE Transactions on Smart Grid [127]. Copyright permissions from the publishers are included in Appendix B.

the increasing stress on the XFRs for XFR maintenance is needed. Utility companies try to flatten the electricity demand curve to decrease the stress on the distribution XFRs by providing day-ahead or hour-ahead dynamic electricity pricing schemes for the customers [107]. Numerous existing works proposed scheduling techniques for the time-shiftable appliances (e.g., Dishwasher, washer dryer, EV charging, etc.) of a household to capitalize the dynamic pricing [126, 160, 129].

Although such DSM techniques can flatten the demand curve to an extent, they do not sufficiently address the increasing stress of EV charging on the distribution XFRs since they lack the utility-side management of the problem. Motivated by this research gap, we take a comprehensive look at the problem of increasing EV charging stress on the distribution XFRs. Specifically, we consider both the demand-side (i.e., EV charge scheduling) and the utility-side (i.e., XFR maintenance) management of the problem. While the proposed DSM technique helps with load flattening to minimize transformer aging, the proposed USM technique enables timely (proactive) maintenance of distribution transformers to prevent costly transformer failures and blackouts.

### 5.1.1 DSM for EV Charge Scheduling

Centralized collaboration among EV users served by the same distribution XFR may provide the most effective way of minimizing the peak demand of the XFR [109, 131, 130]. The work [109] shows that coordination among the EV chargers under a distribution XFR minimizes peak demand to extend the XFR lifetime at the expense of consumers' arbitrage benefit. However, their approach lacks consumer comfort, ignoring that delayed EV charging may compromise user comfort. Another work [131] opts to minimize the EV owner's energy arbitrage benefit and distribution network maintenance cost through an optimal charging schedule. However, the objective function of this work also lacks user discomfort due to delayed charging. The chapter [130] proposes a fuzzy logic system for the demand-side operator

to devise a centralized EV charging schedule. This approach is too strict to accommodate user preferences and needs more adaptability to serve different types of customers.

In short, these techniques lack integrating customer preferences into their objective functions, hence may suffer in real-life implementation. We address this shortcoming by directly considering customer preference for charging duration and amount, and by introducing a monetary incentive to the customers based on their charging preferences (see Section II-C for details).

### 5.1.2 USM for XFR Maintenance

The distribution grid, especially the customer-end XFRs, is susceptible to overloading and costly maintenance. Replacement of gasoline cars by EVs urges installing charging stations in place of gas stations and home charging arrangements. So, the power system needs more energy generation, transmission, and distribution capacity at all levels. Many works provide charging station assessment, capacity, installation, and optimization techniques [86, 69]. In this work, we focus on EV charging at home, which EV users typically prefer due to the convenience and cheaper charging cost [92]. Home charging may significantly burden the customer-end distribution XFRs, as modern EVs take more than 7kW power from type-2 home chargers, higher than the average cumulative demand from all other loads in a household. Overloading an XFR leads to overheating and electrical insulation breakdown of an XFR [87]. IEEE guidelines provide estimation for effective aging due to overheating of the insulation [37]. The work [64] develops a probabilistic failure distribution that depends on the effective age of an XFR.

Transformer selection for replacement/upgrade is naturally a sequential decision making problem, requiring a solution that is adaptive to the observed states. Hence, dynamic programming (DP) techniques, which can optimize transformer selections according to the changing environmental factors such as EV charging stress, suit better to this problem than static optimization techniques. Since it is not tractable to model the future state transi-

tions (probabilistically or deterministically) as the network consists of many transformers and each action creates another branch of possible states, the model-based DP techniques like value iteration and policy iteration are not suitable. Reinforcement learning (RL) is a model-free DP approach that utilizes a data-driven technique of approximating a solution through sampling. Furthermore, deep RL (DRL) methods capitalize neural network-based function approximation to deal with the continuous-valued large input state (i.e., the current age and load of each transformer for our problem). Recent advances in neural network-based deep RL algorithms lead to widespread applications, including gaming [85], finance [72], energy systems [124], transportation [60], communications [90], environmental systems [119], and healthcare systems [123].

### 5.1.3 Contributions

We propose an EV integration policy for the utility company that aims to minimize the long-term maintenance costs for the electrical distribution grid. Our contributions can be summarized as follows.

- *The first comprehensive study of the problem of increasing stress on the distribution XFRs due to EV charging. Specifically, a combination of novel DSM and USM techniques is proposed for flattening the load curve and making timely maintenance of the distribution XFRs, respectively.*
- *A novel utility-driven EV charging scheme to flatten the load curve of the XFR. Different from the existing EV charging methods, our method directly considers customer preference for charging duration and amount, and a proportional monetary incentive.*
- *A novel DRL-based policy for XFR replacement and capacity upgradation to minimize the maintenance cost.*

The remainder of the chapter is organized as follows. Section 5.2 presents the proposed utility-driven EV charging method. Section 5.3 formulates the Markov decision process

(MDP) for the proposed DRL-based XFR maintenance policy. Experimental results and analysis are presented in Section 5.4 for a distribution XFR feeder. We conclude the chapter in Section 5.5.

## 5.2 EV Charge Scheduling for DSM

Our utility-driven EV charge scheduling offers a reasonable balance between peak load reduction and customer satisfaction. Utility companies offer lower electricity prices during off-peak hours to encourage consumers to shift their load towards those hours. However, this can create extensive peak demand during “off-peak” hours for a distribution XFR that serves many EVs, especially when EV owners employ smart charging to exploit low tariffs. Overloading the XFR results in expedited aging and subsequent risk of expensive XFR maintenance and power outage. So, we propose a utility-driven charging technique that aims to minimize the maintenance cost by flattening the load curve for the XFR while ensuring customer satisfaction. The proposed DSM considers the other household devices as base loads and schedules EV charging based on the available power after providing power for the base loads. As a result, the utility company faces fewer maintenance costs thanks to peak load reduction. It incentivizes the consumers using the profit it makes from reduced maintenance costs to participate in the scheduling program.

### 5.2.1 Proposed Technique

In our proposed technique, as shown in Fig. 5.1 (left blue box), the utility employs one charging agent for each XFR to schedule and control the charging of the EVs. Note that the time units for DSM (hourly) and USM (monthly) are different. Whenever an EV is plugged in for charging (EV arrival), the agent collects the battery charge level  $E_n$ , the target charge level  $E_{tgt}$ , and calculates the charging time window,

$$T_w = \left\lceil \frac{E_{tgt} - E_n}{\tau \times P_{max}} \right\rceil + T_b,$$

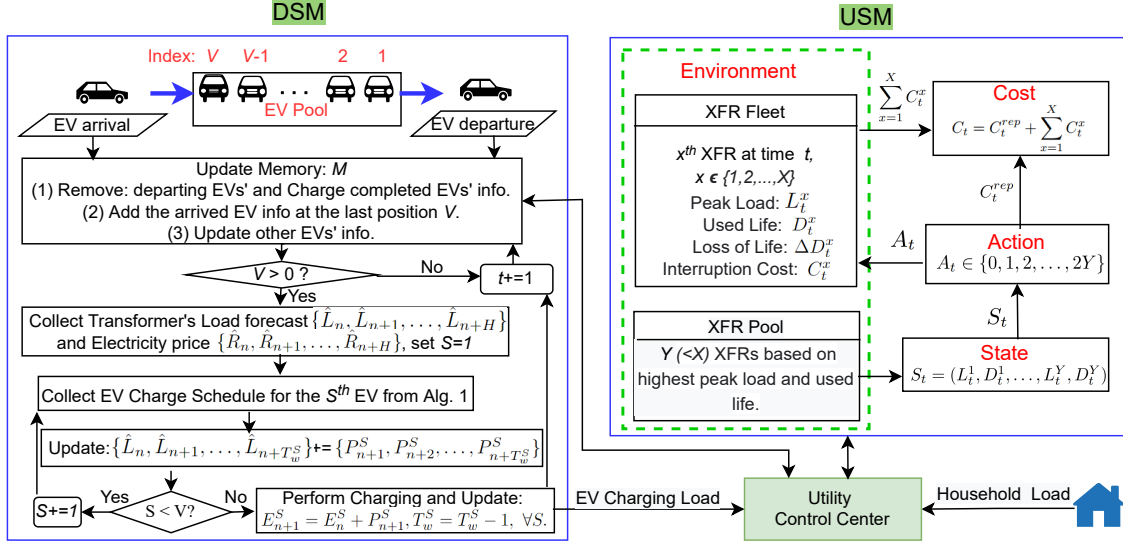


Figure 5.1: Proposed DSM flowchart (left) and DRL model for USM (right).

where  $P_{max}$  is the maximum charging capacity of the particular EV and  $\tau$  is the duration of timestep in hours. The ceiling  $\lceil \cdot \rceil$  of the fraction indicates the minimum number of time units for completing the target charging, and  $T_b$  provides buffer time to the agent to do the scheduling.

Then the agent updates its memory  $M$  by removing the departed EVs and charged EVs ( $E_n = E_{tgt}$ ) information and puts the arrived EV at the last position,  $V$ . Here,  $V$  indicates the number of EVs awaiting charge in the pool, and also the length of Memory  $M = \{(E_n^1, E_{tgt}^1, P_{max}^1, T_w^1), \dots, (E_n^V, E_{tgt}^V, P_{max}^V, T_w^V)\}$ .

Next, the agent proceeds to charge scheduling if there is any EV in the pool ( $V > 0$ ). The agent gathers electricity price and household load forecast for the next  $H$  timesteps, a decision time horizon which is bigger than the charging time window of any EV (e.g.,  $H = 16$  hours). LSTM algorithm fits well for the household sequential load prediction [125]. We take the temperature forecast  $\theta_n$ , the holiday flag  $H_n \in \{0, 1\}$ , and the load data of the last  $m$  time steps  $\{L_{n-m}, \dots, L_{n-2}, L_{n-1}\}$ , to predict the XFR load forecast  $\hat{L}_n, \hat{L}_{n+1}, \dots, \hat{L}_{n+H}$ . Similarly, the agent uses LSTM to forecast electricity prices  $\{\hat{R}_n, \hat{R}_{n+1}, \dots, \hat{R}_{n+H}\}$  from the last  $m$  time steps data  $\{R_{n-m}, \dots, R_{n-2}, R_{n-1}\}$ .



The agent follows the first-come-first-serve approach and starts scheduling the first EV ( $S = 1$ ) in the pool. Algorithm 5.1 shows the EV charge scheduling technique for the  $S^{th}$  EV in the pool. If the charge time window is not over, i.e.,  $T_w^S > 0$ , the agent sorts electricity prices  $R$  in ascending order and store the indices in vector  $I$ . The agent schedules charging for the next  $T_w^S$  timesteps, starting from the cheapest electricity tariff hours to the costlier ones. The available power forecast  $\mathcal{L}$  is the difference between load capacity  $L_{cap}$  and load forecast of the XFR for the corresponding hour (Line 5). Notably, the utility decides on the load capacity  $L_{cap}$  of the XFR, typically between 100-120 % of the nameplate kVA rating (e.g., 25-30 kVA for a 25 kVA rated XFR). The agent reads the battery charge status  $E_t$  from  $M$  and calculates the required charging  $E_R$  (Line 6). So, the EV charge allocation for the cheapest hour is

$$P_{n+I(1)}^s = \min\{P_{max}^S, E_R, \mathcal{L}\}.$$

The agent updates the charge level  $E_L$  for the following schedule step (Line 8). This process continues for charge allocation for all the time steps from the second cheapest,  $P_{n+I(2)}^S$  till the costliest one  $P_{n+I(T_w^S)}^S$ . Finally, the algorithm outputs charge allocation for the next  $T_w^S$  time steps as  $\{P_{n+1}^s, P_{n+2}^s, \dots, P_{n+T_w^S}^s\}$  (line 10). However, if the charging window is over (i.e.,  $T_w \leq 0$ ), the agent implements charging  $\{P_{n+1}^s\}$  for the immediate time step, as explained next.

