OPTIMAL & GAME THEORETIC FEEDBACK DESIGN FOR EFFICIENT HUMAN PERFORMANCE IN HUMAN-SUPERVISED AUTONOMY

By

Piyush Gupta

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Electrical Engineering—Doctor of Philosophy

2023

ABSTRACT

Human-in-the-loop systems play a pivotal role in numerous safety-critical applications, ensuring both safety and efficiency in complex operational environments. However, these systems face a significant challenge stemming from the inherent variability in human performance, influenced by factors such as workload, fatigue, task learning, expertise, and individual differences. Therefore, effective management of human cognitive resources is paramount in designing efficient human-in-the-loop systems.

To address this challenge, it is critical to design robust and adaptive systems capable of continuously adapting models of human performance, and subsequently providing tailored feedback to enhance it. Effective feedback mechanisms play a pivotal role in improving the overall system performance by optimizing human workload, fostering skill development, and facilitating efficient collaboration among individuals within diverse human teams, each with their unique skill sets and expertise.

In this dissertation, the primary focus lies in exploring optimal and game-theoretic approaches for feedback design to enhance system performance, particularly in scenarios where humans are integral components. We begin by studying the problem of optimal fidelity selection for a human operator servicing a stream of homogeneous tasks, where fidelity refers to the degree of exactness and precision while servicing the task. Initially, we assume a known human service time distribution model, later relaxing this assumption. We design a human decision support system that recommends optimal fidelity levels based on the operator's cognitive state and queue length. We evaluate our methods through human experiments involving participants searching for underwater mines.

We extend the optimal fidelity selection problem by incorporating uncertainty into the human service-time distribution. This extension involves the development of a robust and adaptive framework that accurately learns the human service-time model and adapts the policy while ensuring robustness under model uncertainty. However, a major challenge in designing adaptive and robust systems arises from the conflicting objectives of exploration and robustness. To mitigate system uncertainty, an agent must explore high-uncertainty state space regions, while robust policy optimization seeks to avoid these regions due to poor worst-case performance. To address this trade-off, we introduce an efficient Deterministic Sequencing of Exploration and Exploitation (DSEE) algorithm for model-based reinforcement learning. DSEE interleaves exploration and exploitation epochs with increasing lengths, resulting in sub-linear cumulative regret growth over time.

In addition to cognitive resource management, enhancing human performance can also be achieved through tutoring for skill development. In this context, we study the impact of evaluative feedback on human learning in sequential decision-making tasks. We conduct experiments on Amazon Mechanical Turk, where participants engage with the Tower of Hanoi puzzle and receive AI-generated feedback during their problem-solving. We examine how this feedback influences their learning and skill transfer to related tasks. Additionally, we explore computational models to gain insights into how individuals integrate evaluative feedback into their decision-making processes.

Lastly, we expand our focus from a single human operator to a team of heterogeneous agents, each with diverse skill sets and expertise. Within this context, we delve into the challenge of achieving efficient collaboration among heterogeneous team members to enhance overall system performance. Our approach leverages a game theoretic framework, where we design utility functions to incentivize decentralized collaboration among these agents.

This dissertation is dedicated to my beloved family.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the support and encouragement of several people. First and foremost, my advisor, Dr. Vaibhav Srivastava, deserves special recognition. He has been there for me every step of the way, guiding me and supporting me in my academic journey. He has helped me improve my writing and analytical skills significantly. I am thankful for his unwavering support and for always making time for me during my PhD. I would also like to extend my gratitude to my graduate committee members, Dr. Xiaobo Tan, Dr. Hayder Radha, and Dr. Vishnu Boddeti, for their constant encouragement and guidance throughout my time in graduate school. They have played a significant role in my academic development.

I extend my gratitude to Dr. Shaunak D. Bopardikar for his unwavering support during my PhD journey and invaluable insights into common resource pool games and analysis. Additionally, I would like to thank Dr. Joshua E. Siegel for our engaging discussions on various aspects of autonomous vehicles. These discussions inspired me to pursue internship opportunities at the Honda Research Institute (HRI), US, in 2022-2023.

During my three internships at HRI, I had the privilege of working on problems related to interaction-aware planning and applying the skills and knowledge I gained during my PhD in the field of autonomous driving. I am deeply appreciative of my mentors, Sangjae Bae and David Isele from HRI, for their unwavering support throughout these internships. Their guidance not only broadened my research horizons but also played a significant role in refining my coding and writing abilities.

I would like to express my gratitude to my friends and lab mates, Ankur, Dong Hae, Pradyumna Reddy, and Abhishek, for the enduring sense of camaraderie in the lab and the close friendships we've cultivated. Additionally, I want to thank my friends in California, Faizan Tariq, Sophia Newton, Aakriti Kumar, Snehal Dikhale, Faizan Siddiqui, and Avinash Singh with whom I formed a special bond during my internships at HRI. Our weekend hiking adventures, shared laughter, and the wisdom I gained from their experiences have continually

inspired me to aim higher in life.

No amount of words can adequately convey my gratitude for the unwavering love and wholehearted support of my family. I want to express my heartfelt thanks to my parents for instilling in me the values that have guided my path. Each day, I gain a deeper understanding of the sacrifices they made while raising me. My elder sister has been a constant source of guidance and support throughout the years, and from my younger brother, I have learned the importance of taking breaks and finding positivity during tough times. I am immensely grateful for their unwavering presence during challenging moments. My family has been a guiding light in my life, inspiring me on my journey. I can never thank them enough for everything they've done for me.

I am deeply grateful to both my high school mathematics teacher, Mr. Sanjay Singh, and my IIT Delhi professor, Dr. Supreet Singh Bahga, for their early influence that ignited my appreciation for mathematics. They have collectively been unwavering sources of support and inspiration throughout my academic journey. I want to express my heartfelt thanks to both of them for consistently being there to provide guidance whenever I needed it.

The work reported in this dissertation was supported in parts by the NSF awards IIS-1734272, CMMI-1940950, ECCS-2024649, and ONR award N00014-22-1-2813.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Modern-day computers are getting faster and more efficient; allowing automation to perform complex tasks [1]. Nevertheless, in numerous safety-critical domains, human involvement remains essential to guarantee both safety and efficiency. These human-in-the-loop systems have become ubiquitous, finding application in fields such as search and rescue [2–4], semi-autonomous vehicle systems [5, 6], robot-assisted surgery [7], and flight control [2].

The role of humans within these systems is notably application-dependent, ranging from mere supervision to direct teleoperation of robots. For instance, in a search and rescue scenario [2], human operators might be tasked with remotely controlling robots to conduct search operations, aid in victim rescue efforts, and verify the findings of autonomous agents.

Nonetheless, these systems encounter a noteworthy challenge rooted in the inherent variability of human performance, which can be influenced by factors including workload, fatigue, task learning, expertise, and individual differences. Furthermore, the prevailing goal of maximizing the ratio of robots to human operators, often driven by cost considerations, frequently results in increased workload conditions for human operators [8]. As a consequence, the effective management of human cognitive resources becomes paramount in the design of efficient human-in-the-loop systems.

To address this challenge, it is imperative to develop robust and adaptive systems that can continuously learn and adapt their models of human performance. These systems should utilize these learned models to actively provide customized feedback to humans, thus enhancing their performance. Effective feedback mechanisms can elevate the overall system performance by optimizing the human workload, fostering the development of skills, and facilitating efficient collaboration among individuals within diverse human teams, each with their unique skill sets and expertise.

This dissertation primarily focuses on the exploration of optimal and game-theoretic approaches for feedback design, with a specific emphasis on enhancing system performance, particularly in scenarios where humans are integral components. The key objectives include:

(i) *Feedback design for managing human cognitive resources:* A central aspect of this research focuses on designing feedback strategies that enable human operators to manage their cognitive resources efficiently. Specifically, these strategies offer optimal recommendations, based on the system state and cognitive state of the human operator, that optimize the utilization of human cognitive resources for improved system performance.

(ii) *Adaptive Algorithms for Robust Feedback under Model Uncertainty:* Human performance is inherently agent-specific, influenced by a variety of factors such as workload, fatigue, and trust. Therefore, we delve into the design of robust and adaptable algorithms. These algorithms possess the ability to continuously learn and refine their model of human behavior, adapting and changing feedback based on updated human models. Importantly, the proposed adaptive recommendations optimize worst-case performance, and consequently, remain robust in the face of uncertainties in the human model.

(iii) *Role of feedback in fostering skill-development:* Enhancing system performance can be achieved through effective tutoring and skill development in individuals performing the task. Therefore, we explore the role of feedback, extending beyond its immediate impact on current performance, to nurture skill development among human operators. Our objective is to use feedback as a means of providing effective tutoring, thereby fostering continuous growth and improvement over time.

(iv) *Feedback for Collaboration Across Skill sets:* In scenarios where multiple humans with diverse skill sets are involved, we explore how feedback incentives can be harnessed to improve system performance. This involves achieving efficient team collaboration, ensuring that each individual's unique expertise contributes to the collective success of the system.

To achieve these objectives, we begin by studying the problem of optimal fidelity selection for a human operator servicing a stream of homogeneous tasks, where fidelity refers to the degree of exactness and precision while servicing the task. Initially, we assume a known human service time distribution model, later relaxing this assumption. We design a human decision support system that recommends optimal fidelity levels based on the operator's cognitive state and queue length. We evaluate our methods through human experiments involving participants searching for underwater mines.

We extend the optimal fidelity selection problem by incorporating uncertainty into the human service-time distribution. This extension involves the development of a robust and adaptive framework that accurately learns the human service-time model and adapts the policy while ensuring robustness under model uncertainty. However, a major challenge in designing adaptive and robust systems arises from the conflicting objectives of exploration and robustness. To mitigate system uncertainty, an agent must explore high-uncertainty state space regions, while robust policy optimizes worst-case performance and consequently avoids these regions. To address this trade-off, we introduce an efficient Deterministic Sequencing of Exploration and Exploitation (DSEE) algorithm for model-based reinforcement learning (RL). DSEE interleaves exploration and robust exploitation epochs with increasing lengths, resulting in sub-linear cumulative regret growth over time.

In addition to cognitive resource management, enhancing human performance can also be achieved through tutoring for skill development. In this context, we study the impact of evaluative feedback on human learning in sequential decision-making tasks. We conduct experiments on Amazon Mechanical Turk, where participants engage with the Tower of Hanoi puzzle and receive AI-generated feedback during their problem-solving. We examine how this feedback influences their learning and skill transfer to related tasks. Additionally, we explore computational models to gain insights into how individuals integrate evaluative feedback into their decision-making processes.

Lastly, we expand our focus from a single human operator to a team of heterogeneous agents, each with diverse skill sets and expertise. Within this context, we delve into the challenge of achieving efficient collaboration among heterogeneous team members to enhance overall system performance. Our approach leverages a game-theoretic framework, where we design utility functions to incentivize decentralized collaboration among these agents. We show the existence of a unique Pure Nash Equilibrium (PNE) and establish the convergence of the best response dynamics to this unique PNE. Additionally, we establish an analytical upper bound on measures of PNE inefficiency, shedding light on the effectiveness of our collaborative strategies.

In summary, this dissertation seeks to advance feedback design techniques, particularly in the context of human-in-the-loop systems, by introducing optimal and game-theoretic approaches. By focusing on enhancing human performance, fostering skill development, and optimizing collaboration, this research aims to make significant contributions to the design and implementation of more efficient and effective human-machine systems.

1.1 Literature review

In this section, we review the literature in areas relevant to this dissertation. We organize the literature according to the broad topics of interest in this dissertation.

1.1.1 Control of Queues with State-dependent Servers

We start by studying the problem of optimal fidelity selection for a human operator servicing a queue of homogeneous tasks. We model it as a control of a queue problem, where the human acts as a server with its own cognitive dynamics.

Recent years have seen significant efforts in integrating human knowledge and perception skills with autonomy [9]. A key research theme within this area concerns the systematic allocation of human cognitive resources for efficient performance. Therein, some of the fundamental questions studied include optimal scheduling of the tasks to be serviced by the operator [10], enabling shorter operator reaction times by controlling the task release [11], and determining optimal operator attention allocation [12]. In contrast to the aforementioned works, we consider a semi-Markov decision process (SMDP) [13] formulation to deal with general (non-memoryless) service time distributions of the human operator. Furthermore, while the above works propose heuristic algorithms, we focus on establishing the structural properties of the optimal policy.

Some interesting recent studies with state-dependent queues are considered in [14, 15]. In these works, authors design scheduling policies that stabilize a queueing system and decrease the utilization rate of a non-preemptive server that measures the proportion of time the server is working. The performance of the server degrades with the increase in server utilization and improves when the server is allowed to rest. In contrast to monotonic server performance with the utilization rate in [14, 15], we model the service time of the human operator as a unimodal function of its cognitive state. Our model for service time is inspired by experimental psychology literature [16] and incorporates the influence of cognitive state and fidelity level on service time.

The optimal control of queueing systems [17] is a classical problem in queueing theory. Of particular interest are the works [18, 19], where authors study the optimal policies for an M/G/1 queue using an SMDP formulation and describe its qualitative features. In contrast to a standard control of queues problem, the server in our problem is a human operator with its cognitive dynamics that must be incorporated into the problem formulation.

Our mathematical techniques to establish the structural properties of the optimal policy are similar to [20]. In [20], the authors establish structural properties of an optimal transmission policy for transmitting packets over a point-to-point channel in communication networks. The optimal policy of their Markov decision process (MDP) [21] depends on the queue length, the number of packet arrivals, and the channel fading state. In [22], authors study structural properties of the optimal resource allocation policy for a single-queue system in which a central decision-maker assigns servers to each job. In contrast to [20, 22], a major challenge in our problem arises due to SMDP formulation for non-memoryless service time distribution and its unimodal dependence on the cognitive state.

1.1.2 Robust Control Policies with Uncertain Dynamics

In this dissertation, we also study the optimal fidelity selection problem in the presence of uncertainty in human models. Specifically, we assume that the service time distribution of the human operator is unknown a priori and therefore, formulate a robust adaptive SMDP. An SMDP accounts for the system uncertainty through probabilistic state transitions. However, the obtained policy is sensitive to errors in the stochastic models [23, 24]. The large uncertainty in the service time models, especially in the initial stage with limited observation data may lead to sub-optimal policies. Existing methods formulate such control problems with model uncertainty as an MDP with an uncertain state transition model. In [25], authors propose a constrained MDP framework with risk constraints. They optimize Conditional Value-at-Risk (CVaR) [26] and propose an iterative offline algorithm to find the risk-constrained optimal control policy. In [27], a chance-constrained MDP [28] is proposed that provides a probabilistic framework for handling uncertainty in the transition probabilities. Robust MDPs are studied in [29, 30] that solve for policies that optimize a min-max criterion when the unknown stochastic model is assumed to lie within an uncertainty set. The policies obtained from solving robust MDPs can be overly conservative when implemented on the nominal system [27]. In our work, we consider a human agent with non-memoryless service time distribution and thus formulate a robust adaptive SMDP to deal with general service time distributions [18, 19] and learn a policy that is robust in the transient learning phase and converges to the optimal policy asymptotically.

In this dissertation, we show that the solution of the synchronous and asynchronous value iteration (VI) methods [31] for robust adaptive SMDP converges to the optimal solution for the uncertainty-free SMDP. While there exists a convergence analysis for the robust [32, 33] and adaptive [34] MDPs, such an analysis is missing for robust adaptive SMDPs to the best of our knowledge. A key challenge that we address in comparison to MDPs is the time dependence of the robust adaptive Bellman operator for SMDPs that requires a careful comparison between optimal value functions for intermediate SMDPs at different time steps, and the optimal value function for the uncertainty-free SMDP.

1.1.3 Efficient Algorithms for Model-Based Reinforcement Learning

In this dissertation, we introduce an efficient DSEE algorithm for model-based RL. RL is used in solving complex sequential decision-making tasks in uncertain environments such as motion planning for robots [35, 36], personalized web services [37, 38], and the design of decision-support systems for human-supervisory control [39–41]. MDPs [31] provide a natural framework for optimal decision-making under uncertainty and are used to model and solve numerous model-based RL problems. The objective of these problems is to simultaneously learn the system model and the optimal policy. While MDP formulation accounts for environment uncertainty by using stochastic models, MDP policies are known to be sensitive to errors in these stochastic models [17, 42].

In many safety-critical systems, robust MDPs [29, 43] are used to mitigate performance degradation due to uncertainty in the learned MDP. However, to reduce the system uncertainty, the agent must explore the environment and visit parts of the state space associated with high estimation uncertainty. Most often, RL algorithms use simple randomized methods to explore the environment, e.g. applying ϵ-greedy policies [44, 45] or adding random noise to continuous actions [46]. The objective of the robust MDPs conflicts with the exploration objective. To mitigate system uncertainty, an agent must explore high-uncertainty state space regions, while robust policy optimizes worst-case performance and consequently avoids these regions. Therefore, to balance the trade-off between learning the MDP and designing a robust policy, we design a DSEE algorithm, in which exploration and exploitation epochs of increasing lengths are interleaved.

There exist efficient algorithms for solving RL problems with provable bounds on the sample complexity [47, Definition 1]. In [47], authors analyze the Model-based Interval Estimation (MBIE) algorithm that applies confidence bounds to compute an optimistic policy and show that the algorithm is PAC-optimal [47, Definition 2]. They provide an upper bound on the algorithm's sample complexity given by $O\left(\frac{|\mathcal{S}|^2|\mathcal{A}|}{(1-\gamma)^6\epsilon^3}\log(\delta^{-1})\right)$ which is the maximum number of time steps until when the MBIE policy is not $\epsilon-$optimal with at least probability $1-\delta$, where $|\mathcal{S}|$, $|\mathcal{A}|$, are the cardinality of the state space and action space, respectively, γ is the discount factor, and $\epsilon, \delta \in (0,1)$ are pre-defined constants. A similar bound on the sample complexity is obtained for the R-max algorithm [48] which distinguishes the “known" and “unknown" states based on how often they have been visited. It explores by acting to maximize rewards under the assumption that unknown states deliver the maximum reward. UCRL2 algorithm [49] relies on optimistic bounds on the reward functions and probability density functions and enjoys near-optimal regret bounds. A review of model-based RL algorithms with provable finite time guarantees can be found in [50, Chapter 38]. A major drawback of these algorithms is that they consider optimism in the face of uncertainty and hence, are not robust to the estimation uncertainties. Furthermore, these algorithms with random exploration might lead to a bad user experience in applications in which the RL agent seeks to learn human preferences for system optimization.

To address these shortcomings, we propose a DSEE algorithm for model-based RL in which we design a deterministic sequence of exploration and exploitation epochs. The DSEE approach has been used in multi-arm bandit problems [51–54] and multi-robot coordination problems [55]. It allows for differentiation between exploration and exploitation epochs. The announced exploration may lead to a better user experience for the agents (especially for human agents) than random exploration at any time. For example, many personalized web services calibrate their recommendations intermittently by announced exploration, i.e., through surveys and user selection. Another advantage of the DSEE algorithm is that it allows for efficient exploration of the environment in multi-agent systems. Specifically, in multi-agent systems, exploration can be well-planned to cover all regions of the state-space through agent coordination which can be easily arranged due to the deterministic structure of exploration and exploitation.

1.1.4 Resource Sharing Games

Another key focus of this dissertation is to incentivize collaboration in a team of heterogeneous agents. We connect the class of problems involving human-team-supervised autonomy [3] with the common-pool resource (CPR) games [56, 57], and design utilities that yield the desired behavior. CPR games [56, 57] is a class of resource-sharing games in which players jointly manage a common pool of resource and make strategic decisions to maximize their utilities. Our CPR formulation has features similar to the CPR game studied in [57,58]. In these works, authors utilize prospect theory to capture the risk aversion behavior of the players investing in a fragile CPR [59] that fails if there is excessive investment in the CPR. In the case of CPR failure, no player receives any return from the CPR. While our design of the common review pool is similar to the fragile CPR, our failure model incorporates the constraint that only serviced tasks can be reviewed. In contrast to the agent heterogeneity due to prospect-theoretic risk preferences in [57], heterogeneity in our model arises due to differences in the agents' mean service and review times.

While we use human-team-supervised autonomy as a motivating example, our problem formulation can be applied to a broad range of problems involving tandem queues [60], where servicing and reviewing of tasks can be considered as the subsequent stages of the queueing network. Tandem queues are utilized to study problems such as resource allocation, inventory management, process optimization, and quality control [61]. Existing game theoretic approaches [62, 63] to service rate control in tandem queues assume that a single server is present at each stage of the tandem queue and each server has its independent resources. In contrast, in our setup, multiple heterogeneous agents allocate their time at different stations based on their skill sets and maximize the system throughput. Additionally, our mathematical techniques are applicable to many problems involving the dual-screening process. For example, in human-in-the-loop systems which are pervasive in areas such as search-and-rescue, semi-autonomous vehicle systems, surveillance, etc., humans often supervise (review) the actions (service) performed by the autonomous agents. In such settings, our framework incentivizes collaboration among heterogeneous agents.

Game-theoretic approaches have been utilized for problems in distributed control [64], wherein the overall system is driven to an efficient equilibrium strategy that is close to the social optimum through an appropriate design of utility functions [65]. Price of Anarchy (PoA) [66] is often used to characterize efficiency of the equilibrium strategies in a game. Associated analysis techniques utilize smoothness property of the utility functions [67], leverage submodularity of the welfare function [68], or solve an auxiliary optimization problem [69,70]. These approaches do not immediately apply to our setup. Instead, we follow a new line of analysis to obtain bounds on PoA by constructing a homogeneous CPR game, for which we show that the equilibrium strategy is also the social optimum (PoA=1), and relating its utility to the original game.

1.2 Organization and Contribution

In this section, we present the organization of the remainder of the thesis and the contributions of the work in each chapter.

Chapter 2: In this chapter, we study optimal fidelity selection for a human operator servicing a queue of homogeneous tasks. The agent can service a task with a normal or high fidelity level, where fidelity refers to the degree of exactness and precision while servicing the task. Therefore, high-fidelity servicing results in higher-quality service but leads to larger service times and increased operator tiredness. We treat the cognitive state of the human operator as a lumped parameter that captures psychological factors such as workload, and fatigue. The service time distribution of the human operator depends on her cognitive dynamics and the level of fidelity selected for servicing the task. Her cognitive dynamics evolve as a Markov chain in which the cognitive state increases with high probability whenever she is busy, and decreases while resting. The tasks arrive according to a Poisson process and each task waiting in the queue loses its value at a fixed rate. We address the trade-off between high-quality service of the task and consequent loss in value of subsequent tasks using an SMDP framework. We numerically determine an optimal policy and the corresponding optimal value function. Finally, we establish the structural properties of an optimal fidelity policy and provide conditions under which the optimal policy is a threshold-based policy. The material in this chapter is from [41] and [39].

The major contributions of this work are threefold. First, we pose the fidelity selection problem in an SMDP framework and compute an optimal policy. We formulate a control of queue problem, where in contrast to a standard queue, the server is a human operator with her own cognitive dynamics. Our model for service time distribution of the human operator incorporates the influence of cognitive state and fidelity level on the service time. Second, we numerically show the influence of cognitive dynamics on the optimal policy. In particular, we show that servicing the tasks with high fidelity is not always optimal due to larger service times and increased tiredness of the human operator. In fact, we determine the optimal policy as a function of the queue length as well as the cognitive state of the human operator. Our results provide insight into the efficient design of human decision support systems. Third, we establish structural properties of the optimal fidelity selection policy and provide sufficient conditions under which, for each cognitive state, there exist thresholds on queue lengths at which optimal policy switches fidelity levels.

Chapter 3: In this chapter, we study the problem of optimal fidelity selection for a human operator performing an underwater visual search task. Human performance depends on various cognitive factors such as workload and fatigue. We perform human experiments in which participants perform two tasks simultaneously: a primary task, which was subject to evaluation, and a secondary task to estimate their workload. The primary task requires participants to search for underwater mines in videos, while the secondary task involves a simple visual test where they respond when a green light displayed on the side of their screens turns red. Videos arrive as a Poisson process and are stacked in a queue to be serviced by the human operator. The operator can choose to watch the video with either normal or high fidelity, with normal fidelity videos playing at three times the speed of high fidelity ones. Participants receive rewards for their accuracy in mine detection for each primary task and penalties based on the number of videos waiting in the queue. We consider the workload of the operator as a hidden state and model the workload dynamics as an Input-Output Hidden Markov Model (IOHMM). We use a Partially Observable Markov Decision Process (POMDP) to learn an optimal fidelity selection policy, where the objective is to maximize total rewards. Our results demonstrate improved performance when videos were serviced based on the optimal fidelity selection policy compared to a baseline where humans chose the fidelity level themselves. The material in this chapter is from [71].

The major contributions of this work are threefold. First, we address the optimal fidelity selection challenge by framing it as a control of queue problem, incorporating a hidden workload state. We employ IOHMM and POMDP to derive the optimal fidelity selection policy. Second, we compare the human fidelity selection policy with the optimal policy and draw valuable insights into human behavioral patterns. Third, we illustrate that by recommending the optimal policy, autonomous systems can effectively aid human decision-making, leading to a substantial improvement in system performance.

Chapter 4: In this chapter, we relax the assumption of the known human service time model in the optimal fidelity selection problem. We assume the parameters of the human's service time distribution depend on the selected fidelity level and her cognitive state and are assumed to be unknown a priori. These parameters are learned online through Bayesian parameter estimation. We formulate a robust adaptive SMDP to solve our optimal fidelity selection problem. We extend the results on the convergence of robust-adaptive MDP to robust-adaptive SMDPs and show that the solution of the robust adaptive SMDP converges to the optimal solution for the uncertainty-free SMDP. Furthermore, we numerically illustrate the convergence of the synchronous and asynchronous robust adaptive policy to the uncertainty-free optimal policy. The material in this chapter is from [17].

The major contributions of this work are fourfold. First, we pose the optimal fidelity selection problem with uncertain human service time distribution in a robust adaptive SMDP framework. Second, we continuously improve the service time distribution estimates using

Bayesian parametric estimation [72] and utilize it to obtain a robust policy. Third, we formally show that the solution of both synchronous and asynchronous value iteration methods for the robust adaptive SMDP converges to the optimal solution for the uncertainty-free SMDP. Fourth, we provide numerical illustrations that show the convergence of the robust adaptive SMDP solution to the uncertainty-free SMDP.

Chapter 5: In this chapter, we propose a DSEE algorithm with interleaving exploration and exploitation epochs for model-based RL problems that aim to simultaneously learn the system model, i.e., an MDP, and the associated optimal policy. During exploration, DSEE explores the environment and updates the estimates for expected reward and transition probabilities. During exploitation, the latest estimates of the expected reward and transition probabilities are used to obtain a robust policy with high probability. We design the lengths of the exploration and exploitation epochs such that the cumulative regret grows as a sub-linear function of time. The material in this chapter is from [73].

The major contributions of this work are twofold: (i) we propose a DSEE algorithm for model-based RL problems and (ii) we design the lengths of the exploration and exploitation epochs such that the cumulative regret for the DSEE algorithm grows as a sub-linear function of time.

Chapter 6: In this chapter, we investigate the role of feedback in fostering human learning in sequential decision-making tasks. Cognitive rehabilitation, STEM skill acquisition, and coaching games such as chess often require tutoring decision-making strategies. The advancement of AI-driven tutoring systems for facilitating human learning requires an understanding of the impact of evaluative feedback on human decision-making and skill development. To this end, we conduct human experiments using Amazon Mechanical Turk to study the influence of evaluative feedback on human decision-making in sequential tasks. In these experiments, participants solve the Tower of Hanoi puzzle and receive AI-generated feedback while solving it. We examine how this feedback affects their learning and skill transfer to related tasks. We also explore various computational models to understand how people incorporate evaluative feedback into their decision-making processes. The material in this chapter is from [74].

There are three major contributions of this work. (i) We investigate the impact of different evaluative feedback strategies on the performance of individuals learning to solve ToH, a widely studied sequential decision-making task. Furthermore, we explore how individuals trained with different feedback strategies transfer their skills to a more challenging task. (ii) Treating humans as noisy optimal agents, we study how various evaluative feedback strategies affect their reward functions. Our research highlights the influence of different forms of evaluative feedback on the implicit reward structure that explains human decisions. (iii) We create a set of candidate computational models that may explain how humans integrate evaluative feedback into their sequential decision-making processes. Our goal is to identify the model that best explains human decision-making under evaluative feedback conditions.

Chapter 7: In this chapter, we consider a team of heterogeneous agents that is collectively responsible for servicing, and subsequently reviewing, a stream of homogeneous tasks. Each agent has an associated mean service time and a mean review time for servicing and reviewing the tasks, respectively. Agents receive a reward based on their service and review admission rates. The team objective is to collaboratively maximize the number of “serviced and reviewed" tasks. We formulate a Common-Pool Resource (CPR) game and design utility functions to incentivize collaboration among heterogeneous agents in a decentralized manner. We show the existence of a unique Pure Nash Equilibrium (PNE), and establish convergence of best response dynamics to this unique PNE. Finally, we establish an analytic upper bound on three inefficiency measures of the PNE, namely the price of anarchy (PoA), the ratio of the total review admission rate (TRI), and the ratio of latency (LI). The material in this chapter is from [75] and [76].

The major contributions of this work are fivefold. First, we present a novel formulation of team backup behavior and design incentives, within the CPR game formalism, to facilitate such behavior. Second, we show that there exists a unique PNE for the proposed game. Third, we show that the proposed game is a best response potential game as defined in [77], for which both sequential best response dynamics [78] and simultaneous best reply dynamics [79] converge to the PNE. Thus, the best response of self-interested agents in a decentralized team converge to the PNE. Fourth, we provide the structure of the social welfare solution and numerically quantify different measures of the inefficiency for the PNE, namely the PoA, the ratio of the total review admission rate (TRI), and the ratio of latency (LI), as a function of a measure of heterogeneity. While PoA is a widely used inefficiency metric, we define TRI and LI as other relevant measures for our setup based on the total review admission rate and latency (inverse of throughput), respectively. Finally, we provide an analytic upper bound for all three measures of the inefficiency.

CHAPTER 2

OPTIMAL FIDELITY SELECTION FOR HUMAN-IN-THE-LOOP QUEUES

In this chapter, we study optimal fidelity selection for a human operator servicing a queue of homogeneous tasks. The agent can service a task with a normal or high fidelity level, where fidelity refers to the degree of exactness and precision while servicing the task. Therefore, high-fidelity servicing results in higher-quality service but leads to larger service times and increased operator tiredness. We treat the cognitive state of the human operator as a lumped parameter that captures psychological factors such as workload, and fatigue. The service time distribution of the human operator depends on her cognitive dynamics and the level of fidelity selected for servicing the task. Her cognitive dynamics evolve as a Markov chain in which the cognitive state increases with high probability whenever she is busy, and decreases while resting. The tasks arrive according to a Poisson process and each task waiting in the queue loses its value at a fixed rate. We address the trade-off between high-quality service of the task and consequent loss in value of subsequent tasks using an SMDP framework. We numerically determine an optimal policy and the corresponding optimal value function. Finally, we establish the structural properties of an optimal fidelity policy and provide conditions under which the optimal policy is a threshold-based policy.

2.1 Background and Problem Formulation

We now discuss our problem setup, formulate it as an SMDP, and solve it to obtain an optimal fidelity selection policy.

2.1.1 Problem Setup

We consider a human supervisory control system in which a human operator is servicing a stream of homogeneous tasks. The human operator may service these tasks with different levels of fidelity. The servicing time of the operator depends on the fidelity level with which she services the task as well as her cognitive state. We assume that the mean service time of the operator increases with the selected fidelity level. For example, when the operator

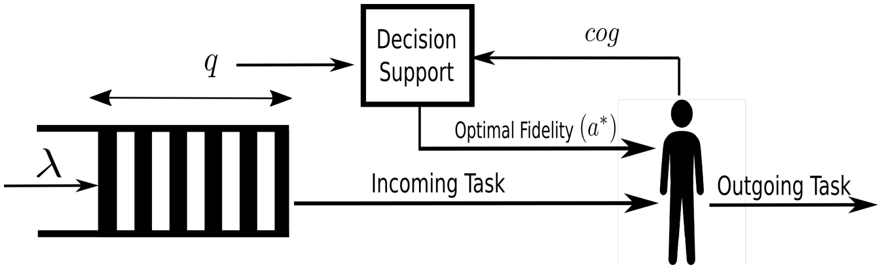


Figure 2.1 Overall schematic of the problem setup. The incoming tasks arrive as a Poisson process with rate λ. The tasks are serviced by the human operator based on the recommended fidelity level by the decision support system. Each task loses its value at a fixed rate while waiting in the queue.

services the task with high fidelity, she may look into deeper details of the task, and consequently take a longer time to service.

In addition to the fidelity level, the human service time may depend on their cognitive state. We treat the cognitive state as a lumped parameter that can capture various physiological measures. It can be a function of stress, workload, arousal rate, operator utilization ratio, etc. Such lumped representation can be obtained by classifying these psychological measurements into different service time distribution parameters. Inspired by the Yerkes-Dodson law, for a fixed level of fidelity, we model the service time as a unimodal function of the human cognitive state. Specifically, the mean service time is minimal corresponding to an intermediate optimal cognitive state (later referred to as the optimal cognitive state cog*) as shown in Fig. 2.2c.

We are interested in the optimal fidelity selection policy for the human operator. To this end, we formulate a control of queue problem, where in contrast to a standard queue, the server is a human operator with her cognitive dynamics. The incoming tasks arrive according to a Poisson process at a given rate $\lambda \in \mathbb{R}_{>0}$ and are serviced by the operator based on the fidelity level recommended by a decision support system (Fig. 2.1). We consider a dynamic queue of homogeneous tasks with a maximum capacity $L \in \mathbb{N}$. The operator is penalized for each task waiting in the queue at a constant rate $c \in \mathbb{R}_{>0}$ per unit delay in its servicing. The set of possible actions available for the operator corresponds to (i) ***Waiting (W)***, when the queue is empty, (ii) ***Resting (R)***, which allows the operator to rest and reach the optimal cognitive state, (iii) ***Skipping (S)***, which allows the operator to skip a task to reduce the queue length and thereby focus on newer tasks, (iv) ***Normal Fidelity (N)*** for servicing the task with normal fidelity, and (v) ***High Fidelity (H)*** for servicing the task more carefully with high precision. The skipping action ensures the stability of the queue by allowing the operator to reduce the queue length by skipping some tasks. Ideally, through appropriate control of the arrival rate, the system designer should ensure that skipping is not an optimal action.

Let $s \in \mathcal{S}$ be the state of the system and $\mathcal{A}_s$ be the set of admissible actions in state s, which we define formally in Section 2.1.2. The human receives a reward $r : \mathcal{S} \times \mathcal{A}_s \mapsto \mathbb{R}_{\geq 0}$ defined by

$$r(s,a) = \begin{cases} r_H, & \text{if } a = H, \\ r_N, & \text{if } a = N, \\ 0, & \text{if } a \in \{W,\ R,\ S\}, \end{cases} \tag{2.1}$$

where, $r_H, r_N \in \mathbb{R}_{\geq 0}$ and $r_H > r_N$. We intend to design a decision support system that assists the operator by recommending optimal fidelity level to service each task[1]. The recommendation is based on the queue length and the operator's cognitive state which we assume to have real-time access using, e.g., Electroencephalogram (EEG) measurements (see [80] for measures of cognitive load from EEG data) or eye-tracking and pupillometry [81]. We assume that the noisy data from these devices can be clustered into a finite number of bins to estimate the cognitive state. We study the optimal policy under the perfect knowledge of the cognitive state[2].

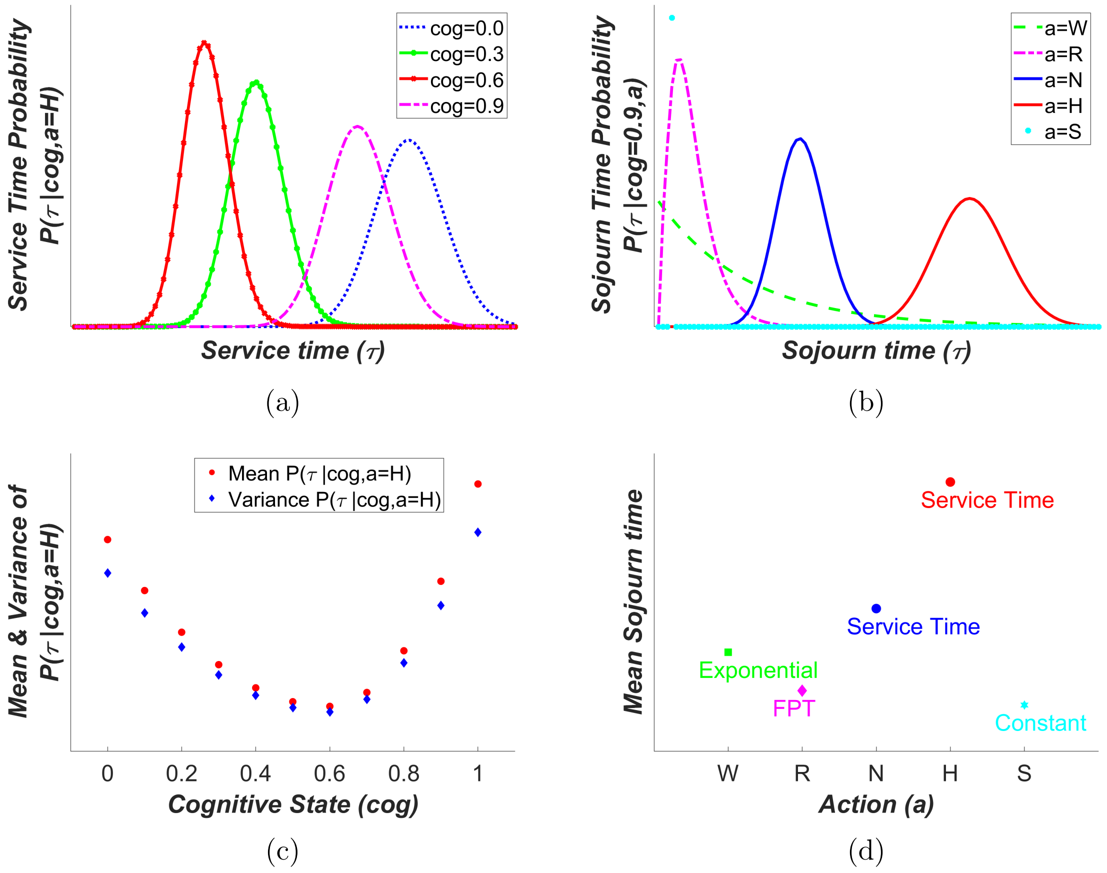


Figure 2.2 Service time distribution of the human operator with (a) varying cognitive state and high fidelity, (b) varying action and fixed cognitive state, cog = 0.9. (c) Mean and variance of the service time distribution are unimodal functions of the cognitive state. (d) The mean sojourn time distribution takes on different forms based on the selected action.

	Forward	***Backward***	***Stay Probability***[c]
Action	***Probability***[a] ($\lambda_f \delta t$)	***Probability***[b] ($\lambda_b \delta t$)	***(1-***$\lambda_f \delta t - \lambda_b \delta t$***)***
W	$\lambda_f = 0.02$ (Noise)	$\lambda_b = 0.5$	$1 - 0.52\delta t$
R	$\lambda_f = 0.02$ (Noise)	$\lambda_b = 0.5$	$1 - 0.52\delta t$
N	$\lambda_f = 0.6$	$\lambda_b = 0.02$ (Noise)	$1 - 0.62\delta t$
H	$\lambda_f = 1.1$	$\lambda_b = 0.02$ (Noise)	$1 - 1.12\delta t$
S	$\lambda_f = 0$	$\lambda_b = 0$	1

[a] Forward Probability does not exist for cog = 1 (reflective boundary)

[b] Backward Probability does not exist for cog = 0 (reflective boundary)

[c] Stay Probability is $1 - \lambda_f \delta t$ for cog = 0 and $1 - \lambda_b \delta t$ for cog = 1

Table 2.1 Cognitive Dynamics modeled as Markov chain.

2.1.2 Mathematical Modeling

We formulate the control of queue problem as a discrete-time SMDP Γ defined by the following six components:

(i) A finite state space $\mathcal{S} := \{(q, \text{cog})|\ q \in \{0, 1, ..., L\},\ \ \text{cog} \in \mathcal{C} := \{i/N\}_{i\in\{0,\cdots,N\}}\}$, for some $N \in \mathbb{N}$, where q is the queue length and cog represents the lumped cognitive state, which increases (decreases) when the operator is busy (idle).

(ii) A set of admissible actions $\mathcal{A}_s$ for each state $s \in \mathcal{S}$ which is given by: (i) $\mathcal{A}_s := \{\, W \mid s \in \mathcal{S},\ q = 0\}$ when queue is empty, (ii) $\mathcal{A}_s := \{\{R,\ S,\ N,\ H\ \}|\ s \in \mathcal{S},\ q \neq 0\}$ when queue is non-empty and cog $>$ cog*, where cog$^* \in \mathcal{C}$ is the optimal cognitive state associated with minimum mean service time, and (iii) $\mathcal{A}_s := \{\{S,\ N,\ H\ \}|\ s \in \mathcal{S},\ q \neq 0\}$ when queue is non-empty and cog $\leq$ cog*.

(iii) A state transition distribution $\mathbb{P}\,(s'|\ \tau, s, a)$ from state s to s' for each action $a \in A_s$ conditioned on the discrete sojourn time $\tau \in \mathbb{R}_{>0}$ (time spent in state s before transitioning into next state s'). The state transition from $s = (q,\ \text{cog}) \to s' = (q',\ \text{cog}')$ consists of two independent transition processes which are given by (i) a Poisson process for transition from $q \to q'$ (ii) human cognitive dynamics for the transition from cog $\to$ cog$'$. We model the cognitive dynamics of the human operator as a Markov chain in which, while servicing the task, the probability of an increase in cognitive state in small time $\delta t \in \mathbb{R}_{>0}$ is greater than the probability of a decrease in cognitive state. Furthermore, the probability of the increase in the cognitive state increases with the level of fidelity selected for servicing the task. Similarly, while waiting or resting, the probability of a decrease in cognitive state in small time δt

[1] We assume compliance of the operator with the recommendations. To account for non-compliance, we can introduce p as the probability of compliance and $1-p$ as the probability that the operator will deviate and follow a different behavioral policy. This deviation can be incorporated by using a mixed service time distribution with probabilities p and $1-p$ for the recommended and behavioral actions respectively.

[2] If the cognitive state is not perfectly known, then our policy can be used within algorithms such as Q_{MDP} [82], to derive approximate solutions to the associated partially observable Markov decision process [83].

is higher than the probability of an increase in cognitive state. Sample parameters of the model used in our numerical simulations are shown in Table 2.1. This model of cognitive state dynamics is a stochastic equivalent of deterministic models of the utilization ratio considered in [11]. It is assumed that the cognitive state remains unchanged when the human operator chooses to skip the task.

(iv) Sojourn time distribution $\mathbb{P}(\tau|\ s, a)$ of (discrete) time $\tau \in \mathbb{R}_{>0}$ spent in state s until the next action is chosen takes on different forms depending on the selected action (Fig. 2.2d). The sojourn time is the service time while servicing the task (normal/ high fidelity), resting time while resting, constant time of skipping $t_s \in \mathbb{R}_{>0}$ while skipping, and time until the next task arrival while waiting in case of an empty queue. We model the rest time as the time required to reach from the current cognitive state to the optimal cognitive state cog*. In our numerical illustrations, we model the service time distribution while servicing the task using a hypergeometric distribution (Fig. 2.2a and 2.2b), where the parameters of the distribution are chosen such that the mean service time has the desired characteristics, i.e., it increases with the fidelity level (Fig. 2.2d) and is a unimodal function of the cognitive state (Fig. 2.2c). While resting, sojourn time distribution is the first passage time (FPT) distribution for transitioning from the current cognitive state cog to cog*. We determine this distribution using matrix methods [84] applied to the Markov chain used to model the cognitive dynamics. Finally, to ensure the stability of the queue, we assume that the constant time of skip is less than $\frac{1}{\lambda}$, i.e., queue length decreases on average while skipping tasks.

(v) For selecting action a at state s, the human receives a bounded reward $r(s, a)$ defined in (2.1). Additionally, the human incurs a penalty at a constant cost rate of c due to each task waiting in the queue, and consequently, the cumulative expected cost for choosing action a at state $s = (q, \text{cog})$ is given by:

$$\sum_{\tau} \mathbb{P}(\tau|s, a) c\tau \left(\mathbb{E}\left[\frac{q + q'}{2} \middle| \tau, s, a \right] \right) = \sum_{\tau} \mathbb{P}(\tau|s, a) c\tau \left(\frac{2q + \lambda\tau}{2} \right),$$

which is obtained by using $\mathbb{E}[q|\tau, s, a] = q$ and $\mathbb{E}[q'|\tau, s, a] = q + \lambda\tau$. The expected net immediate reward received by the operator for selecting an action a in state s is given by:

$$\begin{aligned} R(s,a) &= r(s,a) - \sum_{\tau} \mathbb{P}(\tau|s,a) c \left(\frac{2q+\lambda\tau}{2} \right) \tau \\ &= r(s,a) - c\,\mathbb{E}\left[\tau|s,a\right] q - \frac{c\lambda}{2}\,\mathbb{E}\left[\tau^2|s,a\right], \end{aligned} \tag{2.2}$$

where $\mathbb{E}\left[\tau|\ s,a\right]$ and $\mathbb{E}\left[\tau^2|s,a\right]$ represent the first and the second conditional moment of the sojourn time distribution, respectively.

(vi) A discount factor $\gamma \in [0, 1)$, which we choose as 0.96 for our numerical illustration.

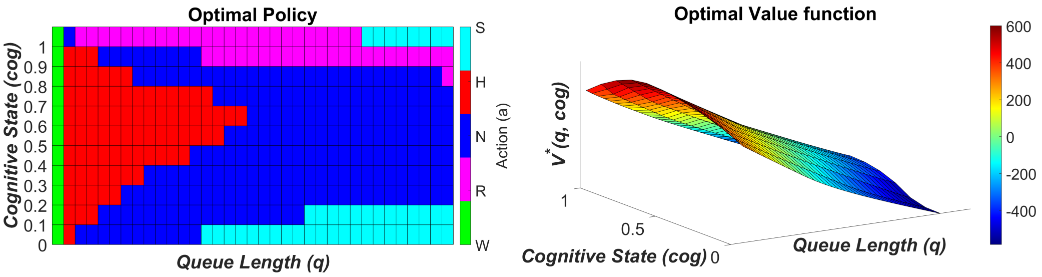
Remark 1. *Although we assume a finite skip time, an alternative approach is to incorporate a penalty for the skip action. Note that, unlike a fixed penalty, a finite skip time results in a penalty that increases with queue length (see* (2.2)*). Consequently, the current approach is less inclined to skip tasks as the queue length increases compared to a model with a constant penalty.*

Remark 2. *The reward $R(s, a)$ formulation can be interpreted as an unconstrained SMDP corresponding to a constrained SMDP that maximizes $r(s, a)$ subject to a constraint on the average queue length for the stability of the queue. Therefore, the penalty rate c acts as the Lagrange multiplier for the unconstrained problem, and hence, can be obtained by primal-dual methods that use dual ascent for finding the Lagrange multiplier [20].*

2.1.3 Solving SMDP for Optimal Policy

For SMDP Γ, the optimal value function $V^* : \mathcal{S} \to \mathbb{R}$ satisfies the following Bellman equation [85]:

$$V^*(s) = \max_{a \in A_s} \left[R(s,a) \ + \sum_{s',\tau} \gamma^{\tau} \mathbb{P}\left(s', \tau | s, a\right) V^*\left(s'\right) \right], \tag{2.3}$$



(a)

(b)

Figure 2.3 (a) Optimal Policy π^* and (b) Optimal Value Function V^* for SMDP Γ where time required to skip the tasks is not too small compared to the mean service time required to process the task.

