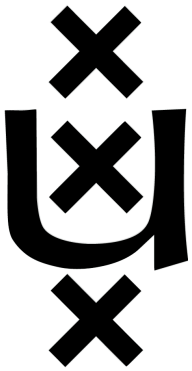

Bachelor's Thesis

Global Warming Mitigation: An Empirical Multi-Agent Reinforcement Learning Approach

written by
Anton X. Pham

under the guidance of
Dr. Maurice A.L. Koster

in partial fulfillment of the requirements for the degree of
BSc Econometrics & Data Science

at the *Universiteit van Amsterdam.*

Date of the submission:
July 31, 2024

Examiners of the Thesis:
Dr. Maurice A.L. Koster (supervisor)
Dr. Kees Jan van Garderen (coordinator)
Dr. Patrick R. Stastra (lecturer)

This page is intentionally left blank

Dedication

To my beloved grandparents and parents,

Without your unwavering support and belief in me, none of my accomplishments, including this thesis, would have been possible. There were times when I was not the best son and grandchild, yet you never gave up on me. Your encouragement, persistence, and trust have guided me, and I strive every day to live up to your faith in me. You have heard it before, but once more – thank you so much for everything!

Acknowledgment

I extend my deepest gratitude to the remarkable individuals whose phenomenal support has nurtured both my academic pursuits and personal development. Your guidance has propelled me to strive for excellence and continuously engage in intellectual challenges.

Jared Frazier – thank you for giving me so much of your experience and knowledge in the thesis, and for being the role model in various aspects of my life.

Dr. Maurice Koster – thank you for inspiring the topic of this research and for using your time to give me an in-depth and comprehensive understanding of the topic right from the start. Furthermore, thank you so much for being understanding and always granting me more than enough support and extension to complete my thesis.

Dr. Cláudia Custodio, Dr. Emilia Bunea, Dr. David Stolin, and Dr. Albert Menkveld – thank you for giving me valuable feedback and encouragement to further improve my work.

Furthermore, I really want to thank the support from the managers at the Universiteitsbibliotheek Singel who allowed me to use as many computer units as needed for the model training and evaluation. Therefore, last but not least, this thesis could not be completed without the knowledge instilled everyone who taught me new things or helped me throughout my time in the Netherlands.

Statement of Originality

This document is written by Student Anton X. Pham who declares to take full responsibility for the contents of this document. I declare that the text and the work presented in this document are original and that no sources other than those mentioned in the text and its references have been used in creating it. UvA Economics and Business is responsible solely for the supervision of completion of the work and submission, not for the contents.

Universiteit van Amsterdam
Amsterdam School of Economics

Bachelor's Thesis

Global Warming Mitigation: An Empirical Multi-Agent Reinforcement Learning Approach

July 31, 2024

Student:
Anton X. Pham (13179543)
BSc Econometrics & Data Science
Faculty of Economics & Business

Supervisor:
Maurice A.L. Koster
Section Quantitative Economics
Faculty of Economics & Business

Abstract

This thesis investigates the efficacy of serial cost-sharing taxation policies in regulating emission levels from petroleum industry production, which is crucial for meeting the Paris Agreement's goal of limiting global temperature rise to below 2 degrees Celsius. Utilizing an innovative heterogeneous-agent reinforcement learning model, the study identifies key factors influencing firms' economic viability while achieving environmental targets. Findings reveal that linear cost functions offer simplicity and predictability, facilitating stable market conditions but may fall short in driving substantial emissions reductions. Convex cost functions effectively promote sustainable practices by imposing higher costs with increased production, encouraging firms to remain within environmentally safe limits. However, concave cost functions, while enhancing efficiency, can lead to market volatility and are suitable for highly regulated environments. Firm heterogeneity plays a critical role, with diverse firms showing superior adaptability and efficiency in dynamic markets. Additionally, a competitive, atomistic market structure fosters balanced and efficient outcomes, optimizing production and emissions reductions. These insights highlight the necessity for nuanced regulatory frameworks that integrate appropriate cost functions, firm diversity, and market structure to manage emissions sustainably and maintain economic stability.

Keywords: Cost-Sharing Mechanisms, Multi-Agent System, Heterogeneous-Agent Reinforcement Learning, Climate Change, Emissions Regulation, U.S. Gasoline.

The code for this thesis is published in this git repository.

Table of Contents

1 Introduction

Global warming, driven by a large collection of indicators and predominantly by human activities such as fossil fuel combustion, poses one of the most significant challenges of the 21st century (Forster et al., 2023). The Paris Agreement, an international treaty adopted in 2015, aims to combat climate change by limiting the global temperature rise to well below 2 degrees Celsius above pre-industrial levels, with efforts to limit the increase to 1.5 degrees Celsius (Agreement, 2015). Achieving this goal is crucial to mitigating the severe impacts of climate change, which include rising sea levels, more frequent and intense extreme weather events, loss of biodiversity, and significant economic disruptions. Failure to meet these targets could result in catastrophic consequences for natural ecosystems and human societies alike, leading to massive financial losses and social instability.

The economic costs of global warming are substantial. According to estimates from the research of Meinshausen et al. (2009), unmitigated climate change could reduce global GDP by up to 18% by 2050 if temperatures rise by 3.2 degrees Celsius. This includes direct impacts such as damage to infrastructure from extreme weather events, as well as indirect costs like reduced agricultural yields, health care expenses due to increased prevalence of climate-related diseases, and forced migration due to habitable land loss. Moreover, the environmental degradation resulting from global warming can lead to irreversible losses in biodiversity and ecosystem services, further exacerbating economic and social challenges.

In light of these profound risks, it is imperative for policymakers to develop strategies that effectively manage the production of environmentally harmful products while ensuring the economic viability of industries critical to societal advancement. The evolution of environmental concerns towards pressing global issues has driven the development of policies and regulations at national and international levels. Governments have implemented regulatory measures to distribute compliance burdens to protect natural resources and reduce pollution. Internationally, multilateral agreements facilitated by organizations like the UNEP and IPCC highlight the need for collective action on global environmental challenges. Meanwhile, the economic impact of environmental externalities, including costs to public health and ecosystems, underscores the importance of implementing a production limitation policy to promote sustainable development. Luqman et al. (2018) emphasize the necessity for fair distribution of pollution reduction costs among countries, emphasizing international cooperation's significance. As environmental challenges persist, collaborative efforts remain crucial in mitigating environmental damage and fostering sustainability.

The petroleum industry, in particular, is at the heart of this dilemma. Petroleum products are fundamental to modern technological and social development, providing essential energy resources that power transportation, manufacturing, and numerous other sectors. However, the industry's production processes are major contributors to greenhouse gas emissions, necessitating a delicate balance between environmental sustainability and economic vitality. Consequently, policy and regulatory frameworks shape cost-sharing mechanisms and incentivize firms to reduce environmental externalities through carbon pricing schemes, cap-and-trade systems, and subsidies. Furthermore, the effectiveness of emissions trading systems (ETSs) highlight institutional learning and administrative improvements across jurisdictions, as reviewed by Narassimhan et al. (2018). Cap-and-trade systems and subsidies provide flexibility and financial incentives for emissions reduction and technology adoption, fostering innovation and sustainable development. These frameworks drive environmental progress, internalize external costs, and stimulate collaboration and innovation essential for transitioning to sustainable economies.

This thesis explores the potential of cost-sharing taxation policies to regulate emission levels from firm production activities in the petroleum industry. Among the cost-sharing mechanisms, serial cost-sharing is an algorithm which distributes systematically larger portions of the cost pool to large producers. As firms with lower production levels would incur lower costs than the actual costs of the externality they will have created, by using this mechanism to distribute the costs of environmental damage among firms, policymakers may be able to incentivize reductions in harmful emissions without stifling economic activity. Consequently, the question guiding this investigation is: How can serial cost-sharing taxation policy regulate the emission levels from firm production activities in the petroleum industry, and what factors influence the likelihood of firms finding optimal strategies to maintain economic viability?

Understanding the dynamics of cost-sharing taxation and its effectiveness in promoting environmentally responsible production is critical. Firms' responses to these policies are influenced by various factors, including their production costs, market demand, technological capabilities, and regulatory environment. This thesis employs a multi-agent reinforcement learning (MARL) framework to model the interactions between firms and regulators, capturing the complexities of decision-making processes in a competitive market. The MARL approach allows for the simulation of different scenarios, providing insights into the conditions under which firms are likely to adopt optimal strategies that align with both economic and environmental objectives.

The findings of this research will contribute to the ongoing discourse on sustainable industrial practices, offering practical recommendations for policymakers seeking to balance environmental protection with economic growth. By advancing our understanding of cost-sharing mechanisms and their impact on firm behavior, this thesis aims to support the development of effective policies that drive the transition towards a more sustainable and resilient global economy.

2 Theoretical Framework

2.1 Historical Overview

The origin of environmental externalities and their recognition as market failures finds its roots in early economic theories and sociological perspectives. Marshall's (1890) pioneering work on externalities highlighted the significant effects that economic activities, such as production and consumption, can have beyond the immediate parties involved in a transaction. His insights established a foundational understanding of how these externalities—whether positive or negative—can influence broader societal welfare. Marshall's contributions enabled economists to explore the ramifications of externalities on public policy and resource allocation, shaping modern economic thought and practice. Additionally, Coase (1960) challenges the traditional view of business liability for harm caused to others, proposing a reciprocal problem-solving approach. Coase advocates for solutions that maximize production value while considering the costs of market transactions and legal rights in addressing social costs.

Arthur Pigou's work on externalities laid the groundwork for understanding the concept of internalizing external costs through taxation (Pigou, 1920). Pigou introduced "Pigovian taxes," corrective measures designed to internalize externalities and align private costs with social costs. His research underscores the importance of government intervention in promoting economic efficiency and social welfare. Furthermore, Dahlman (1979) delves into the relationship between externalities and transaction costs, challenging the notion that externalities necessarily indicate market failure. Dahlman argues that transaction costs are the fundamental cause of persistent externalities and questions the effectiveness of government intervention in correcting market failures attributed to externalities. This historical trajectory underscores the significance of cost-sharing mechanisms in addressing environmental challenges.

Equivalently, the history of economic studies of environmental issues can be traced back to the eighteenth century, as evidenced by Condorcet's citation of agricultural activities causing air pollution that led to illnesses in neighboring homes (Rothschild, 2001; Sandmo, 2015). He reasons that the air pollution that led to illnesses in neighboring homes nececssitates intervention from the government to prohibit harmful activities or undertake public works to restore air quality. Additionally, Princen (2001) argues for an ecologically informed "consumption angle" on economic activity, challenging the traditional supply-demand dichotomy and emphasizing the role of excess consumption in environmental degradation. Against this backdrop, cost sharing emerges as a mechanism to allocate the costs of externalities among firms and society, with Aadland and Kolpin (2004) highlighting the significant influence of environmental factors on cost-sharing rule selection. This historical trajectory underscores the significance of cost-sharing mechanisms in addressing environmental challenges.

2.2 Cost-Sharing Models and Strategies

Cost-sharing models are pivotal in addressing environmental externalities by distributing the costs of mitigation efforts among stakeholders. As early as the 1950s, the conceptual underpinning of these models traces back to the seminal work of Shapley et al. (1953), as documented in "A value for n-person games," where he introduced the Shapley value. This value allocates outcomes uniquely among coalitions of n players, representing a substantial advancement in economic theory. Over time, the Shapley value has been applied extensively in cost allocation problems. In this context, such problems are viewed through the lens of cooperative game theory, which interprets them as games that can utilize the Shapley value to distribute costs among players. Further advancements in the field, notably in the study of non-atomic games from (Aumann & Shapley, 1975), have expanded on these concepts, providing deeper insights into the implications of the Shapley value for cost-sharing mechanisms within environmental economics.

Drawing further inspiration from Shapley value and other seminal works in game theory and allocation procedures, researchers have developed diverse approaches to cost sharing that cater to varying economic contexts. Billera and Heath (1982) lay the foundation for modern cost allocation procedures by establishing axiomatic properties that yield a unique method, providing a robust framework for equitable distribution of shared costs. Building upon this groundwork, Moulin and Shenker (1992) introduce the serial cost-sharing rule, offering a systematic mechanism for allocating costs among a fixed group of agents. This rule ensures fairness and efficiency in cost allocation by incrementally assigning shares based on individual demands, culminating in a unique Nash equilibrium that withstands coalitional deviations. Koster et al. (1998) then further extend the serial cost-sharing method to accommodate heterogeneous goods demand, introducing a class of serial extensions based on preordering principles. This advancement underscores the adaptability and scalability of cost-sharing mechanisms, enabling their application across diverse economic landscapes. At the same time, Moulin (1999) expands on the concept with incremental cost sharing, which further enhances stability and fairness in allocation, particularly in settings with convex preferences and idiosyncratic consumption patterns. These models equip policymakers and stakeholders with versatile tools to internalize external costs, incentivize pollution reduction, and foster sustainable development, thereby promoting more equitable and efficient environmental management. Furthermore, Moulin (2002) effectively connects the cost-sharing frameworks with the models concerning rationing and taxation, establishing a conceptual linkage between these two issues. Nevertheless, while existing theoretical frameworks provide valuable insights into cost-sharing mechanisms and their environmental impact, there remains a notable gap in translating these mathematical concepts into actionable solutions.

