

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Learning-Augmented Optimization for Next-Generation Operational Systems

Permalink

<https://escholarship.org/uc/item/42g544sf>

ISBN

9798189270796

Author

Lee, Joo Seung

Publication Date

2026-05-01

Peer reviewed|Thesis/dissertation

Learning-Augmented Optimization for Next-Generation Operational Systems

By

Joo Seung Lee

A dissertation submitted in partial satisfaction of the
requirements for the degree of

Doctor of Philosophy

in

Engineering - Industrial Engineering and Operations Research

and the Designated Emphasis

in

Development Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Anil Aswani, Chair

Professor Zeyu Zheng

Professor Mark W. Mueller

Spring 2026

Learning-Augmented Optimization for Next-Generation Operational Systems

Copyright 2026

by

Joo Seung Lee

Abstract

Learning-Augmented Optimization for Next-Generation Operational Systems

by

Joo Seung Lee

Doctor of Philosophy in Engineering - Industrial Engineering and Operations Research

and the Designated Emphasis in

Development Engineering

University of California, Berkeley

Professor Anil Aswani, Chair

This dissertation presents three projects that develop data-driven optimization frameworks for sequential decision-making under uncertainty. Each project addresses a distinct application domain—healthcare, environmental resource management, and robotics—but shares a common methodological foundation: combining statistical learning with mathematical optimization to derive actionable policies from observational or streaming data.

The first project develops an interpretable reinforcement learning methodology for mechanical ventilation control in intensive care units. Using a causal, matching-based off-policy evaluation framework, we assess ventilator policies for their ability to improve patient-specific oxygenation outcomes while avoiding lung injury. Experiments on real-world ICU data demonstrate that an interpretable decision tree policy achieves performance comparable to deep reinforcement learning methods. The second project proposes a differential tax mechanism for groundwater management, modeled as a Stackelberg game between a water agency and water consumers. We derive a single-level reformulation of the resulting bilevel optimization problem, characterize the structure of optimal tax policies, and establish that the proposed mechanism can replicate the first-best social welfare outcome. Numerical simulations calibrated to the Western La Mancha aquifer validate the approach. The third project addresses profit maximization for a Robotics-as-a-Service operator who must jointly decide on job pricing and robot replacement under stochastic equipment degradation. We formulate this as a semi-Markov decision process and develop an online learning framework that integrates survival analysis and inverse optimization to estimate customer and degradation models. Numerical experiments show that the learned policy converges to near-optimal profitability.

Contents

List of Figures	iv
1 Introduction	1
1.1 Contributions and Outline	1
2 Matching-Based Off-Policy Evaluation for Reinforcement Learning Applied to Mechanical Ventilation	4
2.1 Introduction	4
2.1.1 Reinforcement Learning for Dynamic Treatment	4
2.1.2 Contributions and Outline	6
2.2 Data	6
2.3 Model	9
2.3.1 Markov Decision Process (MDP) Definition	9
2.3.2 RL Problem Definition	9
2.3.3 Policy Learning Scheme	11
2.4 Matching-Based Off-Policy Evaluation	13
2.4.1 Generation of Simulated Trajectories	14
2.5 Numerical Results	15
2.5.1 Experimental Setup	15
2.5.2 Results	18
2.6 Conclusion	23
3 Game-Theoretic Incentive Design for Sustainable Groundwater Management	24
3.1 Introduction	24
3.1.1 Differential Taxation as a Policy Mechanism	25
3.1.2 Our Contributions and Chapter Outline	25
3.2 The Model	26
3.2.1 Water Consumers' Usage and Costs	26
3.2.2 The Water Agency	26
3.2.3 Aquifer Dynamics	27
3.2.4 Water Consumers' Optimization Problem	27

3.2.5	Water Agency's Optimization Problem	28
3.3	Analysis of the Optimal Solution	29
3.4	Characteristics of the Optimal Policy	33
3.4.1	Formal Results on the Optimal Policy	33
3.4.2	Reporting Decision	35
3.4.3	Economic Intuition Behind the Optimal Policy	35
3.5	Numerical Experiment	37
3.5.1	Agent Responses and Social Welfare	37
3.5.2	Comparison with Single-Tax Regime	37
3.6	First-Best Implementation	38
3.6.1	The Centralized Planner's Problem	38
3.6.2	Structural Equivalence	39
3.7	Bounded Penalties	43
3.8	Robustness to Monitoring Effort	46
3.9	Conclusion	47
4	Profit Maximization for a Robotics-as-a-Service Model	48
4.1	Introduction	48
4.1.1	Predictive Maintenance	48
4.1.2	Failure Modeling	48
4.1.3	Inverse Optimization	49
4.1.4	Contribution and Outline	50
4.2	Model and Problem Statement	50
4.2.1	Customer and Jobs Model	50
4.2.2	Robot Degradation and Failure Model	51
4.2.3	Model of RaaS Operator	52
4.2.4	Problem Statement	53
4.3	Solution Framework	54
4.3.1	Phase 1: Learning the Utility Parameters u	54
4.3.2	Phase 2: Learning Degradation Parameters $(\theta, \lambda_0(\cdot))$	55
4.3.3	Phase 3: Solving the SMDP Problem	55
4.4	Numerical Experiments	56
4.4.1	Experimental Setup	56
4.4.2	Learning Scheme	57
4.4.3	Baseline Policies	57
4.4.4	Results	57
4.5	Conclusion	62
5	Conclusion	63
	Bibliography	65

A	Supplementary Material for Chapter 2	76
A.1	List of ICD-9 Codes Relevant to Sepsis	76
A.2	Action Space Binning Intervals	76
A.3	Transition Model Feature Grouping	77
A.4	Transition Model Hyperparameters	77

List of Figures

2.1	Overview of the pipeline for mechanical ventilation policy learning and evaluation	8
2.2	Decision tree policy with maximum depth of 3, learned from the clinicians' behavioral data	12
2.3	Overview of the matching-based off-policy evaluation framework	14
2.4	Sample trajectory using Nadaraya-Watson estimator transition model. This particular patient is a 50-year-old female without prior ICU admission who survived for at least 90 days	15
2.5	Left-child view of the decision tree policy trained by CQI algorithm with the maximum depth of 4 and the penalty coefficients (α^*, β^*) , i.e. CQI*-D4. The first splitting point corresponds to $\text{SpO}_2 < 97.6$. The units for $V_{t_{\text{set}}}$ and PEEP are mL/kg and cm H ₂ O, respectively	17
2.6	Comparison of OPE methods with bootstrap confidence intervals: (Upper) OPE estimates for various policies evaluated using WIS, FQE, and our non-parametric model-based OPE (Ours). (Lower) OPE estimates zoomed in to focus exclusively on the OPE estimates from our method	18
2.7	Distribution of actions across different RL and BC policies. The x-axis refers to action bins introduced in Table A.2	20
3.1	(Top) Partition of the (τ_0, τ_1) space showing water consumers' optimal decisions with the yellow line marking the combination that maximizes social welfare. (Bottom) The water agency's utility over the tax-rate space.	36
3.2	Numerical validation of Theorem 3 and Remark 1 for the Western La Mancha calibration ($\eta = 0.75$, $\sigma = 4$, giving $\lambda = 11/12 \approx 0.9167$). The colored regions reproduce the empirical partition from the top panel of Figure 3.1. The solid green line plots the closed-form compliance boundary $\tau_0 = \lambda\tau_1$; the dashed orange line marks the Pigouvian rate $\tau_0^{FB} = (1-\gamma)\delta = 8,000 \text{ €/Mm}^3$; the dashed purple line marks the compliance threshold $\bar{\tau}_1 = \tau_0^{FB}/\lambda \approx 8,727 \text{ €/Mm}^3$. The empirical self-reporting / truthfully-reporting boundary tracks the green ray, and the welfare-maximizing grid point lies at their intersection, in agreement with the theoretical prediction.	42

4.1	An example timeline illustrating the operational flow for the Robotics-as-a-Service (RaaS) provider. Upward arrows indicate the arrival of new customers (x_i). The robot's status alternates between being under rental ('R' in a green box) and idle ('I' in a purple box). The timeline also depicts key events like a price rejection from an arriving customer (diamond symbol), an unexpected robot failure during a job (star symbol), and a voluntary replacement initiated by the operator (heart symbol). The rewards associated with each period are shown above the timeline, reflecting revenues $\hat{u}^\top x_i$, holding costs $h\tau_i$, and costs for failure F and replacement R	51
4.2	A diagram of the three-phase learning and control framework. Phase 1: The Projected Volume algorithm is run to obtain a reliable estimate of the customer utility vector, $\hat{u}$. Phase 2: This estimate is used to learn the degradation parameters ($\hat{\theta}, \hat{\lambda}_0$) and train an initial control policy, $\hat{\pi}$. Phase 3: The system operates using this policy until a predefined number of new robot failures (K) occur. The data from these failures updates the dataset $\mathcal{D}_\theta$, which is then used to re-estimate the parameters and retrain the policy in a repeating cycle.	53
4.3	Rolling average profit rate per time unit for the oracle, online learning, myopic heuristic, and fully myopic policies, computed over a 2,000-unit window and averaged across 100 simulation runs. Shaded regions indicate ± 2 standard errors.	58
4.4	ℓ_2 -norm estimation errors for utility vector $\hat{u}$ (top) and degradation parameter $\hat{\theta}$ (bottom) versus number of customers processed, showing means and standard deviations over 100 runs.	59
4.5	Decision regions for job acceptance (purple) or rejection (yellow) in the arrival phase, plotted against desired rental duration T and cumulative degradation $\theta^\top X + \theta^\top x$, at selected percentiles of customer revenue $u^\top x$	60
4.6	Sensitivity of the interior departure threshold c^* to the cost parameters (R, F, h, λ). Each panel sweeps two parameters on a 60×60 grid while holding the remaining two at their default values from Table 4.1; the red star marks the default operating point. Grey cells denote parameter combinations under which the optimal policy never voluntarily replaces.	61

Acknowledgments

First, I want to express my gratitude to my advisor, Anil Aswani. His unwavering support, sharp technical insights, and patience have been essential for my doctoral studies. He gave me the freedom to explore complex problems while always guiding me back to rigorous, impactful methodologies. I could not have asked for a better mentor to shape my growth as an independent researcher.

I also extend my sincere thanks to Zeyu Zheng and Mark Mueller for their time and valuable feedback as members of my dissertation committee. Additionally, I am grateful to Chiwei Yan for serving on my qualifying exam committee.

My academic journey began long before this dissertation, and I owe a great deal to Uri Wilensky, Oleksii Mostovyi, Christos Dimoulas, Han Gan, and Daniel Cuzzocreo for their early guidance and support during my undergraduate years.

Beyond the university, my internships provided invaluable professional perspective, and I sincerely thank Zeyad Kassem, Arman Jabbari, Wen Cheng, Sylvain Flamant, Michael Troy, and Ashvin Veligandla for their mentorship. Their leadership and guidance throughout my internships greatly enriched my professional development.

I feel incredibly fortunate to have shared my years at Berkeley with such a wonderful group of friends. Runhan Xie made the long PhD years lighter with our lunches and easy conversations about research, job searches, and life—and the occasional terrible jokes. I am also grateful to my groupmates under Anil’s supervision, Yoon Lee, Sukanya Kudva, and Muyan Jiang, for the camaraderie of being advised by the same generous mentor. Finally, I want to thank the Korean community in IEOR—Hyunki Im, Jihun Kim, Suho Kang, Jaewon Jeong, and Tonghun Lee—for the squash games that became a quiet tradition passed down each year, and for making the department feel a little more like home. Additionally, my thanks go to my Intersection neighbors—Juliana Lee, Min Jee Lee, and Subin Jeong—for their kindness, friendship, and the many wonderful meals we shared.

Beyond Berkeley, I am grateful for my friends from my undergraduate years. Even as we moved to different cities and pursued different paths, their support has remained a constant presence outside of academia. I especially want to thank Seongsik Kim and Mark Jung, whose advice and encouragement were absolutely invaluable during my job search.

I would also like to acknowledge the IEOR department staff, whose hard work and administrative support kept our daily operations running smoothly throughout my doctoral studies.

Finally, I owe my deepest thanks to my mother, father, and brother. Your quiet sacrifices, patience, and constant encouragement made this milestone possible. Reaching this point was never a solitary effort, and I am grateful to everyone who supported me along the way.

Chapter 1

Introduction

Many important problems in engineering and operations share a common structure: a decision-maker must repeatedly choose actions over time, facing uncertainty about the environment, the behavior of other agents, or both. In healthcare, a clinician adjusts ventilator settings in response to a patient’s evolving physiology. In environmental policy, a regulatory agency sets tax rates to influence the extraction behavior of self-interested water consumers. In robotics-as-a-service, an operator prices jobs and manages equipment replacement as customers arrive with unknown preferences and robots degrade stochastically. In each case, the quality of decisions depends critically on the ability to learn relevant models from data and to optimize within the resulting framework.

This dissertation develops data-driven optimization methodologies for three such problems. While the application domains differ, the projects are connected by a shared set of methodological concerns: formulating sequential decision problems under uncertainty, estimating models of agent or system behavior from observational data, and computing policies that are both near-optimal and practically implementable. The specific tools vary across chapters (reinforcement learning and causal inference, game theory and bilevel optimization, stochastic control and survival analysis) but the underlying goal is consistent: to bridge the gap between data availability and principled decision-making.

1.1 Contributions and Outline

In this dissertation, we present significant contributions in three key areas, each addressed in a separate chapter:

- **Chapter 2: Matching-Based Off-Policy Evaluation for Reinforcement Learning Applied to Mechanical Ventilation**

In Chapter 2, we address the challenge of developing data-driven control strategies for mechanical ventilation, a critical life-support intervention that delivers controlled air and oxygen to patients’ lungs. While several reinforcement learning (RL) approaches have been proposed to optimize ventilator control, they often lack interpretability and

alignment with clinical domain knowledge, hindering their adoption in practice. Our primary contribution lies in presenting a methodology for interpretable RL aimed at improving mechanical ventilation control. Using a causal, nonparametric model-based off-policy evaluation, we assess RL policies for their ability to enhance patient-specific outcomes: specifically, increasing blood oxygen levels while avoiding aggressive ventilator settings that may cause ventilator-induced lung injuries. Through numerical experiments on real-world ICU data from the MIMIC-III database, we demonstrate that our interpretable decision tree policy achieves performance comparable to state-of-the-art deep RL methods while outperforming standard behavior cloning approaches. This chapter is based on work from [62].

- **Chapter 3: Game-Theoretic Incentive Design for Sustainable Groundwater Management**

Chapter 3 addresses the mismanagement of groundwater resources, a classic example of the tragedy of the commons where individual users overextract a shared resource to maximize their private benefit. We propose a novel incentive mechanism that leverages a differential tax structure to regulate groundwater consumption. Our framework models the interaction between a water agency and water consumers as a Stackelberg game, where the agency sets two tax rates: a lower rate for users who voluntarily install water meters and a higher rate for those who choose not to. By incorporating strategic reporting behavior and uncertainty in compliance monitoring, we identify optimal tax policies that maximize social welfare while minimizing environmental harm. Our contributions include the formulation of the bilevel optimization problem, derivation of single-level reformulations through KKT conditions, and characterization of comparative statics for the optimal policy. We demonstrate the efficacy of this approach through numerical simulations based on Western La Mancha aquifer usage data. This chapter is based on work from [60].

- **Chapter 4: Profit Maximization for a Robotics-as-a-Service Model**

Chapter 4 addresses the emerging operational challenges in Robotics-as-a-Service (RaaS), particularly in optimizing business decisions involving pricing and equipment management. We formulate a stochastic control problem where jobs arrive sequentially, and for each, the provider must decide on a price, which the customer can accept or reject. Upon job completion, the robot undergoes stochastic degradation, increasing its probability of failure in future tasks. The operator must then decide whether to replace the robot, balancing replacement costs against future revenue potential and holding costs. To solve this sequential decision-making problem under uncertainty, we develop a framework that integrates data-driven estimation techniques inspired by survival analysis and inverse optimization to learn models of customer behavior and robot failure. These models are used within a semi-Markov decision process (SMDP) framework to compute an optimal policy for joint pricing and replacement. Numerical experiments demonstrate the efficacy of our approach in maximizing profit by adap-

tively managing pricing and robot lifecycle decisions. This chapter is based on work from [61].

Chapter 2

Matching-Based Off-Policy Evaluation for Reinforcement Learning Applied to Mechanical Ventilation

2.1 Introduction

Mechanical ventilation is a critical life-support intervention involving the use of a ventilator, a machine that delivers a controlled mixture of air and oxygen to the patient’s lungs to facilitate proper gas exchange and to assist or replace the spontaneous breathing of patients who are unable to maintain adequate respiratory function. Respiratory failure requiring mechanical ventilation can result from various medical conditions such as infection, surgery, or trauma.

Improper administration of mechanical ventilation poses a considerable risk, potentially leading to ventilator-induced lung injuries and other complications [13, 92]. Specifically, the choice of high tidal volumes is known to cause inflammation and pulmonary edema [82, 92], and a high fraction of inspired oxygen is known to cause hyperoxia [85], all of which have deleterious outcomes in patients.

2.1.1 Reinforcement Learning for Dynamic Treatment

Data-driven decision-making has emerged as a powerful paradigm in healthcare, where clinicians routinely face complex dynamic treatment regimes. Reinforcement Learning (RL) methods are increasingly explored for their ability to leverage historical patient trajectories to inform sequential treatment strategies that could optimize patient outcomes. Compared to traditional static treatment guidelines, RL offers a data-centered framework for learning personalized, time-varying interventions that adapt to the evolving condition of each patient.

However, successful application of RL in healthcare requires careful consideration of several challenges. From a methodological standpoint, RL for dynamic treatment regimes

involves learning policies from high-dimensional and noisy observational data. The state space consists of continuous physiological signals and clinical variables, which must be faithfully represented to avoid introducing modeling biases. Unlike controlled laboratory settings, in the clinical context we rely solely on retrospective data, making RL inherently “offline.” This offline setting must address issues such as confounding, selection bias, and distribution shifts to ensure that learned policies are both reliable and robust.

Research in RL for healthcare has shown promise across diverse clinical domains, including HIV management [103, 78], sepsis treatment [52, 73], insulin dosing [76, 38], and mechanical ventilation [79, 53]. By framing treatment as a sequential decision-making process, RL agents aim to optimize patient outcomes over time through a policy that suggests which therapeutic action to take given the current state of the patient. As most RL approaches in clinical care rely on offline data, recent advances in offline RL have been instrumental in defining robust learning strategies under distribution constraints.

Despite these advances, key obstacles remain. A major barrier to clinical deployment is that state-of-the-art RL policies, often realized through deep neural networks, tend to be “black boxes” with limited interpretability. Clinicians require transparency and clear rationale for suggested interventions—particularly critical for life-supporting treatments—before they can trust and adopt RL-guided strategies in practice. Similarly, commonly used Off-Policy Evaluation (OPE) methods in RL often produce high-variance or opaque estimates of expected performance, further hindering understanding and acceptance by clinicians. Improved methods are needed that not only produce accurate quantitative evaluations, but also enable intuitive, clinically relevant interpretations of how candidate policies alter patient trajectories.

Reinforcement Learning for Mechanical Ventilation

RL-based decision support lends itself well to the mechanical ventilation setting, as clinicians must continually adjust ventilator settings such as tidal volume ($V_{t_{\text{set}}}$) and fraction of inspired oxygen (FiO_2) in response to changing patient conditions. The complexity and high stakes of these decisions have motivated recent efforts to apply RL to optimize ventilation strategies [79, 53, 83, 59, 30]. For instance, tabular Q-learning and deep RL methods have been explored to reduce mortality and other complications, while providing automated or semi-automated guidance on setting $V_{t_{\text{set}}}$, positive end-expiratory pressure (PEEP), and FiO_2 . Although these methods demonstrate potential improvements over standard clinical practice, key challenges persist. The learned policies remain difficult to interpret due to reliance on large-scale neural networks. Moreover, the OPE methods used to validate policies often fail to provide in-depth, clinically meaningful insights about how these policies will affect patient physiology, beyond coarse metrics like mortality.

A recurring issue in these applications is the choice of state representation. Some earlier attempts discretized the state space by clustering patient measurements, aiming to derive a Markov Decision Process (MDP) with manageable complexity [79]. However, as lung physiology and patient trajectories are inherently continuous and heterogeneous, state dis-

cretization can induce biases and lose critical granularity. Recent perspectives suggest that continuous state representations are essential for reliable modeling and policy learning, as state discretization can degrade policy performance and consistency [67, 64].

2.1.2 Contributions and Outline

In contrast to prior work, which often relies on black-box deep RL models that lack interpretability or fail to address clinical constraints, this chapter proposes a novel approach to optimizing mechanical ventilation strategies using interpretable RL and a robust off-policy evaluation (OPE) framework. Our contributions are as follows:

1. **Interpretable RL for Mechanical Ventilation:** We develop an interpretable RL method based on the Conservative Q-Improvement (CQI) algorithm [86]. Unlike deep RL approaches, our method produces decision tree-based policies that are transparent, enabling clinicians to understand and trust the suggested ventilation strategies.
2. **A Matching-Based Off-Policy Evaluation Method:** We propose a causal, non-parametric OPE method that generates counterfactual patient trajectories, allowing for deeper policy evaluation beyond traditional value functions. This includes clinically meaningful metrics such as overall SpO₂ gain and the proportion of time aggressive ventilation settings are chosen.
3. **Clinically-Informed Reward Function:** We design a reward function that directly optimizes for SpO₂ improvements while explicitly penalizing aggressive ventilator settings (e.g., high tidal volumes and high inspired oxygen levels) to avoid ventilator-induced complications, which better aligns the RL objective with clinical best practices [91].

To demonstrate the effectiveness of our approach, we compare the performance of our interpretable RL policy (CQI) against a supervised behavior cloning baseline and a state-of-the-art deep RL policy (Conservative Q-Learning). Using both traditional OPE methods and our novel OPE framework, we evaluate the policies on real-world ICU data from the MIMIC-III database. Our results show that the interpretable CQI policy achieves competitive performance while adhering to clinical constraints and providing transparent decision-making pathways.

2.2 Data

This study leverages data from MIMIC-III, a large-scale and openly accessible critical care database [49]. We extracted patient stays involving mechanical ventilation, recording patient characteristics, lab values, vital signs, and ventilator settings at 4-hour intervals. If multiple mechanical ventilation events occurred within the same ICU stay, we selected the earliest event that lasted for more than 24 hours.

Inclusion Criteria Following criteria similar to [79], we included patients who:

1. Were at least 18 years old at admission.
2. Had documented 90-day mortality outcomes.
3. Had recorded tidal volumes, ensuring the use of volume-controlled ventilation (thus excluding pressure-controlled modes).

These criteria yielded a total of 10,739 mechanical ventilation events for analysis, processed through the pipeline summarized in Figure 2.1.

Preprocessing Steps Continuous features were winsorized at the 0.3th and 99.7th percentiles to mitigate the influence of extreme outliers. We truncated any ventilation event longer than 72 hours (i.e., beyond the 18th time step) to the first 18 time steps. There are four hours between each time step. Missing values were handled in a stepwise manner. Initially, we used forward and backward filling for up to five consecutive time steps. Any residual missing data were then imputed using K-nearest neighbors (KNN). Patient stays requiring more than 50% imputation for a particular action parameter over the time course were excluded.

To provide additional clinical context, we retrieved ICD-9 codes for each patient and flagged a ventilation event as sepsis-related if the admission included at least one sepsis-associated code. Given that sepsis frequently necessitates mechanical ventilation due to acute organ failure and circulatory compromise [75], identifying these events helps differentiate respiratory failure arising from severe infections versus other causes like drug intake or post-surgical recovery.

Data Variables Each patient episode includes six time-invariant demographic and outcome-related variables (e.g., age, sex, 90-day mortality) and 20 time-dependent variables measured at each 4-hour interval. We categorize these variables as follows:

1. Demographics (6): Sex, Sepsis, Weight, Age, 90 days mortality, ICU readmission¹
2. Lab Values (3): PaCO₂, PaO₂, SpO₂
3. Vital Signs (11): Systolic BP, Diastolic BP, Mean BP, Glasgow Coma Scale (GCS), Heart Rate, PaO₂FiO₂ Ratio (P/F Ratio), Temperature, Respiratory Rate, Spontaneous Tidal Volume, Shock index, Urine output
4. Others (6): Total IV fluids administered, Cumulative fluid balance (CBA), SIRS, SOFA, Mean Airway Pressure (MAP), Oxygenation Index

¹A variable indicating if the patient has prior ICU admission history

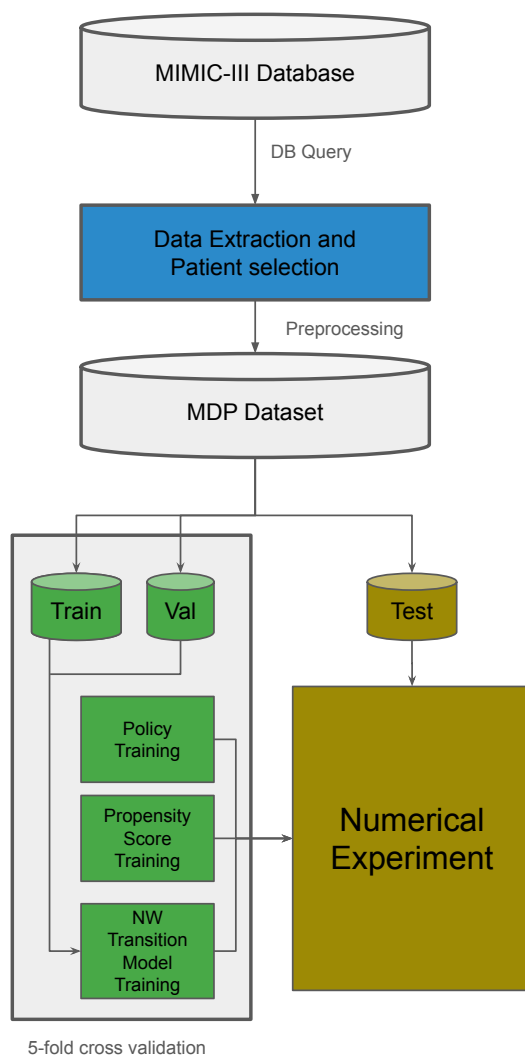


Figure 2.1: Overview of the pipeline for mechanical ventilation policy learning and evaluation

2.3 Model

2.3.1 Markov Decision Process (MDP) Definition

Mechanical ventilation is a critical intervention in the ICU, where the clinician must continuously adjust the ventilator settings based on a patient’s evolving condition. This forms a finite time-horizon sequential decision-making problem, as each adjustment to the ventilator affects future patient outcomes. To capture this, we model the mechanical ventilation control problem as a Markov decision process (MDP), represented by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}$ is the transition dynamics, r is the reward function, and γ is the discount factor. A policy $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$ maps patient states to distributions over actions.