If  $T_w \leq 0$ , but the target charge level is not achieved (Line 12), the agent offers compensation charging at a fixed rate based on the battery charge status  $E_n^S$ . We define two more user input  $E_{safe}^S$  and  $E_{crit}^S$  that each consumer can initiate and update as required. As the EV is expected to leave anytime soon ( $T_w \leq 0$ ), Algorithm 5.1 outputs allocated compensation

charging for the next time step as:

$$P_{n+1}^S = \begin{cases} \min\{P_{max}^S, \mathcal{L}\}, & E_n^S \leq E_{crit}^S \\ \min\{0.5 \times P_{max}^S, \mathcal{L}\}, & E_{crit}^S < E_n^S \leq E_{safe}^S \\ 0, & E_n^S > E_{safe}^S \end{cases} \quad (5.1)$$

where,  $\mathcal{L} = L_{cap} - \hat{L}_{n+1}$ , is the estimated available power. After the completion of charge scheduling for each EV, the agent updates the load forecast by adding the scheduled EV charges. This charge scheduling continues till all the  $V$  EVs are scheduled through Algorithm 5.1. Upon completion of scheduling, the charging for the  $n^{th}$  time step is implemented. Although the actual load may differ from the prediction, with an appropriate method, the prediction error will be within the range that causes insignificant aging difference to the XFR. So, the agent implements the charging as per scheduled and update the memory as:

$$E_{n+1}^S = E_n^S + P_{n+1}^S, T_w^S = T_w^S - 1, \forall S.$$

The agent moves to the next time step with its memory update, and this recursive loop continues as shown in Fig. 5.1.

### 5.2.2 Consumer Incentive

The consumers receive free smart EV charging service and a monetary incentive for participating in the DSM. The monetary incentive,

$$I = \frac{\kappa}{100} \times \frac{T_b}{E_{tgt}} \times EV \text{ charging bill}, \quad (5.2)$$

depends on their preferences for  $T_b$  and  $E_{tgt}$ . Customers who provide longer buffer time  $T_b$  and smaller target charge level  $E_{tgt}$  get more incentive. On the contrary, customers who prioritize comfort by selecting  $T_b = 0$  ensure the fastest possible EV charging without any

---

**Algorithm 5.1** Utility-Driven EV Charge Scheduling

---

```
1: Input:  $E_L = E_n^S$ ,  $T_b^S$ ,  $E_{tgt}^S$ ,  $P_{max}^S$ ,  $R = \{R_{n+1}, \dots, R_{n+H}\}$ ,  $\hat{L} = \{\hat{L}_{n+1}, \dots, \hat{L}_{n+H}\}$ .
2: if  $T_w^S > 0$  then
3:   Sort electricity prices  $R$  in ascending order and store the indices in vector  $I$ .
4:   for  $\tau = 1, 2, \dots, T_w^S$  do
5:     Estimate available power,  $\mathcal{L} = L_{cap} - \hat{L}_{n+I(\tau)}$ .
6:     Remaining charge,  $E_R = E_{tgt}^S - E_L$ .
7:     Scheduled charge,  $P_{n+I(\tau)}^S = \min\{P_{max}^S, E_R, \mathcal{L}\}$ .
8:     Update  $E_L = E_L + P_{n+I(\tau)}^S$ .
9:   end for
10:  Output: EV charge schedule  $\{P_{n+1}^S, P_{n+2}^S, \dots, P_{n+T_w^S}^S\}$ .
11: else
12:  Output: EV charge compensation  $\{P_{n+1}^S\}$  from Eq. (5.1).
13: end if
```

---

incentive. EV owners set  $T_b$  and  $E_{tgt}$  during system setup and can update their choices from time to time. This setup offers the customer control over their preferences and gives our method an edge over the existing techniques. The utility selects the incentive coefficient  $\kappa$  based on the savings in maintenance cost.

### 5.3 DRL Based XFR Replacement for USM

We propose an RL framework shown in Fig. 5.1 (right blue box), where the electricity utility company is the RL agent that makes replacement and upgradation decisions for the distribution XFRs.

#### 5.3.1 Environment

The RL environment is the distribution feeder with  $X$  customer-end XFRs and their connected loads. The XFRs can be of different capacities (kVA rating), serving different household numbers. The environment provides the peak load,  $L_t^x$ , and the loss of life,  $\Delta D_t^x$ , for the  $x^{th}$  XFR during the  $t^{th}$  time step. We calculate the effective aging as per the IEEE

standard [37] as:

$$\Delta D_t^x = \int_t^{t+1} 9.8 \times 10^{-18} e^{\frac{15000}{T_H+273}} dt, \quad (5.3)$$

where  $T_H$  denotes the hotspot temperature of the XFR which depends on the ambient temperature and the electrical load. We approximate the effective ageing integral equation through fine granularity (per minute) estimation. Apart from the scheduled maintenance, the utility also bears unscheduled interruption costs, mainly due to XFR failure and fuse blowing events. XFR failure occurs due to insulation breakdown, which depends on the used life

$$D_t^x = D_0^x + \sum_{n=1}^t \Delta D_n^x = D_{t-1}^x + \Delta D_t^x$$

of the XFR, where  $D_0^x$  is the initial age of the XFR in days. Weibull distribution is popular for forecasting the insulation failure of a XFR [87]. Our preliminary work [124] shows the XFR failure probability during the  $t^{th}$  timestep is

$$P_t^x = 1 - \exp \left[ \left( \frac{D_t^x}{\alpha} \right)^\beta - \left( \frac{D_t^x + 1}{\alpha} \right)^\beta \right] \quad (5.4)$$

where  $\alpha$  and  $\beta$  are the scale and the shape parameters of the Weibull distribution, respectively. XFR failure brings interruption cost  $C_t^x$  to cover XFR replacement, required labor, and unplanned outages.

Fuse-blowing events are deterministic and protect the XFR by disconnecting the circuit whenever the load exceeds the rating of the fuse; typically, 180% of the XFR's rated load [97]. Since fuse is meant to protect the XFR, its replacement brings minor labor and outage costs. Hence, the interruption cost for fuse replacement is smaller than that of XFR failure. While the monetary value of  $C_t^x$  varies with time and place, a utility company can have a proper estimate of  $C_t^x$  for XFR failure or blown fuse.

### 5.3.2 State

The agent makes replacement decisions based on the used life (hereinafter, age) and peak load of a XFR. However, XFRs with low age and peak load are not suitable candidates for change; hence eliminating them from the RL decision process creates a smaller state space and faster algorithm convergence without performance compromise. So, the agent takes the most loaded  $Y_l$  and most aged  $Y_a$  XFRs to make a pool of size  $Y = Y_l + Y_a$ . The load and age of these XFRs create the state for time step  $t$ ,

$$S_t = (L_t^1, D_t^1, L_t^2, D_t^2, \dots, L_t^Y, D_t^Y).$$

The percentile load  $L_t^y$ , which is the ratio of peak load and capacity of the  $y^{th}$  XFR, does not require normalization. We divide the age by the IEEE recommended lifetime of a XFR (7500 days) for normalization.

### 5.3.3 Action

The utility needs to replace the overloaded and older XFR to avoid failure and outage-related costs. However, under budgetary constraints, the RL agent chooses one XFR for replacement from the pool at each time step. Replacing the XFR with a bigger one is more cost-effective if the existing peak load is significantly higher than the capacity. So, our RL agent's action includes replacing the XFR with the same-sized or double-sized (kVA) one. If there is no overloaded or old XFR in the fleet, the optimal action might skip replacement ( $A_t=0$ ). As a result, our action space contains  $2Y + 1$  options

$$A_t \in \{0, 1, 2, \dots, 2y - 1, 2y, \dots, 2Y - 1, 2Y\}$$

where,  $y \in \{1, 2, \dots, Y\}$  is the index of the XFR in the pool;  $2y - 1$  and  $2y$  represent replacing the  $y^{th}$  XFR with the same and double capacity one, respectively.

### 5.3.4 Cost

The monetary cost for maintenance constitutes the RL framework's cost function  $C_t$  (negative reward). The maintenance cost includes the replacement (or upgrade) cost  $C_t^{rep}$  (or  $C_t^{upg}$ ) and emergency interruption cost  $\sum_{x=1}^X C_t^x$  as in

$$C_t = C_t^{rep} + \sum_{x=1}^X C_t^x.$$

Notably, as failure brings emergency labor and unplanned longer outages, XFR failure is way costlier than a scheduled replacement for a same-sized XFR. Furthermore, an under-sized XFR brings huge maintenance costs by multiple interruptions through fuse blows and eventual failure, which can be negated by upgrading its size.

### 5.3.5 Next State

The selected action installs a new XFR (zero aged) in place of the previous one. The replaced XFR's index gets attributed to the new one. The age of a XFR for the next time step is

$$D_{t+1}^x = \begin{cases} 0, & A_t \in \{2x - 1, 2x\} \\ D_t^x + \Delta D_t^x, & otherwise \end{cases}$$

where the environment provides the effective aging in time step  $t$ ,  $\Delta D_t^x$ , from Eq. (5.3). Furthermore, the environment provides the maximum load of the  $x^{th}$  XFR during time step  $t$ , which is used to estimate the peak load of the XFR as

$$L_{t+1}^x = \frac{Maximum\ Load}{Rated\ Capacity}.$$

The rated capacity of the XFR is updated whenever it is upgraded by a double-sized one.

## 5.4 Experiments

### 5.4.1 Experimental Setup

From the 2009 RECS dataset for the Midwest region of the United States [141], Muratori generates 200 household load profiles, along with 348 predicted EV charging loads connected to those households in [88]. The households vary in size, occupancy, electricity consumption. Lisha et al. [135] present an EV diffusion model for feeder level distribution system that considers different socioeconomic factors of the neighborhood. They provide an EV inclusion model for a 30-year timeline based on the car age, neighborhood, economy, and other critical features for an urban distribution feeder in North Carolina [135]. We combine the load and EV charging profile of [88] with the EV diffusion model of [135] to obtain the load profile in our case study. Distribution feeder data are summarized in Table 5.1.

The distribution feeder maintenance includes scheduled and unscheduled replacement (due to failure) of the XFR and protective fuses in our setup. Based on our study of the equipment cost and labor, we set the total cost for different types of maintenance as shown in Table 5.2. Fault-based maintenance brings emergency outages and customer inconvenience cost. We take the inconvenience cost for XFR failure and for fuse blows as \$1.3 per kWh and \$2 per kWh, respectively, according to the service value assessment in [134]. Since the customers are notified beforehand, the inconvenience cost for scheduled replacement is zero.

For the neural networks, we take the discount factor,  $\gamma = 0.95$ , and learning rate  $3 \times 10^{-4}$ . The actor and the critic networks have three hidden layers, each with 30, 120, and 48 neurons. The LSTM network for the load prediction has two LSTM layers. We use Adam optimizer for both the LSTM and DRL networks.