2.3 Application of Reinforcement Learning

The integration of advanced machine learning techniques holds promise for bridging this gap in multiple economic disciplines, offering avenues for empirical validation and optimization in real-world settings. Among these techniques, reinforcement learning (RL) aligns well with many economic scenarios where various entities must continuously adapt to changing conditions, decisions are interdependent, and outcomes evolve over time. This is an interdisciplinary machine learning technique where an agent learns to make decisions by taking actions in an environment to maximize its cumulative reward. Deep learning, on the other hand, involves neural networks with many layers (hence "deep") that can capture complex patterns in data. When combined, deep reinforcement learning (DRL) leverages deep neural networks to enable RL agents to handle high-dimensional state and action spaces, thus providing a powerful framework for tackling complex decision-making problems.

For a comprehensive picture of RL and DRL research, please refer to the recent reviews of the literature from Charpentier et al. (2021) and Mosavi et al. (2020). Charpentier et al. (2021) underscore the applicability of RL techniques across various economic disciplines, including economics, game theory, operations research, and finance, demonstrating their potential for empirical validation and optimization in real-world economic scenarios. In particular, their research highlights the versatility of RL frameworks in addressing optimal control problems by leveraging advancements in computational science and deep learning algorithms. Similarly, Mosavi et al. (2020) provide an in-depth review of deep reinforcement learning (DRL) methods and their application to economics,

emphasizing the scalability and robustness of DRL in handling high-dimensional and nonlinear economic data, and highlighting its superior performance in optimizing complex economic systems.

Additionally, there have been a plethora of development and applications of MARL techniques in economics disciplines. During the same time when he introduced the Shapley value, Shapley (1953) formalized the concept of stochastic game, then also called a Markov game by Littman (1994), which represents a fully cooperative multi-agent task. In the context of the Prisoner's Dilemma game, Banerjee and Sen (2007) have shown that although the Nash Equilibrium is Pareto-dominated, conditional joint action learners (CJAL) can achieve Pareto-optimal outcomes by maximizing social welfare through cooperation. Safiri et al. (2022) then introduce a multi-agent distributed reinforcement learning (MADRL) algorithm for optimizing the incremental cost of units in different economic dispatch problems. By incorporating consensus control mechanisms, their approach showcases the utility of RL in addressing complex optimization problems with distributed decision-making dynamics.

In a broad economic context, Atashbar and Shi (2023) contribute to the application of RL in macroeconomic modeling by constructing a reinforcement learning-based AI-macroeconomic simulator. Utilizing a deep deterministic policy gradient (DDPG) approach within a real business cycle macroeconomic model, their study demonstrates the efficacy of RL in approximating optimal decision-making in both deterministic and stochastic environments. The findings suggest the potential of RL-based approaches in enhancing macroeconomic simulations and policy analysis, paving the way for further exploration and refinement of RL algorithms in macroeconomic research. Furthermore, in tandem with the cost-sharing problem, Zheng et al. (2020) introduce a data-driven approach to dynamic tax policy design using RL techniques. Their two-level deep RL framework learns dynamic tax policies in active economies without relying on traditional economic modeling assumptions. By leveraging AI-driven tax policies, their study showcases improvements in the trade-off between economic equality and productivity, highlighting the potential of RL algorithms in shaping policy interventions for socioeconomic optimization.

Overall, these studies collectively demonstrate the growing significance of RL techniques in addressing diverse economic challenges and optimizing decision-making processes in dynamic and complex economic environments. Therefore, taking inspiration from these research, the objective of this study is to employ multi-agent reinforcement learning to explore the problem under serial cost-sharing rule and determine its efficacy in incentivizing firms to mitigate environmental externalities and achieve sustainable outcomes in a diverse set of scenarios.

3 Methodology

The purpose of this section is to provide a comprehensive description of the model to be used in investigating the efficacy of serial cost-sharing in mitigating emissions resulting from firm production activities. The detailed explanation of the mathematical validation behind this mechanism and other most common cost-sharing algorithms is established in Appendix C.

First of all, the research investigates the application of cost sharing in a motor gasoline market of the United States. The reason behind this choice as well the the design of the model are illustrated in Appendices A and F. A deep multi-agent reinforcement learning (MARL) framework is utilized to investigate the efficacy of cost-sharing mechanisms in mitigating emissions resulting from firm production activities. The utilization of MARL facilitates the examination of intricate interactions among autonomous agents operating within a dynamic economic environment. The new heterogeneous-agent reinforcement learning (HARL) framework with multitude of HARL learning algorithms is utilized as the solver for the environment. A concise overview of reinforcement learning, MARL, and HARL is provided in Appendix D. In order to investigate the research comprehensively, three different experiments are executed, in which their details are comprehensively covered in Appendix G.

- The first experiment involves inspecting into the characteristics of of the cost function used by the policymaker to penalize firms' heavy production. It does this by alternating between the use of convex, linear, and concave functions as the cost function. In this experiment, the default environment is used which consists of two big firms which comprise a sole duopoly in the market.
- The second experiment investigates the impact of price fluctuation on the production of the firms, to aid in the understanding of the price characteristics in a cost-sharing environment. This experiment is designed through investigating environments with either a constant- or a dynamic-price market, consisting of either homogeneous or heterogeneous firms. The dynamic-price market is assumed to resemble a competitive market where the price can be largely influenced by the total supply and the demand of the gasoline market.
- The third experiment finally considers the change in the number of firms in the environment and examines the effective of cost allocation on these firms' production decisions. This experiment is categorized into four categories:

  1. A monopoly, which is essentially the primary single-agent reinforcement learning model
  2. A duopoly, the most basic multi-agent reinforcement learning model
  3. A tripoly with three firms
  4. A oligopoly with four firms
  5. A representation of a competitive settings with twelve firms

  This experiment also considers the impact of heterogeneity of firms on the varying density of agents in the market.

3.1 Cost-Sharing Environment Model

In this research, the environment encapsulates variables representing the firms' production capacities and consumer demand fluctuations at each timestep, facilitating the simulation of diverse scenarios. The complete explanation of the demand model is found in Represented as a whole value of number of barrels of motor gasoline (in thousands), the production capacity is represented by the continuous range of possible output a firm can produce in a single week. The environment also incorporates consumer demand dynamics, reflecting the changes in market preferences and purchasing behaviors.These variables collectively define the economic landscape in which firms operate, providing a comprehensive framework for evaluating the effectiveness of different cost-sharing strategies in addressing environmental externalities.

3.2 Agents' Algorithm

In our MARL framework, each agent represents a firm engaged in production activities contributing to environmental pollution. The number of agents corresponds to the number of firms under investigation, with each agent equipped with an proximal policy optimization (PPO) algorithm, as initially introduced by the OpenAI researchers (Schulman et al., 2017). The most appropriate method for this research is considered to be the heterogeneous-agent reinforcement learning (HARL) framework (Zhong et al., 2024) which allows for the use of agents which exhibit varying characteristics from each other. The HARL framework can be further understood from Appendix D and in the works (Zhong et al., 2024**<empty citation>**).

PPO stands as a prominent reinforcement learning algorithm following their properties of an on-policy gradient method which combines the interaction of both the actors which execute the policies and the critic which evaluates the current value (Sutton & Barto, 2018). It is designed to address scenarios with continuous action spaces, a common feature in many real-world applications. Furthermore, the HAPPO it can address multi-dimensional state and continuous action spaces.

Adapted for multi-agent scenarios, the multi-agent proximal policy optimization (MAPPO) algorithm enables each agent to operate independently with its own PPO algorithm. This independence allows agents to sample data and learn without direct interaction with other agents, enhancing scalability and efficiency in multi-agent environments. While MAPPO supports optional information sharing between agents to improve learning performance, it does not impose mandatory collaboration, ensuring adaptability to diverse scenarios where collaboration may vary. To facilitate optional information sharing, o is used to denote local observations specific to each agent and s to represent the global state shared among agents. This design enables agents to exchange relevant information selectively, optimizing learning while preserving autonomy. To elevate the algorithm further, the heterogeneous-agent proximal policy optimization (HAPPO) can be applied in order to allow the use of agents with diverse sets of characteristics, roles, observation spaces, and actions spaces.

3.3 Reward and Value Functions

The reward function within the environment serves to quantify the financial and environmental ramifications of firms' actions, factoring in their individual profits while considering the cost-sharing regulations under review. The explanation for each cost-sharing mechanism are listed in Appendix C and the design of the rewarding mechansim of the environment is described in the section F.2 of Appendix F.

$$R_{a_i}(s,\ s') = q_i p_i - \xi_i^m(C,\ \boldsymbol{q}) \tag{1}$$

where

a_i : action taken by agent i
q_i : quantity produced by agent i
p_i : profit per item of agent i
ξ_i^m : cost share of agent i using cost-sharing mechanism m
C : cost function of pollution
$\boldsymbol{q}$: vector of production of n agents

Although it is possible to enhance the performance of the HAPPO algorithm by applying reward rescaling techniques as observed in past studies (Duan et al., 2016; Gu et al., 2016), the rewards of this environment are not normalized as this does not compromise the performance of the HARL algorithm (Zhong et al., 2024).

3.4 Assumptions

Firstly, the cost function employed to quantify the financial and environmental implications of firms' actions adheres to the following criteria. It is assumed that the default cost function is strictly convex, accurately reflecting the increasing marginal cost of pollution over the course of production activities. Additionally, the cost function is designed to capture the intricate trade-offs between firms' profits and the societal costs associated with environmental damage, ensuring a comprehensive assessment of the implications of various cost-sharing mechanisms.

Furthermore, the behavior of agents operating within the HARL framework is governed by specific constraints and characteristics. It is assumed that agents possess bounded rationality, making decisions based on the shared state and their individualized observations. Additionally, as delinated in Moulin and Shenker (1992), agents are constrained by certain operational limitations; for instance, the input and output of the production process are deemed non-transferable, and agents are unable to merge into a single entity or split demand into smaller units nor change to a different good. These assumptions allow the result of the simulation to be comparable across all cost-sharing mechanisms.

Moreover, the utilization of the HAPPO algorithm introduces further considerations regarding information sharing among agents. While information sharing is facilitated within the HARL framework to enhance learning performance, it is essential to ensure that such sharing does not compromise the independence of individual agents. The algorithm represents a stepwise policy update scheme, in which all agents are able to observe the policy update of their previous agents and use this information to improve their policies acoordingly. This proves to allow the minimization of the L2 norm of the policy gradients of all agents and can be argued to mimmic the behaviours of firms in the market in their development of the production stategies. Hence, its attempt to achieve this assumption is by randomizing the order of the agents at every time-step at the start of the policy update (Zhong et al., 2024). This assumption underscores the balance between collaboration and autonomy in multi-agent systems, reflecting real-world dynamics where firms may engage in strategic interactions while retaining their independence.

The cost-sharing environment is formulated and tesed based on the structure of a multi-agent parallel PettingZoo environment (Terry et al., 2021). Essentially, all agents receive the global state space and their individualized observations generated at the exact time-step, and use these information update their policies. On the other hand, the stepwise policy update process implies that the learning process is not completely parallel, at least in the updating procedure of the algorithm. However, this does not compromise the parallelization nature of this environment because each agent only have access to the current policy of a finite number of earlier agents. As the agents density grows, asymptotically the portion of policy observation of any individual agent approaches zero. In conclusion, this MARL model employs a HARL algorithm which is executed inside a parallel environment from which it receives the shared states and individual observations of all agents, and then uses those information to sequentially update the policies of each agents (Zhong et al., 2024).

Lastly, the economic and policy context within which the simulation unfolds remains stable throughout the simulation period. During the entire training in a scenario, the context is represented as a detailed set of complicated hyperparameters to the environment. This assumption enables the evaluation of the long-term effects of cost-sharing mechanisms on firms' behavior and environmental outcomes, providing valuable insights into the sustainability and efficacy of various policy interventions. By maintaining stability in the policy environment, the research can effectively analyze the impacts of cost-sharing mechanisms under consistent conditions, facilitating robust conclusions and policy recommendations.

4 Results

To investigate the questions, eighteen different scenarios with altering algorithm configurations, types of algorithm, environment configurations, and firms characteristics are used . The results of the eighteen runs that are used to answer the three areas of interest are reported in the following corresponding sections.

4.1 Cost Function Characteristics

In this experiment which encompasses three scenarios with different cost functions, as the final time-steps are approached, firms continue to search for a joint optimal policy. Notably, equilibrium has not yet been achieved; in the case of the convex cost function, firms are still adjusting strategies, in the case of linear and concave cost functions, firms have opted into abstaining from further action altogether. The training outputs from the algorithm, depicted in Figure 1 further illustrate these ongoing dynamics and outcomes.