The state-value and action-value functions under π are defined as:

$$V_\pi(s) = \mathbb{E}_\pi \left[\sum_{t=1}^T \gamma^t r(s_t, a_t, s_{t+1}) \mid s_1 = s \right]$$

and

$$Q_\pi(s, a) = \mathbb{E}_\pi \left[\sum_{t=1}^T \gamma^t r(s_t, a_t, s_{t+1}) \mid s_1 = s, a_1 = a \right],$$

where $\mathbb{E}_\pi$ denotes expectation under π and the transition dynamics. The goal of reinforcement learning is to find an optimal policy π^* that maximizes the expected cumulative reward V_{π^*} . In offline reinforcement learning, the task is to find an optimal policy π^* using only a pre-gathered dataset $\mathcal{D} = \left\{ \left(s_t^{(i)}, a_t^{(i)}, r_t^{(i)}, s_{t+1}^{(i)} \right)_{t=1}^T \right\}_{i=1}^N$, without the option of real-time interaction. This is essential for safety-critical applications like mechanical ventilation, where direct experimentation is not feasible.

2.3.2 RL Problem Definition

State Space We begin with 26 candidate features and select 15 as directly observable to the RL agent². These features provide a comprehensive view of a patient’s respiratory, hemodynamic, and mental status, while weight is included because tidal volume adjustments typically scale with patient weight.

To address potential confounding in the transition dynamics, we introduce a non-time-varying score z derived from a set of 10 “type” features³. These type features include composite indices like SOFA, SIRS, and shock index, which are partially composed of vari-

²Sepsis, Weight, Age, Heart rate, Respiratory rate, SpO₂, PaO₂FiO₂ Ratio, Spontaneous tidal volume, MAP, PaCO₂, Systolic BP, Diastolic BP, Cumulative fluid balance, PaO₂, GCS

³SOFA, SIRS, Shock Index, Total IV fluids administered, Urine output, Mean BP, Sex, ICU Readmission, Temperature, Oxygenation Index

ables already observed by the RL agent, yet still provide important additional cues about a patient’s underlying condition. Other type variables, such as mean blood pressure and urine output, offer indirect or redundant signals and are thus grouped here to capture confounding influences not directly represented in the primary observable set. We estimate z by training a logistic regression model at the initial time step to predict 90-day mortality, using these type features as inputs. Although the patient’s mortality probability might evolve in response to clinical interventions during ventilation, we treat z as a proxy for the patient’s *type*, a set of relatively stable traits (e.g., comorbidities, genetics) that are unlikely to change significantly within the short duration of mechanical ventilation [16]. Consequently, z is held constant throughout each ventilation episode, reflecting persistent patient characteristics rather than the immediate effects of treatment. This approach is consistent with prior work on policy evaluation in healthcare [74, 8], and z will be used later in the transition model through propensity score matching.

The final state space is therefore $\mathcal{S} \subseteq \mathbb{R}^{16}$, consisting of 15 observable patient features plus the propensity score z .

Action Space The ventilator has three settings:

- Ideal weight adjusted tidal volume ($V_{t_{\text{set}}}$)
- Positive end-expiratory pressure (PEEP)
- Fraction of inspired oxygen (FiO_2)

Although these settings are inherently continuous, we discretize the action space, yielding 7 levels for $V_{t_{\text{set}}}$, 4 for PEEP, and 7 for FiO_2 , based on a clustering analysis that revealed clinicians tend to choose from a relatively small set of commonly used ranges. This choice is further motivated by the more established theoretical foundations of off-policy evaluation in discrete action spaces [67]. Table A.2 in the Appendix provides details of how these levels were defined and bin boundaries selected.

Reward The primary objective of operating a mechanical ventilator is to facilitate the delivery of oxygen and remove carbon dioxide from the body [91]. The effectiveness of oxygen delivery can be assessed by monitoring SpO_2 , which represents blood oxygen levels [88]. Although setting high values of $V_{t_{\text{set}}}$ and FiO_2 is correlated with immediate increases in SpO_2 , such severe ventilator settings can increase the risk of trauma and mortality [1, 15, 26]. However, in severe cases of respiratory failure or injury, these “aggressive” settings may be necessary to maintain adequate oxygenation [26, 40]. Therefore, an RL policy that entirely avoids these aggressive values is not feasible, as it would fail to address the needs of patients in critical conditions. Nevertheless, it is undesirable to maintain patients on such settings for prolonged periods due to associated risks.

To balance these considerations, we define a reward function $r(s_t, s_{t+1}, a_t)$ that encourages moderate improvements in SpO_2 , while penalizing overly aggressive ventilator settings. The agent incurs a penalty in the form of subtracted rewards when $V_{t_{\text{set}}}$ is larger than 10 mL/kg or when FiO_2 is larger than 0.60, which are considered high [82, 7].

Furthermore, we design the reward function to avoid rewarding increases in SpO₂ when blood oxygen saturation is already at or above 95%. This decision reflects the clinical observation that excessively high oxygen saturation can be harmful, leading to potential complications [97]. As such, we set the reward from SpO₂ gain to zero if SpO₂ before or after the time step is above 95%, aligning with the clinical goal of keeping SpO₂ within a safe range. Therefore, the final reward function $r(\cdot)$ is expressed as:

$$r^{spo}(s_t, s_{t+1}) = \begin{cases} s_{t+1}(\text{SpO}_2) - s_t(\text{SpO}_2) & \text{if } s_{t+1}(\text{SpO}_2) < 95 \text{ and } s_t(\text{SpO}_2) < 95 \\ 0 & \text{otherwise} \end{cases}$$

$$r^a(a_t) = -\alpha \cdot \mathbb{1}\{a_t(\text{Vt}_{\text{set}}) \geq 10\} - \beta \cdot \mathbb{1}\{a_t(\text{FiO}_2) \geq 0.6\}$$

$$r(s_t, s_{t+1}, a_t) = r^{spo}(s_t, s_{t+1}) + r^a(a_t),$$

where $\mathbb{1}$ is the indicator function and $\alpha, \beta > 0$ are constants chosen via cross-validation. The binary nature of the penalties for Vt_{set} and FiO₂ simplifies hyperparameter selection, as it avoids introducing additional parameters, making the optimization process more efficient. We use the discount factor of $\gamma = 0.99$, reflecting the fact that prolonged mechanical ventilation is associated with increased risk of complications [66].

2.3.3 Policy Learning Scheme

Behavior Cloning (BC) Behavior cloning is a baseline approach that treats policy learning as a supervised classification problem with the objective of imitating the historical actions taken by clinicians in the dataset. We use a random forest (a non-parametric model) to learn a behavior cloning policy. Out of 196 possible combinations of actions, 164 action triplets were observed in the data. The random forest model was trained with a classification objective, which was sufficient to achieve a highly accurate behavior cloning policy. Figure 2.2 shows a depth-3 decision tree fit to the same clinician trajectories, included as an interpretable reference for later comparisons.

Conservative Q-Learning (CQL) The CQL algorithm [54] is designed to mitigate overestimation of Q-values in offline learning by introducing a regularization mechanism. Two regularization terms are added to the standard Bellman error in the training objective $\mathcal{R}(\theta)$ to minimize the overestimated Q-value at unseen actions while maximizing it under the data distribution:

$$\mathcal{R}(\theta) = \max_{\mu} \mathbb{E}_{s \sim \mathcal{D}, a \sim \mu(\cdot|s)} [Q(s, a)] - \mathbb{E}_{s \sim \mathcal{D}, a \sim \hat{\pi}_{\beta}(\cdot|s)} [Q(s, a)] + \mathcal{R}(\mu)$$

The d3rlpy module was used to train a CQL policy [90].

Conservative Q-Improvement (CQI) Conservative Q-Improvement [86] learns a decision tree policy directly. We implemented CQI but modified it to constrain the decision tree to a specified maximum depth. In contrast to CART algorithms that split a node based on

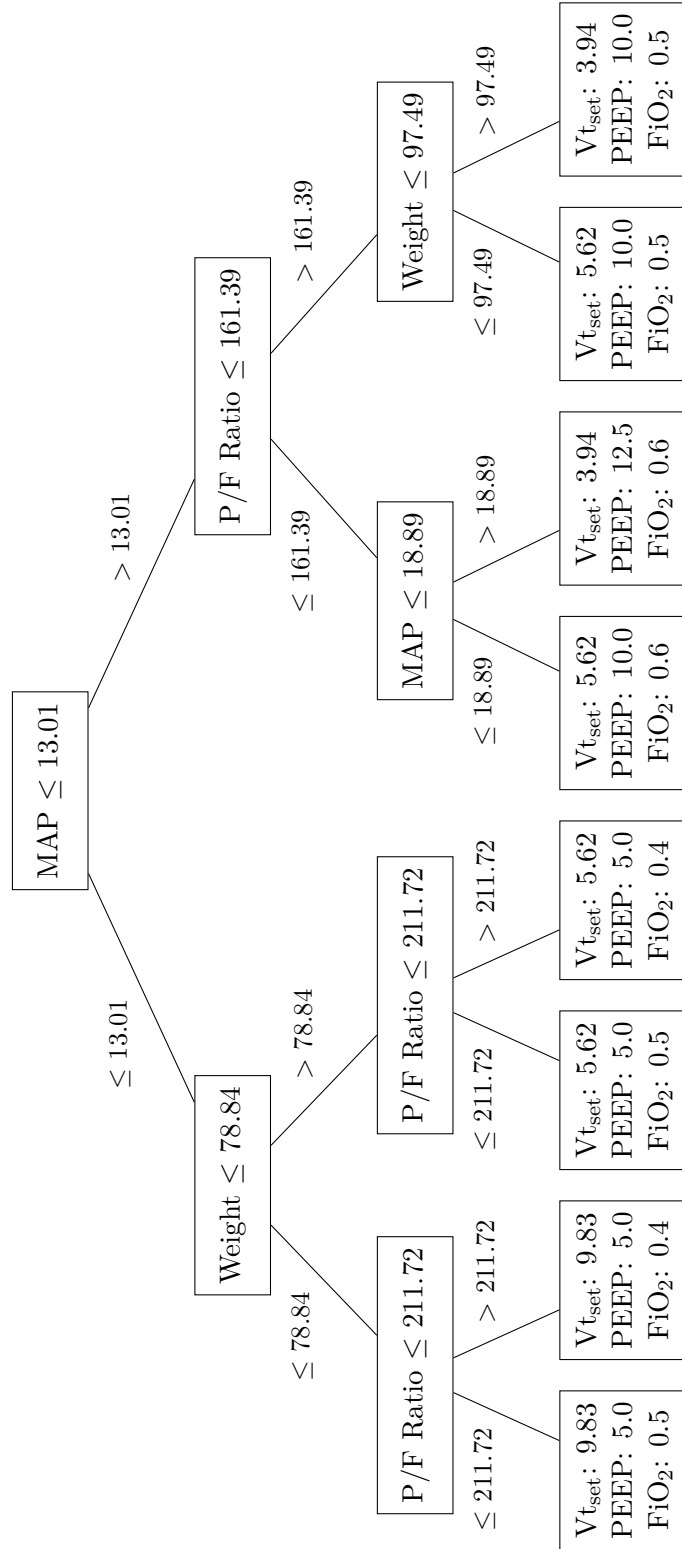


Figure 2.2: Decision tree policy with maximum depth of 3, learned from the clinicians' behavioral data

information gain, the CQI algorithm splits a node based on gain of Q-values. Each leaf node L represents an abstract state space (intersections of half-spaces created by the decision path), and the Q-value at each leaf is updated by the Q-learning update equation:

$$Q(L(s), a) \leftarrow (1 - \alpha)Q(L(s), a) + \alpha(r + \gamma \max_{a'} Q(L(s_{t+1}), a')),$$

where $\alpha \in (0, 1)$ is a learning rate and $L(s)$ is the leaf node corresponding to state s . Each branching splits the state space into two half spaces on a specific feature that leads to the Q-value gain, and the leaf node corresponds to the action with the highest Q-value in each state.

2.4 Matching-Based Off-Policy Evaluation

We use the Nadaraya-Watson estimator (NWE) [72, 101], a kernel-density-based non-parametric regression model, to estimate next-step states for off-policy evaluation. Although neural network approaches can capture complex dynamics, they often require substantially larger datasets for stable training. By contrast, NWE is balanced between flexibility and data requirements, making a suitable choice given our dataset size. Formally, for a current state-action pair (s, a) and z , we model the next state as:

$$s' = g(s, a, z) + \epsilon,$$

where $g(\cdot)$ is an unknown nonlinear function and ϵ is a noise term. To estimate $g(\cdot, \cdot, \cdot)$ from offline data $\{(s_i, a_i, r_i, s'_i)\}_{i=1}^n$, the NWE takes a weighted average of next states s'_i from similar patient transitions (s_i, a_i) and propensity score z in the dataset. Formally, for a given state-action pair (s, a) and z , the NWE is defined as:

$$\hat{g}(s, a, z) = \frac{\sum_{i=1}^n K(\|s_i - s\|/h)K(\|a_i - a\|/h)K(\|z_i - z\|/h) \cdot s'_i}{\sum_{i=1}^n K(\|s_i - s\|/h)K(\|a_i - a\|/h)K(\|z_i - z\|/h) + \lambda},$$

where $K(\cdot)$ is a kernel function, λ is a regularization constant, and h is a bandwidth parameter.

In the context of patient state transition modeling, we apply separate bandwidths for the state variables, action variables, and the propensity score. More specifically, the 16 visible state variables are divided into four clinically oriented groups: hemodynamic, respiratory, blood gas, and miscellaneous. Each group is assigned its own bandwidth parameter $(h_{s,h}, h_{s,r}, h_{s,b}, h_{s,m})$, allowing the model to capture the distinct rates of physiological change and dependencies within each category more effectively.

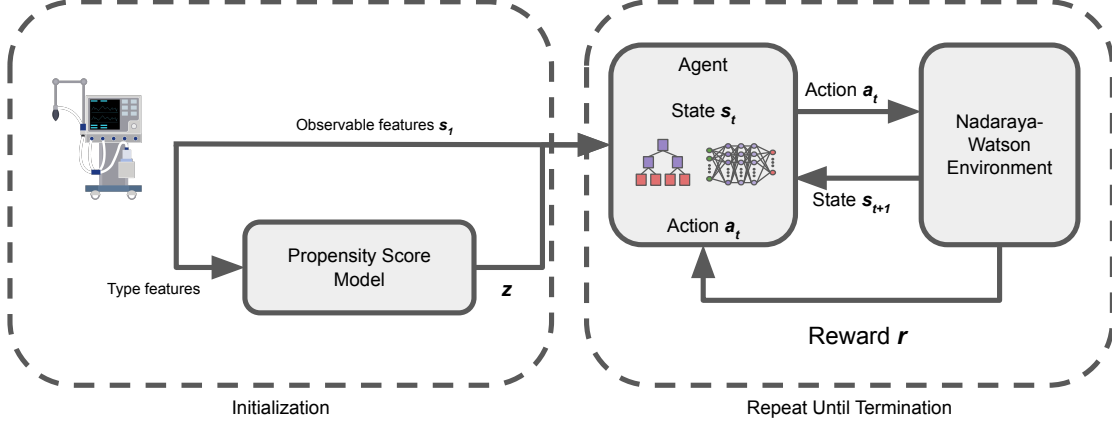


Figure 2.3: Overview of the matching-based off-policy evaluation framework

$$\hat{s}' = \frac{\sum_{i=1}^n \left(\prod_{k \in \{h, r, b, m\}} K\left(\frac{\|s_{i,k} - s_k\|}{h_k}\right) \prod_{t \in \{v, p, f\}} K\left(\frac{\|a_{i,t} - a_t\|}{h_{a,t}}\right) K\left(\frac{\|z_i - z\|}{h_z}\right) \right) \cdot s'_i}{\sum_{i=1}^n \left(\prod_{k \in \{h, r, b, m\}} K\left(\frac{\|s_{i,k} - s_k\|}{h_k}\right) \prod_{t \in \{v, p, f\}} K\left(\frac{\|a_{i,t} - a_t\|}{h_{a,t}}\right) K\left(\frac{\|z_i - z\|}{h_z}\right) \right) + \lambda}. \quad (2.1)$$

We employ the Epanechnikov kernel, which has finite support. Therefore, training samples that fall outside the bandwidth for any of the state, action, or propensity score variables are not counted towards the Nadaraya-Watson estimate. That is, given the bandwidth parameters h_s , h_a , and h_z , the current state s_0 , action taken a_0 , propensity score z_0 , our estimate of the patient state at the next time step $\hat{s}'_0$ is given by Equation 2.1.

2.4.1 Generation of Simulated Trajectories

In the medical context, confounders can influence patient transitions. To mitigate this, we incorporate a patient-specific propensity score z —estimated from a logistic regression model at the start of each episode—into our kernel matching. This score remains constant during the patient’s stay and effectively stratifies the population into more homogeneous subgroups. By conditioning on z , we help ensure that transitions are matched not only on observable current states and actions but also on persistent latent factors, yielding more reliable counterfactual estimates.

To evaluate evaluation policy π_e , we proceed as follows:

As data preparation, the training dataset is prepared in the form of MDP transitions, i.e., $\{(s_t^i, a_t^i, r_t^i, s_{t+1}^i, z_i)\}$, for the transition at time t for the i^{th} episode. Since age, weight,

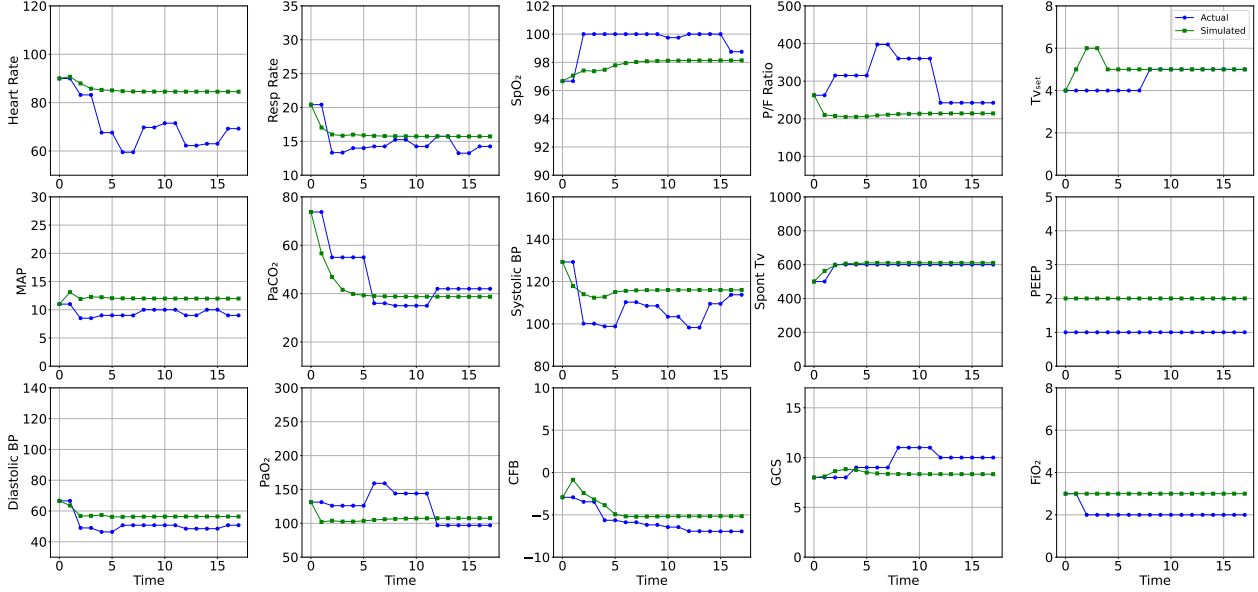


Figure 2.4: Sample trajectory using Nadaraya-Watson estimator transition model. This particular patient is a 50-year-old female without prior ICU admission who survived for at least 90 days

and sepsis flag are assumed to not vary in a single episode, the target variable s_{t+1} has the three features removed. The propensity score z is computed with the initial state of the episode. Given an evaluation policy π_e , an action $a \sim \pi_e(s_1)$ is drawn, and the Nadaraya-Watson estimator transition model predicts the next state and iterates until the specified time horizon, which is set to equal the actual duration of the patient episode. This is also illustrated in Figure 2.3.

A sample trajectory rolled out on the initial state drawn from the testing dataset with the chosen hyperparameters is shown in Figure 2.4.

2.5 Numerical Results

2.5.1 Experimental Setup

Training Scheme

We split the mechanical ventilation events into training and testing subsets in an 80:20 ratio. Within the training set, a 5-fold cross-validation procedure was used to select hyperparameters for Behavior Cloning (BC), Conservative Q-Learning (CQL), Conservative Q-Improvement (CQI), and the Nadaraya-Watson (NW) transition model. In addition, a separate grid search was conducted to identify the penalty coefficients α and β for discour-

aging aggressive ventilator settings.

A behavior cloning policy was trained using a random forest to replicate clinician-chosen discretized actions. Its final testing mean absolute errors (MAEs) in predicting Vt_{set} , PEEP, FiO_2 were 0.383, 0.755, and 0.825, respectively. Concurrently, a logistic regression model achieved a testing ROC-AUC of 0.658 for estimating each patient’s propensity score, which remained fixed throughout their ventilation episode. This logistic regression score was incorporated into our off-policy evaluation to address latent confounding. For our transition model, we employed the Nadaraya-Watson (NW) estimator with separate bandwidth parameters for state features, action features, and the propensity score. Using an Epanechnikov kernel and cross-validation, the selected bandwidths $(h_{s,h}, h_{s,r}, h_{s,b}, h_{s,m}, h_a, h_z)$ were (3.04, 2.8, 2.53, 2.03, 1.0, 1.5), and the regularization constant λ was 10^{-3} . Under the BC policy, the tuned NW model attained a mean absolute error 1.56 when predicting SpO_2 time step by time step, and an AUC-ROC of 0.716 for distinguishing whether SpO_2 exceeded 94%. A separate grid search over the penalty coefficients $\alpha, \beta \in [\frac{3}{20}, \frac{3}{21}, \frac{3}{22}, \frac{3}{23}, \frac{3}{24}]$ and the CQL conservativeness parameter $\alpha_{CQL} \in [4, 2, 1, 0.5, 0.25]$ was carried out to balance maximizing SpO_2 improvements and minimizing aggressive ventilator settings. Aggressive actions were defined as those with action index $Vt_{\text{set}} \geq 6$ or $FiO_2 \geq 4$, corresponding to the higher bins in our discretized space. The optimal values were $\alpha^* = 0.375, \beta^* = 0.75$, and $\alpha_{CQL} = 0.25$. With these settings, the CQL policy achieved an average increase of SpO_2 of 0.070, with Vt_{set} deemed aggressive in only 1.13% of time steps and FiO_2 aggressive in 3.46% of time steps.

Off Policy Evaluation

Since direct real-time experimentation with new policies is infeasible, we rely on OPE to estimate each policy’s performance. OPE leverages retrospective data to approximate the expected returns of a policy different from the one that generated the dataset, circumventing safety risks. We considered three widely used OPE methods:

1. *Importance Sampling* [84] Importance sampling reweighs observed trajectories based on the ratio of the evaluation policy to the behavior policy at each step. We use per-step weighted IS estimator (PSWIS), which is given by:

$$\hat{V}_{\text{step-WIS}} = \frac{1}{n} \sum_{i=1}^n \sum_{t=1}^T \gamma^t \frac{\rho_{1:t}^{(i)}}{w_t} r_t^{(i)}.$$

where $\rho_{1:t}^{(i)} = \prod_{s=1}^t \frac{\pi_e(a_s^{(i)} | s_s^{(i)})}{\pi_b(a_s^{(i)} | s_s^{(i)})}$ and $w_t = \sum_{i=1}^{|\mathcal{D}|} \frac{\rho_{1:t}^{(i)}}{|\mathcal{D}|}$.