### 5.4.2 EV Charging (DSM)

We select the buffer time,  $T_b = 3$  and target charge level,  $E_{tgt} = 0.9$  for all the customers. [88] shows the impact of uncoordinated charging on the distribution grid, in which the EVs

Table 5.1: Feeder data (Numbers in parentheses indicate the number of XFRs serving that many homes).

Number of private homes	1116
Number of XFRs	232
Number of homes per XFR	7 (30), 6 (30), 5 (80), 4 (50), 3 (42)
XFR rating	25 kVA, 480/208 V, 1 phase, ONAN
XFR characteristics	Taken from [124]

Table 5.2: Utility company’s equipment and labor cost for different maintenance types.

Maintenance Work	Cost (\$)		Outage Time
	25 kVA	50 kVA	
XFR scheduled replacement	1500	3000	1 hr
XFR failure replacement	3000	4500	24 hr
Fuse blow restoration	500		6 hr

get charged to full capacity without any schedule. Our proposed DSM reduces the peak load significantly, as shown in Fig. 5.2. Out of the 232 XFRs, XFR-4 receives the most EVs (14) during the 30-year timeline. For the 1st year, the proposed and uncoordinated load profiles are the same, as there is no EV inclusion in the beginning. With growing time and EV inclusion, the proposed charging method reduces the peak load increasingly. For the first week of 30<sup>th</sup> year, uncoordinated charging results in a peak load above 33 kV compared to the peak load of around 21.1 kV with the proposed utility-driven charging. Similarly, for all the XFRs for the 30-year timeline, uncoordinated charging yields as much as 49.73 kVA load, compared to the 32.07 kVA max load of the proposed charging method. This indicates that uncoordinated charging incurs a significantly higher cost for XFR replacement and upgradation compared to the proposed charging method.

Apart from uncoordinated charging, we examine the following smart EV charging techniques from the literature.

(1) Rule-based in [109]: Sarker et al. [109] present a centralized strategy for EV charging by co-optimization of distribution XFR aging and energy arbitrage. The objective is to minimize the total cost of electricity consumption and the damage cost to the XFR. They estimate the damage cost by multiplying the price of the XFR by its loss of life, using Eq.



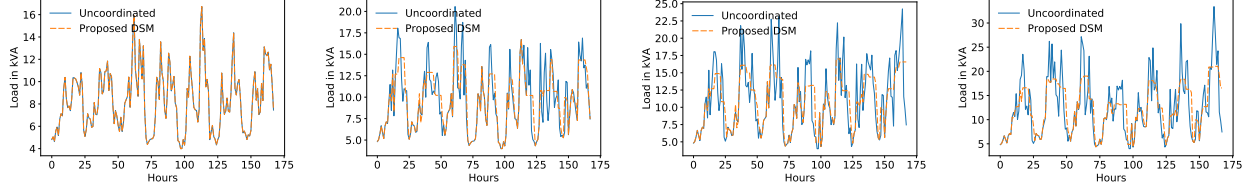


Figure 5.2: Comparison between the proposed utility-driven DSM and uncoordinated EV charging in terms of hourly load for XFR-4 for the first week of Year-1, Year-10, Year-20, and Year-30 (from left to right).

(5.3). The utility pays incentives to compensate the customers, as charging often happens during higher tariffs to minimize damage costs to the XFR. This constraint optimization strategy for EV charging satisfies constraints related to the battery’s state of charge to represent user preference, which is too basic to capture a user’s driving traits and routine.

(2) MARL in [76]: Li et al. [76] proposed a Multi-Agent Reinforcement Learning (MARL) based EV charging strategy. Each EV under a distribution XFR is an individual agent that minimizes the total cost due to electricity bills and XFR damage cost under a central agent, i.e., the distribution XFR. The MARL state is defined by real-time electricity price, XFR hotspot temperature, load forecast, EV state of charge, and other parameters. The reward function includes the customer’s EV range anxiety cost, representing the inconvenience cost due to delaying charging to utilize lower tariff hours. The authors model three different types of range anxiety (RA) cost as a function of the EV’s state of charge at departure time, of which we select Type-1 RA for the comparative analysis.

(3) CIBECS in [125]: The consumer input based EV charge scheduling (CIBECS) [125] for a residential home can be achieved by following Algorithm 5.1 with one modification of making the scheduled charging free of estimated available power  $\mathcal{L}$  from Line 7 as:

$$P_{n+1}^S(\tau) = \min\{P_{max}^S, E_R\}.$$

Table 5.3 shows the cost comparison among the different charging methods for XFR-4 for two representative years, the 15<sup>th</sup> and 30<sup>th</sup> years. The customer cost represents the

Table 5.3: Yearly cost (\$) for XFR-4 to customers and the utility for different charging techniques.

Charging Technique	15 <sup>th</sup> Year Cost			30 <sup>th</sup> Year Cost		
	Cust.	Utility	Total	Cust.	Utility	Total
Uncoordinated [88]	14562	123	14685	18023	3285	21308
Rule based [109]	12765	130	12895	15980	1869	17849
MARL [76]	12640	147	12787	16674	487	17161
CIBECS [125]	12290	216	12506	15339	3630	18969
Proposed DSM	12297	104	12401	15221	772	15993

electricity cost, and the utility cost represents the XFR loss of life, fuse-blowing costs, and customer incentive (if any). The proposed charging method estimates the maintenance savings with respect to the utility cost of the uncoordinated charging method. We select the incentive coefficient  $\kappa = 1$  for the 30<sup>th</sup> year, which correspond to 3.33% (\$524) discount on the customers' EV charging bill. The Uncoordinated charging [88] and CIBECS [125] prioritize EV charging, hence resulting in high utility costs (due to frequent fuse blows). On the contrary, MARL in [76] and Rule-based method in [109] maintain strict peak load constraints to minimize the utility cost. However, they are susceptible to undercharged EVs, which is not a desirable solution for customers. The Rule-based method in [109] provides the customer with an incentive from its maintenance savings, which contributes to its utility cost. Our proposed DSM method capitalizes low-price hours, accommodates customer preference, and maintains load flattening simultaneously. As a result, the customer cost for the proposed DSM technique is the least among all the methods, and the utility cost is only marginally higher than the MARL in [76]. Lastly, as there are no fuse blow events and negligible utility cost saving for the 15<sup>th</sup> year, the proposed DSM offers zero incentive for that year.

#### 5.4.3 DRL Based Maintenance (USM)

Based on the comparative analysis for EV charging in the previous section, we focus on the proposed utility-driven charging technique to implement our DRL-based XFR replacement policy. Fig. 5.3 shows that our method learns the optimal policy within 3000 episodes.

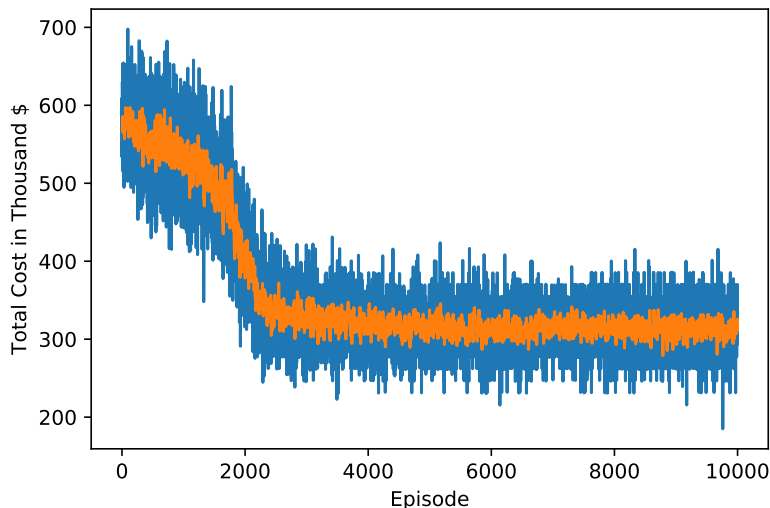


Figure 5.3: Convergence of the DRL based maintenance policy (with proposed DSM).

Table 5.4: Computational details for the experiments.

Hardware	Software	Task	Computation time
Intel <sup>®</sup> Core i7 3.60GHz 16 GB RAM	Python 3.7 Pytorch 1.8.1	EV Charge Scheduling	2.4 sec
		DRL Convergence	180 min
		DRL Decision	0.1 sec

Table 5.4 shows the computation time for the proposed method. It takes 2.4 seconds for the proposed EV charge scheduling by using an Intel<sup>®</sup> Core i7, 3.60 GHz, 16 GB RAM computer. The DRL algorithm needs 180 minutes to perform the 3000 episodes for convergence. Notably, the computational time for each decision is 0.01 second, negligible compared to maintenance policy-making steps (i.e., 1 month).

We compare our method with an idle policy, two rule-based methods from [145] and [43], and the popular statistical Markov Chain Monte Carlo (MCMC) [19] method.

(1) Idle policy: In this policy, the utility waits till the failure of a XFR for replacement. The utility would replace the XFR with double capacity if it endured more than five fuse blowing events during the previous twelve months; otherwise, replace it with the same capacity one.

(2) Ranking-based method [145]: Vasquez et al. [145] proposes a ranking-based approach for XFR replacement. The ranking score is calculated based on the XFR’s probability of

Table 5.5: Cumulative EV charging and maintenance cost for all 232 XFRs over a 30-year timeline.

DSM USM	Uncoordinated Charging					Proposed DSM				
	Idle Policy	Ranking based [145]	Risk Score based [43]	MCMC [19]	Prop. DRL	Idle Policy	Ranking based [145]	Risk Score based [43]	MCMC [19]	Prop. DRL
Fuse Blow	237	125	22	45	16	0	0	0	0	0
XFR Failure	130	102	127	100	95	125	99	125	98	89
XFR Replace	0	28	10	26	23	0	26	0	25	22
XFR Upgrade	0	2	3	3	3	0	0	0	0	0
Outage (hr)	4534	3555	3399	2670	2376	3003	2383	3005	2352	2136
Outage (kWh)	36.49	28.10	22.34	20.67	17.04	18.03	14.36	18.03	14.15	13.29
Cost \$	571,738	494,273	472,728	427,851	386,356	398,772	356,212	399,159	349,797	317,308

failure (from Eq. (5.4)) and its failure replacement cost  $\xi_t^x$  (from Table 5.2). The ranking score of the  $x^{th}$  XFR for the  $t^{th}$  time step is given by

$$R_t^x =_t^x \times \xi_t^x.$$

The highest-ranked XFR is replaced if the ranking score exceeds the threshold set through trial and error. The new XFR will be double-sized if the peak load is more than 1.5 times, otherwise same sized as the replaced one. Notably, this method portrays aggressive XFR replacement, hence functions opposite the above-mentioned idle policy.

(3) Risk score based method [43]: The following equation is used in [43] to estimate the risk score for a XFR,

$$\mathfrak{R} = Cond \times \frac{60 + age \text{ (in yr)}}{60} \times \frac{Peak Load}{Capacity} \times EF.$$

Since all the XFRs serve under similar environmental factor ( $EF$ ) and have similar characteristic conditions ( $Cond$ ), we remove these two parameters when estimating the risk factor  $\mathfrak{R}$  for each XFR. At the end of the month, the XFR with the highest risk factor  $\mathfrak{R}$  is replaced. If the risk factor value is lower than a threshold, no replacement occurs. We found 1.85 as the optimal threshold in our experiments. If the XFR's peak load is more than 150% of its capacity, it is replaced with a double-sized one; otherwise, with a same-sized XFR.