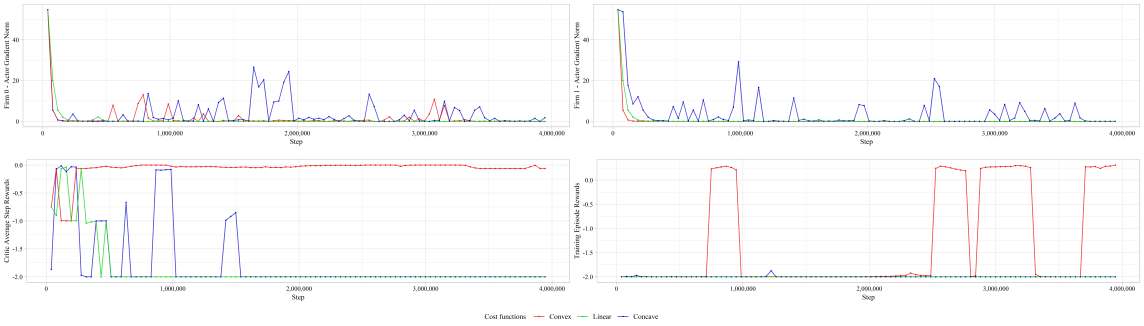

Figure 1: *The results of HARL models with different types of cost function.*

4.1.1 Gradient Norm of the Agents

The gradient norm plots for agents operating under a linear cost function exhibit a notably smooth trajectory. This smoothness underscores the straightforward relationship between production decisions and costs, where incremental changes in production result in proportional adjustments in net earnings and subsequent rewards. Such clarity in incentive structure facilitates stable policy optimization, potentially enabling quicker convergence to optimal strategies. Conversely, under a convex cost function, the gradient norm plots reveal frequent, minor spikes in policy updates. These fluctuations occur when firms' production levels diverge significantly, prompting adjustments aimed at minimizing incurred costs. The convex nature of the cost function introduces non-linear penalties, incentivizing firms to moderate their production to avoid escalating costs. This dynamic fosters a learning process characterized by gradual adjustments as firms navigate the cost landscape to optimize their strategies. For the concave cost function, the gradient norm plots display intermittent, substantial spikes in policy updates. These spikes indicate that firms' policies are highly sensitive to variations in costs, particularly under scenarios where cost impacts fluctuate non-linearly with production levels. This heightened sensitivity suggests a more volatile learning process as firms react to abrupt changes in cost conditions, potentially leading to slower convergence or less stable policy optimization.

4.1.2 Average Step and Episode Rewards

Examining average step rewards from the critic across all cost functions reveals convergence over the course of training. Specifically, the convex cost function stabilizes at step rewards, indicating that firms have learned to maintain profitable production levels below critical thresholds. Conversely, the linear cost function consistently yields net zero earnings across all production levels due to a unitary taxation mechanism, implying a persistent loss scenario exacerbated by fixed

costs. The concave cost function, while initially promising higher earnings at high production levels, faces instability due to the cascading impact of cost sharing, where one firm's reduced output disproportionately burdens the other.

4.1.3 Policy Implications and Market Dynamics

The linear cost function emerges as the most straightforward and predictable incentive structure, facilitating clear policy optimization pathways. Its proportional impact on firms' net earnings simplifies decision-making processes and encourages stable learning dynamics. Policymakers may find merit in adopting such a scheme for its transparency and potential for rapid policy convergence, contingent upon accurate cost estimations prior to penalty application. On the other hand, despite requiring extended training periods to achieve optimal production-cost balance, the convex cost function incentivizes cooperative behavior among firms to maintain production below harmful levels. This mechanism supports environmental goals by curbing negative externalities associated with excessive production. In contrast, the concave cost function demonstrates instability in policy optimization, particularly under competitive market conditions with few firms. Its sensitivity to cost fluctuations and the resulting variability in policy adjustments highlight challenges in achieving stable market outcomes. Moreover, the concave cost structure may inadvertently promote market volatility by intensifying competition among firms.

In conclusion, the choice of cost function significantly influences the behavior and performance of firms in a competitive market environment under MARL. While linear costs offer clarity and efficiency in policy optimization, convex costs promote cooperative behavior toward environmental sustainability. Conversely, concave costs present challenges in achieving stable market conditions due to their sensitivity to cost dynamics and potential for exacerbating competitive pressures. These findings underscore the nuanced interplay between economic incentives and market outcomes, providing valuable insights for policymakers and stakeholders seeking to design effective regulatory frameworks and promote sustainable market practices.

4.2 Economy Attributes

In this experiment, the degree of influence from variations in environmental characteristics, specifically the nature of the prices, on the convergence of outcomes is explored. The results, depicted in Figure 2, highlight several instances of convergence dynamics. The comparison of heterogeneous and homogeneous agents is made to distinguish the degree of impact of changing prices on the firms, and the introduction of production quotas is there to examine how imposing barriers against firms taking no action mitigates the tendency towards unfavorable Nash equilibria.

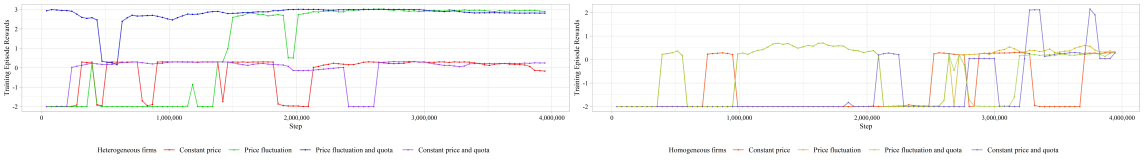

Figure 2: *The results of HARL models with either price fluctuation or constant price.*

4.2.1 Firms Differentiation

The role of firm homogeneity or heterogeneity proves pivotal in determining the ability of agents to converge towards optimal policies within the two environments in question. A surprising result from this experiment is that heterogenous agents, when engaged in the experiment, exhibit a notable tendency to stabilize around a Nash equilibrium, where rewards hover around a total of 0 or 2 units (excluding profits). Across 700,000 initial steps, amidst price fluctuations, these agents demonstrate a capacity for survival and minimal emissions. This phase allows them ample opportunity to learn optimal production strategies aimed at maximizing profits while collaboratively sharing production responsibilities to minimize costs. By the final time step, the average discrepancy in production output between agents within an episode approximates 4,500 thousand barrels, representing a slight deviation from the ideal maximum output of 31,500 thousand barrels.

In scenarios of constant pricing, however, heterogeneous firms face constraints, earning only the minimum unit price of 1 due to price floors. This limits their earnings compared to more dynamic market conditions where prices are buoyed by constrained supply. Nonetheless, these firms manage to sustain operations by learning cooperative production strategies aimed at survival and profitability. The inherent requirement to align production closely with ideal levels, particularly within the default serial cost-sharing environment, reinforces behaviors observed in literature on serial cost-sharing mechanisms.

Conversely, homogeneous agents exhibit marked differences in decision-making processes despite the same firm characteristics. For a reason, their similarities lead to distinct policy update patterns, resulting in divergent production levels across instances. This variability contributes to a prolonged learning phase where firms struggle to consistently achieve profitable outcomes within the environment. Surprisingly, in contrast to heterogeneous counterparts, a constant-price environment facilitates homogeneous firms' ability to attain desired production ranges conducive to profitability and environmental sustainability. This context allows them to effectively manage quotas, leading to their advancement into higher reward bands after extensive training periods, notably by 3,000,000 time-steps.

4.2.2 Price Fluctuation

The influence of market dynamics, whether characterized by dynamic or price-stable conditions, manifests differently depending on firm characteristics. Heterogeneous firms leverage market dynamics favorably, capitalizing on profit increases resulting from heightened demand and subsequent price spikes. In contrast, an inelastic market environment with stable price regimes offers agents a structured and predictable economic landscape. Here, agents adeptly navigate stable cost and reward mechanisms, effectively aligning individual and collective objectives regardless of any possible initial disparities in firm characteristics.

In the situation where high production is heavily punished and hence demand always exceeds supply, when price is constant, agents only need to find a high enough optimal production output that allows it to earn profits while not exceeding the ideal production set out by the policymakers/environment. However, when price is dependent on both supply and demand, the agents need to utilize their forecasted demand level and the possible occurring events to predict the demand before determining the output. Therefore, from the obtained output of the monopoly in a price-fluctuating environment, the firm either was not able to learn the optimal policy for this or it could not handle the change in demand of the entire country alone to consistently earn profits.

In summary, the differentiation between heterogeneous and homogeneous firms significantly shapes their adaptation and performance within the studied environments. Heterogeneous firms demonstrate adaptability across variable pricing scenarios, leveraging market dynamics to optimize profitability and operational efficiency. Conversely, homogeneous firms face challenges in aligning policy updates and production decisions, leading to extended learning phases and variable performance outcomes. Understanding these dynamics provides valuable insights into optimizing economic strategies and policy frameworks conducive to sustainable market practices and efficient resource allocation.

4.3 Number of Firms

The results presented here in Figure 3 shed light on how market structure, characterized by the density and diversity of competing and collaborating firms, shapes the strategic decisions and collective performance of firms within the studied environment.

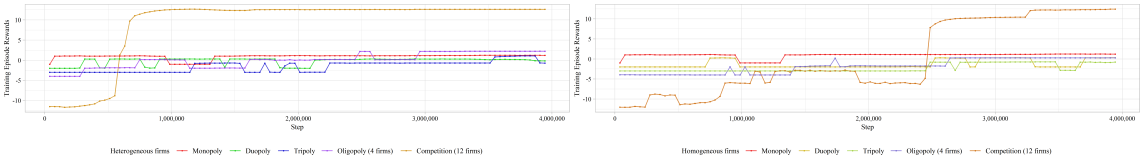

Figure 3: *The results of HARL models with different numbers of firms in the market (the data for the 4-firm and 12-firm scenarios are incomplete due to limited computing resources and time for training).*

The plots of the rewards imply the existence of multiple cases of prisoner's dilemma, as firms tend converge to certain rewards levels, with the lowest level being the most prominent. Notably, this tendency is particularly pronounced in the the duopoly scenarios which are the default choice for most of the scenarios in the previous experiments.

As the number of firms increases, a trend emerges where firms tend to converge towards similar production outputs. For instance, in scenarios involving four firms, each firm produces within a range of 5,000-7,000 thousand barrels. Similarly, with twelve firms, production levels hover above 2,000 thousand barrels per firm, closely approximating the optimal total production target. This convergence is notable given the complexity introduced by higher agent density and extended training durations.

Mirroring previous findings, heterogeneous firms demonstrate a notable advantage in synchronizing their production levels compared to homogeneous counterparts. Even with varied characteristics influencing policy updates and gradient norms, heterogeneous firms exhibit early synchronization in production outputs. In the context of twelve fifrms, diverse entities achieve the highest reward band in under 700,000 training steps, whereas homogeneous firms achieve this milestone only by the 2,500,000th step.

The benefit of increasing the number of firms in the environment, especially evident among heterogeneous groups, underscores the dynamism of competitive, atomistic markets. In such settings, no single firm possesses sufficient influence to significantly sway market conditions. Instead, firms focus on optimizing production based on their individual capabilities, contributing to a balanced and efficient market equilibrium. This decentralized approach fosters a more balanced and efficient market equilibrium, where firms continually adjust their strategies in response to prevailing market signals and competitive pressures.

Consequently, varying the number of firms within a market environment unveils critical insights into competitive dynamics and strategic interactions among agents. The observed convergence patterns and production synchronization underscore the influence of market structure on economic behaviors. This empirical validation of market dynamics provides invaluable insights for policymakers aiming to foster robust competition and sustainable economic growth within complex market ecosystems.

5 Conclusion and Discussion

5.1 Summary of Findings

The intricate interplay between serial cost-sharing mechanism and market dynamics reveals that although this sharing process stimulate firms to reduce and adjust their production levels to similar amounts, designing effective regulatory frameworks for emissions control necessitates a nuanced approach. The findings from this research suggest that the effectiveness of serial cost-sharing taxation policy depends on several factors, hence, the optimal strategy to implement cost-sharing mechanisms to regulate emission levels should incorporate the following key principles:

5.1.1 Type of Cost Function

The nature of the total and external costs of a product carry diverse effect on firms' production. Linear cost functions, characterized by their simplicity and proportionality, offer significant advantages in terms of clarity and predictability. These functions provide firms with a straightforward incentive structure, where costs increase linearly with production levels. This transparency aids firms in optimizing their production strategies efficiently, resulting in stable market conditions and consistent policy convergence. However, while linear costs are beneficial for ease of implementation and compliance, they may not sufficiently incentivize reductions in emissions beyond a certain point, as firms might only reduce emissions to the level where marginal cost equals marginal benefit.

Conversely, convex cost functions are particularly effective in promoting cooperative behavior among firms. These functions escalate costs more rapidly as production increases, creating strong incentives for firms to moderate their output to avoid disproportionately high costs. This dynamic can be particularly effective in contexts where collective action is required to mitigate negative externalities such as emissions. By encouraging firms to operate below certain thresholds, convex costs support sustainable production practices and environmental stewardship.

However, when the external cost of the product is modeled using concave cost functions, policymaking should be approached with caution. While they can theoretically encourage high levels of production efficiency by imposing steep penalties at lower production levels, their inherent sensitivity can lead to volatility in market behavior. Firms may experience significant fluctuations in costs, leading to unstable market conditions and challenging policy adherence. Therefore, concave cost functions might be more suitable in highly regulated environments where firms can reliably predict and manage their production levels.