2. *Fitted Q-Evaluation* Fitted Q-Evaluation [57] is a regression-based direct method to perform OPE. It iteratively improves the estimate of the Q-function by computing the targets $y_i = r_t + \gamma Q_{k-1}(s_{t+1}, \pi_e(s_{t+1}))$ and solving the supervised problem $Q_k =$

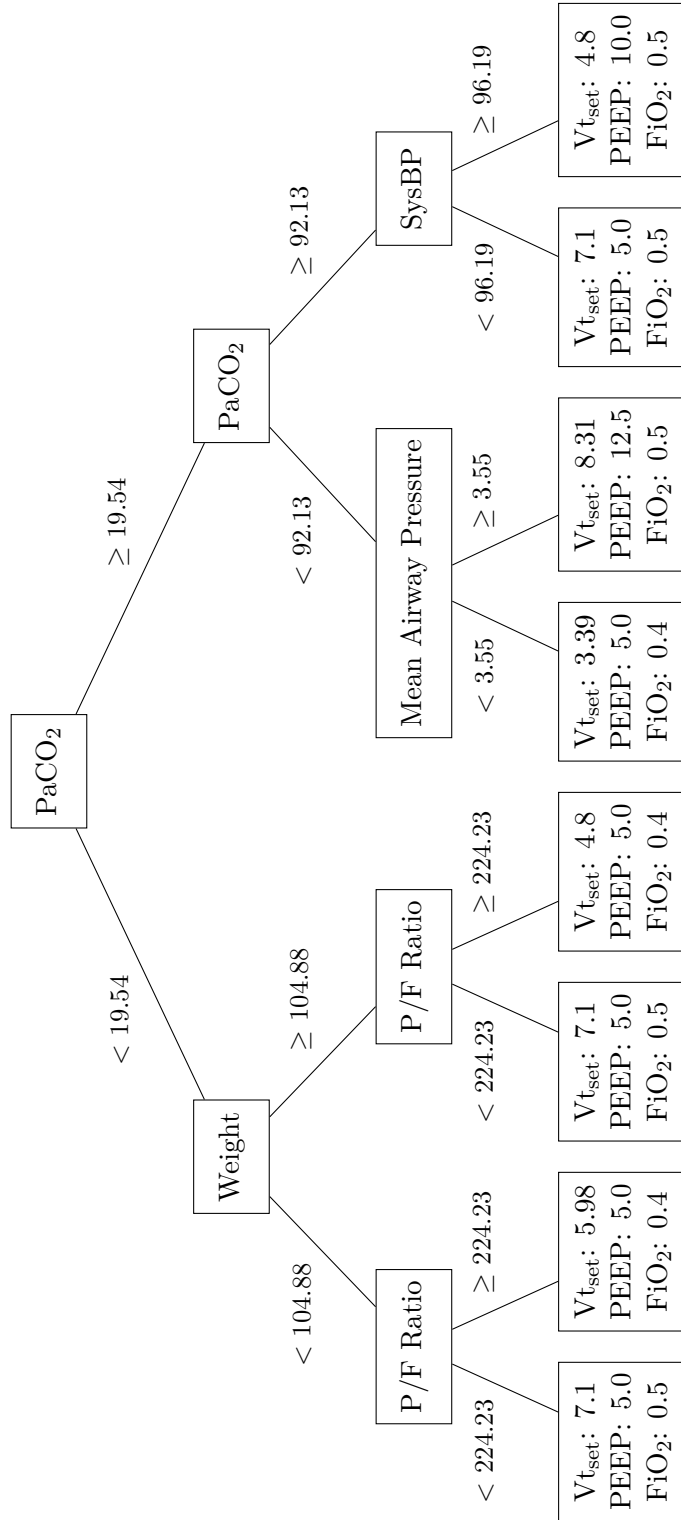


Figure 2.5: Left-child view of the decision tree policy trained by CQI algorithm with the maximum depth of 4 and the penalty coefficients (α^*, β^*) , i.e. CQI*-D4. The first splitting point corresponds to $\text{SpO}_2 < 97.6$. The units for $V_{t_{\text{set}}}$ and PEEP are mL/kg and cm H₂O, respectively

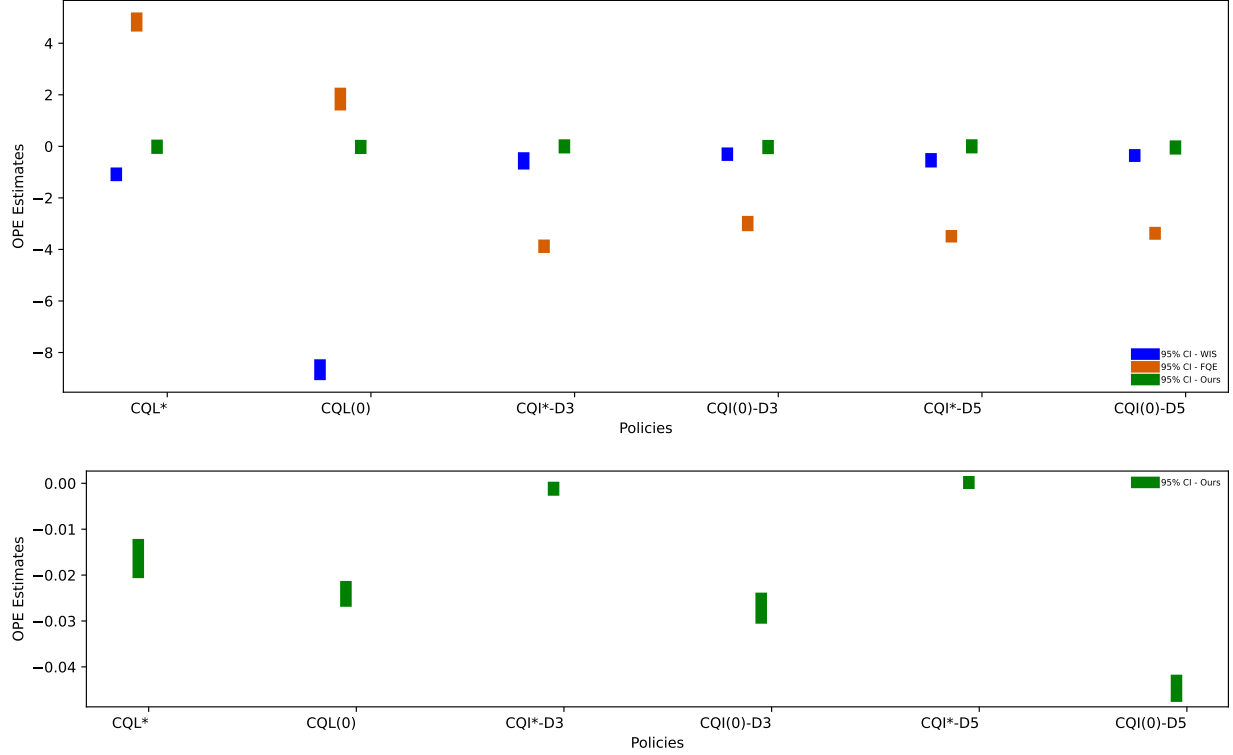


Figure 2.6: Comparison of OPE methods with bootstrap confidence intervals: (Upper) OPE estimates for various policies evaluated using WIS, FQE, and our nonparametric model-based OPE (Ours). (Lower) OPE estimates zoomed in to focus exclusively on the OPE estimates from our method

$\arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (f(s_t, a_t) - y_t)^2$, where $\mathcal{F}$ is a learnable function class, such as the set of neural networks.

2.5.2 Results

Let π_{BC} denote the behavior cloning policy, $\pi_{CQL(0,0)}$ and π_{CQL^*} each refer to CQL policy with $\alpha = \beta = 0$ and α^*, β^* , and $\pi_{CQI(0,0)-Dn}$ and π_{CQI^*-Dn} refer to CQI policy with maximum depth of n , each with $\alpha = \beta = 0$ and α^*, β^* .

Comparison of various OPE methods

We empirically compare the performances of various OPE methods in Figure 2.6. The results reveal several key trends and distinctions across the evaluated methods: WIS, FQE, and our proposed nonparametric model-based OPE method (“Ours”).

Algorithm	% $\text{SpO}_2 > 95$	$\Delta(\text{SpO}_2)$	% Vt_{set}	% FiO_2
BC	98.10	-0.47	24.19	25.57
CQL(0,0)	98.99	0.04	15.63	37.68
CQL*	99.87	0.63	0.76	11.75
CQI(0,0)-D3	99.75	-0.32	1.22	60.86
CQI*-D3	99.03	-1.19	0.00	0.00
CQI(0,0)-D5	100	N/A	0.86	99.25
CQI*-D5	99.70	-1.87	0.00	0.00

Table 2.1: Comparisons of the different RL and BC policies. % $\text{SpO}_2 > 95$ indicates the proportion of episodes that ended with terminal SpO_2 of 95 or above. $\Delta(\text{SpO}_2)$ indicates the average increase in SpO_2 per event, only among those whose terminal SpO_2 is less than 95. Proportion of patients starting with $\text{SpO}_2 \leq 95\%$ is 13%

FQE consistently rates policies from the CQL family as performing better than those from the CQI family. This observation aligns with our intuition that more complex models, such as those enabled by deep reinforcement learning, achieve better outcomes by leveraging their capacity to capture nuanced patient trajectories. In contrast, WIS identifies CQL(0,0) as particularly poor, suggesting that this method is sensitive to the aggressive policies characteristic of unpenalized algorithms.

Our method tends to favor policies trained with penalties (e.g., π_{CQL^*} , $\pi_{\text{CQI}^*-\text{D3}}$, $\pi_{\text{CQI}^*-\text{D5}}$) over those without penalties. This trend is consistent across all policy pairs. A plausible interpretation is that more aggressive policies, which avoid penalization, tend to perform worse under our nonparametric OPE framework. This may reflect an inherent property of the method: its ability to account for penalties imposed on aggressive action choices, which are more likely with unpenalized policies. Notably, this alignment with clinical intuition underscores the clinical relevance of our OPE in prioritizing cautious policies that balance performance with patient safety.

The comparison also illustrates the balance between complexity and interpretability. While deep RL methods such as π_{CQL^*} achieve strong performance, our method’s preference for interpretable policies like $\pi_{\text{CQI}^*-\text{D3}}$ and $\pi_{\text{CQI}^*-\text{D5}}$, when penalties are applied, suggests that interpretable models can remain competitive when designed with appropriate constraints. This balance is critical for clinical adoption, where the transparency of decision-making is as vital as quantitative performance metrics.

Distribution of Action Choices and Performance Comparison

The distribution of actions selected by different policies reveals significant differences in their behavior and alignment with clinical goals. As shown in Figure 2.7, policies trained with penalties on aggressive actions demonstrate a marked reduction in the frequency of such actions compared to both the behavior cloning policy and unregularized policies. Specifically, the CQI policies constrained by positive penalties entirely avoid aggressive actions,

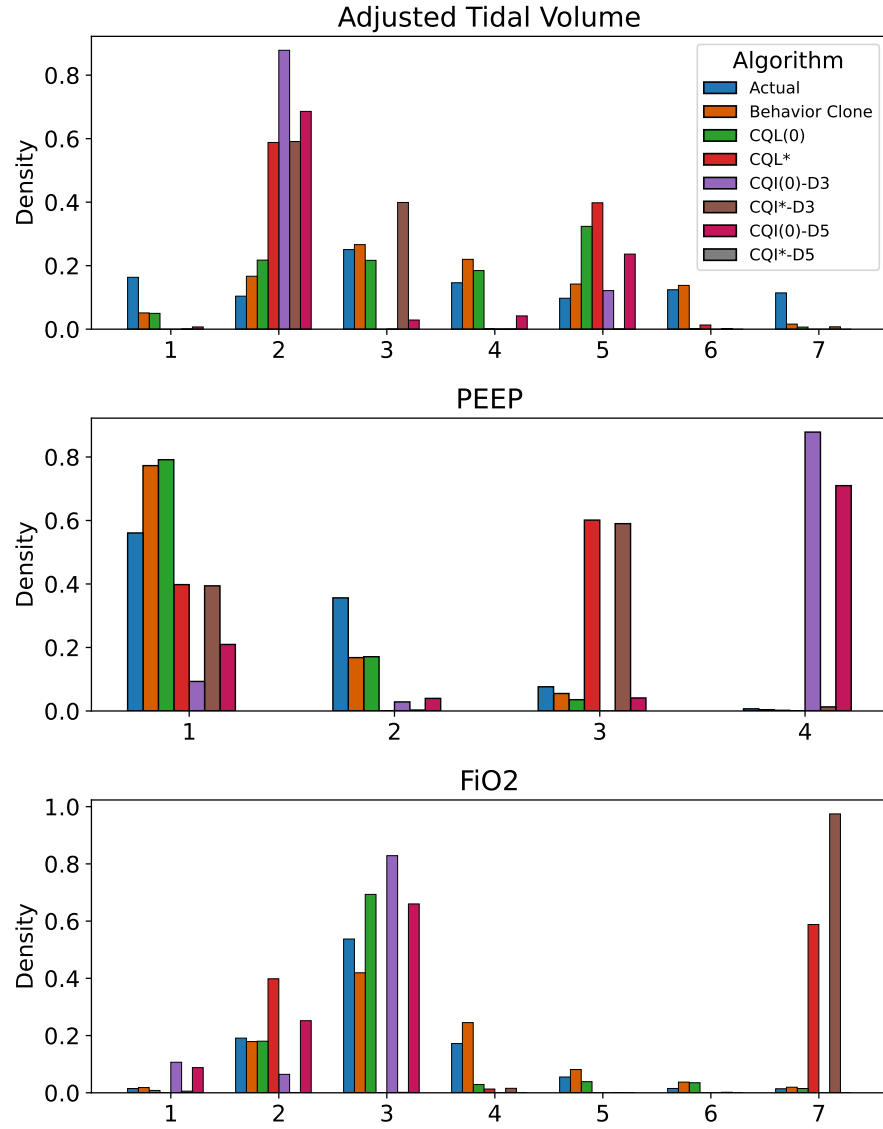


Figure 2.7: Distribution of actions across different RL and BC policies. The x-axis refers to action bins introduced in Table A.2

as evidenced by 0% aggressiveness in both tidal volume ($V_{t_{\text{set}}} \geq 10$ mL/kg) and fraction of inspired oxygen ($\text{FiO}_2 \geq 0.6$). This outcome reflects the structural constraints imposed by the limited number of decision tree leaves and the explicit penalty in the reward function, which collectively encourage the selection of non-aggressive actions across all leaves.

In contrast, CQL policies, which leverage a more flexible neural network architecture, achieve a balance between reducing aggressive actions and improving SpO_2 (Table 2.1). Among these, the regularized π_{CQL^*} stands out, achieving the lowest frequency of aggressive actions while maintaining a positive $\Delta(\text{SpO}_2)$ for patients whose terminal SpO_2 falls below 95%. This outcome highlights the ability of π_{CQL^*} to optimize oxygenation without relying excessively on high-risk settings. However, the unregularized $\pi_{\text{CQL}(0,0)}$, while maintaining a positive $\Delta(\text{SpO}_2)$, exhibits higher aggressiveness in FiO_2 , with 37.68% of time steps classified as aggressive. This trade-off underscores the importance of regularization in achieving a clinically appropriate balance between performance and safety.

The two policies π_{CQI^*-D3} and π_{CQI^*-D5} , while interpretable due to their decision tree structure, exhibit certain limitations. Both policies achieve 0% aggressiveness due to the combined effect of their shallow tree depth and the reward penalties, which effectively constrain their action space. However, this strict avoidance of aggressive actions comes at the expense of performance in improving SpO_2 for critical patients, as evidenced by their negative $\Delta(\text{SpO}_2)$. The negative values indicate that these policies are less effective at enhancing oxygenation in patients who require aggressive interventions.

Interestingly, $\pi_{\text{CQI}(0,0)-D5}$ achieves 100% of episodes terminating with $\text{SpO}_2 > 95\%$, the highest among all policies. However, this comes at the cost of relying on a broader range of actions, including those that may not align with the penalty structure used in π_{CQI^*} . By contrast, π_{CQI^*-D5} achieves comparable results (99.70%) while strictly adhering to the non-aggressiveness constraint, demonstrating that interpretability and safety can coexist, albeit with some trade-offs in $\Delta(\text{SpO}_2)$.

The Cost of Interpretability

The progression from CQI-D3 to CQI-D5 to CQL reflects a fundamental trade-off in RL policy design: increasing model complexity enhances performance but reduces interpretability. CQI-D3, the shallowest decision tree, provides simple, threshold-based rules but struggles with complex scenarios, achieving a negative $\Delta(\text{SpO}_2)$. Its strict avoidance of aggressive ventilator settings ensures safety but sacrifices performance in critical cases requiring aggressive interventions.

Increasing the tree depth to CQI-D5 marginally improves performance and retains interpretability relative to more complex models. CQI(0,0)-D5 achieves nearly perfect % $\text{SpO}_2 > 95$ outcomes without aggressive $V_{t_{\text{set}}}$ actions, but has severe aggressiveness in FiO_2 strategy. Also, π_{CQI^*-D5} has the worst outcome in $\Delta(\text{SpO}_2)$.

The deep neural network policy (CQL) excels in $\Delta(\text{SpO}_2)$ and minimizing aggressive actions due to its ability to model complex, nonlinear relationships. However, this performance gain comes at the cost of transparency, as its decisions are not readily explainable.

Decision Tree

The decision tree presented in Figure 2.5 represents a distilled policy learned via reinforcement learning (RL) from mechanical ventilation data documented in the electronic health record, guided by a custom reward structure. While some decision level splits were associated with clinical reasoning typically used at the bedside, many decision level splits were based on data demarcations that are not consistent with expected measures of sepsis physiology.

A striking observation is that the policy heavily relies on PaCO_2 as its primary decision node, subdividing patients into those with very low PaCO_2 (<19.5 mmHg), moderate PaCO_2 (19.5–92.1 mmHg), and very high PaCO_2 (≥ 92.1 mmHg). While titrating the ventilator’s respiratory rate and $V_{t\text{set}}$ to goal PaCO_2 levels is an important part of mechanical ventilator management, clinicians typically titrate PaCO_2 in sepsis management to far different ranges of PaCO_2 [91]. In contrast, the other two influential split features— PaO_2FiO_2 Ratio and mean airway pressure, commonly used to assess the severity of acute respiratory distress—yield split values more in line with clinical decision making at the bedside [98, 68].

Inspection of the terminal nodes shows that the RL policy selects distinct tidal volumes (ranging from approximately 3.4 to 8.3 mL/kg), PEEP values (5–12.5 cmH₂O), and FiO_2 (0.4–0.5), depending on the interplay between CO_2 levels, oxygenation (PaO_2FiO_2 Ratio), and hemodynamic status (systolic blood pressure). In scenarios with very high mean airway pressure and moderately elevated PaCO_2 , the policy increases both tidal volume and PEEP. A plausible clinical rationale is that high PEEP may recruit alveoli when the lung mechanics are poor, whereas a larger tidal volume might aid in CO_2 clearance [46]. Conversely, when mean airway pressure is low and PaCO_2 is moderate, the policy opts for a markedly reduced tidal volume, suggesting an emphasis on lung-protective ventilation and prevention of overdistension. In the case of extremely high PaCO_2 (≥ 92.1 mmHg), the tree further partitions patients by systolic blood pressure. Higher PEEP (10 cmH₂O) is recommended only when systolic blood pressure is adequate (≥ 96 mmHg). This indicates a learned behavior that offsets the risk of compromised venous return and hypotension often associated with increased intrathoracic pressure [35]. When blood pressure is lower, the tree prescribes moderate tidal volume (7.1 mL/kg) and minimal PEEP (5 cmH₂O), presumably to avoid hemodynamic deterioration.

The policy thus appears to balance hypercapnia and hypoxemia by adapting tidal volume and PEEP based on PaCO_2 and PaO_2FiO_2 Ratio, and it distinguishes heavier patients (above 105 kg) for potential differences in lung compliance or the need to approximate ideal body weight. Hemodynamic considerations also play a role when PaCO_2 is extremely high, as indicated by a separate decision node for systolic blood pressure, demonstrating the RL model’s capacity to integrate respiratory support with cardiovascular stability. These patterns illustrate how interpretable RL policy representations can provide insight into data-driven ventilator adjustments that incorporate both oxygenation and ventilation goals, while accounting for body habitus and systemic perfusion pressures.

2.6 Conclusion

In conclusion, we present an interpretable RL policy for mechanical ventilation control and a novel causal non-parametric model-based approach for evaluating RL policies in an offline setting. By comparing various policies (behavior cloning, CQI and CQL), we empirically establish the superior performance of the latter two policies over the behavior cloning of the clinicians as observed in MIMIC-III database in two ways: achieving greater increases in SpO_2 on average while taking aggressive actions less frequently. This contributes to the broader goals of connected health by providing interpretable decision-making systems that can integrate into ICU monitoring systems and enhance clinical workflows.

However, it is crucial to acknowledge the inherent limitations of offline RL and offline off-policy evaluation (OPE) methods, which may suffer from issues such as data bias, distribution shift, and lack of exploration. Although the model may outperform the clinicians for the goals set in this study, the goals set by the clinicians are unknown as the clinical context is not included in the predictors, as in previous RL approaches to control ventilators. Therefore, while the results are promising, additional verification on more diverse datasets outside of MIMIC-III is necessary before clinical deployment. Nevertheless, the interpretable decision tree policies learned by CQI have the potential to provide useful clinical insights in future studies by highlighting the important features and decision rules. Moving forward, integrating more domain knowledge and considering safety constraints will be crucial steps towards developing trustworthy RL systems for safety-critical healthcare applications like mechanical ventilation control.

Chapter 3

Game-Theoretic Incentive Design for Sustainable Groundwater Management

3.1 Introduction

The increasing global demand for water, coupled with unsustainable groundwater extraction, has led to a critical challenge in resource management. Groundwater sources are being depleted at an alarming rate, with potentially irreversible consequences, including land subsidence, ecosystem degradation, and reduced agricultural yields [12]. Environmental shifts exacerbate the problem by decreasing surface water availability, forcing water consumers to rely even more on groundwater. Without effective regulation, this over-extraction can lead to severe long-term economic and environmental damage.

Despite the urgency of groundwater management, existing regulations have struggled to prevent excessive usage. One widely discussed strategy is irrigation control, which seeks to limit water withdrawal through technological interventions. However, this has proven ineffective due to the “paradox of irrigation efficiency” [44] (or more generally the Jevons paradox): when farmers improve irrigation efficiency, they often expand irrigated land or switch to more water-intensive crops, offsetting conservation benefits. A similar phenomenon, known as induced demand, occurs in traffic engineering, where expanding road capacity to alleviate congestion encourages more driving, negating intended benefits. In both cases, improving efficiency in resource use leads to behavioral adaptation that sustains or even exacerbates overall consumption. As a result, groundwater depletion remains largely a consequence of poor management rather than a fundamental shortage of water.

To address this issue, economic instruments such as taxation have been proposed as a solution. Historically, groundwater has been treated as a free resource under the “Absolute Ownership Doctrine”, but there is increasing support for introducing a tax to regulate its use [58, 43]. This aligns with Pigouvian taxation, where users are charged to internalize the

negative externalities of resource depletion [81]. Empirical evidence from the Pajaro Valley, the first region in the United States to implement a groundwater tax, has shown promising results with a 22% reduction in water usage over five years [17]. Revenue from this tax has been reinvested into groundwater restoration projects, such as rainwater capture systems and water recycling plants [28].

However, implementing a groundwater tax presents significant challenges. Measuring individual water usage accurately is impractical, and water consumers, as rational agents, may have incentives to underreport consumption to minimize their tax burden [71]. Enforcing strict monitoring through physical measurements, such as tracking aquifer levels on private property, may be met with resistance or evasion. Illegal extraction is another concern, making effective taxation a complex problem that requires careful incentive design. Nevertheless, advancements in digital transformation technologies, such as IoT-enabled water meters and aerial drones, offer promising avenues for more accurate groundwater monitoring.

3.1.1 Differential Taxation as a Policy Mechanism

A single, uniform tax rate on groundwater usage may not adequately address water consumers' strategic behavior or incentivize voluntary compliance. In response, we explore a differential tax rate scheme where incentives encourage truthful reporting and sustainable water use. Differential taxation is a well-established concept widely applied in various policy domains. For example, congestion pricing adjusts toll rates based on time-of-day or vehicle type to optimize road usage, effectively reducing traffic congestion by encouraging off-peak travel [93]. Similarly, excise taxes on goods such as tobacco and sugary drinks are structured to discourage excessive consumption by imposing higher rates on products with greater social costs [20]. Inspired by these mechanisms, we propose a dual-tax scheme where users who adopt transparent water usage practices are taxed at a lower rate, while others face a higher penalty.

3.1.2 Our Contributions and Chapter Outline

This chapter presents a model where a public water agency sets groundwater tax rates while accounting for water consumers' strategic behavior. Specifically, we model the interaction between the regulatory agency and water consumers as a Stackelberg game (which is expressed as a bilevel optimization problem), where the agency aims to maximize water conservation while water consumers optimize agricultural profits under imperfect monitoring conditions. We analyze the resulting equilibrium and derive closed-form conditions under which the differential tax implements the first-best (centralized planner's) outcome, even when legal or enforcement constraints bind.

The rest of this chapter is structured as follows: Section 2 develops the formal model of the groundwater taxation problem. Section 3 presents the solution methodology for computing optimal tax rates. Section 4 provides characteristics of the impact of tax policy on

water consumers' responses. Section 5 presents numerical experiments illustrating the effectiveness of differential taxation. Sections 6–8 then develop analytical results on first-best implementation, the price of bounded penalties, and the substitution between monitoring capacity and penalty headroom.

3.2 The Model

We formulate the groundwater taxation problem as a Stackelberg game between a public water agency (leader) and water consumers (followers). The agency aims to regulate groundwater extraction through a differential tax scheme, while water consumers strategically decide their water usage and whether to comply with taxation rules. The interaction is modeled over a discrete multi-time-step horizon, incorporating the dynamics of the aquifer and the economic incentives of the agents.

3.2.1 Water Consumers' Usage and Costs

We adopt the notation from [11] to model the water demand and extraction costs of a representative water consumer. Given a collective of N homogeneous water consumers, the demand for water follows a linear price-demand function:

$$w = \alpha - \beta p, \quad (3.1)$$

where p is the price per unit of water, and $\alpha, \beta > 0$ are parameters capturing the demand elasticity. Let the total cost of extraction be a function of water table H and water pumped w :

$$C(w, H) = (c_0 - c_1 H)w \quad (3.2)$$

where $c_0, c_1 > 0$ represent base and depth-dependent extraction costs, respectively. A lower water table increases extraction costs due to increased energy requirements.

3.2.2 The Water Agency

To regulate groundwater consumption, the water agency imposes a tax on water withdrawal. Unlike existing approaches that rely on a single tax rate [11], we propose a dual tax system in which each water consumer decides whether or not to install a water meter. Let $i \in \{0, 1\}$ denote this binary decision: For $i = 0$, water consumers install water meters, truthfully reporting usage, and are subject to tax rate τ_0 . For $i = 1$, water consumers do not install water meters and face the higher tax rate τ_1 , as their true extraction cannot be verified.

Water consumers who do not install water meters ($i = 1$) may underreport usage by an evasion share $\theta \in [0, 1]$, the proportion of water they choose to not report. The agency, in

turn, conducts monitoring efforts to detect evasion attempts. If a water consumer is caught, they face a sanction multiplier $\sigma > 1$, applied to the unpaid tax. The detection probability ϕ is given by $\phi = \eta\theta$, where $\eta \in (0, 1]$ represents the agency's monitoring effort. The expected cost of tax evasion for a water consumer with $i = 1$ is $-(1 - \theta)\tau_1 w - \phi \cdot \theta\tau_1 w \cdot \sigma$, where the first term represents the actual tax paid, and the second term is the expected sanction for underreporting.

3.2.3 Aquifer Dynamics

Groundwater extraction affects the water table level over time. We model evolution of H_t using the standard aquifer dynamics equation from [42]:

$$H' = \frac{1}{K_a}[R - (1 - \gamma)W - \mu(H - H_{\min})], \quad (3.3)$$

where R is the natural recharge rate of the aquifer, $\gamma \in (0, 1)$ is the return flow coefficient, K_a is the product of aquifer area and storativity, $\mu = \frac{R}{H_{\max} - H_{\min}}$ is the natural discharge coefficient, and $H_{\min}$ is the minimum water table level.