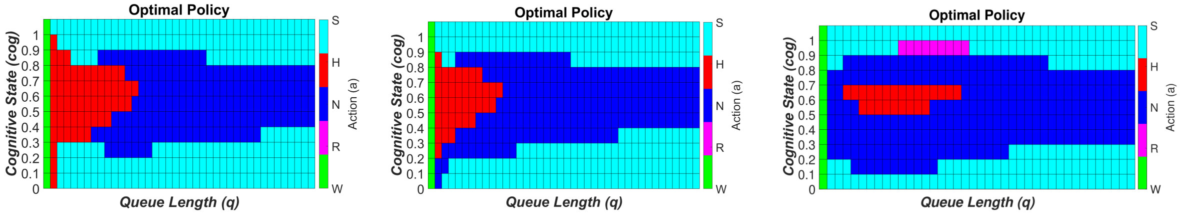
where $\mathbb{P}(s', \tau | s, a)$, which is the joint probability that a transition from state s to state s' occurs after time τ when action a is selected can be rewritten as:

$$\mathbb{P}(s', \tau | s, a) = \mathbb{P}(s' | \tau, s, a) \, \mathbb{P}(\tau | s, a), \tag{2.4}$$

where $\mathbb{P}(s' | \tau, s, a)$ and $\mathbb{P}(\tau | s, a)$ are given by the state transition probability distribution and the sojourn time probability distribution, respectively. An optimal policy $\pi^* : \mathcal{S} \to \mathcal{A}_s$ at each state s selects an action that achieves the maximum in (2.3). We utilize the value iteration algorithm [31] to compute an optimal policy.

2.2 Numerical Illustration of Optimal Fidelity Selection

We now numerically illustrate the optimal value function and an optimal policy for SMDP Γ.



(a)

(b)

(c)

Figure 2.4 Optimal policy π^* for different values of arrival rate λ. In case (a) and (b) the action S in the optimal policy does not have a unique threshold for some cognitive states. Similarly in case (c) actions S and N in the optimal policy do not have unique thresholds.

Fig. 2.3a and 2.3b show an optimal policy π^*, and the optimal value function V^*, respectively, for the case in which the skip time is not too small compared to the mean service time. If the skip time is too small, the action S is the optimal action almost everywhere to reduce the queue length. For a sufficiently high arrival rate λ such that there is always a task in the queue after servicing the current task, we observe that for any given cog, V^* is monotonically decreasing with q.

Additionally, we observe that for a given q, V^* is a unimodal function of cog, with its maximum value corresponding to the optimal cognitive state ($\text{cog}^* = 0.6$ for numerical illustrations). We observe that π^* selects the high fidelity level around the cog* for low queue length, and thereafter transitions to a normal fidelity level for higher queue lengths. We also observe that in low cognitive states, the optimal policy is to keep skipping the tasks until the queue length becomes small, and then start servicing the tasks. In higher cognitive states, we observe that resting is the optimal action at smaller queue lengths while skipping tasks is the optimal choice at larger queue lengths. Additionally, we observe the effect of cog on π^*. In particular, we observe that π^* switches from H to N, N to R, and R to S at certain thresholds on q, and these thresholds appear to be a unimodal function of cog. This behavior can be attributed to the mean service time being unimodal w.r.t cog.

Fig. 2.4 shows some examples of π^* for certain parameters. We observe that for some cognitive states, π^* does not have a unique threshold and the same action reappears after switching to another action. For example, in Fig. 2.4a, action S is observed between actions H and N, as well as after action N. In the following section, we provide sufficient conditions under which π^* has unique transition thresholds at which actions switch, and the previous action does not re-appear for the same cog.

2.3 Structural Properties of the Optimal Policy

We establish the structural properties of the optimal infinite-horizon value function by considering the finite horizon case and then extending the results to the infinite horizon by taking the infinite step limit.

Let $V_n^*(s_0), n \geq 0$, be the discounted n-step optimal expected reward when the initial state is s_0, where $V_0^*(q, \text{cog}) = -Cq$ is the terminal cost for the finite-horizon case for a non-negative constant C. Each step size $k \in \{0, \dots, n-1\}$ is based on the sojourn time τ_k, spent in a state s_k when action a_k is selected. Let $J_{n,\pi}(s_0)$ denote the discounted n-step expected reward with initial state s_0 under a given policy π. Henceforth, for the brevity of notation, we denote the conditional expectation $\mathbb{E}[\cdot|s_0, \pi]$ by $\mathbb{E}_\pi[\cdot]$. $J_{n,\pi}(s_0)$ is given by:

$$J_{n,\pi}(s_0) = \mathbb{E}_\pi \left[\sum_{i=0}^{n-1} \gamma^{\zeta_i} R(s_i, a_i) - \gamma^{\zeta_n} C q_n \right], \tag{2.5}$$

where $\zeta_i := \sum_{j=0}^{i-1} \tau_j$ for $i > 0$ and $\zeta_0 := 0$. The discounted n-step optimal expected reward $V_n^*(s)$ is given by:

$$V_n^*(s_0) = J_{n,\pi^*}(s_0), \tag{2.6}$$

where π^* is the optimal policy that maximizes $J_{n,\pi}(s_0)$ at each s_0.

Let $\mu^1 : \mathcal{S} \times \mathcal{A}_s \mapsto \mathbb{R}_{>0}$ and $\mu^2 : \mathcal{S} \times \mathcal{A}_s \mapsto \mathbb{R}_{>0}$ be function defined by $\mu^1(s, a) = \mathbb{E}\left[\tau|s, a\right]$ and $\mu^2(s, a) = \mathbb{E}\left[\tau^2|s, a\right]$, where τ is the sojourn time. We study the structural properties of the optimal policy for a large queue capacity, i.e. in the limit $L \to +\infty$, and under the following assumptions:

(A1) The task arrival rate λ is sufficiently high so that the queue is never empty with high probability[3].

(A2) For any state $s = (q, \mathrm{cog})$[4]:

$$\begin{aligned} &\mu^1(s, S) < \mu^1(s, R) < \mu^1(s, N) < \mu^1(s, H), \text{ and} \\ &\mu^2(s, S) < \mu^2(s, R) < \mu^2(s, N) < \mu^2(s, H). \end{aligned} \tag{2.7}$$

(A3) We assume that $\mathbb{E}_\pi[\gamma^\tau] \leq f(\mathbb{E}_\pi[\tau], \mathrm{Var}_\pi(\tau)) < 1$, where $\mathrm{Var}_\pi(\tau) = \mathrm{Var}(\tau|s_0, a = \pi(s_0))$ is the variance of τ in any initial state s_0 under a given policy π, and f is a monotonic function such that $f(\cdot, \mathrm{Var}_\pi(\tau))$ is monotonically decreasing and $f(\mathbb{E}_\pi[\tau], \cdot)$ is monotonically increasing.

[3] Given the service time distributions and the Poisson arrival rate, we can precisely determine the distribution of the number of arrivals that occur between each state transition. Therefore, Chernoff bounds [86] can be utilized to characterize the high probability that the number of arrivals between state transitions while servicing a task exceeds one.

[4] The action R is only available for states with $\mathrm{cog} > \mathrm{cog}^*$.

We make the assumption (A1) for convenience. Indeed, if the queue is allowed to be empty, then we will need to deal with an extra “waiting" action. Also, high arrival rates are the most interesting setting to study optimal fidelity selection. Assumption (A2) is true for a broad range of interesting parameters that define sojourn time distribution(s). Assumption (A3) holds for a class of light-tail distributions with non-negative support for τ, for example, when the moment generating function (MGF) of τ is upper bounded by the MGF of Gamma distribution, i.e.,

$$\mathbb{E}_\pi[e^{t\tau}] \leq \left(1 - \frac{\mathrm{Var}_\pi(\tau)t}{\mathbb{E}_\pi[\tau]}\right)^{\frac{-\mathbb{E}_\pi[\tau]^2}{\mathrm{Var}_\pi(\tau)}}, \text{for all } t < \frac{\mathbb{E}_\pi[\tau]}{\mathrm{Var}_\pi(\tau)}.$$

In this scenario, substituting $t = \ln(\gamma) < 0 < \frac{\mathbb{E}_\pi[\tau]}{\mathrm{Var}_\pi(\tau)}$, we get

$$\mathbb{E}_\pi[\gamma^\tau] \leq \left(1 - \frac{\mathrm{Var}_\pi(\tau)\ln(\gamma)}{\mathbb{E}_\pi[\tau]}\right)^{\frac{-\mathbb{E}_\pi[\tau]^2}{\mathrm{Var}_\pi(\tau)}}$$
$$=: f(\mathbb{E}_\pi[\tau], \mathrm{Var}_\pi(\tau)).$$

Let $\rho := \max_{\mathrm{cog},a} f(\mathbb{E}[\tau|\mathrm{cog}, a], \mathrm{Var}(\tau|\mathrm{cog}, a))$. Therefore, $\mathbb{E}_\pi[\gamma^\tau] \leq \rho$. For the class of distributions of τ satisfying assumption (A3), and any initial state s_0 and policy π, we have

$$\mathbb{E}_\pi[\gamma^{\zeta_k}] \overset{(1)^*}{=} \prod_{i=0}^{k-1} \mathbb{E}_\pi[\gamma^{\tau_i}] \leq \prod_{i=0}^{k-1} f(\mathbb{E}_\pi[\tau_i], \mathrm{Var}_\pi(\tau_i)) \leq \rho^k,$$

where $(1)^*$ follows from the independence of τ_i and τ_j, for $i \neq j$. Therefore, we have

$$\lim_{n\to\infty} \sum_{k=0}^{n-1} \mathbb{E}_\pi[\gamma^{\zeta_k}] \leq \lim_{n\to\infty} \sum_{k=0}^{n-1} \rho^k = \frac{1}{1-\rho}.$$

We will now establish that the optimal policy for SMDP Γ is a threshold-based policy if the following condition holds for each cognitive state cog:

$$\min\{\mathbb{E}[\tau|\mathrm{cog}, H] - \mathbb{E}[\tau|\mathrm{cog}, N],\ \mathbb{E}[\tau|\mathrm{cog}, N] - \mathbb{E}[\tau|\mathrm{cog}, R],\ \mathbb{E}[\tau|\mathrm{cog}, R] - t_s\} + \frac{t_s \gamma^{\mathbb{E}[\tau|\mathrm{cog},H]}}{1-\gamma^{t_{\max}}} \geq \frac{t_{\max}}{1-\rho} \max_{a\in\mathcal{A}_s} \mathbb{E}[\gamma^\tau|\mathrm{cog}, a], \quad (2.8)$$

where $t_{\max} = \mathbb{E}[\tau|\mathrm{cog} = 1, a = H]$ is the maximum expected sojourn time (assuming the largest mean service time in the highest cognitive state), and t_s is the constant time for the skip.

Remark 3. *For tasks with large differences in expected sojourn times, i.e.,* $0 \ll t_s \ll \mathbb{E}[\tau|\text{cog}, R] \ll \mathbb{E}[\tau|\text{cog}, N] \ll \mathbb{E}[\tau|\text{cog}, H]$, $\max_{a\in\mathcal{A}_s} \mathbb{E}[\gamma^\tau|\text{cog}, a] \to 0$, *and* (2.8) *always holds.*

We introduce the following notation. Let $q_j^* : \mathcal{C} \mapsto \mathbb{Z}_{\geq 0} \cup \{+\infty\}$, for $j \in \{1, 2, 3\}$ be some functions of cog.

Theorem 1 (***Structure of optimal policy***)**.** *For SMDP* Γ *under assumptions (A1-A3) and an associated optimal policy* π^**, if the difference in the expected sojourn times is sufficiently large such that* (2.8) *holds for any cognitive state* cog*, then the following statements hold:*

(i) there exists unique threshold functions $q_1^*(\text{cog})$*,* $q_2^*(\text{cog})$*, and* $q_3^*(\text{cog})$ *such that for each* $\text{cog} > \text{cog}^*$*:*

$$\pi^*(s = (q, \text{cog})) = \begin{cases} H, & q \leq q_1^*(\text{cog}), \\ N, & q_1^*(\text{cog}) < q \leq q_2^*(\text{cog}), \\ R, & q_2^*(\text{cog}) < q \leq q_3^*(\text{cog}), \\ S, & q > q_3^*(\text{cog}); \end{cases}$$

(ii) there exists unique threshold functions $q_1^*(\text{cog})$ *and* $q_2^*(\text{cog})$ *such that for any* $\text{cog} \leq \text{cog}^*$*:*

$$\pi^*(s = (q, \text{cog})) = \begin{cases} H, & q \leq q_1^*(\text{cog}), \\ N, & q_1^*(\text{cog}) < q \leq q_2^*(\text{cog}), \\ S, & q > q_2^*(\text{cog}). \end{cases}$$

We prove Theorem 1 using the following lemmas.

Lemma 1. ***(Immediate Reward)****: For SMDP* Γ*, the immediate expected reward* $R(s, a)$*, for each* $a \in A_s$

(i) is linearly decreasing with queue length q *for any fixed cognitive state* cog*;*

(ii) is a unimodal function[5] *of the cognitive state* cog *for any fixed queue length q with its maximum value achieved at the optimal cognitive state* cog^*.

Proof. The proof follows by noting that (2.2) is linearly decreasing in q and the coefficients $\mathbb{E}\left[\tau|\ s,a\right]$ and $\mathbb{E}\left[\tau^2|s,a\right]$ are unimodal w.r.t cog. Interested readers can refer to [87] for detailed proof. □

We now provide important mathematical results in Lemma 2 which we use to establish Lemma 3.

Lemma 2. *For the SMDP Γ, the following equations hold for any initial state s_0 and policy π:*

(i) $\mathbb{E}_\pi\left[\gamma^{\zeta_k}\,\mathbb{E}[\tau_k^2|\text{cog}_k,a_k]\right] = \mathbb{E}_\pi\left[\gamma^{\zeta_k}\tau_k^2\right]$;

(ii) $\mathbb{E}_\pi\left[\gamma^{\zeta_k}\,\mathbb{E}[\tau_k|\text{cog}_k,a_k]q_k\right] = \mathbb{E}_\pi\left[\gamma^{\zeta_k}\tau_k\,\mathbb{E}_\pi\left[q_k|s_0,\zeta_k\right]\right]$

Proof. The proof utilizes the properties of the expectation operator, and independence of the transition processes for q_k and cog_k. Interested readers can refer to [87] for detailed proof. □

Lemma 3. *(**Value function bounds**): For SMDP Γ under assumptions (A1-A3), for any $\tilde{q}_0 \geq q_0$, $0 \leq \frac{ct_s\Delta q}{1-\gamma^{t_{\max}}} \leq V^*(q_0,\text{cog}_0) - V^*(\tilde{q}_0,\text{cog}_0) \leq \frac{ct_{\max}\Delta q}{1-\rho}$, where $\Delta q = \tilde{q}_0 - q_0$, ρ is an upper bound on $\mathbb{E}_\pi[\gamma^\tau]$, $t_{\max} = \mathbb{E}[\tau|\text{cog}=1, a=H]$ is the maximum expected sojourn time, and t_s is the constant time for skip.*

Proof. See Appendix A: Chapter 2 for the proof. □

Remark 4. *It follows from Lemma 3, that for SMDP Γ under assumptions (A1-A3), the optimal value function $V^*(q,\cdot)$ is monotonically decreasing with queue length q.*

[5] The expected immediate reward under action S is a constant, which we treat as a unimodal function.

Lemma 4. *(**Thresholds for low cognitive states**): For the SMDP* Γ *under assumptions (A1-A3), and an associated optimal policy* π^**, the following statements hold for each* $\text{cog} \leq \text{cog}^*$*:*

(i) there exists a threshold function $q_1^*(\text{cog})$*, such that N strictly dominates H, for each* $q > q_1^*(\text{cog})$ *if*

$$\mathbb{E}[\tau|\text{cog}, H] - \mathbb{E}[\tau|\text{cog}, N] + \frac{t_s \gamma^{\mathbb{E}[\tau|\text{cog},H]}}{1-\gamma^{t_{\max}}} \geq \frac{t_{\max}}{1-\rho}\,\mathbb{E}[\gamma^{\tau}|\text{cog}, N];$$

(ii) there exists a threshold function $q_2^*(\text{cog})$*, such that for each* $q > q_2^*(\text{cog})$*, action S is optimal if*

$$\mathbb{E}[\tau|\text{cog}, N] - t_s + \frac{t_s \gamma^{\mathbb{E}[\tau|\text{cog},H]}}{1-\gamma^{t_{\max}}} \geq \gamma^{t_s}\frac{t_{\max}}{1-\rho}.$$

Proof. See Appendix A: Chapter 2 for the proof. □

Lemma 5. *(**Thresholds for high cognitive states**): For the SMDP* Γ *under assumptions (A1-A3), and an associated optimal policy* π^**, the following statements hold for each* $\text{cog} > \text{cog}^*$*:*

(i) there exists a threshold function $q_1^*(\text{cog})$*, such that N strictly dominates H, for each* $q > q_1^*(\text{cog})$ *if*

$$\mathbb{E}[\tau|\text{cog}, H] - \mathbb{E}[\tau|\text{cog}, N] + \frac{t_s \gamma^{\mathbb{E}[\tau|\text{cog},H]}}{1-\gamma^{t_{\max}}} \geq \frac{t_{\max}}{1-\rho}\,\mathbb{E}[\gamma^{\tau}|\text{cog}, N];$$

(ii) there exists a threshold function $q_2^*(\text{cog})$*, such that R strictly dominates H & N, for each* $q > q_2^*(\text{cog})$ *if*

$$\mathbb{E}[\tau|\text{cog}, N] - \mathbb{E}[\tau|\text{cog}, R] + \frac{t_s \gamma^{\mathbb{E}[\tau|\text{cog},H]}}{1-\gamma^{t_{\max}}} \geq \frac{t_{\max}}{1-\rho}\mathbb{E}[\gamma^{\tau}|\text{cog}, R].$$

(iii) there exists a threshold function $q_3^*(\text{cog})$*, such that for each* $q > q_3^*(\text{cog})$*, action S is optimal if*

$$\mathbb{E}[\tau|\text{cog}, R] - t_s + \frac{t_s \gamma^{\mathbb{E}[\tau|\text{cog},H]}}{1-\gamma^{t_{\max}}} \geq \gamma^{t_s}\frac{t_{\max}}{1-\rho}.$$

Proof. Recall that $\mathcal{A}_s := \{\{R,\ S,\ N,\ H\ \}|\ s \in \mathcal{S},\ q \neq 0\}$ when queue is non-empty and $\text{cog} > \text{cog}^*$. The proof of Lemma 5 follows analogously to the proof of Lemma 4. □

Proof of Theorem 1: The proof follows by finding the intersection of the sufficient conditions from Lemmas 4 and 5 to obtain condition (2.8) for a threshold-based π^*.

2.4 Conclusions and Future Directions

We studied optimal fidelity selection for a human operator servicing a stream of homogeneous tasks using an SMDP framework. In particular, we studied the influence of human cognitive dynamics on an optimal fidelity selection policy. We presented numerical illustrations of the optimal policy and established its structural properties. These structural properties can be leveraged to tune the design parameters, deal with the model uncertainty, or determine a minimally parameterized policy for specific individuals and tasks.

There are several possible avenues for future research. An interesting direction is to conduct experiments with human subjects, measure EEG signals to assess their cognitive state and test the benefits of recommending optimal fidelity levels. It is of interest to extend this work to a team of human operators servicing a stream of heterogeneous tasks. A preliminary setup is considered in [75, 88], where authors study a game-theoretic approach to incentivize collaboration in a team of heterogeneous agents. In such a setting, finding the optimal routing and scheduling strategies for these heterogeneous tasks is also of interest.

CHAPTER 3

OPTIMAL FIDELITY SELECTION FOR IMPROVED PERFORMANCE IN HUMAN-IN-THE-LOOP QUEUES FOR UNDERWATER SEARCH

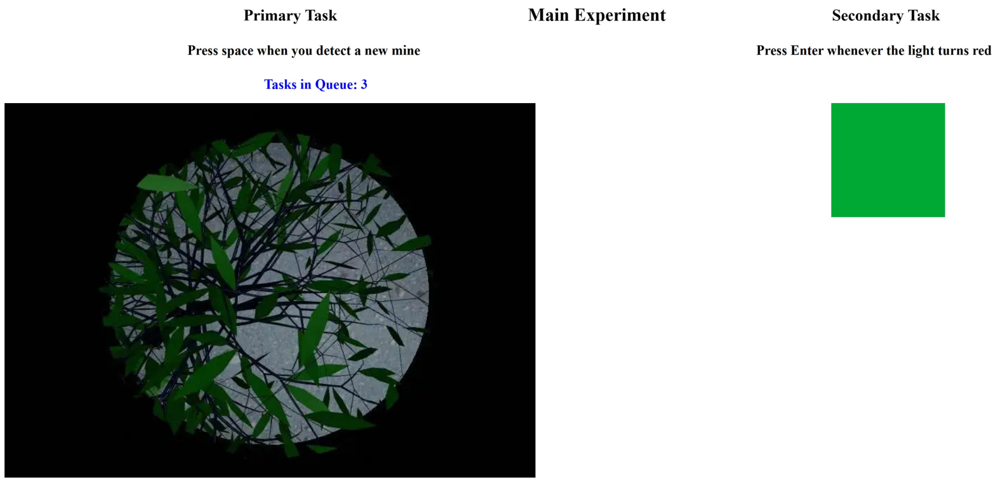
In this chapter, we study the problem of optimal fidelity selection for a human operator performing an underwater visual search task. Human performance depends on various cognitive factors such as workload and fatigue. We perform human experiments in which participants perform two tasks simultaneously: a primary task, which was subject to evaluation, and a secondary task to estimate their workload. The primary task requires participants to search for underwater mines in videos, while the secondary task involves a simple visual test where they respond when a green light displayed on the side of their screens turns red. Videos arrive as a Poisson process and are stacked in a queue to be serviced by the human operator. The operator can choose to watch the video with either normal or high fidelity, with normal fidelity videos playing at three times the speed of high fidelity ones. Participants receive rewards for their accuracy in mine detection for each primary task and penalties based on the number of videos waiting in the queue. We consider the workload of the operator as a hidden state and model the workload dynamics as an Input-Output Hidden Markov Model (IOHMM). We use a Partially Observable Markov Decision Process (POMDP) to learn an optimal fidelity selection policy, where the objective is to maximize total rewards. Our results demonstrate improved performance when videos were serviced based on the optimal fidelity selection policy compared to a baseline where humans chose the fidelity level themselves.

3.1 Background and Problem Formulation

We now discuss our problem setup, formulate it as POMDP, and solve it to obtain the optimal fidelity selection policy.

3.1.1 Problem Setup

We study the problem of optimal fidelity selection for a human operator performing a visual search task. The human operator performs two tasks simultaneously, a primary task and a secondary task. The primary task involves searching for underwater mines in videos



Primary Task

Main Experiment

Secondary Task

Press space when you detect a new mine

Press Enter whenever the light turns red

Tasks in Queue: 3

Figure 3.1 Human experiment interface. The participants press the spacebar key whenever a new mine is detected in the primary task video. Additionally, the green light (secondary task) randomly turns red once for each primary task and the participant responds by pressing the Enter key as early as possible. The queue length (tasks waiting in the queue) is displayed on top of the primary task.

generated from an underwater simulation designed using Gazebo [89] and ROS [90]. The operator watches videos and responds by pressing a key whenever a mine is spotted. On the other hand, the secondary task involves a simple visual exercise, where participants press a key as early as possible when a green light located at the side of the screen changes to red. During each primary task, the green light undergoes this transition randomly, occurring between the 25% and 75% mark of the video. We record the participants' reaction time in the secondary tasks, and in a rare event when a participant misses the red light, the reaction time is set to the total time the light stays red until the end of the primary task.

Fig. 3.1 shows the experiment interface. The videos for the primary task arrive as a Poisson process with an arrival rate of $\lambda \in \mathbb{R}_{>0}$ and get stacked in a queue awaiting service by the human operator. The operator has the option to select either high or normal fidelity levels for servicing each video. In normal fidelity, the video is presented at a speed of three times faster compared to high-fidelity processing. Additionally, operators can choose to delegate a task for autonomous processing, even though the accuracy of the autonomous system may be lower. This delegation or “skip" action serves as a means to maintain queue stability, especially in situations with large queue lengths.

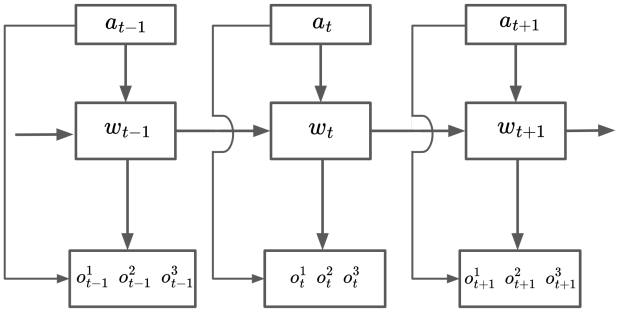

a_{t-1}
a_t
a_{t+1}
w_{t-1}
w_t
w_{t+1}
o^1_{t-1} o^2_{t-1} o^3_{t-1}
o^1_t o^2_t o^3_t
o^1_{t+1} o^2_{t+1} o^3_{t+1}

Figure 3.2 Input-output hidden Markov model. The input a represents the fidelity level, the hidden state w signifies the workload, and the three observations o^1, o^2, and o^3 correspond to the fraction of correctly detected mines in the primary task, the count of false alarms in the primary task, and the reaction time recorded during the execution of the secondary task, respectively.

The performance of a human operator relies on their workload, making it a crucial factor in our problem formulation. To address this, we formulate the problem as a POMDP, with the workload of the operator treated as a latent or hidden variable. We estimate this hidden workload through a combination of reaction time measurements from the secondary task and performance metrics obtained from the primary task. Specifically, we model the workload dynamics of the human operator using an IOHMM (see Fig. 3.2). In this model, the fidelity level a serves as the input, the workload w operates as the hidden state, and we observe three distinct output measures o^1, o^2, o^3. These output measures correspond to the fraction of correctly detected mines in the primary task, the count of false alarms in the primary task, and the reaction time recorded during the execution of the secondary task, respectively. This modeling approach helps us understand how fidelity, workload, and task performance are interconnected in one unified framework.

We utilize the extended Baum-Welch algorithm [91] to train the IOHMM model, which provides the transition probabilities $p(w'|w, a)$, observation probabilities $p(o|w, a)$, and the initial state distribution $p(w_0)$ through expectation maximization. These probabilities are utilized in the POMDP formulation to obtain the optimal fidelity selection policy.

To determine the most suitable number of hidden workload states, we use the Akaike information criterion (AIC) [92] and Bayesian information criterion (BIC) [93] for model selection. For a given number of hidden states, the AIC and BIC are defined as:

$$AIC = 2p - 2\log(\hat{\mathcal{L}}), \quad BIC = p\log(n_o) - 2\log(\hat{\mathcal{L}}), \tag{3.1}$$

where p represents the number of learned parameters, n_o stands for the number of observation trajectories, and $\hat{\mathcal{L}}$ signifies the maximized value of the log-likelihood function derived from the trained model. The model with the lowest AIC (or BIC) value is considered the optimal choice, as determined by the AIC (or BIC) criteria. Based on these criteria (presented in Section 3.2.3), we train an IOHMM model with two hidden states, which we refer to as normal and high workload states.

3.1.2 Mathematical Modeling

We formulate our problem as a POMDP $\mathcal{P} = \{\mathcal{S}, \mathcal{A}, \Omega, \mathcal{T}, \mathcal{O}, r, \gamma\}$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ in the action space, Ω is the set of observations, $\mathcal{T}$ is the set of conditional transition probabilities between states, $\mathcal{O}$ is the set of conditional observation probabilities, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor. The state space of the system is defined using a tuple $\mathcal{S} = (q, w)$, where $q \in \{0, 1, \ldots, L\}$ and $w \in \mathcal{W} = \{0, 1\}$ denotes the number of tasks waiting in the queue (with maximum queue length L) and the hidden discrete workload of the human operator, respectively. We define $w = 0$ as the normal workload state and $w = 1$ as the high workload state. The action space is defined as $\mathcal{A} = \{N, H, D\}$, with N and H representing normal and high-fidelity task processing, respectively. The action D corresponds to task delegation, allowing the task to be handled by autonomous systems. This delegation action is particularly useful for maintaining queue stability when dealing with a large queue length. To discourage excessive use of the delegation action, we assume that the autonomy's accuracy in mine detection is significantly lower which results in a lower immediate reward.

The observation space is defined by a tuple $\Omega = (o^1, o^2, o^3)$, where o^1, o^2, and o^3 correspond to the fraction of correctly detected mines in the primary task, the count of false alarms in the primary task, and the reaction time recorded during the execution of the secondary task, respectively. The state transition probability $p(s'|s,a) \in \mathcal{T}$ is derived from the queue dynamics $p(q'|q,a)$ given by the Poisson distribution with arrival rate λt, where t is the duration of the task, and workload dynamics $p(w'|w,a)$ obtained by training the IOHMM. The observation probabilities $p(o|s',a) = p(o^1, o^2, o^3|w',a) \in \mathcal{O}$ are also obtained from the trained IOHMM model. The observation probabilities are assumed to be independent of the queue length, and therefore, only depend on the workload of the operator, i.e., $p(o|s',a) = p(o|w',a)$. We define the reward function as $r(s,a) = \alpha_1 o^1 - \alpha_2 o^2 - \alpha_3 q$, where α_i for $i \in \{1,2,3\}$ are positive constants, which rewards high accuracy in the primary task and penalizes for the number of tasks waiting in the queue.

We convert the POMDP to a belief MDP defined by $\mathcal{M} = \{\mathcal{B}, \mathcal{A}, \tau, r, \gamma\}$. Here, $\mathcal{B} := \{(q, b_H)|\ q \in \{0,1,\ldots,L\}, b_H \in \Delta_D\}$ is the new state space, where q is the original queue length, b_H is the discrete belief probability for being in the high workload state, and Δ_D is a discretization of the interval $[0,1]$. Therefore, the belief probability for being in the normal workload state is given by $1 - b_H$. For our experiments, we discretize $[0,1]$ with a step size of 0.1, i.e., $\Delta_D = \{0, 0.1, \ldots, 1\}$. Note that the discretization of b_H results in a finite state space $\mathcal{B}$. Let $b : \mathcal{W} \to \Delta_D$ denote the belief vector, where $b(0) = 1 - b_H$, and $b(1) = b_H$. From a current belief $b(w)$, taking an action a and observing o, the updated belief $b'(w')$ is given by:

$$b'(w') = \eta p(o|w',a) \sum_{w \in \mathcal{W}} p(w'|w,a) b(w), \tag{3.2}$$

where $\eta = \frac{1}{p(o|b,a)}$ is the normalizing constant with

$$p(o|b,a) = \sum_{w' \in \mathcal{W}} p(o|w',a)) \sum_{w \in \mathcal{W}} p(w'|w,a) b(w). \tag{3.3}$$

The updated belief probabilities in $b'(w')$, where $b'(0) = 1 - b'_H$ and $b'(1) = b'_H$ obtained from (3.2) are mapped to the closest discrete states such that $b'_H \in \Delta_D$. The action set $\mathcal{A}$ is the original action space. The transition probabilities $\tau(q', b'|q, b, a) = p(q'|q,a)p(b'|b,a)$ is composed of the Poisson process for queue dynamics $p(q'|q,a)$, and the workload dynamics

$p(b'|b,a)$, which is given by:

$$p(b'|b,a) = \sum_{o \in \mathcal{O}} p(b'|b,a,o)p(o|b,a), \tag{3.4}$$

where

$$p(b'|b,a,o) = \begin{cases} 1, & \text{if belief update in (3.2) returns } b', \\ 0, & \text{otherwise,} \end{cases} \tag{3.5}$$

and $p(o|b,a)$ is defined in (3.3). The reward function $r : \mathcal{B} \times \mathcal{A} \to \mathbb{R}$ is given by:

$$r(q, b_H, a) = \sum_{s \in \mathcal{S}|w=0} r(s,a)(1-b_H) + \sum_{s \in \mathcal{S}|w=1} r(s,a)b_H, \tag{3.6}$$

where $r(s,a)$ is the original reward function for the POMDP. The discount factor γ is the original discount factor of the POMDP. For the belief MDP $\mathcal{M}$, the expected value for policy π starting from an initial state $(q_0, b_{H,0})$ is defined as:

$$\begin{aligned} V^{\pi}(q_0, b_{H,0}) &= \sum_{t=0}^{\infty} \gamma^t r(q_t, b_{H,t}, a_t) \\ &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}\left[r(s_t, a_t)|q_0, b_{H,0}, \pi\right], \end{aligned} \tag{3.7}$$

where the expectation is computed over $(q_t, b_{H,t}, w_t)$. The optimal fidelity selection policy maximizes the value in each belief state, i.e.,

$$\pi^* = \arg\max_{\pi} V^{\pi}(q_0, b_{H,0}).$$

We utilize the value iteration algorithm to solve the belief MDP and obtain the optimal fidelity selection policy.

3.2 Human Experiments

In this section, we discuss the design of our human experiments conducted using Prolific (`www.prolific.com`).

3.2.1 Experimental Setup

We developed an underwater mine search experiment within the Robot Operating System (ROS) framework using Gazebo models. For our simulation, we employed the "UUV simulator" [94], a comprehensive package encompassing Gazebo plugins and ROS nodes specifically

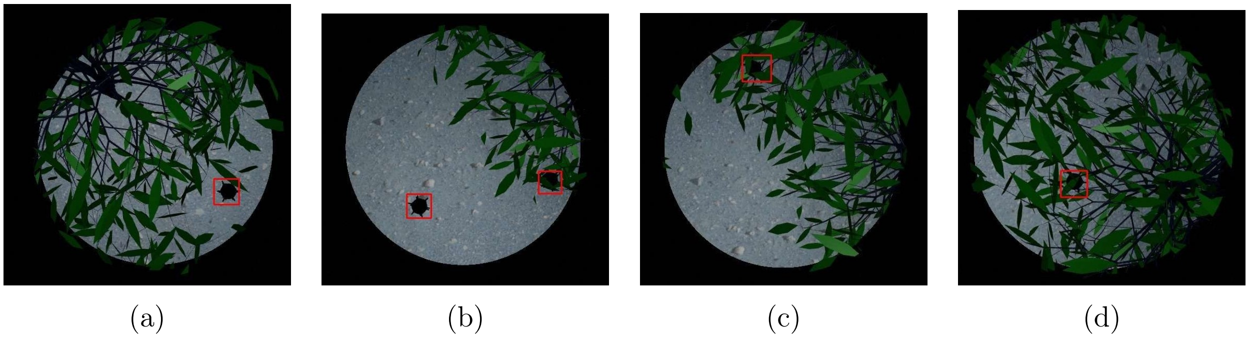


Figure 3.3 Example image frames containing mines recorded from ROV. The mines' positions are highlighted within red bounding boxes in the frames.

tailored for simulating unmanned underwater vehicles, including remotely operated vehicles (ROVs) and autonomous underwater vehicles (AUVs).

In our experimental setup, we designed an underwater scenario where we placed underwater mines randomly throughout the environment following a uniform distribution within a predefined area. To increase the complexity of the task, we also randomly introduced a significant amount of underwater vegetation, which added an extra layer of difficulty in detecting the mines.

We captured underwater videos by deploying an ROV into the environment. The ROV followed a predetermined circular trajectory as it moved through the underwater space. Equipped with a downward-facing camera, the ROV recorded images. To ensure optimal recording conditions, we maintained complete darkness in the environment. The sole source of illumination was a single light attached to the ROV, directing its beam vertically downward. Consequently, only the area directly beneath the ROV was within the recording scope, resulting in a focused and well-illuminated view. We recorded a total of 5600 images which are used as 56 videos, with each video consisting of 100 frames. Fig. 3.3 shows example image frames containing underwater mines recorded from ROV. The mines' positions are highlighted within red bounding boxes in the frames.

We performed 5 set of experiments, with each experiment group consisting of 20 participants. In each experiment, the participants first performed 8 practice tasks followed by 48 main tasks, where each task consists of a primary and a secondary task. Following is the list of experiments.

- *Experiment* 1*:* The first experiment was the base experiment where we recorded data to train the IOHMM model and learn the optimal fidelity selection policy by solving the POMDP. In this experiment, the 48 videos were evenly divided, with half of them being displayed at normal fidelity and the other half at high fidelity. To maintain consistency and balance, the tasks were presented in alternating blocks of four tasks each. Specifically, the sequence of 48 tasks was organized as $\{N, N, N, N, H, H, H, H, \ldots\}$ for half of the participants and $\{H, H, H, H, N, N, N, N, \ldots\}$ for another half of the participants. This reversal was thereby used to prevent ordering bias in our estimates.

- *Experiment 2:* In this experiment, the participants were allowed to choose a fidelity level for servicing each task. Before releasing each task, the participant decided on their desired fidelity level by pressing a key. This served as a baseline, where the human operator received no decision support. In this experiment, we do not allow task delegation, and the participants were only allowed to choose to service each task with either a normal or high fidelity level.

- *Experiment 3:* In this experiment, the decision support system determined the appropriate fidelity level for each task by considering both the current operator performance and the queue length. Specifically, the decision support system monitored the operator's performance in each task, updated its belief using (3.2), and selected the optimal action in accordance with the optimal fidelity selection policy. For this experiment, we utilized an optimal policy that comprises only two available actions: normal and high fidelity.

- *Experiment 4:* This experiment was similar to Experiment 2, with the modification that the participants were provided an additional action of delegating the task to the autonomous system.

- *Experiment 5:* This experiment was similar to Experiment 3, with the modification that the optimal policy consisting of three actions: normal fidelity, high fidelity, and task delegation was used to choose the fidelity level for each task.

3.2.2 Methods

After receiving the IRB consent (MSU IRB #9452) from Michigan State University's IRB office, we recruited 100 participants using Prolific for the study. Inclusion criteria were established as having completed a minimum of 500 prior studies and maintaining a 99% approval rate on the platform. Participants were compensated with a base payment of \$6 and had the opportunity to earn additional performance-based bonuses ranging from $\$0-\$4$.

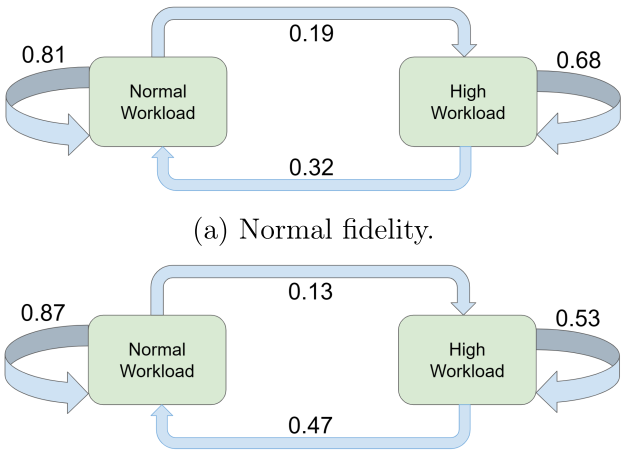
3.2.3 IOHMM results

We use the data from Experiment 1 to train IOHMM models with different numbers of hidden states. Table 3.1 illustrates the AIC and BIC values (normalized with the number of observation trajectories) for the trained IOHMM models with 2, 3, and 4 hidden workload states, respectively. Based on the AIC and BIC criterion, we utilize a trained IOHMM model with two hidden states, which we refer to as normal and high workload states.

Hidden States	2	3	4
AIC	**735.29**	735.91	736.28
BIC	**736.39**	737.85	739.26

Table 3.1 AIC and BIC values for model selection.

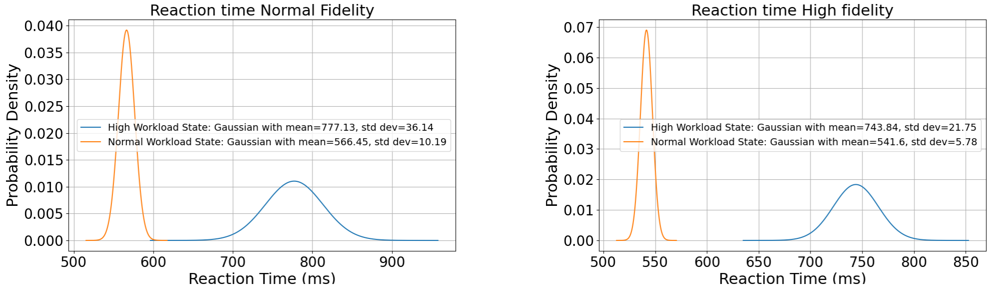
Fig. 3.4a and 3.4b show the workload transition diagram for the trained IOHMM model under normal and high fidelity, respectively. Under high fidelity, due to the slower speed of videos, there is a higher probability of transitioning from the high workload state to a lower workload state as compared to under normal fidelity. Similarly, the probability of transitioning into a high workload state from the normal workload state is higher under normal fidelity as compared to the high fidelity servicing. In the case of task delegation action, the task is instantaneously removed from the queue and hence, we assume that the workload remains the same under task delegation.



(a) Normal fidelity.

(b) High fidelity.

Figure 3.4 Workload transition diagram under (a) normal and (b) high fidelity servicing of tasks.



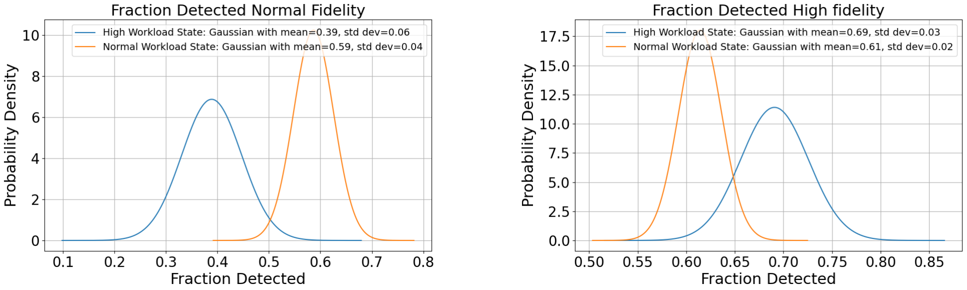
(a) Normal fidelity.

(b) High fidelity.

Figure 3.5 Reaction time diagram under (a) normal and (b) high fidelity servicing of tasks.

For observations o^1, o^2, and o^3, we learn a normal distribution with unknown mean and standard deviation for each state-action pair. Fig. 3.5a and 3.5b present the reaction time distributions for normal and high fidelity, respectively. We identify state 0 as a normal workload state, while state 1 represents the high workload state. The mean and the variance of the reaction time in the secondary task are larger in the high workload state than in the normal workload state.

Fig. 3.6a and 3.6b depict the distributions for the fraction of detected mines in the primary task for normal and high fidelity, respectively. Notably, we observe a substantial difference in the means of the fraction of detected mines between normal and high workload states under normal fidelity. In contrast, the means are relatively similar between these

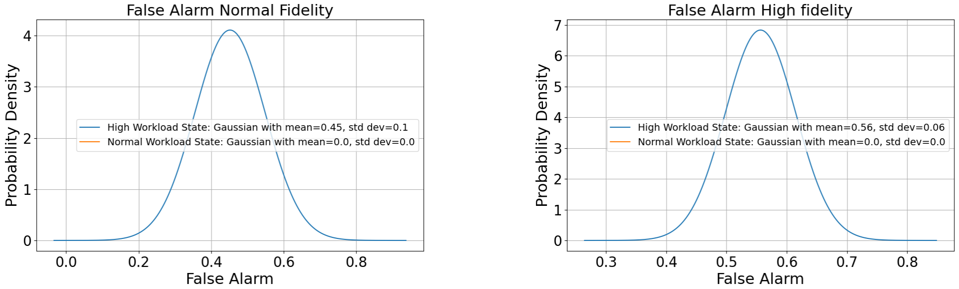


(a) Normal fidelity.

(b) High fidelity.

Figure 3.6 Fraction of mines detected in primary task under (a) normal and (b) high fidelity servicing of tasks.

states under high fidelity. This observation suggests that when we do not take into account the penalty associated with the queue length, choosing high fidelity (slower videos) could be a suitable action during high workload conditions.



(a) Normal fidelity.

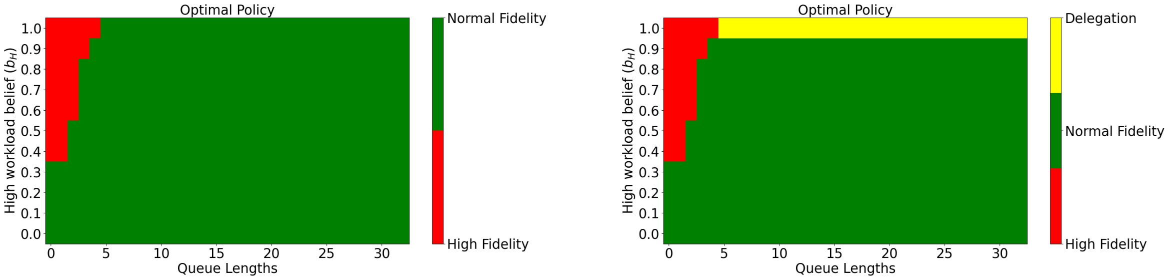
(b) High fidelity.

Figure 3.7 Number of false alarms in primary task under (a) normal and (b) high fidelity servicing of tasks.

Fig. 3.7a and 3.7b illustrate the distributions for the number of false alarms in the primary task for normal and high fidelity, respectively. Notably, in the normal workload state, no false alarms were recorded. Therefore, in the normal workload state, we replace the normal distribution with a Dirac delta function that yields a probability of 1 at 0 false alarms and 0 everywhere else.

Finally, the initial distribution for the workload was determined as $[0.662,\ 0.338]$, where 0.662 is the probability of starting in a normal workload state. To solve the POMDP, we discretize the distributions of the reaction time, fraction of mines detected, and the false alarms, with a step size of 25 ms, 0.05, and 0.5, respectively.

3.2.4 Optimal Policy



(a) Task delegation unavailable.

(b) Task delegation available.

Figure 3.8 Optimal fidelity selection policy where the action space (a) does not include task delegation and (b) includes task delegation.