5.1.2 Firm Heterogeneity

The heterogeneity of firms plays a crucial role in the effectiveness of cost-sharing mechanisms. Heterogeneous firms, with their varied capacities and strategic approaches, exhibit greater adaptability to dynamic market conditions. This adaptability is crucial in environments with fluctuating prices and complex cost structures. Heterogeneous firms can leverage their diverse attributes to optimize production and emissions reductions more effectively than homogeneous firms.

5.1.3 Market Structure and Competition

The number of firms within the market significantly impacts the effectiveness of cost-sharing mechanisms. As the market's density increases, its behavior tends to converge towards more balanced and efficient outcomes. A competitive, atomistic market structure, where no single firm has the power to significantly influence market conditions, promotes optimal production and emissions efficiency. In such environments, firms focus on individual production based on their capabilities, leading to an equitable distribution of resources and enhanced market efficiency.

5.2 Implications

This research not only advances machine learning techniques in economic research but also provides critical insights for designing effective regulatory mechanisms and understanding complex competitive market dynamics, with the focus on serial cost-sharing as the fundamental component.

5.2.1 Advanced Economic and Policy Analysis

Utilizing a novel heterogeneous-agent reinforcement learning (HARL) framework represents a significant leap in applying advanced machine learning techniques to economic and policy analysis. By accommodating diverse agent characteristics, observation spaces, action spaces, and roles, HARL surpasses traditional MARL algorithms, enabling more accurate simulations of complex market dynamics. This approach enhances the modeling of heterogeneous agent interactions, reflecting real-world variations in firm behavior, capabilities, and strategies. The research showcases HARL's capability to uncover intricate economic interdependencies and emergent phenomena, offering deeper insights into regulatory policy effects and market structure dynamics. Policymakers can leverage HARL to design more responsive and effective regulatory frameworks, revolutionizing economic research with a nuanced understanding of market behaviors and policy impacts.

5.2.2 Emission Control Mechanism Design

The study underscores the necessity of tailoring emission control mechanisms to firms' characteristics and market conditions. Linear cost functions, straightforward in nature, are suitable for immediate compliance scenarios but may require supplementation with other measures for deeper emission cuts. Convex cost functions prove effective in fostering cooperation and reducing emissions collectively, ideal for industries requiring collaborative efforts. Meanwhile, challenges associated with concave cost functions highlight the need for cautious application in stable, regulated environments or exploration of hybrid models combining convex and concave elements for balanced incentives and market stability.

5.2.3 Effective Cost Allocation Policy Design

Firm heterogeneity is pivotal in optimizing cost allocation policies. The adaptability of heterogeneous firms in dynamic pricing environments emphasizes the importance of policies accommodating diverse firm capabilities and strategies. Supporting SMEs and fostering innovation can enhance the efficacy of cost-sharing mechanisms. Insights from varying numbers of firms stress the significance of competitive market structures in preventing monopolistic behaviors and promoting equitable cost distribution. Policymakers should prioritize creating competitive conditions through antitrust measures and incentives for market entry to foster sustainable economic growth.

5.3 Limitations and Recommendations for Future Research

5.3.1 Simplified Modeling of Economic Factors

Despite the effort to fully capture the most fundamental aspects of the emissions and demand, there still exist far more contributors to these areas. Oil companies are responsible for higher emissions than what is proposed. For instance, when indirect Scope 3 emissions are taken into account, they significantly complicate the accuracy and realism of emission calculations (Hertwich & Wood, 2018). Stemming from sources beyond a company's direct control like customer product usage, Scope 3 emissions present a significant challenge in precisely assessing environmental impact. Simplifying emissions modeling in this study may lead to underestimations of the genuine environmental costs linked to corporate operations, potentially constraining the depth of policy suggestions regarding emission regulation and carbon pricing strategies.

5.3.2 Environmental Reward Design

The design choice to heavily incentivize firms with a global reward upon reaching the final time-step without exceeding emissions limits may distort the behavioral incentives of agents. While intended to promote environmental stewardship, this approach could artificially shift firms' priorities away from profit maximization towards emissions minimization. Such a scenario may not accurately reflect real-world market dynamics, where firms balance economic objectives with regulatory compliance and environmental responsibilities. Future research should carefully consider the trade-offs and unintended consequences of reward structures to ensure they align with realistic market behaviors and policy objectives.

5.3.3 Cost-Sharing Mechanism Options

While this research exclusively employs the serial cost-sharing mechanism to allocate costs among stakeholders, it is imperative to recognize the limitations of relying on a singular model. Serial cost sharing, although effective in ensuring each agent pays between their stand-alone and unanimous costs, may fail to capture the intricate and strategic behaviors present in diverse or dynamic markets. Future research should investigate alternative mechanisms, such as average cost sharing and marginal cost sharing, which offer unique benefits like resistance to manipulation and alignment with marginal cost variations. Exploring these methods can yield a more nuanced understanding of cost allocation dynamics, particularly in markets with heterogeneous demand structures and production costs. This broader analysis is vital for developing more equitable and efficient cost-sharing frameworks, tailored to specific market conditions and policy objectives.

5.3.4 Algorithmic Hyperparameter Choices

The sensitivity of MARL algorithms in general and HARL algorithms in particular to hyperparameters, such as learning rates and exploration-exploitation trade-offs, poses a significant challenge in achieving robust and reliable results. The initial high learning rates used in this study may have expedited early convergence to deterministic policies but could limit the model's ability to adapt to dynamic market conditions over extended training periods.

Moreover, the default configuration and limited fine-tuning of neural network architectures may not optimize model performance for specific economic contexts or agent behaviors. The adoption of a four-layer, 128-node MLP neural network architecture may not fully capture the intricacies of firm behaviors and market responses. The lack of customized network architecture tuning for specific economic environments and agent characteristics limits the fidelity and accuracy of simulation outcomes.

5.3.5 Risk of Catastrophic Interference

The phenomenon of catastrophic interference, where new learning disrupts previously acquired knowledge, poses a risk in long-term policy simulations. In HARL applications, this risk is exacerbated by the complexity of economic environments and the continuous adaptation of agents' policies. The observed convergence to suboptimal strategies or stalemates in certain scenarios may indicate instances of catastrophic interference, where agents struggle to balance exploration and exploitation effectively. Mitigating this risk requires ongoing refinement of learning algorithms, regularization techniques, and neural network architectures to preserve learning stability and enhance policy relevance in dynamic economic simulations.

5.3.6 Model Complexity and Resource Constraints

Furthermore, the computational demands and resource constraints inherent in HARL simulations impose practical limitations on the scale and scope of economic models. The use of simplified neural network architectures and constrained learning resources may compromise the model's ability to capture the full complexity of market dynamics and policy interactions accurately. Future research should explore advanced computational strategies, such as distributed computing and parallel processing, to enhance model scalability and robustness across diverse economic scenarios.

5.3.7 Computational Complexity and Scalability

Implementing HARL algorithms in realistic economic scenarios involves significant computational complexity and resource requirements. The scalability of the model may be limited when simulating larger markets or integrating a greater number of heterogeneous agents with diverse observation and action spaces. Addressing computational constraints and optimizing algorithm efficiency are crucial for extending the applicability of HARL to broader economic analyses and policy simulations.

6 References

Aadland, D., & Kolpin, V. (2004). Environmental determinants of cost sharing. *Journal of Economic Behavior & Organization*, *53*(4), 495–511.

Agreement, P. (2015). Paris agreement. *report of the conference of the parties to the United Nations framework convention on climate change (21st session, 2015: Paris). Retrived December*, *4*(2017), 2.

Albrecht, S. V., Christianos, F., & Schäfer, L. (2024). *Multi-agent reinforcement learning: Foundations and modern approaches*. MIT Press.

Atashbar, T., & Shi, R. A. (2023). *Ai and macroeconomic modeling: Deep reinforcement learning in an rbc model*. International Monetary Fund.

Aumann, R. J., & Shapley, L. S. (1975). *Values of non-atomic games*. Princeton University Press.

Banerjee, D., & Sen, S. (2007). Reaching pareto-optimality in prisoner's dilemma using conditional joint action learning. *Autonomous Agents and Multi-Agent Systems*, *15*, 91–108.

Billera, L. J., & Heath, D. C. (1982). Allocation of shared costs: A set of axioms yielding a unique procedure. *Mathematics of Operations Research*, *7*(1), 32–39.

Billera, L. J., Heath, D. C., & Raanan, J. (1978). Internal telephone billing rates—a novel application of non-atomic game theory. *Operations Research*, *26*(6), 956–965.

Box, G. E., & Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American statistical Association*, *65*(332), 1509–1526.

Charpentier, A., Elie, R., & Remlinger, C. (2021). Reinforcement learning in economics and finance. *Computational Economics*, 1–38.

Coase, R. H. (1960). The problem of social cost. *Journal of Law and Economics*, *3*(1), 2.

Dahlman, C. J. (1979). The problem of externality. *The journal of law and economics*, *22*(1), 141–162.

Duan, Y., Chen, X., Houthooft, R., Schulman, J., & Abbeel, P. (2016). Benchmarking deep reinforcement learning for continuous control. *International conference on machine learning*, 1329–1338.

Eggleston, H., Buendia, L., Miwa, K., Ngara, T., & Tanabe, K. (2006). 2006 ipcc guidelines for national greenhouse gas inventories.

EPA. (2023). Greenhouse gas emissions from a typical passenger vehicle [Retrieved June 05 2024 at https://www.epa.gov/greenvehicles/greenhouse-gas-emissions-typical-passenger-vehicle].

Espey, M. (1996). Explaining the variation in elasticity estimates of gasoline demand in the united states: A meta-analysis. *The Energy Journal*, *17*(3), 49–60.

Forster, P. M., Smith, C. J., Walsh, T., Lamb, W. F., Lamboll, R., Hauser, M., Ribes, A., Rosen, D., Gillett, N., Palmer, M. D., et al. (2023). Indicators of global climate change 2022: Annual update of large-scale indicators of the state of the climate system and human influence. *Earth System Science Data*, *15*(6), 2295–2327.

Gu, S., Lillicrap, T., Ghahramani, Z., Turner, R. E., & Levine, S. (2016). Q-prop: Sample-efficient policy gradient with an off-policy critic. *arXiv preprint arXiv:1611.02247*.

Hertwich, E. G., & Wood, R. (2018). The growing importance of scope 3 greenhouse gas emissions from industry. *Environmental Research Letters*, *13*(10), 104013.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts.

Institute, W. R. (2022). Climate watch historical country greenhouse gas emissions data [Retrieved June 05, 2024 at https://www.climatewatchdata.org/ghg-emissions].

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research*, *4*, 237–285.

Koster, M., Tijs, S., & Borm, P. (1998). Serial cost sharing methods for multi-commodity situations. *Mathematical Social Sciences*, *36*(3), 229–242.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994* (pp. 157–163). Elsevier.

Luqman, M., Peng, S., Huang, S., Bibi, A., & Najid, A. (2018). Cost allocation for the problem of pollution reduction: A dynamic cooperative game approach. *Economic research-Ekonomska istraživanja*, *31*(1), 1717–1736.

Marshall, A. (1890). *Principles of economics: Unabridged eighth edition*. Cosimo, Inc.

Meinshausen, M., Meinshausen, N., Hare, W., Raper, S. C., Frieler, K., Knutti, R., Frame, D. J., & Allen, M. R. (2009). Greenhouse-gas emission targets for limiting global warming to 2 c. *Nature*, *458*(7242), 1158–1162.

Mosavi, A., Faghan, Y., Ghamisi, P., Duan, P., Ardabili, S. F., Salwana, E., & Band, S. S. (2020). Comprehensive review of deep reinforcement learning methods and applications in economics. *Mathematics*, *8*(10), 1640.

Moulin, H. (1999). Incremental cost sharing: Characterization by coalition strategy-proofness. *Social Choice and Welfare*, *16*(2), 279–320.

Moulin, H. (2002). Axiomatic cost and surplus sharing. *Handbook of social choice and welfare*, *1*, 289–357.

Moulin, H., & Shenker, S. (1992). Serial cost sharing. *Econometrica: Journal of the Econometric Society*, 1009–1037.

Narassimhan, E., Gallagher, K. S., Koester, S., & Alejo, J. R. (2018). Carbon pricing in practice: A review of existing emissions trading systems. *Climate Policy*, *18*(8), 967–991.

NHTSA et al. (2010). Light-duty vehicle greenhouse gas emission standards and corporate average fuel economy standards; final rule. *Federal Register*, *40*, 25323–25728.

Pigou, A. (1920). *The economics of welfare*. Routledge.

Princen, T. (2001). Consumption and its externalities: Where economy meets ecology. *Global Environmental Politics*, *1*(3), 11–30.

Rothschild, E. (2001). *Economic sentiments*. harvard university Press.

Safiri, S., Nikoofard, A., Khosravy, M., & Senjyu, T. (2022). Multi-agent distributed reinforcement learning algorithm for free-model economic-environmental power and chp dispatch problems. *IEEE Transactions on Power Systems*, *38*(5), 4489–4500.

Sandmo, A. (2015). The early history of environmental economics. *Review of Environmental Economics and Policy*.