For homogeneous water consumers, total extraction is $W_t = Nw_t$, leading to the discrete-time update rule:

$$H_t - H_{t-1} = \frac{1}{S_a}[R - (1 - \gamma)Nw_t - \mu(H_{t-1} - H_{\min})] \quad (3.4)$$

where $S_a = K_a/\Delta t$ and Δt is the time between consecutive periods, and H_0 is the given initial water table level. This captures the depletion of groundwater as a function of extraction and recharge rates.

3.2.4 Water Consumers' Optimization Problem

Each water consumer maximizes net profit while deciding: how much water to pump each period w_t , whether to install a water meter $i \in \{0, 1\}$, and if unmetered, how much to

evade taxation θ_t . Each water consumer solves the problem:

$$\begin{aligned}
& \max_{\mathbf{w}, \mathbf{H}, \boldsymbol{\theta}, i} \mathcal{P}(\mathbf{w}, \mathbf{H}, \boldsymbol{\theta}, i; \tau_0, \tau_1) \\
& = \max_{\mathbf{w}, \mathbf{H}, \boldsymbol{\theta}, i} \sum_{t=1}^T \frac{\alpha}{\beta} w_t - \frac{1}{2\beta} w_t^2 - (c_0 - c_1 H_t) w_t - (1 - i) \tau_0 w_t - i \cdot [(1 - \theta_t) \tau_1 w_t + \eta \theta_t^2 \sigma \tau_1 w_t] \\
& \text{s.t.} \quad H_t - H_{t-1} = \frac{1}{S_a} [R - (1 - \gamma) N w_t - \mu (H_{t-1} - H_{\min})] \\
& \quad 0 \leq \theta_t \leq 1 \\
& \quad H_t \geq 0 \\
& \quad w_t \geq 0 \\
& \quad i \in \{0, 1\},
\end{aligned} \tag{3.5}$$

where $\mathbf{w}, \mathbf{H}, \boldsymbol{\theta}$ are each $(w_1, \dots, w_T)$, $(H_1, \dots, H_T)$, and $(\theta_1, \dots, \theta_T)$. The objective function includes the revenue (first two terms), pumping cost (third term), tax payment for truthful water consumers (fourth term), and tax and expected penalty for untruthful water consumers (fifth term).

3.2.5 Water Agency's Optimization Problem

The agency sets the tax rates (τ_0, τ_1) to maximize social welfare, which consists of social surplus (SS), which is the profit from legal portion of water extraction, and environmental damage (ED), the negative impact of excessive water withdrawal proportional to the change in volume of aquifer. Following [39] and [11], these are:

$$\begin{aligned}
\text{SS}_t &= \left[\frac{\alpha}{\beta} w_t^* - \frac{1}{2\beta} \cdot w_t^{*2} - (c_0 - c_1 H_t^*) w_t^* \right] \cdot (1 - \theta_t^*) N \\
\text{ED}_t &= [(1 - \gamma) w_t^* N - R] \delta,
\end{aligned} \tag{3.6}$$

where w_t^*, θ_t^* are the water consumers' best response decision given the tax rates. The agency's optimization problem is:

$$\begin{aligned}
& \max_{\tau_0, \tau_1} \sum_{t=1}^T \text{SS}_t(\mathbf{w}^*, \mathbf{H}^*, \boldsymbol{\theta}^*, i^*) - \text{ED}_t(\mathbf{w}^*) \\
& \text{s.t.} \quad \{\mathbf{w}^*, \mathbf{H}^*, \boldsymbol{\theta}^*, i^*\} = \arg \max_{(\mathbf{w}, \mathbf{H}, \boldsymbol{\theta}, i) \in \mathcal{X}} \mathcal{P}(\mathbf{w}, \mathbf{H}, \boldsymbol{\theta}, i; \tau_0, \tau_1) \\
& \quad \tau_0, \tau_1 \geq 0,
\end{aligned} \tag{3.7}$$

where $\mathcal{X}$ is the set of constraints in problem (3.5). This represents a Stackelberg game, where:

1. The water agency (leader) selects tax rates to optimize social welfare.
2. Water consumers (followers) respond by choosing their water usage, reporting, and evasion strategy.

By optimizing (τ_0, τ_1) , the agency seeks to reduce environmental harm while preserving economic viability.

3.3 Analysis of the Optimal Solution

In this section, we give a single-level reformulation of the bilevel optimization problem (3.7). To solve the bilevel problem, we first decompose it into two distinct cases based on the water consumers' reporting decision i^* , which determines whether they install a water meter and truthfully report their water consumption.

In the first case, where $i^* = 1$, the water consumer does not install a meter and may engage in tax evasion. The water consumer's optimization problem then includes an additional evasion proportion θ_t , which determines the fraction of extracted water that is underreported. If the evasion is detected, a penalty is imposed based on the monitoring effort η and the sanction multiplier σ . At each time step, any detected underreporting immediately incurs the penalty. The water consumer's objective function in this case is:

$$\max_{\mathbf{w}, \mathbf{H}, \theta} F_1(\mathbf{w}, \mathbf{H}, \theta) = \sum_{t=1}^T \frac{\alpha}{\beta} w_t - \frac{1}{2\beta} w_t^2 - (c_0 - c_1 H_t) w_t - (1 - \theta_t) \tau_1 w_t - \eta \sigma \tau_1 \theta_t^2 w_t. \quad (3.8)$$

The second case, where $i^* = 0$, corresponds to the scenario where the water consumer installs a meter and truthfully reports their water usage. Here, tax evasion is not possible, and the optimization problem simplifies to:

$$\max_{\mathbf{w}, \mathbf{H}} F_0(\mathbf{w}, \mathbf{H}) = \sum_{t=1}^T \frac{\alpha}{\beta} w_t - \frac{1}{2\beta} w_t^2 - (c_0 - c_1 H_t) w_t - \tau_0 w_t. \quad (3.9)$$

In both cases, the water withdrawal decisions w_1 and w_0 are constrained by the aquifer dynamics, which follow the recurrence relation in (3.4).

To reformulate the bilevel problem as a single-level problem, we first establish that the water consumers' optimization problems are convex, allowing us to characterize their best responses using the Karush-Kuhn-Tucker (KKT) conditions.

Proposition 1. *When $i^* = 1$, the water consumers' subproblem reduces to a convex quadratic program, which is convex with respect to w_t .*

Proof. When $i^* = 1$, the optimal evasion share decision is $\theta_t^* = \min\{1/(2\eta\sigma), 1\}$, by observing that the objective is quadratic in θ_t and differentiating with respect to θ_t yields the

optimal evasion share. Substituting θ_t^* into the objective and rewriting H_t using its recurrence relation, $H_t = C_t - \frac{(1-\gamma)N}{S_a} \sum_{s=1}^t A^{t-s} w_s$, where $A = 1 - \frac{\mu}{S_a}$ and $C_t = A^t(H_0 - H_{\max}) + H_{\max}$, we obtain a quadratic optimization problem in w_t :

$$\begin{aligned} & \max_{\mathbf{w}} F_1(\mathbf{w}) \\ & = \max_{\mathbf{w}} \sum_{t=1}^T \left(\frac{\alpha}{\beta} + c_1 C_t - c_0 - \left(\frac{4\eta\sigma - 1}{4\eta\sigma} \right) \tau_1 \right) w_t - K \sum_{s=1}^t A^{t-s} w_s w_t - \frac{1}{2\beta} w_t^2, \end{aligned} \quad (3.10)$$

where $K = \frac{c_1(1-\gamma)N}{S_a}$. To establish convexity, we examine the Hessian $\mathcal{H}$ of F_1 . The second derivative with respect to w_t is given by:

$$\mathcal{H}_{ij} = \begin{cases} -\frac{1}{\beta} - 2K & \text{if } i = j \\ -KA^{|i-j|} & \text{if } i \neq j \end{cases} \quad (3.11)$$

Define a matrix M by $M_{ij} = A^{|i-j|}$. Since $A \in (0, 1)$, the matrix M is positive semidefinite, as it corresponds to the covariance matrix of an AR(1) process with autocorrelation A . Then, $\mathcal{H} = -\frac{1}{\beta}I - KI - KM$. It follows immediately that $\mathcal{H}$ is negative definite, confirming that F_1 is concave and the problem is convex. $\square$

Corollary 1. *By applying the KKT conditions to the water consumers' subproblem when $i^* = 1$, we obtain the following equations, which are necessary and sufficient for optimality:*

$$\begin{aligned} & \frac{w_t}{\beta} + \left(c_0 - c_1 C_t - \frac{\alpha}{\beta} + \left(\frac{4\eta\sigma - 1}{4\eta\sigma} \right) \tau_1 \right) \\ & + K \left(w_t + \sum_{s=1}^T A^{|s-t|} w_s \right) - u_{1,t}^{(1)} + \frac{K}{c_1} \sum_{s=t}^T A^{s-t} u_{2,s}^{(1)} = 0 \\ & u_{1,t}^{(1)} \cdot w_t = 0, \\ & u_{2,t}^{(1)} \cdot \left(C_t - \frac{(1-\gamma)N}{S_a} \sum_{s=1}^t A^{t-s} w_s \right) = 0, \\ & u_{1,t}^{(1)}, u_{2,t}^{(1)} \geq 0, \forall t = 1, \dots, T, \end{aligned} \quad (3.12)$$

where $u_{1,t}^{(1)}$ and $u_{2,t}^{(1)}$ are the dual variables associated with nonnegativity constraints on w_t and H_t , respectively. The KKT conditions are necessary and sufficient by Slater's condition, since the problem is concave maximization over a nonempty polyhedral feasible region by Proposition 1.

Proposition 2. *When $i^* = 0$, the water consumers' subproblem also reduces to a quadratic program and is convex in w_t .*

Proof. By substituting out H_t , the objective simplifies to

$$\begin{aligned}
& \max_{\mathbf{w}} F_0(\mathbf{w}) \\
&= \max_{\mathbf{w}} \sum_{t=1}^T \frac{\alpha}{\beta} w_t - \frac{1}{2\beta} w_t^2 - (c_0 - c_1 H_t) w_t - \tau_0 \cdot w_t \\
&= \max_{\mathbf{w}} \sum_{t=1}^T \left[\left(\frac{\alpha}{\beta} - c_0 - \tau_0 + c_1 C_t \right) w_t - \frac{1}{2\beta} w_t^2 - K \sum_{s=1}^t A^{t-s} w_s w_t \right].
\end{aligned} \tag{3.13}$$

The Hessian is identical to (3.11), establishing convexity. $\square$

Corollary 2. *The following are necessary and sufficient for optimality of the water consumers' subproblem when $i^* = 0$:*

$$\begin{aligned}
& \frac{w_t}{\beta} + \left(c_0 - c_1 C_t - \frac{\alpha}{\beta} + \tau_0 \right) \\
& \quad + K \left(w_t + \sum_{s=1}^T A^{|s-t|} w_s \right) - u_{1,t}^{(0)} + \frac{K}{c_1} \sum_{s=t}^T A^{s-t} u_{2,s}^{(0)} = 0 \\
& u_{1,t}^{(0)} \cdot w_t = 0, \\
& u_{2,t}^{(0)} \cdot \left(C_t - \frac{(1-\gamma)N}{S_a} \sum_{s=1}^t A^{t-s} w_s \right) = 0, \\
& u_{1,t}^{(0)}, u_{2,t}^{(0)} \geq 0, \forall t = 1, \dots, T.
\end{aligned} \tag{3.14}$$

Proof. The derivation and equivalence are similar to that of Corollary 1. $\square$

Having established convexity, we now transform the bilevel optimization problem into a single-level formulation by substituting the KKT conditions for both reporting cases into the water agency's problem.

Theorem 1. *For $M > 0$ sufficiently large, the bilevel optimization problem in (3.7) is equiv-*

alently expressed as the single-level problem below:

$$\begin{aligned}
\max \quad & \sum_{t=1}^T SS_t(\mathbf{w}', \boldsymbol{\theta}') - ED_t(\mathbf{w}') \\
s.t. \quad & \frac{w_{t,1}}{\beta} + \left(c_0 - c_1 C_t - \frac{\alpha}{\beta} + \left(\frac{4\eta\sigma - 1}{4\eta\sigma} \right) \tau_1 \right) \\
& + K \left(w_{t,1} + \sum_{s=1}^T A^{|s-t|} w_{s,1} \right) - u_{1,t}^{(1)} + \frac{K}{c_1} \sum_{s=t}^T A^{s-t} u_{2,s}^{(1)} = 0 \\
& u_{1,t}^{(1)} \leq M z_{1,t}^{(1)} \\
& w_{t,1} \leq M(1 - z_{1,t}^{(1)}) \\
& u_{2,t}^{(1)} \leq M z_{2,t}^{(1)} \\
& C_t - \frac{(1-\gamma)N}{S_a} \sum_{s=1}^t A^{t-s} w_{s,1} \leq M(1 - z_{2,t}^{(1)}) \\
& C_t - \frac{(1-\gamma)N}{S_a} \sum_{s=1}^t A^{t-s} w_{s,1} \geq 0 \\
& \frac{w_{t,0}}{\beta} + \left(c_0 - c_1 C_t - \frac{\alpha}{\beta} + \tau_0 \right) \\
& + K \left(w_{t,0} + \sum_{s=1}^T A^{|s-t|} w_{s,0} \right) - u_{1,t}^{(0)} + \frac{K}{c_1} \sum_{s=t}^T A^{s-t} u_{2,s}^{(0)} = 0 \\
& u_{1,t}^{(0)} \leq M z_{1,t}^{(0)} \\
& w_{t,0} \leq M(1 - z_{1,t}^{(0)}) \\
& u_{2,t}^{(0)} \leq M z_{2,t}^{(0)} \\
& C_t - \frac{(1-\gamma)N}{S_a} \sum_{s=1}^t A^{t-s} w_{s,0} \leq M(1 - z_{2,t}^{(0)}) \\
& C_t - \frac{(1-\gamma)N}{S_a} \sum_{s=1}^t A^{t-s} w_{s,0} \geq 0
\end{aligned} \tag{3.15}$$

$$\begin{aligned}
w_{t,1} - M \cdot (1 - i) &\leq w'_t \leq w_{t,1} + M \cdot (1 - i) \\
w_{t,0} - M \cdot i &\leq w'_t \leq w_{t,0} + M \cdot i \\
&\text{for all } t = 1, \dots, T \\
u_{j,t}^{(k)} &\geq 0, z_{j,t}^{(k)} \in \{0, 1\} \text{ for all } k \in \{1, 0\}, j \in \{1, 2\}, t \in [T] \\
w_{t,k} &\geq 0 \quad \text{for all } k \in \{1, 0\}, t \in [T] \\
\theta' &= i \cdot \min \{1/(2\eta\sigma), 1\} \\
i &\in \{0, 1\} \\
F_0(\mathbf{w}_0) - F_1(\mathbf{w}_1) &\leq M \cdot (1 - i) \\
F_1(\mathbf{w}_1) - F_0(\mathbf{w}_0) &\leq M \cdot i
\end{aligned} \tag{3.16}$$

Proof. To ensure the principal (i.e., water agency) considers the correct reporting decision made by water consumers, we use the big-M method to enforce the correct value of i :

$$\begin{aligned}
F_0(\mathbf{w}_0) - F_1(\mathbf{w}_1) &\leq M \cdot (1 - i) \\
F_1(\mathbf{w}_1) - F_0(\mathbf{w}_0) &\leq M \cdot i \\
i &\in \{0, 1\}
\end{aligned} \tag{3.17}$$

This ensures that the chosen reporting decision corresponds to the strategy that maximizes water consumer utility when M is sufficiently large. In (3.15), $\mathbf{w}_1$ and $\mathbf{w}_0$ represent the optimal water extraction decisions in each reporting case, and the leader (water agency) selects tax rates that maximize social welfare under the assumption of rational water consumer responses $\mathbf{w}'$. The rest of the constraints follow from the KKT conditions (3.12) and (3.14). $\square$

3.4 Characteristics of the Optimal Policy

Here, we analyze the structure of the optimal groundwater taxation policy. We first present theoretical results regarding the water consumers' reporting decision and the tax rate thresholds that drive different behavioral responses. We provide economic intuition and policy implications.

3.4.1 Formal Results on the Optimal Policy

Define the set $S = \{\mathbf{w}, \mathbf{H} : w_t \geq 0, H_t \geq 0, H_t - H_{t-1} = \frac{1}{S_a}[R - (1 - \gamma)Nw_t - \mu(H_{t-1} - H_{\min})]\}$. For a fixed τ_1 , increasing τ_0 makes water consumers more likely to choose self-reporting.

Proposition 3. *Let $i^*(\tau_0, \tau_1) \in \{0, 1\}$ be the binary evasion decision by the water consumers when given the two tax rates. Then for a fixed τ_1 , i^* is non-decreasing in τ_0 . That is, $i^*(\tau_{0,a}, \tau_1) \leq i^*(\tau_{0,b}, \tau_1)$ if $\tau_{0,a} \leq \tau_{0,b}$ for any $\tau_1 \geq 0$.*

Proof. Let $F_0^*(\cdot)$ and $F_1^*(\cdot)$ be defined by:

$$\begin{aligned} F_0^*(\tau_0) &= \max_{\mathbf{w}, \mathbf{H} \in S} F_0(\tau_0, \mathbf{w}, \mathbf{H}) = \max_{\mathbf{w} \in S} \sum_{t=1}^T \frac{\alpha}{\beta} w_t - \frac{1}{2\beta} w_t^2 - (c_0 - c_1 H_t) w_t - \tau_0 \cdot w_t \\ F_1^*(\tau_1) &= \max_{\mathbf{w}, \mathbf{H} \in S} \sum_{t=1}^T \frac{\alpha}{\beta} w_t - \frac{1}{2\beta} w_t^2 - (c_0 - c_1 H_t) w_t - (1 - \theta_t^*) \tau_1 w_t - \eta \sigma \tau_1 \theta_t^{*2} w_t. \end{aligned} \quad (3.18)$$

Since i is defined by

$$i = \begin{cases} 0 & \text{if } F_0^*(\tau_0) \geq F_1^*(\tau_1) \\ 1 & \text{if } F_0^*(\tau_0) < F_1^*(\tau_1) \end{cases}, \quad (3.19)$$

and $F_1^*(\tau_1)$ is fixed, it suffices to show $F_0^*(\tau_{0,a}) \geq F_0^*(\tau_{0,b})$ if $\tau_{0,a} \leq \tau_{0,b}$. Since $F_0(\tau_{0,a}, \mathbf{w}, \mathbf{H}) - F_0(\tau_{0,b}, \mathbf{w}, \mathbf{H}) = \sum_{t=1}^T (\tau_{0,b} - \tau_{0,a}) w_t$ for every feasible solution $\mathbf{w}, \mathbf{H} \in S$, and $\tau_{0,b} - \tau_{0,a}, w_t \geq 0$ by assumption and constraint respectively, by monotonicity we have $F_0^*(\tau_{0,a}) \geq F_0^*(\tau_{0,b})$. $\square$

Among the truthfully reporting water consumers, a higher τ_0 leads to lower optimal water extraction.

Proposition 4. *Let $w_t^*(\tau_0)$ be the optimal water consumption decision at time t by the water consumers who choose to self-report ($i = 0$) when presented the two tax rates (τ_0, τ_1) . Then $w_t^*(\tau_0)$ is non-increasing in τ_0 , meaning $w_t^*(\tau_{0,a}) \geq w_t^*(\tau_{0,b})$ if $\tau_{0,a} \leq \tau_{0,b}$ for any $\tau_1 \geq 0$ and $t = 1, \dots, T$.*

Proof. Due to the nonnegative water height constraint, the set of feasible $\mathbf{w}$ is bounded. By strict concavity of the objective and Berge's Maximum Theorem (BMT), $\mathbf{w}_t^*(\tau_0)$ exists and is unique. We rewrite the objective as

$$\begin{aligned} F_0^*(\tau_0) &= \max_{\mathbf{w}, \mathbf{H}} \sum_{t=1}^T \frac{\alpha}{\beta} w_t - \frac{1}{2\beta} w_t^2 - (c_0 - c_1 H_t) w_t - \tau_0 w_t \\ &= \max_{\mathbf{w}} \sum_{t=1}^T \left[\left(\frac{\alpha}{\beta} - c_0 - \tau_0 + c_1 C_t \right) w_t - \frac{1}{2\beta} w_t^2 - K \sum_{s=1}^t A^{t-s} w_s w_t \right]. \end{aligned} \quad (3.20)$$

For each t , let $f(w_t) = -\tau_0 w_t$ and let $g(w_1, \dots, w_t, \dots, w_T)$ include all remaining terms. Applying Corollary 2 of [3], $w_t^*(\tau_{0,a}) \geq w_t^*(\tau_{0,b})$. $\square$

Corollary 3 (Comparative Statics). *Fix τ_1 . As we decrease τ_0 , i^* is non-increasing and w_t^* is non-decreasing.*

Proof. The result follows from Propositions 3 and 4. $\square$

3.4.2 Reporting Decision

We show that the τ_0 threshold varies continuously with τ_1 .

Theorem 2. *Let $\tau'_0(\tau_1) = \max_{\tau_0} \{\tau_0 : i^*(\tau_0, \tau_1) = 0\}$. Then, for appropriate $\underline{\tau}_1$, $\tau'_0(\tau_1)$ is continuous on $[\underline{\tau}_1, \infty)$.*

Proof. We establish the continuity of $\tau'_0(\tau_1)$ by first identifying the domain where it is well-defined and then proving the continuity of the relevant functions.

We begin by defining the best response actions for each player, and practical upper bounds $\bar{\tau}_i$ for τ_i . For each i , define $w_i^*(\tau_i) = \arg \max_{\mathbf{w}_i \in S} F_i(\mathbf{w}_i, \tau_i)$, and the optimal value function when $i^* = 1$: $F_1^*(\tau_1) = F_1(\tau_1, \mathbf{w}_1^*(\tau_1))$. Define $\bar{\tau}_i$ by $\bar{\tau}_i = \min_{\tau_i} \{\tau_i : \max_{\mathbf{w}, \mathbf{H}} F_i(\tau_i, \mathbf{w}, \mathbf{H}) = 0\}$. By the definition of i^* , we express $\tau'_0(\tau_1) = \max_{\tau_0} \{\tau_0 : F_0^*(\tau_0) \geq F_1^*(\tau_1)\}$. Define auxiliary functions $V(\zeta, \mathbf{w}) = \max_{\tau_0} \{\tau_0 : F_0(\tau_0, \mathbf{w}) \geq \zeta, \bar{\tau}_0 \geq \tau_0 \geq 0\}$ and $W(\tau_1) = \max_{\mathbf{w} \in S} V(F_1^*(\tau_1), \mathbf{w})$. Note $\tau'_0(\tau_1)$ is precisely $W(\tau_1)$. We wish to show that F_1^* , V , and W are each continuous with respect to their arguments. Now, we determine the domain where $V(\zeta, \mathbf{w})$ is well-defined. Specifically, we find the interval $[0, \bar{\zeta}]$ such that the feasible set in the definition of V is nonempty. The upper bound is $\bar{\zeta} = \max_{\tau_1} \{F_1(\mathbf{w}_1^*(\tau_1), \tau_1) | \exists \tau_0 \geq 0 : F_0(\mathbf{w}_0^*(\tau_0), \tau_0) \geq F_1(\mathbf{w}_1^*(\tau_1), \tau_1)\}$. First, we note that $F_1^*(\tau_1)$ is continuous as $\max_{\mathbf{w}} F_1(\mathbf{w}; \tau_1)$ meets the conditions for BMT. Choosing $\underline{\tau}_1$ such that $F_1^*(\underline{\tau}_1) = \bar{\zeta}$ and noting that $F_1^*(\tau_1)$ is strictly decreasing and continuous in τ_1 , we conclude that $\underline{\tau}_1$ is unique. This ensures $V(\zeta, \mathbf{w})$ is well-defined for $\zeta \in [0, \bar{\zeta}]$.

Next, we establish continuity of V and W . The constraints are linear and the constraint set of $V(\zeta, \mathbf{w})$ is compact. From Lemma 4 of [105], the constraint set is jointly continuous in ζ and $\mathbf{w}$. Since $F_1^*(\tau_1)$ is continuous in τ_1 and $V(\zeta, \mathbf{w})$ is jointly continuous in $(\zeta, \mathbf{w})$, the function $W(\tau_1)$ is also continuous in τ_1 by BMT. $\square$

3.4.3 Economic Intuition Behind the Optimal Policy

Corollary 3 highlights a fundamental trade-off that the water agency must manage when designing tax policies. We prove that the threshold tax rate for truthful reporting is continuous, ensuring that small adjustments in taxation do not lead to abrupt shifts in reporting behavior, making the policy design process more stable and predictable.

Theorem 2 formalizes the idea that increasing τ_0 makes truthful reporting less attractive, since water consumers must pay more tax if they truthfully report their water usage. Consequently, some water consumers will opt out of installing meters and instead evade taxation. This highlights a fundamental policy trade-off: While a lower τ_0 encourages truthful reporting, it also increases water extraction among truthfully reporting water consumers. The water agency must balance these two competing effects to achieve the best social outcome.