(4) MCMC [19]: Markov Chain Monte Carlo (MCMC) simulation is a popular tabular RL technique for problems with discrete and tractable state and action spaces. We discretize

the state space (as opposed to the continuous-valued DRL states) as the MCMC utilizes a tabular method to learn the value function for the state. The granularity of the discretization is a trade-off between the optimization results and computation time. We discretized each input variable in  $m = 10$  equally spaced states for a manageable computation burden, which requires the convergence for  $m^r = 10^{12}$  states, where  $r = 12$  is the number of input variables (i.e., age and load of the 3 oldest and the 3 most loaded XFRs in the network).

#### 5.4.4 Comparative Analysis

We implement the above mentioned maintenance policies for both uncoordinated and the proposed utility-driven charging for an ablation study. Table 5.5 shows the cumulative maintenance cost to the utility for different EV charging and maintenance policy combinations for the distribution feeder over a 30-year timeline. Table 5.6 further elaborates the results in terms of the following metrics.

##### 5.4.4.1 Fuse Blow

As the load (EV charging) grows, fuse blow events occur more frequently during the late part of the simulation timeline. Without any planned capacity upgrade (as in Idle policy), it accumulates 237 such events in the 30-year timeline. The Ranking method [145] ignores the peak load in its decision criteria and performs worse than the other methods. The Risk score method [43] puts significance on peak load and reduces fuse blow events through XFR upgrades. The proposed DRL method learns the correlation and minimizes the fuse blows; however, the MCMC method lags due to discretized state space. The proposed DSM approach flattens the load to such an extent that none of the policies experience any fuse-blowing events.

#### 5.4.4.2 XFR Failure

The XFR failure events can not be nullified as it follows the Weibull distribution in (5.4). However, the proposed DRL method minimizes XFR failure by approximately 30% followed by the MCMC method. The Ranking method performs well as it prioritizes XFR age in its maintenance decision. On the contrary, the Risk score method underestimate XFR age in risk calculation to reduce XFR failure.

#### 5.4.4.3 Planned Maintenance

The DRL method implements 23 replacements and 3 upgrades in the Uncoordinated charging case. In the proposed DSM case, the proposed DRL requires 22 replacements and no upgrades. Its optimal selections yield minimum XFR failure, outage, and cost compared to the benchmark methods.

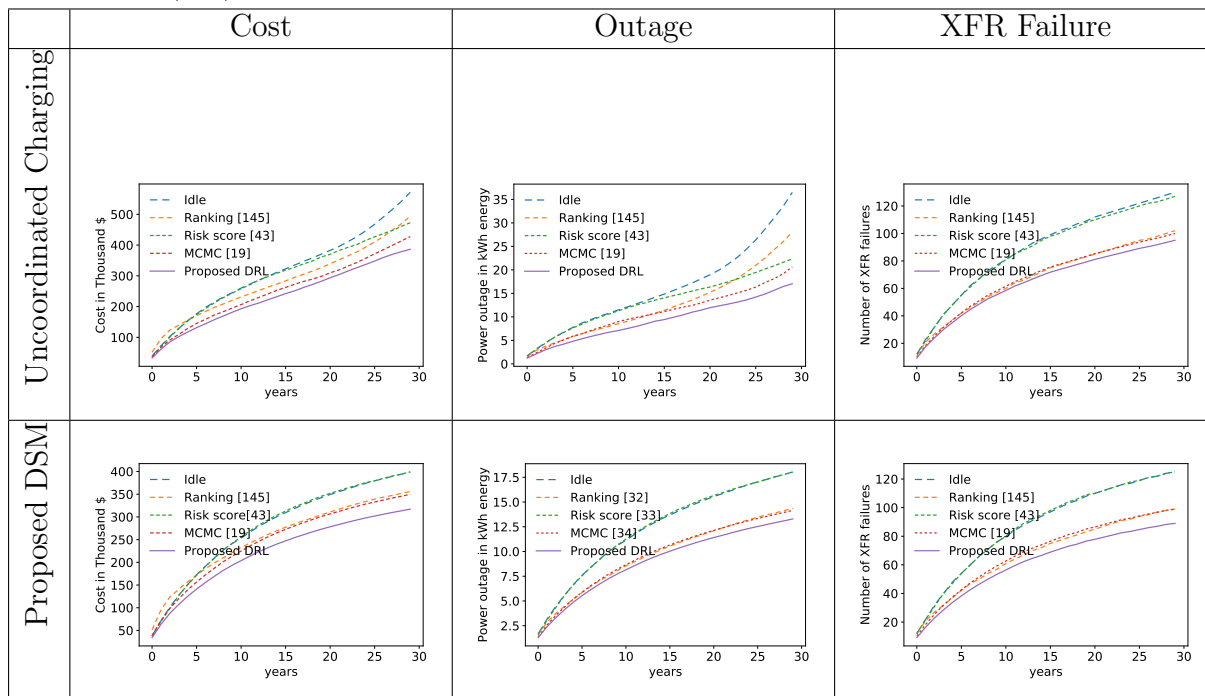
#### 5.4.4.4 Monetary Cost

Cumulative cost includes the planned and unplanned maintenance costs, which is the actual objective of the utility company to minimize. Our proposed DRL, accompanied by the proposed DSM, is the best performing combination.

#### 5.4.5 Key Insights

- In MCMC-based RL, discretized states for feasible training time result in significant performance degradation with respect to the DRL method.
- The rule-based methods are too simple to set the appropriate balance between XFR age and load in decision-making. Hence, they either suffer many fuse blows (Ranking [145]) or XFR failures (Risk score [43]).
- The DRL policy learns the optimal weight of age and peak load of the candidate XFRs for selecting the most appropriate XFR for maintenance, which is evident by the

Table 5.6: Comparison among XFR maintenance policies for uncoordinated charging (top) and proposed DSM (bottom) in terms of cumulative cost (left), power outage (middle), and XFR failure (left).



reduction in XFR failure, fuse blows, and subsequent outages. As there are many aged XFRs in the network initially, our policy aggressively replaces the aged and overloaded XFRs with newer ones. These proactive actions reduce the number of XFR failures and fuse blows.

- The proposed EV charging technique substantially boosts the DRL-based policy to minimize the long-term maintenance cost.

## 5.5 Conclusion

This work offers insight and solutions for maintaining the distribution system to accommodate EV charging load. It demonstrates a complete EV adoption strategy for the utility company considering long-term planning for both demand side management (DSM) and utility side management (USM). For DSM, the proposed utility-driven EV charge scheduling

based on customer preferences offers a reasonable balance between peak load reduction and customer satisfaction. Consequently, the utility company faces less maintenance costs due to peak load reduction. The utility compensates the customers using its profit from reduced maintenance costs to keep them interested in participating in the scheduled EV charging program. For USM, our DRL-based XFR maintenance policy chooses the best XFR for replacement or upgrade. Experiments show that the combination of the proposed DSM and USM methods outperforms the existing optimization techniques by a wide margin in terms of long-term maintenance cost and power outage.



## Chapter 6: Conclusions

With the revolutionary success of data-driven Machine Learning (ML) techniques, social, economic, and environmental planners are embracing these state-of-the-art methods. Deep Reinforcement Learning (DRL) is the most popular ML technique for optimization tasks, particularly for systems that require sequential decisions. With the enormous flow of data, human expectations from policymakers are going up simultaneously. However, DRL-based optimization requires both domain knowledge and algorithmic skills. The domain expert often opts to learn this optimization technique. However, this DRL-based research field is moving too fast for them to catch up. More precisely, DRL research is thriving to fit different applications and has generated a huge literature and tree of methods. Inherently, people with advanced AI skills and knowledge can provide the policy maker critical edge in selecting, developing, and applying an appropriate model for their task. This can save money, ensure consumer satisfaction, and enhance the reputation of the planning authority. This dissertation investigated DRL-based resource allocation to provide optimal solutions for several complex systems.

In the NSF-funded project "Infrastructure development Against Sea-Level Rise (SLR)," we investigated appropriate mathematical models and historical data for natural disasters for applying Multiagent DRL. Sea-level rise (SLR) problem, which is a major outcome of climate change, has been well documented and studied. Although it is globally observed due to climate change, local projections are needed to plan SLR adaptation strategies accurately. Since SLR is a community-wide multi-stakeholder problem at the local level, adaptation strategies can be more successful if the main stakeholders, e.g., government, residents, businesses, collaborate in shaping them. Simulating the local socioeconomic system around SLR,

including the interactions between essential stakeholders and nature, can be an effective way of evaluating different adaptation strategies and planning the best strategy for the local community. This project presents how such an SLR socioeconomic system can be modeled as a Markov decision process (MDP) and simulated using multi-agent reinforcement learning (RL). The proposed multi-agent RL framework serves two purposes. It provides a general scenario planning tool to investigate the cost-benefit analysis of natural events (e.g., flooding, hurricane) and agents' investments (e.g., infrastructure improvement). It also shows how much the total cost due to SLR can be reduced over time by optimizing the adaptation strategies. We demonstrate the proposed scenario planning tool using available economic data and sea-level projections for Pinellas County, Florida, in a case study.

We extend the DRL framework to the Home Energy Recommendation System (HERS) project, the first smart home energy management approach based on residents' feedback and activity. Smart home appliances can take command and act intelligently, making them suitable for implementing optimization techniques. Artificial intelligence (AI) based control of these smart devices enables demand-side management (DSM) of electricity consumption. By integrating human feedback and activity in the decision process, this project proposes a deep Reinforcement Learning (RL) method for managing smart devices to optimize electricity cost and comfort residents. Our contributions are twofold. Firstly, we incorporate human feedback in the objective function of our DSM technique. Secondly, we include human activity data in the RL state definition to enhance energy optimization performance. We perform comprehensive experimental analyses to compare the proposed deep RL approach with existing approaches that lack the aforementioned critical decision-making features. The proposed model is robust to varying resident activities and preferences and applicable to a broad spectrum of homes with different resident profiles.

We implement multi-objective RL (MORL) for our hospital capacity expansion planning, which is critical for a healthcare authority, especially in regions with a growing diverse population. Policymaking to this end often requires satisfying two conflicting objectives,

minimizing capacity expansion cost and minimizing the number of denial of service (DoS) for patients seeking hospital admission. The uncertainty in hospital demand, especially considering a pandemic event, makes expansion planning even more challenging. This project presents a multi-objective reinforcement learning (MORL) based solution for healthcare expansion planning to optimize expansion cost and DoS simultaneously for pandemic and non-pandemic scenarios. Importantly, our model provides a simple and intuitive way to set the balance between these two objectives by only determining their priority percentages, making it suitable across policymakers with different capabilities, preferences, and needs. Specifically, we propose a multi-objective adaptation of the popular Advantage Actor-Critic (A2C) algorithm to avoid forced conversion of DoS discomfort cost to a monetary cost. Our case study for the state of Florida illustrates the success of our MORL based approach compared to the existing benchmark policies, including a state-of-the-art deep RL policy that converts DoS to economic cost to optimize a single objective.