Using the distributions from trained IOHMM, we convert the POMDP into a belief MDP as detailed in Sec. 3.1.2. We utilize the following reward function:

$$r(s,a) = \begin{cases} 100o^1 - 30o^2 - 2q, & \text{for } a \in \{N, H\}, \\ 30 - 2(q-1), & \text{for } a = D, \end{cases} \tag{3.8}$$

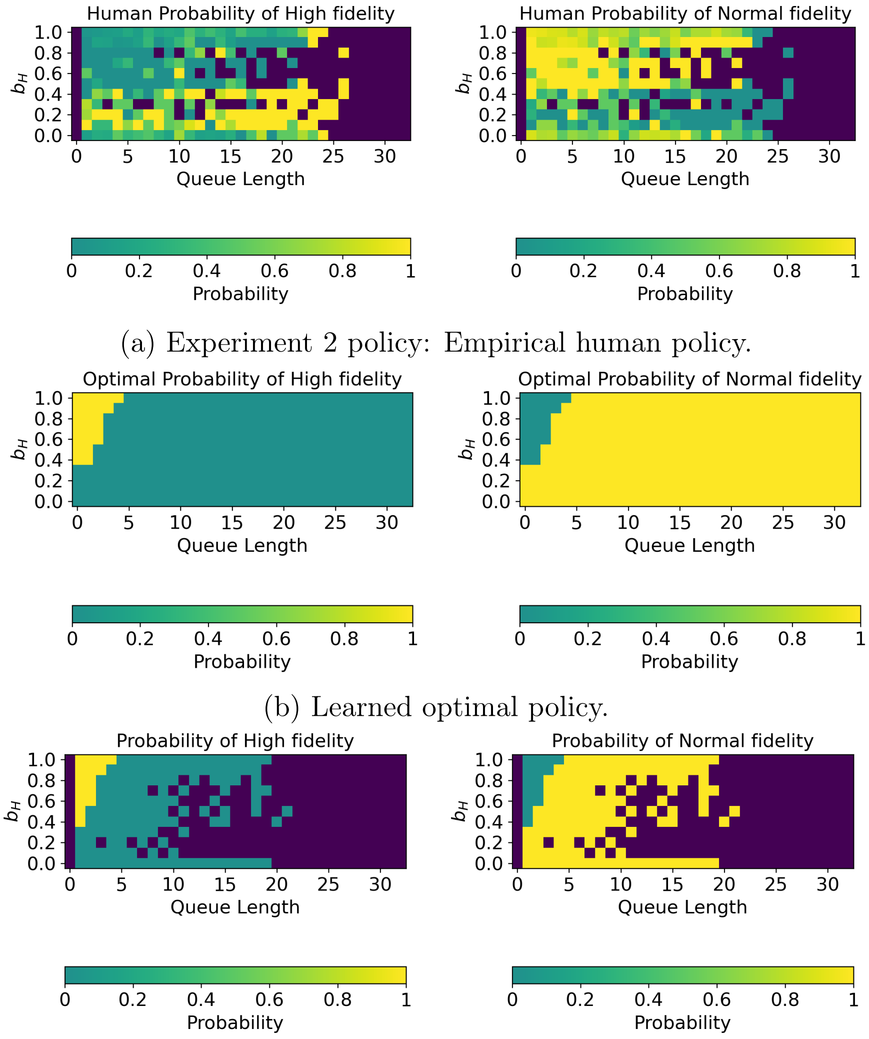
where we assumed the accuracy of autonomous servicing (task delegation) to be just 30%. We employed the value iteration algorithm to derive an optimal fidelity selection policy.

We developed two optimal policies: one with only two available actions and another with three available actions, including task delegation.

Fig. 3.8a and 3.8b illustrate the optimal policy with two and three available actions, respectively. In situations characterized by low queue lengths and a higher level of belief that workload is high, opting for high-fidelity servicing emerges as the optimal action. For other regions of the state space, normal fidelity servicing is the optimal action. When task delegation is a viable option, it becomes the optimal action only when there is a near certainty of being in the high workload state, as indicated by a belief close to 1, and the queue length is substantial.

3.3 Results

We now discuss the results of the experiments.



(a) Experiment 2 policy: Empirical human policy.

(b) Learned optimal policy.

(c) Experiment 3 policy: Empirical optimal policy.

Figure 3.9 (a) Empirical human policy in Experiment 2, (b) Learned optimal policy by solving POMDP with two available actions, and (c) Empirical policy obtained using data from Experiment 3 that deploys optimal fidelity selection policy.

Fig. 3.9 compares the policy utilized in Experiments 2 and 3 that only allowed servicing a task with either normal or high fidelity. Fig. 3.9a, 3.9b, and 3.9c show the empirical human policy obtained from Experiment 2 data, complete optimal policy learned from POMDP, and the empirical policy obtained using data from Experiment 3 that deploys optimal fidelity selection policy, respectively. The two columns of the plots show the probability of choosing high-fidelity action and normal-fidelity action, respectively. The dark purple region in the plots represents the portions of the state space that have not been visited during the

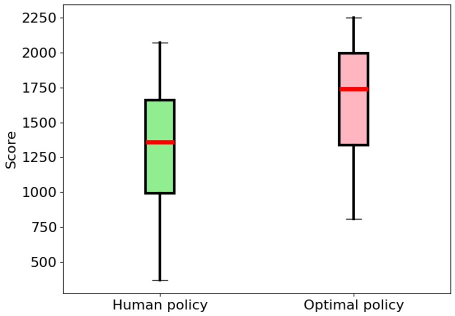


Figure 3.10 Box plots for the participants' scores in Experiment 2 (human policy) and Experiment 3 (optimal policy). Within each box plot, the median is represented by the red horizontal line, while the lower and upper edges of the box signify the 25th and 75th percentiles, respectively. Whiskers extend to encompass the most extreme data points. The p value in the two-sample t-test comparing the outcomes of Experiments 2 and 3 was computed as 0.017, indicating a high level of statistical significance.

experiment.

By comparing the human policy with the optimal policy, we can gain important insights into human behavioral patterns. It is evident that the human policy exhibits minimal variation in fidelity with respect to the queue length. This suggests that humans struggle to effectively balance the trade-off between their ongoing tasks and the pending tasks in the queue.

Furthermore, human policy reveals a tendency for high-fidelity servicing in low-workload states and normal-fidelity servicing in high-workload states. This implies that humans tend to have difficulty switching between actions. Participants who prioritize high accuracy in primary tasks opt for high-fidelity servicing, resulting in normal workload conditions. Conversely, those who prioritize speed choose normal fidelity servicing, which can lead to operating under high workload conditions. Consequently, humans appear to have difficulty accurately assessing their workload and performance, hampering their ability to switch actions effectively.

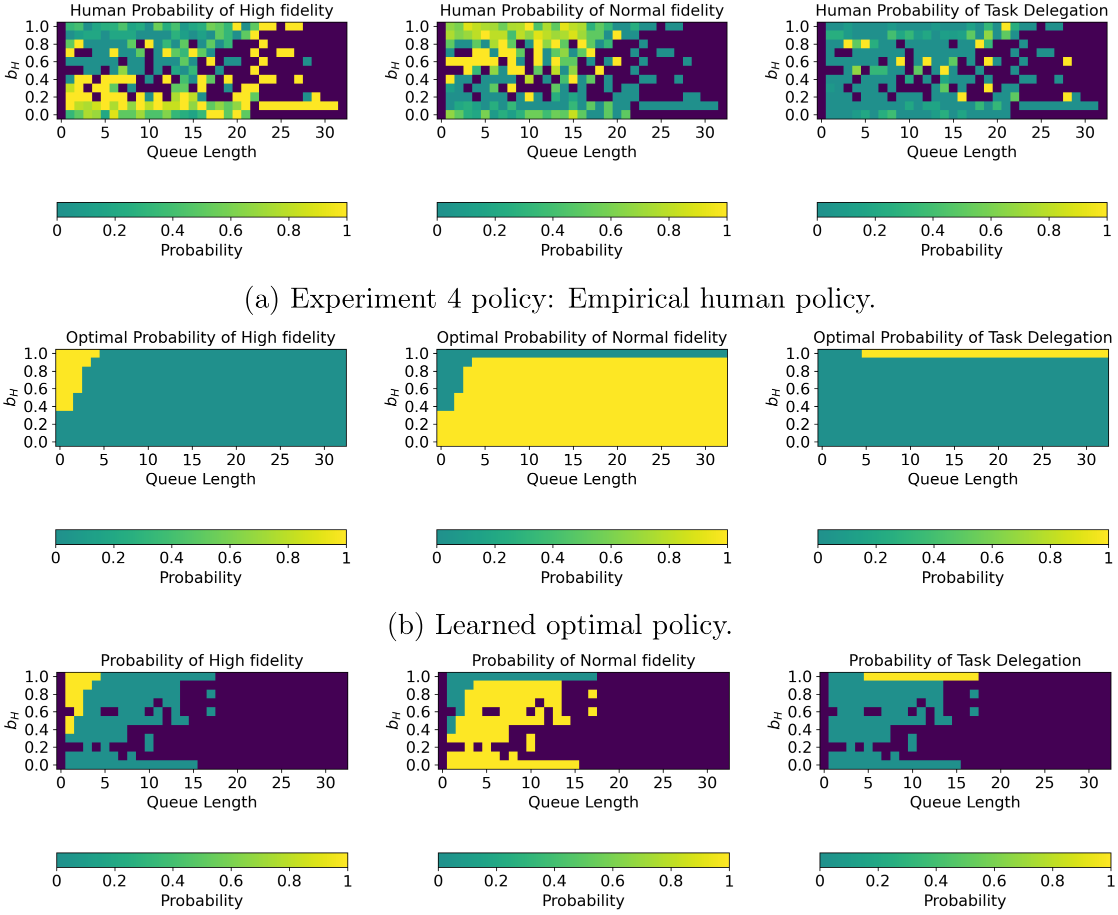
Lastly, the empirical policy obtained using data from Experiment 3 indicates that the optimal policy effectively manages the queue length, thereby keeping the region of high queue length unexplored during the experiment.

Fig. 3.10 presents the comparative box plots for the scores obtained in Experiment 2 (human policy) and Experiment 3 (optimal policy). Here, the score refers to the cumulative reward accrued by a participant over tasks, i.e., Score $= \sum_{t=1}^{48} r_t$, where r_t denotes the reward obtained for servicing the task t given by (3.8). It can be seen that the performance under optimal policy is much better than the performance under human policy. Furthermore, under the optimal policy, we observe an improvement of 26.54% in the average total score as compared to the human policy.

To assess the statistical significance of these findings, we conducted a two-sample t-test comparing the outcomes of Experiments 2 and 3. Remarkably, the p value was computed as 0.017, indicating a notably high level of significance. In line with the widely accepted significance threshold of 0.05, a p value below this threshold prompts us to reject the null hypothesis. This implies that the data from the two experiments do not stem from the same distribution at a 5% significance level. These results underscore the substantial impact of the optimal policy in enhancing human performance.

Next, we allow an additional action of task delegation under which a task is instantaneously removed from the queue to be serviced by the autonomy. Fig. 3.11 compares the policy utilized in Experiments 4 and 5 that allowed all three actions. Fig. 3.11a, 3.11b, and 3.11c show the empirical human policy obtained using data from Experiment 4, complete optimal policy learned from POMDP, and the empirical policy obtained using data from Experiment 3 that deploys optimal fidelity selection policy, respectively. The three columns of the plots show the probability of choosing high-fidelity action, normal-fidelity action, and task delegation, respectively. The dark purple region in the plots represents the portions of the state space that have not been visited during the experiment.

Similar observations to those depicted in Fig. 3.9 can also be made for high and normal fidelity servicing in Fig. 3.11. Moreover, it is noticeable that the human policy exhibits a somewhat random utilization of task delegation. This observation suggests that humans struggle with workload management and effective task delegation. Lastly, the optimal policy



(a) Experiment 4 policy: Empirical human policy.

(b) Learned optimal policy.

(c) Experiment 5 policy: Empirical optimal policy.

Figure 3.11 (a) Empirical human policy in Experiment 4, (b) Learned optimal policy by solving POMDP with three available actions, and (c) Empirical policy obtained using data from Experiment 5 that deploys optimal fidelity selection policy.

in Experiment 5 keeps the queue length under check, and therefore, regions of higher queue lengths remain unvisited in the experiment.

Fig. 3.12 shows the box plots for the participants' scores in Experiment 4 (human policy) and Experiment 5 (optimal policy). It can be seen that the performance under optimal policy is much better than the performance under human policy. Furthermore, under the optimal policy, we observe an improvement of 50.3% in the average total score as compared to the human policy. Lastly, the p-value in the two-sample t-test is given by 0.001, showing that the results are statistically significant.

Remark 5. *The average score achieved with the optimal policy in Experiment* 5 *was marginally*

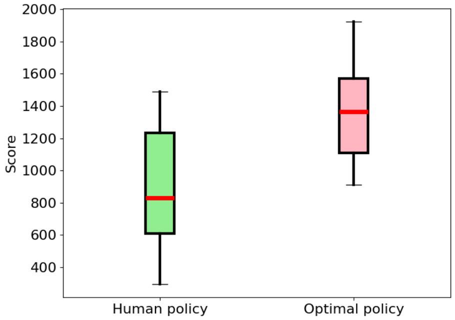


Figure 3.12 Box plots for the participants' scores in Experiment 4 (human policy) and Experiment 5 (optimal policy). Within each box plot, the median is represented by the red horizontal line, while the lower and upper edges of the box signify the 25th and 75th percentiles, respectively. Whiskers extend to encompass the most extreme data points. The p value in the two-sample t-test comparing the outcomes of Experiments 4 and 5 was computed as 0.001, indicating a high level of statistical significance.

less than that obtained with the optimal policy in Experiment 3*. This discrepancy could be due to the relatively smaller reward obtained from task delegation, as opposed to servicing the task with normal fidelity. Although the optimal policy in Experiment* 5 *may default to task delegation under the presumption of diminished performance during periods of high workload, certain participants may demonstrate the capability to execute tasks with high accuracy even under high workload conditions in Experiment* 3*.*

3.4 Conclusions

We studied the problem of optimal fidelity selection for a human operator engaged in a visual search task. The human participants performed two tasks simultaneously: a primary mine search task and a secondary visual task used for estimating the human workload. We treated the human workload as a hidden state and modeled the workload dynamics using an Input-Output Hidden Markov Model (IOHMM). Leveraging the probability distributions derived from the IOHMM, we formulated a Partially Observable Markov Decision Process (POMDP) and solved it to derive an optimal fidelity selection policy. The results of our experiments offer valuable insights into human behavioral patterns and underscore the substantial performance enhancements achievable through the application of the optimal policy.

CHAPTER 4

ROBUST AND ADAPTIVE FIDELITY SELECTION FOR HUMAN-IN-THE-LOOP QUEUES

In this chapter, we relax the assumption of the known human service time model in the optimal fidelity selection problem. We assume the parameters of the human's service time distribution depend on the selected fidelity level and her cognitive state and are assumed to be unknown a priori. These parameters are learned online through Bayesian parameter estimation. We formulate a robust adaptive SMDP to solve our optimal fidelity selection problem. We extend the results on the convergence of robust-adaptive MDP to robust-adaptive SMDPs and show that the solution of the robust adaptive SMDP converges to the optimal solution for the uncertainty-free SMDP. Furthermore, we numerically illustrate the convergence of the synchronous and asynchronous robust adaptive policy to the uncertainty-free optimal policy.

4.1 Background and Problem Formulation

We now present our problem setup, and formulate the optimal fidelity selection problem as a robust adaptive SMDP.

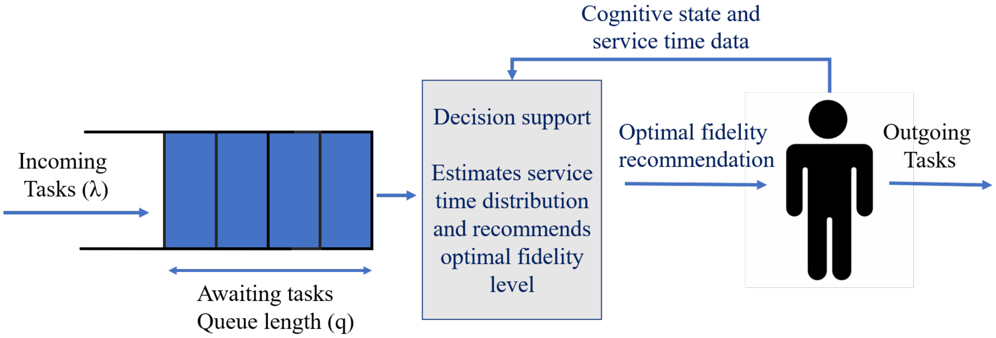

Cognitive state and service time data
Incoming Tasks (λ)
Awaiting tasks Queue length (q)
Decision support
Estimates service time distribution and recommends optimal fidelity level
Optimal fidelity recommendation
Outgoing Tasks

Figure 4.1 Schematic of our problem setup. The incoming tasks arrive at a constant arrival rate λ and gets stored in a queue. The decision support learns the human service-time distribution based on the online observations and recommends an optimal fidelity level to the human agent based on the system state (queue length and cognitive state).

4.1.1 Problem Setup

We consider a human agent with unknown service time distribution servicing a queue of homogeneous tasks. We assume the availability of approximate model for human service time distribution from human subject experiments, which we adapt online to estimate the agent's distribution using Bayesian parametric estimation.

The homogeneous tasks arrive according to a Poisson process with a constant arrival rate $\lambda \in \mathbb{R}_{>0}$, and are stored in a dynamic queue with a maximum capacity $L \in \mathbb{N}$, until serviced by the human agent in a first-come-first-serve discipline. These tasks continuously lose value at a constant rate $c \in \mathbb{R}_{>0}$ while waiting in the queue. The agent can choose to service the tasks with normal or high fidelity level. When the agent services a task with high fidelity, she meticulously looks into the details, which results in higher-quality service but leads to larger service times and increased operator tiredness. We treat the cognitive state of the agent as a lumped parameter that captures psychological factors such as fatigue, stress and situational awareness. We assume that the unknown mean service time of the agent increases with the fidelity level and is a unimodal function of the cognitive state, which is inspired from the experimental psychology literature. For example, according to Yerkes-Dodson law [16], excessive stress overwhelms the operator and too little stress leads to reduction in vigilance. Hence, the human performance is optimal for some intermediate cognitive state.

Fig. 4.1 shows a schematic of our problem setup. We are interested in design of a decision support system that continuously learns the human service-time distribution and assists the agent by recommending an optimal fidelity level to service each task. The recommendation is made based on the robust adaptive policy learned by the decision support for a given queue length $q \in \mathbb{Z}_{\geq 0}$ and the human cognitive state. We assume to have real time access to the human cognitive state using, e.g., Electroencephalogram (EEG) measurements (see [80] for measures of cognitive load from EEG data).

4.1.2 Robust Adaptive SMDP formulation

We now model our problem as a robust adaptive SMDP Γ^{RA}. We focus on the unknown service time distributions and refer the interested readers to [41] for more details on modeling of uncertainty-free distributions.

We consider a finite state space $\mathcal{S} := \{(q, \text{cog})|\ q \in \{0, 1, ..., L\},\ \text{cog} \in \mathcal{C} := \{i/N\}_{i \in \{0,...,N\}}$, for some $N \in \mathbb{N}$, where cog represents the lumped cognitive state. We consider five possible actions for the agent given by: (i) ***Waiting (W)***, when the queue is empty, (ii) ***Resting (R)***, which provides the resting time for the human operator to reach the optimal cognitive state when tired, (iii) ***Skipping (S)***, which allows the operator to skip a task to reduce the queue length and thereby focus on newer tasks, (iv) ***Normal Fidelity (N)*** for servicing the task with normal fidelity, and (v) ***High Fidelity (H)*** for servicing the task more carefully with high precision. Hence, a set of admissible actions $\mathcal{A}_s$ for each state $s \in \mathcal{S}$ is given by: (i) $\mathcal{A}_s := \{\, W \,|\ s \in \mathcal{S},\ q = 0\}$ when queue is empty, (ii) $\mathcal{A}_s := \{\{R,\ S,\ N,\ H\ \}|\ s \in \mathcal{S},\ q \neq 0\}$ when queue is non-empty and cog $>$ cog*, where cog$^* \in \mathcal{C}$ is the optimal cognitive state, and (iii) $\mathcal{A}_s := \{\{S,\ N,\ H\ \}|\ s \in \mathcal{S},\ q \neq 0\}$ when queue is non-empty and cog $\leq$ cog*.

Let τ be the sojourn time spent in state s. The sojourn time distribution $\mathbb{P}(\tau|\ s, a)$ represents the service time while servicing the task with normal or high fidelity, resting time, constant time of skip, and time until the next task arrival while waiting. We model service time distributions in Section 4.1.3 and refer the readers to [41] for details on resting and waiting time. We define a state transition distribution $\mathbb{P}(s'|\ \tau, s, a)$ from state s to s' conditioned on an action $a \in \mathcal{A}_s$ and sojourn time τ spent in state s. This distribution involves the transition in queue length which is given by Poisson distribution and the cognitive dynamics which we model as a Markov chain such that the cognitive state cog increases with high probability when the operator is busy ($a \in \{N, H\}$), and decreases when the operator is idle ($a \in \{R, W\}$). For a detailed description of the modeling of cognitive dynamics, we refer the interested readers to [41].

For each task, the human agent receives a high (low) immediate reward for servicing the task with high (normal) fidelity, and no reward for not servicing the task. Furthermore, the agent incurs a penalty at a constant rate $c \in \mathbb{R}_{>0}$ for each task waiting in the queue. Hence, it can be shown that the expected net immediate reward received by the agent for selecting an action a in state s is given by:

$$R(s, a) = r(s, a) - \sum_{\tau} \mathbb{P}(\tau | s, a) c \Big(\frac{2q + \lambda \tau}{2} \Big) \tau, \tag{4.1}$$

where $r : \mathcal{S} \times \mathcal{A}_s \mapsto \mathbb{R}_{\geq 0}$ is the reward defined by: (i) $r(s, a) = r_H$, if $a = H$; (ii) $r(s, a) = r_N$, if $a = N$; and $r(s, a) = 0$, if $a \in \{W, R, S\}$, with $r_H, r_N \in \mathbb{R}_{\geq 0}$ and $r_H > r_N$, and $\sum_\tau \mathbb{P}(\tau|s, a) c \tau \Big(\mathbb{E} \left[\frac{q+q'}{2} \middle| \, \tau, s, a \right] \Big)$ is the expected penalty due to tasks waiting in the queue.

4.1.3 Modeling human service time and uncertainity set

There are many approximate models used for modeling service time distribution $\mathbb{P}(\tau|s, a)$ for the human agents, most common being lognormal [95] and inverse Gaussian distribution [96]. We assume that the agent's service time distribution follows a log-normal distribution Lognormal(μ, σ^2) with unknown parameters μ and σ^2, that are the functions of the cognitive state and fidelity-level. We utilize the Bayesian parameter estimation with a normal-inverse-chi-squared prior [97] to estimate the distribution parameters using online observations. Using the prior distribution $NI\chi^2(\mu_0, \kappa_0, \nu_0, \sigma_0^2) = \mathcal{N}(\mu | \mu_0, \sigma^2/\kappa_0) \times \chi^{-2}(\sigma^2 | \nu_0, \sigma_0^2)$ for the parameters μ and σ^2, and $n \in \mathbb{N}$ realizations from Lognormal(μ, σ^2), the posterior distribution of (μ, σ^2) is given by:

$$p(\mu, \sigma^2) = NI\chi^2(\mu_n, \kappa_n, \nu_n, \sigma_n^2), \quad \text{where} \tag{4.2}$$

$$\begin{aligned} \mu_n &= \frac{\kappa_0 \mu_0 + n \overline{x}}{\kappa_n}, \; \kappa_n = \kappa_0 + n, \; \nu_n = \nu_0 + n, \\ \sigma_n^2 &= \frac{1}{\nu_n} \left(\nu_0 \sigma_0^2 + \sum_i (x_i - \overline{x})^2 + \frac{n \kappa_0}{\kappa_0 + n} (\mu_0 - \overline{x})^2 \right), \end{aligned}$$

$x_i, \; i \in \{1, \dots, n\}$ are the service time samples and $\bar{x}$ is the sample mean. Hence, the posterior distribution at any time t can be computed to recursively estimate the model parameters μ and σ^2 by using the online observations. Let $\hat{\mathbb{P}}_t(\tau|s, a)$ be the estimate of the service time distribution at time t. Note that the estimate $\hat{\mathbb{P}}_t(\tau|s,a)$ can be used to estimate $\mathbb{P}(s',\tau|s,a)$ and $R(s,a)$, resulting in estimates $\hat{\mathbb{P}}_t(s',\tau|s,a)$ and $\hat{R}_t(s,a)$ at time t.

The uncertainty in the human service time models could be large, especially in the initial stage with limited observation data, which may lead to suboptimal policies. This can be mitigated through the use of robust SMDP. The robust SMDP optimizes the worst-case performance to obtain robust policy when the joint distribution $\mathbb{P}(s',\tau|s,a)$ lies in an uncertainty set $\mathcal{P}^a$, i.e, $\mathbb{P}(s',\tau|s,a) \in \mathcal{P}^a$. Note that the robust SMDP formulation is not inherently adaptive in nature and does not explicitly use improved transition models that can be learned using online observations to obtain less conservative policy. The robust adaptive SMDP utilizes the latest improved estimates $\hat{\mathbb{P}}_t$ and $\hat{R}_t$ for the joint probability $\mathbb{P}(s',\tau|s,a)$ and reward $R(s,a)$ at time t, respectively.

The choice of uncertainty set $\mathcal{P}^a$ is critical for the performance of the robust algorithm. A poor modeling of the uncertainty set increases the computational complexity and could lead to highly conservative robust policy. Hence, a choice of uncertainty set $\mathcal{P}^a$ is typically made such that the robust policy is not overly conservative and the optimization can be performed in a computationally tractable manner. Let $\mathcal{D}$ be the observation data up to time t. To construct an uncertainty set $\mathcal{P}_t^a$ at time t, random samples are generated from the posterior distribution of (μ,σ^2) to construct a set Δ_t comprised of matrices $\hat{\mathbb{P}}_t(s',\tau|s,a)$ for each $s \in \mathcal{S}$ and $a \in \mathcal{A}_S$. Finally, a Ψ_t-confidence level subset of the transition probabilities Δ_t defined by:

$$\mathcal{P}_t^a(\Psi_t) = \{p_t^{sa} \in \Delta_t : \|p_t^{sa} - \overline{p}_t^{sa}\|_1 \leq \Psi_t,\ s \in \mathcal{S}\}, \tag{4.3}$$

where $\overline{p}_t^{sa}$ is the nominal transition given by $\overline{p}_t^{sa} = \mathbb{E}[p_t^{sa}|\mathcal{D}]$, is used as a choice for the uncertainty set. We seek to find Ψ_t-confidence sets for state transition probability vector for every state-action pair at time t. We choose $\Psi_t = \frac{6\alpha}{|\mathcal{S}||\mathcal{A}_S|\pi^2 t^2}$ such that union bounds applied over each state-action pair and time yield that all state transition probabilities belong to respective confidence sets with at least probability α. In the following we choose $\alpha = 0.95$.

Using the uncertainty set $\mathcal{P}_t^a$ constructed at time t based on the latest improved estimates $\hat{\mathbb{P}}_t$, the robust adaptive SMDP solves the following robust Bellman equation (4.4),

$$V^*(s) = \max_{a \in \mathcal{A}_\mathcal{S}} \min_{\hat{\mathbb{P}}_t \in \mathcal{P}_t^a} \left\{ \hat{R}_t(s,a) + \sum_{\tau} \sum_{s'} \gamma^\tau \hat{\mathbb{P}}_t(s', \tau | s, a) V^*(s') \right\}, \tag{4.4}$$

where $0 < \gamma < 1$ is the discount factor, to obtain a robust policy $\pi^* = \text{argmax}_{a \in \mathcal{A}_\mathcal{S}} V(s)$, which optimizes the worst-case performance through minimization with respect to the uncertainty set $\mathcal{P}_t^a$. In the next section, we show that the learned policy through robust adaptive SMDP converges to an optimal policy for the uncertainty-free formulation.

4.2 Convergence of Robust Adaptive SMDP

We study the convergence properties of the robust adaptive SMDP under the following assumptions.

(A1) State space $\mathcal{S}$ and actiona space $\mathcal{A}_\mathcal{S}$ are finite.

(A2) $\hat{\mathbb{P}}_t$ and $\hat{R}_t$ remains bounded for any t.

(A3) $\hat{\mathbb{P}}_t$ and $\hat{R}_t$ converges to their true values $\mathbb{P}$ and R, respectively, with probability 1.

(A4) The uncertainty set $\mathcal{P}_t^a$ converges to a singleton estimate $\mathbb{P}$ with probability 1.

(A5) Each admissible action is executed from every state infinitely often.

Since agent's service time distribution is independent of the queue length, assumption (A5) can be relaxed to executing each action at every cognitive state. Furthermore, assumptions (A3) and (A5) can be satisfied by adopting exploration strategies such as Gibbs/Boltzman distribution method for action selection [31], wherein an action is selected with probability proportional to the current estimate of the state-action value function divided by a temperature parameter. The temperature parameter is annealed to ensure that the action selection rule converges over time to a greedy policy with respect to the value function estimate, while ensuring each cognitive state-action pair is selected infinitely often.

Let $T : \mathbb{R}^{|\mathcal{S}|} \mapsto \mathbb{R}^{|\mathcal{S}|}$ be the Bellman operator for an uncertainty-free SMDP Γ defined by:

$$T(V(s)) = \max_{a \in \mathcal{A}_{\mathcal{S}}} \left\{ R(s,a) + \sum_{\tau} \sum_{s'} \gamma^{\tau} \mathbb{P}(s', \tau | s, a) V(s') \right\}. \tag{4.5}$$

We first show the convergence of the adaptive asynchronous VI method followed by the convergence of the robust approach. In asynchronous VI, an effective approach to deal with large state spaces, the value of only a subset of states $s \in B_t \subseteq \mathcal{S}$ are updated at any time t. The adaptive VI method adapts to the latest estimates of the model to compute the control policy, which is continuously improved through online observations. Hence, the adaptive asynchronous VI update at time t is given by:

$$V_{t+1}(s) = \begin{cases} \hat{T}_t(V_t), & \text{if } s \in B_t, \\ V_t(s), & \text{otherwise,} \end{cases} \tag{4.6}$$

where $B_t \subseteq \mathcal{S}$ is the subset of states that are updated at time t, and $\hat{T}_t$ is the Bellman operator for the SMDP estimate $\hat{\Gamma}_t$ at time t, that utilizes the estimates $\hat{\mathbb{P}}_t$ and $\hat{R}_t$ in (4.5). The set B_t can be chosen using prioritized sweeping [98] for improved computational performance.

Theorem 2. *Under Assumptions A1-A5, the adaptive asynchronous VI converges to the optimal value function V^* for the uncertainty-free SMDP Γ with probability* 1.

Proof. See Appendix B: Chapter 4 for the proof. □

Let $\mathcal{P}^a$ be the uncertainty set for the probability $\mathbb{P}(s', \tau | s, a)$ for a given action $a \in \mathcal{A}_{\mathcal{S}}$. Then, the robust adaptive asynchronous VI update at time t is given by:

$$V_{t+1}(s) = \begin{cases} \max_{a \in \mathcal{A}_{\mathcal{S}}} \min_{\mathbb{P} \in \mathcal{P}^a} \left\{ \mathcal{J}^a(V_t) \right\}, & \text{if } s \in B_t, \\ V_t(s), & \text{otherwise,} \end{cases} \tag{4.7}$$

where for a given $a \in \mathcal{A}_{\mathcal{S}}$, $\mathcal{J}^a : \mathbb{R}^{|\mathcal{S}|} \mapsto \mathbb{R}^{|\mathcal{S}|}$ is given by

$$\mathcal{J}^a(V(s)) = R(s,a) + \sum_{\tau} \sum_{s'} \gamma^{\tau} \mathbb{P}(s', \tau | s, a) V(s'). \tag{4.8}$$

Let $T_r : \mathbb{R}^{|\mathcal{S}|} \mapsto \mathbb{R}^{|\mathcal{S}|}$ be the robust Bellman operator for SMDP Γ defined by:

$$T_r(V(s)) = \max_{a \in \mathcal{A}_{\mathcal{S}}} \min_{\mathbb{P} \in \mathcal{P}^a} \{\mathcal{J}^a(V(s))\}. \tag{4.9}$$

Theorem 3. *Under Assumptions A1-A5, the robust adaptive asynchronous VI converges to the optimal value function V^* for the uncertainty-free SMDP Γ with probability 1. In addition, for $\Psi_t = \frac{6\alpha}{|\mathcal{S}||\mathcal{A}_{\mathcal{S}}|\pi^2 t^2}$, the union bounds applied over each state-action pair and time yield that at any time, the obtained policy is robust with respect to uncertainty in service time distributions with at least probability α.*

Proof. See Appendix B: Chapter 4 for the proof. □

4.3 Numerical Illustrations

Fig. 4.2a and 4.2b shows an optimal policy and the optimal value function obtained using VI algorithm for the uncertainty-free SMDP Γ with $\text{cog}^* = 0.6$ as the optimal cognitive state. The optimal policy selects high fidelity around the cog^* for small queue lengths, and then transitions to normal fidelity as the queue length increases in sub-optimal cognitive states. The optimal action is to rest when the queue length is large and cognitive state is high. Similarly, skipping of tasks is an optimal action for large queue lengths in sub-optimal cognitive states. The corresponding optimal value function is a decreasing function of q and is a uni-modal function of the cognitive state, the maximum for which occurs at cog^* for each q; see [41] for more details.

Fig. 4.2c and 4.2d shows a robust adaptive policy, and the corresponding optimal value function for the uncertain SMDP Γ^{RA} obtained from asynchronous VI algorithm. In our numerical illustrations, we only estimate the parameter μ of the Lognormal(μ, σ^2) distribution using Bayesian estimation. In case when σ^2 is unknown, Markov chain Monte Carlo (MCMC) methods [99] can be used to sample from the posterior inverse-chi-squared-distribution to create the uncertainty set $\mathcal{P}_t^a$ at time t. An identical policy and value function is obtained in case of synchronous VI algorithm. Hence, the solution of the synchronous and asynchronous VI converges to the optimal solution of uncertainty-free SMDP.

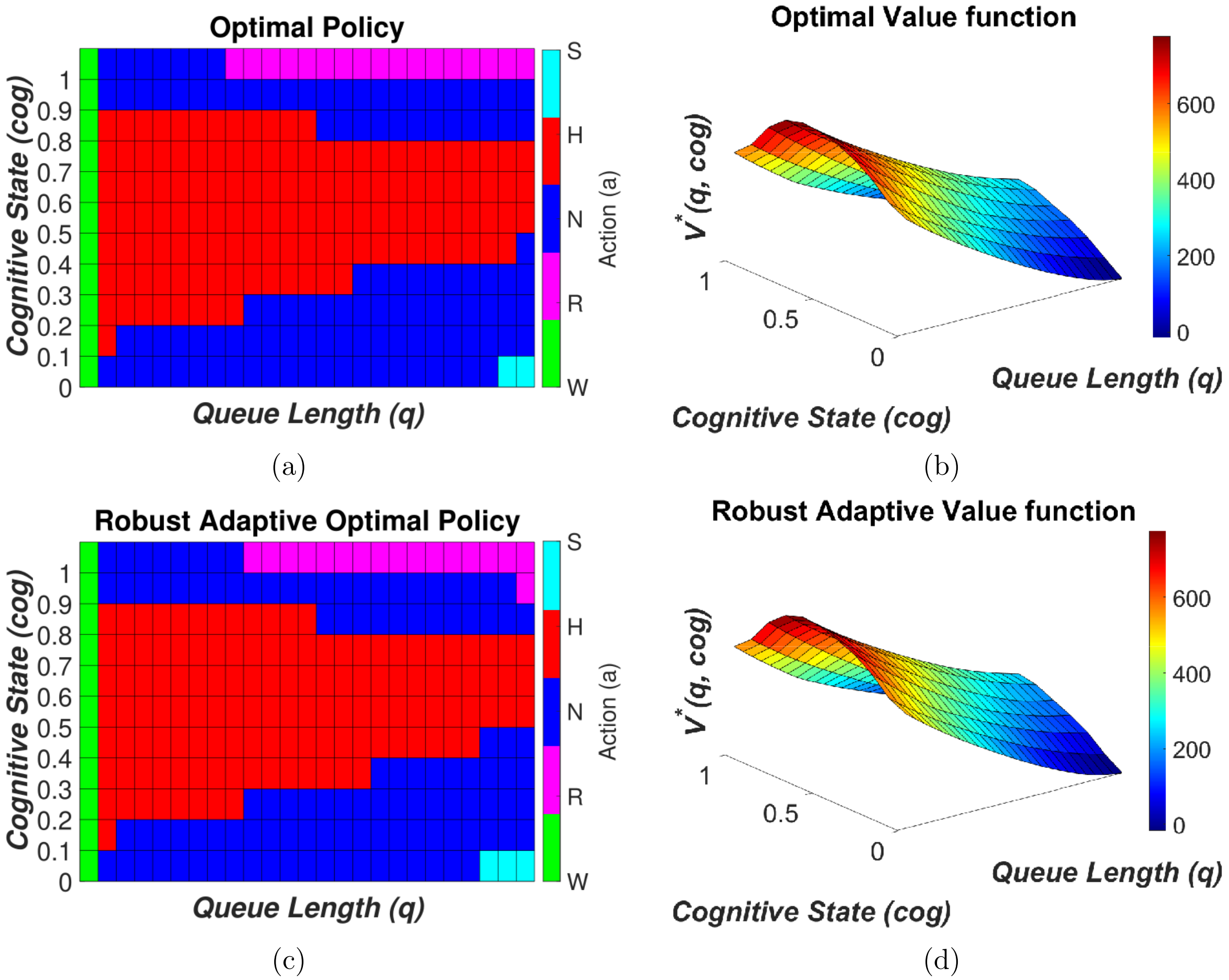


Figure 4.2 (a) Optimal policy and (b) optimal value function for the uncertainty-free SMDP Γ. (c) Robust adaptive optimal policy, and corresponding (d) optimal value function for the uncertain SMDP Γ^{RA}.

Fig. 4.3 shows the policy updates for the synchronous and asynchronous robust adaptive algorithm after 4, 8, 32 and 76 iterations, respectively. The synchronous robust adaptive algorithm performs a VI update at each time step in (4.6) for all states, i.e. $B_t = \mathcal{S}$, while in the asynchronous robust adaptive algorithm, the VI update is performed on a randomly chosen subset $B_t \in S$ at each time step. The asynchronous robust adaptive algorithm converges (80 iterations) much faster than the synchronous robust adaptive algorithm (404 iterations), while both eventually converge to the optimal policy for the uncertainty-free SMDP Γ as shown in Fig. 4.2.

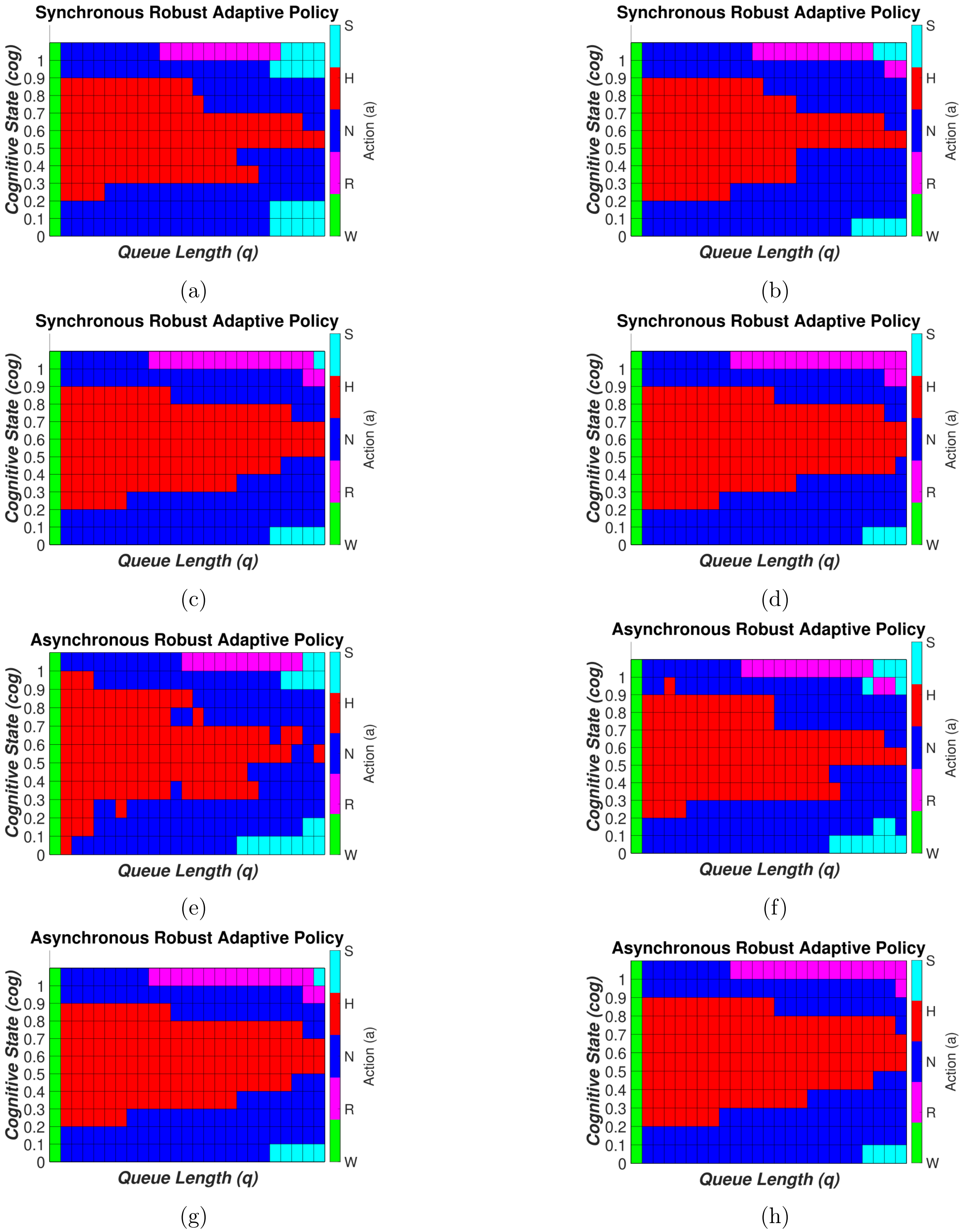


Figure 4.3 Policy updates for the synchronous ((a)-(d)) and asynchronous ((e)-(h)) robust adaptive algorithm after 4, 8, 32 and 76 iterations, respectively.

4.4 Conclusions

We studied the optimal fidelity selection problem for a human agent servicing a stream of homogeneous tasks. The parameters of the service time distribution for the agent are unknown a priori which are learned online through Bayesian parametric estimation. We utilize the robust-adaptive SMDP approach which adapts to the latest estimates of the distribution model, while obtaining robust policy towards the worst-case performance. We formally extend the convergence results of the robust adaptive MDP to robust adaptive SMDP, and show that the solution of the robust adaptive SMDP converges to the optimal solution for the uncertainty-free SMDP. Furthermore, we numerically illustrate the convergence of the synchronous and asynchronous robust adaptive policy to the uncertainty-free optimal policy.

CHAPTER 5

DETERMINISTIC SEQUENCING OF EXPLORATION AND EXPLOITATION FOR REINFORCEMENT LEARNING

In this chapter, we propose a DSEE algorithm with interleaving exploration and exploitation epochs for model-based RL problems that aim to simultaneously learn the system model, i.e., an MDP, and the associated optimal policy. During exploration, DSEE explores the environment and updates the estimates for expected reward and transition probabilities. During exploitation, the latest estimates of the expected reward and transition probabilities are used to obtain a robust policy with high probability. We design the lengths of the exploration and exploitation epochs such that the cumulative regret grows as a sub-linear function of time.

5.1 Background and Problem Formulation

We focus on the model-based RL problems which aim to simultaneously learn the system model, i.e., a Markov decision process (MDP), and the associated optimal policy. We seek to design policies that are robust to uncertainty in the learned MDP. However, learning the MDP requires visiting parts of the state space associated with high uncertainty in estimates and has exactly the opposite effect of a robust policy. Therefore, to balance the trade-off between learning the MDP and designing a robust policy, we design a DSEE algorithm, in which exploration and exploitation epochs of increasing lengths are interleaved. In exploration epochs, the algorithm learns the MDP, while in exploitation epochs, it uses a robust policy based on the learned MDP and the associated uncertainty.

Consider an MDP $(\mathcal{S}; \mathcal{A}, R, \mathbb{P}, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, reward $R(s, a)$, for each $(s, a) \in \mathcal{S} \times \mathcal{A}$, is a random variable with support $[0, R_{\max}]$, $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta_{|\mathcal{S}|}$ is the transition distribution, and $\gamma \in (0, 1)$ is the discount factor. Here, $\Delta_{|\mathcal{S}|}$ represents probability simplex in $\mathbb{R}^{|\mathcal{S}|}$, $|\cdot|$ represents the cardinality of a set. Let $\overline{R}(s, a)$ be the expected value of $R(s, a)$. We consider a finite MDP setting in which $|\mathcal{S}|$ and $|\mathcal{A}|$ are finite.

We assume that the rewards R and the state transition distribution $\mathbb{P}$ are unknown a priori. Hence, during exploration, we estimate $\overline{R}$ and $\mathbb{P}$ using online observations. Let (s, a) be any state-action pair where $s \in \mathcal{S}$ and $a \in \mathcal{A}$. At any time t, let $n_t(s, a)$ be the number of times state-action pair (s, a) is observed until time t. For each (s, a), the empirical mean estimates $\hat{R}_t(s, a)$ and $\hat{\mathbb{P}}_t(s'|s, a)$, $s' \in \mathcal{S}$ are:

$$\hat{R}_t(s,a) = \frac{1}{n_t(s,a)} \sum_{i=1}^{n_t(s,a)} r_i(s,a), \text{and} \tag{5.1}$$

$$\hat{\mathbb{P}}_t(s'|s,a) = \frac{n_t(s,a,s')}{n_t(s,a)}, \tag{5.2}$$

respectively, where $r_i(s, a)$ is the immediate reward obtained in (s, a) during observation $i \in \{1, \dots, n_t(s, a)\}$ until time t and $n_t(s, a, s')$ is the number of times the next state s' is observed from (s, a) out of $n_t(s, a)$ times.

Oftentimes, the uncertainty in probability transition matrices and mean reward function can be large, especially in the initial stages of learning due to limited observation data, which may lead to sub-optimal policies. Robust MDPs [29] mitigate the sub-optimal performance arising from this uncertainty by optimizing the worst-case performance over given uncertainty sets for reward function and probability transition matrices to obtain a robust policy. Given, at time t, uncertainty sets $\mathcal{R}_t^U$ and $\mathcal{P}_t^U$ containing $\overline{R}$ and $\mathbb{P}$, respectively, the robust MDP solves the following robust Bellman equation:

$$V_t^R(s) = \max_{a \in \mathcal{A}} \min_{\tilde{R}_t \in \mathcal{R}_t^U,\ \tilde{\mathbb{P}}_t \in \mathcal{P}_t^U} \Big\{ \tilde{R}_t(s,a) + \gamma \sum_{s'} \tilde{\mathbb{P}}_t(s'|s,a) V_t^R(s') \Big\}, \tag{5.3}$$

to obtain a robust policy $\hat{\pi}_t^R = \text{argmax}_{a \in \mathcal{A}} V_t^R$, which optimizes the worst-case performance through minimization with respect to the uncertainty sets $\mathcal{R}_t^U$ and $\mathcal{P}_t^U$, where V_t^R is the robust value function.

The choice of these uncertainty sets are critical for the performance of the robust algorithm. A poor modeling choice can increase the computational complexity and result in a highly conservative policy [17,30]. To avoid these issues, during the exploitation epoch of the DSEE, we construct these uncertainty sets based on the estimates $\hat{R}_t$ and $\hat{\mathbb{P}}_t$ from the previous exploration epochs and Hoeffding bounds [100] for $\hat{R}_t$ (Lemma 6) and $\hat{\mathbb{P}}_t$ (Lemma 7).

Subsequently, we utilize robust MDP to learn a policy that is robust to the estimation uncertainties with high probability. The convergence of the robust MDP with uncertain transition matrices to the uncertainty-free MDP can be shown under the assumption that the uncertainty sets converge to singleton estimates almost surely [17,32].