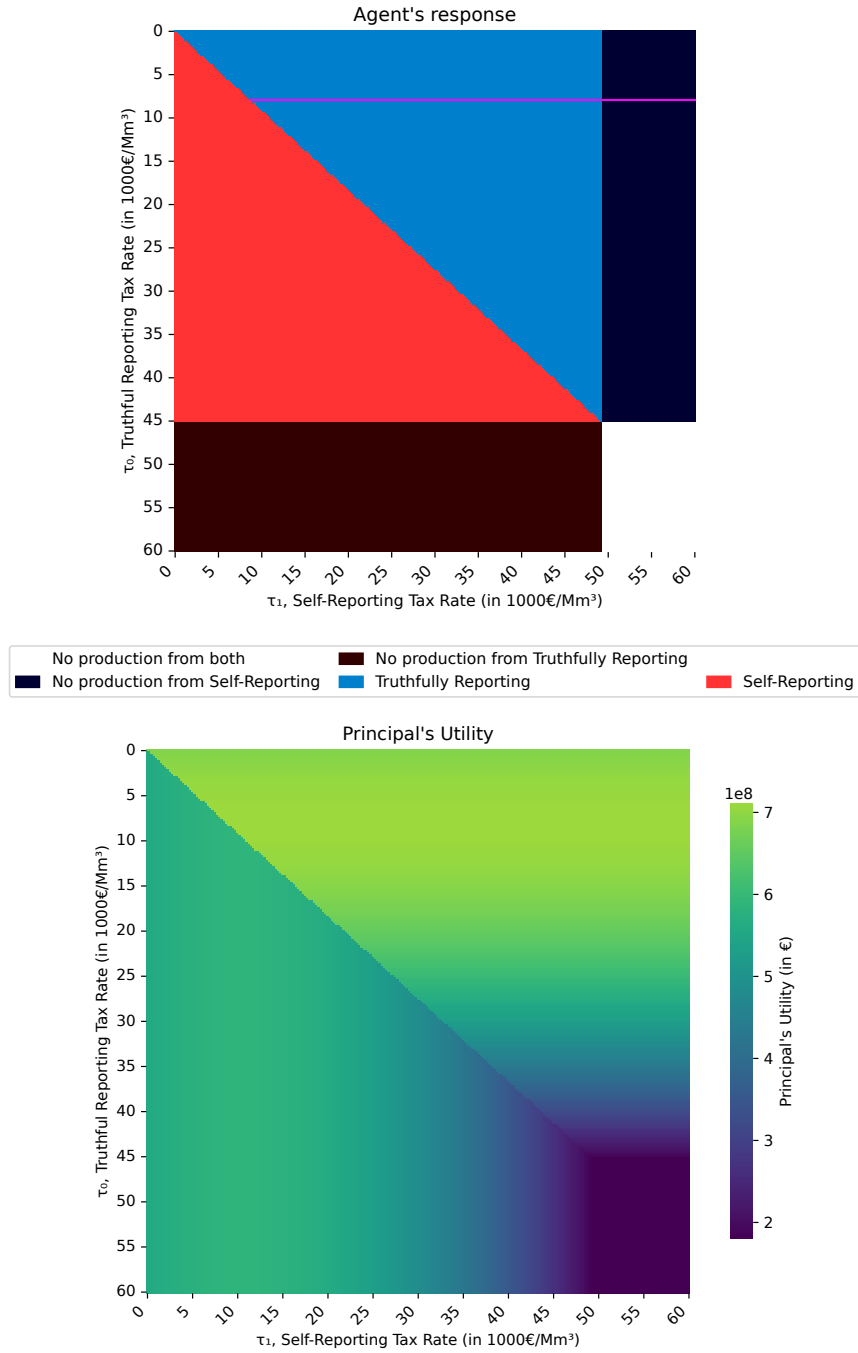


Figure 3.1: (Top) Partition of the (τ_0, τ_1) space showing water consumers' optimal decisions with the yellow line marking the combination that maximizes social welfare. (Bottom) The water agency's utility over the tax-rate space.

3.5 Numerical Experiment

We take parameter values from the Western La Mancha aquifer (extracted for agricultural purposes), a widely studied setting in the literature [39, 29, 80, 11]:

Params.	Description	Units	Value
α	Intercept of the water demand	€/Mm ³	4400.73
β	Slope of the water demand	€/Mm ³	0.097
c_0	Intercept of the pumping cost	€/Mm ³	266000
c_1	Slope of the pumping cost	€/(Mm ³ .m)	400
γ	Return flow coefficient	-	0.2
S_a	Aquifer area	Mm ²	126.5
R	Natural recharge	Mm ³	360
δ	Ecosystem Damage cost	€	10000
$H_0, H_{\max}$	Initial & Maximum height	m	665
$H_{\min}$	Minimum height	m	600
μ	Drainage term	Mm ² /yr	5.538
η	Monitoring effort	-	0.75
σ	Administrative sanction	-	4
N	Number of water consumers	-	100
T	Time period length	-	50

Table 3.1: Calibration parameters used in the numerical experiment, taken from prior studies of the Western La Mancha aquifer.

Simulations were performed using Gurobi with Python 3.9 on an Apple M4 Pro [45].

3.5.1 Agent Responses and Social Welfare

The top panel of Figure 3.1 illustrates how water consumers' reporting decisions i^* vary across tax rate combinations (τ_0, τ_1) . The plot partitions the decision space; the yellow line marks optimal tax combinations that induce truthful reporting ($i^* = 0$) in equilibrium. Tax rates outside this set may trigger evasion (self-reporting region), discourage production due to excessive taxation, or fail to maximize social welfare despite compliance.

The bottom panel of Figure 3.1 shows that social welfare declines when τ_0 is too low (over-extraction) or too high (production losses), with $\tau_0 = 8000$ maximizing welfare under the differential scheme. These results demonstrate that a differential tax scheme promotes compliance while balancing economic productivity and environmental sustainability.

3.5.2 Comparison with Single-Tax Regime

To benchmark our approach, we replicated our experiment under a uniform tax regime [11]. The optimal single tax rate was $\tau = 10,472.7$, yielding a water agency utility value of €585.0 million, which is lower than that of €710.4 million with $\tau_0 = 8000$ and $\tau_1 \geq 8,750$ achieved by our differential tax mechanism. A uniform tax loses 17.6% of welfare relative to the dual tax we propose, highlighting the effectiveness of flexible tax structures in aligning incentives for sustainable groundwater management.

It is also interesting to compare our taxation policy to the (unimplementable) policy with perfect monitoring. This centrally controlled regime yields (in this case) identical water consumption to the differential tax we propose, achieving the same social welfare. While the policy with perfect monitoring represents an upper bound on welfare, it is unimplementable; our implementable dual tax policy thus achieves this upper bound in this example.

3.6 First-Best Implementation

We formalize the observation from Section 5 that the differential tax mechanism can replicate the centralized planner's outcome. We first define the planner's problem, then show the differential tax achieves the same allocation.

3.6.1 The Centralized Planner's Problem

The social planner directly chooses water extraction $\mathbf{w} = (w_1, \dots, w_T)$ to maximize social welfare, defined as the aggregate social surplus minus environmental damage, subject to aquifer dynamics. Under full compliance ($\theta_t = 0$ for all t), the planner solves:

$$\begin{aligned}
 W^{FB} = & \\
 \max_{\mathbf{w}, \mathbf{H}} & \sum_{t=1}^T N \left[\frac{\alpha}{\beta} w_t - \frac{1}{2\beta} w_t^2 - (c_0 - c_1 H_t) w_t \right] \\
 & - [(1 - \gamma) N w_t - R] \delta \\
 \text{s.t.} & H_t - H_{t-1} = \frac{1}{S_a} [R - (1 - \gamma) N w_t - \mu(H_{t-1} - H_{\min})] \\
 & w_t \geq 0, \quad H_t \geq 0, \quad \forall t = 1, \dots, T.
 \end{aligned} \tag{3.21}$$

Since each term in the environmental damage $\text{ED}_t = [(1 - \gamma) N w_t - R] \delta$ is linear in w_t , the planner's objective can be decomposed on a per-consumer basis by dividing by N (noting $R\delta/N$ is a constant that does not affect the optimal $\mathbf{w}$):

$$\begin{aligned}
 \max_{\mathbf{w}, \mathbf{H}} & \sum_{t=1}^T \left[\frac{\alpha}{\beta} w_t - \frac{1}{2\beta} w_t^2 - (c_0 - c_1 H_t) w_t - (1 - \gamma) \delta \cdot w_t \right] \\
 \text{s.t.} & H_t - H_{t-1} = \frac{1}{S_a} [R - (1 - \gamma) N w_t - \mu(H_{t-1} - H_{\min})] \\
 & w_t \geq 0, \quad H_t \geq 0, \quad \forall t.
 \end{aligned} \tag{3.22}$$

3.6.2 Structural Equivalence

We now state and prove the main result. Recall that when $i^* = 0$ (meter installed, truthful reporting), each water consumer solves:

$$\max_{\mathbf{w}, \mathbf{H}} F_0(\mathbf{w}, \mathbf{H}) = \sum_{t=1}^T \frac{\alpha}{\beta} w_t - \frac{1}{2\beta} w_t^2 - (c_0 - c_1 H_t) w_t - \tau_0 w_t \quad (3.23)$$

subject to the same aquifer dynamics and nonnegativity constraints. When $i^* = 1$ (no meter, evasion), the optimal evasion share is $\theta^* = \min\{1/(2\eta\sigma), 1\}$ (Proposition 1), and the reduced objective is:

$$\begin{aligned} \max_{\mathbf{w}, \mathbf{H}} F_1(\mathbf{w}, \mathbf{H}) = \sum_{t=1}^T \frac{\alpha}{\beta} w_t - \frac{1}{2\beta} w_t^2 - (c_0 - c_1 H_t) w_t \\ - \underbrace{\frac{4\eta\sigma - 1}{4\eta\sigma}}_{\triangleq \lambda} \cdot \tau_1 \cdot w_t \end{aligned} \quad (3.24)$$

subject to the same constraints. Here, $\lambda = (4\eta\sigma - 1)/(4\eta\sigma) \in (0, 1)$ is the *effective tax rate coefficient*, representing the fraction of τ_1 that the evading consumer effectively bears after optimally evading a share θ^* .

We define the *common value function* that captures the optimal payoff as a function of any effective tax rate τ :

$$V^*(\tau) \triangleq \max_{(\mathbf{w}, \mathbf{H}) \in S} \sum_{t=1}^T \frac{\alpha}{\beta} w_t - \frac{1}{2\beta} w_t^2 - (c_0 - c_1 H_t) w_t - \tau \cdot w_t, \quad (3.25)$$

where S is the feasible set defined by aquifer dynamics, $w_t \geq 0$, and $H_t \geq 0$. By construction:

$$F_0^*(\tau_0) = V^*(\tau_0), \quad F_1^*(\tau_1) = V^*(\lambda\tau_1). \quad (3.26)$$

Lemma 1. *The common value function $V^*(\tau)$ is strictly decreasing, continuous, and convex in τ on $[0, \bar{\tau}]$, where $\bar{\tau} = \min\{\tau : V^*(\tau) = 0\}$. Moreover, the optimal extraction trajectory $\mathbf{w}^*(\tau)$ is continuous and componentwise strictly decreasing in τ on the interior.*

Proof. Continuity of $V^*(\tau)$ follows from Berge's Maximum Theorem, since the feasible set S is compact and independent of τ , and the objective is continuous and jointly concave. Strict decrease follows because for $\tau_a < \tau_b$, $V^*(\tau_a) \geq V^*(\tau_b) + (\tau_b - \tau_a) \sum_t w_t^*(\tau_b) > V^*(\tau_b)$ whenever $w_t^*(\tau_b) > 0$ for some t . Convexity of $V^*(\tau)$ follows from the fact that $V^*(\tau)$ is the pointwise maximum of a family of functions that are affine (hence convex) in τ . Continuity and monotonicity of $\mathbf{w}^*(\tau)$ follow from BMT and the strict concavity of the objective in $\mathbf{w}$ (established in Propositions 1 and 2). $\square$

Theorem 3 (First-Best Implementation). *Suppose $\eta\sigma > 1/2$ and the environmental damage function is linear, i.e., $\text{ED}_t = [(1 - \gamma)Nw_t - R]\delta$. Define:*

$$\tau_0^{FB} \triangleq (1 - \gamma)\delta \quad (3.27)$$

and define the compliance threshold:

$$\bar{\tau}_1 \triangleq \frac{4\eta\sigma}{4\eta\sigma - 1} \cdot \tau_0^{FB} = \frac{4\eta\sigma}{4\eta\sigma - 1} \cdot (1 - \gamma)\delta. \quad (3.28)$$

If the agency sets $\tau_0 = \tau_0^{FB}$ and $\tau_1 \geq \bar{\tau}_1$, then:

- (i) In equilibrium, all water consumers install meters: $i^* = 0$.
- (ii) The equilibrium extraction $\mathbf{w}^*$ solves the centralized planner's problem (3.21).
- (iii) The resulting social welfare equals the first-best: $\text{SW}(\mathbf{w}^*, i^* = 0) = W^{FB}$.

Proof. The proof proceeds in three steps.

Step 1: Matching the first-order conditions. Comparing the per-consumer planner's problem (3.22) with the compliant consumer's problem (3.23), the two objectives are identical when $\tau_0 = (1 - \gamma)\delta = \tau_0^{FB}$.

Crucially, both problems share the same feasible set S (aquifer dynamics, nonnegativity constraints), and both objectives are strictly concave in $\mathbf{w}$ by Proposition 1 (the Hessian (3.11) depends only on β , K , and A , none of which vary between the two problems). By strict concavity, the solution to each problem is unique. Since the objectives coincide, the optimal extraction trajectories coincide:

$$\mathbf{w}^*(\tau_0^{FB}) = \mathbf{w}^{FB}. \quad (3.29)$$

This establishes that *conditional on compliance* ($i^* = 0$), the equilibrium extraction replicates the planner's allocation.

Step 2: Enforcing compliance. It remains to show that water consumers prefer $i = 0$ to $i = 1$. By (3.26), the consumer chooses $i = 0$ if and only if:

$$V^*(\tau_0) \geq V^*(\lambda\tau_1). \quad (3.30)$$

Since V^* is strictly decreasing (Lemma 1), this holds if and only if $\tau_0 \leq \lambda\tau_1$, i.e.:

$$\tau_1 \geq \frac{\tau_0}{\lambda} = \frac{4\eta\sigma}{4\eta\sigma - 1} \cdot \tau_0. \quad (3.31)$$

Setting $\tau_0 = \tau_0^{FB}$ yields the threshold $\bar{\tau}_1$ defined in (3.28). For any $\tau_1 \geq \bar{\tau}_1$, the effective evasion tax $\lambda\tau_1 \geq \tau_0^{FB}$, so the consumer (weakly) prefers compliance. This establishes (i).

Step 3: Social welfare equivalence. Under $i^* = 0$, we have $\theta^* = 0$ and $\mathbf{w}^* = \mathbf{w}^{FB}$. The social welfare is:

$$\text{SW}(\mathbf{w}^{FB}, 0) = \quad (3.32)$$

$$\begin{aligned} & \sum_{t=1}^T N \left[\frac{\alpha}{\beta} w_t^{FB} - \frac{1}{2\beta} (w_t^{FB})^2 - (c_0 - c_1 H_t^{FB}) w_t^{FB} \right] \\ & - [(1 - \gamma) N w_t^{FB} - R] \delta = W^{FB}, \end{aligned} \quad (3.33)$$

where the last equality holds because this is precisely the planner's objective evaluated at the planner's optimal solution. This establishes (ii) and (iii). $\square$

Remark 1 (Relationship to Theorem 2). *Theorem 3 subsumes the continuity result of Theorem 2. The compliance threshold $\tau'_0(\tau_1) = \max\{\tau_0 : i^*(\tau_0, \tau_1) = 0\}$ studied in Theorem 2 can now be computed in closed form. Since V^* is strictly monotone (Lemma 1), the indifference condition $V^*(\tau_0) = V^*(\lambda\tau_1)$ yields:*

$$\tau'_0(\tau_1) = \lambda\tau_1 = \frac{4\eta\sigma - 1}{4\eta\sigma} \cdot \tau_1. \quad (3.34)$$

That is, the compliance boundary in the (τ_0, τ_1) space is a ray through the origin with slope λ . Continuity (and linearity) follows immediately, without recourse to Berge's Maximum Theorem or the auxiliary functions $V(\zeta, \mathbf{w})$ and $W(\tau_1)$ employed in the proof of Theorem 2. This closed form also provides an analytical validation of the partition geometry observed numerically in Figure 3.1 in the earlier numerical experiment: the boundary separating the compliance and evasion regions should be exactly this ray. Figure 3.2 overlays the ray, the Pigouvian rate τ_0^{FB} , and the compliance threshold $\bar{\tau}_1$ on the empirical partition, confirming that the numerical boundary coincides with the closed-form prediction.

In extensions where this closed form may not obtain (e.g., under convex damage functions where the optimal τ_0 becomes state-dependent or heterogeneous agents where the value functions are no longer related by a simple rescaling), the Berge's Maximum Theorem argument of Theorem 2 would recover continuity of the threshold in the general case.

Remark 2 (Role of $\eta\sigma$). *The assumption $\eta\sigma > 1/2$ ensures that the interior evasion solution $\theta^* = 1/(2\eta\sigma) < 1$ obtains, so that the effective tax coefficient $\lambda = (4\eta\sigma - 1)/(4\eta\sigma) \in (0, 1)$ is well-defined. If $\eta\sigma \leq 1/2$, then $\theta^* = 1$ (full evasion): the consumer reports zero water and pays no declared tax. However, the expected penalty is $\eta\sigma\tau_1\theta^{*2}w_t = \eta\sigma\tau_1w_t$, so the effective tax coefficient in this corner regime is $\lambda = \eta\sigma > 0$. Compliance can still be enforced by setting $\tau_1 \geq \tau_0^{FB}/(\eta\sigma)$, which is finite. Thus, $\eta\sigma > 1/2$ is not strictly necessary for the mechanism to function, but is assumed throughout to ensure the interior evasion solution and the closed-form expressions derived in Theorem 3.*

Remark 3 (Tinbergen's Rule Interpretation). *Theorem 3 can be understood through the lens of Tinbergen's Rule: The agency has two economic policy targets (internalizing the environmental externality and enforcing compliance) and two policy instruments (τ_0 and τ_1). The*

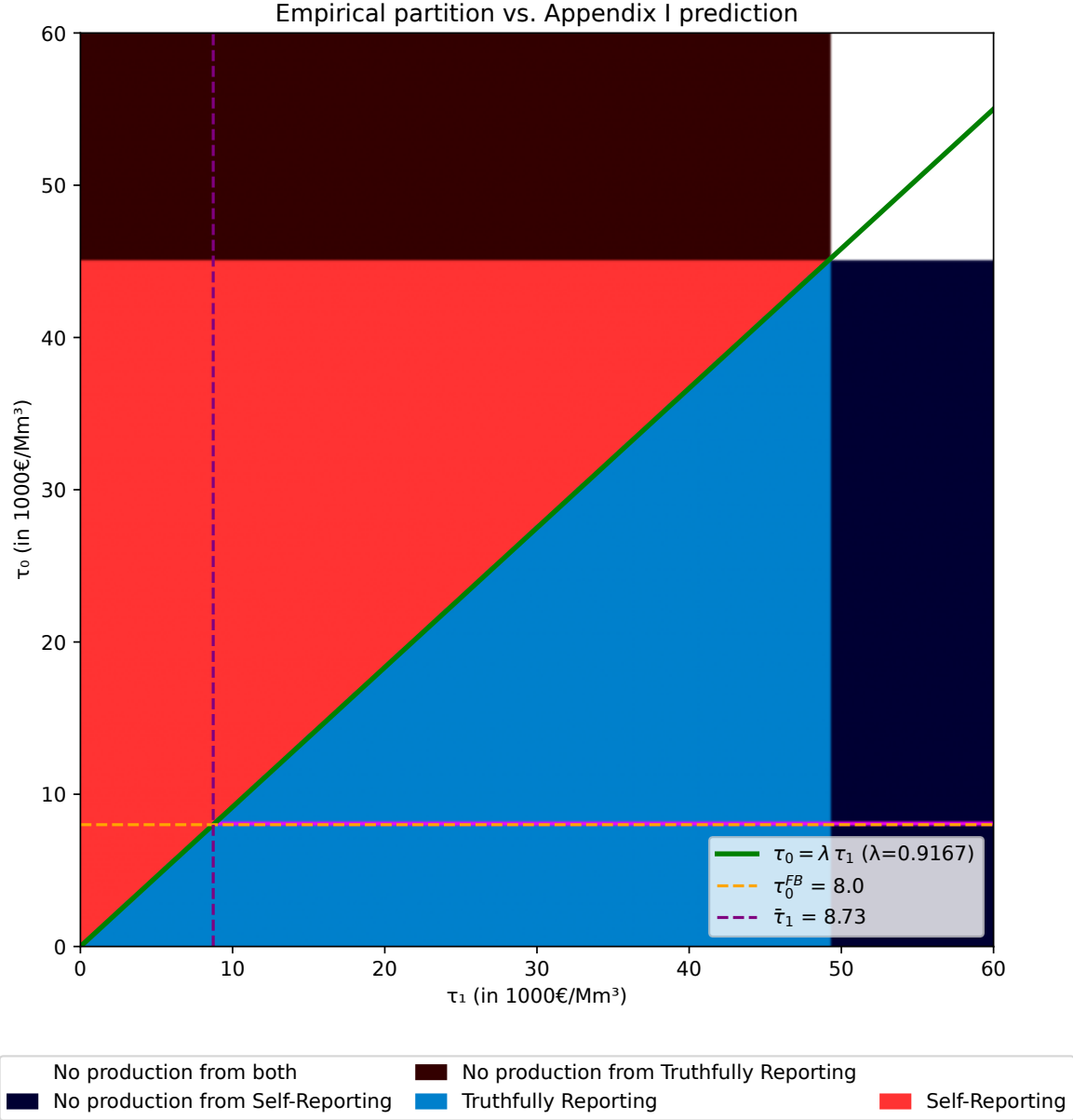


Figure 3.2: Numerical validation of Theorem 3 and Remark 1 for the Western La Mancha calibration ($\eta = 0.75$, $\sigma = 4$, giving $\lambda = 11/12 \approx 0.9167$). The colored regions reproduce the empirical partition from the top panel of Figure 3.1. The solid green line plots the closed-form compliance boundary $\tau_0 = \lambda \tau_1$; the dashed orange line marks the Pigouvian rate $\tau_0^{FB} = (1 - \gamma)\delta = 8,000 \text{ €/Mm}^3$; the dashed purple line marks the compliance threshold $\bar{\tau}_1 = \tau_0^{FB}/\lambda \approx 8,727 \text{ €/Mm}^3$. The empirical self-reporting / truthfully-reporting boundary tracks the green ray, and the welfare-maximizing grid point lies at their intersection, in agreement with the theoretical prediction.

tax $\tau_0 = (1 - \gamma)\delta$ handles the Pigouvian correction perfectly, while $\tau_1 \geq \bar{\tau}_1$ serves purely as a compliance fence. In the single-tax regime, one instrument τ must simultaneously serve both roles, and the resulting compromise is generically suboptimal.

3.7 Bounded Penalties

Theorem 3 assumes τ_1 can be set arbitrarily high. In practice, legal or political constraints may impose an upper bound $\tau_{\max}$ on admissible tax rates. We now characterize what happens when this bound is binding.

Theorem 4 (Price of Limited Enforcement). *Suppose $\eta\sigma > 1/2$ and let $\tau_{\max}$ be the maximum admissible tax rate. Let τ_0^{FB} , $\bar{\tau}_1$, and λ be as defined in Theorem 3. Then:*

- (i) **(Achievability.)** *If $\tau_{\max} \geq \bar{\tau}_1$, the first-best is achievable: the agency sets $\tau_0 = \tau_0^{FB}$ and any $\tau_1 \in [\bar{\tau}_1, \tau_{\max}]$ to obtain $SW = W^{FB}$.*
- (ii) **(Compliance-constrained regime.)** *If $\tau_{\max} < \bar{\tau}_1$, the first-best is not achievable under compliance ($i^* = 0$). The best achievable compliant policy sets $\tau_1 = \tau_{\max}$ and*

$$\tau_0^c = \lambda \cdot \tau_{\max} = \frac{4\eta\sigma - 1}{4\eta\sigma} \cdot \tau_{\max} < \tau_0^{FB}, \quad (3.35)$$

yielding compliance ($i^ = 0$) but with over-extraction ($w_t^c > w_t^{FB}$ for all t at which $w_t^{FB} > 0$).*

- (iii) **(Welfare loss characterization.)** *The welfare loss under the compliance-constrained policy is:*

$$\Delta W(\tau_{\max}) \triangleq W^{FB} - SW(\tau_0^c) > 0, \quad (3.36)$$

and ΔW is strictly decreasing and convex in $\tau_{\max}$ on $[0, \bar{\tau}_1]$, with $\Delta W(\bar{\tau}_1) = 0$.

Proof. **Part (i)** follows immediately from Theorem 3.

Part (ii). Under compliance ($i^* = 0$), the incentive compatibility constraint requires $V^*(\tau_0) \geq V^*(\lambda\tau_1)$, i.e., $\tau_0 \leq \lambda\tau_1$ (since V^* is strictly decreasing). With $\tau_1 \leq \tau_{\max}$, the highest admissible τ_0 that maintains compliance is:

$$\tau_0^c = \lambda \cdot \tau_{\max}. \quad (3.37)$$

Since $\tau_{\max} < \bar{\tau}_1 = \tau_0^{FB}/\lambda$, we have $\tau_0^c = \lambda\tau_{\max} < \lambda \cdot (\tau_0^{FB}/\lambda) = \tau_0^{FB}$.

The agency wants to set τ_0 as high as possible (up to τ_0^{FB}) to deter over-extraction. We now verify this by computing $dSW/d\tau_0$. Substituting out H_t using the aquifer recurrence as

in Proposition 1, the social welfare under compliance ($i^* = 0$, $\theta^* = 0$) can be written purely as a function of $\mathbf{w}$:

$$\begin{aligned} \text{SW}(\mathbf{w}) = N \sum_{t=1}^T & \left[\left(\frac{\alpha}{\beta} - c_0 + c_1 C_t - (1 - \gamma)\delta \right) w_t \right. \\ & \left. - \frac{1}{2\beta} w_t^2 - K \sum_{s=1}^t A^{t-s} w_s w_t \right] + TR\delta, \end{aligned} \quad (3.38)$$

where C_t , K , A are as defined in Proposition 1 and $TR\delta$ is a constant (independent of $\mathbf{w}$). The compliant consumer's reduced objective (Proposition 2) is:

$$\begin{aligned} F_0(\mathbf{w}; \tau_0) = \sum_{t=1}^T & \left[\left(\frac{\alpha}{\beta} - c_0 + c_1 C_t - \tau_0 \right) w_t \right. \\ & \left. - \frac{1}{2\beta} w_t^2 - K \sum_{s=1}^t A^{t-s} w_s w_t \right]. \end{aligned} \quad (3.39)$$

Comparing (3.38) and (3.39) term by term, SW and $N \cdot F_0$ share all the quadratic and cross terms and differ only in the linear coefficient: SW has $-(1 - \gamma)\delta$ where F_0 has $-\tau_0$. Therefore, the per-component gradient satisfies:

$$\frac{1}{N} \frac{\partial \text{SW}}{\partial w_t} = \frac{\partial F_0}{\partial w_t} + [\tau_0 - (1 - \gamma)\delta], \quad \forall t. \quad (3.40)$$

Note that τ_0 does not appear directly in SW —the tax is a transfer from consumer to agency that cancels in social welfare. The entire effect of τ_0 on social welfare is *indirect*, operating through the consumer's extraction response $\mathbf{w}^*(\tau_0)$. We therefore apply the chain rule:

$$\frac{d \text{SW}}{d \tau_0} = \sum_{t=1}^T \frac{\partial \text{SW}}{\partial w_t} \bigg|_{\mathbf{w}^*(\tau_0)} \cdot \frac{\partial w_t^*}{\partial \tau_0}. \quad (3.41)$$

At the consumer's optimum, $\partial F_0 / \partial w_t|_{\mathbf{w}^*} = 0$ (first-order stationarity, which holds at interior solutions; at corner solutions $w_t^* = 0$ the corresponding term vanishes since $\partial w_t^* / \partial \tau_0 = 0$). Substituting (3.40):

$$\frac{d \text{SW}}{d \tau_0} = N \sum_{t=1}^T [\tau_0 - (1 - \gamma)\delta] \cdot \frac{\partial w_t^*}{\partial \tau_0}. \quad (3.42)$$

Since $\partial w_t^* / \partial \tau_0 \leq 0$ (Proposition 4, with strict inequality when $w_t^* > 0$), we have $d \text{SW} / d \tau_0 > 0$ when $\tau_0 < (1 - \gamma)\delta = \tau_0^{FB}$, confirming that social welfare is increasing in τ_0 below the first-best. Therefore, the agency optimally sets $\tau_0 = \tau_0^c$ (the highest compliant rate), and since $\tau_0^c < \tau_0^{FB}$, the compliant consumers face a lower effective tax than socially optimal,

leading to over-extraction: $w_t^c = w_t^*(\tau_0^c) > w_t^*(\tau_0^{FB}) = w_t^{FB}$ by monotonicity (Proposition 4).