Schulman, J., Moritz, P., Levine, S., Jordan, M., & Abbeel, P. (2015). High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shapley, L. S. (1953). Stochastic games. *Proceedings of the national academy of sciences*, *39*(10), 1095–1100.

Shapley, L. S., et al. (1953). A value for n-person games.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Terry, J., Black, B., Grammel, N., Jayakumar, M., Hari, A., Sullivan, R., Santos, L. S., Dieffendahl, C., Horsch, C., Perez-Vicente, R., et al. (2021). Pettingzoo: Gym for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, *34*, 15032–15043.

U.S. Energy Information Administration, E. (2024). Petroleum & other liquids: U.s. weekly product supplied [Retrieved June 01, 2024 from EIA at https://www.eia.gov/dnav/pet/pet_cons_wpsup_k_w.htm].

Van Hasselt, H. P., Guez, A., Hessel, M., Mnih, V., & Silver, D. (2016). Learning values across many orders of magnitude. *Advances in neural information processing systems*, *29*.

Yang, Y., & Wang, J. (2020). An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint arXiv:2011.00583*.

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., & Wu, Y. (2022). The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, *35*, 24611–24624.

Zheng, S., Trott, A., Srinivasa, S., Naik, N., Gruesbeck, M., Parkes, D. C., & Socher, R. (2020). The ai economist: Improving equality and productivity with ai-driven tax policies. *arXiv preprint arXiv:2004.13332*.

Zhong, Y., Kuba, J. G., Feng, X., Hu, S., Ji, J., & Yang, Y. (2024). Heterogeneous-agent reinforcement learning. *Journal of Machine Learning Research*, *25*(1-67), 1.

A Global Greenhouse Gas Emissions Cost Model

The total global cost of greenhouse gases (GHG), particularly carbon dioxide (CO_2), is modeled by considering both the emissions resulting from the consumption of petroleum and the approximate emissions from the operations of oil companies. Given that motor gasoline accounts for nearly half of the consumption of petroleum products (U.S. Energy Information Administration, 2024), it is used as the primary commodity produced by firms in this model .

According to the United Nations Intergovernmental Panel on Climate Change (IPCC) and the U.S. National Highway Traffic Safety Administration (NHTSA) and Environmental Protection Agency (EPA), a common conversion factor of 8,887 grams of CO_2 emissions per gallon of gasoline consumed is used (Eggleston et al., 2006; EPA, 2023; NHTSA et al., 2010). This conversion factor is assumed to be linear, meaning that the emissions are directly proportional to the amount of gasoline consumed.

In the global context, the United States is the second-largest emitter of CO_2 following China (Institute, 2022). Over the past 24 years, the U.S. has contributed approximately 11% of global emissions . Notably, CO_2 emissions from motor gasoline consumption alone accounted for around 22% of the U.S. total CO_2 emissions. Incorporating all these changes into consideration, the results of the ARIMA model used to forecast the global emissions and the emissions of the gasoline market in the United States are shown in Figures 4 and 5.

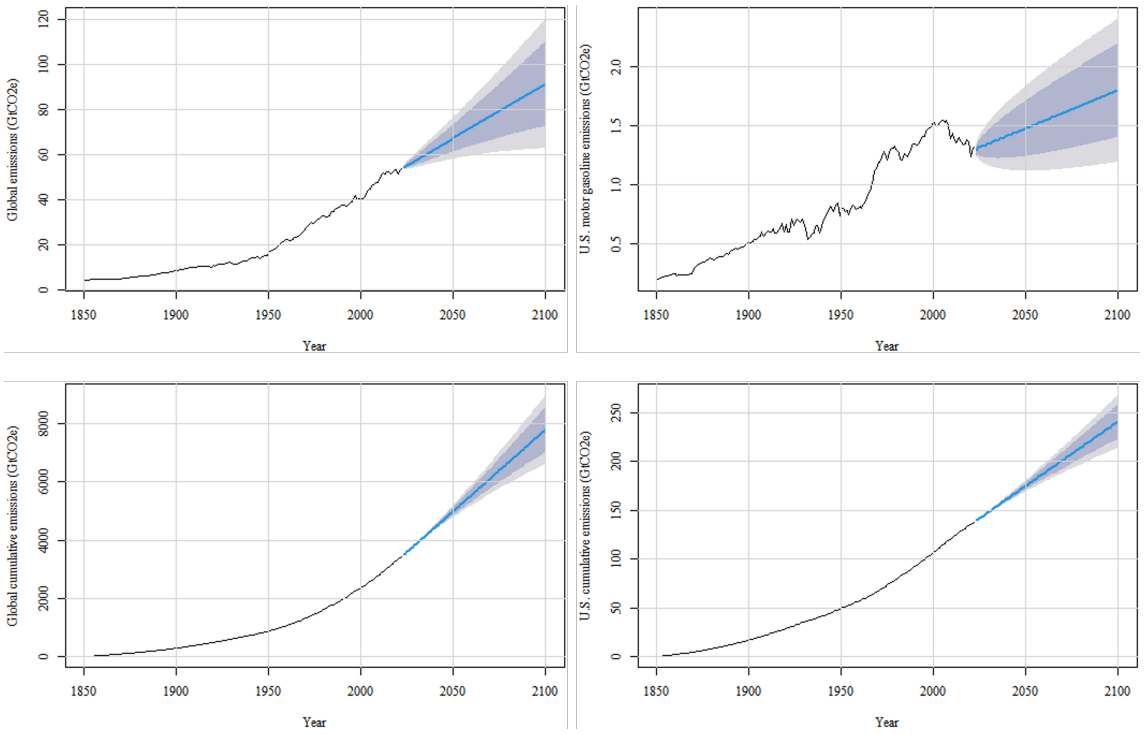

Figure 4: *The emissions of gasoline production in the global landscape (left) and in the U.S. (right).*

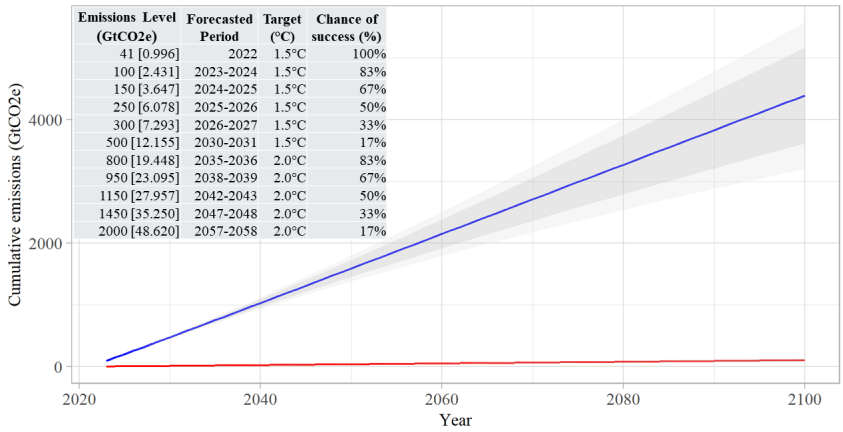

Figure 5: *The cumulative GHG emissions forecast relative to the pre-industrial levels of the world and the U.S. and estimates of the chance of preventing global climate and adhering to the Paris Agreement.*

B United States Motor Gasoline Supply-Demand Model

On the demand side, a time series model based on real-life data forecasts the movement along the demand curve. The demand curve is further adjusted by stochastic demand shocks, simulating real-world fluctuations in consumer behavior. The details of the forecasting model of the demand is illustrated in Appendix F.

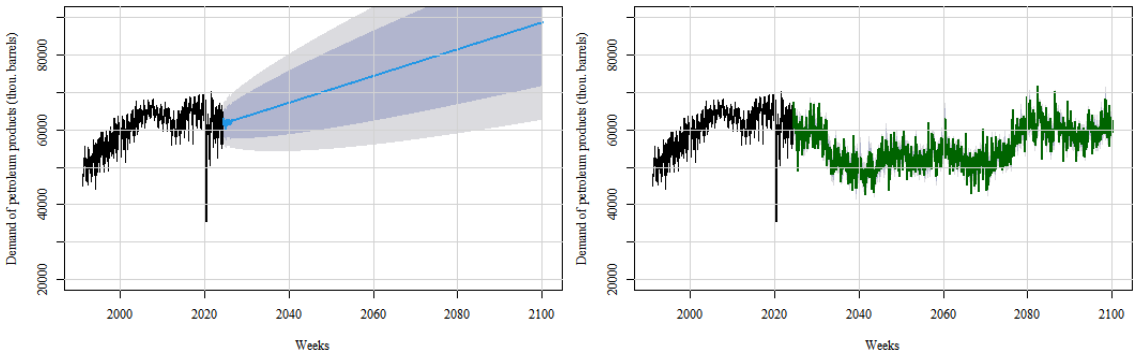

Figure 6: *The initial forecasts of the time series model and the instances of the forecast of motor petroleum demand derived from the final hybrid model.*

Then, the supply-demand model of the U.S. motor gasoline market used in this thesis is derived from the meta analysis of Espey (1996) on the differences in elasticity of the gasoline demand in the U.S., providing a robust framework to analyze the interplay between supply, demand, and pricing under various economic conditions. Initially, the price is set to a unit price of 1, serving as a baseline influenced by subsequent movements in supply and demand. The supply model incorporates the total inventory of gasoline barrels and additional weekly production, dynamically adjusting the supply curve based on real-time inventory levels. These design steps are illustrated in Figure 7.

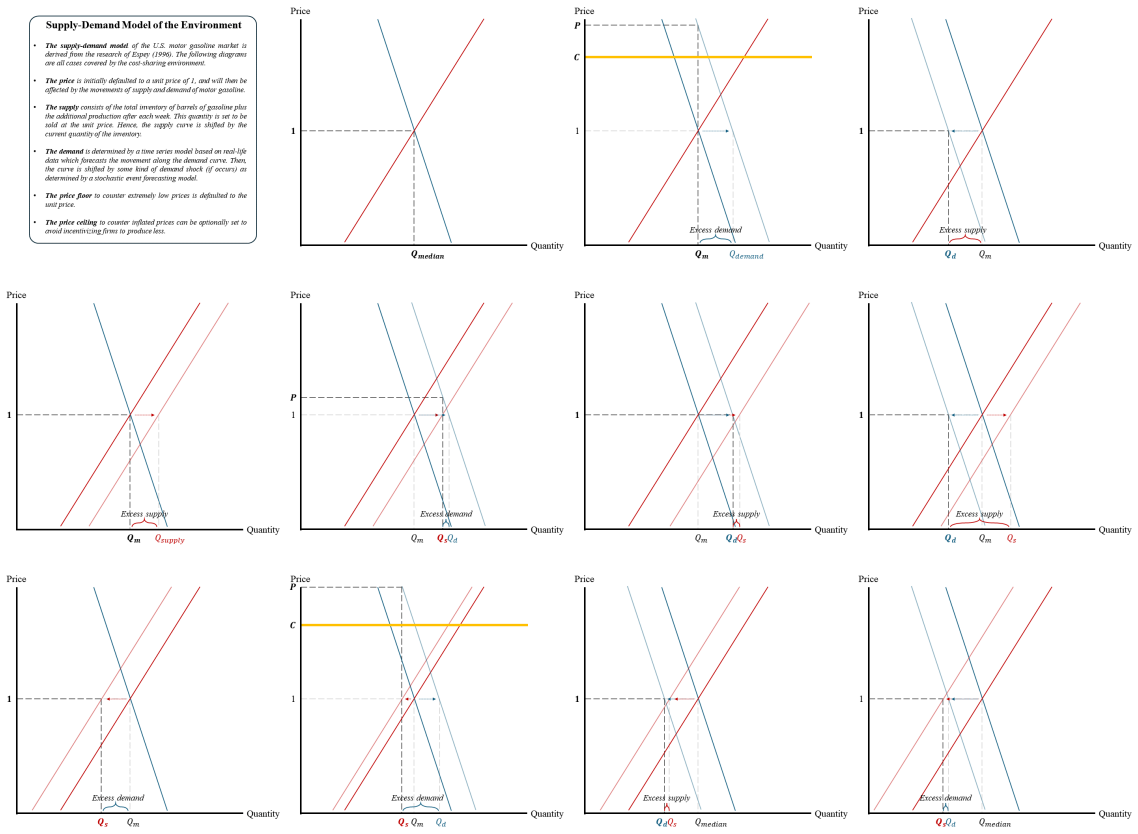

Figure 7: *All possible cases of demand and supply that is covered by the model and their resulting price in the market of motor gasoline.*

To prevent extreme price volatility, the model includes mechanisms for both a price floor, set at the unit price to counteract destabilizing low prices, and an optional price ceiling to prevent excessively high prices that could disincentivize production. This comprehensive approach ensures a balanced and dynamic understanding of market behavior, crucial for policymakers aiming to manage environmentally harmful production while maintaining economic viability. By incorporating real-time data and stochastic elements, this model effectively captures the complexities of the U.S. motor gasoline market, facilitating informed decision-making to achieve sustainable economic and environmental goals.