Definition 1 (***Instantaneous and Cumulative Regret***)**.** *For a discounted and ergodic RL [21], consider an algorithm $\mathbb{A}$ that, at the end of the $(t-1)$-th step, returns a policy π_t to be applied in the t-th step. For any state $s \in \mathcal{S}$, let $V^*(s)$ and $V^{\pi_t}(s)$ be the optimal value of the state and its value under the policy π_t, respectively. At any time t, the instantaneous regret $\mathfrak{R}(t)$ of the algorithm $\mathbb{A}$ is given by:*

$$\mathfrak{R}(t) = \|V^*(s) - V^{\pi_t}(s)\|_\infty, \tag{5.4}$$

where $\|\cdot\|_\infty$ denotes the L^∞-norm of a vector, and the cumulative regret $\mathbf{R}_T$ until time horizon T is given by:

$$\mathbf{R}_T = \sum_{t=1}^{T} \mathfrak{R}(t) = \sum_{t=1}^{T} \|V^*(s) - V^{\pi_t}(s)\|_\infty. \tag{5.5}$$

We design the exploration and exploitation epochs of the DSEE algorithm such that its cumulative regret grows as a sub-linear function of time. In the next section, we provide an overview of the DSEE algorithm.

5.2 DSEE Algorithm

We design the DSEE algorithm for model-based RL under the following assumptions:

(A1) State space $\mathcal{S}$ and action space $\mathcal{A}$ are finite sets.

(A2) The MDP is ergodic under the uniform policy π, i.e., under a policy π that, in every state s, randomly selects the actions from $\mathcal{A}$ with equal probability, the MDP admits a unique stationary distribution $\phi_\pi(s) : \mathcal{S} \to \Delta_{|\mathcal{S}|}$, with $\phi_\pi(s) > 0$ for all s.

Ergodic MDP [21] (assumption (A2)) is a common assumption. It ensures that the stationary distribution is independent of the initial distribution and all states are recurrent,

```
**Input:** Set of states $\mathcal{S}$, Set of actions $\mathcal{A}$, Initial State $s_0$;
**Set:** $\eta > 1$, Sequences $\{\epsilon_j\}_{j\in\mathbb{N}}$, $\{\delta_j\}_{j\in\mathbb{N}}$, $s_0^{\text{end}} = s_0$, $s = s_0$;
**Set:** $t = 0$, $n(s,a) = 0, n(s,a,s') = 0, \mathrm{S}(s,a) = 0, \forall s, a, s'$;
for epoch $j = 1, 2, \ldots$ do
        % Exploration phase:
    $\rho_j \leftarrow \frac{\epsilon_j}{4+\frac{2R_{\max}\gamma}{(1-\gamma)^2}}$;
    $\mu \leftarrow 2\left[\log(2^{|\mathcal{S}|}-2) + \log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta_j}\right)\right]$;
    $U_j \leftarrow \max\left\{\frac{(R_{\max})^2\log(\frac{4|\mathcal{S}||\mathcal{A}|}{\delta_j})}{2\rho_j^2}, \frac{\mu}{\rho_j^2}\right\}$
    while $n(s,a) < U_j, \ \forall(s,a)$
        $t \leftarrow t+1$;
        Pick $a \sim \mathrm{UNIF}(\mathcal{A})$ in current state $s$ do
        Observe reward R and the next state $s'$
        if $s_{j-1}^{\text{end}}$ has been visited in epoch $j$ then
            $n(s,a) \leftarrow n(s,a)+1$;
            $n(s,a,s') \leftarrow n(s,a,s')+1$;
            $\mathrm{S}(s,a) = \mathrm{S}(s,a) + \mathrm{R}$;
        end if
        $s \leftarrow s'$;
    end while
    $s_j^{\text{end}} \leftarrow s$;
    $\hat{R}_t(s,a) = \frac{\mathrm{S}(s,a)}{n(s,a)}, \ \forall(s,a)$;
    $\hat{\mathbb{P}}_t(s'|s,a) = \frac{n(s,a,s')}{n(s,a)}, \ \forall(s,a)$;
        % Exploitation phase:
    Construct uncertainty sets $\mathcal{R}_t^U$ and $\mathcal{P}_t^U$ using (5.11);
    Compute $V_t^R(s)$ and $\hat{\pi}_t^R$ using (5.3);
    Implement $\hat{\pi}_t^R$ for $\lceil\eta^j\rceil$ time steps;
    $t \leftarrow t + \lceil\eta^j\rceil$;
end for
```

Algorithm 5.1 Deterministic Sequencing of Exploration and Exploitation (DSEE).

i.e., each state s is visited infinitely often and $\phi_\pi(s) > 0$. We use this assumption to estimate the number of times each state is visited in N time steps.

Algorithm 5.1 shows an overview of the DSEE algorithm. In the DSEE algorithm, we design a sequence of alternating exploration and exploitation epochs. Let α_i and β_i be the lengths of the i-th exploration and exploitation epoch, respectively, where $i \in \mathbb{N}$. During an exploration epoch, we uniformly sample the action in the current state and update the estimates $\hat{R}_t$ and $\hat{\mathbb{P}}_t$. For a given sequence of $\{\epsilon_i\}_{i\in\mathbb{N}}$ and $\{\delta_i\}_{i\in\mathbb{N}}$ that we design in Section 5.3, the length of the exploration epoch α_i is determined to reduce the estimation uncertainty such that $\mathbf{P}(\|V^*(s) - V^{\hat{\pi}_t^R}(s)\|_\infty \leq \epsilon_i) \geq 1 - \delta_i$ after the epoch, where $\mathbf{P}(\cdot)$ denotes the probability measure, and $V^{\hat{\pi}_t^R}(s)$ is the value of state s under the robust policy $\hat{\pi}_t^R$. In DSEE, we choose exponentially increasing lengths of the exploitation epochs β_i. During the exploitation epoch, we utilize the estimates $\hat{R}_t$ and $\hat{\mathbb{P}}_t$ from previous exploration epochs and construct the uncertainty sets $\mathcal{R}^U$ and $\mathcal{P}^U$ at time t. We use these uncertainty sets with a robust Bellman equation to learn a policy that is robust to the estimation uncertainties with high probability. In next section, we analyze the DSEE algorithm, and design the sequence of $\{\epsilon_i\}_{i\in\mathbb{N}}$ and $\{\delta_i\}_{i\in\mathbb{N}}$, such that the cumulative regret (5.5) grows as a sub-linear function of time.

5.3 Analysis of DSEE algorithm

We now characterize the regret of the DSEE algorithm under the assumptions (A1-A2) and design the exploration and exploitation epochs. The optimal value $V^*(s_t)$ of the state s_t is given by:

$$V^*(s_t) = \overline{R}(s_t, \pi^*(s_t)) + \gamma \mathbb{E}[V^*(s_{t+1})|s_t, \pi^*(s_t)], \tag{5.6}$$

where π^* is an optimal policy that satisfies:

$$\pi^*(s_t) = \operatorname*{argmax}_{a_t} \left\{ \overline{R}(s_t, a_t) + \gamma \mathbb{E}[V^*(s_{t+1})|s_t, a_t] \right\}. \tag{5.7}$$

We define an approximate optimal value function $\hat{V}_t$ that utilizes the estimates $\hat{R}_t$ and $\hat{\mathbb{P}}_t$ at time t. Therefore, $\hat{V}_t(s_t)$ is given by:

$$\hat{V}_t(s_t) = \hat{R}_t(s_t, \hat{\pi}_t(s_t)) + \gamma \hat{\mathbb{E}}\Big[\hat{V}_t(s_{t+1}) | s_t, \hat{\pi}_t(s_t)\Big], \tag{5.8}$$

where $\hat{\mathbb{E}}\Big[\hat{V}_t(s_{t+1})|s_t, \hat{\pi}_t(s_t)\Big]$ is used to denote $\sum_{s_{t+1}} \hat{\mathbb{P}}_t(s_{t+1}|s_t, \hat{\pi}_t(s_t))\hat{V}_t(s_{t+1})$ and $\hat{\pi}_t$ is an optimal policy for the approximate optimal value function given by:

$$\hat{\pi}_t(s_t) = \underset{a_t}{\operatorname{argmax}} \left\{ \hat{R}_t(s_t, a_t) + \gamma \hat{\mathbb{E}}\Big[\hat{V}_t(s_{t+1})|s_t, a_t\Big] \right\}. \tag{5.9}$$

Theorem 4 (***Concentration of robust value function***)**.** *Let $\|\cdot\|_1$ denote the L^1-norm of a vector. For any given $\epsilon_t, \delta_t \in (0,1)$, there exists an $n \in O\left(\frac{|\mathcal{S}|}{\epsilon_t^2} + \frac{1}{\epsilon_t^2}\log\left(\frac{|\mathcal{S}||\mathcal{A}|}{\delta_t}\right)\right)$ such that if each state-action pair (s,a) is observed $n_t(s,a) \geq n$ times until time t, then for each state s, the following inequality holds:*

$$\mathbf{P}\left(\|V^*(s) - V^{\hat{\pi}_t^R}(s)\|_\infty \leq \epsilon_t\right) \geq 1 - \delta_t, \tag{5.10}$$

where $V^{\hat{\pi}_t^R}(s)$ is the value of state s under the robust policy $\hat{\pi}_t^R = \operatorname{argmax}_{a \in \mathcal{A}} V_t^R$. The robust value function V_t^R is defined in (5.3) with $\rho_t = \frac{\epsilon_t}{2}\left(2 + \frac{R_{\max}\gamma}{(1-\gamma)^2}\right)^{-1}$ and

$$\begin{aligned} \mathcal{R}_t^U &= \left\{ R^U(s,a) : |R^U(s,a) - \hat{R}_t(s,a)| \leq \rho_t, \ \forall (s,a) \right\}, \\ \mathcal{P}_t^U &= \left\{ \mathbb{P}^U(s,a) : \left\| \mathbb{P}^U(s,a) - \hat{\mathbb{P}}_t(s,a) \right\|_1 \leq \rho_t, \ \forall (s,a) \right\}. \end{aligned} \tag{5.11}$$

We prove Theorem 4 using the following Lemmas 6-9.

Lemma 6 (***Concentration of rewards***)**.** *Suppose until time step t, the state-action pair (s,a) is observed $n_t(s,a)$ times and bounded immediate rewards $r_i(s,a)$, $i \in \{1, \ldots, n_t(s,a)\}$, are obtained at these instances. Then the following inequality holds:*

$$\mathbf{P}\left(\left|\overline{R}(s,a) - \hat{R}_t(s,a)\right| \leq \epsilon_t^R\right) \geq 1 - \delta_t^R, \tag{5.12}$$

where $\hat{R}_t(s,a)$ is the empirical mean reward defined in (5.1) and $\epsilon_t^R = \sqrt{\frac{(R_{\max})^2 \log(2/\delta_t^R)}{2n_t(s,a)}}$.

Proof. For brevity of notation, let ϵ_R and δ_R denote ϵ_t^R and δ_t^R, respectively. For bounded random variables $r_i(s,a)$, using the Hoeffding bounds [100], we have

$$\mathbf{P}\left(\left|\overline{R}(s,a) - \hat{R}_t(s,a)\right| \leq \epsilon_R\right) \geq 1 - 2e^{-\frac{2n_t(s,a)\epsilon_R^2}{(R_{\max})^2}}. \tag{5.13}$$

Choosing $\delta_R = 2e^{-\frac{2n_t(s,a)\epsilon_R^2}{(R_{\max})^2}}$, we get the desired result. □

Lemma 7 (***Concentration of transition probabilities***)**.** *Suppose until time step t, the state-action pair (s,a) is observed $n_t(s,a)$ times and let $\mathbb{P}(s,a) \in \Delta_{|\mathcal{S}|}$ be the true transition probability distribution for (s,a). Then for any (s,a), the following inequality holds:*

$$\mathbf{P}\left(\left\|\mathbb{P}(s,a) - \hat{\mathbb{P}}_t(s,a)\right\|_1 \leq \epsilon_t^P\right) \geq 1 - \delta_t^P, \tag{5.14}$$

where $\|\cdot\|_1$ is the L^1 norm of a vector and $\hat{\mathbb{P}}_t(s,a)$ is the empirical transition probability vector with components $\hat{\mathbb{P}}_t(s'|s,a)$ defined in (5.2)*, and $\epsilon_t^P = \sqrt{\frac{2[\log(2^{|\mathcal{S}|}-2)-\log(\delta_t^P)]}{n_t(s,a)}}$.*

Proof. For brevity of notation, let ϵ_P and δ_P denote ϵ_t^P and δ_t^P, respectively. Using [101, Theorem 2.1], we have:

$$\mathbf{P}\left(\left\|\mathbb{P}(s,a) - \hat{\mathbb{P}}_t(s,a)\right\|_1 \leq \epsilon_P\right) \geq 1 - (2^{|S|} - 2)e^{-\frac{n_t(s,a)\epsilon_P^2}{2}}. \tag{5.15}$$

Setting $\delta_P = (2^{|S|} - 2)e^{-\frac{n_t(s,a)\epsilon_P^2}{2}}$, yields the desired result. □

Lemmas 6 and 7 provide concentration bounds on the reward and transition probability based on how often a state-action pair is visited.

Lemma 8. ***(Concentration of reward and transition probability functions)*** *Let $\delta_t^R = \delta_t^P = \frac{\delta_t}{2|\mathcal{S}||\mathcal{A}|}$. Then, for any $\rho_t > 0$, there exists an $n \in O\left(\frac{|\mathcal{S}|}{\rho_t^2} + \frac{1}{\rho_t^2}\log\left(\frac{|\mathcal{S}||\mathcal{A}|}{\delta_t}\right)\right)$ such that when each state-action pair (s,a) is observed $n_t(s,a) \geq n$ times, then the following statements hold for any (s,a):*

(i) $\mathbf{P}\left(|\overline{R}(s,a) - \hat{R}_t(s,a)| \leq \rho_t\right) \geq 1 - \frac{\delta_t}{2|\mathcal{S}||\mathcal{A}|}$,

(ii) $\mathbf{P}\left(\left\|\mathbb{P}(s,a) - \hat{\mathbb{P}}_t(s,a)\right\|_1 \leq \rho_t\right) \geq 1 - \frac{\delta_t}{2|\mathcal{S}||\mathcal{A}|}$.

Proof. Using Lemmas 6 and 7, we know that $|\overline{R}(s,a) - \hat{R}_t(s,a)| \leq \rho_t$ and $\left\|\mathbb{P}(s,a) - \hat{\mathbb{P}}_t(s,a)\right\|_1 \leq \rho_t$ holds for any (s,a) with at least probability $1 - \frac{\delta_t}{2|\mathcal{S}||\mathcal{A}|}$ for $\rho_t \geq \sqrt{\frac{(R_{\max})^2 \log(\frac{4|\mathcal{S}||\mathcal{A}|}{\delta_t})}{2n_t(s,a)}}$ and $\rho_t \geq \sqrt{\frac{2\left[\log(2^{|\mathcal{S}|}-2) - \log\left(\frac{\delta_t}{2|\mathcal{S}||\mathcal{A}|}\right)\right]}{n_t(s,a)}}$, respectively. Hence, we have that

$$n_t(s,a) \geq \max\left\{\frac{(R_{\max})^2 \log(\frac{4|\mathcal{S}||\mathcal{A}|}{\delta_t})}{2\rho_t^2}, \frac{\mu}{\rho_t^2}\right\}, \tag{5.16}$$

where $\mu = 2\left[\log(2^{|\mathcal{S}|} - 2) + \log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta_t}\right)\right]$, is sufficient to guarantee that $|\overline{R}(s,a) - \hat{R}_t(s,a)| \leq \rho_t$ and $\left\|\mathbb{P}(s,a) - \hat{\mathbb{P}}_t(s,a)\right\|_1 \leq \rho_t$ holds for any (s,a) with at least probability $1 - \frac{\delta_t}{2|\mathcal{S}||\mathcal{A}|}$. Hence, we can choose $n \in O\left(\frac{|\mathcal{S}|}{\rho_t^2} + \frac{1}{\rho_t^2}\log\left(\frac{|\mathcal{S}||\mathcal{A}|}{\delta_t}\right)\right)$ such that (5.16) holds, and hence, the lemma follows. □

Remark 6 (***Concentration inequalities***)**.** *The concentration inequalities in Lemmas 6, 7, and 8 use a deterministic value of $n_t(s,a)$. However, these bound also apply if $n_t(s,a)$ is a realization of a random process that is independent of $\hat{R}_t(s,a)$ and $\hat{\mathbb{P}}_t(s,a)$, which would be the case in this paper.*

Remark 7 (***Uncertainty set***)**.** *Using union bounds over all (s,a) and Lemma 8, it follows that uncertainty sets $\mathcal{R}_t^U$ and $\mathcal{P}_t^U$ stated in Theorem 4 are ρ_t-level uncertainty sets for $R(s,a)$ and $\mathbb{P}(s,a)$, respectively, for each $(s,a) \in \mathcal{S} \times \mathcal{A}$ with at least probability $1 - \delta_t$. Thus, the policy obtained at time t in an exploitation epoch is robust to estimation uncertainties with at least probability $1 - \delta_t$.*

Lemma 9 (***Loss in robust value function***)**.** *Suppose $|\overline{R}(s,a) - \hat{R}_t(s,a)| \leq \rho_t$ and $\left\|\mathbb{P}(s,a) - \hat{\mathbb{P}}_t(s,a)\right\|_1 \leq \rho_t$, for any (s,a), at time t. Then $\mathcal{L}_t(s) = V^*(s) - V^{\hat{\pi}_t^R}(s)$ satisfies:*

$$\mathcal{L}_t(s) \leq 2\rho_t\left(2 + \frac{R_{\max}\gamma}{(1-\gamma)^2}\right), \tag{5.17}$$

where $V^{\hat{\pi}_t^R}(s)$ is the value of state s under the robust policy $\hat{\pi}_t^R$ at time t.

Proof [Sketch]: The upper bound on the loss is obtained by following a similar analysis as in [102] that provides an upper bound on the loss by considering the uncertainty in $\mathbb{P}$ only.

The analysis in [102] can be extended by considering the uncertainty in $\overline{R}$ as well. The details of the proof can be found in [103].

Lemma 9 provides the bounds on the loss in robust value function w.r.t. the optimal value function using the concentration bounds on the rewards and transition probabilities.

Proof of Theorem 1: Using Lemma 8, we know that when each state-action pair (s,a) is sampled $n \in O\left(\frac{|\mathcal{S}|}{\rho_t^2} + \frac{1}{\rho_t^2}\log\left(\frac{|\mathcal{S}||\mathcal{A}|}{\delta_t}\right)\right)$ times, then the following inequalities holds for any (s,a):

$$\mathbf{P}\left(|\overline{R}(s,a) - \hat{R}_t(s,a)| \le \rho_t\right) \ge 1 - \frac{\delta_t}{2|\mathcal{S}||\mathcal{A}|}, \tag{5.18}$$

$$\mathbf{P}\left(\left\|\mathbb{P}(s,a) - \hat{\mathbb{P}}_t(s,a)\right\|_1 \le \rho_t\right) \ge 1 - \frac{\delta_t}{2|\mathcal{S}||\mathcal{A}|}. \tag{5.19}$$

Hence, using Lemma 9 and applying union bounds, we obtain that the following holds with at least probability $1 - (\delta_t^R + \delta_t^P)|\mathcal{S}||\mathcal{A}|$

$$V^*(s) - V^{\hat{\pi}_t^R}(s) \le 2\rho_t\left(2 + \frac{R_{\max}\gamma}{(1-\gamma)^2}\right). \tag{5.20}$$

Setting $\rho_t = \frac{\epsilon_t}{2}\left(2 + \frac{R_{\max}\gamma}{(1-\gamma)^2}\right)^{-1}$ and $\delta_t^R = \delta_t^P = \frac{\delta_t}{2|\mathcal{S}||\mathcal{A}|}$,

$$\mathbf{P}\left(V^*(s) - V^{\hat{\pi}_t^R}(s) \le \epsilon_t\right) \ge 1 - \delta_t,\ \forall s \in \mathcal{S}$$
$$\implies \mathbf{P}\left(\|V^*(s) - V^{\hat{\pi}_t^R}(s)\|_\infty \le \epsilon_t\right) \ge 1 - \delta_t. \tag{5.21}$$

Additionally, the order of n in terms of ϵ_t becomes $n \in O\left(\frac{|\mathcal{S}|}{\epsilon_t^2} + \frac{1}{\epsilon_t^2}\log\left(\frac{|\mathcal{S}||\mathcal{A}|}{\delta_t}\right)\right)$. ■

In Theorem 4, we obtain the number of times n each state-action pair needs to be visited to reduce the estimation uncertainty in rewards and transition probabilities to obtain an ϵ_t-optimal policy with probability at least $1 - \delta_t$. Now we estimate the total number of exploration steps that are needed to ensure that each state-action pair is visited at least n times.

Lemma 10 (***Adapted from [104, Theorem 3]***). *For an ergodic Markov chain with state space $\mathcal{S}$ and stationary distribution ϕ_{ss}, let $\tau = \tau(\sigma)$ be the σ-mixing time*[1] *with $\sigma \le \frac{1}{8}$. Let*

[1] σ-mixing time for an ergodic Markov chain in the minimal time until the distribution of Markov chain is σ-close in total variation distance to its steady state distribution [105].

ϕ_0 be the initial distribution on $\mathcal{S}$ and let $\|\phi_0\|_{\phi_{\text{ss}}} = \sqrt{\sum_{s\in\mathcal{S}} \frac{\phi_0(s)^2}{\phi_{\text{ss}}(s)}}$. Let $n_{\text{vis}}(s_i, N)$ be the number of times state $s_i \in \mathcal{S}$ is visited until time N. Then, for any $0 \leq \kappa \leq 1$, there exists a constant $c > 0$ (independent of σ and κ) such that:

$$\mathbf{P}\left(n_{\text{vis}}(s_i, N) \geq (1-\kappa)N\phi_{\text{ss}}(s_i)\right) \geq 1 - c\|\phi_0\|_{\phi_{\text{ss}}} e^{-\frac{\kappa^2 N\phi_{\text{ss}}(s_i)}{72\tau}}. \tag{5.22}$$

Proof. See [104, Theorem 3] for the proof. □

In the exploration epoch, at any state $s_i \in \mathcal{S}$, we choose actions uniformly randomly. Consider the Markov chain on $\mathcal{S}$ that is associated with the uniform action selection policy. Let $\{\phi_0(s)\}_{s\in\mathcal{S}}$ and $\{\phi_{\text{ss}}(s)\}_{s\in\mathcal{S}}$, respectively, be the associated initial and stationary distribution. We can also consider an equivalent lifted Markov chain on $\mathcal{S} \times \mathcal{A}$ with states (s_i, a_j) such that $s_i \in \mathcal{S}$ and $a_j \in \mathcal{A}$. The lifted Markov chain has the initial and stationary distribution, $\phi_0(s, a) = \frac{\phi_0(s_i)}{|\mathcal{A}|}$ and $\phi_{\text{ss}}(s, a) = \frac{\phi_{\text{ss}}(s_i)}{|\mathcal{A}|}$, respectively. Hence, we can apply Lemma 10 to obtain the probability of visiting a state-action pair (s, a) at least $(1-\kappa)N\phi_{\text{ss}}(s, a)$ times after N time steps under the uniform action selection policy.

We now design a sequence of exploration and exploitation epochs. Let α_i and β_i be the lengths of the i-th exploration and exploitation epoch, respectively. Let $\phi_{\text{ss}}^{\min} := \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \phi_{\text{ss}}(s, a)$ and $N_i = \frac{\bar{N}_i}{(1-\kappa)\phi_{\text{ss}}^{\min}}$, where $\bar{N}_i$ is the upper bound in (5.16) associated with (ϵ_i, δ_i). Let $\delta^{\alpha_i} := c\|\phi_0\|_{\phi_{\text{ss}}} e^{-\frac{\kappa^2 N_i \phi_{\text{ss}}^{\min}}{72\tau}}$. Note that the desired values of $(1-\kappa)N\phi_{\text{ss}}(s_i, a_j)$ and δ^{α_i} can be obtained by tuning N and κ in (5.22).

Theorem 5 (***Regret bound for DSEE algorithm***)**.** *Let the length of exploitation epochs in DSEE be exponentially increasing, i.e. $\beta_i = \eta^i$, $\eta > 1$. Let $\epsilon_i = \eta^{-\frac{i}{3}}$ and $\delta_i = \eta^{-\frac{i}{3}}$ such that $\mathbf{P}(\|V^*(s) - V^{\hat{\pi}_i^R}(s)\|_\infty \leq \epsilon_i) \geq 1 - \delta_i$ after exploration epoch i. For any $\delta \in (0, 1)$, set $\delta^{\alpha_i} = \frac{6\delta}{|S||A|\pi^2 i^2}$. Then, the cumulative regret for the DSEE algorithm $\mathbf{R}_T \in O((T)^{\frac{2}{3}} \log(T))$ grows sub-linearly with time T with probability at least $1 - \delta$.*

Proof. We note that the system state at the start of the i-th exploration epoch might be different from the final state at the end of the $(i-1)$-th exploration epoch. Therefore, we remember the final state of the previous exploration epoch and wait for the same state to restart the new exploration epoch. For the ergodic MDP under the uniform action selection policy (assumption A2), we know that the expected hitting time is finite [106]. Let $U \in \mathbb{R}_{>0}$ be a constant upper bound on the expected hitting time to reach the final state in the previous exploration epoch from an arbitrary initial state in the current exploration epoch. Hence, the cumulative regret during the i-th exploration epoch of length α_i is upper-bounded by $(U + \alpha_i)\mathfrak{R}_{\max}$, where $\mathfrak{R}_{\max} = \frac{R_{\max}}{1-\gamma}$ is the maximum instantaneous regret.

Since at start of the exploitation epoch i of length β_i, $\mathbf{P}(\|V^*(s) - V^{\hat{\pi}_t^R}(s)\|_\infty \leq \epsilon_i) \geq 1 - \delta_i$, the expected cumulative regret during the exploitation epoch is $(1 - \delta_i)\beta_i\epsilon_i + \delta_i\beta_i\mathfrak{R}_{\max}$. Therefore, the total cumulative regret after k sequences of exploration and exploitation each is upper bounded by:

$$\begin{aligned}\mathbf{R}_{T_k} &\leq \sum_{i=1}^{k} \left((\alpha_i + U)\mathfrak{R}_{\max} + (1 - \delta_i)\beta_i\epsilon_i + \delta_i\beta_i\mathfrak{R}_{\max}\right) \\ &\leq \sum_{i=1}^{k} \left((\alpha_i + U)\mathfrak{R}_{\max} + \beta_i\epsilon_i + \delta_i\beta_i\mathfrak{R}_{\max}\right). \qquad (5.23)\end{aligned}$$

Let T_i be the time at the end of the i-th exploitation epoch. Then, $\sum_{j=1}^{k} \beta_j < T_k \leq \sum_{j=1}^{k}(\alpha_j + U) + \sum_{j=1}^{k} \beta_j$. We design the length of the exploitation epochs to be exponentially increasing, i.e., $\beta_i = \eta^i$, for $\eta > 1$. Thus, $T_k \in O(\sum_{j=1}^{k} \eta^j) = O(\eta^k)$. Let $\epsilon_i = \eta^{-di}$ and $\delta_i = \eta^{-gi}$, where $d \in (0, 1)$ and $g \in (0, 1)$ are constants that we design later. Thus, (5.23) can be written as:

$$\mathbf{R}_{T_k} \leq \mathfrak{R}_{\max}\left(\sum_{i=1}^{k} \alpha_i + kU\right) + \sum_{i=1}^{k} \eta^{i(1-d)} + \mathfrak{R}_{\max}\sum_{i=1}^{k} \eta^{i(1-g)}. \qquad (5.24)$$

For a state-action pair (s, a), where $s \in \mathcal{S}$ and $a \in \mathcal{A}$, let $\delta_{(s,a)}^{\alpha_i} := c\|\phi_0\|_{\phi_{\text{ss}}} e^{-\frac{\kappa^2 N_i \phi_{\text{ss}}(s,a)}{72\tau}}$. Therefore, using Lemma 10, at the end of the i-th epoch,

$$\mathbf{P}\left(n_{\text{vis}}(s, a, N_i) \geq (1 - \kappa)N_i\phi_{\text{ss}}(s, a)\right) \geq 1 - \delta_{(s,a)}^{\alpha_i}. \qquad (5.25)$$

Recall $\delta^{\alpha_i} := c\|\phi_0\|_{\phi_{\text{ss}}} e^{-\frac{\kappa^2 N_i \phi_{\text{ss}}^{\min}}{72\tau}}$, where $\phi_{\text{ss}}^{\min} := \min_{(s,a)} \phi_{\text{ss}}(s, a)$. Substituting $N_i =$

$\frac{\bar{N}_i}{(1-\kappa)\phi_{\text{ss}}^{\min}}$ in (5.25),

$$\mathbf{P}\left(n_{\text{vis}}(s,a,N_i) \geq \bar{N}_i\right) \geq 1 - \sum_{m=1}^{|\mathcal{S}||\mathcal{A}|} \delta^{\alpha_i}, \tag{5.26}$$

for each state-action pair (s,a). Therefore, in N_i time steps of the lifted Markov chain, each (s,a) is visited at least $\bar{N}_i$ times with high probability. Thus, N_i is an upper bound on $\sum_{j=1}^{i} \alpha_j$ with probability in (5.26). Therefore, using union bounds, with high probability $1 - \sum_{j=1}^{k} \sum_{m=1}^{|\mathcal{S}||\mathcal{A}|} \delta^{\alpha_j}$, $\sum_{j=1}^{k} \alpha_j \leq N_k = \frac{\bar{N}_k}{(1-\kappa)\phi_{\text{ss}}^{\min}}$, and hence,

$$\mathbf{R}_{T_k} \leq \frac{\mathfrak{R}_{\max}\bar{N}_k}{(1-\kappa)\phi_{\text{ss}}^{\min}} + kU\mathfrak{R}_{\max} + \sum_{i=1}^{k} \eta^{i(1-d)} + \mathfrak{R}_{\max} \sum_{i=1}^{k} \eta^{i(1-g)}. \tag{5.27}$$

Using Theorem 4, $\bar{N}_k \in O\left(\frac{|\mathcal{S}|}{\epsilon_k^2} + \frac{1}{\epsilon_k^2} \log\left(\frac{|\mathcal{S}||\mathcal{A}|}{\delta_k}\right)\right)$. Therefore,

$$\begin{aligned}
\mathbf{R}_{T_k} &\leq \frac{\mathfrak{R}_{\max}\lambda}{(1-\kappa)\phi_{\text{ss}}^{\min}} \left(\frac{|\mathcal{S}|}{\epsilon_k^2} + \frac{1}{\epsilon_k^2} \log\left(\frac{|\mathcal{S}||\mathcal{A}|}{\delta_k}\right)\right) + kU\mathfrak{R}_{\max} + \sum_{i=1}^{k} \eta^{i(1-d)} + \mathfrak{R}_{\max} \sum_{i=1}^{k} \eta^{i(1-g)} \\
&\leq \frac{\mathfrak{R}_{\max}\lambda}{(1-\kappa)\phi_{\text{ss}}^{\min}} \left(\eta^{2dk}|\mathcal{S}| + \eta^{2dk} \log\left(\eta^{gk}|\mathcal{S}||\mathcal{A}|\right)\right) + kU\mathfrak{R}_{\max} + \sum_{i=1}^{k} \eta^{i(1-d)} + \mathfrak{R}_{\max} \sum_{i=1}^{k} \eta^{i(1-g)},
\end{aligned} \tag{5.28}$$

for some constant λ. Recall that $T_k \in O(\eta^k)$, which implies $k \in O(\log(T_k))$. Let Z be the right-hand side of (5.28). Then, we have:

$$\begin{aligned}
Z &\in O\left((T_k)^{2d} + (T_k)^{2d}\log(T_k) + (T_k)^{(1-d)} + (T_k)^{(1-g)}\right) \\
&\in O((T_k)^{\frac{2}{3}} \log(T_k)),
\end{aligned} \tag{5.29}$$

by choosing $d = g = \frac{1}{3}$. Hence, the cumulative regret $\mathbf{R}_{T_k} \in O((T_k)^{\frac{2}{3}} \log(T_k))$ grows sublinearly with time T_k with probability at least $1 - \sum_{i=1}^{k} \sum_{m=1}^{|\mathcal{S}||\mathcal{A}|} \delta^{\alpha_i}$.

Setting $\delta^{\alpha_i} = \frac{6\delta}{|S||A|\pi^2 i^2}$, we have $\sum_{i=1}^{k} \sum_{m=1}^{|\mathcal{S}||\mathcal{A}|} \delta^{\alpha_i} \leq \delta$. □

5.4 Conclusions

We proposed a DSEE algorithm with interleaving exploration and exploitation epochs for model-based RL problems that aims to simultaneously learn the system model, i.e., an MDP, and the associated optimal policy. During exploration, we uniformly sample the action in each state and update the estimates of the mean rewards and transition probabilities. These estimates are used in the exploitation epoch to obtain a robust policy with high probability. We designed the length of the exploration and exploitation epochs such that the cumulative regret grows as a sub-linear function of time.

CHAPTER 6

FOSTERING HUMAN LEARNING IN SEQUENTIAL DECISION-MAKING: UNDERSTANDING THE ROLE OF EVALUATIVE FEEDBACK

In this chapter, we investigate the role of feedback in fostering human learning in sequential decision-making tasks. Cognitive rehabilitation, STEM skill acquisition, and coaching games such as chess often require tutoring decision-making strategies. The advancement of AI-driven tutoring systems for facilitating human learning requires an understanding of the impact of evaluative feedback on human decision-making and skill development. To this end, we conduct human experiments using Amazon Mechanical Turk to study the influence of evaluative feedback on human decision-making in sequential tasks. In these experiments, participants solve the Tower of Hanoi puzzle and receive AI-generated feedback while solving it. We examine how this feedback affects their learning and skill transfer to related tasks. We also explore various computational models to understand how people incorporate evaluative feedback into their decision-making processes.

6.1 Background and Problem Formulation

We investigate the influence of evaluative feedback on human performance in a sequential decision-making task through experimental evaluations and computational modeling. To this end, we conducted experiments, where the participants were asked to solve the ToH puzzle. ToH is a puzzle in which disks with a priority order are placed on three pegs. The priority order determines which disk can be placed on top of another disk and each instance of admissible disk placement is referred to as a configuration. Thus, for a four-disk and a five-disk ToH, there are $3^4 = 81$ and $3^5 = 243$ possible configurations, respectively. The goal is to move one disk at a time and reach the desired configuration while maintaining the priority order at each time.

Consider the ToH puzzle with n disks, where the disks are numbered $\{0, 1, \ldots, n-1\}$ in ascending order of size, and the three pegs are numbered $\{0, 1, 2\}$ from left to right. The state of the n-disk ToH can be represented as $S_n = (s_0 s_1 \ldots s_{n-1})$, where $s_i \in \{0, 1, 2\}$ denotes the

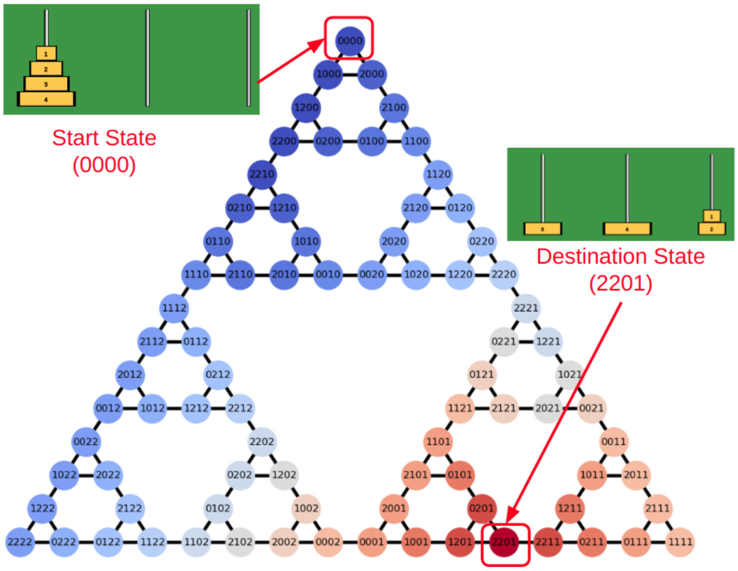


Figure 6.1 State space of a 4-disk ToH with 81 states. Each state corresponds to a unique configuration of the disks on three pegs and edges encode allowed transitions between states. The task is to reach the configuration associated with a randomly selected target state (for example 2201 in this figure). Warmer colors are associated with the higher value function (see Sec. 6.1.1 for discussion).

peg on which disk i is placed, for $0 \leq i \leq n-1$. Each state in an n-disk ToH has either two or three possible state transitions as can be seen by the state space of a 4-disk ToH shown in Fig. 6.1.

6.1.1 Evaluative Feedback

We train an RL agent that is capable of optimally solving the ToH puzzle. The ToH is a finite state space and finite action puzzle, and thus, an optimal policy can be derived using tabular RL methods as described in [31, 107]. In Figure 6.1, we demonstrate the optimal value function for the standard 4-disk ToH. To obtain the optimal value function for a given target state, we utilize the value iteration algorithm [31], where the reward function $r(s) : \mathcal{S} \rightarrow \mathbb{R}$ is designed as follows:

$$r(s) = \begin{cases} 1, & \text{if } s \text{ is the target state,} \\ 0, & \text{otherwise.} \end{cases} \tag{6.1}$$

Using the reward function in (6.1) results in an optimal value function that is proportional to the length of the shortest path for each state to the target state. The obtained optimal value

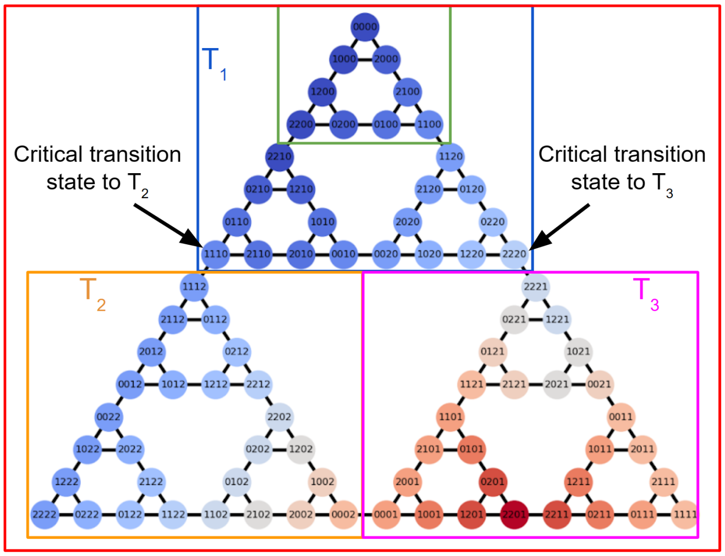


Figure 6.2 State space of a 4-disk ToH with 81 states. Each state corresponds to a unique configuration of the disks on three pegs and edges encode allowed transitions between states. The state space can be visualized as comprising three triangular structures. The states that connect different triangular structures are critical states to transition between triangles.

function is utilized to provide evaluative feedback to the human player based on the change in the value at states before and after the move. We deploy several feedback mechanisms as detailed in Section 6.2.1 and systematically explore how human decision-making is influenced by different feedback mechanisms.

The state space of the ToH problem exhibits a recursive structure. Specifically, the state space of a ToH puzzle with n disks can be effectively illustrated using three interlocking triangles. Each of these triangles symbolizes the state space of a ToH puzzle with $n-1$ disks. To illustrate this concept, let's examine the state space of a 4-disk ToH in Figure 6.2, which is highlighted in red. In the same figure, the blue and green squares are employed to represent the state spaces of 3-disk and 2-disk ToH puzzles, respectively. Hence, the state space for the ToH with $n-1$ disks can be simply achieved by removing the last digit from each state in the upper triangle of the n-disk ToH. This digit corresponds to the position of the largest disk.

As illustrated in Fig. 6.2, the state space of the 4-disk ToH puzzle can be decomposed into three triangles labeled as T_1, T_2, and T_3. Throughout the remainder of the manuscript, we will consistently refer to the regions of the state space as follows: the top triangle will be denoted as T_1, the lower left triangle as T_2, and the lower right triangle as T_3. These triangles are interconnected at their vertices through single edges. These vertex states are critical states, transitioning from one triangle to another necessitates passing through these states. For instance, starting from an initial state in T_1, the optimal path to reach a desired state in T_2 or T_3 must involve the state transitions $1110 \rightarrow 1112$ and $2220 \rightarrow 2221$, respectively. Indeed, to master the art of solving the ToH puzzle effectively, one must grasp its inherent recursive structure. Success in solving the puzzle relies on systematically working towards reaching the crucial critical states within the state space.

6.1.2 Human Rewards using Maximum Entropy Inverse Reinforcement Learning

In the context of human participants solving the ToH puzzle, we can perceive them as noisy optimal agents striving to optimize an implicit reward function. Utilizing their demonstrations, we can leverage Inverse Reinforcement Learning (IRL) techniques to deduce a reward function [108]. This reward function is designed to align the optimal policy with the observed human demonstrations.

The maximum entropy IRL [109, 110] assumes human demonstrations are not perfect and allows us to learn from sub-optimal demonstrations by incorporating a probabilistic model that captures the variability in human behavior. Maximum entropy IRL has gained significant traction in the literature as a means to effectively learn from human demonstrations [111].

Consider a Markov Decision Process [112] $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma, \mathbf{r}\}$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{T}_{s'}^{sa}$ is the probability of transition from state $s \in \mathcal{S}$ to state $s' \in \mathcal{S}$ under the action $a \in \mathcal{A}$, $\gamma \in [0, 1)$ is the discount factor, and $\mathbf{r} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function. In the case of the n-disk ToH, each edge originating from a given state s in the state transition graph can be considered a unique action. Under these actions, the transition probability $\mathcal{T}_{s'}^{sa}$ equals 1 for the transition from state s to s' if they are connected through an edge. Let $\mathcal{D} = \{\zeta_1, \ldots, \zeta_N\}$ be the set of N demonstrations, where each demonstration ζ_i is a path $\zeta_i = \{(s_{i,0}, a_{i,0}), \ldots, (s_{i,T}, a_{i,T})\}$. The unknown reward function $\mathbf{r}$ is expressed as a linear combination of a set of predefined features denoted as $f : \mathcal{S} \to \mathbb{R}$. The weights associated with these features are learned through the maximum entropy IRL algorithm.

In the framework of maximum entropy IRL, the probability of following a particular trajectory ζ is directly proportional to the exponential of the accumulated rewards experienced along that path. This leads to a stochastic behavior model, where the probability of taking a specific action a in a given state s is determined by the exponential of the expected total reward subsequent to taking that action, i.e., $P(a|s) \propto \exp(Q^r_{sa})$, where Q^r_{sa} is computed as $Q^r_{sa} = \mathbf{r} + \gamma \mathcal{T} V^r_s$. The value function V^r_s is computed using a "soft" variant of the familiar Bellman operator: $V^r_s = \log \sum_a \exp(Q^r_{sa})$. Consequently, the probability of action a in state s is normalized by $\exp(V^r_s)$, yielding $P(a|s) = \exp(Q^r_{sa} - V^r_s)$.

The complete log-likelihood of the observed data under the reward function r can be expressed as:

$$\log P(\mathcal{D}|\mathbf{r}) = \sum_i \sum_t \log P(a_{i,t}|s_{i,t}) = \sum_i \sum_t \left(Q^r_{s_{i,t} a_{i,t}} - V^r_{s_{i,t}} \right). \qquad (6.2)$$

Interested readers are referred to [111] for detailed derivations.

We employ maximum entropy IRL to infer the reward functions associated with human behavior. Detailed results are presented in Section 6.3.2.

6.1.3 Modeling human sequential decision-making under feedback

A central and challenging aspect of designing efficient tutoring systems lies in understanding the impact of evaluative feedback from AI on human decision-making. Precisely, the question of how humans incorporate feedback into their decision-making processes is of paramount importance. In modeling this process, a foundational challenge is understanding how they interpret the feedback, including whether it's seen as an immediate reward or an evaluation of long-term impacts. Further, it's important to explore if feedback relates only to the current action or spans the sequence of actions. Additionally, understanding if evaluative feedback affects the assessment of value functions over time or momentarily influences action choices is vital. To tackle these questions, we develop candidate models that embody different mechanisms for incorporating feedback into human decision-making processes. Our models are inspired by the Training an Agent Manually via Evaluative Reinforcement (TAMER) framework [113–116] developed to incorporate human feedback into the policy of an artificial RL agent.

Let $\hat{H}(s,a)$ and $|f|$ denote the evaluative feedback and number of predefined features f employed to define human rewards, respectively. We study four different models detailed as follows:

- Model 1 - Ignore feedback: This baseline model operates under the assumption that evaluative feedback isn't directly integrated into human decision-making processes. Instead, individuals are postulated to focus on maximizing the long-term value derived from their personal reward functions. In this framework, evaluative feedback plays an indirect role by shaping and refining these reward functions. This is the default model studied in Sec. 6.1.2. The model encompasses $|f|$ learned parameters.

- Model 2 - Update $Q(s,a)$: In this model, we postulate that humans interpret evaluative feedback as an indicator of the long-term effectiveness of their strategic actions, serving as an approximation of Q^r_{sa}. The model integrates this feedback to update the Q-estimate as follows:
$$Q'(s,a) = Q(s,a) + k\hat{H}(s,a), \tag{6.3}$$
where k is a parameter to be learned. Consequently, the policy gets updated as $P(a|s) = \exp\left({Q_{sa}}' - {V_s}'\right)$, where $V'_s = \log \sum_a \exp\left(Q'_{sa}\right)$ denotes the newly adjusted value function. The model encompasses $|f| + 1$ learned parameters.

- Model 3 - Update $r(s,a)$: In this model, we postulate that humans perceive the evaluative feedback as a measure of the myopic effectiveness of the strategy, serving as an approximation of $r(s, a)$. The model updates the human rewards as follows:

$$r'(s, a) = r(s, a) + k\hat{H}(s, a), \tag{6.4}$$

where k is a parameter to be learned. The updated reward function is used to estimate the Q-values and the policy. The model encompasses $|f| + 1$ learned parameters.

- Model 4 - Feedback as a measure of $Q(s, a)$: In this model, we assume that humans ignore learning by interaction and treat evaluative feedback as a fixed measure of $Q(s, a)$. Therefore,

  $$Q'(s, a) = k\hat{H}(s, a), \tag{6.5}$$

  where k is the only parameter to be learned.

Remark 8. *It's worth noting that the* log*-likelihood of the maximum entropy IRL depends solely on $Q(s, a)$ through $P(a|s)$. Consequently, Model 2 can be considered equivalent to another model where humans utilize feedback solely to influence their action selection. In this alternate model, humans do not incorporate evaluative feedback into their estimation of $Q(s, a)$; rather, they use it exclusively to bias their action selection, i.e., $P(a|s) \propto \exp(Q(s, a) + k\hat{H}(s, a))$. In the context of maximum entropy IRL, this model is tantamount to Model 2, where humans employ evaluative feedback to update their Q-estimates as $Q'(s, a) = Q(s, a) + k\hat{H}(s, a)$.*

We investigate these models in Sec 6.3.3 to understand how humans incorporate evaluative feedback in their decision-making.

6.2 Human Experiments

In this section, we discuss the human experiments conducted using AMT.