Part (iii). Define $g(\tau_0) = \text{SW}(\tau_0)$. From the derivative (3.42), $g'(\tau_0) > 0$ for $\tau_0 < \tau_0^{FB}$, so $\Delta W(\tau_{\max}) = g(\tau_0^{FB}) - g(\lambda\tau_{\max})$ is strictly positive for $\tau_{\max} < \bar{\tau}_1$ and equals zero at $\tau_{\max} = \bar{\tau}_1$.

To show ΔW is strictly decreasing in $\tau_{\max}$:

$$\frac{d\Delta W}{d\tau_{\max}} = -\lambda \cdot g'(\lambda\tau_{\max}) < 0, \quad (3.43)$$

since $g'(\lambda\tau_{\max}) > 0$ for $\lambda\tau_{\max} < \tau_0^{FB}$.

To show convexity:

$$\frac{d^2\Delta W}{d\tau_{\max}^2} = -\lambda^2 \cdot g''(\lambda\tau_{\max}). \quad (3.44)$$

Differentiating (3.42):

$$g''(\tau_0) = N \sum_{t=1}^T \frac{\partial w_t^*}{\partial \tau_0} + [\tau_0 - (1 - \gamma)\delta] \cdot \frac{\partial^2 w_t^*}{\partial \tau_0^2}. \quad (3.45)$$

At $\tau_0 < \tau_0^{FB}$, the first term is strictly negative (monotonicity), and the second term has $[\tau_0 - (1 - \gamma)\delta] < 0$ multiplied by the second derivative of the extraction response. For the quadratic structure of the model, $w_t^*(\tau_0)$ is piecewise affine in τ_0 (the Hessian is constant, but the slope changes at values of τ_0 where nonnegativity constraints $w_t \geq 0$ or $H_t \geq 0$ bind or unbind). Within each piece, $\partial^2 w_t^* / \partial \tau_0^2 = 0$, and therefore $g''(\tau_0) = N \sum_t \partial w_t^* / \partial \tau_0 < 0$. Since $w_t^*(\tau_0)$ is continuous (by Berge's Maximum Theorem), $g(\tau_0)$ is globally C^1 and piecewise strictly concave, hence globally strictly concave. This gives $d^2\Delta W / d\tau_{\max}^2 = -\lambda^2 g''(\lambda\tau_{\max}) > 0$ on each piece, establishing convexity. $\square$

Remark 4 (Interpretation of the Welfare Loss). *The convexity of ΔW implies that the marginal cost of tightening the penalty bound is increasing as $\tau_{\max}$ decreases: small restrictions on $\tau_{\max}$ near $\bar{\tau}_1$ have minimal welfare impact, but severe restrictions cause rapidly escalating losses.*

Remark 5 (Comparison with Evasion Regime). *When $\tau_{\max} < \bar{\tau}_1$, the agency also has the option of allowing evasion ($i^* = 1$), which corresponds to the single-tax regime analyzed in Section 3.5.2. In this case, the social surplus is discounted by $(1 - \theta^*) = 1 - 1/(2\eta\sigma)$ due to unreported extraction. The agency should compare $\text{SW}(\tau_0^c)$ from the compliant regime with the evasion-regime welfare and choose the higher of the two. For moderate values of $\tau_{\max}$, the compliant regime dominates because it avoids the deadweight loss from evasion.*

Remark 6 (Welfare Discontinuity at the Compliance Boundary). *The agency's utility panel of Figure 3.1 exhibits a sharp welfare drop along the compliance boundary $\tau_0 = \lambda\tau_1$. On the compliant side (below the boundary), welfare is uniformly high. Immediately across the boundary in the evasion region, welfare drops abruptly. This discontinuity reflects two*

compounding sources of deadweight loss from evasion: (i) the social surplus is discounted by the factor $(1 - \theta^*) = 1 - 1/(2\eta\sigma)$ because the unreported share of extraction does not contribute to measured welfare, and (ii) the evasion friction $\eta\sigma\tau_1\theta^{*2}w_t$ represents pure waste that benefits neither the consumer nor the agency.

3.8 Robustness to Monitoring Effort

Corollary 4 (Monitoring Invariance). *Under the differential tax mechanism with unconstrained τ_1 , the maximum attainable social welfare W^{FB} is independent of the monitoring parameter η , provided $\eta\sigma > 1/2$.*

Proof. By Theorem 3, the first-best is achieved by setting $\tau_0 = \tau_0^{FB} = (1 - \gamma)\delta$, which does not depend on η , and $\tau_1 \geq \bar{\tau}_1(\eta) = \frac{4\eta\sigma}{4\eta\sigma - 1} \cdot (1 - \gamma)\delta$. While the compliance threshold $\bar{\tau}_1$ depends on η (lower η requires higher τ_1), as long as τ_1 is unconstrained, the agency can always meet this threshold. Since the equilibrium extraction $\mathbf{w}^{FB}$ depends only on $\tau_0 = \tau_0^{FB}$ and the aquifer parameters, it is invariant to η . $\square$

Corollary 5 (Monitoring–Penalty Substitution). *Suppose $\tau_{\max} < \infty$. For the first-best to be achievable, the monitoring capacity and penalty bound must satisfy:*

$$\eta\sigma \geq \frac{\tau_{\max}}{4(\tau_{\max} - \tau_0^{FB})}, \quad (3.46)$$

where $\tau_0^{FB} = (1 - \gamma)\delta$. Equivalently, the minimum monitoring effort required to achieve the first-best is:

$$\eta_{\min}(\tau_{\max}) = \frac{\tau_{\max}}{4\sigma(\tau_{\max} - \tau_0^{FB})}, \quad (3.47)$$

which is valid for $\tau_{\max} > \tau_0^{FB}$ and is strictly decreasing in $\tau_{\max}$.

Proof. The first-best requires $\tau_{\max} \geq \bar{\tau}_1 = \frac{4\eta\sigma}{4\eta\sigma - 1} \cdot \tau_0^{FB}$. Rearranging:

$$\tau_{\max}(4\eta\sigma - 1) \geq 4\eta\sigma \cdot \tau_0^{FB} \quad (3.48)$$

$$4\eta\sigma(\tau_{\max} - \tau_0^{FB}) \geq \tau_{\max} \quad (3.49)$$

$$\eta\sigma \geq \frac{\tau_{\max}}{4(\tau_{\max} - \tau_0^{FB})}. \quad (3.50)$$

Note $\tau_{\max} > \tau_0^{FB}$ is necessary since $\bar{\tau}_1 > \tau_0^{FB}$ always. Solving for η gives the stated formula, which is manifestly decreasing in $\tau_{\max}$. $\square$

Remark 7 (Policy Interpretation of Corollary 5). *Corollary 5 quantifies the substitution between monitoring capacity (η) and legal penalty headroom ($\tau_{\max}$). An agency with*

weak enforcement (low η) can compensate by ensuring the legal framework permits sufficiently high penalty rates. Conversely, strict legal caps on penalties require stronger monitoring to achieve the same outcome. The hyperbolic relationship $\eta_{\min} \propto 1/(\tau_{\max} - \tau_0^{FB})$ shows that the required monitoring effort diverges as $\tau_{\max} \rightarrow \tau_0^{FB}$ from above, confirming that some penalty premium above the Pigouvian rate is essential for the mechanism to function.

3.9 Conclusion

We proposed a framework for groundwater management where a public water agency sets differential tax rates to regulate extraction while accounting for water consumers' strategic behavior. Modeling the interaction as a Stackelberg game, we characterized the trade-off between incentivizing truthful reporting and limiting groundwater depletion. Numerical experiments showed that the differential tax scheme maximizes social welfare while achieving better compliance than a uniform tax regime. We further showed that the differential tax implements the first-best allocation under unbounded penalties (Theorem 3). When legal caps on penalties bind, monitoring capacity and statutory penalty headroom substitute in a closed-form way (Corollary 5).

Chapter 4

Profit Maximization for a Robotics-as-a-Service Model

4.1 Introduction

Rapid advances in robotics and artificial intelligence (AI) are enabling the growth of *cloud robotics* and *robotics-as-a-service* (RaaS) [22, 48, 51, 18]. The applications of RaaS to urban mobility [94, 23] and drone delivery [36, 89] are in particular finding rapid development and planned adoption, although concepts in agriculture [31, 69], manufacturing [102], and other domains [48, 95] are also being considered. Yet despite growing attention to the RaaS paradigm, questions related to business aspects (e.g., pricing, maintenance, or replacement decisions) have been less well explored.

4.1.1 Predictive Maintenance

One important business aspect of RaaS is decisions related to maintenance. Because equipment maintenance is a major source of costs and delays in industry, there has been substantial interest in developing data-driven approaches for streamlining maintenance decisions. Predictive maintenance [104, 19, 100] uses sensors that collect streaming data from equipment in order to predict degradation and identify maintenance that may be needed. Approaches based on more classical statistical modeling [104], machine learning [19], and neural-network-based AI [100] have been proposed. Such approaches have found high-visibility success in the aviation context, where predictive maintenance is used for commercial aircraft and their constituent components [41].

4.1.2 Failure Modeling

Another important business aspect of RaaS is decisions related to replacement of robots. Whereas questions of maintenance are related to understanding tradeoffs between the timing

and amount of maintenance with uncertain knowledge about the level of equipment degradation, decisions regarding replacement are related to managing tradeoffs between holding costs (e.g., costs related to storage), failure costs (e.g., costs related to failures while in operation), and replacement costs (i.e., costs for purchasing new robots).

Though there is relatively little work on managing replacements, there has been considerably more research on modeling rates of survival or failure. Component-based models for predicting failure rates in electronics have been developed [96, 47] and found use in practice. Survival analysis [99, 50] uses data to empirically estimate rates of failure and their sensitivity to predictive factors, and it is most commonly used within healthcare [99, 24, 37] with less frequent use in other domains like predicting machine failure [50, 77]. The mainstay approach for survival analysis is the proportional hazards model [25, 55], which is able to successfully handle the right-censoring of survival data (i.e., life-span data is not available for items or people that have survived at the time of data collection).

4.1.3 Inverse Optimization

An important business aspect of RaaS is decisions related to pricing. For a fleet of robots that performs different jobs arriving sequentially, a relevant business decision is how to price each job in order to maximize profit while accounting for equipment degradation and other costs. The most related line of work is the topic of data-driven contract design. In the area of contract design [56], the *principal* is a leader in making decisions while the *agent* is a follower that makes decisions knowing the decision of the principal. The principal seeks to make an optimal decision knowing the *leader-follower* information structure.

One set of approaches to data-driven contract design makes use of inverse optimization [4, 21], which are techniques that input decision data from agents and estimate utility function models for which the decision data are optimal. These estimated utility function models can then be input into bilevel optimization problems (i.e., optimization problems in which some constraints are optimization problems themselves) in order to design an optimal contract knowing the leader-follower information structure [5, 70, 63].

Another approach to data-driven contract design revolves around multi-armed bandit models [33, 34, 87] in order to better understand exploration-exploitation issues. Based upon multi-armed bandit models in which choosing an action provides a stochastic reward, in these models the agent chooses an action in order to maximize their own expected rewards while considering an incentive amount provided by the principal for each action. The central issue is in how the principal can indirectly estimate the agent's expected rewards for each action having observations of only the sequence of actions chosen and not directly measuring the agent's realized rewards. Viewed from a certain perspective, this estimation problem is analogous to inverse optimization.

4.1.4 Contribution and Outline

To the best of our knowledge, this is the first paper to consider the problem of maximizing profit for robots operating as part of a RaaS business where pricing and replacement decisions are made. Potential customers arrive sequentially, where each customer's job has a vector-valued context and a model for (a) the maximum price willing to pay for the job and (b) the stochastic degradation of a robot completing the job. The operator of the RaaS business must decide what price to charge for servicing the job, and the customer gets to decide whether or not to accept the price. If a customer rejects a price, then the job is lost forever. If a customer accepts a price, then a robot services the job. The robot fails during the job with a stochastic rate depending upon its current degradation level. If the robot completes the job, then the operator gets to decide whether or not to replace the robot considering the tradeoffs between robot replacement costs, holding (e.g., storage) costs, and opportunity costs (i.e., possible future revenue without replacing the robot).

The chapter is outlined as follows: Section 4.2 formally defines the elements of the quantitative model that we consider in this chapter, and it presents a stochastic control formulation of the profit maximization problem as a semi-Markov decision process (SMDP). Section 4.3 describes our proposed approaches for optimizing the RaaS operator's decisions within the SMDP. The chapter concludes with numerical experiments in Section 4.4 that demonstrate the potential efficacy of our proposed approach.

4.2 Model and Problem Statement

This section defines the system model and formulates the operator's decision problem as an average-reward SMDP.

4.2.1 Customer and Jobs Model

In our online setting, customers arrive sequentially at times indexed by $k = 1, 2, \dots$. The k -th customer is characterized by a tuple (x_k, T_k, τ_k) , where $x_k \in \mathbb{R}_+^d$ ($\|x_k\|_2 \leq 1$) is a context vector representing the job, $T_k > 0$ is the desired rental duration, and $\tau_k > 0$ is the interarrival time.

Upon arrival of a customer, the RaaS operator observes (x_k, T_k) and offers a price p_k to the customer.

We assume that a customer's decision to rent a robot for their job is deterministic and governed by an unknown (to the RaaS operator) utility vector $u \in \mathbb{R}_+^d$, with $\|u\|_2 \leq 1$. The customer accepts the offered price if their utility meets or exceeds the price, i.e., $u^\top x_k \geq p_k$. One of the RaaS operator's challenges is learning u in order to be able to set profit-maximizing prices.

Remark 8. *There are two key aspects to the above assumption. The first is that we are assuming that a customer's utility for their job is a linear function of the context x_k . Such a*

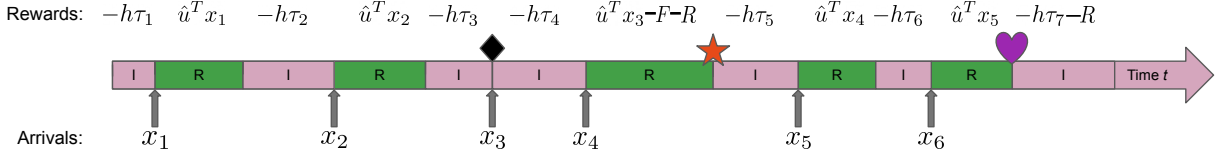


Figure 4.1: An example timeline illustrating the operational flow for the Robotics-as-a-Service (RaaS) provider. Upward arrows indicate the arrival of new customers (x_i). The robot's status alternates between being under rental ('R' in a green box) and idle ('I' in a purple box). The timeline also depicts key events like a price rejection from an arriving customer (diamond symbol), an unexpected robot failure during a job (star symbol), and a voluntary replacement initiated by the operator (heart symbol). The rewards associated with each period are shown above the timeline, reflecting revenues $\hat{u}^\top x_i$, holding costs $h\tau_i$, and costs for failure F and replacement R .

linear context assumption is common in the literature involving online models [2, 32, 6]. The second is that we are assuming a customer rents a robot as long as their utility for completing the job is greater than or equal to the price. In the case where the utility equals the price, presumably the customer receives no surplus value for the completion of the job since they must pay an amount equal to their utility. However, we note that such an assumption is common in the contract design literature [56], and the corresponding inequality is often referred to as a participation constraint. We also note that this assumption can be relaxed to an alternative model where a customer rents a robot if $u^\top x_k \geq p_k + \nu$ where ν is a constant that describes the minimum surplus value a customer must receive for accepting a rental, but for conciseness we assume the above.

4.2.2 Robot Degradation and Failure Model

A robot's condition is modeled by its cumulative operational history. Let X be the sum of context vectors from all prior successful rentals since the last replacement, and let t_{age} be the sum of their rental durations. During a new rental with context x_k , the robot is subject to breakdown. We model this using a Cox proportional hazards model [25, 55]: The *hazard rate* at time $t \in [0, T_k]$ into the rental, which can be interpreted as the probability the robot will not survive in an infinitely small period of time beyond $t_{\text{age}} + t$, conditioned on the robot having survived up to time $t_{\text{age}} + t$, is

$$\lambda(t|X, x_k) = \lambda_0(t + t_{\text{age}}) \exp(\theta^\top (X + x_k)) \quad (4.1)$$

where

- $\lambda_0(\cdot)$ is an unknown (to the RaaS operator), non-negative baseline hazard function;

- $\theta \in \mathbb{R}_+^d$ is an unknown (to the RaaS operator) degradation parameter vector, with normalization $\|\theta\|_\infty \leq 1$, that modulates the impact of cumulative context on failures.

A breakdown incurs a failure cost F and a replacement cost R , after which the robot's cumulative context and cumulative rental duration are reset to states corresponding to a new robot (i.e., $X = 0$ and $t_{\text{age}} = 0$). The second challenge for the RaaS operator is to learn θ and $\lambda_0(\cdot)$ to effectively manage the risk of breakdown.

Remark 9. *Note that we are assuming that the hazard rate depends upon a linear function of the cumulative contexts X , which is the sum of the contexts from prior successful rentals since the last replacement. This is a modest generalization of the linear context assumption that is common in the literature involving online models [2, 32, 6], but it is arguably reasonable since we would expect failure rates to depend upon the history of completed jobs.*

4.2.3 Model of RaaS Operator

We formulate the RaaS operator's decision problem as an infinite-horizon SMDP defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$.

State Space $\mathcal{S}$ The state $s \in \mathcal{S}$ must capture the robot's condition and the current decision epoch. We define $s = (X, t_{\text{age}}, \phi)$, where $X \in \mathbb{R}_+^d$ is the cumulative context vector since the last replacement, $t_{\text{age}} \in \mathbb{R}_+$ is the cumulative operational time conditioning the baseline hazard, and $\phi \in \{0, 1\}$ is a phase indicator ($\phi = 0$ for a pricing decision upon customer arrival, and $\phi = 1$ for a replacement decision when idle).

Action Space $\mathcal{A}(s)$ The available actions depend on the phase: during arrival ($\phi = 0$), the operator chooses $a \in \{\text{Accept}, \text{Reject}\}$; during the idle phase ($\phi = 1$), the operator chooses $a \in \{\text{Continue}, \text{Replace}\}$. The 'Replace' action incurs cost R and resets the state.

Reward Function $\mathcal{R}$ & Transition Dynamics $\mathcal{P}$ Let the current state be $s = (X, t_{\text{age}}, \phi)$. The rewards and state transitions are determined by the system's current state $s = (X, t_{\text{age}}, \phi)$ and the operator's chosen action. Because sojourn times vary with the policy, we describe the raw (un-normalized) rewards below; the per-unit-time weighting is handled by continuous-time discounting in the SMDP Bellman equations.

When the system is in the arrival phase (i.e., Case $\phi = 0$), a customer with characteristics (x_k, T_k) has just arrived. If the RaaS operator chooses to 'Reject' the rental, the system transitions to $(X, t_{\text{age}}, 0)$ with zero reward and zero elapsed time.

If the RaaS operator chooses to 'Accept', the customer makes an acceptance decision. If $p_k > u^\top x_k$ (i.e., the price is greater than the customer's utility for the job), then the customer rejects the offered price and the system transitions to $(X, t_{\text{age}}, 0)$ with zero reward and zero elapsed time. Otherwise, the customer accepts the offered price and the rental begins. The outcome is stochastic, and it is specifically governed by the probability of robot failure $p_{\text{fail}} = 1 - \exp(-\exp(\theta^\top (X + x_k)) \int_0^{T_k} \lambda_0(t_{\text{age}} + t) dt)$. With probability $1 - p_{\text{fail}}$, the rental is successful, yielding a reward of p_k . The system state is updated to reflect the accumulated usage, transitioning to the idle phase at state $(X + x_k, t_{\text{age}} + T_k, 1)$. With

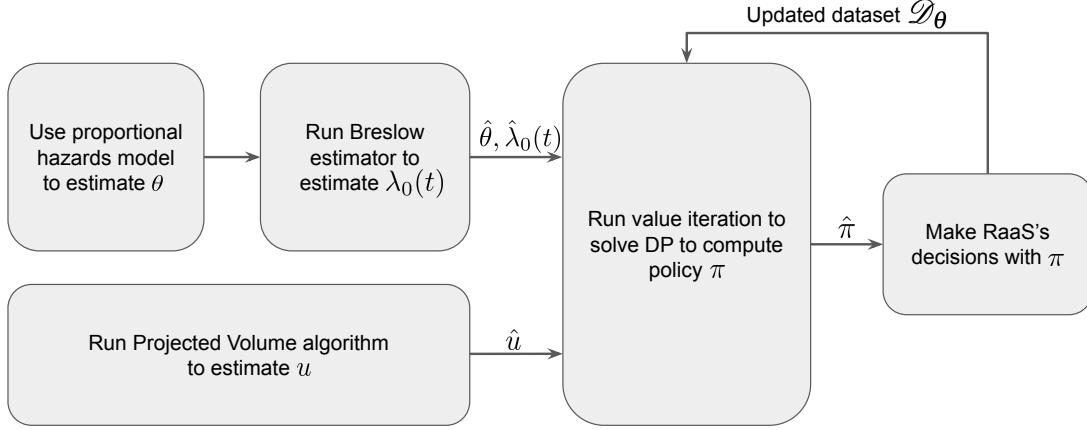


Figure 4.2: A diagram of the three-phase learning and control framework. **Phase 1:** The Projected Volume algorithm is run to obtain a reliable estimate of the customer utility vector, $\hat{u}$. **Phase 2:** This estimate is used to learn the degradation parameters $(\hat{\theta}, \hat{\lambda}_0)$ and train an initial control policy, $\hat{\pi}$. **Phase 3:** The system operates using this policy until a predefined number of new robot failures (K) occur. The data from these failures updates the dataset $\mathcal{D}_\theta$, which is then used to re-estimate the parameters and retrain the policy in a repeating cycle.

probability p_{fail} , the robot breaks down during the rental at a random time $f_k < T_k$. This results in a reward of $p_k - F - R$, accounting for revenue, failure cost, and replacement cost. The robot is replaced, and the state resets to $(0, 0, 1)$ because the RaaS operator is now using a new robot.

When the system is in the idle phase (i.e., Case $\phi = 1$), the RaaS operator decides whether or not to replace the robot. If the action is to ‘Continue’, then no replacement occurs. The system incurs a holding cost at rate h over the interarrival time τ_{k+1} , resulting in a reward of $-h\tau_{k+1}$, and transitions to state $(X, t_{\text{age}}, 0)$ when the next customer arrives. If the action is to ‘Replace’, then the RaaS operator pays the replacement cost R in addition to the holding cost, resulting in a reward of $-R - h\tau_{k+1}$. The state is reset to $(0, 0, 0)$ for the next customer arrival.

4.2.4 Problem Statement

We assume that the RaaS operator’s objective is to maximize their average profit per unit of time. Formally, let π denote a policy that specifies, for each arriving customer k , whether to offer a rental (and at what price p_k) and, after each rental or rejection, whether to voluntarily replace the robot. Under policy π , the system generates a sequence of revenues and costs over an infinite horizon. Figure 4.1 provides an example timeline of this process, illustrating the sequence of events and their associated rewards. Let $\mathcal{T}[K]$ denote the total

time elapsed after the first K customers:

$$\mathcal{T}[K] = \sum_{k=1}^K \tau_k + \sum_{k=1}^K T_k \cdot \mathbb{I}\{a_k = \text{accept}\}, \quad (4.2)$$

accounting for both idle and rental periods (adjusted for partial rentals due to failures). The profit consists of accepted prices $\{p_k : a_k = \text{accept}\}$, where a_k is the customer's acceptance decision. The costs include failure penalties F for each mid-rental breakdown, replacement costs R for each breakdown or voluntary replacement, and holding costs $h \cdot \tau_k$ for each interarrival idle period τ_k .

The fundamental optimization problem that the RaaS operator seeks to solve is thus given by

$$\max_{\pi} \lim_{K \rightarrow \infty} \frac{1}{\mathcal{T}[K]} \left[\sum_{k=1}^K p_k \cdot \mathbb{I}\{a_k = \text{accept}\} - F \cdot n_F[K] - R \cdot n_R[K] - h \sum_{k=1}^K \tau_k \right], \quad (4.3)$$

where $n_F[K]$ is the number of failures and $n_R[K]$ is the number of replacements (including those triggered by failures) up to customer K , and $\mathbb{I}\{\cdot\}$ is the indicator function. All terms depend on π by its influence on acceptances, failures, and replacements. Since the unknown parameters (u, θ, λ_0) must first be learned, the operator has to manage the exploration-exploitation tradeoff between exploration to estimate these parameters and exploitation to maximize profit.

4.3 Solution Framework

To tractably compute a policy that optimizes (4.3), we propose an algorithm that decouples the optimization problem into three sequential phases: (1) estimation of the customer utility vector, (2) estimation of the robot degradation parameters, and (3) computation of an optimal control policy. This iterative learning and control process is illustrated in Figure 4.2. We manage the exploration-exploitation tradeoff by making use of a standard ϵ -greedy approach [27]. This section describes the algorithms used for each phase, and implementation details are provided in the next section.