C Cost-Sharing Mechanisms

Cost-sharing mechanisms are fundamental to the equitable distribution of costs among participants in collaborative manner, particularly in addressing environmental externality mitigation efforts. This appendix delves into three prominent cost-sharing methods: average cost sharing, marginal cost sharing, and serial cost sharing. These methodologies are essential in understanding how costs can be equitably distributed among agents with differing demands for a product or service. Each of these cost-sharing mechanisms offers unique advantages and potential drawbacks. Average cost sharing provides simplicity and resistance to strategic behavior, marginal cost sharing aligns payments with incremental costs, while serial cost sharing ensures fairness in dynamic allocation settings.

C.1 Average Cost Sharing

Average Cost Sharing (ACS) is a method that distributes the total cost of a project or service evenly among all participants. The primary principle behind ACS is fairness, ensuring that each

participant bears an equal share of the cost, regardless of their individual contribution or benefit derived from the project. In average cost sharing, the total cost of producing a given quantity $q_N = \sum_{i=1}^{n} q_i$ of a product is distributed among the agents based on their respective demands. The cost $\xi_i(C, q)$ distributed to each agent i is calculated as:

$$\xi_i(C, q) = \frac{C(q_N) \cdot q_i}{q_N} \tag{2}$$

This approach ensures that each agent contributes proportionally to their demand. One significant advantage of average cost sharing is its robustness against strategic manipulations, such as agents merging their demands or splitting them into multiple sub-agents. Additionally, this method ensures that each agent pays at least their stand-alone cost, the cost they would incur if they were the only agent being served. However, a potential drawback is that an agent may end up paying more than their unanimous cost, the cost they would pay if all other agents had identical demands.

C.2 Marginal Cost Sharing

Marginal Cost Sharing (MCS) focuses on allocating costs based on the incremental cost of serving each additional unit demanded by an agent to be imposed on each participant. This method aligns the cost burden with the incremental impact each participant has on the total cost. The individual cost $\xi_i(C, q)$ is given by:

$$\xi_i(C, q) = q_i \cdot C'(q_N) + \frac{1}{n}\left(C(q_N) - q_N \cdot C'(q_N)\right) \tag{3}$$

This method ensures that no agent pays more than their unanimous cost, promoting fairness in the cost distribution. However, an agent might pay less than their stand-alone cost, which could lead to situations where some agents are effectively subsidized by others. This scenario can occur when the marginal cost structure results in a lower incremental cost for smaller demands.

C.3 Serial Cost Sharing

Serial cost sharing is a dynamic process that distributes costs based on the sequence of agents exiting a collective arrangement once their demands are met. For instance, consider a scenario where multiple firms produce motor gasoline, and a regulatory body imposes a tax on the total production. A representative procedure of the algorithm can be outlined as follows:

1. All firms are initially subject to the taxation of their production of gasoline barrels.
2. The production starts, and the tax on each additional barrel, or n additional barrels where each firm produces 1 barrel, is equally shared among all firms.
3. When this week's production requirement of a firm is satisfied, it can exit the production chain and the cost-distributive policy does not apply to it.
4. The other agents remain to share the costs of subsequent output until they decide to exit the production chain.

Formally, if agents are ordered by ascending production output, agent i needs to pay (Moulin & Shenker, 1992):

$$\xi_i(C, q)(C, q) = \frac{C(q^i)}{n - i + 1} - \sum_{k=1}^{i-1} \frac{C(q^k)}{(n - k + 1)(n - k)} \tag{4}$$

where q^n is the new quantity in which the cost will be taxed on after the agents $(1, ..., n$ have

exited the production chain.

$$
\begin{aligned}
q^0 &= 0 \\
q^1 &= nq_1 \\
q^2 &= q_1 + (n-1)q2 \\
&... \\
q^i &= q_1 + ... + q_{i-1} + (n+1-i)q_i \\
&... \\
q^n &= \sum_i q_i
\end{aligned} \tag{5}
$$

This method guarantees that each agent pays at least their stand-alone cost and at most their unanimous cost. However, serial cost sharing is susceptible to manipulations such as splitting or merging demands, and transferring costs and products among agents. Therefore, it is most effective in scenarios where such manipulations are impractical or impossible, for instance, in the case of cost allocation of telephone services (Billera et al., 1978).

D Reinforcement Learning

This appendix provides a comprehensive overview of reinforcement learning (RL) with a focus on the development of RL as the number of agents increases. It starts by discussing the fundamental concepts and mathematical foundations of single-agent RL, progresses to the complexities of multi-agent RL, introduces the emerging heterogeneous-agent RL framework, and concludes with the advancements brought by deep learning techniques.

D.1 Single-Agent Reinforcement Learning

Single-agent reinforcement learning (SARL), commonly known as reinforcement learning (RL), focuses on how an individual agent can learn to make decisions in an environment to maximize cumulative rewards (Kaelbling et al., 1996; Sutton & Barto, 2018; Mosavi et al., 2020) (). In this framework, the two key components are the agent and the environment. The agent's primary objective is to achieve a specific goal by interacting with its environment. This interaction involves taking actions and receiving feedback in the form of rewards, which helps the agent learn the optimal strategy for reaching its goal. At each time-step, the agent selects an action based on its current state and receives a reward from the environment, along with an updated state. This process continues until a terminating condition or the final time-step is reached.

Hence, as the essence of reinforcement learning lies in its ability to capture the dynamic and temporal nature of real-world problems, allowing the agent to adapt its strategy over time to optimize performance, reinforcement learning is deeply rooted in the theories of dynamical systems and optimization. It is particularly effective in dealing with the optimal control of unknown or partially observable Markov decision processes (POMDPs). By leveraging these mathematical foundations, SARL provides a robust framework for solving complex decision-making problems where the environment's dynamics are not fully known or are subject to change. Through iterative learning and adaptation, the agent gradually improves its policy, which is the mapping from states to actions, to achieve the highest possible cumulative reward in the given environment.

D.1.1 Markov Decision Process

A Markov Decision Process (MDP) serves as the mathematical foundation for SARL. An MDP provides a formal framework for modeling decision-making problems where outcomes are partly random and partly under the control of the decision-maker. Based on the notations from Sutton and Barto (2018), this process is defined by the components below.

- **States $\mathcal{S}$.** It is the set of all possible condition or state the environment can be in. The random variable $S = s \in \mathcal{S}$ represents the stochastic process of reaching a state s in the environment. Each state provides the agent with information about the current situation.

- **Actions $\mathcal{A}$.** It is the set of all possible moves the agent can take. Every action $A = a$ is the decision made by the agent to interact with the environment. Furthermore, the set of all actions the agent can take in a certain state s is defined as $\mathcal{A}(s)$.
- **Transition probabilities function ($P(s'|s, a)$).** This function defines the probability of transitioning from state s to state s' given action a. This captures the dynamics of the environment.
- **Reward function ($R(s, a) \to \mathbb{R} \in \mathcal{R}$).** This function provides immediate feedback to the agent by assigning a numerical reward for taking action a in state s. This guides the agent towards favorable outcomes.
- **Discount factor (γ).** The discount factor $\gamma \in [0, 1]$ determines the importance of future rewards. A lower discount factor emphasizes immediate rewards, while a higher discount factor places more value on long-term gains.

The process begins with the agent in an initial state s_0, drawn from a predefined distribution. At each time-step t, the agent perceives the current state of the environment, denoted as $S_0 \in \mathcal{S}$, and selects an action $A_0 = A(s_0) = a_0$ based on this state. The selection of action a_0 causes the environment to transition to a new state s_{t+1}, determined by the probability transition function $P(s_{t+1}|s_t, a_t)$. Subsequently, the agent receives a numerical reward $r_{t+1} \in \mathcal{R}$ as feedback, reflecting the immediate impact of its action. This sequence of interactions—state, action, new state, reward—creates a trajectory or episode that evolves over time: $s_0, a_0, r_1, s_1, a_1, r_2, s_2, a_2, r_3, \ldots$.

Consequently, an MDP is defined by the tuple $(\mathcal{S}, \mathcal{A}, P, \mathcal{R}, \gamma)$, and through continuously selecting actions and receiving feedback, the agent iteratively refines its strategy to maximize cumulative rewards, navigating through the MDP to achieve optimal decision-making. By using the MDP framework, reinforcement learning models the sequential decision-making process in a structured manner, allowing the agent to learn optimal behaviors through trial and error and feedback from the environment.

D.1.2 Optimal Control

Optimal control theory constitutes a pivotal domain within control theory, specifically aimed at devising a strategic control methodology for dynamic systems to maximize a predefined objective function over a duration. Optimal control is a fundamental concept in reinforcement learning, focusing on determining the best possible actions an agent can take to maximize its cumulative reward over time. This involves solving the MDP to find an optimal policy $\pi : \mathcal{S} \to \mathcal{A}$ that maximizes the total rewards the agent obtained from the environment. For discrete time-steps, with $i_t \in (1, ..., |\mathcal{A}(s)|)$, the total rewards or return G at time-step t is defined as:

$$G_t = R(s_0, a_{i_1}) + \gamma R(s_1, a_{i_2}) + \gamma^2 R(s_3, a_{i_3}) + ... = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{6}$$

At the heart of optimal control in reinforcement learning is the value function, which estimates the expected cumulative reward an agent can obtain starting from a particular state and following a specific policy thereafter. To maximize its total rewards, the agent needs to estimate these two value functions:

1. State value function $v^\pi(s)$ represents the expected return starting from state s and following policy π:

$$v^\pi(s) = \mathbb{E}^\pi \left[G_t \mid S_t = s \right] = \mathbb{E} \left[\sum_{a_t \in \mathcal{A}(s_t)} \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \right] \tag{7}$$

2. Action value function ($q^\pi(s, a)$) represents the expected return starting from state s, taking action a, and following policy π:

$$q^\pi(s, a) = \mathbb{E}^\pi \left[G_t \mid S_t = s, A_t = a \right] = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \right] \tag{8}$$

Then, optimal control is used to find the optimal policy $\pi*$ that maximizes these value functions. The optimal state value function $V^*(s)$ and the optimal action value function $Q^*(s,a)$ satisfy the Bellman optimality equations:

$$\begin{aligned} v^*(s) &= \max_{\pi} v_\pi(s) \qquad (9) \\ &= \max_{a} \mathbb{E}\left[R_{t+1} + \gamma v^*(S_{t+1}) \mid S_t = s, A_t = a\right] \\ Q^*(s,a) &= \max_{\pi} q_\pi(s,a) \qquad (10) \\ &= \mathbb{E}\left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a\right] \\ &= \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} Q^*(S_{t+1}, a') \mid S_t = s, A_t = a\right] \end{aligned}$$

Hence, all values for each pair of state-action s–a can be stored in a value-updating table. These equations form the basis for all elementary reinforcement learning algorithms, such as Q-learning and SARSA, which iteratively update estimates of the value functions to converge towards the optimal policy. However, finding the exact solutions to these equations in practice can be computationally infeasible for large state and action spaces, leading to the use of function approximation methods to generalize value functions across states and actions. Nevertheless, by solving the optimal control problem, reinforcement learning agents can learn to navigate complex environments and make decisions that maximize long-term rewards, effectively balancing the trade-off between exploration and exploitation.

D.1.3 Value Function Approximation Methods

Reinforcement learning (RL) confronts challenges when applied to environments with vast state or action spaces, rendering exact methods like dynamic programming impractical. To surmount this, RL employs function approximation techniques to generalize value functions or policies across states and actions. Rather than iteratively updating tabular values, these methods estimate the state-value function V^π from on-policy data generated under an initially known policy π.

Central to these methods is the representation of v^π through parameterized functional forms denoted as $\hat{v}(s;\boldsymbol{w})$, where $\boldsymbol{w} \in \mathbb{R}^d$ represents a weight vector of lower dimensionality compared to the tabular approach. This parameterization enhances storage efficiency significantly, as it necessitates storing only the weights $\boldsymbol{w}$ rather than storing individual values for each state.

Formally, the state-value function approximation can be expressed as $\hat{V}(s;\boldsymbol{w}) \approx V^\pi(s)$, where $\hat{V}(s;\boldsymbol{w})$ is the estimated value of state s using parameters $\boldsymbol{w}$. The process involves learning $\boldsymbol{w}$ from experience to minimize the error between $\hat{V}(s;\boldsymbol{w})$ and $V^\pi(s)$.

This approach trades off some accuracy for scalability and efficiency in storage and computation. Advanced techniques, such as neural networks, enable the representation of highly complex value functions by learning intricate mappings from states to values. These methods not only broaden RL's applicability to real-world problems but also contribute to ongoing advancements in autonomous decision-making systems.

D.1.4 Exploration-Exploitation Dilemma

Among the most critical challenges in reinforcement learning is the exploration-exploitation trade-off. This dilemma requires an agent to strike a balance between two conflicting objectives: exploiting known actions that yield high rewards and exploring new actions to discover potentially superior strategies. Exploitation leverages past experiences to maximize immediate rewards by selecting actions with the highest known payoff. In contrast, exploration involves trying untested actions to gather information that could improve long-term performance. The core of the problem lies in the fact that relying exclusively on exploitation may cause the agent to miss out on better strategies, while excessive exploration can result in suboptimal performance due to a lack of focus on established high-reward actions.