6.2.1 Experiment Design

We examine the effect of evaluative feedback on sequential decision-making using ToH task. To achieve this, we designed five separate experiments, each featuring a different type of feedback. Participants for each experiment were recruited randomly through AMT. Each participant first solved a 4-disk ToH task ten times (training task) and then a 5-disk ToH task five times (transfer task) to evaluate their skill transfer to a more challenging task. The initial state of each puzzle was standardized with all the disks located on the first peg. Considering the state-space of the puzzle as comprising of three interlocking triangles, the target state was randomly selected from the states within the triangles that did not include the initial state, i.e., triangles T_2 and T_3. The participants were given a maximum number of moves m_{allowed} to solve the puzzle, calculated as :

$$m_{\text{allowed}} = \lceil 1.5 \times m_{\text{min}} \rceil,$$

where m_{min} represents the minimum number of moves from the initial configuration to the final configuration, as determined by the minimum path length in the state graph (Fig. 6.1). The only difference among the experiments was the feedback provided during the 4-disk ToH training task. No feedback was provided during the 5-disk ToH transfer task in any of the experiments. In each experiment, participants were asked to try their best to get the highest scores. The feedback and scoring metrics used during the training task for the five experiments were:

(i. *Experiment* 1 - *No feedback:* In this experiment, the participants solved the 4-disk ToH puzzle without any feedback. The scoring metric for these tasks was selected as:

$$S = 10(m_{\text{allowed}} - m_{\text{used}} + 1), \tag{6.6}$$

where m_{used} is the total moves used to solve the puzzle. The participant receives a score of 0 if the puzzle remains unsolved after exhausting the allowed number of moves. The same scoring metric was used in the 5-disk transfer tasks in all the experiments.

(ii. *Experiment* 2 - *Numeric feedback:* The participants in this experiment received visual feedback on each move they made while solving the 4-disk ToH. The feedback was in the form of text that reads “good move +2" or “bad move −2", indicating whether the move increased or decreased the value of the state, respectively. The scoring metric for the training tasks in this experiment was selected as:

$$S = 10(m_{\text{allowed}} - m_{\text{used}} + 1) + 2(m_{\text{good}} - m_{\text{bad}}), \tag{6.7}$$

where $m_{\text{good}} \in \mathbb{N}$ and $m_{\text{bad}} \in \mathbb{N}$ denote the number of good and bad moves, respectively.

(iii. *Experiment* 3 - *Optional feedback:* In this experiment, the participants did not receive visual feedback automatically but had the option to request it by pressing a button, which came at the cost of a small penalty. If the participant requested feedback, they would receive the same visual feedback as in Experiment 2, which evaluated their last move. The scoring metric for the training tasks in this experiment was as follows:

$$S = 10(m_{\text{allowed}} - m_{\text{used}} + 1) - f_{\text{optional}}, \tag{6.8}$$

where $f_{\text{optional}} \in \mathbb{N}$ denotes the number of times the participant requests feedback.

(iv. *Experiment* 4 - *Sub-goal:* In the state graph of the 4-disk ToH task (as illustrated in Fig. 6.1), states 1110 and 2220 are critical in reaching the target states efficiently in triangles T_2 and T_3, respectively. In this experiment, based on the target configuration, the participants were presented with an intermediate sub-goal configuration (1110 or 2220) in addition to the target configuration. The participant was instructed to try to reach the intermediate sub-goal first. The scoring metric for the training tasks in this experiment was as follows:

$$S = 10(m_{\text{allowed}} - m_{\text{used}} + 1) + 5s_{\text{subgoal}}, \tag{6.9}$$

where $s_{\text{subgoal}} \in \{0, 1\}$ was set to 1 if the participant successfully reaches intermediate sub-goal configuration, and 0 otherwise.

(v. *Experiment* 5 - *Sub-goal with numeric feedback:* In this experiment, the participants received both the visual feedback as in Experiment 2 and the intermediate sub-goal configuration as in Experiment 4. The scoring metric for the training tasks in this experiment was calculated as follows:

$$S = 10(m_{\text{allowed}} - m_{\text{used}} + 1) + 2(m_{\text{good}} - m_{\text{bad}}) + 5s_{\text{subgoal}}. \tag{6.10}$$

This experiment provided the participants with the maximum amount of evaluative feedback.

6.2.2 Methods

After receiving the IRB consent (MSU IRB #8421) from Michigan State University's IRB office, we recruited 238 participants using AMT for the study. Inclusion criteria were established as having completed a minimum of 500 prior studies and maintaining a 98% approval rate on the platform. Participants were compensated with a base payment of \$6 and had the opportunity to earn additional performance-based bonuses ranging from \$0 − \$4. Of the recruited participants, 78 participants were excluded due to self-reported prior experience with the ToH task.

6.3 Results

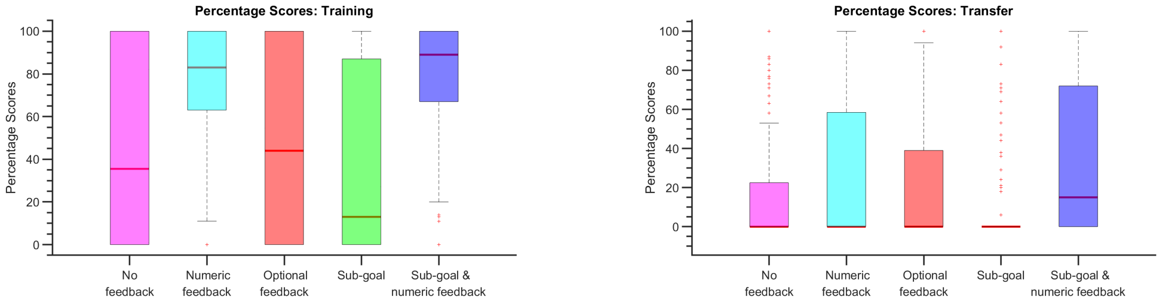
In this section, we discuss the results of the experiments conducted on AMT.

6.3.1 Performance under Evaluative Feedback

First, we collect the data of 20 participants each for the 5 set of experiments detailed in Section 6.2.1. In every experiment, we assess participants' performance by calculating their percentage scores for both the training and transfer tasks as follows:

$$100 \times \left(\frac{m_{\text{allowed}} - m_{\text{used}} + 1}{m_{\text{allowed}} - m_{\text{min}} + 1} \right).$$

In Fig. 6.3a, we present box plots illustrating the percentage scores achieved in the training tasks (4-disk ToH). Notably, participants who underwent training with evaluative feedback in Experiment 2 (numeric feedback) and Experiment 5 (sub-goal with numeric



(a) Training.

(b) Transfer.

Figure 6.3 Box plots displaying percentage scores for both training (a) and transfer (b) tasks. Within each box plot, the median is represented by the red horizontal line, while the lower and upper edges of the box signify the 25th and 75th percentiles, respectively. Whiskers extend to encompass the most extreme data points that are not classified as outliers, and individual outliers are plotted using the symbol '+'.

feedback) exhibited significantly improved performance during these training tasks compared to participants in Experiment 1 (no feedback), who received no evaluative feedback.

In Experiment 3 (optional feedback), participants seldom requested feedback to avoid the feedback penalty, resulting in performance levels akin to those observed in Experiment 1. Experiment 4 (sub-goal) introduced a unique approach, where participants were exclusively exposed to sub-goal configuration (1110 or 2220) crucial for reaching the desired target state. In the absence of evaluative feedback, this method resembled the conditions of Experiment 1, where the sub-goal can be effectively thought as a target state until the sub-goal state is reached. We hypothesize that supplying solely sub-goal configurations without evaluative feedback may induce confusion, as participants may now consider two target states simultaneously—the sub-goal and the target state. Consequently, participants in Experiment 4 exhibited a marginal decrease in performance compared to those in Experiment 1.

In Fig. 6.3b, we present box plots illustrating the percentage scores achieved in the transfer tasks involving the 5-disk ToH. It's important to note that solving the 5-disk ToH, with its 243 states, presents a significantly greater challenge compared to the training task, which involved the 4-disk ToH with 81 states. Furthermore, participants had no prior experience with the 5-disk ToH and relied solely on their training with the 4-disk ToH. Consequently, the transfer tasks yielded relatively lower scores, with many trials failing to solve the puzzle within the allotted number of moves, which can make it challenging to interpret the box plots in Fig. 6.3b.

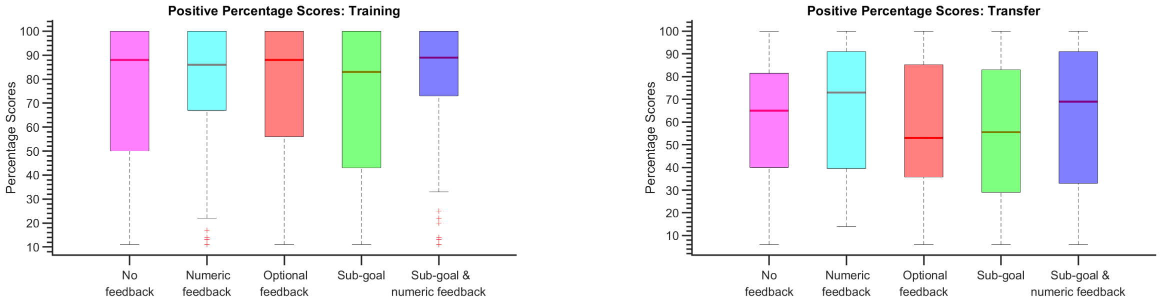
To focus on successful outcomes, we filtered for positive percentage scores in each experiment, representing the trials where participants successfully solved the ToH puzzle. Table 6.1 provides an overview of the percentage of successful trials for each experiment, both in the training and transfer tasks. Notably, Experiment 2 and Experiment 5 demonstrated a substantial improvement in successful trials, showing increases of 33.5% and 36%, respectively, compared to Experiment 1 in the training tasks. In the transfer tasks, Experiments 2 and 5 also showed notable improvements, with success rates increasing by 13% and 26%, respectively, compared to Experiment 1.

To assess the statistical significance of these findings, we conducted a two-sample t-test comparing the results of Experiments 2 and 5 with the data from Experiment 1. Remarkably, the p values for Experiment 2 (in comparison to Experiment 1) and Experiment 5 (relative to Experiment 1) are 1.59×10^{-12} and 1.71×10^{-17}, respectively, in the training tasks, indicating highly significant differences. In the transfer tasks, the p values are 3.9×10^{-2} and 7.17×10^{-4} for Experiments 2 and 5 compared to Experiment 1, respectively. Consistent with the commonly accepted significance level of 0.05, a p value below this threshold leads us to reject the null hypothesis, indicating that the data from the two experiments do not arise from the same distribution at a 5% significance level. These results underscore the substantial impact of evaluative feedback on performance, both in the training and transfer tasks.

	No feedback	**Numeric feedback**	**Optional feedback**	**Sub-goal**	**Sub-goal with numeric feedback**
Training	58%	91.5%	62%	52.5%	94%
Transfer	28%	41%	35%	22%	54%

Table 6.1 Percentage of successful trials in the training and transfer tasks.

Fig. 6.4a and 6.4b display box plots representing successful trials after filtering for positive scores. Notably, the medians of these box plots closely align with each other, suggesting that

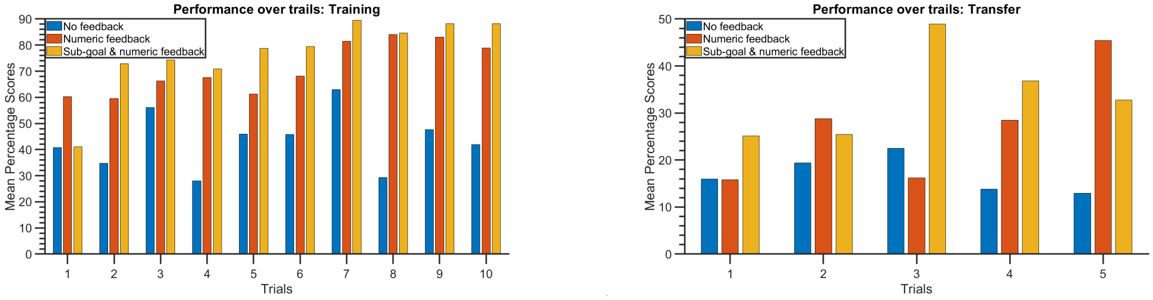


(a) Training.

(b) Transfer.

Figure 6.4 Box plots displaying positive percentage scores for both training (a) and transfer (b) tasks.

participants' performances in the experiments can be effectively compared solely through the percentage of successful trials. Once participants have successfully learned to solve the ToH puzzle, their scores exhibit relatively little variation across experiments during successful trials. This observation highlights the stability and consistency of participants' performance once they have mastered the task.



(a) Training.

(b) Transfer.

Figure 6.5 Bar plots displaying the mean percentage scores for different trials for both training (a) and transfer (b) tasks.

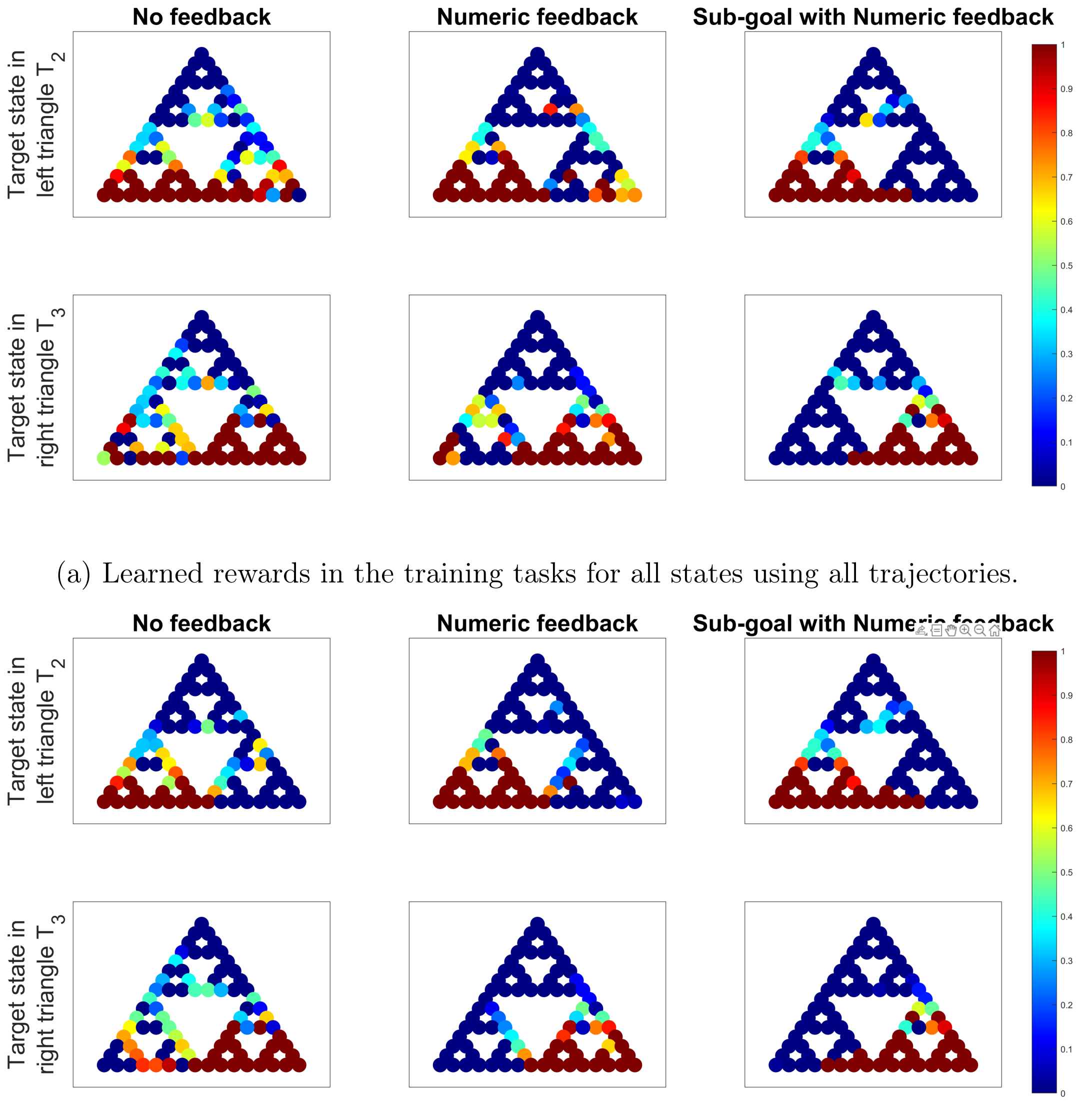
Recall that each participant completed 10 trials of training and 5 trials of transfer tasks. In Figures 6.5a and 6.5b, bar plots represent the mean percentage scores for different trials in the training and transfer tasks, respectively. It's evident that participants who received no feedback exhibited relatively low scores compared to those who received either numeric feedback or sub-goals with numeric feedback. Furthermore, while there is no consistent improvement over the trials for participants who did not receive feedback, participants who received evaluative feedback demonstrated performance enhancement with increasing scores across trials. Similar trends are observable in the transfer tasks, indicating that participants who received evaluative feedback found it easier to transfer their skills to related tasks and showed improvement across trials.

The results in Table 6.1 underscore significant improvements in human decision-making attributed to evaluative feedback during training tasks, along with effective skill transfer to related tasks. We employ maximum entropy IRL [110] to investigate the pivotal role of evaluative feedback in shaping human decision-making, as detailed in Sections 6.3.2 and 6.3.3. To enable this analysis, we conducted additional data collection sessions with 20 participants each, encompassing experiments devoid of feedback (Experiment 1) and those involving evaluative feedback (Experiments 2 and 5).

6.3.2 Human Rewards under Evaluative Feedback

In this section, we treat humans solving the ToH puzzle as noisy optimal agents striving for optimal play with some implicit reward structure. We examine participants from three sets of experiments: (a) No feedback (Experiment 1), (b) Numeric feedback (Experiment 2), and (c) Sub-goal with numeric feedback (Experiment 5). To gain insights into human learning under these varying feedback conditions, we employ maximum entropy IRL analysis to uncover the underlying human reward structures. Visualizing these human rewards can offer valuable insights into the learning process with and without evaluative feedback.

Recall that for each experiment, the initial state is standardized with the starting state represented by the top vertex of triangle T_1, and the target state is randomly selected from either triangles T_2 or T_3 (see Fig. 6.2). For each experiment, we partition the experimental data into two sets, one with target states in T_2 and the other with target states in T_3. In each of these sets, we learn the human rewards expressed as a linear combination of predefined features. By modifying these predefined features, we consider two different settings where the human rewards are learned for all the states and for a subset of 8 states. To estimate the rewards, we maximize the log-likelihood as defined in (6.2), while applying an $\mathcal{L}_1$ penalty to promote sparse rewards. To determine the coefficient of the $\mathcal{L}_1$ penalty $\lambda \in \mathbb{R}_{\geq 0}$, we consider



(a) Learned rewards in the training tasks for all states using all trajectories.

(b) Learned rewards in the training tasks for all states using only successful trajectories.

Figure 6.6 IRL plots displaying learned human rewards in the training tasks for all states, using trajectory datasets (from trials 6-10 for each participant) from each experiment that encompass (a) all available trajectories and (b) only successful trajectories, where success is defined by reaching the target state.

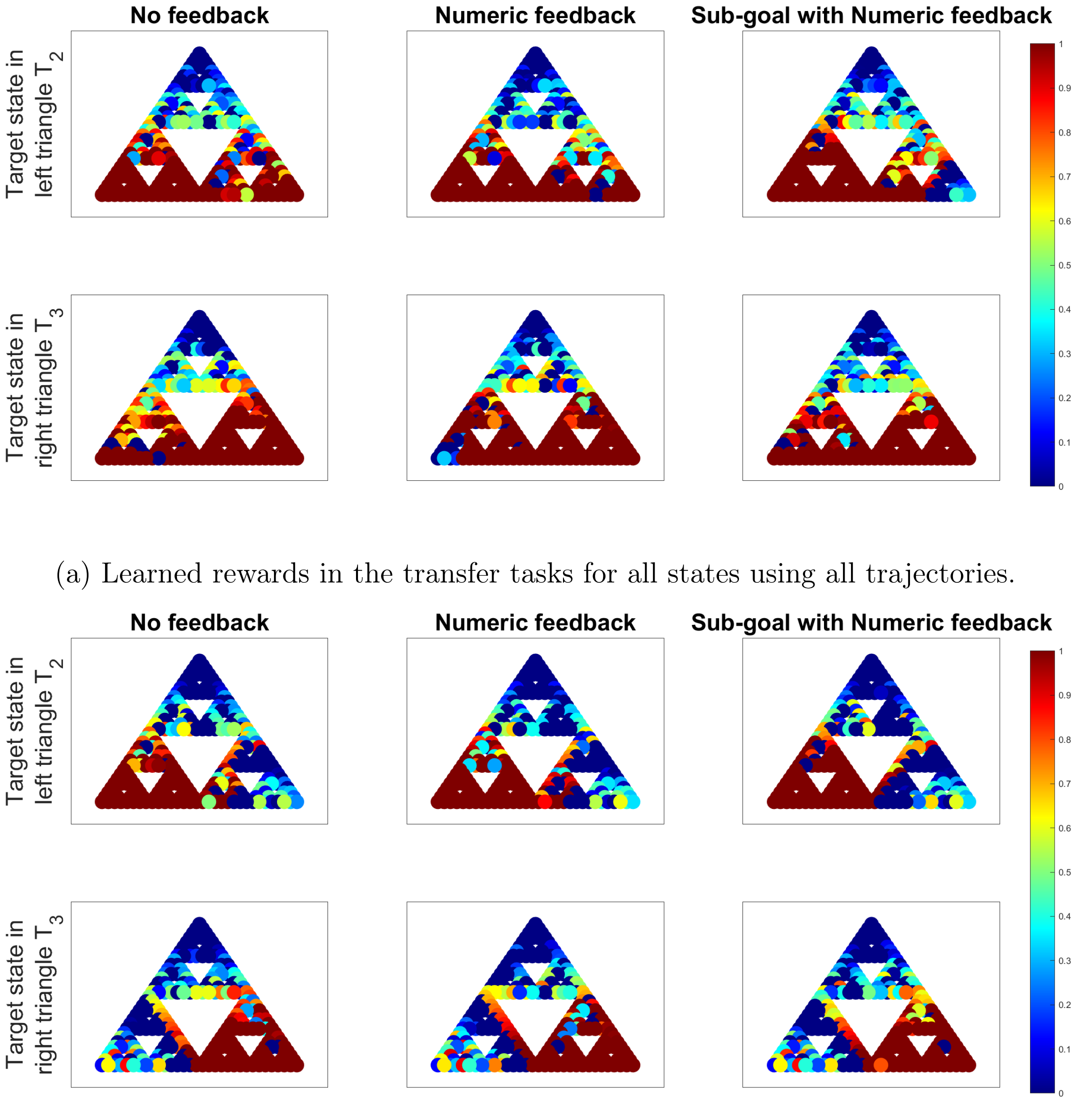
$\lambda \in \{0, 0.1, \ldots, 2\}$ and perform 5-fold cross-validation on the data and select the coefficient that yields the maximum mean log-likelihood across the 5-fold validation sets.

Fig. 6.6a and 6.6b display the learned IRL rewards in the training tasks for all states. While IRL typically assumes expert demonstrations, it's important to note that participants may still be learning the task during the initial trials. Since the performance does not vary significantly in the latter half of the trials (see Fig. 6.5a), we assume that the human rewards are relatively stationary from trials 6 to 10 and, therefore, exclusively utilize these trials for our IRL analysis. From these latter trajectories, we derive IRL rewards, considering both (a) all available trajectories and (b) only the successful ones, where success is defined by reaching the target state.

Each of these plots is organized into a grid with 2 rows and 3 columns. The top row represents trajectories with the target state in triangle T_2, while the bottom row represents trajectories with the target state in triangle T_3. The columns correspond to the three sets of experiments: no feedback, numerical feedback, and sub-goal with numerical feedback arranged from left to right.

In Fig. 6.6a, it becomes apparent that participants' rewards in the experiment with no feedback (first column) exhibit a distribution across all states, encompassing both T_2 and T_3, despite the target state's placement in T_2 for the first row and in T_3 for the second row. The occurrence of high rewards in T_3 (respectively T_2) when the target state resides in T_2 (respectively T_3) primarily stems from the unsuccessful attempts to solve the ToH puzzle in each experiment. Consequently, we observe that as participants' performance improves across experiments from left to right, rewards increase within the triangle containing the target state while decreasing in the opposing triangle. Another noteworthy observation is the presence of high rewards at the critical states (vertices of the target triangle), which serve as pivotal entry points to the target triangle. These rewards become more pronounced as performance enhances from left to right.

Fig. 6.6b depicts the learned IRL rewards derived exclusively from successful trajectories in each experiment. Due to the absence of failed trajectories in each experiment, the disparities in IRL rewards across experiments, from left to right, become less pronounced. In each experiment, states within the target triangle and critical states exhibit higher rewards compared to the opposing triangles. In Experiments 1 and 2, the elevated rewards along the edge in the opposite triangle, which is closer to the target triangle, suggest that participants in these experiments occasionally complete the puzzle by opting for suboptimal routes. In contrast, participants in Experiment 5 predominantly solve the puzzle utilizing the optimal trajectory.



(a) Learned rewards in the transfer tasks for all states using all trajectories.

(b) Learned rewards in the transfer tasks for all states using only successful trajectories.

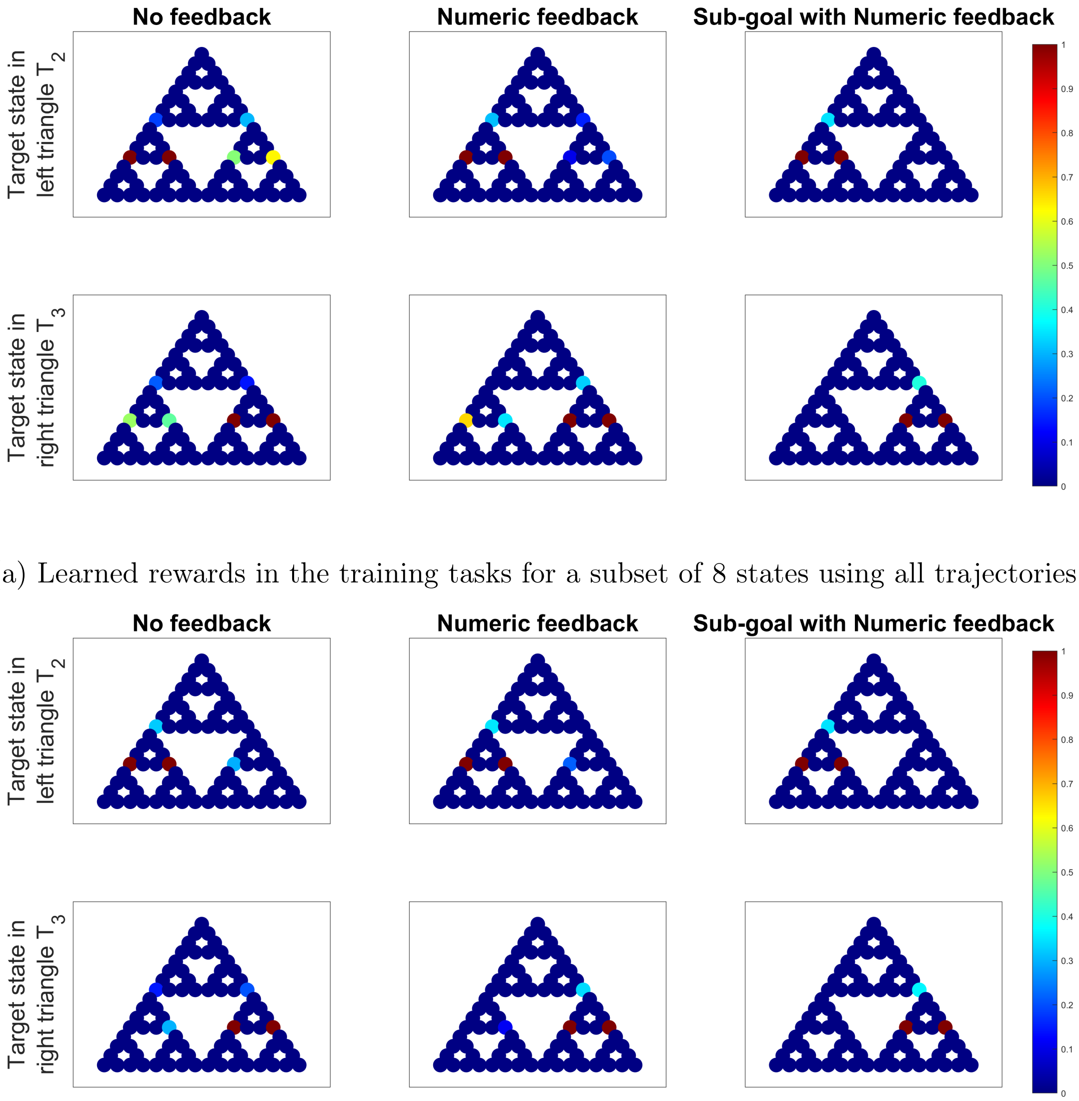
Figure 6.7 IRL plots displaying learned human rewards in the transfer tasks for all states, using trajectory datasets from each experiment that encompass (a) all available trajectories and (b) only successful trajectories, where success is defined by reaching the target state.

Fig. 6.7a and 6.7b present the learned IRL rewards for all states within the transfer tasks, utilizing trajectory datasets that encompass (a) all available trajectories and (b) only successful trajectories. It is important to note that the transfer tasks pose significant challenges, with none of the participants receiving any feedback. Consequently, the trajectories for the transfer tasks in each experiment comprise numerous failed trajectories.

However, a noticeable trend emerges: participants from Experiment 5, who were trained using sub-goals with numeric feedback, exhibit faster learning in solving the transfer tasks compared to participants from Experiments 1 and 2, who received no feedback and only numerical feedback, respectively. This is evident from the higher rewards within the target triangle and lower rewards in the opposite triangle for Experiment 5. When considering only successful trajectories to derive the IRL rewards in Fig. 6.7b, the differences across experiments become less pronounced due to the exclusion of failed trajectories in all experiments.

The results presented in Fig. 6.6 and 6.7 offer valuable insights into how humans acquire puzzle-solving skills under various evaluative feedback strategies. However, it's worth noting that the learned rewards appear less sparse due to the predefined features, which permit non-zero rewards in all states. Consequently, while these learned IRL rewards for all states offer insights into critical states, they can complicate the comparison between experiments. Furthermore, most of the RL rewards are often sparse. To this end, we modify the predefined features to encourage sparser rewards, allowing non-zero rewards in only 8 states for both the training and transfer tasks. These 8 states were thoughtfully selected as the vertices of the smaller triangles within the state space. In Fig. 6.2, these states correspond to $2200, 1100, 1110, 2220, 0012, 2212, 1121, 0021$.

Fig. 6.8a and 6.8b illustrate the learned IRL rewards for a specific subset of 8 states during the training tasks. These rewards are derived from trajectory datasets obtained from the latter half of the trials (trials 6 to 10) for each participant. We consider two scenarios: (a) using all available trajectories and (b) using only the trajectories that resulted in successful task completion. It is evident that participants from Experiment 5 demonstrate non-zero rewards exclusively within the target triangle and the corresponding critical states. As we progress from left to right, the non-zero rewards in the opposite triangle diminish due to



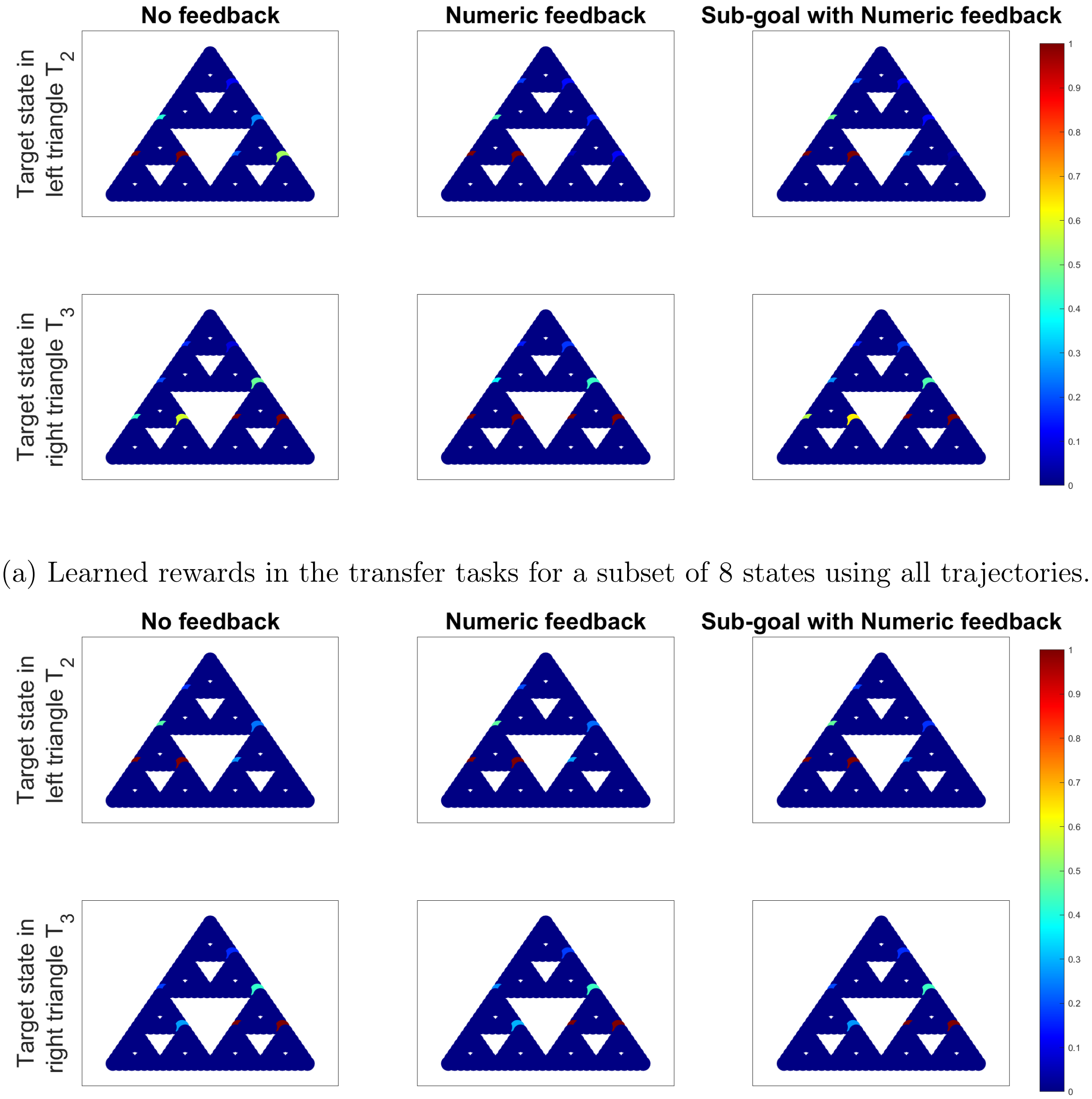
(a) Learned rewards in the training tasks for a subset of 8 states using all trajectories.

(b) Learned rewards in the training tasks for a subset of 8 states using only successful trajectories.

Figure 6.8 IRL plots displaying learned human rewards in the training tasks for a subset of 8 states, using trajectory datasets (from trials 6-10 for each participant) from each experiment that encompass (a) all available trajectories and (b) only successful trajectories, where success is defined by reaching the target state.

fewer instances of failure. These differences become less pronounced when we solely consider successful trajectories in Fig. 6.8b.

Fig. 6.9a and 6.9b depict the learned IRL rewards for a selected subset of 8 states within the transfer tasks, using trajectory datasets that encompass (a) all available trajectories and



(a) Learned rewards in the transfer tasks for a subset of 8 states using all trajectories.

(b) Learned rewards in the transfer tasks for a subset of 8 states using only successful trajectories.

Figure 6.9 IRL plots displaying learned human rewards in the transfer tasks for a subset of 8 states, using trajectory datasets from each experiment that encompass (a) all available trajectories and (b) only successful trajectories, where success is defined by reaching the target state.

(b) only successful trajectories. While the distinctions are somewhat less pronounced due to the presence of numerous failure attempts in all experiments, the lower rewards in the opposite triangle indicate swifter learning when participants are trained with feedback, in contrast to participants who receive no feedback. These differences become less noticeable when we exclusively consider successful trajectories in Fig. 6.9b, effectively eliminating most of the non-zero rewards in the opposite triangle.

The results of the max entropy IRL analysis underscore the significance of critical states and demonstrate how human learning in sequential decision-making tasks can be organized more effectively when evaluative feedback is provided, in contrast to participants solely learning through exploration without any feedback. The results further indicate that the participants trained with evaluative feedback exhibit an ability to transfer their learning to newer, related, and more demanding tasks at a significantly accelerated pace compared to those who learn without feedback.

6.3.3 Modeling Human Decision-Making under evaluative feedback

In Sec. 6.3.1 and 6.3.2, we have demonstrated the pivotal role of evaluative feedback in enhancing learning and performance within the context of the ToH puzzle. In this section, we delve into exploring models that aim to elucidate the mechanisms through which humans integrate evaluative feedback into their decision-making processes.

In our analysis, we explore four distinct models for incorporating feedback into human decision-making, as detailed in Sec. 6.1.3. For each of these models, we calculate both the Akaike information criterion (AIC) [92] and Bayesian information criterion (BIC) [93] to identify the most suitable model. For a given model, the AIC and BIC are defined as:

$$\mathrm{AIC} = 2p - 2\log(\hat{L}), \quad \mathrm{BIC} = p\log(o) - 2\log(\hat{L}), \tag{6.11}$$

where p, o, and $\hat{L}$ denote the number of learned parameters, the number of observations, i.e., the sample size, and the maximized value of the likelihood function of the model, respectively. The model with the lowest AIC (or BIC) is deemed the optimal choice according to AIC (or BIC) criteria. For this analysis, we leverage the experimental data gathered during the training tasks of Experiment 2, where participants received numeric feedback.

It is important to note that this numeric feedback is determined based on the change in state value before and after the state transition. Consequently, it is intrinsically tied to the target state, given that the value function is contingent upon the target state.

Since the target state is subject to randomization in triangles T_2 and T_3, we further segment these triangles into three sub-triangles each. This subdivision allows us to categorize the experimental data into six distinct groups, based on the location of the target state within these six sub-triangles. Within each group, we select the top vertex of the sub-triangle as the designated target state and truncate the trajectories to the point at which they initially enter the target sub-triangle.

Upon completing this partitioning process for the 200 trajectories obtained from the training tasks, we arrived at six groups, each containing a respective number of trajectories: $41, 34, 27, 31, 32, 35$. Within each group, we subject all four models to testing, making appropriate modifications to either the Q-function or the reward function as discussed in Sec. 6.1.3. To estimate the unknown parameters for each model, we employ maximum entropy IRL, optimizing the log-likelihood as defined in (6.2) while applying an $\mathcal{L}_1$ penalty.

To determine the coefficient for the $\mathcal{L}_1$ penalty in each model, we consider $\lambda \in \{0, 0.2, \ldots, 1\}$ and perform 5-fold cross-validation. This allows us to select the coefficient that results in the highest mean log-likelihood across the five validation sets.

		Model 1		**Model** 2		**Model** 3		**Model** 4	
Num. of parameters		81		82		82		1	
Group	**Num. of obs.**	**AIC**	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**
1	41	23.65	27.03	23.02	26.45	23.70	27.12	**22.89**	**22.93**
2	34	27.76	31.40	**22.70**	26.38	27.82	31.50	24.66	**24.70**
3	27	30.84	34.73	**26.94**	30.87	30.52	34.46	28.49	**28.54**
4	31	21.65	25.40	**20.68**	24.47	21.72	25.51	22.16	**22.20**
5	32	27.69	31.40	**23.49**	27.25	27.76	31.51	24.30	**24.34**
6	35	30.42	34.02	**26.56**	30.21	30.47	34.12	27.620	**27.66**

Table 6.2 AIC, and BIC values (normalized by the number of observations) for different models allowing non-zero rewards for all the 81 states.

Similar to Sec. 6.3.2, we investigate two settings for the predefined features: one that allows non-zero rewards in all states and another that restricts rewards to a subset of 8 states.

Table 6.2 presents the AIC and BIC values (normalized by the number of observations) for different models within each group when non-zero rewards are permitted for all 81 states. It's worth noting that, while Model 2, where $Q'(s, a) = Q(s, a) + k\hat{H}(s, a)$, emerges as the best fit according to AIC for the majority of the groups, Model 4, where $Q'(s, a) = k\hat{H}(s, a)$, is selected as the best fit under the BIC criterion. This preference for Model 4 under BIC is attributed to the significant difference in the number of learned parameters between the two models, with BIC favoring the model with fewer learned parameters. Indeed, when considering non-zero rewards for all 81 states, it leads to a preference for the model with just a single learned parameter.

		Model 1		**Model** 2		**Model** 3		**Model** 4	
Num. of parameters		8		9		9		1	
Group	**Num. of obs.**	**AIC**	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**
1	41	31.84	32.18	**22.54**	**22.92**	31.89	32.27	22.92	22.96
2	34	41.72	42.08	**24.03**	**24.43**	41.77	42.18	24.76	24.80
3	27	44.04	44.43	**27.74**	**28.17**	44.12	44.55	28.43	28.48
4	31	30.49	30.86	**21.26**	**21.68**	30.56	30.97	22.20	22.25
5	32	42.78	43.15	**23.62**	**24.03**	42.84	43.26	24.41	24.45
6	35	43.53	43.89	**27.25**	27.65	43.59	43.99	27.57	**27.62**

Table 6.3 AIC, and BIC values (normalized by the number of observations) for different models allowing non-zero rewards for a sub-set of 8 states.

Table 6.3 presents the AIC and BIC values (normalized by the number of observations) for different models within each group when non-zero rewards are allowed for only a subset of 8 states. This setting represents a more realistic scenario with sparse rewards. Notably, in this context, Model 2 consistently emerges as the best fit according to both the AIC and BIC criteria. This suggests that humans tend to interpret evaluative feedback as a strong indicator of the long-term effectiveness of their strategic actions.

Remark 9. *Even though Model* 2 *stands out as the preferred model according to both AIC and BIC criteria, there is a small evidence of support for Model* 4 *as well (in case of non-sparse rewards). This suggests that there might be instances where some individuals do not primarily* *learn through interaction but instead focus on maximizing their evaluative feedback directly. Such individuals could potentially encounter challenges in transfer tasks where evaluative feedback is not available.*

6.4 Conclusions

In this work, we study the influence of AI-generated evaluative feedback on human decision-making, with a specific focus on sequential decision-making tasks exemplified by the Tower of Hanoi. Our study demonstrates that individuals who receive training with evaluative feedback not only experience significant improvements in their decision-making abilities but also excel in transferring these enhanced skills to similar tasks. Through an analysis utilizing the maximum entropy inverse reinforcement learning framework, we show that human learning exhibits a more structured and organized implicit reward pattern when evaluative feedback is provided during the training process. This highlights the critical role played by AI-generated feedback in improving the cognitive and strategic abilities of individuals.

Furthermore, our investigation explores various models to better comprehend how humans integrate feedback into their decision-making processes. Our findings provide substantial evidence suggesting that individuals tend to interpret evaluative feedback as a valuable indicator of the long-term effectiveness of their strategic actions. This valuable insight can be leveraged to design intelligent IoT devices, capable of enriching human learning experiences and shaping human decision-making.

CHAPTER 7

INCENTIVIZING COLLABORATION IN HETEROGENEOUS TEAMS VIA COMMON-POOL RESOURCE GAMES

In this chapter, we address the challenge of achieving efficient collaboration in a team of heterogeneous agents with different skill-sets and expertise. We consider a team of heterogeneous agents that is collectively responsible for servicing, and subsequently reviewing, a stream of homogeneous tasks. Each agent has an associated mean service time and a mean review time for servicing and reviewing the tasks, respectively. Agents receive a reward based on their service and review admission rates. The team objective is to collaboratively maximize the number of “serviced and reviewed" tasks. We formulate a Common-Pool Resource (CPR) game and design utility functions to incentivize collaboration among heterogeneous agents in a decentralized manner. We show the existence of a unique Pure Nash Equilibrium (PNE), and establish convergence of best response dynamics to this unique PNE. Finally, we establish an analytic upper bound on three inefficiency measures of the PNE, namely the price of anarchy (PoA), the ratio of the total review admission rate (TRI), and the ratio of latency (LI).

7.1 Background and Problem Formulation

In this section, we describe the problem setup and formulate the problem using a game-theoretic framework. We also present some definitions that will be used in the paper.

7.1.1 Problem Description

We consider a heterogeneous team of $N \in \mathbb{N}$ agents tasked with servicing a stream of homogeneous tasks. These agents could be autonomous systems or human operators. Each task, after getting serviced by a team-member, gets stored in a common review pool for a second review. This second review is a feedback process in which any team-member can re-examine the serviced task from the common review pool for performance monitoring and quality assurance purposes. Each agent $i \in \mathcal{N} = \{1, \dots, N\}$ may choose to spend a portion of her time to review the tasks from the common pool while spending her remaining time

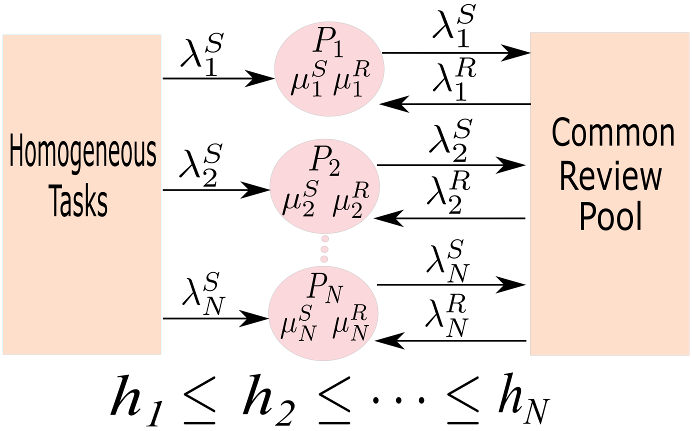


Figure 7.1 Player i devotes her time to service homogeneous tasks (at a constant service admission rate λ_i^S) while reviewing serviced tasks from the common review pool (at a constant review admission rate λ_i^R). The maximum admission rate for player i for servicing and reviewing the tasks is given by μ_i^S and μ_i^R, respectively.

to service the incoming tasks. We consider heterogeneity among the operators due to the difference in their level of expertise and skill-sets in servicing and reviewing the tasks. This heterogeneity is captured by the average service time $(\mu_i^S)^{-1} \in \mathbb{R}_{>0}$ and average review time $(\mu_i^R)^{-1} \in \mathbb{R}_{>0}$ spent by operator $i \in \mathcal{N}$ on servicing and reviewing a task, respectively.

Let $\lambda_i^S \in [0,\ \mu_i^S]$ and $\lambda_i^R \in [0,\ \mu_i^R]$ be the deterministic service and review admission rates, i.e., the rates at which agent i chooses to admit tasks for servicing and reviewing, respectively. Each agent i can choose their service and review admission rate independent of other agents. The range of λ_i^S and λ_i^R have been chosen to satisfy the stability conditions (including marginal stability) for the service and review queues for operator $i \in \mathcal{N}$ [117, Chapter 8].

Suppose agent i selects λ_i^S and λ_i^R as their service and review admission rates, then

$$\frac{\lambda_i^S}{\mu_i^S} + \frac{\lambda_i^R}{\mu_i^R} \leq 1,$$

where $\frac{\lambda_i^S}{\mu_i^S}$ (respectively, $\frac{\lambda_i^R}{\mu_i^R}$) is the average time the agent spends on servicing (respectively, reviewing) the tasks within a unit time. Thus, if the agent has selected a review admission rate λ_i^R, then the service admission rate satisfies

$$\lambda_i^S \leq \mu_i^S - h_i \lambda_i^R, \tag{7.1}$$

where $h_i := \frac{\mu_i^S}{\mu_i^R}$ is the heterogeneity measure for the player i.