4.3.1 Phase 1: Learning the Utility Parameters u

To learn u , we employ the projected volume algorithm [87, 65] for multidimensional binary search. We maintain a convex uncertainty set $S_k \subset \{s \in \mathbb{R}^d : \|s\|_2 \leq 1\}$ guaranteed to contain u . At arrival k with context x_k , the algorithm computes an approximate centroid $\hat{s}_k$ derived from S_k to offer a query price $p_k = \hat{s}_k^\top x_k$. The customer's binary accept/reject decision yields a linear constraint, which updates the set via a volume-cutting halfspace:

$S_{k+1} = S_k \cap \{s : s^\top x_k \geq p_k\}$. This phase terminates when S_k is sufficiently small, outputting the final centroid as our high-confidence estimate $\hat{u}$.

4.3.2 Phase 2: Learning Degradation Parameters $(\theta, \lambda_0(\cdot))$

Once a reliable estimate $\hat{u}$ is obtained, we enter a data collection phase to learn the degradation model. Using a fixed pricing policy (e.g., $p_k = \hat{u}^\top x_k$), we gather rental data tuples of (cumulative context, rental duration, cumulative operational time, breakdown indicator).

Estimating θ : We estimate the parameter θ by maximizing the partial log-likelihood of the Cox model [99, 25]:

$$\mathcal{L}(\theta) = \sum_{t \in \mathcal{F}} \left[\sum_{k \in \mathcal{D}(t)} \theta^\top (X_k + x_k) - |\mathcal{D}(t)| \log \left(\sum_{j \in R(t)} \exp(\theta^\top (X_j + x_j)) \right) \right], \quad (4.4)$$

where $\mathcal{F}$ is the set of unique failure times, $\mathcal{D}(t)$ is the set of failures at time t , and $R(t)$ is the risk set at t (rentals active at t). This conveniently eliminates the dependency on the baseline hazard $\lambda_0(t)$. We impose the constraint $\theta \geq 0$ to ensure that usage is strictly degrading under the estimated $\hat{\theta}$.

Estimating $\lambda_0(\cdot)$: With the estimate $\hat{\theta}$, we then estimate the cumulative baseline hazard $\Lambda_0(t) = \int_0^t \lambda_0(s) ds$ using the non-parametric Breslow estimator [14]. The instantaneous baseline hazard $\hat{\lambda}_0(t)$ is recovered by smoothing the increments of $\hat{\Lambda}_0(t)$ and then differentiating.

4.3.3 Phase 3: Solving the SMDP Problem

If the parameters $(u, \theta, \lambda_0(\cdot))$ are known, then in principle the optimal policy can be computed using dynamic programming [10, 9]. The relevant computational issue is whether or not the dimensionality of the state space is small enough to employ an algorithm that directly computes an optimal policy via dynamic programming.

Though the state space (see Section 4.2.3) is continuous and high-dimensional (because of the cumulative context vector $X \in \mathbb{R}_+^d$), there exists a lower-dimensional projection that preserves all information relevant for decision-making. The key insight is that the rewards and dynamics depend only on scalar projections of the context vectors. We thus redefine the state for the arrival phase (i.e., Case $\phi = 0$) as a 5-dimensional tuple $s' = (c_c, c_x, c_u, T, t)$, where:

- $c_c = \theta^\top X$ is the cumulative degradation context;
- $c_x = \theta^\top x_k$ is the customer-specific degradation context;
- $c_u = u^\top x_k$ is the customer-specific revenue context;
- T is the job's rental duration;

Table 4.1: Simulation parameters.

Symbol	Value	Description
d	4	Context dimension
$\mathbb{E}[\tau_k]$	5	Mean interarrival time
$\mathbb{E}[T_k]$	10	Mean rental duration
h	0.02	Holding cost per unit time
F	1.2	Failure cost
R	1.5	Replacement cost
u	(0.37, 0.11, 0.34, 0.71)	True utility vector
θ	(0.5, 0.2, 0.4, 0.3)	True degradation parameter
λ	0.001	Baseline hazard rate

- t is the robot's cumulative active lifetime.

For the idle phase (i.e., Case $\phi = 1$), the state is fully described by the cumulative degradation c_c and t . This projected state space is small enough that dynamic programming can be applied computationally in a direct manner.

Our approach to computing a policy for the SMDP, which optimizes the RaaS operator's objective (4.3), is informed by the above state space projection. Specifically, we use the approaches described in Phases 1 and 2 to estimate the parameters $(u, \theta, \lambda_0(\cdot))$. Value iteration [9], over a finite, discretized grid, is used to compute an optimal policy for the SMDP with model parameters set to be equal to the estimated parameters $(\hat{u}, \hat{\theta}, \hat{\lambda}_0)$. The flow between the various phases of our solution framework is summarized in Figure 4.2.

4.4 Numerical Experiments

4.4.1 Experimental Setup

We validate our online learning and control framework using an event-driven simulation environment.

We simulate a horizon of $K = 20000$ sequential customer arrivals. Customer context vectors $x \in \mathbb{R}_+^d$ are drawn uniformly from the positive orthant of a d -dimensional unit ball. Interarrival times τ_k and rental durations T_k are drawn independently from exponential distributions. The robot's failure mechanism is governed by a Cox proportional hazards model with a constant baseline hazard rate. All simulation parameters are summarized in Table 4.1.

4.4.2 Learning Scheme

We compute a policy in our simulation using the algorithm outlined in Section 4.3. Experiment-specific details are as follows:

Phase 1: Utility Vector Estimation. Building on the Projected Volume algorithm, we allocate an initial exploration period that typically spans the first ~ 40 customer arrivals, terminating dynamically when the diameter of the uncertainty set S_k (measured along the normalized context direction $x_k/\|x_k\|_2$) falls below 10^{-4} . The centroid of S_k is approximated using hit-and-run sampling. During this phase, any accepted rentals contribute initial degradation data to $\mathcal{D}_\theta$ if failures occur.

Interleaved Phases 2 and 3: Degradation Estimation and Policy Learning. Immediately after Phase 1 convergence, we estimate $\hat{\theta}$ by maximizing the partial log-likelihood under the constraint $\theta \geq 0$. For the baseline hazard, we apply the Breslow estimator to compute $\hat{\Lambda}_0(t)$, followed by smoothing its increments via a Gaussian kernel density estimator to obtain a continuous $\hat{\lambda}_0(t)$. If insufficient failures have been observed (fewer than 2), we continue data collection using near-optimal pricing $p_k = \hat{u}^\top x_k - \varepsilon_p$ ($\varepsilon_p = 10^{-2}$) to encourage accepts until the threshold is met, then proceed to estimating θ and $\lambda_0(\cdot)$.

With the system dynamics and reward models specified by the learned parameters $(\hat{u}, \hat{\theta}, \hat{\lambda}_0)$, we compute the optimal policy by solving the SMDP directly. This is accomplished using value iteration on a discretized representation of the five-dimensional state space (c_c, c_x, c_u, T, t) . To manage the uncertainty from initial parameter estimates, the agent begins with a decaying ϵ -greedy policy derived from the initial value function with initial $\epsilon = 0.10$ and decay factor $\gamma_d = 0.95$. This encourages exploration and mitigates the risk of a greedy policy exploiting inaccuracies in the early estimates $\hat{\theta}$ and $\hat{\lambda}_0(\cdot)$, which might otherwise lead to suboptimal, premature ‘Reject’ actions. The policy runs until $K_{\text{fail}} = 100$ additional failures accumulate, at which point we re-estimate $\hat{\theta}$ and $\hat{\lambda}_0(\cdot)$ from the updated $\mathcal{D}_\theta$ and retrain the policy. This batch update process repeats throughout the horizon, adapting to new degradation data.

4.4.3 Baseline Policies

We compare our framework against two myopic baselines using the same estimated parameters $(\hat{u}, \hat{\theta}, \hat{\lambda}_0)$. The Myopic Heuristic accepts a rental if the immediate expected profit is positive ($\hat{u}^\top x_k - p_{\text{fail}} \cdot (F + R) > 0$), and replaces the robot when the expected profit from an average customer turns negative. The Fully Myopic baseline uses the same acceptance rule but only replaces the robot upon failure, never voluntarily.

4.4.4 Results

Profitability

To benchmark our learning algorithms, we also train an oracle optimal policy which assumes *perfect* knowledge of u, θ , and $\lambda_0(t)$ and run simulations under an identical envi-

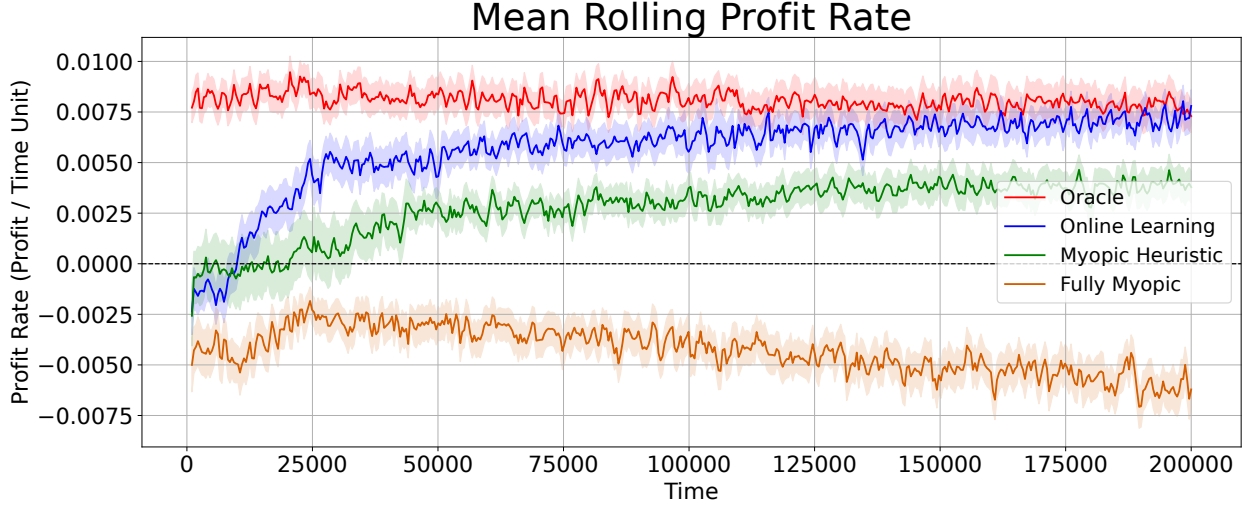


Figure 4.3: Rolling average profit rate per time unit for the oracle, online learning, myopic heuristic, and fully myopic policies, computed over a 2,000-unit window and averaged across 100 simulation runs. Shaded regions indicate ± 2 standard errors.

ronment, representing an upper bound on performance. We evaluate the profitability of our online learning framework by comparing its rolling average profit rate (with window size 2,000) to that of the oracle optimal policy with full knowledge of the true parameters u , θ , and $\lambda_0(t)$. Results are aggregated over 100 independent simulation runs, with means and ± 2 standard errors reported. As shown in Figure 4.3, the oracle policy maintains a consistent positive profit rate from the outset, reflecting its perfect alignment with the environment. In contrast, the online policy initially yields negative rates due to exploratory pricing for estimating u and suboptimal decisions from early, inaccurate estimates of θ . However, as learning progresses through batch updates, the online policy’s profit rate recovers, becoming positive and converging toward the optimal rate. The myopic heuristic achieves a positive but substantially lower steady-state profit rate, and the fully myopic policy converges to a negative profit rate, as it is unable to voluntarily replace a degraded robot. This demonstrates the framework’s ability to adaptively achieve near-optimal performance.

Convergence of $\hat{\theta}$ and $\hat{u}$

To assess parameter estimation, we track the ℓ_2 -norm error between the estimates $\hat{u}$ and $\hat{\theta}$ and their ground-truth values, averaged over 100 runs with standard errors. Figure 4.4 shows rapid convergence for $\hat{u}$, reaching near-zero error after approximately 40 customers, satisfying the termination criterion for Phase 1. Convergence for θ is slower and asymptotic, reflecting the need for accumulating failure observations. Though estimation error decreases slowly after an initial rapid decrease, Figure 4.3 indicates that this does not substantially

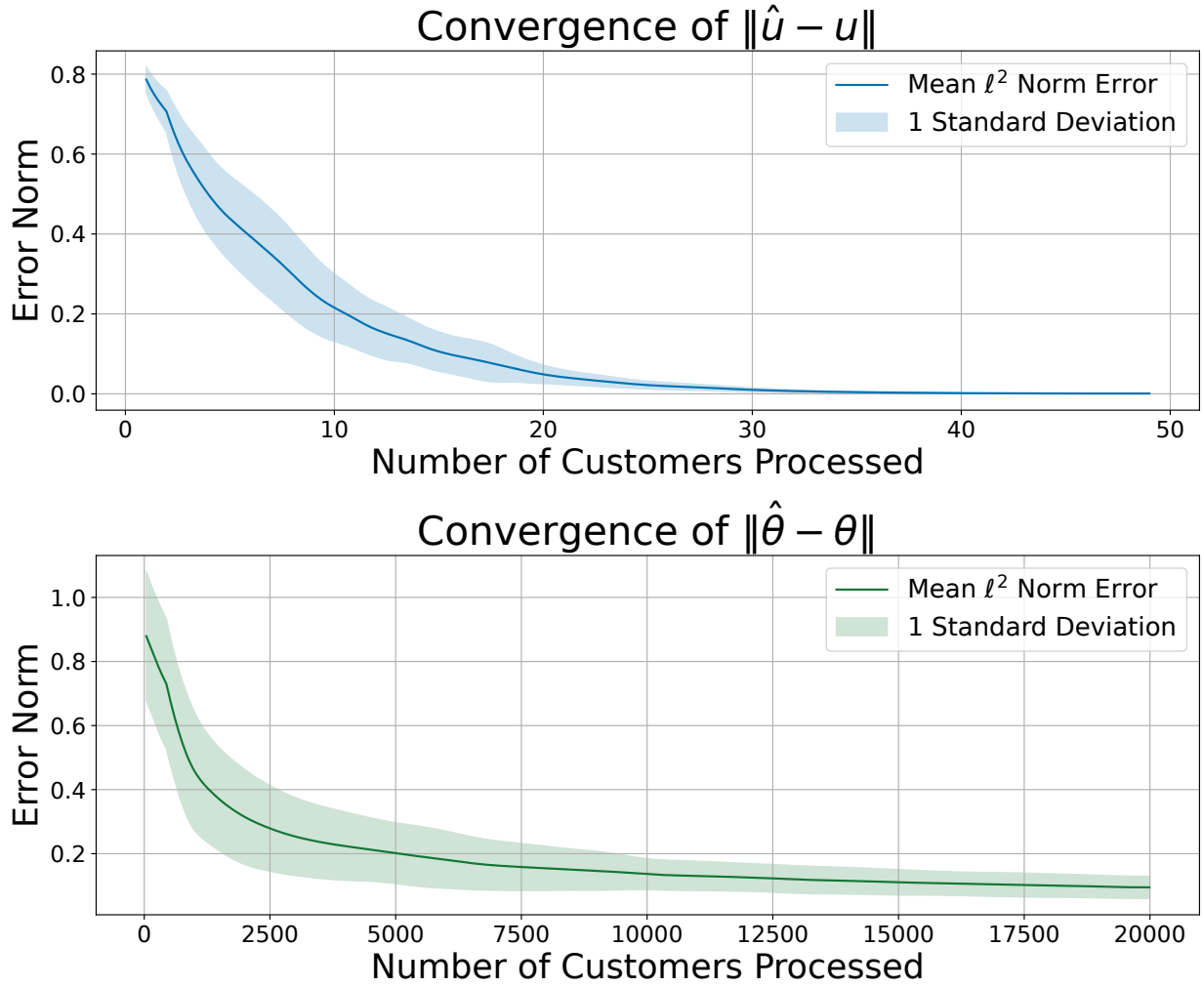


Figure 4.4: ℓ_2 -norm estimation errors for utility vector $\hat{u}$ (top) and degradation parameter $\hat{\theta}$ (bottom) versus number of customers processed, showing means and standard deviations over 100 runs.

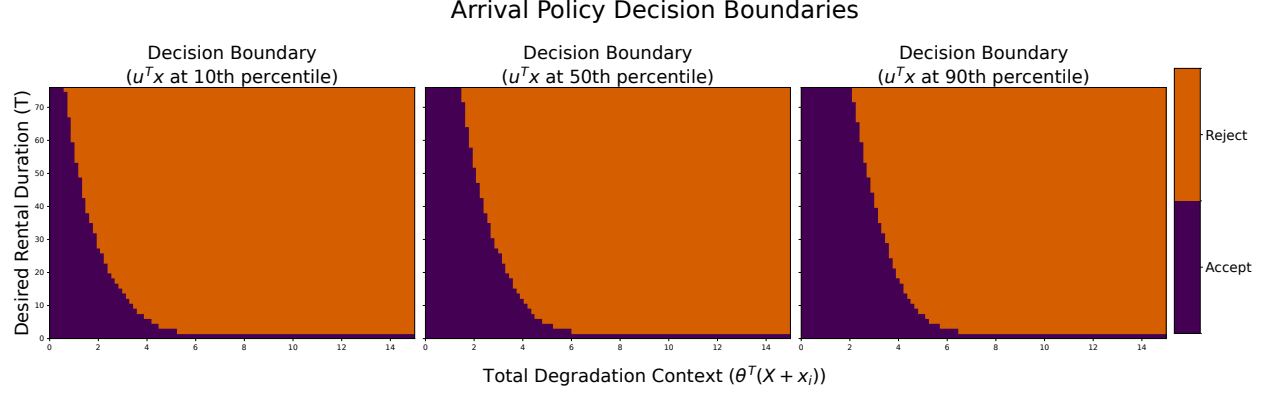


Figure 4.5: Decision regions for job acceptance (purple) or rejection (yellow) in the arrival phase, plotted against desired rental duration T and cumulative degradation $\theta^\top X + \theta^\top x$, at selected percentiles of customer revenue $u^\top x$.

harm the profit rate.

Policy Structure

The computed optimal policy exhibits intuitive structures in its arrival and departure phases, balancing revenue potential against degradation risks, failure costs, holding costs, and replacement expenses.

In the arrival phase (see Figure 4.5), decision boundaries for accepting or rejecting jobs are visualized in the plane of desired rental duration T and projected cumulative degradation $\theta^\top(X + x)$, with slices at the 10th, 50th, and 90th percentiles of projected customer revenue $u^\top x$. Higher rental price percentiles expand the acceptance region, as increased expected payoffs offset degradation-related risks for a given failure probability (determined by T and $\theta^\top(X + x)$).

In the departure phase ($\phi = 1$), a threshold at $\theta^\top X \approx 4$ governs replacement decisions. Below this, retaining the robot is optimal to capitalize on potential future high-revenue jobs; above it, replacement is preferred to preempt likely failures (incurring additional costs beyond replacement) or excessive holding costs while awaiting suitable customers.

To assess whether this threshold structure is specific to the cost parameters of Table 4.1, we recompute the optimal policy via value iteration on a 60×60 grid for each pair of (R, F, h, λ) , holding the remaining two parameters at their default values. As shown in Figure 4.6, the interior departure threshold c^* varies smoothly and monotonically across more than an order of magnitude in each parameter: c^* rises with R , falls with F and λ , and is essentially insensitive to h , consistent with the economic interpretation of each parameter. Grey regions correspond to parameter combinations under which the optimal policy never voluntarily replaces, either because failure costs are too low to justify replacement or because

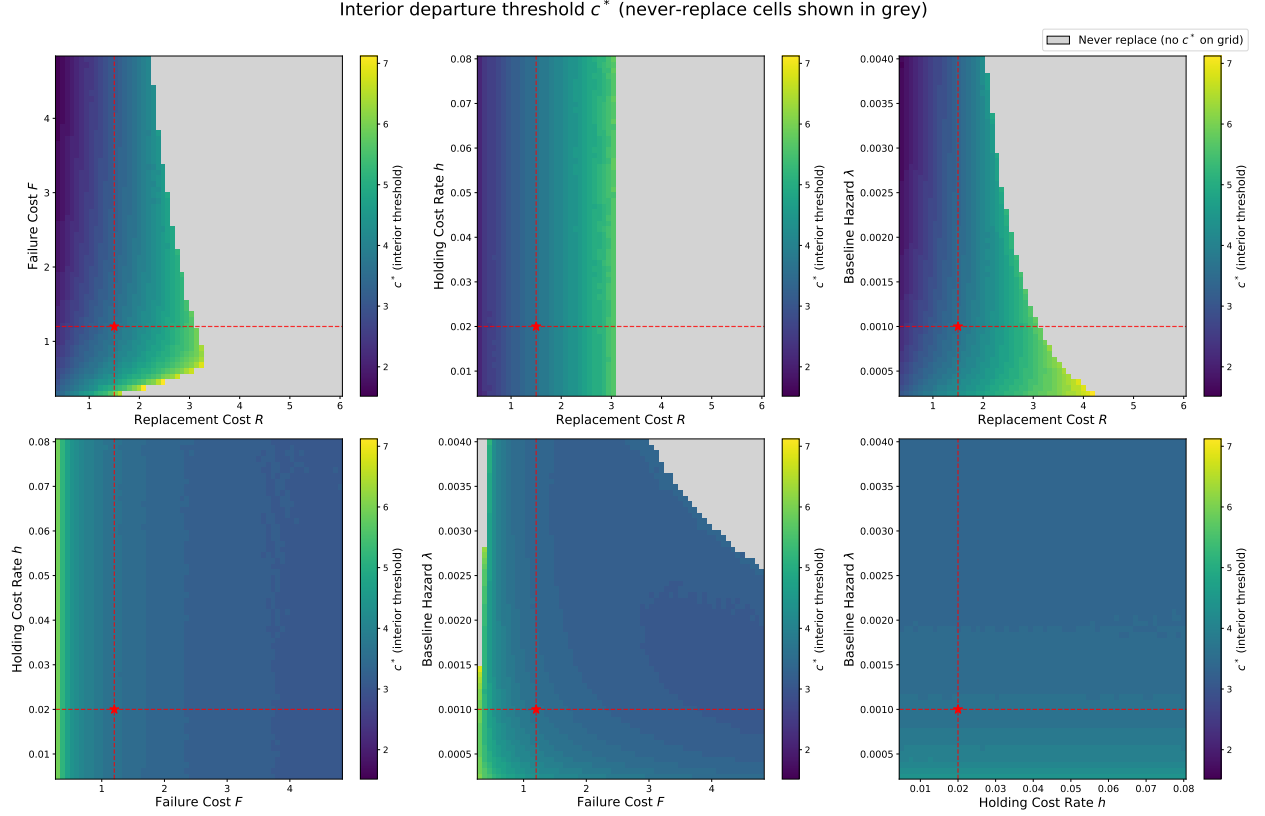


Figure 4.6: Sensitivity of the interior departure threshold c^* to the cost parameters (R, F, h, λ) . Each panel sweeps two parameters on a 60×60 grid while holding the remaining two at their default values from Table 4.1; the red star marks the default operating point. Grey cells denote parameter combinations under which the optimal policy never voluntarily replaces.

robots degrade too rapidly for a fresh robot to recover the replacement cost R over its useful life.

4.5 Conclusion

Our framework enables RaaS operators to learn customer utilities and robot degradation models online, yielding policies that converge to near-optimal profitability through adaptive pricing and replacement decisions. Numerical results validate its efficacy in balancing revenue against operational risks. Future extensions could consider more realistic settings such as stochastic customer acceptance behaviors (see for example [33]), which would necessitate more complex models.

Chapter 5

Conclusion

This dissertation has presented three projects that develop data-driven optimization frameworks for sequential decision-making under uncertainty, spanning healthcare, environmental policy, and robotics.

In Chapter 2, we addressed the problem of optimizing mechanical ventilation control in intensive care units. We proposed an interpretable reinforcement learning methodology that combines decision tree policies with a causal, matching-based off-policy evaluation framework. Using real-world ICU data from the MIMIC-III database, we demonstrated that our interpretable policy achieves performance competitive with deep reinforcement learning approaches while offering transparency that is important for clinical adoption. A natural extension of this work, currently being pursued, is to augment the state representation with information extracted from clinical notes using language model embeddings, which could provide richer patient context and improve decision-making at each ventilation step.

In Chapter 3, we proposed a differential tax mechanism for sustainable groundwater management, formulated as a Stackelberg game between a public water agency and strategic water consumers. We derived a single-level reformulation of the bilevel optimization problem, established structural properties of the optimal policy including comparative statics and continuity of the compliance threshold, and proved that the differential tax can replicate the centralized planner’s first-best outcome. Numerical experiments calibrated to the Western La Mancha aquifer showed that the dual tax scheme achieves substantially higher social welfare than a uniform tax regime. Future work could incorporate stochastic water availability, dynamic enforcement strategies, and crop growth dynamics into the model.

In Chapter 4, we developed a framework for profit maximization in a Robotics-as-a-Service business, where the operator must jointly optimize pricing and robot replacement decisions under uncertainty about customer preferences and equipment degradation. We formulated the problem as a semi-Markov decision process and proposed an online learning framework that integrates survival analysis for degradation modeling with inverse optimization for estimating customer utility. Numerical experiments demonstrated that the learned policy converges to near-optimal profitability and substantially outperforms myopic baselines. A possible future direction is to consider stochastic customer acceptance models,

which would require more complex estimation methods.

Across these three projects, the common thread has been the use of mathematical optimization in conjunction with statistical learning to derive actionable policies from data. Each chapter addresses a setting where classical approaches fall short, and develops methods that combine the strengths of both. We hope this work contributes to the broader effort of making data-driven decision-making more rigorous and more practically useful across application domains.