The fourth project provides charging management for Electric vehicles (EV) for a distribution grid. Electricity authorities need capacity assessment and expansion plans for efficiently charging the growing EV fleet. Specifically, the distribution grid needs significant capacity expansion as it faces the most impact to accommodate the high variant residential EV charging load. Utility companies employ different scheduling policies for the maintenance of their distribution transformers (XFR). However, they lack scenario-based plans to cope with the varying EV penetration across locations and time. The contributions of this project are twofold. First, we propose a customer feedback-based EV charging scheduling to simultaneously minimize the peak load for the distribution XFR and satisfy the customer needs. Second, we present a deep reinforcement learning (DRL) method for XFR maintenance, which focuses on the XFR's effective age and loading to periodically choose the best candidate XFR for replacement. Our case study for a distribution feeder shows the adaptability and success of our EV load scheduling method in reducing the peak demand to extend the XFR life. Furthermore, our DRL-based XFR replacement policy outperforms

the existing rule-based policies. Together, the two approaches provide a complete capacity planning tool for efficient XFR maintenance to cope with the increasing EV charging load.

## References

- [1] HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2010-2017.
- [2] Marin County Community Development Agency, Game of Floods. <https://www.marincounty.org/depts/cd/divisions/planning/csmart-sea-level-rise/game-of-floods>, 2016.
- [3] Sea-Level Change Curve Calculator (Version 2021.12). [https://cwbi-app.sec.usace.army.mil/rccslc/slcc\\_calc.html](https://cwbi-app.sec.usace.army.mil/rccslc/slcc_calc.html), 2017.
- [4] The Cost Of Doing Nothing, Economic Impacts of Sea Level Rise in the Tampa Bay Area, Tampa Bay Regional Planning Council. [http://www.tbrpc.org/wp-content/uploads/2018/11/2017-The\\_Cost\\_of\\_Doing\\_Nothing\\_Final.pdf](http://www.tbrpc.org/wp-content/uploads/2018/11/2017-The_Cost_of_Doing_Nothing_Final.pdf), 2017.
- [5] PINELLAS COUNTY COASTAL MANAGEMENT PROGRAM SUMMARY PLANNING DOCUMENT. [http://www.pinellascounty.org/environment/coastalMngmt/pdfs/Pinellas\\_County\\_CMP\\_Sum\\_Document.pdf](http://www.pinellascounty.org/environment/coastalMngmt/pdfs/Pinellas_County_CMP_Sum_Document.pdf), 2018.
- [6] Cost of Flood Insurance in Florida and How Coverage Works. <https://www.valuepenguin.com/flood-insurance/florida>, 2020.
- [7] NOAA National Centers for Environmental Information (NCEI), U.S. Billion-Dollar Weather and Climate Disasters. <https://www.ncdc.noaa.gov/billions/time-series>, 2020.
- [8] QuickFacts of Pinellas County, Florida. <https://www.census.gov/quickfacts/fact/table/pinellascountyflorida/RHI225219>, 2020.

- [9] QuickFacts of Pinellas County, Florida. <https://www.census.gov/quickfacts/fact/table/stpetersburgcityflorida/POP060210>, 2020.
- [10] St. petersburg tourism. [https://www.stpete.org/economic\\_development/stpete\\_advantage/tourism.php#:~:text=Petersburg%20offers%2077%20hotel%2Fmotel,billion%20to%20the%20local%20economy.](https://www.stpete.org/economic_development/stpete_advantage/tourism.php#:~:text=Petersburg%20offers%2077%20hotel%2Fmotel,billion%20to%20the%20local%20economy.), 2021.
- [11] Real personal income and regional price parities, Accessed: 2021-09-29.
- [12] Coronavirous resource center: Florida. *Johns Hopkins University & Medicine*, Accessed: 2022-01-20.
- [13] FEMA Opens \$660 Million Grant Application Process. <https://www.fema.gov/press-release/20200929/fema-opens-660-million-grant-application-process>, September 29, 2020.
- [14] Jorge A Acuna, José L Zayas-Castro, and Hadi Charkhgard. Ambulance allocation optimization model for the overcrowding problem in us emergency departments: A case study in florida. *Socio-Economic Planning Sciences*, 71:100747, 2020.
- [15] International Energy Agency. *Global EV Outlook*. OECD-ilibrary, 2019.
- [16] Hafzullah Aksoy. Use of gamma distribution in hydrological analysis. *Turkish Journal of Engineering and Environmental Sciences*, 24(6):419–428, 2000.
- [17] Mehmet Aktukmak, Yasin Yilmaz, and Ismail Uysal. A probabilistic framework to incorporate mixed-data type features: Matrix factorization with multimodal side information. *Neurocomputing*, 367:164–175, 2019.
- [18] Hande Alemdar, Halil Ertan, Ozlem Durmaz Incel, and Cem Ersoy. Aras human activity datasets in multiple homes with multiple residents. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, pages 232–235. IEEE, 2013.

- [19] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50(1):5–43, 2003.
- [20] American Hospital Association et al. Trendwatch chartbook 2018: trends affecting hospitals and health systems, 2018, 2019.
- [21] Stef Baas, Sander Dijkstra, Aleida Braaksma, Plom van Rooij, Fieke J Snijders, Lars Tiemessen, and Richard J Boucherie. Real-time forecasting of covid-19 bed occupancy in wards and intensive care units. *Health care management science*, pages 1–18, 2021.
- [22] Shahab Bahrami, Vincent WS Wong, and Jianwei Huang. An online learning algorithm for demand response in smart grid. *IEEE Transactions on Smart Grid*, 9(5):4712–4725, 2017.
- [23] Beau Grant Barnes and Nancy L. Harp. The U.S. Medicare Disproportionate Share Hospital program and capacity planning. *Journal of Accounting and Public Policy*, 37(4):335–351, 2018.
- [24] Timothy Beatley. *Planning for coastal resilience: best practices for calamitous times*. Island Press, 2012.
- [25] Richard Bellman. A Markovian decision process. *Journal of mathematics and mechanics*, pages 679–684, 1957.
- [26] Philip R Berke and Mark R Stevens. Land use planning for climate adaptation: Theory and practice. *Journal of Planning Education and Research*, 36(3):283–289, 2016.
- [27] Petra Berkholz, Rainer Stamminger, Gabi Wnuk, Jeremy Owens, and Simone Bernarde. Manual dishwashing habits: an empirical analysis of uk consumers. *International journal of consumer studies*, 34(2):235–242, 2010.

- [28] Heider Berlink and Anna HR Costa. Batch reinforcement learning for smart home energy management. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.
- [29] Bloomberg New Energy Finance. Electric vehicle Outlook. Technical report, BNEF, 2021.
- [30] Bosch. Bosch 500 Series- Stainless steelSHP65T55UC instruction manual. [https://media3.bosch-home.com/Documents/9001218494\\_A.pdf](https://media3.bosch-home.com/Documents/9001218494_A.pdf).
- [31] Maya K Buchanan, Robert E Kopp, Michael Oppenheimer, and Claudia Tebaldi. Allowances for evolving coastal flood risk under uncertain local sea-level rise. *Climatic Change*, 137(3-4):347–362, 2016.
- [32] Arlan Burdick. Strategy guideline: accurate heating and cooling load calculations. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2011.
- [33] Maya Burke, Libby Carnahan, Kelli Hammer-Levy, and Gary Mitchum. Recommended projections of sea level rise in the tampa bay region (update). 2019.
- [34] Ana C Cebrián, Michel Denuit, and Philippe Lambert. Generalized pareto fit to the society of actuaries large claims database. *North American Actuarial Journal*, 7(3):18–36, 2003.
- [35] Chao Chen, Diane J Cook, and Aaron S Crandall. The user side of sustainability: Modeling behavior and energy usage in the home. *Pervasive and Mobile Computing*, 9(1):161–175, 2013.



- [36] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 456–464, 2019.
- [37] Transformers Committee, Of The, and IEEE Power & Energy Society. IEEE Guide for Loading Mineral- Oil-Immersed Transformers and Step-Voltage Regulators. *IEEE Std C*, 57:1–106, 2011.
- [38] Diane J Cook, Maureen Schmitter-Edgecombe, Linus Jönsson, and Anne V Morant. Technology-enabled assessment of functional health. *IEEE reviews in biomedical engineering*, 12:319–332, 2018.
- [39] Alison Evans Cuellar and Paul J Gertler. How the expansion of hospital systems has affected consumers. *Health affairs*, 24(1):213–219, 2005.
- [40] Anthony Daspit and K Das. The generalized pareto distribution and threshold analysis of normalized hurricane damage in the united states gulf coast. In *Joint Statistical Meetings (JSM) Proceedings, Statistical Computing Section, Alexandria, VA: American Statistical Association*, pages 2395–2403, 2012.
- [41] Derek DeLia. Annual bed statistics give a misleading picture of hospital surge capacity. *Annals of Emergency Medicine*, 48(4):384–388.e2, 2006.
- [42] Alexis Drogoul. Agent-based modeling for multidisciplinary and participatory approaches to climate change adaptation planning. In *Proceedings of RFCC-2015 Workshop, AIT*, 2015.
- [43] E Duarte, D Falla, J Gavin, M Lawrence, T McGrail, D Miller, P Prout, and B Rogan. A practical approach to condition and risk based power transformer asset replacement. In *2010 IEEE International Symposium on Electrical Insulation*, pages 1–4. IEEE, 2010.

- [44] M Irfan Fahmi, M Zarlisb, and Syahril Efendi Tulus. An optimization model for solving integrated hospital capacity planning problem. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(6):4816–4824, 2021.
- [45] Diana Farrell and Fiona E. Greig. Paying out-of-pocket: The healthcare spending of 2 million us families. *SSRN Electronic Journal*, 2017.
- [46] Jerzy Filar and Koos Vrieze. *Competitive Markov decision processes*. Springer Science & Business Media, 2012.
- [47] Fixr. Cost to Build a Hospital — Hospital Construction Cost.
- [48] Barry E Flanagan, Edward W Gregory, Elaine J Hallisey, Janet L Heitgerd, and Brian Lewis. A social vulnerability index for disaster management. *Journal of homeland security and emergency management*, 8(1), 2011.
- [49] Xinyu Fu, Mohammed Gomaa, Yujun Deng, and Zhong-Ren Peng. Adaptation planning for sea level rise: A study of US coastal cities. *Journal of Environmental Planning and Management*, 60(2):249–265, 2017.
- [50] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.
- [51] Mahshid Ghanbari, Mazdak Arabi, and Jayantha Obeysekera. Chronic and acute coastal flood risks to assets and communities in southeast florida. *Journal of Water Resources Planning and Management*, 146(7):04020049, 2020.
- [52] Mahshid Ghanbari, Mazdak Arabi, Jayantha Obeysekera, and William Sweet. A coherent statistical model for coastal flood frequency analysis under nonstationary sea level conditions. *Earth’s Future*, 7(2):162–177, 2019.

- [53] Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sonntag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1):16–18, jan 2019.
- [54] Linda V Green. How many hospital beds? *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, 39(4):400–412, 2002.
- [55] Eastern Research Group et al. *What Will Adaptation Cost?: An Economic Framework for Coastal Community Infrastructure*. National Oceanic and Atmospheric Administration Coastal Services Center, 2013.
- [56] Stephane Hallegatte, Colin Green, Robert J Nicholls, and Jan Corfee-Morlot. Future flood losses in major coastal cities. *Nature climate change*, 3(9):802–806, 2013.
- [57] Stéphane Hallegatte, Nicola Ranger, Olivier Mestre, Patrice Dumas, Jan Corfee-Morlot, Celine Herweijer, and Robert Muir Wood. Assessing climate change impacts, sea level rise and storm surge risk in port cities: a case study on copenhagen. *Climatic change*, 104(1):113–137, 2011.
- [58] Gary W Harrison, Andrea Shafer, and Mark Mackay. Modelling variability in hospital bed occupancy. *Health Care Management Science*, 8(4):325–334, 2005.
- [59] Mathew E Hauer, Jason M Evans, and Deepak R Mishra. Millions projected to be at risk from sea-level rise in the continental united states. *Nature Climate Change*, 6(7):691–695, 2016.
- [60] Ammar Haydari and Yasin Yilmaz. Deep reinforcement learning for intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2020.