We consider self-interested agents that receive a utility based on their service and review admission rates. Hence, we will assume that agents operate at their maximum capacity, and equality holds in (7.1). Fig. 7.1 shows the schematic of our problem setup. Note that only serviced tasks are available for review, and therefore,

$$\sum_{i=1}^{N} \lambda_i^R \leq \sum_{i=1}^{N} \lambda_i^S. \tag{7.2}$$

By substituting (7.1) in (7.2), we obtain,

$$\sum_{i=1}^{N} a_i \lambda_i^R \leq \sum_{i=1}^{N} \mu_i^S, \tag{7.3}$$

where $a_i := (1 + h_i)$. Eq. (7.3) represents the system constraint on the review admission rates chosen by agents.

We are interested in incentivizing collaboration among the agents for the better team performance. Towards this end, we propose a game-theoretic setup defined below.

7.1.2 A Common-Pool Resource Game Formulation

We now formulate our problem as a Common-Pool Resource (CPR) game. Henceforth, we would refer to each agent as a player. A maximum service admission rate μ_i^S and a maximum review admission rate μ_i^R are associated with each player i, based on her skill-set and level of expertise. Without loss of generality, let the players be labeled in increasing order of their heterogeneity measures, i.e., $h_1 \leq \cdots \leq h_N$.

Let $S_i := [0,\ \mu_i^R]$ be the strategy set for each player i, from which the player chooses her review admission rate for reviewing the tasks from the common review pool. Since we have assumed (7.1) holds with equality, once player i decides her review admission rate $\lambda_i^R \in S_i$, her service admission rate for servicing the tasks λ_i^S is given by the right hand side of (7.1). Let $S = \prod_{i \in \mathcal{N}} S_i$ be the joint strategy space of all the players, where $\prod$ denotes the Cartesian product. Furthermore, we define $S_{-i} = \prod_{j \in \mathcal{N}, j \neq i} S_j$ as the joint strategy space of all the players except player i.

For brevity of notation, we denote the total service admission rate and the total review admission rate by $\lambda_T^S = \sum_{i=1}^N \lambda_i^S$ and $\lambda_T^R = \sum_{i=1}^N \lambda_i^R$, respectively. Similarly, $\mu_T^S = \sum_{i=1}^N \mu_i^S$

and $\mu_T^R = \sum_{i=1}^N \mu_i^R$ denote the aggregated sum of the maximum service admission rates and maximum review admission rates of all the players, respectively.

Let $x \in \mathbb{R}$, defined by

$$x = \lambda_T^S - \lambda_T^R = \mu_T^S - \sum_{i=1}^N a_i \lambda_i^R, \tag{7.4}$$

be the slackness parameter for system constraint (7.3). The constraint (7.3) is violated for negative values of x, i.e., when total review rate exceeds the total service rate. In such an event, some players commit to review more tasks than that are available in the common review pool. The slackness parameter characterizes the gap between the total service admission rate and the total review admission rate for all the players. In order to maximize high quality team throughput, i.e., the number of tasks that are both serviced and reviewed, we seek to incentivize the team to operate close to $x = 0$.

Each player i receives a constant reward $r^S \in \mathbb{R}_{>0}$ for servicing each task. Hence, the service utility $u_i^S : S_i \mapsto \mathbb{R}_{>0}$ for player i servicing the tasks at the service admission rate λ_i^S is given by:

$$u_i^S = \lambda_i^S r^S. \tag{7.5}$$

To incentivize collaboration among the agents, we design the review utility $u_i^R : S \mapsto \mathbb{R}_{>0}$ received for reviewing the tasks from the common review pool using two functions: a rate of return, $r^R : S \mapsto \mathbb{R}_{>0}$ for each reviewed task and a constraint probability $p : S \mapsto [0, 1]$ of the common review pool. The constraint probability p is a soft penalty on the violation of system constraint (7.3).

We model the rate of return r^R and the constraint probability p in terms of the strategy of all the players through slackness parameter x. Furthermore, we assume that r^R is strictly decreasing in x. Therefore, for each $x \in [0,\ \mu_T^S]$, system constraint (7.3) is satisfied, and the rate of return is maximized at $x = 0$. The rate of return can be interpreted as the perks that the employer provides to all the employees for high quality service. For example, an employer generates higher revenue based on the high quality throughput of her company, i.e., based on the number of “serviced and reviewed" tasks, which she redistributes among her employees as perks as per their contribution to the review process. Highest quality throughput is achieved by the company when the team efficiently reviews all the serviced tasks, i.e., when $x = 0$.

We introduce the constraint probability p as a soft penalty of the violation of system constraint (7.3), and therefore, we let $p = 1$ if the constraint gets violated, i.e. when $x < 0$. We assume that the constraint probability p is non-increasing in x, and approaches 1 as x approaches 0. The class of sharply decreasing exponential functions, $p(\lambda_i^R, \lambda_{-i}^R) = \exp(-Ax)$, where $x \in [0,\ \mu_T^S]$ and $A \in \mathbb{R}_{>0}$, can be a good choice to effectively model the system constraint. While the system constraint (7.3) ensures stability in the asymptotic limit, the deviation of empirical mean service/review times from the true mean service/review times during the transient may lead to large queue lengths. Hence, the constraint probability p can be interpreted as the deviation probability of the empirical mean service/review times from the true mean service/review times used in (7.3). If the constraint is violated with probability p, then $u_i^R = 0$ for each player i. Therefore, we define the utility u_i^R by

$$u_i^R(\lambda_i^R, \lambda_{-i}^R) = \begin{cases} 0, & \text{with probability } p(\lambda_i^R, \lambda_{-i}^R), \\ \lambda_i^R r^R(\lambda_i^R, \lambda_{-i}^R), & \text{otherwise.} \end{cases} \tag{7.6}$$

Let $u_i(\lambda_i^R, \lambda_{-i}^R) = u_i^S + u_i^R$ be the total utility of player $i \in \mathcal{N}$. Each player i tries to maximize her expected utility $\tilde{u}_i : S \mapsto \mathbb{R}$ defined by

$$\begin{aligned} \tilde{u}_i(\lambda_i^R, \lambda_{-i}^R) &= \mathbb{E}[u_i^S(\lambda_i^R, \lambda_{-i}^R) + u_i^R(\lambda_i^R, \lambda_{-i}^R)] \\ &= \lambda_i^S r^S + \lambda_i^R r^R(\lambda_i^R, \lambda_{-i}^R)(1 - p(\lambda_i^R, \lambda_{-i}^R)), \end{aligned} \tag{7.7}$$

where the expectation is computed over the constraint probability p. Since r^R and p depend on the review admission rates of all the players only through the slackness parameter x, with a slight abuse of notation, we express $r^R(\lambda_i^R, \lambda_{-i}^R)$ and $p(\lambda_i^R, \lambda_{-i}^R)$ by $r^R(x)$ and $p(x)$, respectively. Substituting (7.1) in (7.7), we get:

$$\begin{aligned} \tilde{u}_i &= \mu_i^S r^S + \lambda_i^R \left[r^R(x)(1 - p(x)) - h_i r^S \right] \\ &=: \mu_i^S r^S + \lambda_i^R f_i(x), \end{aligned} \tag{7.8}$$

where $f_i : S \mapsto \mathbb{R}$ is defined by

$$f_i(\lambda_i^R, \lambda_{-i}^R) = f_i(x) = r^R(x)(1 - p(x)) - h_i r^S. \tag{7.9}$$

The function f_i is the incentive for player i to review the tasks. Note that player i will choose a non-zero λ_i^R if and only if she has a positive incentive to review the tasks, i.e., $f_i(x) > 0$. Otherwise, player i drops out without reviewing any task ($\lambda_i^R = 0$) and focuses solely on servicing of tasks ($\lambda_i^S = \mu_i^S$), thereby maximizing her expected utility given by $\tilde{u}_i = \mu_i^S r^S$.

In the following, we will refer to the above CPR game by $\Gamma = (\mathcal{N}, \{S_i\}_{i\in\mathcal{N}}, \{\tilde{u}_i\}_{i\in\mathcal{N}})$. In this paper, we are interested in equilibrium strategies for the players that constitute a PNE defined below.

Definition 2 (***Pure Nash Equilibrium***)**.** *A PNE is a strategy profile $\lambda^{R^*} = \{\lambda_i^{R^*}\}_{i\in\mathcal{N}} \in S$, such that for each player $i \in \mathcal{N}$, $\tilde{u}_i(\lambda_i^{R^*}, \lambda_{-i}^{R^*}) \geq \tilde{u}_i(\lambda_i^R, \lambda_{-i}^{R^*})$, for any $\lambda_i^R \in S_i$.*

Let $b_i : S_{-i} \mapsto S_i$ defined by

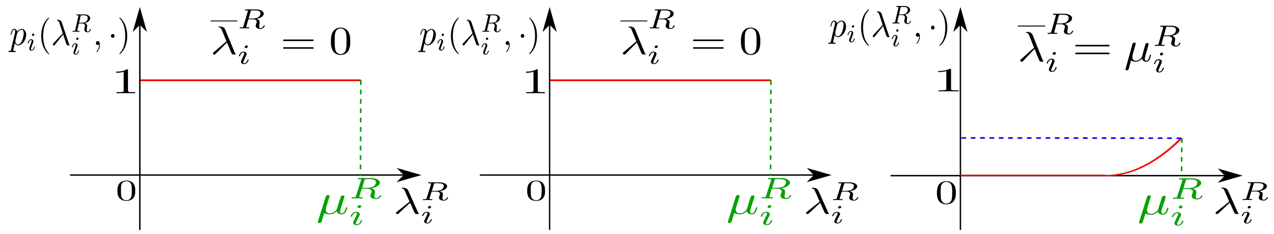
$$b_i(\lambda_{-i}^R) \in \underset{\lambda_i^R \in S_i}{\operatorname{argmax}}\ \tilde{u}_i(\lambda_i^R,\ \lambda_{-i}^R),$$

be a *best response* of player i to the review admission rates of other players λ_{-i}^R. A PNE exists if and only if there exists an invariant strategy profile, $\lambda^{R^*} = \{\lambda_i^{R^*}\}_{i\in\mathcal{N}} \in S$, such that $\lambda_i^{R^*} = b_i(\lambda_{-i}^{R^*})$, for each $i \in \mathcal{N}$.

7.2 Existence and Uniqueness of PNE

In this section, we study the existence and uniqueness of the PNE for the CPR game Γ under the system constraint (7.3). Each player $i \in \mathcal{N}$ chooses a review admission rate from her strategy set $S_i = [0,\ \mu_i^R]$ and receives an expected utility $\tilde{u}_i$ given by (7.8). For any given $\lambda_{-i}^R \in S_{-i}$, we obtain an upper bound $\overline{\lambda}_i^R : S_{-i} \mapsto S_i$ on λ_i^R defined by

$$\overline{\lambda}_i^R = \begin{cases} 0, & \text{if } \Lambda_i < 0, \\ \Lambda_i, & \text{if } 0 \leq \Lambda_i \leq \mu_i^R, \\ \mu_i^R, & \text{if } \Lambda_i > \mu_i^R, \end{cases}$$



(a) (b) (c)

Figure 7.2 Constraint probability of player i as $\overline{\lambda}_i^R$ varies from 0 to μ_i^R. a) For $\overline{\lambda}_i^R = 0$, $p_i(\lambda_i^R,\cdot) = 1,\ \forall \lambda_i^R \in S_i$, b) for $\overline{\lambda}_i^R \in (0,\ \mu_i^R)$, $p_i(\lambda_i^R,\cdot)$ is convex for $\lambda_i^R \in [0,\ \overline{\lambda}_i^R)$, with $p_i(\lambda_i^R,\cdot) \mapsto 1$ as $\lambda_i^R \mapsto \overline{\lambda}_i^R$, and $p_i(\lambda_i^R,\cdot) = 1,\ \forall \lambda_i^R \in [\overline{\lambda}_i^R,\ \mu_i^R]$, and c) for $\overline{\lambda}_i^R = \mu_i^R$, $p_i(\lambda_i^R,\cdot)$ is convex in λ_i^R and $p_i(\lambda_i^R,\cdot) < 1,\ \forall \lambda_i^R \in S_i$.

where $\Lambda_i := \frac{\mu_T^S - \sum_{j\in\mathcal{N}, j\neq i} a_j\lambda_j^R}{a_i}$, such that for $\lambda_i^R \in [0,\ \overline{\lambda}_i^R) \subset S_i$, constraint (7.3) is automatically satisfied, and for $\lambda_i^R \in (\overline{\lambda}_i^R,\ \mu_i^R] \subset S_i$, constraint (7.3) is violated. For $\lambda_i^R = \overline{\lambda}_i^R$, constraint (7.3) is satisfied if $\overline{\lambda}_i^R \in (0,\ \mu_i^R]$, and is violated if $\overline{\lambda}_i^R = 0$.

We study the properties of game Γ under following assumptions. Recall that $x = \lambda_T^S - \lambda_T^R = \mu_T^S - \sum_{i=1}^N a_i\lambda_i^R$.

(A1) For a given $\lambda_{-i}^R \in S_{-i},\ i \in \mathcal{N}$, we assume that the rate of return $r^R(\lambda_i^R,\cdot)$ for reviewing the tasks is continuously differentiable, strictly increasing and strictly concave for $\lambda_i^R \in S_i$, with $r^R(0,0) = 0$. Equivalently, $x \mapsto r^R(x)$ is continuously differentiable, strictly decreasing and strictly concave for $x \in [0,\ \mu_T^S]$, with $r^R(\mu_T^S) = 0$.

(A2) For a given $\lambda_{-i}^R \in S_{-i},\ i \in \mathcal{N}$, we assume that the constraint probability $p(\lambda_i^R,\cdot)$ is (i) continuous on S_i; (ii) is continuously differentiable, non-decreasing and convex for $\lambda_i^R \in (0,\ \overline{\lambda}_i^R) \subset S_i$; and (iii) is equal to 1, for $\lambda_i^R \in (\overline{\lambda}_i^R,\ \mu_i^R]$. See Fig. 7.2 for an illustration. Equivalently, $p(x)$ is continuously differentiable, non-increasing and convex for $x \in (0,\ \mu_T^S]$, and $p(x) \to 1$, as $x \to 0$. Furthermore, $p = 1$, for every $x < 0$.

(A3) We assume $f_i(\mu_i^R, 0) = r^R(\mu_i^R, 0)(1 - p(\mu_i^R, 0)) - h_i r^S > 0$, for each $i \in \mathcal{N}$, i.e., if no other player reviews any task, then each player i has a positive incentive to review tasks with maximum admission rate μ_i^R.

Remark 10. *The rate of return r^R and the constraint probability p can be easily designed to accommodate (A1-A3). Under Assumptions (A1) and (A2), the incentive function $f_i(\lambda_i^R, \cdot)$ is strictly concave in λ_i^R, which means for a fixed λ_{-i}^R, the player i has diminishing marginal incentive to review tasks. We make Assumption (A3) to provide positive incentives for players to review tasks with their maximum review admission rate μ_i^R, if no other player chooses to review any task. We can design game Γ to satisfy Assumption (A3) by ensuring that the following conditions hold:*

(i) $r^R(\mu_i^R, 0) > r^S$, *and* $\mu_i^S \leq \mu_i^R$, *for each* $i \in \mathcal{N}$, *and*

(ii) $\mu_i^R \ll \mu_T^S / a_i$, *or equivalently* $\sum_{j \in \mathcal{N},\ j \neq i} \mu_j^S \gg \mu_i^R$, *for each* $i \in \mathcal{N}$.

If the latter condition holds, then x is large, and consequently, the constraint probability $p(\mu_i^R, 0) \approx 0$, for each $i \in \mathcal{N}$. For most practical purposes, servicing a task requires more time than reviewing it, i.e., $\mu_i^S \leq \mu_i^R$. Therefore, condition (i) can be easily satisfied by designing rewards such that $r^R(\mu_i^R, 0) > r^S$, for each $i \in \mathcal{N}$. If the total service admission rate of all the players except player i is much higher than the maximum review admission rate of player i, i.e. $\sum_{j \in \mathcal{N},\ j \neq i} \mu_j^S \gg \mu_i^R$, for each $i \in \mathcal{N}$, then condition (ii) holds. Notice that for a large team of agents where a single agent does not have much impact on the overall service rate, condition (ii) is true. We refer the reader to Section 7.5 for an example. □

Theorem 6 (***Existence of PNE***)**.** *The CPR game Γ, under Assumptions (A1-A3), admits a PNE.*

Proof. See Appendix C: Chapter 7 for the proof. □

Let $f_i'(\lambda_i^R, \lambda_{-i}^R)$ be the first partial derivative of $f_i(\lambda_i^R, \lambda_{-i}^R)$ with respect to λ_i^R. We now provide a corollary that characterizes a PNE of CPR game Γ.

Corollary 1 (***PNE***)**.** *For the CPR game Γ, under Assumptions (A1-A3), the following statements hold for a PNE $\lambda^{R^*} = [\lambda_1^{R^*}, \ldots, \lambda_N^{R^*}]$ with $x^* = \mu_T^S - \sum_{i=1}^{N} a_i \lambda_i^{R^*}$:*

(i) $f_i'(\lambda_i^{R^*}, \lambda_{-i}^{R^*}) < 0$ *(or* $\frac{df_i}{dx}(x^*) > 0$*) for every player;*

(ii) $\lambda_i^{R^*} = 0$*, if and only if,* $f_i(\lambda_i^{R^*}, \lambda_{-i}^{R^*}) \leq 0$ *; and*

(iii) $\lambda_i^{R^*}$ *is non-zero and satisfies the following implicit equation if and only if* $f_i(\lambda_i^{R^*}, \lambda_{-i}^{R^*}) = f_i(x^*) > 0$ *at PNE, where*

$$\lambda_i^{R^*} = \min\left\{\tilde{\lambda}_i^{R^*},\ \mu_i^R\right\}, \tag{7.10}$$

with $\tilde{\lambda}_i^{R^*} = -\frac{f_i(\lambda_i^{R^*}, \lambda_{-i}^{R^*})}{f_i'(\lambda_i^{R^*}, \lambda_{-i}^{R^*})} = \frac{f_i(x^*)}{a_i \frac{df_i}{dx}(x^*)}$.

Proof. See Appendix C: Chapter 7 for the proof. □

Proposition 1 (***Structure of PNE***)**.** *For the CPR game* Γ *with players ordered in increasing order of* h_i*, let* $\lambda^{R^*} = [\lambda_1^{R^*}, \lambda_2^{R^*}, \ldots, \lambda_N^{R^*}]$ *be a PNE. Then, the following statements hold:*

(i) If, for any player k_1*,* $\lambda_{k_1}^{R^*} < \mu_{k_1}^R$*, then* $a_{k_1}\lambda_{k_1}^{R^*} \geq a_{k_2}\lambda_{k_2}^{R^*}$ *and* $\lambda_{k_1}^{R^*} \geq \lambda_{k_2}^{R^*}$*, for each* $k_2 > k_1$*; and*

(ii) if $\lambda_l^{R^*} = 0$*, for any* $l \in \mathcal{N}$*, then* $\lambda_i^R = 0$*, for each* $i \in \{j \in \mathcal{N} \mid j \geq l\}$*.*

Proof. See Appendix C: Chapter 7 for the proof. □

It follows from Proposition 1 that the review admission rate of a player i at a PNE is monotonically decreasing with the ratio h_i. Therefore, at a PNE, as the heterogeneity in terms of h_i among the players becomes very large, players with small (respectively, large) h_i review tasks with high (respectively, zero) review admission rate. We will show in Lemma 12 that the PNE shares these characteristics with the social welfare solution, which we define in Section 7.4. We illustrate this further in Section 7.5.

Theorem 7 (***Uniqueness of PNE***)**.** *The PNE admitted by the CPR game* Γ*, under assumptions (A1-A3), is unique.*

Proof. See Appendix C: Chapter 7 for the proof. □

7.3 Convergence to the Nash Equilibrium

We now show that the proposed CPR game Γ under Assumptions (A1-A3) belong to the class of *Quasi Aggregative games* [118] as defined below.

Definition 3 (***Quasi Aggregative game***)**.** *Consider a set of players $\mathcal{N}$, where each player $i \in \mathcal{N}$ has a strategy set S_i, and a utility function u_i. Let $S = \prod_{i\in\mathcal{N}} S_i$ be the joint strategy space of all the players, and $S_{-i} = \prod_{j\in\mathcal{N}, j\neq i} S_j$ be the joint strategy space of all the players except player i. A game $\Gamma = (\mathcal{N}, \{S_i\}_{i\in\mathcal{N}}, \{u_i\}_{i\in\mathcal{N}})$ is a quasi-aggregative game with aggregator $g : S \mapsto \mathbb{R}$, if there exists continuous functions $F_i : \mathbb{R} \times S_i \mapsto \mathbb{R}$ (the shift functions) and $\sigma_i : S_{-i} \mapsto X_{-i} \subseteq \mathbb{R}$, $i \in \mathcal{N}$ (the interaction functions) such that the utility functions u_i for each player $i \in \mathcal{N}$ can be written as:*

$$u_i(s) = \tilde{u}_i(\sigma_i(s_{-i}), s_i), \tag{7.11}$$

where $\tilde{u}_i : X_{-i} \times S_i \mapsto \mathbb{R}$, and

$$g(s) = F_i(\sigma_i(s_{-i}), s_i), \text{ for all } s \in S \text{ and } i \in \mathcal{N}. \tag{7.12}$$

An alternative, but less general way of defining a quasi-aggregative game replaces (7.11) *in the definition with:*

$$u_i(s) = \overline{u}_i(g(s), s_i), \tag{7.13}$$

where $\overline{u}_i : X \times S_i \mapsto \mathbb{R}$, and $X = \{g(s) \mid s \in S\} \subseteq \mathbb{R}$.

For the CPR game Γ, let $\sigma_i(\lambda^R_{-i}) = \sum_{j=1,j\neq i}^N a_j\lambda_j^R$ and $g(\lambda^R) = F_i(\sigma_i(\lambda^R_{-i}), \lambda_i^R) = \sum_{j=1,j\neq i}^N a_j\lambda_j^R + a_i\lambda_i^R$ be the interaction functions and shift functions, respectively. The expected utility $\tilde{u}_i$, which is defined in (7.8), can be re-written in the form

$$\tilde{u}_i(\lambda_i^R, \lambda_{-i}^R) = \tilde{u}_i(\sigma_i(\lambda_{-i}^R), \lambda_i^R). \tag{7.14}$$

Hence, the CPR game Γ is a quasi-aggregative game.

Specializing [118, Theorem 1] to the CPR game Γ, we obtain that if the best response for all the players is non-increasing in the interaction function $\sigma_i(\lambda^R_{-i}) = \sum_{j=1,j\neq i}^N a_j\lambda_j^R$, the CPR game Γ is a best response pseudo-potential game [119] as defined below.

Definition 4 (***Best response (pseudo)-potential game***)**.** *A game* $\Gamma = (\mathcal{N}, \{S_i\}_{i\in\mathcal{N}}, \{\tilde{u}_i\}_{i\in\mathcal{N}})$ *is a best response pseudo-potential game if there exists a continuous function* $\phi : S \mapsto \mathbb{R}$ *such that for every* $i \in \mathcal{N}$*,*

$$b_i(\lambda_{-i}^R) \supseteq \underset{\lambda_i^R \in S_i}{\operatorname{argmax}}\ \phi(\lambda_i^R,\ \lambda_{-i}^R),$$

where $b_i(\lambda_{-i}^R)$ *is the best response of player* i *to the review admission of other players* λ_{-i}^R*. Furthermore, if*

$$b_i(\lambda_{-i}^R) = \underset{\lambda_i^R \in S_i}{\operatorname{argmax}}\ \phi(\lambda_i^R,\ \lambda_{-i}^R),$$

then the game Γ *is a best response potential game.*

We now establish that the best response for each player is non-increasing in σ_i.

Lemma 11 (***Non-increasing best response***)**.** *For the CPR game* Γ*, under Assumptions (A1-A2), the best response mapping* $b_i(\lambda_{-i}^R)$ *is non-increasing in* $\sigma_i(\lambda_{-i}^R)$*, for each* $i \in \mathcal{N}$*, where* $\sigma_i(\lambda_{-i}^R) = \sum_{j=1,j\neq i}^{N} a_j \lambda_j^R$*.*

Proof. See Appendix C: Chapter 7 for the proof. □

Furthermore, Remark 1 in [78] states that a best response pseudo-potential game with a unique best response, is an instance of best response potential game [77]. Therefore, the CPR game Γ, with its unique (Lemma 16) and non-increasing best response b_i in $\sigma_i(\lambda_{-i}^R)$ (Lemma 11), is a best response potential game. Hence, simple best response dynamics such as sequential best response dynamics [78] and simultaneous best response dynamics [79] converge to the unique PNE.

7.4 Social Welfare and Inefficiency of PNE

In this section, we characterize the social welfare solution and provide analytic upper bounds on inefficiency measures for the PNE.

7.4.1 Social Welfare

Social welfare corresponds to the optimal (centralized) allocation by players with respect to a social welfare function. To characterize the effect of self-interested optimization of each agent, we compare the decentralized solution (PNE of the CPR game) with the centralized optimal solution (social welfare).

We choose a typical social welfare function $\Psi(\lambda^R) : S \mapsto \mathbb{R}$ defined by the sum of expected utility of all players, i.e.,

$$\begin{aligned}\Psi = \sum_{i=1}^{N} \tilde{u}_i &= \sum_{i=1}^{N} [\mu_i^S r^S + \lambda_i^R f_i(x)] \\ &= \mu_T^S r^S + \lambda_T^R r^R(x)(1-p(x)) - r^S \sum_{i=1}^{N} h_i \lambda_i^R \\ &= (\lambda_T^R + x) r^S + \lambda_T^R r^R(x)(1-p\,(x))\,. \end{aligned} \tag{7.15}$$

A *social welfare solution* is an optimal allocation that maximizes the social welfare function.

Lemma 12 (***Social welfare solution***)**.** *For the CPR game Γ with constraint $\sum_{i=1}^{N} a_i \lambda_i^R = c$, for any given $c \in \mathbb{R}_{\geq 0}$, and players ordered in increasing order of h_i, the associated social welfare solution, $\lambda^R \in S$ is given by:*

$$\lambda^R = \left[\mu_1^R,\ \mu_2^R, \ldots,\ \mu_{k-1}^R,\ \frac{1}{a_k}(c - \sum_{i=1}^{k-1} a_i \mu_i^R),\ 0, \ldots,\ 0\right],$$

where k is the smallest index such that $\sum_{i=1}^{k-1} a_i \mu_i^R \leq c < \sum_{i=1}^{k} a_i \mu_i^R$. Furthermore, since $\sum_{i=1}^{N} a_i \lambda_i^R \in [0,\ \mu_T^S + \mu_T^R]$, a bisection algorithm can be employed to compute optimal c and hence, the optimal social welfare solution.

Proof. Under the constraint $\sum_{i=1}^{N} a_i \lambda_i^R = c$ (equivalently, $x = \mu_T^S - c$), Ψ is a strictly increasing function of λ_T^R. Therefore, for a fixed $\sum_{i=1}^{N} a_i \lambda_i^R = c$, λ_T^R is maximized by selecting $k-1$ players with smallest a_i's (equivalently, h_i) to operate at their highest review admission rate, where the value of k is selected such that $\sum_{i=1}^{k-1} a_i \mu_i^R \leq c < \sum_{i=1}^{k} a_i \mu_i^R$. Finally, the k-th player in the ordered sequence is selected to operate at a review admission rate such that the constraint $\sum_{i=1}^{N} a_i \lambda_i^R = \sum_{i=1}^{k} a_i \lambda_i^R = c$, is satisfied. Therefore, the social welfare solution is of the form,

$$\lambda^R = \left[\mu_1^R,\ \mu_2^R, \ldots,\ \mu_{k-1}^R,\ \frac{1}{a_k}(c - \sum_{i=1}^{k-1} a_i \mu_i^R),\ 0, \ldots,\ 0\right].$$

Furthermore, for the function $r^R(x)$ and $p(x)$ satisfying Assumptions (A1-A2), Ψ is strictly concave in x, i.e., $\frac{\partial^2 \Psi}{\partial x^2} = \lambda_T^R \frac{d^2 f_i}{dx^2} < 0$ (Lemma 15). With the known form of the social welfare solution, the value of c, which corresponds to the unique maximizer x of Ψ, can be computed efficiently by employing a bisection algorithm [120]. □

7.4.2 Inefficiency of the PNE

We consider three measures of the inefficiency for the PNE: a) Price of Anarchy (PoA), b) Ratio of total review admission rate (η_{TRI}), and c) Ratio of Latency (η_{LI}), which are described by

$$PoA = \frac{(\Psi)_{SW}}{(\Psi)_{PNE}}, \ \eta_{TRI} = \frac{(\lambda_T^R)_{SW}}{(\lambda_T^R)_{PNE}}, \ \eta_{LI} = \frac{(\sum_{i=1}^N a_i \lambda_i^R)_{PNE}}{(\sum_{i=1}^N a_i \lambda_i^R)_{SW}},$$

respectively. While PoA is a widely used measure of the inefficiency, η_{TRI} and η_{LI} capture the inefficiency of the PNE based on the total review admission rate and the latency (inverse of throughput), respectively. Since incentivizing team collaboration is of interest, all three measures capture the inefficiency of the PNE well.

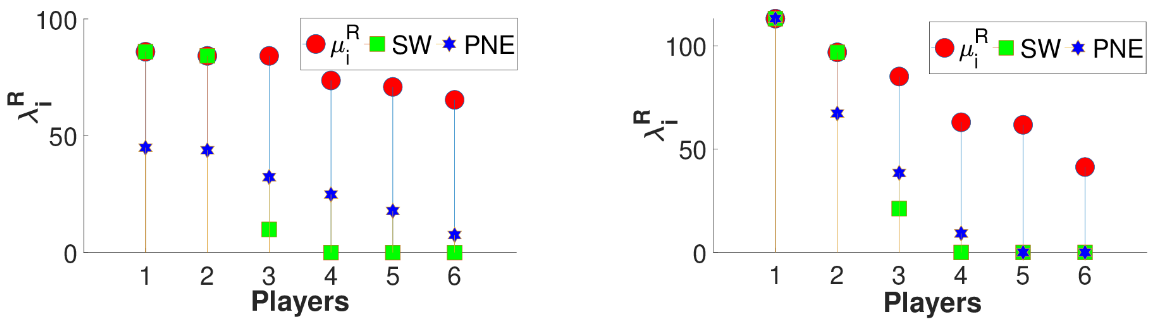
We now provide an analytic upper bound for each of these measures of inefficiency for the PNE. To this end, we assume that $\min_i\{\mu_i^S\} > \frac{\mu_T^S h_N}{N(1+h_N)}$. For scenarios wherein servicing a task requires much more time than reviewing it, i.e., $\mu_N^S \ll \mu_N^R$ ($h_N \to 0$), the assumption reduces to $\min_i\{\mu_i^S\} > 0$.

Theorem 8 (***Analytic bounds on PNE inefficiency***)**.** *For the CPR game Γ, under assumptions (A1-A3), and $\min_i\{\mu_i^S\} > \frac{\mu_T^S h_N}{N(1+h_N)}$, the inefficiency metrics for the PNE are upper bounded by*

$$PoA < \frac{\mu_T^S a_N}{\mu_T^S - \overline{x}}, \ \eta_{TRI} < \frac{\mu_T^S a_N}{(\mu_T^S - \overline{x}) a_1}, \ \eta_{LI} < \frac{\mu_T^S}{\mu_T^S - \overline{x}}, \tag{7.16}$$

where $\overline{x}$ is the unique maximizer of f_i, i.e., $\frac{df_i}{dx}(\overline{x}) = 0$.

Proof. See Appendix C: Chapter 7 for the proof. □



(a)

(b)

Figure 7.3 Social welfare solution (SW) and pure Nash equilibrium for a) low and b) high heterogeneity among players, respectively. Red circles show the maximum review admission rate (μ_i^R) for player i.

Example 1: We show analytic upper bounds on inefficiency measures for the PNE for a specific class of exponential functions $r^R(x) = A[1-\exp\{B(x-\mu_T^S)\}]$ and $p(x) = \exp(-Bx)$, where A and B are positive constants, and $x \in [0,\ \mu_T^S]$.

Setting $\frac{df_i}{dx}(\overline{x}) = 0$, we obtain $\overline{x} = \frac{\mu_T^S}{2}$. Using Theorem 8, we get PoA $< 2a_N, \eta_{TRI} < \frac{2a_N}{a_1}$, and $\eta_{LI} < 2$. For $\mu_N^S \ll \mu_N^R$, PoA $< 2a_N \to 2$, and $\eta_{TRI} < \frac{2a_N}{a_1} < 2a_N \to 2$.

7.5 Numerical Illustrations

In this section, we present numerical examples illustrating the uniqueness of PNE and the variation of inefficiency with increasing heterogeneity among the players.

In our numerical illustrations, we obtain the PNE by simulating the sequential best response dynamics of players with randomized initialization of their strategy. We verify the uniqueness of the PNE for different choices of functions, $r^R(x)$ and $p(x)$ satisfying Assumptions (A1-A2), and by following sequential best response dynamics with multiple random initializations for the strategy of each player. Furthermore, in our numerical simulations, we relax Assumption (A3) and still obtain a unique PNE.

An example illustration is shown in Fig. 7.3, where we show the social welfare solution (obtained using fmincon in MATLAB) and PNE for low and high heterogeneity in terms of variation in h_i among players, respectively. For our numerical illustrations, we choose the number of players, $N = 6$, and choose the functions $r^R(x)$ and $p(x)$, satisfying Assumptions

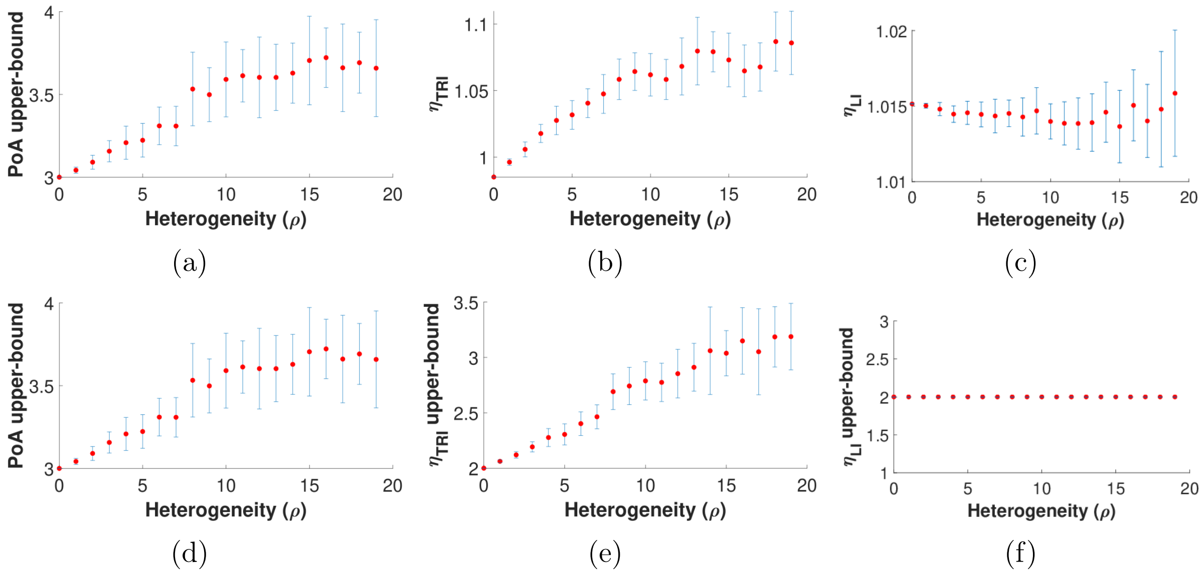


Figure 7.4 Empirical a) PoA, b) η_{TRI}, and c) η_{LI}, along with analytic upper bounds for d) PoA, e) η_{TRI}, and f) η_{LI} with increasing heterogeneity (ρ) among the agents.

(A1-A2) as following:

$$r^R(\lambda_i^R, \lambda_{-i}^R) = r^R(x) = 5[1 - \exp\{0.5(x - \mu_T^S)\}],$$

$$p_i^R(\lambda_i^R, \lambda_{-i}^R) = p(x) = \begin{cases} 1, & \text{if} \quad x \leq 0, \\ \exp(-0.5x), & \text{otherwise}, \end{cases}$$

where $x = \mu_T^S - \sum_{i=1}^N a_i \lambda_i^R$ is the slackness parameter. To characterize the heterogeneity among the players, we sample the player's maximum service admission rate μ_i^S and maximum review admission rate μ_i^R at random from normal distributions with fixed means, $M_{\mu_S} \in \mathbb{R}_{>0}$, and $M_{\mu_R} \in \mathbb{R}_{>0}$, and identical standard deviation, $\rho \in \mathbb{R}_{>0}$. We only consider realizations that satisfy $\mu_i^S \leq \mu_i^R$ for all the players, and hence, $h_i \leq 1$. For most practical purposes, where servicing a task requires much more time than reviewing it, the assumption $\mu_i^S \leq \mu_i^R$ holds true. Any non-positive realizations were discarded. We consider the standard deviation of the distributions as the measure of heterogeneity among the players.

Fig. 7.3 shows that in the social welfare solution, players with low ratio of h_i review the tasks at maximum review admission rate and players with high ratio of h_i drop out of the game. At PNE, the strategy profile of players follow the characteristics described by

Proposition 1. Lastly, with the increase in heterogeneity among the players, the PNE starts to approach the social welfare solution.

Fig. 7.4a-7.4c and Fig. 7.4d-7.4f shows the variation of different measures of inefficiency for PNE, and their corresponding analytic upper bounds (see Theorem 8), with increasing heterogeneity among the players. Fig. 7.4a shows the plot of PoA with increasing heterogeneity. In case of homogeneous players, i.e., $\rho = 0$, we obtain $PoA = 1$, which we establish in Lemma 19. As we initially increase the heterogeneity among the players, PNE starts to deviate from the social welfare solution, resulting in an increase in the PoA. We note that PoA ≤ 1.15, suggesting that the unique PNE is close to the optimal centralized social welfare solution. Fig. 7.4b and 7.4c shows η_{TRI} and η_{LI}, which are other relevant measures of inefficiency for our problem. It is evident from Fig. 7.4, that all three measures of inefficiency are close to 1, therefore suggesting near-optimal PNE solution.

7.6 Conclusions and Future Directions

We studied incentive design mechanisms to facilitate collaboration in a team of heterogeneous agents that is collectively responsible for servicing and subsequently reviewing a stream of homogeneous tasks. The heterogeneity among the agents is based on their skill-sets and is characterized by their mean service time and mean review time. To incentivize collaboration in the heterogeneous team, we designed a Common-Pool Resource (CPR) game with appropriate utilities and showed the existence of a unique PNE. We showed that the proposed CPR game is an instance of the best response potential game and by playing the sequential best response against each other, players converge to the unique PNE. We characterized the structure of the PNE and showed that at the PNE, the review admission rate of the players decreases with the increasing ratio of $h_i = \frac{\mu_i^S}{\mu_i^R}$, i.e., the review admission rate is higher for the players that are “better" at reviewing the tasks than servicing the tasks (characterized by their average service and review time). Furthermore, we consider three different inefficiency metrics for the PNE, including the Price of Anarchy (PoA), and provide an analytic upper bound for each metric. Additionally, we provide numerical evidence of their proximity to unity, i.e., the unique PNE is close to the optimal centralized social welfare solution.

There are several possible avenues of future research. It is of interest to extend the results for a broader class of games with less restrictive choice of utility functions, i.e., games that are not quasi-aggregative or commonly used games of weak strategic substitutes (WSTS) [78] or complements (WSTC) [78]. An interesting open problem is to consider a team of agents processing stream of heterogeneous tasks. In such a setting, incentivizing team collaboration based on the task-dependent skill-set of the agents is also of interest.

CHAPTER 8

CONCLUSIONS

In this dissertation, we focused on human-in-the-loop systems that suffer from inherent variability of human performance that depends on various factors such as cognitive state, task learning, expertise, and individual differences. We explored the optimal and game-theoretic feedback mechanisms for improving human performance in human-supervised autonomy. To this end, we first studied the problem of optimal fidelity selection for effective management of human cognitive resources. We assumed known models for human service time distribution and formulated the problem as a semi-Markov decision process (SMDP). We solved the SMDP to obtain the optimal fidelity selection policy and studied the structural properties of the optimal policy.

Next, we conducted human experiments on optimal fidelity selection to study the effect of the optimal policy on human performance. We assumed the human cognitive state as a hidden state and modeled the cognitive dynamics using an Input-Output Hidden Markov Model (IOHMM). We utilized the trained IOHMM model to formulate a Partially Observable Markov Decision Process (POMDP). We solved the POMDP to obtain the optimal fidelity selection policy and showed that the optimal policy significantly improves human performance.

Next, we extended the optimal fidelity selection problem by incorporating uncertainty into the human service-time distribution. We designed a robust and adaptive framework that accurately learns the human service-time model and adapts the policy while ensuring robustness under model uncertainty. However, a major challenge in designing adaptive and robust systems arises from the conflicting objectives of exploration and robustness. To mitigate system uncertainty, an agent must explore high-uncertainty state space regions, while robust policy optimizes worst-case performance and consequently avoids these regions. To address this trade-off, we introduced an efficient Deterministic Sequencing of Exploration and Exploitation (DSEE) algorithm for model-based reinforcement learning (RL). DSEE interleaves exploration and exploitation epochs with increasing lengths, resulting in sublinear cumulative regret growth over time.

Next, we focused on enhancing human performance through task learning and skill development. In this context, we studied the impact of evaluative feedback on human learning in sequential decision-making tasks. We conducted experiments on Amazon Mechanical Turk, where participants engaged with the Tower of Hanoi puzzle and received AI-generated feedback during their problem-solving. We examined how this feedback influenced their learning and skill transfer to related tasks. Additionally, we explored computational models to gain insights into how individuals integrate evaluative feedback into their decision-making processes.

Lastly, we expanded our focus from a single human operator to a team of heterogeneous agents, each with diverse skill sets and expertise. Within this context, we delved into the challenge of achieving efficient collaboration among heterogeneous team members to enhance overall system performance. Our approach leveraged a game theoretic framework, where we designed utility functions to incentivize decentralized collaboration among these agents. We showed the existence of a unique Pure Nash Equilibrium (PNE) and established the convergence of the best response dynamics to this unique PNE. Additionally, we established an analytical upper bound on measures of PNE inefficiency, shedding light on the effectiveness of our collaborative strategies.

BIBLIOGRAPHY

[1] P. Gupta, D. Isele, D. Lee, and S. Bae, "Interaction-aware trajectory planning for autonomous vehicles with analytic integration of neural networks into model predictive control," in *IEEE International Conference on Robotics and Automation*, pp. 7794–7800, 2023.

[2] I. R. Nourbakhsh, K. Sycara, M. Koes, M. Yong, M. Lewis, and S. Burion, "Human-robot teaming for search and rescue," *IEEE Pervasive Computing*, vol. 4, no. 1, pp. 72–79, 2005.

[3] M. A. Goodrich, J. L. Cooper, J. A. Adams, C. Humphrey, R. Zeeman, and B. G. Buss, "Using a mini-UAV to support wilderness search and rescue: Practices for human-robot teaming," in *IEEE International Workshop on Safety, Security and Rescue Robotics*, pp. 1–6, IEEE, 2007.

[4] A. Rosenfeld, N. Agmon, O. Maksimov, and S. Kraus, "Intelligent agent supporting human–multi-robot team collaboration," *Artificial Intelligence*, vol. 252, pp. 211–231, 2017.

[5] S. A. Seshia, D. Sadigh, and S. S. Sastry, "Formal methods for semi-autonomous driving," in *52nd Design Automation Conference*, pp. 1–5, IEEE, June 2015.

[6] P. Gupta, D. Isele, D. Lee, and S. Bae, "Interaction-aware trajectory planning for autonomous vehicles with analytic integration of neural networks into model predictive control," *arXiv preprint arXiv:2301.05393*, 2023.

[7] A. M. Okamura, "Methods for haptic feedback in teleoperated robot-assisted surgery," *Industrial Robot: An International Journal*, vol. 31, no. 6, pp. 499–508, 2004.

[8] J. Heard, C. E. Harriott, and J. A. Adams, "A survey of workload assessment algorithms," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 5, pp. 434–451, 2018.

[9] J. Peters, V. Srivastava, G. Taylor, A. Surana, M. P. Eckstein, and F. Bullo, "Human supervisory control of robotic teams: Integrating cognitive modeling with engineering design," *IEEE Control System Magazine*, vol. 35, no. 6, pp. 57–80, 2015.

[10] J. R. Peters, A. Surana, and F. Bullo, "Robust scheduling and routing for collaborative human/unmanned aerial vehicle surveillance missions," *Journal of Aerospace Information Systems*, pp. 1–19, 2018.

[11] K. Savla and E. Frazzoli, "A dynamical queue approach to intelligent task management for human operators," *Proceedings of the IEEE*, vol. 100, no. 3, pp. 672–686, 2012.

[12] V. Srivastava, R. Carli, C. Langbort, and F. Bullo, "Attention allocation for decision making queues," *Automatica*, vol. 50, no. 2, pp. 378–388, 2014.

[13] Q. Hu and W. Yue, *Markov Decision Processes with their Applications*, vol. 14. Springer Science & Business Media, 2007.

[14] M. Lin, R. J. La, and N. C. Martins, "Stabilizing a queue subject to activity-dependent server performance," *IEEE Transactions on Control of Network Systems*, vol. 8, no. 4, pp. 1579–1591, 2021.

[15] M. Lin, N. C. Martins, and R. J. La, "Queueing subject to action-dependent server performance: Utilization rate reduction," *arXiv preprint arXiv:2002.08514*, 2020.

[16] R. M. Yerkes and J. D. Dodson, "The relation of strength of stimulus to rapidity of habit-formation," *Journal of Comparative Neurology and Psychology*, vol. 18, no. 5, pp. 459–482, 1908.

[17] P. Gupta and V. Srivastava, "On robust and adaptive fidelity selection for human-in-the-loop queues," in *European Control Conference*, pp. 872–877, 2021.

[18] S. Stidham Jr and R. R. Weber, "Monotonic and insensitive optimal policies for control of queues with undiscounted costs," *Operations Research*, vol. 37, no. 4, pp. 611–625, 1989.

[19] L. I. Sennott, "Average cost semi-Markov decision processes and the control of queueing systems," *Probability in the Engineering and Informational Sciences*, vol. 3, no. 2, pp. 247–272, 1989.

[20] M. Agarwal, V. S. Borkar, and A. Karandikar, "Structural properties of optimal transmission policies over a randomly varying channel," *IEEE Transactions on Automatic Control*, vol. 53, no. 6, pp. 1476–1491, 2008.

[21] M. L. Puterman, "Markov decision processes," *Handbooks in Operations Research and Management Science*, vol. 2, pp. 331–434, 1990.

[22] R. Yang, S. Bhulai, and R. van der Mei, "Structural properties of the optimal resource allocation policy for single-queue systems," *Annals of Operations Research*, vol. 202, no. 1, pp. 211–233, 2013.

[23] C. C. White III and H. K. Eldeib, "Markov decision processes with imprecise transition probabilities," *Operations Research*, vol. 42, no. 4, pp. 739–749, 1994.

[24] S. Mannor, D. Simester, P. Sun, and J. N. Tsitsiklis, "Bias and variance approximation in value function estimates," *Management Science*, vol. 53, no. 2, pp. 308–322, 2007.