To navigate this trade-off, agents must continuously alternate between exploration and exploitation, gradually leaning towards actions that consistently demonstrate higher rewards. This dynamic adjustment is particularly crucial in stochastic environments, where each action must be

tried multiple times to accurately estimate its expected reward. The primary methods which aim to resolve this dilemma include algorithms such as epsilon-greedy, which introduces randomness into action selection to ensure a mix of exploration and exploitation, and upper confidence bound (UCB) approaches that quantify the uncertainty of rewards to guide exploration more strategically. Through these advancements, reinforcement learning agents are better equipped to optimize their decision-making processes, thereby enhancing overall performance in complex and dynamic tasks. However, because this problem is intrinsically connected with and dependent on the characteristics of the environment, there is no one universal algorithm that outperforms all others in every environment. As a result, over the past decade, researchers have devised sophisticated techniques to balance exploration and exploitation effectively.

D.2 Multi-Agent Reinforcement Learning

As RL systems evolve, the complexity increases when multiple agents are involved. Multi-agent reinforcement learning (MARL) represents an advancement in reinforcement learning (RL) that addresses the complexities arising from interactions between multiple decision-making agents within a shared environment. To learn more about multi-agent reinforcemnt learning, please find comprehensive overview and explanation on this subject from Yang and Wang (2020) and Albrecht et al. (2024).

Formally, in a multi-agent environment, the interaction among N agents can be described using a decentralized policy approach, where each agent i selects actions $\mathbf{a}_i$ based on its local observations $\mathbf{o}_i$ and potentially shared information:

$$\mathbf{a}_i = \pi_i(\mathbf{o}_i, \mathbf{o}_{-i}), \tag{11}$$

where π_i denotes the policy of agent i, $\mathbf{o}_i$ represents the observations of agent i, and $\mathbf{o}_{-i}$ denotes the observations of other agents.

In MARL scenarios, each agent interacts with the environment and other agents, influencing and being influenced by their actions and policies. This framework can encompass cooperative scenarios, where agents work towards a common objective, as well as competitive scenarios, where agents pursue individual rewards, potentially conflicting with others.

In cooperative scenarios, agents may employ techniques like joint action learning to align their policies towards a shared goal:

$$\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \mathbb{E}_{\pi_i}[\sum_{t=0}^{\infty} \gamma^t r_i(\mathbf{o}_i^t, \mathbf{a}_i^t)], \tag{12}$$

where $\boldsymbol{\theta}$ represents the parameters of all agents' policies, $\mathbf{o}_i^t$ and $\mathbf{a}_i^t$ denote the observations and actions of agent i at time-step t, and r_i is the reward function for agent i.

Conversely, in competitive scenarios, agents may adopt adversarial training strategies where each agent optimizes its own utility function:

$$\max_{\boldsymbol{\theta}_i} \mathbb{E}_{\pi_i}[\sum_{t=0}^{\infty} \gamma^t r_i(\mathbf{o}_i^t, \mathbf{a}_i^t; \boldsymbol{\theta}_i) - \lambda \sum_{j \neq i} \mathbb{E}_{\pi_j}[\sum_{t=0}^{\infty} \gamma^t r_j(\mathbf{o}_j^t, \mathbf{a}_j^t; \boldsymbol{\theta}_j)]], \tag{13}$$

where $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ are the parameters of agents i and j's policies, respectively, and λ balances the trade-off between individual and collective rewards.

The challenges in MARL are multifaceted, including coordination among agents with potentially conflicting goals, dealing with partial observability where agents have limited views of the environment, and managing concurrent learning processes where agents adapt their policies over time. These challenges necessitate sophisticated strategies for effective learning and coordination to achieve optimal or satisfactory outcomes in both cooperative and competitive settings. Hence, some of the most common issues in MARL are outlined below.

D.2.1 Challenges in MARL

Non-stationarity and Learning Dynamics A fundamental characteristic of MARL is its inherent non-stationarity, driven by agents continuously updating their policies based on evolving experiences. This dynamic adaptation creates a moving target scenario where each agent's learning trajectory is intertwined with and influenced by others. As agents adjust their strategies in response to changing policies of their counterparts, the environment's dynamics undergo continuous shifts. This dynamism can lead to cyclic and unstable learning behaviors, complicating the convergence of learning algorithms. Variations in learning rates among agents, influenced by diverse rewards and local observations, further exacerbate non-stationarity, posing significant challenges in achieving stable and effective learning outcomes.

Partial Observability In MARL settings, agents typically operate with partial observability, where each agent has access only to a limited view of the environment. This partial information complicates decision-making processes, as agents must infer unobserved aspects of the environment to formulate effective policies. The challenge lies in developing strategies that robustly account for incomplete information, requiring sophisticated approaches such as belief states or decentralized policies that synthesize local observations with shared information.

Credit Assignment Problem Temporal credit assignment in MARL involves attributing received rewards to individual actions within the context of joint actions taken by multiple agents. This problem is intricately linked to the interdependence among agents, where the impact of each agent's actions on the collective reward must be accurately assessed. Resolving the credit assignment problem efficiently remains a key research focus, necessitating advanced techniques such as credit assignment mechanisms based on counterfactual reasoning or decentralized learning paradigms that facilitate effective learning amidst complex agent interactions.

Optimality and Equilibrium Selection Unlike single-agent RL, achieving optimality in MARL entails evaluating the effectiveness of policies within the context of multiple interacting agents. This task involves selecting equilibria where agents' strategies collectively lead to desirable outcomes. Determining optimal policies requires sophisticated equilibrium concepts and coordination mechanisms to ensure that agents effectively cooperate or compete to achieve desired performance metrics. Achieving coordination among agents while balancing individual and collective goals remains a pivotal challenge in MARL research.

Scaling with Number of Agents The scalability of MARL algorithms poses significant computational and algorithmic challenges as the number of agents increases. With each additional agent, the complexity of managing interactions and coordinating actions grows exponentially. The combinatorial explosion of possible action combinations amplifies computational demands and algorithmic complexity, necessitating scalable approaches that can efficiently handle large-scale multi-agent systems. Addressing scalability challenges requires innovative algorithmic designs, distributed computing frameworks, and parallelization strategies to ensure effective coordination and learning across numerous agents.

D.3 Heterogeneous-Agent Reinforcement Learning (HARL)

Heterogeneous-Agent Reinforcement Learning (HARL) is a more recent extension of MARL, focusing on scenarios where agents have diverse characteristics and capabilities. Unlike homogeneous agents that are often assumed in MARL, HARL considers agents with different characteristics, capabilities, roles, and/or action sets. For in-depth details, please refer to Zhong et al. (2024).

E Reinforcement Learning Algorithms

E.1 Proximal Policy Optimization (PPO)

Algorithm 1: Pseudocode of the PPO algorithm (Schulman et al., 2017)

Input: steps, actors, policy, advantage estimates, total training steps, surrogate, epochs, mini-batch size

Output: optimized surrogate

while $step \leq step_{max}$ **do**
  for actor in $(1, 2, \ldots, N)$ **do**
    Run policy $\pi_{\theta_{old}}$ in environment for T time steps;
    Compute advantage estimates $\hat{A}_1, \ldots, \hat{A}_T$;
  Optimize surrogate L wrt. θ, with K epochs and minibatch size $M \leq NT$;
  $\theta_{old} \leftarrow \theta$;

E.2 Multi-Agent Proximal Policy Optimization (MAPPO)

Algorithm 2: Pseudocode of the MAPPO algorithm (Yu et al., 2022)

Input: stepsize σ, batch size B, n agents, $step_{max}$ total training steps, T steps per episode, learning rate α

Output: optimized surrogate

Initialize:;
- Global parameters $\boldsymbol{\theta}$ for policy π;
- Global V-value network $\{\phi\}$;

while $step \leq step_{max}$ **do**
  Set data buffer $\mathcal{B} = \{\ \}$;
  for i in $(1, 2, ..., B)$ **do**
    $\tau = [\ ]$ (an empty list;
    Initialize $h^{(1)}_{0,\pi}, ..., h^{(n)}_{0,\pi}$ actor RNN states;
    Initialize $h^{(1)}_{0,V}, ..., h^{(n)}_{0,V}$ critic RNN states;
    for t in $(1, 2, ..., T)$ **do**
      for agent a in $(1, 2, ..., n)$ **do**
        $p^{(a)}_t, h^{(a)}_{t,\pi} = \pi(o^{(a)}_t, h^{(a)}_{t-1,\pi}; \boldsymbol{\theta})$;
        $u^{(a)}_t \sim p^{(a)}_t$;
        $v^{(a)}_t, h^{(a)}_{t,V} = V(s^{(a)}_t, h^{(a)}_{t-1,V}; \phi)$;
      Execute actions u_t, observe $r_t, s_{t+1}, \boldsymbol{o}_{t+1}$;
      $\tau += [s_t, \boldsymbol{o}_t, \boldsymbol{h}_{t,\pi}, \boldsymbol{h}_{t,V}, \boldsymbol{u}_t, r_t, s_{t+1}, \boldsymbol{o}_{t+1}]$;
    Compute advantage estimate $\hat{A}$ via generalized advantage estimation (GAE) (Schulman et al., 2015) on τ, using Pop-Art (Van Hasselt et al., 2016);
    Compute reward-to-go $\hat{R}$ on τ and normalize with Pop-Art;
    Split trajectory τ into chunks of length L;
    for l in $(0, 1, ..., T//L)$ **do**
      $\mathcal{B} = \mathcal{B} \cup (\tau[l : l + T], \hat{A}[l : l + L], \hat{R}[l : l + L])$;
  for mini-batch k in $(1, 2, ..., K)$ **do**
    b = random mini-batch from D with all agent data;
    for each data chunk c in mini-batch b **do**
      update RNN hidden states for π and V from first hidden state in data chunk;
  Adam update $\boldsymbol{\theta}$ on $L(\boldsymbol{\theta})$ with data b;
  Adam update ϕ on $L(\phi)$ with data b;

E.3 Heterogeneous-Agent Proximal Policy Optimization (HAPPO)

Algorithm 3: Pseudocode of the HAPPO algorithm (Zhong et al., 2024)

Input: stepsize σ, batch size B, n agents, $step_{max}$ total training steps, T steps per episode

Output: optimized surrogate

Initialize:;
- Number of episodes $K = step_{max}//T$;
- Actor networks $\{\theta_0^i,\ \forall i \in \mathcal{N}\}$;
- Global V-value network $\{\phi_0\}$;
- Replay buffer $\mathcal{B}$;

for k in $(0, 1, ..., K-1)$ **do**

Collect a set of trajectories by running the joint policy $\pi_{\theta_k} = (\pi^1_{\theta^1_k}, ..., \pi^n_{\theta^n_k})$;

Push transitions $\{(s_t, o^i_t, a^i_t, r_t, s_{t+1}, o^i_{t+1}), \forall i \in \mathcal{N}, t \in T\}$ into $\mathcal{B}$;

Sample a random minibatch of B transitions from $\mathcal{B}$;

Compute advantage function $\hat{A}(s, \boldsymbol{a})$ based on global V-value network with GAE;

Create a random permutation of n agents $i_{1:n}$;

Set $M^{i_1}(s, \boldsymbol{a}) = \hat{A}(s, \boldsymbol{a})$;

for agent i_m in $(i_1, ..., i_n)$ **do**

Update actor i_m with the argmax of the PPO-clip objective:;

$$\theta^{i_m}_{k+1} = \frac{1}{BT}\sum_{b=1}^{B}\sum_{t=0}^{T} \min\left(\frac{\pi^{i_m}_{\theta^{i_m}}(a^{i_m}_t|o^{i_m}_t)}{pi^{i_m}_k(a^{i_m}_t|o^{i_m}_t)} M^{i_{1:m}}(s_t, \boldsymbol{a}_t), clip\left(\frac{\pi^{i_m}_{\theta^{i_m}}(a^{i_m}_t|o^{i_m}_t)}{pi^{i_m}_k(a^{i_m}_t|o^{i_m}_t)}, 1 \pm \epsilon\right) M^{i_{1:m}}(s_t, \boldsymbol{a}_t)\right)$$

if $m \neq n$ **then**

Compute:;

$$M^{i_{1:m+1}}(s, \boldsymbol{a}) = \frac{\pi^{i_m}_{\theta^{i_m}}(a^{i_m}|o^{i_m})}{pi^{i_m}_k(a^{i_m}|o^{i_m})} M^{i_{1:m}}(s, \boldsymbol{a})$$

Update the V-value network:;

$$\phi_{k+1} = \text{argmin}_\phi \frac{1}{BT}\sum_{b=1}^{B}\sum_{t=0}^{T}\left(V_\phi(s_t) - \hat{R}_t\right)^2$$

E.4 Per-Step Process of Cost-Sharing Environment

Algorithm 4: Cost-Sharing Step

Input: Actions of the agents
Output: Observations, rewards, terminations, truncations, and information

foreach *agent in remaining agents* **do**
    if *agent is bankrupt* **then**
        continue
    Update firm's production and sales;
    Calculate net earnings;
    Update firm's cash flow;
    if *firm overproduces and goes bankrupt* **then**
        Set rewards to -1;
    else
        Set terminations to False and append agent to next period's remaining agents;

Check if the game ends;

if *total emissions exceed the limit* **then**
    Set terminations to True and set rewards to -1;
else if *no more agents* **then**
    Set terminations to True;
else if *total emissions are within the limit* **then**
    Set rewards to 1 for all remaining agents;

Update observations;
Set dummy infos;
Update agents;

F Design of the MARL Cost-Sharing Environment of the U.S. Gasoline Market

F.1 Supply-Demand Dynamics

At each timestep (week), the firms produce some quantities of barrels and contribute to the total stock/inventory. This inventory is considered as the total supply at the unit price.