- [61] Jakob Heins, Jan Schoenfelder, Steffen Heider, Axel R Heller, and Jens O Brunner. A scalable forecasting framework to predict covid-19 hospital bed occupancy. *INFORMS Journal on Applied Analytics*, 2022.
- [62] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- [63] A C Hine, D P Chambers, T D Clayton, M R Hafen, and G T Mitchum. *Sea Level Rise in Florida: Science, Impacts, and Options*. University Press of Florida, 2016.
- [64] Yili Hong, William Q Meeker, James D McCalley, and Others. Prediction of remaining life of power transformers based on left truncated and right censored lifetime data. *The Annals of Applied Statistics*, 3(2):857–879, 2009.
- [65] Ronald A Howard. *Dynamic programming and markov processes*. 1960.
- [66] Solomon Hsiang, Robert Kopp, Amir Jina, James Rising, Michael Delgado, Shashank Mohan, DJ Rasmussen, Robert Muir-Wood, Paul Wilson, Michael Oppenheimer, et al. Estimating economic damage from climate change in the united states. *Science*, 356(6345):1362–1369, 2017.
- [67] Rodney P Jones. Does the ageing population correctly predict the need for medical beds? part two: wider implications. *British Journal of Healthcare Management*, 27(10):1–9, 2021.
- [68] Noah S. Kalman, Bradley G. Hammill, Kevin A. Schulman, and Bimal R. Shah. Hospital overhead costs: The neglected driver of health care spending? *Journal of Health Care Finance*, 41(4), 2015.

- [69] Sarah M Kandil, Hany EZ Farag, Mostafa F Shaaban, and M Zaki El-Sharafy. A combined resource allocation framework for pevs charging stations, renewable energy resources and distributed energy storage systems. *Energy*, 143:961–972, 2018.
- [70] Shakil Bin Kashem, Bev Wilson, and Shannon Van Zandt. Planning for climate adaptation: Evaluating the changing patterns of social vulnerability and adaptation challenges in three coastal cities. *Journal of Planning Education and Research*, 36(3):304–318, 2016.
- [71] Murad Khan, Junho Seo, and Dongkyun Kim. Real-time scheduling of operational time for smart home appliances based on reinforcement learning. *IEEE Access*, 8:116520–116534, 2020.
- [72] Petter N Kolm and Gordon Ritter. Modern perspectives on reinforcement learning in finance. *Modern Perspectives on Reinforcement Learning in Finance (September 6, 2019)*. *The Journal of Machine Learning in Finance*, 1(1), 2020.
- [73] A. Kumar, Roger J. Jiao, and S. J. Shim. Predicting bed requirement for a hospital using regression models. In *IEEE International Conference on Industrial Engineering and Engineering Management*, pages 665–669, 2008.
- [74] Ekaterina Kutafina, Istvan Bechtold, Klaus Kabino, and Stephan M Jonas. Recursive neural networks in hospital bed occupancy forecasting. *BMC medical informatics and decision making*, 19(1):39, 2019.
- [75] Paolo Landa, Michele Sonnessa, Elena Tànfani, and Angela Testi. Multiobjective bed management considering emergency and elective patient flows. *International Transactions in Operational Research*, 25(1):91–110, 2018.

- [76] Sichen Li, Weihao Hu, Di Cao, Zhenyuan Zhang, Qi Huang, Zhe Chen, and Frede Blaabjerg. A multi-agent deep reinforcement learning-based approach for the optimization of transformer life using coordinated electric vehicles. *IEEE Transactions on Industrial Informatics*, 2022.
- [77] Weixian Li, Thillainathan Logenthiran, and Wai Lok Woo. Intelligent multi-agent system for smart home energy management. In *2015 IEEE Innovative Smart Grid Technologies-Asia (ISGT ASIA)*, pages 1–6. IEEE, 2015.
- [78] Daniele Liciotti, Michele Bernardini, Luca Romeo, and Emanuele Frontoni. A sequential deep learning application for recognising human activities in smart homes. *Neurocomputing*, 396:501–513, 2020.
- [79] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.
- [80] Fengji Luo, Gianluca Ranzi, Shu Wang, and Zhao Yang Dong. Hierarchical energy management system for home microgrids. *IEEE Transactions on Smart Grid*, 10(5):5536–5546, 2018.
- [81] Guoxuan Ma and Erik Demeulemeester. A multilevel integrative approach to hospital case mix and capacity planning. *Computers & Operations Research*, 40(9):2198–2207, 2013.
- [82] Francesca Marcello and Virginia Pilloni. Smart building energy and comfort management based on sensor activity recognition and prediction. *Sensors*, 1:s2, 2020.
- [83] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR, 2016.

- [84] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.
- [85] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.
- [86] Zeinab Moghaddam, Iftekhar Ahmad, Daryoush Habibi, and Quoc Viet Phung. Smart charging strategy for electric vehicle charging stations. *IEEE Transactions on transportation electrification*, 4(1):76–88, 2017.
- [87] G C Montanari, J C Fothergill, N Hampton, R Ross, and G Stone. IEEE Guide for the statistical analysis of electrical insulation breakdown data. *IEEE standard 930-2004*, 2005.
- [88] Matteo Muratori. Impact of uncoordinated plug-in electric vehicle charging on residential power demand. *Nature Energy*, 3(3):193–201, 2018.
- [89] Matteo Muratori, Michael J Moran, Emmanuele Serra, and Giorgio Rizzoni. Highly-resolved modeling of personal transportation energy consumption in the united states. *Energy*, 58:168–177, 2013.
- [90] Almuthanna Nassar and Yasin Yilmaz. Deep reinforcement learning for adaptive network slicing in 5g for intelligent vehicular systems and smart cities. *IEEE Internet of Things Journal*, 2021.
- [91] NYISO. NYISO REAL-TIME DASHBOARD. <https://www.nyiso.com/real-time-dashboard>.

- [92] Office of Energy Efficiency & Renewable Energy . Electric Vehicles: Charging at Home. Technical report, DOE, 2021.
- [93] Peter Palensky and Dietmar Dietrich. Demand side management: Demand response, intelligent energy systems, and smart loads. *IEEE transactions on industrial informatics*, 7(3):381–388, 2011.
- [94] June Young Park, Thomas Dougherty, Hagen Fritz, and Zoltan Nagy. Lightlearn: An adaptive and occupant centered controller for lighting based on reinforcement learning. *Building and Environment*, 147:397–414, 2019.
- [95] Pascal Peduzzi, Bruno Chatenoux, H Dao, Andréa De Bono, Christian Herold, James Kossin, Frédéric Mouton, and Ola Nordbeck. Global trends in tropical cyclone risk. *Nature climate change*, 2(4):289–294, 2012.
- [96] Virginia Pilloni, Alessandro Floris, Alessio Meloni, and Luigi Atzori. Smart home energy management including renewable sources: A qoe-driven approach. *IEEE Transactions on Smart Grid*, 9(3):2006–2018, 2016.
- [97] CharlesW Plummer, GL Goedde, Elmer L Pettit, Jeffrey S Godbee, and Michael G Hennessey. Reduction in distribution transformer failure rates and nuisance outages using improved lightning protection concepts. *IEEE transactions on power delivery*, 10(2):768–777, 1995.
- [98] Ruth Potts, Lisa Jacka, and Lachlan Hartley Yee. Can we ‘Catch ‘em All’? An exploration of the nexus between augmented reality games, urban planning and urban design. *Journal of Urban Design*, 22(6):866–880, 2017.
- [99] NC Proudlove, K Gordon, and R Boaden. Can good bed management solve the overcrowding in accident and emergency departments? *Emergency Medicine Journal*, 20(2):149–155, 2003.



- [100] Wang Qiang and Zhan Zhongli. Reinforcement learning model, algorithms and its application. In *2011 International Conference on Mechatronic Science, Electric Engineering and Computer (MEC)*, pages 1143–1146. IEEE, 2011.
- [101] Hamid Ravaghi, Saeide Alidoost, Russell Mannion, and Victoria D. Bélorgeot. Models and methods for determining the optimal number of beds in hospitals and regions: A systematic scoping review. *BMC Health Services Research*, 20(1):186, 2020.
- [102] Stefan Rayer and Ying Wang. Projections of florida population by county, 2020–2045, with estimates for 2019. *Bureau of Economics and Business Research, Florida Population Studies, Bulletin 186*, 2020.
- [103] Borja G Reguero, Michael W Beck, David N Bresch, Juliano Calil, and Imen Meliane. Comparing the cost effectiveness of nature-based and coastal adaptation: A case study from the gulf coast of the united states. *PloS one*, 13(4):e0192132, 2018.
- [104] National Research Council. Informing decisions in a changing climate. Panel on Strategies and Methods for Climate-Related Decision Support of the Committee on the Human Dimensions of Global Change, National Research Council of the National Academies, 2009.
- [105] Mathieu Reymond and Ann Nowé. Pareto-dqn: Approximating the pareto front in complex multi-objective decision problems. In *Proceedings of the adaptive and learning agents workshop (ALA-19) at AAMAS*, 2019.
- [106] Felipe Rocha, Lucas Cristiano Dantas, Luís Felipe Santos, Samela Ferreira, Bruna Soares, Alan Fernandes, Everton Cavalcante, and Thais Batista. Energy efficiency in smart buildings: An iot-based air conditioning control system. In *IFIP International Internet of Things Conference*, pages 21–35. Springer, 2019.