[25] V. Borkar and R. Jain, "Risk-constrained Markov decision processes," *IEEE Transactions on Automatic Control*, vol. 59, no. 9, pp. 2574–2579, 2014.

[26] R. T. Rockafellar and S. Uryasev, "Optimization of conditional value-at-risk," *Journal of Risk*, vol. 2, pp. 21–42, 2000.

[27] E. Delage and S. Mannor, "Percentile optimization for Markov decision processes with parameter uncertainty," *Operations Research*, vol. 58, no. 1, pp. 203–213, 2010.

[28] H. N. Nguyen, *Chance-Constrained Optimization: Applications in Game Theory and Markov Decision Processes.* PhD thesis, Université Paris-Saclay, 2023.

[29] W. Wiesemann, D. Kuhn, and B. Rustem, "Robust Markov decision processes," *Mathematics of Operations Research*, vol. 38, no. 1, pp. 153–183, 2013.

[30] L. F. Bertuccelli, A. Wu, and J. P. How, "Robust adaptive Markov decision processes: Planning with model uncertainty," *IEEE Control Systems Magazine*, vol. 32, no. 5, pp. 96–109, 2012.

[31] R. S. Sutton and A. G. Barto, *Reinforcement Learning, Second Edition: An Introduction.* MIT Press, Nov. 2018.

[32] G. N. Iyengar, "Robust dynamic programming," *Mathematics of Operations Research*, vol. 30, no. 2, pp. 257–280, 2005.

[33] A. Nilim and L. El Ghaoui, "Robust control of Markov decision processes with uncertain transition matrices," *Operations Research*, vol. 53, no. 5, pp. 780–798, 2005.

[34] V. Gullapalli and A. G. Barto, "Convergence of indirect adaptive asynchronous value iteration algorithms," in *Advances in Neural Information Processing Systems*, pp. 695–702, 1994.

[35] J. Xin, H. Zhao, D. Liu, and M. Li, "Application of deep reinforcement learning in mobile robot path planning," in *2017 Chinese Automation Congress*, pp. 7112–7116, IEEE, 2017.

[36] P. Gupta, D. Coleman, and J. E. Siegel, "Towards safer self-driving through great PAIN (Physically Adversarial Intelligent Networks)," *arXiv preprint arXiv:2003.10662*, 2020.

[37] A. G. Barto, P. S. Thomas, and R. S. Sutton, "Some recent applications of reinforcement learning," in *Proceedings of the Eighteenth Yale Workshop on Adaptive and Learning Systems*, 2017.

[38] S. Ferretti, S. Mirri, C. Prandi, and P. Salomoni, "Automatic web content personalization through reinforcement learning," *Journal of Systems and Software*, vol. 121, pp. 157–169, 2016.

[39] P. Gupta and V. Srivastava, "Structural properties of optimal fidelity selection policies for human-in-the-loop queues," *Automatica*, vol. 159, p. 111388, 2024. Extended version available at: arXiv preprint arXiv: 2201.09990.

[40] X. Zhou, H. Yue, T. Chai, and B. Fang, "Supervisory control for rotary kiln temperature based on reinforcement learning," in *Intelligent Control and Automation*, pp. 428–437, Springer, 2006.

[41] P. Gupta and V. Srivastava, "Optimal fidelity selection for human-in-the-loop queues using semi-Markov decision processes," in *American Control Conference*, pp. 5266–5271, 2019.

[42] L. F. Bertuccelli, *Robust Decision-Making with Model Uncertainty in Aerospace Systems*. PhD thesis, Massachusetts Institute of Technology, 2008.

[43] A. Nilim and L. El Ghaoui, *Robust Markov Decision Processes with Uncertain Transition Matrices*. PhD thesis, University of California, Berkeley, 2004.

[44] C. J. C. H. Watkins, *Learning from Delayed Rewards*. King's College, Cambridge United Kingdom, 1989.

[45] P. Gupta, D. Coleman, and J. E. Siegel, "Towards Physically Adversarial Intelligent Networks (PAINs) for safer self-driving," *IEEE Control Systems Letters*, vol. 7, pp. 1063–1068, 2022.

[46] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[47] A. L. Strehl and M. L. Littman, "An analysis of model-based interval estimation for Markov decision processes," *Journal of Computer and System Sciences*, vol. 74, no. 8, pp. 1309–1331, 2008.

[48] S. M. Kakade, *On the Sample Complexity of Reinforcement Learning*. PhD thesis, University College London, 2003.

[49] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," *Journal of Machine Learning Research*, vol. 11, no. 51, pp. 1563–1600, 2010.

[50] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge University Press, 2020.

[51] S. Vakili, K. Liu, and Q. Zhao, "Deterministic sequencing of exploration and exploitation for multi-armed bandit problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 5, pp. 759–767, 2013.

[52] H. Liu, K. Liu, and Q. Zhao, "Learning in a changing world: Restless multiarmed bandit with unknown dynamics," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1902–1916, 2013.

[53] L. Wei and V. Srivastava, "On abruptly-changing and slowly-varying multiarmed bandit problems," in *American Control Conference*, pp. 6291–6296, June 2018.

[54] N. Nayyar, D. Kalathil, and R. Jain, "On regret-optimal learning in decentralized multiplayer multiarmed bandits," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 1, pp. 597–606, 2016.

[55] L. Wei, A. McDonald, and V. Srivastava, "Multi-robot Gaussian process estimation and coverage: Deterministic sequencing algorithm and regret analysis," in *IEEE International Conference on Robotics and Automation*, pp. 9080–9085, 2021.

[56] C. Keser and R. Gardner, "Strategic behavior of experienced subjects in a common pool resource game," *International Journal of Game Theory*, vol. 28, no. 2, pp. 241–252, 1999.

[57] A. R. Hota, S. Garg, and S. Sundaram, "Fragility of the commons under prospect-theoretic risk attitudes," *Games and Economic Behavior*, vol. 98, pp. 135–164, 2016.

[58] A. R. Hota and S. Sundaram, "Controlling human utilization of failure-prone systems via taxes," *IEEE Transactions on Automatic Control*, vol. 66, no. 12, pp. 5772–5787, 2020.

[59] E. Ostrom, R. Gardner, J. Walker, and J. Walker, *Rules, Games, and Common-Pool Resources.* University of Michigan Press, 1994.

[60] P. Le Gall, "The theory of networks of single server queues and the tandem queue model," *International Journal of Stochastic Analysis*, vol. 10, no. 4, pp. 363–381, 1997.

[61] N. T. Thomopoulos, *Fundamentals of Queuing Systems: Statistical Methods for Analyzing Queuing Models.* Springer Science & Business Media, 2012.

[62] E. Altman, "Applications of dynamic games in queues," in *Advances in Dynamic Games*, pp. 309–342, Springer, 2005.

[63] L. Xia, "Service rate control of closed Jackson networks from game theoretic perspective," *European Journal of Operational Research*, vol. 237, no. 2, pp. 546–554, 2014.

[64] J. R. Marden and J. S. Shamma, "Game theory and control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, pp. 105–134, 2018.

[65] G. Arslan, J. R. Marden, and J. S. Shamma, "Autonomous vehicle-target assignment: A game-theoretical formulation," *Journal of Dynamic Systems Measurement and Control-Transactions of the ASME*, vol. 129, 09 2007.

[66] T. Başar and G. J. Olsder, *Dynamic Noncooperative Game Theory*, vol. 23. SIAM, 1999.

[67] T. Roughgarden, "Intrinsic robustness of the price of anarchy," in *Annual ACM Symposium on Theory of Computing*, pp. 513–522, 2009.

[68] J. R. Marden and T. Roughgarden, "Generalized efficiency bounds in distributed resource allocation," *IEEE Transactions on Automatic Control*, vol. 59, no. 3, pp. 571–584, 2014.

[69] L. Deori, K. Margellos, and M. Prandini, "Price of anarchy in electric vehicle charging control games: When Nash equilibria achieve social welfare," *Automatica*, vol. 96, pp. 150–158, 2018.

[70] D. Paccagnan, R. Chandan, and J. R. Marden, "Utility design for distributed resource allocation - Part I: Characterizing and optimizing the exact price of anarchy," *IEEE Transactions on Automatic Control*, pp. 1–1, 2019.

[71] P. Gupta and V. Srivastava, “Optimal fidelity selection for improved performance in human-in-the-loop queues for underwater search,” *arXiv preprint arXiv:2311.06381*, 2023.

[72] S. C. Kramer and H. W. Sorenson, “Bayesian parameter estimation,” *IEEE Transactions on Automatic Control*, vol. 33, no. 2, pp. 217–222, 1988.

[73] P. Gupta and V. Srivastava, “Deterministic sequencing of exploration and exploitation for reinforcement learning,” in *61st Conference on Decision and Control*, pp. 2313–2318, IEEE, 2022.

[74] P. Gupta, S. Biswas, and V. Srivastava, “Fostering human learning in sequential decision-making: Understanding the role of evaluative feedback,” *arXiv preprint arXiv:2311.03486*, 2023.

[75] P. Gupta, S. D. Bopardikar, and V. Srivastava, “Achieving efficient collaboration in decentralized heterogeneous teams using common-pool resource games,” in *58th Conference on Decision and Control*, pp. 6924–6929, IEEE, 2019.

[76] P. Gupta, S. D. Bopardikar, and V. Srivastava, “Incentivizing collaboration in heterogeneous teams via common-pool resource games,” *IEEE Transactions on Automatic Control*, vol. 68, no. 3, pp. 1902–1909, 2022.

[77] M. Voorneveld, “Best-response potential games,” *Economics Letters*, vol. 66, no. 3, pp. 289–295, 2000.

[78] P. Dubey, O. Haimanko, and A. Zapechelnyuk, “Strategic complements and substitutes, and potential games,” *Games and Economic Behavior*, vol. 54, no. 1, pp. 77–94, 2006.

[79] M. K. Jensen, “Stability of pure strategy Nash equilibrium in best-reply potential games,” *University of Birmingham, Tech. Rep*, 2009.

[80] R. P. Rao, *Brain-Computer Interfacing: An Introduction.* Cambridge University Press, 2013.

[81] O. Palinko, A. L. Kun, A. Shyrokov, and P. Heeman, “Estimating cognitive load using remote eye tracking in a driving simulator,” in *Symposium on Eye-tracking Research & Applications*, pp. 141–144, 2010.

[82] M. L. Littman, A. R. Cassandra, and L. P. Kaelbling, “Learning policies for partially observable environments: Scaling up,” in *Machine Learning Proceedings 1995*, pp. 362–370, Elsevier, 1995.

[83] M. T. Spaan, “Partially observable Markov decision processes,” in *Reinforcement Learning*, pp. 387–414, Springer, 2012.

[84] A. Diederich and J. R. Busemeyer, “Simple matrix methods for analyzing diffusion models of choice probability, choice response time, and simple response time,” *Journal of Mathematical Psychology*, vol. 47, no. 3, pp. 304–322, 2003.

[85] A. G. Barto and S. Mahadevan, "Recent advances in hierarchical reinforcement learning," *Discrete Event Dynamic Systems*, vol. 13, no. 1-2, pp. 41–77, 2003.

[86] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press, 2013.

[87] P. Gupta and V. Srivastava, "Structural properties of optimal fidelity selection policies for human-in-the-loop queues," *arXiv preprint arXiv:2201.09990*, 2022.

[88] P. Gupta, S. D. Bopardikar, and V. Srivastava, "Incentivizing collaboration in heterogeneous teams via common-pool resource games," *arXiv preprint arXiv: 1908.03938*, Aug. 2019.

[89] N. Koenig and A. Howard, "Design and use paradigms for Gazebo, an open-source multi-robot simulator," in *International Conference on Intelligent Robots and Systems*, vol. 3, pp. 2149–2154, IEEE, 2004.

[90] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "ROS: An open-source robot operating system," in *ICRA Workshop on Open Source Software*, vol. 3, p. 5, Kobe, Japan, 2009.

[91] Y. Bengio and P. Frasconi, "Input-output HMMs for sequence processing," *IEEE Transactions on Neural Networks*, vol. 7, no. 5, pp. 1231–1249, 1996.

[92] Y. Sakamoto, M. Ishiguro, and G. Kitagawa, "Akaike information criterion statistics," *Dordrecht, The Netherlands: D. Reidel*, vol. 81, no. 10.5555, p. 26853, 1986.

[93] A. A. Neath and J. E. Cavanaugh, "The Bayesian information criterion: Background, derivation, and applications," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 2, pp. 199–203, 2012.

[94] M. M. M. Manhães, S. A. Scherer, M. Voss, L. R. Douat, and T. Rauschenbach, "UUV simulator: A Gazebo-based package for underwater intervention and multi-robot simulation," in *OCEANS Monterey*, pp. 1–8, 2016.

[95] L. F. Bertuccelli and M. L. Cummings, "Operator choice modeling for collaborative uav visual search tasks," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 42, no. 5, pp. 1088–1099, 2012.

[96] V. Srivastava, P. Holmes, and P. Simen, "Explicit moments of decision times for single- and double-threshold drift-diffusion processes," *Journal of Mathematical Psychology*, vol. 75, no. 2016, pp. 96–109, 2016. Special Issue in Honor of R. Duncan Luce.

[97] K. P. Murphy, "Conjugate Bayesian analysis of the Gaussian distribution," tech. rep., University of British Columbia, BC, Oct 2007.

[98] L. Li and M. Littman, "Prioritized sweeping converges to the optimal value function," Tech. Rep. DCS-TR-631, Rutgers University, NJ, June 2008.

[99] S. Chib, "Markov chain Monte Carlo methods: computation and inference," *Handbook of Econometrics*, vol. 5, pp. 3569–3649, 2001.

[100] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge University Press, 2019.

[101] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. J. Weinberger, "Inequalities for the L_1 deviation of the empirical distribution," *Hewlett-Packard Labs, Tech. Rep*, 2003.

[102] A. Mastin and P. Jaillet, "Loss bounds for uncertain transition probabilities in Markov decision processes," in *51st IEEE Conference on Decision and Control*, pp. 6708–6715, IEEE, 2012.

[103] P. Gupta and V. Srivastava, "Deterministic sequencing of exploration and exploitation for reinforcement learning," *arXiv preprint arXiv: 2209.05408*, Sept. 2022.

[104] K.-M. Chung, H. Lam, Z. Liu, and M. Mitzenmacher, "Chernoff-Hoeffding bounds for Markov chains: Generalized and simplified," *arXiv preprint arXiv:1201.0559*, 2012.

[105] D. Aldous, L. Lovász, and P. Winkler, "Mixing times for uniformly ergodic Markov chains," *Stochastic Processes and their Applications*, vol. 71, no. 2, pp. 165–185, 1997.

[106] H. Chen and F. Zhang, "The expected hitting times for finite Markov chains," *Linear Algebra and its Applications*, vol. 428, no. 11-12, pp. 2730–2749, 2008.

[107] C. Szepesvári, *Algorithms for Reinforcement Learning.* Springer Nature, 2022.

[108] M. Lopes, F. Melo, and L. Montesano, "Active learning for reward estimation in inverse reinforcement learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 31–46, Springer, 2009.

[109] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning.," in *AAAI*, vol. 8, pp. 1433–1438, Chicago, IL, USA, 2008.

[110] S. Levine, Z. Popovic, and V. Koltun, "Nonlinear inverse reinforcement learning with Gaussian processes," *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[111] B. D. Ziebart, *Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy.* Carnegie Mellon University, 2010.

[112] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons, 1994.

[113] W. B. Knox and P. Stone, "Interactively shaping agents via human reinforcement: The TAMER framework," in *International Conference on Knowledge Capture*, pp. 9–16, 2009.

[114] W. B. Knox and P. Stone, "Augmenting reinforcement learning with human feedback," in *ICML Workshop on New Developments in Imitation Learning*, vol. 855, p. 3, 2011.

[115] W. B. Knox and P. Stone, "Reinforcement learning from simultaneous human and MDP reward.," in *AAMAS*, vol. 1004, pp. 475–482, Valencia, 2012.

[116] W. B. Knox and P. Stone, "Learning non-myopically from human-generated reward," in *International Conference on Intelligent User Interfaces*, pp. 191–202, 2013.

[117] C. G. Cassandras and S. Lafortune, *Introduction to Discrete Event Systems.* Springer Science & Business Media, 2009.

[118] M. K. Jensen, "Aggregative games and best-reply potentials," *Economic Theory*, vol. 43, no. 1, pp. 45–66, 2010.

[119] B. Schipper, "Pseudo-potential games," tech. rep., University of Bonn, Germany, 2004. Working paper.

[120] R. Burden and J. Faires, "The bisection method," *Numerical Analysis*, pp. 48–56, 2011.

[121] F. M. Dekking, C. Kraaikamp, H. P. Lopuhaä, and L. E. Meester, *A Modern Introduction to Probability and Statistics: Understanding Why and How.* Springer Science & Business Media, 2005.

[122] A. Gosavi, *Simulation-Based Optimization: Parametric Optimization Techniques and Reinforcement Learning.* Springer, 2015.

[123] D. G. Luenberger and Y. Ye, *Linear and Nonlinear Programming*, vol. 2. Springer, 1984.

[124] C. Berge, *Topological Spaces: Including a Treatment of Multi-valued Functions, Vector Spaces, and Convexity.* Courier Corporation, 1997.

APPENDIX A: CHAPTER 2

Proof of Lemma 3

Let w_k be the number of tasks that arrive during stage $k \in \{0, \ldots, n-1\}$ with sojourn time τ_k, in which the state transitions from $s_k = (q_k,\ \mathrm{cog}_k) \to s_{k+1} = (q_{k+1},\ \mathrm{cog}_{k+1})$ and action a_k is selected. Let a_k be an optimal action at state s_k and π be the corresponding optimal policy such that $a_k = \pi(s_k)$. The optimal policy π when applied from an initial state s_0 induces a sequence of states $< s_k >$ and sojourn times $< \tau_k >$ or ζ_{k+1}, where $\zeta_{k+1} = \sum_{j=0}^{k} \tau_j$ and $\zeta_0 = 0$.

Similarly, let $\tilde{s}_0 = (\tilde{q}_0, \mathrm{cog}_0)$ be another initial state with the same initial cognitive state, and $\tilde{q}_0 \geq q_0$. Apply a policy $\tilde{\pi}$ from the initial state $\tilde{s}_0$ such that $\tilde{\pi}(\overline{q}, \overline{\mathrm{cog}}) = \pi(\overline{q} + q_0 - \tilde{q}_0, \overline{\mathrm{cog}})$ for any $(\overline{q}, \overline{\mathrm{cog}})$. Note that $\tilde{a}_0 = a_0$. The optimal policy $\tilde{\pi}$ when applied from an initial state $\tilde{s}_0 = (\tilde{q}_0, \mathrm{cog}_0)$ induces a sequence of realizations $< \tilde{s}_k >$ and $< \tilde{\tau}_k >$. Since cognitive state and sojourn time are independent of the current queue length, for the same action sequence applied from the initial states s_0 and $\tilde{s}_0$, the random process associated with the evolution of cognitive state and sojourn time is almost surely the same except for the offset in the queue length. Hence, the probability of observing a sequence of realizations $< \tilde{s}_k = (\tilde{q}_k, \mathrm{cog}_k) >$, $< \tilde{a}_k >$ and $< \tilde{\tau}_k >$ when policy $\tilde{\pi}$ is applied from $\tilde{s}_0$ is equal to the probability of observing a sequence of realizations $< s_k = (q_k, \mathrm{cog}_k) >$, $< a_k >$ and $< \tau_k >$ when policy π is applied from s_0, where $\tilde{q}_k - \tilde{q}_0 = q_k - q_0$, $\tilde{a}_k = a_k$ and $\tilde{\tau}_k = \tau_k$. Therefore, it is easy to show that:

$$\mathbb{E}_{\tilde{\pi}}[\tilde{q}_k | \tilde{s}_0, \zeta_k] - \mathbb{E}_{\pi}[q_k | s_0, \zeta_k] = \tilde{q}_0 - q_0. \tag{8.1}$$

Note that the realization of sequence of actions $< a_k = \pi(s_k) >$, which are optimal for $< s_k = (q_k, \mathrm{cog}_k) >$ might be sub-optimal for $< \tilde{s}_k = (\tilde{q}_k, \mathrm{cog}_k) >$. Recall that $\mathbb{E}_{\pi}[\cdot]$ and $\mathbb{E}_{\tilde{\pi}}[\cdot]$ represents $\mathbb{E}[\cdot | s_0, \pi]$ and $\mathbb{E}[\cdot | \tilde{s}_0, \tilde{\pi}]$, respectively. Let $\Delta q := \tilde{q}_0 - q_0$, and $Z := V_n^*(q_0, \mathrm{cog}_0) - V_n^*(\tilde{q}_0, \mathrm{cog}_0)$. We first show the upper bound on Z.

$$\begin{aligned} Z &\leq V_n^*(q_0, \mathrm{cog}_0) - J_{n,\tilde{\pi}}(\tilde{q}_0, \mathrm{cog}_0) \\ &= \mathbb{E}_{\pi}\left[\sum_{k=0}^{n-1} \gamma^{\zeta_k} R(s_k, a_k) - \gamma^{\zeta_n} C q_n\right] - \mathbb{E}_{\tilde{\pi}}\left[\sum_{k=0}^{n-1} \gamma^{\zeta_k} R(\tilde{s}_k, \tilde{a}_k) - \gamma^{\zeta_n} C \tilde{q}_n\right] \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{\pi}\left[\sum_{k=0}^{n-1}\gamma^{\zeta_k}\{r(a_k) - c\,\mathbb{E}[\tau_k|\mathrm{cog}_k, a_k]q_k - \frac{c\lambda}{2}\,\mathbb{E}[\tau_k^2|\mathrm{cog}_k, a_k]\} - \gamma^{\zeta_n}Cq_n\right] \\
&- \mathbb{E}_{\tilde{\pi}}\left[\sum_{k=0}^{n-1}\gamma^{\zeta_k}\{r(\tilde{a}_k) - c\,\mathbb{E}[\tau_k|\mathrm{cog}_k, \tilde{a}_k]\tilde{q}_k - \frac{c\lambda}{2}\,\mathbb{E}[\tau_k^2|\mathrm{cog}_k, \tilde{a}_k]\} - \gamma^{\zeta_n}C\tilde{q}_n\right].
\end{aligned}
\tag{8.2}
$$

Using statements of Lemma 2, RHS of (8.2) is given by:

$$
\begin{aligned}
&\sum_{k=0}^{n-1}\mathbb{E}_{\pi}[\gamma^{\zeta_k}r(a_k)] - c\sum_{k=0}^{n-1}\mathbb{E}_{\pi}\left[\gamma^{\zeta_k}\tau_k\,\mathbb{E}_{\pi}\left[q_k|s_0,\zeta_k\right]\right] - \frac{c\lambda}{2}\sum_{k=0}^{n-1}\mathbb{E}_{\pi}\left[\gamma^{\zeta_k}\tau_k^2\right] - C\,\mathbb{E}_{\pi}\left[\gamma^{\zeta_n}\,\mathbb{E}_{\pi}\left[q_n|s_0,\zeta_n\right]\right] \\
&- \sum_{k=0}^{n-1}\mathbb{E}_{\tilde{\pi}}[\gamma^{\zeta_k}r(\tilde{a}_k)] + c\sum_{k=0}^{n-1}\mathbb{E}_{\tilde{\pi}}\left[\gamma^{\zeta_k}\tau_k\,\mathbb{E}_{\tilde{\pi}}\left[\tilde{q}_k|\tilde{s}_0,\zeta_k\right]\right] + \frac{c\lambda}{2}\sum_{k=0}^{n-1}\mathbb{E}_{\tilde{\pi}}\left[\gamma^{\zeta_k}\tau_k^2\right] + C\,\mathbb{E}_{\tilde{\pi}}\left[\gamma^{\zeta_n}\,\mathbb{E}_{\tilde{\pi}}\left[\tilde{q}_n|\tilde{s}_0,\zeta_n\right]\right] \\
&\overset{(4)^*}{=} c\sum_{k=0}^{n-1}\mathbb{E}_{\pi}\left[\gamma^{\zeta_k}\tau_k\{\mathbb{E}_{\tilde{\pi}}[\tilde{q}_k\ |\tilde{s}_0,\zeta_k] - \mathbb{E}_{\pi}[q_k\ |s_0,\zeta_k]\}\right] + C\,\mathbb{E}_{\pi}\left[\gamma^{\zeta_n}\{\mathbb{E}_{\tilde{\pi}}[\tilde{q}_n\ |\tilde{s}_0,\zeta_k] - \mathbb{E}_{\pi}[q_n\ |s_0,\zeta_k]\}\right],
\end{aligned}
\tag{8.3}
$$

where $(4)^*$ follows by recalling that the probability of observing a sequence of realizations $< \tilde{s}_k = (\tilde{q}_k, \mathrm{cog}_k) >$, $< \tilde{a}_k >$ and $< \tilde{\tau}_k >$ when policy $\tilde{\pi}$ is applied from $\tilde{s}_0$ is equal to the probability of observing a sequence of realizations $< s_k = (q_k, \mathrm{cog}_k) >$, $< a_k >$ and $< \tau_k >$ when policy π is applied from s_0, where $\tilde{q}_k - \tilde{q}_0 = q_k - q_0$, $\tilde{a}_k = a_k$ and $\tilde{\tau}_k = \tau_k$. Substituting (8.1) in (8.3), we get,

$$
\begin{aligned}
Z &\leq \left\{c\sum_{k=0}^{n-1}\mathbb{E}_{\pi}\left[\gamma^{\zeta_k}\tau_k\right] + C\,\mathbb{E}_{\pi}\left[\gamma^{\zeta_n}\right]\right\}\Delta q \\
&\overset{(5)^*}{=} \left\{c\sum_{k=0}^{n-1}\mathbb{E}_{\pi}\left[\gamma^{\zeta_k}\right]\mathbb{E}_{\pi}\left[\tau_k\right] + C\,\mathbb{E}_{\pi}\left[\gamma^{\zeta_n}\right]\right\}\Delta q \\
&\leq \left\{ct_{\max}\sum_{k=0}^{n-1}\rho^k + C\rho^n\right\}\Delta q,
\end{aligned}
\tag{8.4}
$$

where $(5)^*$ follows due to independence of $\zeta_k = \sum_{i=0}^{k-1}\tau_k$ and τ_k. Taking the infinite time limit in (8.4), we get, $V^*(q_0, \mathrm{cog}_0) - V^*(\tilde{q}_0, \mathrm{cog}_0) \leq \frac{ct_{\max}\Delta q}{1-\rho}$, $\tilde{q}_0 \geq q_0$.

We now show the lower bound on Z. Let $\tilde{a}_k$ be optimal for $\tilde{s}_k = (\tilde{q}_k, \mathrm{cog}_k)$, and choose $< a_k > = < \tilde{a}_k >$ for the sequence $< s_k = (q_k, \mathrm{cog}_k) >$, where $\tilde{s}_0 = (\tilde{q}_0, \mathrm{cog}_0)$ and $s_0 = (q_0, \mathrm{cog}_0)$. Note that (8.1) still holds. Analogous to (8.3), Z is lower-bounded by:

$$Z \geq J_{n,\pi}(q_0, \mathrm{cog}_0) - V_n^*(\tilde{q}_0, \mathrm{cog}_0)$$

$$= c\sum_{k=0}^{n-1}\mathbb{E}_{\tilde{\pi}}\left[\gamma^{\tilde{\zeta}_k}\tilde{\tau}_k\left\{\mathbb{E}_{\tilde{\pi}}[\tilde{q}_k|\tilde{s}_0,\tilde{\zeta}_k]-\mathbb{E}_{\pi}[q_k|s_0,\tilde{\zeta}_k]\right\}\right]+C\,\mathbb{E}_{\tilde{\pi}}\left[\gamma^{\tilde{\zeta}_n}\left\{\mathbb{E}_{\tilde{\pi}}[\tilde{q}_n|\tilde{s}_0,\tilde{\zeta}_k]-\mathbb{E}_{\pi}[q_n|s_0,\tilde{\zeta}_k]\right\}\right]. \tag{8.5}$$

Substituting (8.1) in (8.5), we get,

$$\begin{aligned} Z &\geq \left\{c\sum_{k=0}^{n-1}\mathbb{E}_{\tilde{\pi}}\left[\gamma^{\tilde{\zeta}_k}\right]\mathbb{E}_{\tilde{\pi}}\left[\tilde{\tau}_k\right]+C\,\mathbb{E}_{\tilde{\pi}}\left[\gamma^{\tilde{\zeta}_n}\right]\right\}\Delta q \\ &\overset{(6)^*}{\geq} \left\{ct_s\sum_{k=0}^{n-1}\gamma^{\mathbb{E}_{\tilde{\pi}}[\tilde{\zeta}_k]}+C\gamma^{\mathbb{E}_{\tilde{\pi}}[\tilde{\zeta}_n]}\right\}\Delta q \\ &\geq \left\{\frac{(1-\gamma^{nt_{\max}})ct_s}{1-\gamma^{t_{\max}}}+\gamma^{nt_{\max}}C\right\}\Delta q, \end{aligned} \tag{8.6}$$

where $(6)^*$ follows by applying Jensen's inequality [121] ($\mathbb{E}[g(x)] \geq g(\mathbb{E}[x])$) on the convex function $g(x) = \gamma^x$. Taking the infinite time limit in (8.6), we get, $0 \leq \frac{ct_s\Delta q}{1-\gamma^{t_{\max}}} \leq V^*(q_0, \mathrm{cog}_0) - V^*(\tilde{q}_0, \mathrm{cog}_0),\ \tilde{q}_0 \geq q_0$. □

Proof of Lemma 4

Proof. We start by proving the first statement. In the following, we find conditions under which if action N is the optimal choice at queue length q_1 for a given cognitive state $\mathrm{cog} \leq \mathrm{cog}^*$, then for all $q_2 > q_1$, N dominates H. Let N be the optimal action in state $s_1 = \{q_1, \mathrm{cog}\}$. Let $F(s, a)$ denote the expected future rewards received in state s for taking action a (the second term in the Bellman equation (2.3)). Then, we have

$$\begin{aligned} &R(s_1,N)-R(s_1,H)+F(s_1,N)-F(s_1,H)>0, \\ \Longrightarrow\ &M+\sum_{\tau}\sum_{\mathrm{cog}'}\sum_{\overline{q}}\gamma^{\tau}\mathtt{Pois}(\overline{q}|\tau)V^*(\mathrm{cog}',q_1+\overline{q}-1)(\mathbb{P}(\mathrm{cog}',\tau|\mathrm{cog},N)-\mathbb{P}(\mathrm{cog}',\tau|\mathrm{cog},H))>0, \end{aligned} \tag{8.7}$$

where $M := c(\mathbb{E}[\tau|\mathrm{cog}, H] - \mathbb{E}[\tau|\mathrm{cog}, N])q_1 + r_N - r_H + \frac{c\lambda}{2}(\mathbb{E}[\tau^2|\mathrm{cog}, H] - \mathbb{E}[\tau^2|\mathrm{cog}, N])$, and $\mathbb{P}(q_1 + \overline{q} - 1|q_1, \tau)$ is replaced by $\mathtt{Pois}(\overline{q}|\tau)$, which is the Poisson probability of $\overline{q}$ arrivals during service time τ.

Now for the state $s_2 = \{q_2, \mathrm{cog}\}$, with $q_2 > q_1$ and identical cog, under the assumption (A1) we show that:

$$R(s_2, N) - R(s_2, H) + F(s_2, N) - F(s_2, H) > 0. \tag{8.8}$$

The left-hand side of (8.8) is given by:

$$X+M+\sum_{\tau}\sum_{\text{cog}'}\sum_{\overline{q}}\gamma^{\tau}\texttt{Pois}(\overline{q}|,\tau)V^{*}(\text{cog}',q_2+\overline{q}-1)(\mathbb{P}(\text{cog}',\tau|\text{cog},N)-\mathbb{P}(\text{cog}',\tau|\text{cog},H)), \tag{8.9}$$

where $X := c(\mathbb{E}[\tau|\text{cog},H]-\mathbb{E}[\tau|\text{cog},N])(q_2-q_1)$. To show (8.8), we prove that the difference between LHS of (8.8) and (8.7) is positive. Subtracting LHS of (8.7) from (8.9), we get:

$$X-\sum_{\tau}\sum_{\text{cog}'}\sum_{\overline{q}}\gamma^{\tau}\texttt{Pois}(\overline{q}|\tau)V_D(\mathbb{P}(\text{cog}',\tau|\text{cog},N)-\mathbb{P}(\text{cog}',\tau|\text{cog},H)), \tag{8.10}$$

where $V_D := \Big[V^{*}(\text{cog}',q_1+\overline{q}-1)-V^{*}(\text{cog}',q_2+\overline{q}-1)\Big]$. From Lemma 3, we know that

$$0 \leq \beta := \frac{ct_s(q_2-q_1)}{1-\gamma^{t_{\max}}} \leq V_D \leq \frac{ct_{\max}(q_2-q_1)}{1-\rho} =: \alpha.$$

Therefore, (8.10) is lower bounded by

$$X+\beta\sum_{\tau}\sum_{\text{cog}'}\sum_{\overline{q}}\gamma^{\tau}\texttt{Pois}(\overline{q}|\tau)\mathbb{P}(\text{cog}',\tau|\text{cog},H)-\alpha\sum_{\tau}\sum_{\text{cog}'}\sum_{\overline{q}}\gamma^{\tau}\texttt{Pois}(\overline{q}|\tau)\mathbb{P}(\text{cog}',\tau|\text{cog},N)$$
$$\geq X+\beta\gamma^{\mathbb{E}[\tau|\text{cog},H]}-\alpha\,\mathbb{E}[\gamma^{\tau}|\text{cog},N], \tag{8.11}$$

where we utilized Jensen's inequality on convex function γ^{τ} to obtain $\mathbb{E}[\gamma^{\tau}|\text{cog},H] \geq \gamma^{\mathbb{E}[\tau|\text{cog},H]}$. (8.11) is nonnegative when the condition in the first statement holds.

Now we prove the second statement. Using a similar analysis it can be shown that if action S is the optimal choice at queue length q_1 for a given cognitive state $\text{cog} \leq \text{cog}^{*}$, then for every $q_2 > q_1$, S dominates H and N, respectively, under the following conditions:

$$\mathbb{E}[\tau|\text{cog},H]-t_s+\frac{t_s\gamma^{\mathbb{E}[\tau|\text{cog},H]}}{1-\gamma^{t_{\max}}} \geq \gamma^{t_s}\frac{t_{\max}}{1-\rho}, \tag{8.12}$$

$$\mathbb{E}[\tau|\text{cog},N]-t_s+\frac{t_s\gamma^{\mathbb{E}[\tau|\text{cog},N]}}{1-\gamma^{t_{\max}}} \geq \gamma^{t_s}\frac{t_{\max}}{1-\rho}, \tag{8.13}$$

respectively, where we have used $\mathbb{E}[\tau|\text{cog},S] = t_s$ and $\mathbb{E}[\gamma^{\tau}|\text{cog},S] = \gamma^{t_s}$ due to constant time of skip. Since $\mathbb{E}[\tau|\text{cog},H] > \mathbb{E}[\tau|\text{cog},N]$, (8.12)-(8.13) can be combined to obtain the condition in the second statement under which action S dominates both H and N. □

APPENDIX B: CHAPTER 4

Proof of Theorem 2

A key challenge we address is the time-dependence of the asynchronous adaptive Bellman updates that adapt to the latest estimates of the service-time distribution. Using Lemmas 13 and 14, we upper-bound the difference between optimal value functions for intermediate SMDPs at subsequent time steps, and the optimal value function for the uncertainty-free SMDP, which are used to establish the convergence result.

Theorem 9 (adapted from [122, Chapter 10])**.** *T is a contraction mapping and therefore, there exists a unique fixed point satisfying $T(V^*) = V^*$, where V^* is the optimal value function for the uncertainty-free SMDP Γ.*

Let V_t^* be the optimal value function for SMDP $\hat{\Gamma}_t$ defined by the estimates $\hat{\mathbb{P}}_t$ and $\hat{R}_t$. Therefore, $V_t^* = \hat{T}_t(V_t^*)$. Let $\|\cdot\|$ be the max-norm given by $\|v\| = \max\{|v_1|, |v_2| \dots, |v_n|\}$, for any vector $v = (v_1, v_2, \dots, v_n)$. Let $\tau_{\min} \geq 1$ be the minimum number of time steps spent in any state $s \in \mathcal{S}$, for any action $a \in \mathcal{A}_{\mathcal{S}}$.

Lemma 13. *For any state $s \in \mathcal{S}$, following statements hold:*

(i) $|\hat{T}_t(V_1(s)) - \hat{T}_t(V_2(s))| \leq \gamma^{\tau_{\min}} \|V_1 - V_2\|$ *for any $s \in \mathcal{S}$, where the Bellman operator $\hat{T}_t$ at any time t is applied on value function estimates V_1 and V_2.*

(ii) $|V_{t+1}(s) - V_t^*(s)| \leq \gamma^{\tau_{\min}} \|V_t - V_t^*\|,$ *if $s \in B_t$.*

Proof. We prove the first statement. For any $s \in \mathcal{S}$, let $\mathcal{V} := |\hat{T}_t(V_1(s)) - \hat{T}_t(V_2(s))|$. Therefore,

$$\begin{aligned}\mathcal{V} &\leq \max_{a \in \mathcal{A}_{\mathcal{S}}} \left| \sum_{\tau} \sum_{s'} \gamma^{\tau} \hat{\mathbb{P}}_t(s', \tau | s, a)(V_1(s') - V_2(s')) \right| \\ &\leq \|V_1 - V_2\| \sum_{\tau} \gamma^{\tau} \hat{\mathbb{P}}_t(\tau | s, a) \leq \gamma^{\tau_{\min}} \|V_1 - V_2\|.\end{aligned}$$

The second statement follows from the first by noting that $|V_{t+1}(s) - V_t^*(s)| = |\hat{T}_t(V_t) - \hat{T}_t(V_t^*)|$ if $s \in B_t$. □

Lemma 14. *Under Assumptions A1-A5, for any given $\epsilon > 0$, there exists a time $\tilde{t}$, such that for any $t \geq \tilde{t}$, (i) $\|V_t^* - V^*\| \leq \epsilon$, and (ii) $\|V_{t+1}^* - V_t^*\| \leq 2\epsilon$ with probability 1.*

Proof. Since the estimates $\hat{\mathbb{P}}_t$ and $\hat{R}_t$ are assumed to be bounded at any time t (assumption (A2)), the value function estimate V_t at any time t also remains bounded. Furthermore, using assumption (A3), $\hat{\mathbb{P}}_t$ and $\hat{R}_t$ converge to their true values $\mathbb{P}$ and R, respectively, with probability 1, i.e, for any $\epsilon_1, \epsilon_2 > 0$, there exists a time $\tilde{t_0}$ such that, for any $t \geq \tilde{t_0}$, $|\hat{R}_t - R| \leq \epsilon_1$ and $|\hat{p_t^{ij}} - p^{ij}| \leq \epsilon_2$, where $\hat{p_t^{ij}}$ and p^{ij} are the elements of $\hat{\mathbb{P}}_t$ and $\mathbb{P}$, respectively. In asynchronous VI, the value of only the states $s \in B_t \subseteq \mathcal{S}$ are updated at any time t, however, each state is assumed to be updated infinitely often. Therefore, the sequence (4.6) converges with probability 1, i.e., for any $\epsilon_3 > 0$, there exists a time $\tilde{t_1} \geq \tilde{t_0}$ such that $\|V_{t+1} - V_t\| \leq \epsilon_3$, for $t \geq \tilde{t_1}$.

Consider $s \in B_t$ such that $V_{t+1}(s) = \hat{T}_t(V_t(s))$. Let $\mathcal{V}_t^* := \|V_t^* - V^*\|$, where we only consider the states $s \in B_t$ in the vectors V_t^* and V^*. Therefore,

$$\mathcal{V}_t^* \leq \|V_t^* - V_{t+1}\| + \|V_{t+1} - V^*\| =: \mathcal{Z}_1 + \mathcal{Z}_2. \tag{8.14}$$

$\mathcal{Z}_1 = \|V_t^* - V_{t+1}\|$ is upper bounded by:

$$\mathcal{Z}_1 \leq \|V_t^* - \hat{T}_t(V_{t+1})\| + \|\hat{T}_t(V_{t+1}) - \hat{T}_t(V_t)\| =: \mathcal{Z}_1^1 + \mathcal{Z}_1^2. \tag{8.15}$$

Since $\mathcal{Z}_1^1 = \|V_t^* - \hat{T}_t(V_{t+1})\| = \|\hat{T}_t(V_t^*) - \hat{T}_t(V_{t+1})\|$, from statement (i) of Lemma 13, we get

$$\mathcal{Z}_1^1 \leq \gamma^{\tau_{\min}} \|V_t^* - V_{t+1}\|, \text{ and } \mathcal{Z}_1^2 \leq \gamma^{\tau_{\min}} \|V_{t+1} - V_t\|. \tag{8.16}$$

Substituting (8.16) in (8.15), we get:

$$\mathcal{Z}_1 \leq \frac{\gamma^{\tau_{\min}}}{1 - \gamma^{\tau_{\min}}} \|V_{t+1} - V_t\|. \tag{8.17}$$

$\mathcal{Z}_2 = \|V^* - V_{t+1}\|$ is upper bounded by:

$$\mathcal{Z}_2 \leq \|V^* - T(V_{t+1})\| + \|T(V_{t+1}) - V_{t+1}\| =: \mathcal{Z}_2^1 + \mathcal{Z}_2^2. \tag{8.18}$$

Again using statement (i) of Lemma 13, we have

$$\mathcal{Z}_2^1 \leq \gamma^{\tau_{\min}} \|V^* - V_{t+1}\| = \gamma^{\tau_{\min}} \mathcal{Z}_2. \tag{8.19}$$

Furthermore, $\mathcal{Z}_2^2 = \|T(V_{t+1}) - V_{t+1}\| = \|T(V_{t+1}) - \hat{T}_t(V_t)\|$ is upper bounded by:

$$\mathcal{Z}_2^2 \leq \max_{a\in\mathcal{A}_\mathcal{S}} \|R - \hat{R}_t\| + \max_{a\in\mathcal{A}_\mathcal{S}} \Big\| \sum_\tau \sum_{s'} \gamma^\tau \mathbb{P}(s', \tau|s, a) V_{t+1}(s') - \sum_\tau \sum_{s'} \gamma^\tau \hat{\mathbb{P}}_t(s', \tau|s, a) V_t(s') \Big\|. \tag{8.20}$$

Recall that for any $t \geq \tilde{t}_0$, $|\hat{R}_t - R| \leq \epsilon_1$ and $|\hat{p}_t^{ij} - p^{ij}| \leq \epsilon_2$. Therefore, for $t \geq \tilde{t}_0$, (8.20) is upper bounded by:

$$\begin{aligned}
\mathcal{Z}_2^2 &\leq \epsilon_1 + \max_{a\in\mathcal{A}_\mathcal{S}} \Big\| \sum_\tau \sum_{s'} \gamma^\tau |\mathbb{P}(s', \tau|s, a) - \hat{\mathbb{P}}_t(s', \tau|s, a)| V_{t+1}(s') \Big\| \\
&\quad + \Big\| \sum_\tau \sum_{s'} \gamma^\tau \hat{\mathbb{P}}_t(s', \tau|s, a)(V_{t+1}(s') - V_t(s')) \Big\| \\
&\overset{(1^*)}{\leq} \epsilon_1 + \|V_{t+1}\| \sum_\tau \sum_{s'} \gamma^\tau \epsilon_2 + \gamma^{\tau_{\min}} \|V_{t+1} - V_t\|, \\
&= \epsilon_1 + \frac{\epsilon_2 N \gamma^{\tau_{\min}} \|V_{t+1}\|}{1-\gamma} + \gamma^{\tau_{\min}} \|V_{t+1} - V_t\|,
\end{aligned} \tag{8.21}$$

where N is the size of the finite state-space $\mathcal{S}$, and (1^*) follows from $|\hat{p}_t^{ij} - p^{ij}| \leq \epsilon_2$ and statement (i) of Lemma 13. Substituting (8.19) and (8.21) in (8.18), we get:

$$\mathcal{Z}_2 \leq \frac{\gamma^{\tau_{\min}}}{1-\gamma^{\tau_{\min}}} \|V_{t+1} - V_t\| + f(\|V_{t+1}\|), \tag{8.22}$$

where $f(\|V_{t+1}\|) := \frac{1}{1-\gamma^{\tau_{\min}}} \left(\epsilon_1 + \frac{\epsilon_2 N \gamma^{\tau_{\min}} \|V_{t+1}\|}{1-\gamma}\right)$ is bounded for bounded $\|V_{t+1}\|$ and $f(\|V_{t+1}\|) \mapsto 0$, when $\epsilon_1, \epsilon_2 \mapsto 0$. Substituting (8.17) and (8.22) in (8.14), we get

$$\mathcal{V}_t^* \leq \frac{2\gamma^{\tau_{\min}}}{1-\gamma^{\tau_{\min}}} \|V_{t+1} - V_t\| + f(\|V_{t+1}\|). \tag{8.23}$$

Recall that for any $\epsilon_3 > 0$, there exists $\tilde{t}_1 \geq \tilde{t}_0$ such that $|V_{t+1} - V_t\| \leq \epsilon_3$, for $t \geq \tilde{t}_1$. Choosing $\epsilon_1, \epsilon_2 \mapsto 0$, and $\epsilon_3 = \frac{\epsilon(1-\gamma^{\tau_{\min}})}{2\gamma^{\tau_{\min}}}$, we get that there exists $\tilde{t}_1$, such that $\|V_{t+1} - V_t\| \leq \frac{\epsilon(1-\gamma^{\tau_{\min}})}{2\gamma^{\tau_{\min}}}$, and $\mathcal{V}_t^* = \|V_t^* - V^*\| < \epsilon$, for any $t \geq \tilde{t}_1$.

Recall that $\mathcal{V}_t^*$ only considers states $s \in B_t$. However, since each state is updated infinitely often, there exists $\tilde{t}$ such that $\|V_t^* - V^*\| < \epsilon$, for any $t \geq \tilde{t}$. Furthermore, for $t \geq \tilde{t}$, $\|V_{t+1}^* - V_t^*\| \leq \|V_{t+1}^* - V^*\| + \|V_t^* - V^*\| \leq 2\epsilon$. □

Proof of Theorem 2: For any state $s \in \mathcal{S}$, define a sequence $\{t_i^s\}_{i=1}^{\infty}$ of times at which state s is updated by the asynchronous VI, and consider the updates after time $\tilde{t}$, i.e., consider the sequence $\{t_i^s\}_{i=k}^{\infty}$ such that $t_k^s \geq \tilde{t}$. Let $\mathcal{V}_t(s) := |V_{t+1}(s) - V_t^*(s)|$. Therefore, using statement (ii) of Lemma 13, $\mathcal{V}_{t_{i+1}^s} = |V_{t_{i+1}^s+1}(s) - V_{t_{i+1}^s}^*(s)| \leq \gamma^{\tau_{\min}} \|V_{t_{i+1}^s} - V_{t_{i+1}^s}^*\|$, and therefore upper-bounded by:

$$\begin{aligned}\mathcal{V}_{t_{i+1}^s}(s) &\leq \gamma^{\tau_{\min}}(\|V_{t_{i+1}^s} - V_{t_i^s}^*\| + \|V_{t_i^s}^* - V_{t_{i+1}^s}^*\|)\\ &\overset{(1^*)}{\leq} \gamma^{\tau_{\min}}\left(\|\mathcal{V}_{t_i^s}\| + 2\epsilon)\right),\end{aligned} \tag{8.24}$$

for $i \geq k$, where (1^*) follows from statement (ii) of Lemma 14. From (8.24), we get the following recursion:

$$\|\mathcal{V}_{t_{i+1}^s}\| \leq \gamma^{\tau_{\min}}\left(\|\mathcal{V}_{t_i^s}\| + 2\epsilon\right), \tag{8.25}$$

for $i \geq k$.