On the other hand, the total demand at the unit price is determined by the demand forecast model based on the data obtained from U.S. Energy Information Administration (2024). The model is a combination of a seasonal autoregressive integrated moving average (ARIMA) model (Box & Pierce, 1970) and a stochastic model which encompasses sixteen notable events causing either a demand shock or change in seasonality.

ARIMA$(0,1,1)(0,0,2)_{52}$ with drift (notations adapted from Hyndman and Athanasopoulos (2018))

$$\nabla y_t = c + \epsilon_t - \phi_1 \epsilon_{t-1} + \theta_1 \epsilon_{t-52} + \theta_2 \epsilon_{t-104}$$

Coefficient	Estimate	Standard Error
MA(1)	−0.5894	0.0218
SMA(1)	0.1342	0.0245
SMA(2)	0.1350	0.0241
Drift	9.4226	26.9418

$\sigma^2 = 4738223$ $\log \mathcal{L} = -15823.2$ AIC = 31656.39 AICc = 31656.43 BIC = 31683.7

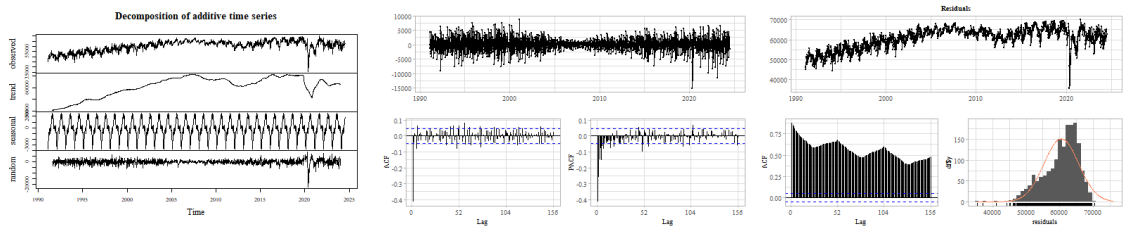

Figure 8: *The results on the tests of the ARIMA model: the decomposition of additive time series, the tests of the differences of demand between consecutive weeks (centre), and the tests of normality of the residuals (right).*

Event forecasting model

The events included in the model are divided into two types: erratic (e.g., hurricanes, new regulations, pandemics) and seasonal (e.g., recessions, severe winter weathers, summer travel seasons). Erratic events are modelled using a Poisson distribution with means based on historical occurrence in the U.S. whereas seasonal events are based on their likelihood of occurrence in each week. A simulation is run to estimate the probabilities of occurring of the sixteen events in each of fifty-two week in a year. The results of the simulation are reported in Table 1.

F.2 Rewards Mechanism

Reward design and normalization are extremely important in allowing agents to learn optimally and converge in cooperative MARL. Hence, to the design a comprehemsive reward signal to promote learning and stabilize convergence, this mixed rewards scheme is used:

- Global reward signal: +1 if the agents reach the last timestep without exceeding the maximum emissions else -1
- Local reward signal: 1/100,000 of the net earnings of the firms after the costs shared among firms are taken into account

Event	Minimum	1st Qua.	Median	Mean	3rd Qua.	Maximum
Normal	0.1828	0.7112	0.8338	0.7112	0.9328	0.9338
Hurricanes	0.0010	0.0010	0.0015	0.0015	0.0020	0.0020
Earthquakes	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010
Recession	0.0030	0.0030	0.0030	0.0030	0.0030	0.0030
Economic Boom	0.0030	0.0030	0.0030	0.0030	0.0030	0.0030
Oil Supply Disruptions	0.0020	0.0020	0.0020	0.0020	0.0020	0.0020
Peace Agreements	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010
Renewable Energy Breakthrough	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010
Improved Fuel Efficiency	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010
New Regulations	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010
Petroleum Products Subsidies	0.0020	0.0020	0.0020	0.0020	0.0020	0.0020
Pandemic	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Large Scale Events	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Severe Winter Weather	0.0000	0.0000	0.0000	0.0346	0.1000	0.1000
Summer Travel Season	0.0000	0.0000	0.0000	0.1875	0.1875	0.7500
Price Speculation	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500

Table 1: *Summary statistics for the probabilities of events.*

Based on the environment's design, the number of winning agents is equal to `round(`*rewards*`)`. For instance, in an environment with 2 firms, a reward of 1.984 implies that both firms have succeeded in staying within the emissions limit although they overall lost some cash flows.

G Design of the Environment's Scenarios

The default set-up of the environment, agents, and algorithm in cost-sharing environment is explained below.

G.1 Cost-Sharing Environment

Inherently, the environment simulates a duopoly market characterized by low fixed costs, no price ceiling, and a constant price, offering a controlled yet challenging scenario that encourages rapid agent adaptation and convergence. This setup is selected to minimize training time while maximizing the likelihood of convergence to optimal strategies.

The temporal span of the experiment is extensive, running from June 1, 2024 (week 28 of 2024) to December 31, 2099 (week 52 of 2099), encompassing a total of 3,944 weeks. This long-term horizon allows for the observation of agent behaviors and market dynamics over an extended period, providing robust data for analysis. A unique feature of the environment is its handling of excess demand: if demand exceeds supply in any given week, 10% of the unmet demand is carried over to the next period. This mechanism introduces a temporal dependency and adds a layer of strategic depth, as agents must consider the future implications of their production decisions.

Environmental constraints are rigorously modeled to reflect real-world concerns, particularly those related to emissions and climate change. The model incorporates data on the U.S. motor gasoline market's contribution to global emissions, setting a stringent upper limit of 47.62 tons of $GtCO_2e$ to align with the Paris Agreement goals. This constraint ensures that agents must balance economic objectives with environmental responsibilities, promoting sustainable decision-making.

The cost-sharing aspect of the environment is designed around a convex cost function that has been appropriately normalized and scaled. This function penalizes overproduction by taxing agents if their combined production exceeds an optimal level, which is calibrated to keep emissions within the specified upper bound. The tax imposed equals the entire price of gasoline barrels for production exceeding this threshold, compelling agents to internalize the environmental cost of their actions.

To distribute this tax burden, the serial cost-sharing mechanism is employed. This method allocates the total cost derived from the convex function among the agents in proportion to their individual contributions to the total production. This approach ensures a fair and equitable distribution of costs, encouraging cooperative behavior among agents while still maintaining competitive market dynamics.

G.2 Petroleum Firms

The agents in this multi-agent reinforcement learning thesis are modeled as two firms operating within the motor gasoline market, designed to reflect realistic economic dynamics while facilitating effective learning and convergence. Both firms begin with identical initial characteristics; each has fixed costs of $1 million, a cash flow of $1 billion, and an initial inventory of 50 million barrels. These initial conditions are selected to provide sufficient resources for strategic decision-making and to simulate a realistic market scenario.

The production capacity of each firm is capped at a maximum of 80 million barrels per week. This constraint reflects practical limitations on production capabilities and introduces a strategic element as firms must optimize their production levels within this boundary. If the number of firms in the environment increases, the production limit should be adjusted accordingly to maintain its relevance and ensure the market remains competitive.

A critical feature of the default environment is the transparency of production information. At the end of each week, firms can observe the production amounts of their competitors. This transparency allows firms to learn and adapt their strategies based on the observed behaviors of others, fostering a dynamic competitive landscape. By analyzing their rivals' production levels, firms can make more informed decisions about their own production quantities in subsequent periods, aiming to optimize their profits while considering market demand and competitive actions.

The design of the agents incorporates both cooperative and competitive elements, encouraging firms to balance their strategies between maximizing individual profits and reacting to the actions of their competitors. This configuration promotes a rich learning environment where agents can develop sophisticated strategies through repeated interactions, observation, and adaptation.

G.3 Learning Algorithm

Following the testing of various algorithms for continuous action spaces, HAPPO was chosen for its superior learning efficacy and robustness. The configuration of the default HAPPO algorithm in this MARL environment is tailored for high effectiveness, stability, and performance within a complex cost-sharing environment.

The training utilizes 10 rollout threads, with each thread representing a different random state, allowing concurrent training of 10 environment instances. The training process involves approximately 3,945,000 steps, divided into 100 episodes, each lasting up to 3,945 steps, covering the entire study period.

Actor and critic updates are conducted over 5 epochs each with single mini-batches, ensuring thorough updates. The clipped value loss technique, with a clip parameter of 0.2, maintains stability, while entropy regularization (coefficient 0.01) encourages exploration. A value loss coefficient of 1 balances policy and value learning. Gradient norm clipping with a maximum norm of 10.0 prevents gradient explosion. Generalized Advantage Estimation (GAE) with λ 0.95 and γ 0.99 optimizes the bias-variance trade-off, and a Huber loss function with delta 10.0 mitigates outliers.

The MLP network features four hidden layers of 128 neurons each, using ReLU activation to enhance non-linearity. Feature normalization stabilizes learning, and orthogonal initialization maintains variance across layers. The output layer's gain is set to 0.01 to prevent destabilizing large initial outputs. The network relies on feedforward mechanisms, avoiding recurrent neural networks to limit computational constraints. Optimization is handled by the Adam optimizer with a learning rate of 0.5 for both actor and critic to promote rapid convergence, ϵ of 0.00001 for numerical stability, and no weight decay to prevent overfitting. This precise and efficient HAPPO configuration is designed to foster effective policy learning, balancing economic and environmental objectives within the cost-sharing framework.

Below are the details of changes in the configurations of the environment, agents, and algorithm used in different research areas.

G.4 Design for Investigation in Cost Function Characteristics

The only change to the default game design is that the convex, linear, and concave cost functions are alternated in different scenarios. The other cost functions that can be used in this experiment but have not been tested are concave, piecewise, step, and sigmoid cost functions. The normalized, scaled version of these cost functions are listed as follows:

1. Convex cost function:

$$C_{\text{convex}}(x) = \frac{0.001x^2}{C_{\text{convex}}(q*)}\sigma, \tag{14}$$

2. Linear cost function:

$$C_{\text{linear}}(x) = \frac{x}{C_{\text{linear}}(q*)}\sigma, \tag{15}$$

3. Concave cost function:

$$C_{\text{concave}}(x) = \frac{1,000,000/(x+1)}{C_{\text{concave}}(q*)}\sigma, \tag{16}$$

where,

- $C(x)$ is the cost function,
- x is the number of barrels (in thousands) produced in a week,
- $q*$ is the maximum total amount that firms can produce in a week to avert global warming,
- $\sigma = \theta p q*$ is the scaling factor of the cost function which is the determined every week, where p is the current price of gasoline barrels and θ is the proportion of that current price to be charged to firms as production penalties.

G.5 Design for Investigation in Economy Attributes

This experiment investigates the impact of market stability, firm heterogeneity, and production quotas on economic behaviors within a simulated environment.

The environment is first alternated between a price-stable market or a volatile market. In the former case, the price is set at the unit price of 1, which acts as both the price floor and price ceiling. In the latter, the price is dependent on the supply-demand model of the U.S. motor gasoline market, as formulated and depicted in Appendix B and section F.1 of Appendix F.

Furthermore, the two firms in these markets are constructed as either heterogeneous or homogeneous firms. Homogeneous firms have identical fixed costs as well as initial inventory levels and cash flows, whereas one firm in the 'heterogeneous' case has twice the values of all characteristics compared to the other.

Lastly, the implementation of a production quota in the markets is also considered. This acdts as the requirement for firms to carry out production regardless, even when their exploration processes fail to find an optimal production level. A market either applies a quota of 10,000 thousand barrels to each firm and requires them to produce at least that amount, or allows firms to free decide their production.

G.6 Design for Investigation in Number of Firms

This study meticulously selects varying numbers of firms across different scenarios–ranging from monopoly to competitive markets–as a fundamental exploration of economic dynamics and strategic interactions. The experiment encompass five distinct scenarios, as the number of firms ranges from one, two, three, four, and twelve, representing a monopolistic, duopolistic, tripolistic, oligopolistic, and competitive market consecutively.

Furthermore, the inclusion of heterogeneous firms in non-monopoly markets enriches the analysis by capturing diverse firm characteristics and strategies. The mechanism for designing the attributes of the firms is similar to how it was done in the previous section. The full details of the differences in heterogeneous firms' characteristics can be found in the environment configuration file 'environment.yaml' in the project's repository. This design choice is important as it allows for a nuanced examination of how market competitiveness, driven by the number and diversity of firms, influences economic behaviors and outcomes. Understanding these dynamics is critical for developing robust economic policies and regulatory frameworks that can effectively navigate complex market environments while promoting fair competition and economic efficiency.