- [107] Hee-Tae Roh and Jang-Won Lee. Residential demand response scheduling with multiclass appliances in the smart grid. *IEEE Transactions on Smart Grid*, 7(1):94–104, 2015.
- [108] Andres Sanchez-Comas, Kåre Synnes, and Josef Hallberg. Hardware for recognition of human activities: A review of smart home and aal related technologies. *Sensors*, 20(15):4227, 2020.
- [109] Mushfiqur R Sarker, Daniel Julius Olsen, and Miguel A Ortega-Vazquez. Co-optimization of distribution transformer aging and energy arbitrage using electric vehicles. *IEEE Trans. on Smart Grid*, 8(6):2712–2722, 2016.
- [110] Charles Scawthorn, Neil Blais, Hope Seligson, Eric Tate, Edward Mifflin, Will Thomas, James Murphy, and Christopher Jones. Hazus-mh flood loss estimation methodology. i: Overview and flood hazard characterization. *Natural Hazards Review*, 7(2):60–71, 2006.
- [111] Julian Schiele, Thomas Koperna, and Jens O. Brunner. Predicting intensive care unit bed occupancy for integrated operating room scheduling via neural networks. *Naval Research Logistics (NRL)*, 68(1):65–88, feb 2021.
- [112] Greg Schrock, Ellen M Bassett, and Jamaal Green. Pursuing equity and justice in a changing climate: Assessing equity in local climate and sustainability plans in US cities. *Journal of Planning Education and Research*, 35(3):282–295, 2015.
- [113] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.
- [114] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

- [115] L. Seematter-Bagnoud, S. Fustinoni, D. H. Dung, B. Santos-Eggimann, V. Koehn, R. Bize, A. Oettli, and J. B. Wasserfallen. Comparison of different methods to forecast hospital bed needs. *European Geriatric Medicine*, 6(3):262–266, jun 2015.
- [116] Philippe Sergent, Guirec Prevot, Giovanni Mattarolo, Jérôme Brossard, Gilles Morel, Fatou Mar, Michel Benoit, François Ropert, Xavier Kergadallan, Jean-Jacques Trichet, et al. Adaptation of coastal structures to mean sea level rise. *La Houille Blanche*, (6):54–61, 2014.
- [117] Sumedha Sharma, Yan Xu, Ashu Verma, and Bijaya Ketan Panigrahi. Time-coordinated multienergy management of smart buildings under uncertainties. *IEEE Transactions on Industrial Informatics*, 15(8):4788–4798, 2019.
- [118] Elham Shirazi and Shahram Jadid. Optimal residential appliance scheduling under dynamic pricing scheme via hemdas. *Energy and Buildings*, 93:40–49, 2015.
- [119] Salman S Shuvo, Yasin Yilmaz, Alan Bush, and Mark Hafen. A markov decision process model for socio-economic systems impacted by climate change. In *International Conference on Machine Learning*. PMLR, 2020.
- [120] Salman Sadiq Shuvo, Helal Uddin Ahmed, et al. Use of machine learning for long term planning and cost minimization in healthcare management. *medRxiv*, 2021.
- [121] Salman Sadiq Shuvo, Md Rubel Ahmed, Sadia Binta Kabir, and Shaila Akter Shetu. Application of Machine Learning based hospital up-gradation policy for Bangladesh. In *ACM International Conference Proceeding Series*, pages 18–24, New York, NY, USA, dec 2020. Association for Computing Machinery.
- [122] Salman Sadiq Shuvo, Md Rubel Ahmed, Hasan Symum, and Yasin Yilmaz. Deep reinforcement learning based cost-benefit analysis for hospital capacity planning. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2021.

- [123] Salman Sadiq Shuvo, Hasan Symum, Md Rubel Ahmed, Yasin Yilmaz, and José L Zayas-Castro. Multi-objective reinforcement learning based healthcare expansion planning considering pandemic events. *IEEE Journal of Biomedical and Health Informatics*, 2022.
- [124] Salman Sadiq Shuvo and Yasin Yilmaz. Predictive maintenance for increasing ev charging load in distribution power system. In *2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (Smart-GridComm)*, pages 1–6. IEEE, 2020.
- [125] Salman Sadiq Shuvo and Yasin Yilmaz. Cibecs: Consumer input based electric vehicle charge scheduling for a residential home. In *2021 North American Power Symposium (NAPS)*, pages 1–6. IEEE, 2021.
- [126] Salman Sadiq Shuvo and Yasin Yilmaz. Home energy recommendation system (hers): A deep reinforcement learning method based on residents’ feedback and activity. *IEEE Transactions on Smart Grid*, 13(4):2812–2821, 2022.
- [127] Salman Sadiq Shuvo and Yasin Yilmaz. Demand-side and utility-side management techniques for increasing ev charging load. *IEEE Transactions on Smart Grid*, 2023.
- [128] Salman Sadiq Shuvo, Yasin Yilmaz, Alan Bush, and Mark Hafen. Modeling and simulating adaptation strategies against sea-level rise using multiagent deep reinforcement learning. *IEEE Transactions on Computational Social Systems*, 9(4):1185–1196, 2021.
- [129] Pierluigi Siano. Demand response and smart grids—a survey. *Renewable and sustainable energy reviews*, 30:461–478, 2014.
- [130] Milad Soleimani and Mladen Kezunovic. Mitigating transformer loss of life and reducing the hazard of failure by the smart ev charging. *IEEE Trans. on Industry Applications*, 56(5):5974–5983, 2020.

- [131] Chitchai Srithapon, Prasanta Ghosh, Apirat Siritaratiwat, and Rongrit Chatthaworn. Optimization of electric vehicle charging scheduling in urban village networks considering energy arbitrage and distribution cost. *Energies*, 13(2):349, 2020.
- [132] Goran Strbac. Demand side management: Benefits and challenges. *Energy policy*, 36(12):4419–4426, 2008.
- [133] Pei-Hao Su, Pawel Budzianowski, Stefan Ultes, Milica Gasic, and Steve Young. Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management. *arXiv preprint arXiv:1707.00130*, 2017.
- [134] Michael Sullivan, Josh Schellenberg, and Marshall Blundell. Updated value of service reliability estimates for electric utility customers in the United States. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2015.
- [135] Lisha Sun and David Lubkeman. Agent-based modeling of feeder-level electric vehicle diffusion for distribution planning. *IEEE Transactions on Smart Grid*, 12(1):751–760, 2020.
- [136] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [137] William VanderVeer Sweet, Robert Kopp, Christopher P Weaver, JTB Obeysekera, Radley M Horton, E Robert Thieler, Chris Eugene Zervas, et al. Global and regional sea level rise scenarios for the united states. 2017.
- [138] Robert B. Tate, Leonard MacWilliam, and Gregory S. Finlayson. A Methodology for Estimating Hospital Bed Need in Manitoba in 2020. *Canadian Journal on Aging / La Revue canadienne du vieillissement*, 24(S1):141–151, 2005.
- [139] John Tribbia and Susanne C Moser. More than information: what coastal managers need to plan for climate change. *Environmental science & policy*, 11(4):315–328, 2008.

- [140] Y Uchiyama and S Watanabe. Modeling Natural Catastrophic Risk and its Application, 2006.
- [141] Energy Information Administration US Department of Energy. Residential energy consumption survey (recs) 2009. <https://www.eia.gov/consumption/residential>.
- [142] William Valladares, Marco Galindo, Jorge Gutiérrez, Wu-Chieh Wu, Kuo-Kai Liao, Jen-Chung Liao, Kuang-Chin Lu, and Chi-Chuan Wang. Energy optimization associated with thermal comfort and indoor air control via a deep reinforcement learning algorithm. *Building and Environment*, 155:105–117, 2019.
- [143] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.
- [144] Kristof Van Moffaert and Ann Nowé. Multi-objective reinforcement learning using sets of pareto dominating policies. *The Journal of Machine Learning Research*, 15(1):3483–3512, 2014.
- [145] Wilson A Vasquez and Dilan Jayaweera. Risk-based approach for power transformer replacement considering temperature, apparent age, and expected capacity. *IET Generation, Transmission & Distribution*, 14(21):4898–4907, 2020.
- [146] José R Vázquez-Canteli and Zoltán Nagy. Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Applied energy*, 235:1072–1089, 2019.
- [147] Thomas Wahl, Shaleen Jain, Jens Bender, Steven D Meyers, and Mark E Luther. Increasing risk of compound flooding from storm surge and rainfall for major US cities. *Nature Climate Change*, 5(12):1093, 2015.

- [148] Tinghan Wang, Yugong Luo, Jinxin Liu, and Keqiang Li. Multi-objective end-to-end self-driving based on pareto-optimal actor-critic approach. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 473–478. IEEE, 2021.
- [149] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*, pages 1995–2003. PMLR, 2016.
- [150] WorldBank. Hospital beds per 1000 people, Accessed: 2021-09-26.
- [151] Worldometer. United states demographic, Accessed: 2021-09-19.
- [152] Rosanna Xia, Swetha Kannan, and Terry Castleman. The Ocean Game: The sea is rising. Can you save your town? <https://www.latimes.com/projects/la-me-climate-change-ocean-game/>, 2019.
- [153] Xu Xu, Youwei Jia, Yan Xu, Zhao Xu, Songjian Chai, and Chun Sing Lai. A multi-agent reinforcement learning-based data-driven method for home energy management. *IEEE Transactions on Smart Grid*, 11(4):3201–3211, 2020.
- [154] Yaodong Yang, Jianye Hao, Yan Zheng, and Chao Yu. Large-scale home energy management using entropy-based collective multiagent deep reinforcement learning framework. In *IJCAI*, pages 630–636, 2019.
- [155] Yasin Yilmaz, Alan Bush, and Mark Hafen. Atlantis: A Game of Sea Level Rise. <https://sites.google.com/site/nsfscsealevelrise/atlantis-a-game-of-sea-level-rise>.
- [156] Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.
- [157] Liang Yu, Tao Jiang, and Yulong Zou. Online energy management for a sustainable smart home with an hvac load and random occupancy. *IEEE Transactions on Smart Grid*, 10(2):1646–1659, 2017.

- [158] Zhiheng Zhao and Chanaka Keerthisinghe. A fast and optimal smart home energy management system: State-space approximate dynamic programming. *IEEE Access*, 8:184151–184159, 2020.
- [159] Jiaohao Zheng, Mehmet Necip Kurt, and Xiaodong Wang. Integrated actor-critic for deep reinforcement learning. In *International Conference on Artificial Neural Networks*, pages 505–518. Springer, 2021.
- [160] Bin Zhou, Wentao Li, Ka Wing Chan, Yijia Cao, Yonghong Kuang, Xi Liu, and Xiong Wang. Smart home energy management systems: Concept, configurations, and scheduling strategies. *Renewable and Sustainable Energy Reviews*, 61:30–40, 2016.
- [161] Zhecheng Zhu, Bee Hoon Hen, and Kiok Liang Teow. Estimating ICU bed capacity using discrete event simulation. *International Journal of Health Care Quality Assurance*, 25(2):134–144, 2012.



## Appendix A: Proof of Theorem 1

In Theorem 1, we analyze the government's policy since it is the most dominant agent with the full observation of other agents' actions. In the first part of the proof, we will show that if  $V(\hat{s}_t, \ell_t) = V_G(s_t, \ell_t, O_t^G)$  is nondecreasing and concave in  $\ell_t$ , then so is

$$F_m(\hat{s}_t, \ell_t) = \mathbb{E} [m\alpha_G + z_{G,t} - f(I_{R,t} + I_{B,t}) + \lambda_g V(\hat{s}_t + m, \ell_t + r_t)], \quad (\text{A.1})$$

for  $m = 0, 1, \dots, A_G$ . Government observes the residents' and businesses' decisions beforehand, thus  $R_t$  and  $B_t$  are considered constant for its decision making. We denote the next infrastructure state that includes the residents' and business' current action with  $\hat{s}_t = s_t + R_t \times \alpha_R / \alpha_G + B_t \times \alpha_B / \alpha_G$ . Assuming  $V(\hat{s}_t, \ell_t)$  is nondecreasing, i.e.,  $\frac{\partial}{\partial \ell_t} V(\hat{s}_t, \ell_t) \geq 0$ , and using the expected value of generalized Pareto-distributed  $z_{G,t}$ , we can write

$$\frac{\partial}{\partial \ell_t} F_m(\hat{s}_t, \ell_t) = \frac{m_G \eta p \ell_t^{p-1}}{(1-\xi) s_t^q} + \lambda_G \mathbb{E} \left[ \frac{\partial}{\partial \ell_t} V(\hat{s}_t + m, \ell_{t+1}) \right] \geq 0.$$

Note that  $I_{R,t}$  and  $I_{B,t}$  are determined by past data, independent of  $\ell_t$ . In the above equation, the derivative can be brought inside the integral due to the monotone convergence theorem. Assuming  $V(\hat{s}_t, \ell_t)$  is concave, i.e.,  $\frac{\partial^2}{\partial \ell_t^2} V(\hat{s}_t, \ell_t) < 0$ , for the second derivative we have

$$\frac{\partial^2}{\partial \ell_t^2} F_m(\hat{s}_t, \ell_t) = \frac{m_G \eta p (p-1) \ell_t^{p-2}}{(1-\xi) s_t^q} + \lambda_G \mathbb{E} \left[ \frac{\partial^2}{\partial \ell_t^2} V(\hat{s}_t + m, \ell_{t+1}) \right] < 0$$

since  $0 < p < 1$ . Hence, it is sufficient to show that  $V(\hat{s}_t, \ell_t)$  is nondecreasing and concave.