Bibliography

- [1] Acute Respiratory Distress Syndrome Network, Roy G. Brower, Michael A. Matthay, Alan Morris, David Schoenfeld, B. Taylor Thompson, and Arthur Wheeler. Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. *New England Journal of Medicine*, 342(18):1301–1308, May 2000.
- [2] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 127–135. PMLR, 2013.
- [3] Anil Aswani and Claire Tomlin. Incentive design for efficient building quality of service. In *Proceedings of the 50th Annual Allerton Conference on Communication, Control, and Computing*, pages 90–97, October 2012.
- [4] Anil Aswani, Zuo-Jun Shen, and Auyon Siddiq. Inverse optimization with noisy data. *Operations Research*, 66(3):870–892, 2018.
- [5] Anil Aswani, Zuo-Jun Max Shen, and Auyon Siddiq. Data-driven incentive design in the Medicare shared savings program. *Operations Research*, 67(4):1002–1026, 2019.
- [6] Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021. doi: 10.1287/mnsc.2020.3605. URL <https://doi.org/10.1287/mnsc.2020.3605>.
- [7] Giacomo Bellani, John G. Laffey, Tai Pham, Eddy Fan, Laurent Brochard, Andres Esteban, Luciano Gattinoni, Frank van Haren, Anders Larsson, Daniel F. McAuley, Marco Ranieri, Gordon Rubenfeld, B. Taylor Thompson, Hermann Wrigge, Arthur S. Slutsky, Antonio Pesenti, LUNG SAFE Investigators, and ESICM Trials Group. Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. *JAMA*, 315(8):788–800, February 2016. ISSN 0098-7484. doi: 10.1001/jama.2016.0291.
- [8] Andrew Bennett and Nathan Kallus. Policy evaluation with latent confounders via optimal balance. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox,

- and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/7c4bf50b715509a963ce81b168ca674b-Paper.pdf.
- [9] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control, Vol. II: Approximate Dynamic Programming*. Athena Scientific, Belmont, MA, 4 edition, 2012.
- [10] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control, Vol. I*. Athena Scientific, Belmont, MA, 4 edition, 2017.
- [11] Marta Biancardi, Gianluca Iannucci, and Giovanni Villani. Inter-temporal decisions, optimal taxation and non-compliant behaviors in groundwater management. *Communications in Nonlinear Science and Numerical Simulation*, 116:106872, 2023. ISSN 1007-5704. doi: <https://doi.org/10.1016/j.cnsns.2022.106872>. URL <https://www.sciencedirect.com/science/article/pii/S1007570422003598>.
- [12] Marc F P Bierkens and Yoshihide Wada. Non-renewable groundwater use and groundwater depletion: a review. *Environmental Research Letters*, 14(6):063002, may 2019. doi: [10.1088/1748-9326/ab1a5f](https://doi.org/10.1088/1748-9326/ab1a5f). URL <https://doi.org/10.1088/1748-9326/ab1a5f>.
- [13] David G. Blauvelt, Hussam S. Inany, Jamie M. Furlong-Dillard, David K. Bailly, Peter Oishi, Martina A. Steurer, and Malini Mahendra. Association of ventilator settings with mortality in pediatric patients treated with extracorporeal life support for respiratory failure. *ASAIO Journal*, 68(12):1536–1543, dec 2022. doi: [10.1097/MAT.0000000000001697](https://doi.org/10.1097/MAT.0000000000001697).
- [14] Norman E. Breslow. discussion on professor cox’s paper. *Journal of the Royal Statistical Society, Series B*, 34:216–217, 1972.
- [15] Martin Britos, Elizabeth Smoot, Kathleen D. Liu, B. Taylor Thompson, William Checkley, and Roy G. Brower. The value of positive end-expiratory pressure and FiO₂ criteria in the definition of the acute respiratory distress syndrome. *Critical Care Medicine*, 39(9):2025–2030, September 2011. ISSN 0090-3493.
- [16] M. Alan Brookhart, Til Stürmer, Robert J. Glynn, Jeremy Rassen, and Sebastian Schneeweiss. Confounding control in healthcare database research: Challenges and potential approaches. *Medical Care*, 48(6):S114–S120, 2010. ISSN 0025-7079.
- [17] Ellen M. Bruno, Katrina K. Jessoe, and W. Michael Hanemann. The dynamic impacts of pricing groundwater. *Journal of the Association of Environmental and Resource Economists*, 11(5):1201–1227, September 2024. doi: [10.1086/728988](https://doi.org/10.1086/728988). URL <https://doi.org/10.1086/728988>.

- [18] Achim Buerkle, William Eaton, Ali Al-Yacoub, Melanie Zimmer, Peter Kinnell, Michael Henshaw, Matthew Coombes, Wen-Hua Chen, and Niels Lohse. Towards industrial robots as a service (IRaaS): Flexibility, usability, safety and business models. *Robotics and Computer-Integrated Manufacturing*, 81:102484, 2023.
- [19] Thyago P. Carvalho, Fabrizzio A. A. M. N. Soares, Roberto Vita, Roberto da P. Francisco, João P. Basto, and Symone G. S. Alcalá. A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, 137:106024, 2019.
- [20] Frank J. Chaloupka, Ayda Yurekli, and Geoffrey T. Fong. Tobacco taxes as a tobacco control strategy. *Tobacco Control*, 21(2):172–180, March 2012.
- [21] Timothy C. Y. Chan, Rafid Mahmood, and Ian Yihang Zhu. Inverse optimization: Theory and applications. *Operations Research*, 73(2):1046–1074, 2025. doi: 10.1287/opre.2022.0382.
- [22] Yinong Chen, Zhihui Du, and Marcos Garcia-Acosta. Robot as a service in cloud computing. In *2010 Fifth IEEE International Symposium on Service Oriented System Engineering*, pages 151–158. IEEE, 2010.
- [23] Adam P. Cohen, Susan A. Shaheen, and Emily M. Farrar. Urban air mobility: History, ecosystem, market potential, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(9):6074–6087, 2021.
- [24] David Collett. *Modelling Survival Data in Medical Research*. Chapman and Hall/CRC, Boca Raton, FL, 3 edition, 2015. doi: <https://doi.org/10.1201/b18041>.
- [25] David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.
- [26] Gerard F. Curley, John G. Laffey, Haibo Zhang, and Arthur S. Slutsky. Biotrauma and ventilator-induced lung injury. *Chest*, 150(5):1109–1117, November 2016. ISSN 0012-3692.
- [27] Chris Dann, Yishay Mansour, Mehryar Mohri, Ayush Sekhari, and Karthik Sridharan. Guarantees for epsilon-greedy reinforcement learning with function approximation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 4666–4689. PMLR, 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/dann22a.html>.
- [28] Coral Davenport. Strawberry case study: What if farmers had to pay for water? *New York Times*, December 2023.

- [29] Julia de Frutos Cachorro, Katrin Erdlenbruch, and Mabel Tidball. Optimal adaptation strategies to face shocks on groundwater resources. *Journal of Economic Dynamics and Control*, 40:134–153, 2014. ISSN 0165-1889. doi: <https://doi.org/10.1016/j.jedc.2014.01.005>. URL <https://www.sciencedirect.com/science/article/pii/S0165188914000141>.
- [30] Floris den Hengst, Martijn Otten, Paul Elbers, Frank van Harmelen, Vincent François-Lavet, and Mark Hoogendoorn. Guideline-informed reinforcement learning for mechanical ventilation in critical care. *Artificial Intelligence in Medicine*, 147:102742, 2024. ISSN 0933-3657. doi: <https://doi.org/10.1016/j.artmed.2023.102742>. URL <https://www.sciencedirect.com/science/article/pii/S0933365723002567>.
- [31] Tharindu Dharmasena, Rajitha de Silva, Nimsiri Abhayasingha, and Pradeep Abeygunawardhana. Autonomous cloud robotic system for smart agriculture. In *2019 Moratuwa Engineering Research Conference (MERCon)*, pages 388–393. IEEE, 2019.
- [32] Maria Dimakopoulou, Zhengyuan Zhou, Susan Athey, and Guido Imbens. Balanced linear contextual bandits. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33013445. URL <https://doi.org/10.1609/aaai.v33i01.33013445>.
- [33] Ilgin Dogan, Zuo-Jun Max Shen, and Anil Aswani. Estimating and incentivizing imperfect-knowledge agents with hidden rewards. *arXiv preprint arXiv:2308.06717*, 2023.
- [34] Ilgin Dogan, Zuo-Jun Max Shen, and Anil Aswani. Repeated principal-agent games with unobserved agent rewards and perfect-knowledge agents. *arXiv preprint arXiv:2304.07407*, 2023.
- [35] Paul M. Dorinsky and Michael E. Whitcomb. The effect of PEEP on cardiac output. *Chest*, 84(2):210–216, August 1983.
- [36] Kevin Dorling, Jordan Heinrichs, Geoffrey G. Messier, and Sebastian Magierowski. Vehicle routing problems for drone delivery. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(1):70–85, 2017. doi: 10.1109/TSMC.2016.2582745.
- [37] S. Eloranta, K. E. Smedby, P. W. Dickman, and T. M. Andersson. Cancer survival statistics for patients and healthcare professionals—a tutorial of real-world data analysis. *Journal of Internal Medicine*, 289(1):12–28, 2021.
- [38] Harry Emerson, Matthew Guy, and Ryan McConville. Offline reinforcement learning for safer blood glucose control in people with type 1 diabetes. *Journal of Biomedical Informatics*, 142:104376, June 2023. ISSN 1532-0464.

- [39] Encarna Esteban and José Albiac. Groundwater and ecosystems damages: Questioning the Gisser–Sánchez effect. *Ecological Economics*, 70(11):2062–2069, September 2011. ISSN 0921-8009.
- [40] Niall D. Ferguson, Eddy Fan, Luigi Camporota, Massimo Antonelli, Antonio Anzueto, Richard Beale, Laurent Brochard, Roy Brower, Andrés Esteban, Luciano Gattinoni, Andrew Rhodes, Arthur S. Slutsky, Jean-Louis Vincent, Gordon D. Rubenfeld, B. Taylor Thompson, and V. Marco Ranieri. The Berlin definition of ARDS: an expanded rationale, justification, and supplementary material. *Intensive Care Medicine*, 38(10):1573–1582, August 2012. ISSN 1432-1238. doi: 10.1007/s00134-012-2682-1.
- [41] GE Aerospace. Maintenance Insight. <https://www.geaerospace.com/systems/saas/maintenance-insight>, 2024. Accessed on 2024-09-14.
- [42] Micha Gisser and Abraham Mercado. Economic aspects of ground water resources and replacement flows in semiarid agricultural areas. *American Journal of Agricultural Economics*, 55(3):461–466, August 1973.
- [43] Mark Gongloff. Taxing groundwater is a sure way to preserve it. *Bloomberg Opinion*, January 2024.
- [44] R. Q. Grafton, J. Williams, C. J. Perry, F. Molle, C. Ringler, P. Steduto, B. Udall, S. A. Wheeler, Y. Wang, D. Garrick, and R. G. Allen. The paradox of irrigation efficiency. *Science*, 361(6404):748–750, August 2018. ISSN 1095-9203.
- [45] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2024. URL <https://www.gurobi.com>.
- [46] Jeffrey M. Halter, Jay M. Steinberg, Louis A. Gatto, Joseph D. DiRocco, Lucio A. Pavone, Henry J. Schiller, Scott Albert, Hsi-Ming Lee, David Carney, and Gary F. Nieman. Effect of positive end-expiratory pressure and tidal volume on lung injury induced by alveolar instability. *Critical Care*, 11(1):R20, 2007.
- [47] Andrew Hinton, Marta Kwiatkowska, Gethin Norman, and David Parker. PRISM: A tool for automatic verification of probabilistic systems. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 441–444. Springer, 2006.
- [48] Guoqiang Hu, Wee Peng Tay, and Yonggang Wen. Cloud robotics: architecture, challenges and applications. *IEEE Network*, 26(3):21–28, 2012.
- [49] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, May 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35.

- [50] Md Rezaul Karim and M. Ataharul Islam. *Reliability and Survival Analysis*. Springer, 2019.
- [51] Ben Kehoe, Sachin Patil, Pieter Abbeel, and Ken Goldberg. A survey of research on cloud robotics and automation. *IEEE Transactions on Automation Science and Engineering*, 12(2):398–409, 2015.
- [52] Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, November 2018. ISSN 1546-170X.
- [53] Flemming Kondrup, Thomas Jiralerspong, Elaine Lau, Nathan de Lara, Jacob Shkrob, My Duc Tran, Doina Precup, and Sumana Basu. Towards safe mechanical ventilation treatment using deep offline reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15696–15702, September 2023. doi: 10.1609/aaai.v37i13.26862.
- [54] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 1179–1191, 2020.
- [55] Dhananjay Kumar and Bengt Klefsjö. Proportional hazards model: a review. *Reliability Engineering & System Safety*, 44(2):177–188, 1994.
- [56] Jean-Jacques Laffont and David Martimort. *The Theory of Incentives: The Principal-Agent Model*. Princeton University Press, 2002. ISBN 9780691091846.
- [57] Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3703–3712. PMLR, 09–15 Jun 2019. URL <https://proceedings.mlr.press/v97/le19a.html>.
- [58] Andrew Leahey. Taxing water is the only way to limit use and reduce depletion. *Bloomberg Tax*, January 2024.
- [59] Hyeonhoon Lee, Hyun-Kyu Yoon, Jaewon Kim, Ji Soo Park, Chang-Hoon Koo, Dongwook Won, and Hyung-Chul Lee. Development and validation of a reinforcement learning model for ventilation control during emergence from general anesthesia. *npj Digital Medicine*, 6(1), August 2023. ISSN 2398-6352.
- [60] Joo Seung Lee and Anil Aswani. Game-theoretic incentive design for sustainable groundwater management. *IFAC-PapersOnLine*, 59(9):259–264, 2025.
- [61] Joo Seung Lee and Anil Aswani. Profit maximization for a robotics-as-a-service model. *arXiv preprint arXiv:2509.26595*, 2025.

- [62] Joo Seung Lee, Malini Mahendra, and Anil Aswani. Matching-based off-policy evaluation for reinforcement learning applied to mechanical ventilation. In *Proceedings of the ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies*, CHASE '25, pages 211–221, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715396. doi: 10.1145/3721201.3721370.
- [63] Qiaomei Li, Kara L Gavin, Corrine I Voils, and Yonatan Mintz. An adaptive optimization approach to personalized financial incentives in behavioral interventions. *Production and Operations Management*, 35(2):546–565, 2026. doi: 10.1177/10591478251349391. URL <https://doi.org/10.1177/10591478251349391>.
- [64] Ying Liu, Brent Logan, Ning Liu, Zhiyuan Xu, Jian Tang, and Yangzhi Wang. Deep reinforcement learning for dynamic treatment regimes on medical registry data. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 380–385. IEEE, August 2017.
- [65] Ilan Lobel, Renato Paes Leme, and Adrian Vladu. Multidimensional binary search for contextual decision-making. *Operations Research*, 66(5):1346–1361, October 2018. ISSN 1526-5463.
- [66] Sérgio Henrique Loss, Roselaine Pinheiro de Oliveira, Juçara Gasparetto Maccari, Augusto Savi, Marcio Manozzo Boniatti, Márcio Pereira Hetzel, Daniele Munaretto Dalleggrave, Patrícia de Campos Balzano, Eubrandio Silvestre Oliveira, Jorge Amilton Höher, André Peretti Torelly, and Cassiano Teixeira. The reality of patients requiring prolonged mechanical ventilation: a multicenter study. *Revista Brasileira de Terapia Intensiva*, 27(1):26–35, 2015.
- [67] Zhiyao Luo, Yangchen Pan, Peter Watkinson, and Tingting Zhu. Position: Reinforcement learning in dynamic treatment regimes needs critical reexamination. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 33432–33465. PMLR, July 2024.
- [68] J. J. Marini and S. A. Ravenscraft. Mean airway pressure: physiologic determinants and clinical importance—part 2: Clinical implications. *Critical Care Medicine*, 20(11): 1604–1616, November 1992.
- [69] Annalisa Milella, Stefan Rilling, Arianna Rana, Rocco Galati, Antonio Petitti, Mark Hoffmann, Jacob Livin Stanly, and Giulio Reina. Robot-as-a-service as a new paradigm in precision farming. *IEEE Access*, 12:47942–47949, 2024.
- [70] Yonatan Mintz, Anil Aswani, Philip Kaminsky, Elena Flowers, and Yoshimi Fukuoka. Behavioral analytics for myopic agents. *European Journal of Operational Research*, 310(2):793–811, 2023.

- [71] François Molle and Alvar Closas. Why is state-centered groundwater governance largely ineffective? a review. *WIREs Water*, 7(1), October 2019. ISSN 2049-1948.
- [72] Elizbar Nadaraya. On estimating regression. *Theory of Probability and Its Applications*, 9:141–142, 1964.
- [73] Thesath Nanayakkara, Gilles Clermont, Christopher James Langmead, and David Swigon. Unifying cardiovascular modelling with deep reinforcement learning for uncertainty aware control of sepsis treatment. *PLOS Digital Health*, 1(2):1–20, February 2022.
- [74] Michael Oberst and David Sontag. Counterfactual off-policy evaluation with Gumbel-Max structural causal models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4881–4890. PMLR, 2019.
- [75] James M. O’Brien, Naeem A. Ali, Scott K. Aberegg, and Edward Abraham. Sepsis. *The American Journal of Medicine*, 120(12):1012–1022, 2007. ISSN 0002-9343.
- [76] Mahsa Oroojeni Mohammad Javad, Stephen Olusegun Agboola, Kamal Jethwani, Abe Zeid, and Sagar Kamarthi. A reinforcement learning-based method for management of type 1 diabetes: Exploratory study. *JMIR Diabetes*, 4(3):e12905, August 2019. ISSN 2371-4379.
- [77] Dimitris Papathanasiou, Konstantinos Demertzis, and Nikos Tziritas. Machine failure prediction using survival analysis. *Future Internet*, 15(5):153, 2023.
- [78] Sonali Parbhoo, Jasmina Bogojenska, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. Combining kernel and model based learning for HIV therapy selection. *AMIA Joint Summits on Translational Science Proceedings*, 2017:239–248, July 2017.
- [79] Arne Peine, Ahmed Hallawa, Johannes Bickenbach, Guido Dartmann, Lejla Begic Fazlic, Anke Schmeink, Gerd Ascheid, Christoph Thiernemann, Andreas Schuppert, Ryan Kindle, Leo Celi, Gernot Marx, and Lukas Martin. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. *npj Digital Medicine*, 4(1):32, February 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00388-6.
- [80] Jean-Christophe Pereau, Alexandre Pryet, and Tina Rambonilaza. Optimality versus viability in groundwater management with environmental flows. *Ecological Economics*, 161:109–120, July 2019. ISSN 0921-8009.
- [81] Arthur Cecil Pigou and Nahid Aslanbeigui. *The economics of welfare*. Routledge, October 2017.

- [82] Roselaine Pinheiro de Oliveira, Márcio Pereira Hetzel, Mauro dos Anjos Silva, Daniele Dallegrave, and Gilberto Friedman. Mechanical ventilation with high tidal volume induces inflammation in patients without lung disease. *Critical Care*, 14(2), March 2010. ISSN 1364-8535.
- [83] Niranjani Prasad, Li-Fang Cheng, Corey Chivers, Michael Draugelis, and Barbara E. Engelhardt. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.
- [84] Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 759–766. Morgan Kaufmann Publishers Inc., 2000.
- [85] Sonal Rachmale, Guangxi Li, Gregory Wilson, Michael Malinchoc, and Ognjen Gajic. Practice of excessive FiO₂ and effect on pulmonary outcomes in mechanically ventilated patients with acute lung injury. *Respiratory Care*, 57(11):1887–1893, November 2012. ISSN 1943-3654.
- [86] Aaron M. Roth, Nicholay Topin, Pooyan Jamshidi, and Manuela Veloso. Conservative Q-improvement: Reinforcement learning for an interpretable decision-tree policy. *arXiv preprint arXiv:1907.01180*, 2019.
- [87] Antoine Scheid, Daniil Tiapkin, Etienne Boursier, Aymeric Capitaine, Eric Moulines, Michael I. Jordan, El-Mahdi El-Mhamdi, and Alain Durmus. Incentivized learning in principal-agent bandit games. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.
- [88] Lynn M. Schnapp and Neal H. Cohen. Pulse oximetry: Uses and abuses. *Chest*, 98(5):1244–1250, 1990. ISSN 0012-3692.
- [89] Judy E. Scott and Carlton H. Scott. Drone delivery models for healthcare. In *Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS)*, 2017.
- [90] Takuma Seno and Michita Imai. d3rlpy: An offline deep reinforcement learning library. *Journal of Machine Learning Research*, 23, October 2022. ISSN 1532-4435.
- [91] Arthur S. Slutsky. Mechanical ventilation. *Chest*, 104(6):1833–1859, 1993. ISSN 0012-3692.
- [92] Arthur S. Slutsky and V. Marco Ranieri. Ventilator-induced lung injury. *New England Journal of Medicine*, 369(22):2126–2136, November 2013. ISSN 1533-4406.

- [93] Kenneth A Small, Erik T Verhoef, and Robin Lindsey. *The economics of urban transportation*. Routledge, London, England, 2 edition, October 2007.
- [94] Anna Straubinger, Raoul Rothfeld, Michael Shamiyeh, Kai-Daniel Büchter, Jochen Kaiser, and Kay Olaf Plötner. An overview of current research and developments in urban air mobility—setting the scene for UAM introduction. *Journal of Air Transport Management*, 87:101852, 2020.
- [95] Manuel Toscano-Moreno, Juan Bravo-Arrabal, Manuel Sánchez-Montero, Javier Serón Barba, Ricardo Vázquez-Martín, Juan Jesús Fernández-Lozano, Anthony Mandow, and Alfonso Garcia-Cerezo. Integrating ROS and Android for rescuers in a cloud robotics architecture: Application to a casualty evacuation exercise. In *2022 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pages 270–276. IEEE, 2022.
- [96] United States Department of Defense. Military handbook: Reliability prediction of electronic equipment. Technical Report MIL-HDBK-217F, Department of Defense, Washington, DC, USA, December 1991.
- [97] Willem van den Boom, Michael Hoy, Jagadish Sankaran, Mengru Liu, Haroun Chahed, Mengling Feng, and Kay Choong See. The search for optimal oxygen saturation targets in critically ill patients: Observational data from large ICU databases. *Chest*, 157(3): 566–573, 2020. ISSN 0012-3692.
- [98] Jesús Villar, Lina Pérez-Méndez, Jesús Blanco, José Manuel Añón, Lluís Blanch, Javier Belda, Antonio Santos-Bouza, Rosa Lidia Fernández, Robert M. Kacmarek, and Spanish Initiative for Epidemiology, Stratification, and Therapies for ARDS (SIESTA) Network. A universal definition of ARDS: the PaO₂/FiO₂ ratio under a standard ventilatory setting—a prospective, multicenter validation study. *Intensive Care Medicine*, 39 (4):583–592, April 2013.
- [99] Eric Vittinghoff, David V. Glidden, Stephen C. Shiboski, and Charles E. McCulloch. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. Springer, New York, NY, 2 edition, 2012.
- [100] Lubing Wang, Zhengbo Zhu, and Xufeng Zhao. Dynamic predictive maintenance strategy for system remaining useful life prediction via deep learning ensemble method. *Reliability Engineering & System Safety*, 245:110012, 2024.
- [101] Geoffrey S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 26(4):359–372, 1964. ISSN 0581572X.
- [102] Hehua Yan, Qingsong Hua, Yingying Wang, Wenguo Wei, and Muhammad Imran. Cloud robotics in smart manufacturing environments: Challenges and countermeasures. *Computers & Electrical Engineering*, 63:56–65, 2017.

- [103] Chao Yu, Yinzhaoh Dong, Jiming Liu, and Guoqi Ren. Incorporating causal factors into reinforcement learning for dynamic treatment regimes in HIV. *BMC Medical Informatics and Decision Making*, 19(2):60, April 2019. ISSN 1472-6947.
- [104] Weiting Zhang, Dong Yang, and Hongchao Wang. Data-driven methods for predictive maintenance of industrial equipment: A survey. *IEEE Systems Journal*, 13(3): 2213–2227, 2019.
- [105] Xiang-Sun Zhang and De-Gang Liu. A note on the continuity of solutions of parametric linear programs. *Mathematical Programming*, 47(1–3):143–153, May 1990. ISSN 1436-4646.

Appendix A

Supplementary Material for Chapter 2

A.1 List of ICD-9 Codes Relevant to Sepsis

The following ICD-9 codes were used to identify patients with sepsis in the MIMIC-III database:

024	036.3	038.0	038.2	038.3	038.40
038.41	038.42	038.43	038.44	038.49	038.8
038.9	098.89	360.00	519.01	522.4	528.3
567.22	599.0	614.9	630	632	634
635	636	637	638	639	659.3
670.2	670.3	672.00	682.9	771.3	771.81
771.89	785.52	791	995.91	995.92	995.94
996.64	998.02	998.59	999.31	999.39	

Table A.1: ICD-9 codes used for sepsis patient identification.

A.2 Action Space Binning Intervals

The continuous action space for mechanical ventilation was discretized into bins as follows:

bin	Vt _{set}	PEEP	FiO ₂
1	[0, 3.9)	[0, 7)	[0, 0.36)
2	[3.9, 5.37)	[7, 11)	[0.36, 0.45)
3	[5.37, 6.55)	[11, 16)	[0.45, 0.55)
4	[6.55, 7.74)	[16, ∞)	[0.55, 0.65)
5	[7.74, 9.12)		[0.65, 0.76)
6	[9.12, 11.11)		[0.76, 0.89)
7	[11.11, ∞)		[0.89, 1]

Table A.2: Action space binning intervals for ventilator settings

Since the action space is discretized into 7 levels, action indices ≥ 6 for Vt_{set} and ≥ 4 for FiO₂ are considered aggressive, as they correspond to the thresholds 10 mL/kg and 0.6, respectively.

A.3 Transition Model Feature Grouping

The state features used in the transition model were grouped as follows:

1. **Respiratory:** Respiratory rate, Spontaneous tidal volume, PaO₂FiO₂ Ratio, Mean airway pressure
2. **Hemodynamic:** Heart rate, Systolic BP, Diastolic BP
3. **Blood Gas:** SpO₂, PaCO₂, PaO₂
4. **Miscellaneous:** Sepsis, Weight, Age, GCS, Cumulative Fluid Balance

A.4 Transition Model Hyperparameters

The following hyperparameter ranges were explored during model training:

1. $h_{s,h}$ (hemodynamic bandwidth): [2.036, 2.536, 3.036]
2. $h_{s,r}$ (respiratory bandwidth): [1.8, 2.3, 2.8]
3. $h_{s,b}$ (blood gas bandwidth): [1.532, 2.032, 2.532]
4. $h_{s,m}$ (miscellaneous bandwidth): [1.532, 2.032, 2.532]
5. h_a (action bandwidth): [1.0, 1.5, 2.0]
6. h_z (additional bandwidth): [0.5, 1.0, 1.5, 2.0]
7. λ (regularization): [10^{-2} , 10^{-3} , 10^{-4}]