Finding the value function iteratively (i.e., value iteration) is a common approach which is known to converge [136]:

$$\lim_{i \rightarrow \infty} V_i(\hat{s}, \ell) = V(\hat{s}, \ell),$$

where, for brevity, we drop the time index from now on. We will next prove that  $V(\hat{s}, \ell)$  is nondecreasing and concave iteratively. Initializing all the state values as zero, i.e.,  $V_0(\hat{s}, \ell) = 0, \forall s, \ell$ , after the first iteration we get

$$\begin{aligned} V_1(\hat{s}, \ell) &= \min_G \left\{ \mathbb{E}[\alpha_G G + z_G(s, \ell) + \lambda_G V_0(\hat{s} + G, \ell + r) + \text{const.}] \right\} \\ &= \text{const.} + \mathbb{E}[z_G(s, \ell)] = \text{const.} + \mu + \frac{m_G \eta \ell^p}{(1 - \xi) s^q}. \end{aligned}$$

Differentiating with respect to  $\ell$ , we get

$$\frac{\partial}{\partial \ell} V_1(\hat{s}, \ell) = m_G \eta p \frac{\ell^{p-1}}{(1 - \xi) s^q} \geq 0, \quad \forall s, \quad (\text{A.2})$$

$$\frac{\partial^2}{\partial \ell^2} V_1(\hat{s}, \ell) = m_G \eta p(p - 1) \frac{\ell^{p-2}}{(1 - \xi) s^q} < 0, \quad \forall s,$$

since  $m_G, \eta > 0, p \in (0, 1), q > 0, \xi < 0$ . Thus,  $V_1(\hat{s}, \ell)$  is nondecreasing and concave in  $\ell$  for all  $s$ . Similarly, the value function after the second iteration becomes

$$\begin{aligned} V_2(\hat{s}, \ell) &= \min_G \left\{ \mathbb{E}[\alpha_G G + z_G(s, \ell) + \lambda_G V_1(\hat{s} + G, \ell + r) + \text{const.}] \right\} \\ &= \min_G \left\{ \alpha_G G + \mu + \frac{m_G \eta \ell^p}{(1 - \xi) s^q} + \mu \lambda_G + \lambda_G \mathbb{E} \left[ \frac{m_G \eta (\ell + r)^p}{(1 - \xi) (\hat{s} + G)^q} \right] + \text{const.} \right\}. \end{aligned}$$

Denoting the optimum action with  $\hat{G}$  we will show that  $V_2(\hat{s}, \ell)$  is nondecreasing and concave for any  $\hat{G}$ . Moreover, the pointwise minimum of nondecreasing and concave functions is also

nondecreasing and concave. Taking the derivative with respect to  $\ell$  we get

$$\begin{aligned}\frac{\partial}{\partial \ell} V_2(\hat{s}, \ell) &= \frac{\partial}{\partial \ell} \left\{ \frac{m_G \eta \ell^p}{(1-\xi)s^q} + \lambda_G \frac{m_G \eta \mathbb{E}[(\ell+r)^p]}{(1-\xi)(\hat{s} + \hat{G})^q} \right\} \\ &= m_G \eta p \frac{\ell^{p-1}}{(1-\xi)s^q} + \lambda_G m_G \eta p \frac{\mathbb{E}[(\ell+r)^{p-1}]}{(1-\xi)(\hat{s} + \hat{G})^q} \geq 0, \quad \forall s \\ \frac{\partial^2}{\partial \ell^2} V_2(\hat{s}, \ell) &= m_G \eta p(p-1) \frac{\ell^{p-2}}{(1-\xi)s^q} \\ &\quad + \lambda_G m_G \eta p(p-1) \frac{\mathbb{E}[(\ell+r)^{p-2}]}{(1-\xi)(\hat{s} + \hat{G})^q} < 0, \quad \forall s.\end{aligned}$$

Hence,  $V_2(\hat{s}, \ell)$  is nondecreasing and concave. Now, for any  $i$ , given that  $V_{i-1}(\hat{s}, \ell)$  is nondecreasing and concave, we can write

$$\begin{aligned}\frac{\partial}{\partial \ell} V_i(\hat{s}, \ell) &= m_G \eta a \frac{\ell^{p-1}}{(1-\xi)s^q} + \lambda_G \mathbb{E} \left[ \frac{\partial}{\partial \ell} V_{i-1}(\hat{s} + \hat{G}, \ell) \right] \geq 0, \quad \forall s \\ \frac{\partial^2}{\partial \ell^2} V_i(\hat{s}, \ell) &= m_G \eta p(p-1) \frac{\ell^{p-2}}{(1-\xi)s^q} \\ &\quad + \lambda_G \mathbb{E} \left[ \frac{\partial^2}{\partial \ell^2} V_{i-1}(\hat{s} + \hat{G}, \ell) \right] < 0, \quad \forall s. \quad (\text{A.3})\end{aligned}$$

Consequently, by mathematical induction,  $V(\hat{s}, \ell)$  is nondecreasing and concave.

The second part of the proof is to show that  $\frac{\partial}{\partial \ell} F_m(\hat{s}, \ell) < \frac{\partial}{\partial \ell} F_{m-1}(\hat{s}, \ell)$ . Similar to the first part, if we show that

$$\frac{\partial}{\partial \ell} V(\hat{s} + m, \ell) < \frac{\partial}{\partial \ell} V(\hat{s} + m - 1, \ell),$$

we can conclude that  $\frac{\partial}{\partial \ell} F_m(\hat{s}, \ell) < \frac{\partial}{\partial \ell} F_{m-1}(\hat{s}, \ell)$  since

$$\begin{aligned}\frac{\partial}{\partial \ell_t} F_m(\hat{s}_t, \ell_t) &= \frac{m_G \eta p \ell_t^{p-1}}{(1-\xi)s_t^q} + \lambda_G \mathbb{E} \left[ \frac{\partial}{\partial \ell_t} V(\hat{s}_t + m, \ell_t + r_t) \right] \\ \frac{\partial}{\partial \ell_t} F_{m-1}(\hat{s}_t, \ell_t) &= \frac{m_G \eta p \ell_t^{p-1}}{(1-\xi)s_t^q} + \lambda_G \mathbb{E} \left[ \frac{\partial}{\partial \ell_t} V(\hat{s}_t + m - 1, \ell_t + r_t) \right].\end{aligned}$$

Starting again with  $V_0(\hat{s}, \ell) = 0, \forall s, \ell$ , from (A.2) we can write the following inequality for the first iteration

$$\frac{\partial}{\partial \ell} V_1(\hat{s}+m, \ell) = m_G \eta p \frac{\ell^{p-1}}{(1-\xi)(\hat{s}+m)^q} < \frac{\partial}{\partial \ell} V_1(\hat{s}+m-1, \ell) = m_G \eta p \frac{\ell^{p-1}}{(1-\xi)(\hat{s}+m-1)^q}.$$


For any  $i$ , given that  $\frac{\partial}{\partial \ell} V_{i-1}(\hat{s}+m, \ell) < \frac{\partial}{\partial \ell} V_{i-1}(\hat{s}+m-1, \ell)$ , from (A.3) we have


$$\begin{aligned} & m_G \eta p \frac{\ell^{p-1}}{(1-\xi)(\hat{s}+m)^q} + \lambda_G \mathbb{E} \left[ \frac{\partial}{\partial \ell} V_{i-1}(\hat{s}+m, \ell) \right] \\ & < m_G \eta p \frac{\ell^{p-1}}{(1-\xi)(\hat{s}+m-1)^q} + \lambda_G \mathbb{E} \left[ \frac{\partial}{\partial \ell} V_{i-1}(\hat{s}+m-1, \ell) \right], \text{ i.e.,} \\ & \frac{\partial}{\partial \ell} V_i(\hat{s}+m, \ell) < \frac{\partial}{\partial \ell} V_i(\hat{s}+m-1, \ell) \end{aligned}$$

As a result, by mathematical induction we conclude that  $\frac{\partial}{\partial \ell} V(\hat{s}+m, \ell) < \frac{\partial}{\partial \ell} V(\hat{s}+m-1, \ell)$ .

## Appendix B: Copyright Permissions

The permission below is for the reproduction of material in Chapter 2.

Home Help Live Chat Sign In Create Account



**Home Energy Recommendation System (HERS): A Deep Reinforcement Learning Method Based on Residents' Feedback and Activity**

Author: Salman Sadiq Shuvo  
Publication: IEEE Transactions on Smart Grid  
Publisher: IEEE  
Date: July 2022

Copyright © 2022, IEEE

**Thesis / Dissertation Reuse**

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

*Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:*

- 1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
- 2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
- 3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

*Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:*

- 1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication].
- 2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
- 3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to [http://www.ieee.org/publications\\_standards/publications/rights/rights\\_link.html](http://www.ieee.org/publications_standards/publications/rights/rights_link.html) to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK

CLOSE WINDOW

The permission below is for the reproduction of material in Chapter 3.



Modeling and Simulating Adaptation Strategies Against Sea-Level Rise Using Multiagent Deep Reinforcement Learning

Author: Salman Sadiq Shuvo  
Publication: IEEE Transactions on Computational Social Systems  
Publisher: IEEE  
Date: August 2022

Copyright © 2022, IEEE

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

- 1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
- 2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
- 3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

- 1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
- 2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
- 3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to [http://www.ieee.org/publications\\_standards/publications/rights/rights\\_link.html](http://www.ieee.org/publications_standards/publications/rights/rights_link.html) to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK

CLOSE WINDOW

The permission below is for the reproduction of material in Chapter 4.



Multi-Objective Reinforcement Learning Based Healthcare Expansion Planning Considering Pandemic Events

Author: Salman Sadiq Shuvo  
Publication: IEEE Journal of Biomedical and Health Informatics  
Publisher: IEEE  
Date: 2022

Copyright © 2022, IEEE

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

- 1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
- 2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
- 3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

- 1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
- 2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
- 3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to [http://www.ieee.org/publications\\_standards/publications/rights/rights\\_link.html](http://www.ieee.org/publications_standards/publications/rights/rights_link.html) to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK

CLOSE WINDOW

The permission below is for the reproduction of material in Chapter 5.



**Demand-side and Utility-side Management Techniques for Increasing EV Charging Load**

Author: Salman Sadiq Shuvo  
Publication: IEEE Transactions on Smart Grid  
Publisher: IEEE  
Date: Dec 31, 1969

Copyright © 1969, IEEE

**Thesis / Dissertation Reuse**

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

*Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:*

- 1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
- 2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
- 3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

*Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:*

- 1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [Year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
- 2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
- 3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to [http://www.ieee.org/publications\\_standards/publications/rights/rights\\_link.html](http://www.ieee.org/publications_standards/publications/rights/rights_link.html) to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK

CLOSE WINDOW