Recursively performing (8.25) to obtain upper-bounds on $\|\mathcal{V}_{t_j^s}\|$, for $j = k, \ldots i$, and substituting in (8.24), we get:

$$\mathcal{V}_{t_{i+1}^s}(s) \leq \gamma^{(i+1)\tau_{\min}}\|\mathcal{V}_{t_k^s}\| + \frac{2\gamma^{\tau_{\min}}(1-\gamma^{(i+1)\tau_{\min}})}{1-\gamma^{\tau_{\min}}}\epsilon,$$

In the limit $i \to \infty$, $\mathcal{V}_{t_{i+1}^s} = |V_{t_{i+1}^s+1}(s) - V_{t_{i+1}^s}^*(s)| \leq \epsilon_4$, where $\epsilon_4 := \frac{2\gamma^{\tau_{\min}}}{1-\gamma^{\tau_{\min}}}\epsilon$, and $\epsilon_4 \mapsto 0$ for $\epsilon \mapsto 0$. Since for each $s \in \mathcal{S}$, $V_{t_{i+1}^s}^*(s)$ converges to $V^*(s)$ (Lemma 14), and ϵ is arbitrary, $V_t(s)$ converges to $V^*(s)$ for any s. ■

Proof of Theorem 3

We prove Theorem 3 using the following Theorem 10.

Theorem 10. *T_r is a contraction mapping, and hence, there exists a unique fixed point satisfying, $T_r(V) = V$.*

Proof. The proof follows similar to the case of robust MDPs [32]. □

Proof of Theorem 3: Since T_r is a contraction mapping (Theorem 10), and each state is updated infinitely often, the robust adaptive asynchronous VI converges to a fixed point

$T_r(V) = V$. Furthermore, bounded $\mathcal{P}^a$ implies that the value function at any time t remains bounded. Once $\mathcal{P}^a$ converges to the singleton estimate $\mathbb{P}$, the robust adaptive asynchronous VI reduces to the adaptive asynchronous VI. Hence, the proof follows using Theorem 2. ■

APPENDIX C: CHAPTER 7

Proof of Theorem 1 [Existence of PNE]

We prove Theorem 6 using Brouwer's fixed point theorem [66, Appendix C] applied to the best response mapping with the help of following lemmas (Lemmas 15-17). Recall that $b_i(\lambda^R_{-i})$ is the best response of player i to the review admission rates of other players λ^R_{-i}. For brevity of notation, we will represent $r^R(\lambda^R_i, \lambda^R_{-i}),\ p(\lambda^R_i, \lambda^R_{-i}),\ f_i(\lambda^R_i, \lambda^R_{-i}),\ \tilde{u}_i(\lambda^R_i, \lambda^R_{-i})$ using $r^R,\ p,\ f_i,\ \tilde{u}_i$, respectively. Furthermore, let q' and q'', respectively, represent the first and the second partial derivatives of a generic function q with respect to λ^R_i.

Lemma 15 (***Strict concavity of incentive***)**.** *For the CPR game Γ, under Assumptions (A1-A2), the incentive function $f_i : S \mapsto \mathbb{R}$ is strictly concave in λ^R_i, for $\lambda^R_i \in [0,\ \overline{\lambda}^R_i]$ and any fixed λ^R_{-i}. Equivalently, $f_i(x)$ is strictly concave in x for $x \in [0,\ \mu^S_T - \sum_{j\in\mathcal{N}, j\neq i} a_j\lambda^R_j]$.*

Proof. Recall from (7.9) that

$$f_i(\lambda^R_i, \lambda^R_{-i}) = f_i(x) = r^R(x)(1 - p(x)) - h_i r^S.$$

The first and the second partial derivative of the incentive function f_i with respect to λ^R_i in the interval $\lambda^R_i \in [0,\ \overline{\lambda}^R_i]$ are given by:

$$f_i' = (r^R)'(1-p) - r^R p' = -a_i \frac{df_i}{dx}, \tag{8.26a}$$

$$f_i'' = (r^R)''(1-p) - 2(r^R)'p' - r^R p'' = a_i^2 \frac{d^2 f_i}{dx^2}. \tag{8.26b}$$

From Assumptions (A1) and (A2), we have $f_i'' < 0$ and $\frac{d^2 f_i}{dx^2} < 0$ in the interval where derivative of f_i exists, thereby proving the strict concavity of f_i in λ^R_i and x. □

Lemma 16 (***Best response mapping***)**.** *For the CPR game Γ, under Assumptions (A1-A2), the best response mapping $b_i(\lambda^R_{-i})$ is unique for any $\lambda^R_{-i} \in S_{-i}$ and is given by:*

$$b_i(\lambda^R_{-i}) = \begin{cases} 0, & \text{if } f_i(\lambda^R_i, \cdot) \le 0,\ \ \forall \lambda^R_i \in S_i, \\ \alpha_i, & \text{if } \exists \alpha_i \in S_i \text{ s.t. } \frac{\partial \tilde{u}_i}{\partial \lambda^R_i}(\alpha_i) = 0,\ \text{and } f_i(\alpha_i, \cdot) > 0, \\ \mu^R_i, & \text{otherwise}\ . \end{cases}$$

Proof. We establish uniqueness of the best response mapping through the following three cases.

Case 1: $f_i(\lambda_i^R, \cdot) \leq 0$, for every $\lambda_i^R \in S_i$.

If for a given $\lambda_{-i}^R \in S_{-i}$, $f_i(\lambda_i^R, \cdot) \leq 0$, for every $\lambda_i^R \in S_i$, then from (7.8), $\tilde{u}_i(\lambda_i^R,\ \lambda_{-i}^R)$ admits a unique maximum at $\lambda_i^R = 0$, and therefore, $b_i(\lambda_{-i}^R) = 0$ is the unique best response.

Case 2: There exists a non-empty interval $\overline{S_i} \subset S_i$, such that $f_i(\lambda_i^R, \cdot) > 0$, and $f_i'(\lambda_i^R, \cdot) < 0$, for every $\lambda_i^R \in \overline{S_i}$.

For any given $\lambda_{-i}^R \in S_{-i}$, recall that the system constraint (7.3) is violated for every $\lambda_i^R \in (\overline{\lambda}_i^R, \mu_i^R] \subset S_i$, and $p(\lambda_i^R, \lambda_{-i}^R) = 1$. Therefore, for every $\lambda_i^R \in (\overline{\lambda}_i^R, \mu_i^R]$, we have

$$f_i = -h_i r^S < 0. \tag{8.27}$$

Therefore, $b_i(\lambda_{-i}^R) \in [0,\ \overline{\lambda}_i^R] \subset S_i$, for any given $\lambda_{-i}^R \in S_{-i}$. Furthermore, for a fixed λ_{-i}^R, since p is continuously differentiable with respect to λ_i^R, for each $\lambda_i^R \in (0,\ \overline{\lambda}_i^R)$, $\tilde{u}_i$ is a smooth function on the set $[0,\ \overline{\lambda}_i^R] \times S_{-i}$. Hence, the best response, which is a global maximizer of $\tilde{u}_i$ on the interval $\lambda_i^R \in S_i$, either occurs at the boundary of S_i or satisfies the first order condition, $\frac{\partial \tilde{u}_i}{\partial \lambda_i^R}(b_i) = 0$ (see [123]).

Let there exist $\alpha_i \in S_i$ such that

$$f_i(\alpha_i, \cdot) > 0, \quad \text{and} \tag{8.28a}$$

$$\frac{\partial \tilde{u}_i}{\partial \lambda_i^R}(\alpha_i) = \alpha_i f_i'(\alpha_i, \cdot) + f_i(\alpha_i, \cdot) = 0. \tag{8.28b}$$

Since $f_i(\alpha_i, \cdot) > 0$ and $\alpha_i > 0$, (8.28b) has a solution only if $f_i'(\alpha_i, \cdot) < 0$. Furthermore, $f_i(\alpha_i, \cdot) > 0$ implies $\alpha_i \in [0,\ \overline{\lambda}_i^R]$ (see (8.27)). Therefore, existence of α_i satisfying (8.28) implies there exists a non-empty set $\overline{S_i} \subset [0,\ \overline{\lambda}_i^R] \subset S_i$, such that for each $\alpha_i \in \overline{S_i}$, $f_i(\alpha_i, \cdot) > 0$ and $f_i'(\alpha_i, \cdot) < 0$. For any $\lambda_i^R \in \overline{S_i}$, such that $f_i(\lambda_i^R, \cdot) > 0$ and $f_i'(\lambda_i^R, \cdot) < 0$, using Lemma 15, we get:

$$\frac{\partial^2 \tilde{u}_i}{\partial {\lambda_i^R}^2} = \lambda_i^R f_i'' + 2 f_i' < 0. \tag{8.29}$$

Hence, for $\lambda_i^R \in \overline{S_i}$, the expected utility $\tilde{u}_i$ is strictly concave with a unique global maximizer $\alpha_i \in \overline{S_i}$ that satisfies $\alpha_i = \min\{-\frac{f_i(b_i, \cdot)}{f_i'(b_i, \cdot)},\ \mu_i^R\}$ (see (8.28b)).

Case 3: There exists a non-empty interval $\tilde{S}_i \subset S_i$, such that $f_i(\lambda_i^R, \cdot) > 0$, for every $\lambda_i^R \in \tilde{S}_i$, and $f_i'(\lambda_i^R, \cdot) \geq 0$, for any $\lambda_i^R \in S_i$.

Finally, consider the case that $f_i'(\lambda_i^R, \cdot) \geq 0$, for every $\lambda_i^R \in S_i$, and there exists an interval $\tilde{S}_i \subset S_i$ where $f_i(\lambda_i^R, \cdot) > 0$, for any $\lambda_i^R \in \tilde{S}_i$. Since $f_i'(\lambda_i^R, \cdot) \geq 0$, for every $\lambda_i^R \in S_i$, i.e., $f_i(\lambda_i^R, \cdot)$ is increasing in λ_i^R, and therefore, $f_i(\lambda_i^R, \cdot)$ is maximized at $\lambda_i^R = \mu_i^R$. Since there exists a non-empty interval $\tilde{S}_i$ such that $f_i(\lambda_i^R, \cdot) > 0$, for every $\lambda_i^R \in \tilde{S}_i$, monotonically increasing $f_i(\lambda_i^R, \cdot)$, it follows $\mu_i^R \in \tilde{S}_i$, and $f_i(\mu_i^R, \cdot) > 0$. Therefore, in the interval $\lambda_i^R \in \tilde{S}_i$, (8.28b) has no solution and the expected utility of player i is strictly increasing in λ_i^R, i.e., $\frac{\partial \tilde{u}_i}{\partial \lambda_i^R} > 0$, for every $\lambda_i^R \in S_i$. Therefore, the best response is the unique maximum of $\tilde{u}_i$ which occurs at the boundary μ_i^R. □

We state some important intermediate results from three cases of Lemma 16 as a corollary for later discussions.

Corollary 2 (***Best response and incentive***)**.** *For the CPR game Γ, under Assumptions (A1-A3), the following statements hold:*

(i) *$b_i = 0$, if and only if, $f_i(\lambda_i^R, \cdot) \leq 0$, for every $\lambda_i^R \in S_i$; furthermore, $f_i(\lambda_i^R, \cdot) \leq 0$, for every $\lambda_i^R \in S_i$ implies $f_i'(\lambda_i^R, \cdot) < 0$, for every $\lambda_i^R \in S_i$;*

(ii) *if there exists an interval $\overline{S_i} \subset S_i$, such that $f_i(\lambda_i^R, \cdot) > 0$, and $f_i'(\lambda_i^R, \cdot) < 0$, $\forall \lambda_i^R \in \overline{S_i}$, then the unique best response for player i satisfies the implicit equation $b_i = \min\{-\frac{f_i(b_i, \cdot)}{f_i'(b_i, \cdot)},\ \mu_i^R\} \in S_i$; and*

(iii) *if $f_i'(\lambda_i^R, \cdot) \geq 0$, for every $\lambda_i^R \in S_i$, then $b_i = \mu_i^R$.*

Proof. We only establish the first statement of the corollary. The other statements are established in the proof of Lemma 16. We have already established in Lemma 16 that if $f_i(\lambda_i^R, \cdot) \leq 0$, then for every $\lambda_i^R \in S_i$, the expected utility $\tilde{u}_i$ is maximized for $b_i = 0$. We now establish the “only if" part. Recall from (7.8) that

$$\tilde{u}_i(\lambda_i^R, \lambda_{-i}^R) = \mu_i^S r^S + \lambda_i^R f_i(\lambda_i^R, \lambda_{-i}^R).$$

Let $b_i = 0$ be the best response for player i for a fixed λ_{-i}^R. If there exists $b \in S_i$, such that $f_i(b, \cdot) > 0$, then $\tilde{u}_i(b, \cdot) > \tilde{u}_i(b_i, \cdot)$, and $b_i = 0$ cannot be a best response. Hence, $b_i = 0$ is the best response for player i, if and only if, $f_i(\lambda_i^R, \cdot) \leq 0$, for every $\lambda_i^R \in S_i$.

We now show that if $f_i(\lambda_i^R, \cdot) \leq 0$, for every $\lambda_i^R \in S_i$, then $f_i'(\lambda_i^R, \cdot) < 0$, for every $\lambda_i^R \in S_i$. Since f_i is strictly concave in x (from Lemma 15) and $f_i(\mu_i^R,\ 0) = f_i(\mu_T^S - a_i\mu_i^R) > 0$ by Assumption (A3), there exist $\gamma_1, \gamma_2 \in \mathbb{R}$ such that $\gamma_1 < \mu_T^S - a_i\mu_i^R < \gamma_2$ and $f_i(x) > 0$ if and only if $x \in (\gamma_1, \gamma_2)$.

If $f_i(\lambda_i^R, \lambda_{-i}^R) = f_i(x) \leq 0$ for each $\lambda_i^R \in S_i$ and for a given λ_{-i}^R, then for each $\lambda_i^R \in S_i$, either $x \leq \gamma_1$, or $x \geq \gamma_2$. Suppose $x \geq \gamma_2$, for each $\lambda_i^R \in S_i$. However, for $\lambda_i^R = \mu_i^R$, $x = \mu_T^S - a_i\mu_i - \sum_{j \neq i} a_j \lambda_j^R \leq \mu_T^S - a_i\mu_i < \gamma_2$, which is a contradiction. Hence, $x \leq \gamma_1$, for each $\lambda_i^R \in S_i$.

Finally, from strict concavity of f_i, f_i is increasing in x for $x \leq \gamma_1$. Equivalently, f_i is decreasing in λ_i^R, i.e., $f_i'(\lambda_i^R, \cdot) < 0$, for every $\lambda_i^R \in S_i$. □

Theorem 11 (***Berge Maximum Theorem, adapted from [124]***)**.** *Let $\tilde{u}_i : S_i \times S_{-i} \mapsto \mathbb{R}$ be a continuous function on $S_i \times S_{-i}$, and $C : S_{-i} \mapsto S_i$ be a compact valued correspondence such that $C(\lambda_{-i}^R) \neq \emptyset$ for all $\lambda_{-i}^R \in S_i$. Define $\tilde{u}_i^* : S_{-i} \mapsto \mathbb{R}$ by*

$$\tilde{u}_i^*(\lambda_{-i}^R) = \max\{\tilde{u}_i(\lambda_i^R, \lambda_{-i}^R) \mid \lambda_i^R \in C(\lambda_{-i}^R)\},$$

and $b_i : S_{-i} \mapsto S_i$ by

$$b_i(\lambda_{-i}^R) = \operatorname{argmax}\{\tilde{u}_i(\lambda_i^R, \lambda_{-i}^R) \mid \lambda_i^R \in C(\lambda_{-i}^R)\}.$$

If C is continuous at λ_{-i}^R, then $\tilde{u}_i^*$ is continuous and b_i is upper hemicontinuous with nonempty and compact values. Furthermore, if $\tilde{u}_i$ is strictly quasiconcave in $\lambda_i^R \in S_i$ for each λ_{-i}^R and C is convex-valued, then $b_i(\lambda_{-i}^R)$ is single-valued, and thus is a continuous function.

Lemma 17 (***Continuity of best response mapping***)**.** *For the CPR game Γ, under Assumptions (A1-A3), the best response mapping $b_i(\lambda_{-i}^R)$ is continuous for each $\lambda_{-i}^R \in S_{-i}$.*

Proof. Let $z(\lambda_{-i}^R) : S_{-i} \mapsto [\frac{\mu_T^S - \sum_{j\in\mathcal{N}, j\neq i} a_j \mu_j^R}{a_i},\ \frac{\mu_T^S}{a_i}]$ be defined by

$$z(\lambda_{-i}^R) := \frac{\mu_T^S - \sum_{j\in\mathcal{N}, j\neq i} a_j \lambda_j^R}{a_i}. \tag{8.30}$$

The mapping $z(\lambda_{-i}^R)$ represents an upper bound on the value of λ_i^R above which the system constraint (7.3) is violated. Therefore, for each $\lambda_i^R \in [z(\lambda_{-i}^R), \infty) \cap S_i$, from (7.9), we get

$$f_i = -h_i r^S < 0, \text{and } f_i' = -r^R p' \leq 0, \tag{8.31}$$

where the latter follows from monotonicity of p (Assumption (A2)).

The mapping $z(\lambda_{-i}^R)$ defined in (8.30) is continuous on S_{-i} and linearly decreasing in λ_j^R, for every $j \in \mathcal{N} \setminus \{i\}$. Therefore, to establish the continuity of the best response mapping $b_i(\lambda_{-i}^R)$ on S_{-i}, it is sufficient to show that $b_i(\lambda_{-i}^R) = \phi(z(\lambda_{-i}^R))$, for some continuous function $\phi : [\frac{\mu_T^S - \sum_{j\in\mathcal{N}, j\neq i} a_j \mu_j^R}{a_i},\ \frac{\mu_T^S}{a_i}] \mapsto [0,\ \mu_i^R]$. To this end, we show that for each fixed value of $z(\lambda_{-i}^R)$, b_i is unique and varies continuously with $z(\lambda_{-i}^R)$.

Let $\hat{\lambda}^+ : S_{-i} \mapsto [0,\ \mu_i^R]$ be defined by

$$\hat{\lambda}^+(\lambda_{-i}^R) = \begin{cases} 0, & \text{if } f_i(\lambda_i^R, \lambda_{-i}^R) \leq 0, \forall \lambda_i^R \in S_i, \\ \sup\{\lambda_i^R \in S_i |\ f_i > 0\}, & \text{otherwise.} \end{cases} \tag{8.32}$$

The mapping $\hat{\lambda}^+(\lambda_{-i}^R)$, when non-zero, represents the maximal admissible review admission rate for player i, that yields her a positive incentive to review the tasks. Fig. C.1 shows the best response of player $i \in \mathcal{N}$ for the three possible cases of $\hat{\lambda}^+$.

Case 1: $z(\lambda_{-i}^R) \leq 0$. From (8.31) and statement (i) of Corollary 2, $b_i(\lambda_{-i}^R) = 0$ is the unique (continuous) best response for player i.

Case 2: $z(\lambda_{-i}^R) > 0$. From (8.31), $f_i < 0$ and $f_i' < 0$ for any $\lambda_i^R \geq z(\lambda_{-i}^R)$. Hence, $\hat{\lambda}^+ < z(\lambda_{-i}^R)$. Now we consider three cases based on the value of $\hat{\lambda}^+$.

Case 2.1: $\hat{\lambda}^+ = 0$. In this case, $f_i \leq 0$, for every $\lambda_i^R \in S_i$, and therefore, from statement (i) of Corollary 2, $b_i = 0$ is the unique best response and continuity holds trivially.

Case 2.2: $\hat{\lambda}^+ \in (0,\ \mu_i^R)$. In this case, for any $\lambda_{-i}^R \in S_{-i}$, there exists an interval $\overline{S_i} \subset [0,\ \hat{\lambda}^+]$ such that $f_i > 0$ and $f_i' < 0$, for every $\lambda_i^R \in \overline{S_i}$. Here, $f_i' < 0$ follows from the fact that

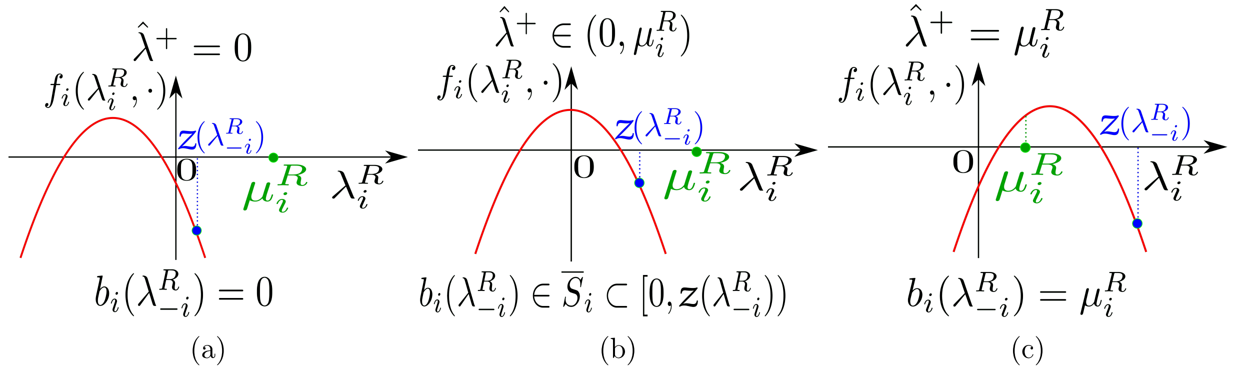

$\hat{\lambda}^+ = 0$
$f_i(\lambda_i^R, \cdot)$
$z(\lambda_{-i}^R)$
0
μ_i^R
λ_i^R
$b_i(\lambda_{-i}^R) = 0$
(a)
$\hat{\lambda}^+ \in (0, \mu_i^R)$
$f_i(\lambda_i^R, \cdot)$
$z(\lambda_{-i}^R)$
0
μ_i^R
λ_i^R
$b_i(\lambda_{-i}^R) \in \overline{S}_i \subset [0, z(\lambda_{-i}^R))$
(b)
$\hat{\lambda}^+ = \mu_i^R$
$f_i(\lambda_i^R, \cdot)$
$z(\lambda_{-i}^R)$
0
μ_i^R
λ_i^R
$b_i(\lambda_{-i}^R) = \mu_i^R$
(c)

Figure C.1 Best response of player i with varying $\hat{\lambda}^+$. The red curve shows different possibilities for strictly concave incentive function $f_i(\lambda_i^R)$ w.r.t $\lambda_i^R \in S_i = [0, \mu_i^R]$, based on the value of λ_{-i}^R. In (a), $f_i < 0$ and $f_i' < 0$ for all $\lambda_i^R \in S_i$; in (b), there exists a subset of S_i where $f_i > 0$ and $f_i' < 0$; and in c) $f_i' \geq 0$ for any λ_{-i}^R. At $z(\lambda_i^R)$ (represented by blue), $f_i < 0$ and $f_i' < 0$. a) For $\hat{\lambda}^+ = 0$, $b_i(\lambda_{-i}^R) = 0$; b) for $\hat{\lambda}^+ \in (0,\ \mu_i^R)$, $b_i(\lambda_{-i}^R) \in \overline{S_i}$; and c) for $\hat{\lambda}^+ = \mu_i^R$, $b_i(\lambda_{-i}^R) = \mu_i^R$.

the supremum in (8.32) corresponds to the decreasing segment of f_i. From statement (ii) of Corollary 2, there exists a unique $b_i(\lambda_{-i}^R) \in \overline{S_i}$ that maximizes $\tilde{u}_i$. Application of Berge maximum theorem [124], yields the continuity of the unique maximizer.

Since, $b_i < \hat{\lambda}^+$, it follows that if $\hat{\lambda}^+ \to 0^+$, then $b_i \to 0^+$. Hence, the continuity holds at $\hat{\lambda}^+ = 0$.

Case 2.3: $\hat{\lambda}^+ = \mu_i^R$. Since $\hat{\lambda}^+ < z(\lambda_{-i}^R)$, $z(\lambda_{-i}^R) \in (\mu_i^R, \frac{\mu_T^S}{a_i}]$. If $f'(\mu_i^R, \lambda_{-i}^R) < 0$, then the continuity follows analogously to Case 2.2. Now consider the case $f'(\mu_i^R, \lambda_{-i}^R) = -\delta$, for $\delta > 0$. Since f_i is concave in λ_i^R and its derivative is decreasing, there exists $\epsilon > 0$ such that $f' < 0$ for $\lambda_i \in (\mu_i^R - \epsilon, \mu_i^R]$. Since $f_i(\lambda_i^R, \cdot)$ is strictly concave in λ_i^R (Lemma 15), there exists at most one point λ_i^R, such that $f'(\lambda_i^R, \cdot) = 0$. Therefore, in the limit $\delta \to 0^+$, $\epsilon \to 0^+$. Hence, in this limiting case $\overline{S_i} = (\mu_i^R - \epsilon,\ \mu_i^R]$, where $\epsilon \to 0^+$, and the best response $b_i(\lambda_{-i}^R) \in \overline{S_i} = (\mu_i^R - \epsilon,\ \mu_i^R]$ converges to μ_i^R.

If $f'(\mu_i^R, \lambda_{-i}^R) \geq 0$, then it follows from strict concavity of f_i that $f'(\lambda_i^R, \lambda_{-i}^R) \geq 0$, for every $\lambda_i^R \in S_i$. Using statement (iii) of Corollary 2, $b_i(\lambda_{-i}^R) = \mu_i^R$ is the unique (continuous) best response.

Note that when $z(\lambda_{-i}^R) = \frac{\mu_T^S}{a_i}$, i.e., when no other player reviews any task ($\lambda_{-i}^R = 0$), from Assumption (A3), $f_i(\mu_i^R, 0) > 0$ and therefore $b_i(\lambda_{-i}^R) = \mu_i^R$. Hence, $b_i(\lambda_{-i}^R)$ is continuous for every $z(\lambda_{-i}^{R})$, and therefore, is continuous for $\lambda_{-i}^{R} \in S_{-i}$. □

Proof of Theorem 6: To prove the existence of a PNE, define a mapping $M : S \mapsto S$ as follows:

$$M(\lambda_1^R,\ \lambda_2^R, ...,\ \lambda_N^R) = (b_1(\lambda_{-1}^R),\ b_2(\lambda_{-2}^R), ...,\ b_N(\lambda_{-N}^R)). \tag{8.33}$$

The mapping M is unique (Lemma 16) and continuous (Lemma 17), and maps the compact convex set S (S_i is convex and compact, $\forall i \in \mathcal{N}$) to itself. Hence, application of Brouwer's fixed point theorem [66, Appendix C] yields that there exists a strategy profile $\lambda^R = \{\lambda_i^{R^*}\}_{i \in \mathcal{N}} \in S$ which is invariant under the best response mapping and therefore is a PNE of the game. □

Proof of Corollary 1 [PNE]:

Since PNE is a best response which remains invariant under the best-response mapping M given by (8.33), Corollary 1 is a direct consequence of Corollary 2 with a simplification that $f_i'(\lambda_i^{R^*}, \lambda_{-i}^{R^*}) < 0$ at PNE. Therefore, to prove Corollary 1, it is sufficient to show statement (i), i.e. $f_i'(\lambda_i^{R^*}, \lambda_{-i}^{R^*}) < 0$ for any player $i \in \mathcal{N}$ at PNE, which we prove by contradiction. Let there exist a player j such that $f_j'(\lambda_j^{R^*}, \lambda_{-j}^{R^*}) \geq 0$ at PNE. From (8.26a), it can be seen that the sign of f_i' remains the same for all players at a PNE. Therefore, $f_j' \geq 0$ implies $f_i' \geq 0$ for all $i \in \mathcal{N}$ at that PNE. In such a case, (8.28b) implies that the expected utility of each player with $\lambda_i^{R^*} > 0$ (therefore, $f_i > 0$) is increasing in λ_i^R at that PNE, and therefore, each of these players can improve their expected utility by unilaterally increasing their review admission rate. Therefore λ^{R^*} cannot be a PNE, which is a contradiction. Hence, $f_i'(\lambda_i^{R^*}, \lambda_{-i}^{R^*}) < 0$ for any player $i \in \mathcal{N}$ at a PNE, and the corollary follows. □

Proof of Proposition 1 [Structure of PNE]

Let $\lambda_{k_1}^{R^*}$ and $\lambda_{k_2}^{R^*}$ be the review admission rates at a PNE for players k_1 and k_2, respectively, with $h_{k_1} \leq h_{k_2}$. By proving $a_{k_1}\lambda_{k_1}^{R^*} \geq a_{k_2}\lambda_{k_2}^{R^*}$, $\lambda_{k_1}^{R^*} \geq \lambda_{k_2}^{R^*}$ is established trivially since $a_{k_1} \leq a_{k_2}$. We assume $a_{k_1}\lambda_{k_1}^{R^*} < a_{k_2}\lambda_{k_2}^{R^*}$ and prove the first statement by establishing a contradiction argument using two cases discussed below. Furthermore, the proof of the second statement is contained within Case 1 below.

Case 1: $\lambda_{k_1}^{R^*} = 0$.

From statement (ii) of Corollary 1, $f_{k_1}(\lambda_{k_1}^{R^*}, \lambda_{-k_1}^{R^*}) \leq 0$. From (7.9), the incentives f_{k_1} and f_{k_2} for players k_1 and k_2 at a PNE satisfies:

$$f_{k_2} = f_{k_1} + (h_{k_1} - h_{k_2})r^S \leq 0.$$

Therefore, utilizing statement (ii) of Corollary 1 again implies $\lambda_{k_2}^{R^*} = 0$, which is a contradiction. This case also proves the second statement.

Case 2: $\lambda_{k_1}^{R^*} > 0$.

By assumption, $a_{k_1}\lambda_{k_1}^{R^*} < a_{k_2}\lambda_{k_2}^{R^*}$, from statement (iii) of Corollary 1, $\lambda_i^{R^*}$, where $i \in \{k_1, k_2\}$, satisfy the implicit equation

$$\lambda_i^{R^*} = \min\left\{ -\frac{f_i(\lambda_i^{R^*},\ \lambda_{-i}^{R^*})}{f_i'(\lambda_i^{R^*},\ \lambda_{-i}^{R^*})},\ \mu_i^R \right\}.$$

We assume that $\lambda_{k_1}^{R^*} < \mu_{k_1}^R$, and therefore, $\lambda_{k_1}^{R^*} = -\frac{f_{k_1}}{f'_{k_1}}$. Using (7.9) and (8.26a), we get

$$\begin{aligned} a_{k_2}\lambda_{k_2}^{R^*} &= \min\left\{ -a_{k_2}\frac{f_{k_2}}{f'_{k_2}},\ a_{k_2}\mu_{k_2}^R \right\} \\ &\leq -a_{k_2}\frac{f_{k_2}}{f'_{k_2}} = -a_{k_1}\frac{f_{k_1} + (h_{k_1} - h_{k_2})r^S}{f'_{k_1}} \leq a_{k_1}\lambda_{k_1}^{R^*}, \end{aligned}$$

which is a contradiction. Hence, if $\lambda_{k_1}^{R^*} < \mu_{k_1}^R$, then $a_{k_1}\lambda_{k_1}^{R^*} \geq a_{k_2}\lambda_{k_2}^{R^*}$ and $\lambda_{k_1}^{R^*} \geq \lambda_{k_2}^{R^*}$ for each $k_2 > k_1$. □

Proof of Theorem 2 [Uniqueness of PNE]

Suppose that the CPR game Γ has multiple PNEs. We define the support of a PNE as the total number of players with non-zero review admission rate. Let $\text{PNE}_1 = \lambda^1 = [\lambda_1^1,\ \lambda_2^1, \ldots,\ \lambda_N^1]$ and $\text{PNE}_2 = \lambda^2 = [\lambda_1^2,\ \lambda_2^2, \ldots,\ \lambda_N^2]$, be two different PNEs with distinct supports m_1 and m_2, respectively. For brevity of notation, we have removed the superscript R from the two PNEs and replaced it by their unique identifier. Without loss of generality, let $m_2 > m_1$. Let $x^1 = \mu_T^S - \sum_{i=1}^N a_i\lambda_i^1$ and $x^2 = \mu_T^S - \sum_{i=1}^N a_i\lambda_i^2$ be the slackness parameters at PNE_1 and PNE_2, respectively.

We prove the uniqueness of PNE using a six step process.

Step 1: *We first show that if there exists two different PNEs with distinct supports* m_1 *and* m_2 *(*$m_1 < m_2$*), then* $x^1 < x^2$.

If m_1 and m_2 are the supports of PNE$_1$ and PNE$_2$, respectively, then $\lambda_i^1 = 0$ and $\lambda_j^2 = 0$, for each $i > m_1$, and $j > m_2$, respectively (Proposition 1). Additionally, $\lambda_i^1 > 0$, and $\lambda_j^2 > 0$, for each $i \leq m_1$, and $j \leq m_2$. Hence, $m_2 > m_1$ implies $\lambda_{m_2}^1 = 0$, while $\lambda_{m_2}^2 > 0$.

From statement (i) of Corollary 2, $b_i = 0$, if and only if $f_i \leq 0$ and $f_i{}' < 0$ (equivalently $\frac{df_i}{dx} > 0$) for all $\lambda_i^R \in S_i$. Therefore, $\lambda_{m_2}^1 = 0$ implies $f_{m_2}^1 := f_{m_2}(\lambda^1) \leq 0$ and $\frac{df_{m_2}}{dx} > 0$ everywhere, while $\lambda_{m_2}^2 > 0$ implies $f_{m_2}^2 := f_{m_2}(\lambda^2) > 0$. Since $f_{m_2}^2 > 0 > f_{m_2}^1$ and $\frac{df_{m_2}}{dx} > 0$ everywhere, it follows that $x^1 < x^2$.

Step 2: *We now show that* $x^1 > x^2$ *using Steps 2-5, which is a contradiction to the result of Step 1*, and consequently $m_1 = m_2$.

From statement (iii) of Corollary 1, the review admission rate of any player i, $i \leq m_1$, at PNE$_k$, $k \in \{1, 2\}$, satisfies

$$\lambda_i^k = \min\left\{ -\frac{f_i^k}{f_i^{k\prime}},\ \mu_i^R \right\}. \tag{8.34}$$

Step 3: *We show that* $f_i^2 > f_i^1$ *for any player* i, $i \leq m_1$.

From (7.9), the incentives f_i and f_j for any two distinct players i and j with $j > i$ at a PNE$_k$, $k \in \{1, 2\}$ satisfies:

$$f_i^k - f_j^k = (h_j - h_i)r^S > 0,\ \forall j > i.$$

Notice that the right hand side of above equation is independent of λ_i^R and therefore, a constant for both PNEs. Hence, for every $i < m_2$

$$f_i^1 - f_{m_2}^1 = f_i^2 - f_{m_2}^2.$$

Therefore, $f_{m_2}^2 > f_{m_2}^1$ implies $f_i^2 > f_i^1$, for every $i \leq m_1 < m_2$.

Step 4: *We show that* $f'^1_i < f'^2_i$*, for every player* i, $i \leq m_1$.

Recall that f_i is strictly concave in x (Lemma 15). Therefore, $x^1 < x^2$ (Step 1) implies $\frac{df_i^1}{dx} > \frac{df_i^2}{dx}$. Therefore, from (8.26a), $f'^1_i < f'^2_i$, for any player i, $i \leq m_1$.

Step 5: *We now show that $x^1 > x^2$, which is a contradiction to result of Step 1, and consequently $m_1 = m_2$.*

Since for all players i, $i \leq m_1$, $f_i^2 > f_i^1$ (Step 3) and $-f'^1_i > -f'^2_i$ (Step 4), (8.34) implies $\lambda_i^2 \geq \lambda_i^1$, for each $i \leq m_1$. Therefore, $\sum_{i=1}^{N} a_i \lambda_i^2 > \sum_{i=1}^{m_1} a_i \lambda_i^2 \;\geq\; \sum_{i=1}^{m_1} a_i \lambda_i^1 = \sum_{i=1}^{N} a_i \lambda_i^1$, which implies $x^1 > x^2$, which is a contradiction to result of Step 1. Hence, $m_1 = m_2$

Step 6: *We now show the value of slackness parameter x at any PNE is unique.*

Steps 1 to 5 show that, at a PNE, the number of players with non-zero review admission rate are unique. Therefore, let m be the identical support for PNE$_1$ and PNE$_2$. Without loss of generality, let $x_1 > x_2$.

Let $g_i : \mathbb{R} \mapsto \mathbb{R}$, for $i \leq m$, be defined by

$$g_i(x) = -\frac{f(x)}{f'(x)}.$$

Differentiating $g_i(x)$ w.r.t x, we get

$$\frac{dg_i(x)}{dx} = \frac{\left(\frac{df_i(x)}{dx}\right)^2 - f_i(x)\frac{d^2 f_i(x)}{dx^2}}{a_i\left(\frac{df_i(x)}{dx}\right)^2}.$$

Recall from statement (iii) of Corollary 1 that players have non-zero review admission rate at PNE, if and only if, $f_i > 0$ at PNE. Strict concavity of f_i (Lemma 15) implies $\frac{dg_i(x)}{dx} > 0$. Consequently, at PNE, the review admission rate for any player i, $i \leq m$, is increasing with x. Therefore, assumption $x^1 > x^2$ implies $\lambda_i^1 \geq \lambda_i^2$, for each player $i \leq m$. Consequently, $x^1 = \mu_T^S - \sum_{i=1}^{m} a_i \lambda_i^1 \leq \mu_T^S - \sum_{i=1}^{m} a_i \lambda_i^2 = x^2$, which is a contradiction. Therefore, $x^1 = x^2$. We now show the uniqueness of PNE. Steps 1 to 6 show that, at a PNE, the number of players with non-zero review admission rate and the slackness parameter x are unique. Therefore, the first order conditions (8.34) give the unique review admission rate for each player i for unique slack parameter x, thereby implying uniqueness of PNE. □

Proof of Lemma 11 [Non-increasing best response]

We prove this lemma by considering the three cases of the best response mapping in Lemma 16 (Appendix 13):

Case 1: $b_i = 0$. Recall that $x = \mu_T^S - \sum_{i=1}^N a_i \lambda_i^R$. In this case, from statement (i) of Corollary 2, $f_i \leq 0$ and $f_i' < 0$ (equivalently, $\frac{df_i}{dx} > 0$), for all $\lambda_i^R \in S_i$. Since x can be re-written as $x = \mu_T^S - a_i\lambda_i^R - \sigma_i(\lambda_{-i}^R)$, therefore $\frac{df_i}{dx} > 0$ implies $\frac{\partial f_i}{\partial \sigma_i} < 0$. Hence, with increase in $\sigma_i(\lambda_{-i}^R)$, $b_i = 0$ remains the best response.

Case 2: $b_i = -\frac{f_i(b_i,\sigma_i(\lambda_{-i}^R))}{f_i'(b_i,\sigma_i(\lambda_{-i}^R))}$. In this case, from statement (ii) of Corollary 2, $b_i \in \overline{S_i}$ such that $f_i > 0$ and $f_i' < 0$, for every $\lambda_i^R \in \overline{S_i}$. Thus,

$$\frac{db_i}{d\sigma_i} = \frac{-f_i'^2 + f_i'' f_i}{a_i f_i'^2} < 0. \tag{8.35}$$

Hence, b_i is strictly decreasing in $\sigma_i(\lambda_{-i}^R)$.

Case 3: $b_i = \mu_i^R$. Since $b_i \in S_i = [0,\ \mu_i^R]$, b_i either decreases or remains constant with increase in $\sigma_i(\lambda_{-i}^R)$. □

Proof of Theorem 8 [Analytic bounds on PNE Inefficiency]

We first establish the analytic upper bound on PoA, followed by upper bounds on η_{TRI} and η_{LI}.

Let $\mathcal{G}$ be the family of CPR games parameterized by the ratios of the maximum service and review admission rates of each player $i \in \mathcal{N}$. Therefore, the CPR game $\Gamma \in \mathcal{G}$, with the corresponding ratio for player i given by h_i. Define a set of homogeneous CPR games $\mathcal{G}^H \subset \mathcal{G}$, in which each player has a constant ratio $\frac{\{\mu_i^S\}^H}{\{\mu_i^R\}^H} =: h$, and $\min_i\{\{\mu_i^R\}^H\} \geq \frac{\{\mu_T^S\}^H}{N(1+h)}$. The superscript H is used to distinguish maximum service and review admission rates of the homogeneous game Γ^H from the CPR game Γ. We obtain the bounds on PoA by comparing the social utility of the heterogeneous CPR game with a corresponding homogeneous game in which all players have the largest heterogeneity measure, i.e., $h_i = h_N$, for every $i \in \mathcal{N}$.

For any CPR game in $\mathcal{G}$, PoA is given by:

$$PoA = \frac{(\Psi)_{SW}}{(\Psi)_{PNE}} = \frac{[\sum_{i=1}^N \mu_i^S r^S + \sum_{i=1}^N \lambda_i^R f_i(x)]_{SW}}{[\sum_{i=1}^N \mu_i^S r^S + \sum_{i=1}^N \lambda_i^R f_i(x)]_{PNE}}. \tag{8.36}$$

We now provide an analytic upper bound on the PoA for the CPR game Γ using following Lemmas.

Lemma 18 (***PNE solution for homogeneous CPR game***)**.** *For any homogeneous CPR game $\Gamma^H \in \mathcal{G}^H$, such that for each player $i \in \mathcal{N}$, $\frac{\{\mu_i^S\}^H}{\{\mu_i^R\}^H} = h$, and $\min_i\{\{\mu_i^R\}^H\} \geq \frac{\{\mu_T^S\}^H}{N(1+h)}$, each player participates in the review process with equal review admission rate $\lambda_i^H = \lambda_H$ at PNE. Let λ_T^H be the total review admission rate at PNE for Γ^H. The unique PNE solution is given by $\lambda_H = \frac{\lambda_T^H}{N}$, where $\lambda_T^H = \frac{f(x)}{(1+h)\frac{df}{dx}}$ and $x = \{\mu_T^S\}^H - (1+h)\lambda_T^H$.*

Proof. For the homogeneous CPR game Γ^H, each player has equal incentive $f(x)$ to review the tasks. If $f(x) \leq 0$ at PNE, all players have $\lambda_i^H = 0$ (statement (ii) of Corollary 1) which contradicts assumption (A3). Hence, at PNE, each players has $\lambda_i^H > 0$.

Let $\min_i\{\{\mu_i^R\}^H\} \geq \frac{\{\mu_T^S\}^H}{N(1+h)}$ for Γ^H. At PNE, $x > 0$. Let $\mathcal{D} \subseteq \mathcal{N}$ be a non-empty set of player indices such that for any $i \in \mathcal{D}$, $\{\mu_i^R\}^H \leq -\frac{f(x)}{f'(x)}$. At PNE,

$$\begin{aligned}\lambda_T^H &= \sum_{i\in\mathcal{D}}\{\mu_i^R\}^H + \sum_{i\in\mathcal{N}\setminus\mathcal{D}}\frac{-f(x)}{f'(x)}\\ &\geq N\min_i\{\{\mu_i^R\}^H\}\\ &\geq \frac{\{\mu_T^S\}^H}{(1+h)}.\end{aligned}$$

Therefore, at PNE,

$$\begin{aligned}x &= \{\mu_T^S\}^H - (1+h)\lambda_T^H\\ &\leq \{\mu_T^S\}^H - (1+h)N\min_i\{\{\mu_i^R\}^H\} \leq 0,\end{aligned}$$

which is a contradiction. Hence, $\mathcal{D}$ is an empty set and each player has equal review admission rate at PNE, given by $\lambda_i^H = \lambda_H = \frac{\lambda_T^H}{N}$. Hence, each player being a maximizer of their expected utility maximizes:

$$\tilde{u}_i = \{\mu_i^S\}^H r^S + \frac{\lambda_T^H}{N}f(x), \tag{8.37}$$

where $x = \{\mu_T^S\}^H - (1+h)\lambda_T^H$. Setting $\frac{\partial\tilde{u}_i}{\partial\lambda_T^H} = 0$, we get $\lambda_T^H = \frac{f(x)}{(1+h)\frac{df}{dx}}$. □

Lemma 19 (***PoA=1 for homogeneous CPR game***)**.** *For any homogeneous CPR game $\Gamma^H \in \mathcal{G}^H$, such that for each player $i \in \mathcal{N}$, $\frac{\{\mu_i^S\}^H}{\{\mu_i^R\}^H} = h$, and $\min_i\{\{\mu_i^R\}^H\} \geq \frac{\{\mu_T^S\}^H}{N(1+h)}$, PoA=1.*

Proof. For homogeneous CPR game Γ^H, social welfare function Ψ^H in (7.15) only depends on λ_T^R and is given by:

$$\Psi^H = \{\mu_T^S\}^H r^S + \lambda_T^R f(x), \tag{8.38}$$

where $x = \{\mu_T^S\}^H - (1+h)\lambda_T^R$, and $f(x)$ is the uniform incentive function for each player. Note that $\frac{d\Psi^H}{d\lambda_T^R} > 0$ when $\frac{df}{dx} \leq 0$, and $\frac{d^2\Psi^H}{d\lambda_T^{R^2}} > 0$ in the interval where $\frac{df}{dx} > 0$. It is easy to show that Ψ^H is maximized by any λ_T^R satisfying $\lambda_T^R = \frac{f(x)}{(1+h)\frac{df}{dx}}$ obtained by setting $\frac{d\Psi^H}{d\lambda_T^R} = 0$, and $\frac{df}{dx} > 0$ at the maximizer.

Let λ_T^H be the total review admission rate at PNE for Γ^H. The unique PNE satisfies $\lambda_T^H = \frac{f(x)}{(1+h)\frac{df}{dx}}$ (Lemma 18), and hence, maximizes social welfare utility resulting in PoA= 1. □

Corresponding to the CPR game Γ, construct a homogeneous game $\Gamma_N^H \in \mathcal{G}^H$ with $\{\mu_i^S\}^H = \mu_i^S$ and $\{\mu_i^R\}^H = \frac{\mu_i^S}{h_N}$, for each player $i \in \mathcal{N}$. Note that for homogeneous players in Γ_N^H, $\frac{\{\mu_i^S\}^H}{\{\mu_i^R\}^H} = h = h_N$, and $\sum_{i=1}^N \{\mu_i^S\}^H = \mu_T^S$. Furthermore, the assumption $\min_i\{\mu_i^S\} > \frac{\mu_T^S h_N}{N(1+h_N)}$ implies $\min_i\{\{\mu_i^R\}^H\} > \frac{\{\mu_T^S\}^H}{N(1+h)}$. Hence, $PoA = 1$ for Γ_N^H (Lemma 19).

To obtain analytic bounds on PoA, we now compute a lower bound on the social utility obtained at the unique PNE Ψ_{PNE}^Γ for the CPR game Γ. In Lemma 20, we show that the social utility obtained at the PNE Ψ_{PNE}^H for the homogeneous game Γ_N^H lower bounds Ψ_{PNE}^Γ. For any $x \in [0,\ \mu_T^S]$, homogeneous players with ratio h_N in Γ_N^H have a lower cumulative incentive ($\sum f$) to review tasks than players in Γ, and therefore, have a lower social utility at PNE, i.e., $\Psi_{PNE}^\Gamma \geq \Psi_{PNE}^H$. We further lower bound Ψ_{PNE}^Γ by computing a lower bound on Ψ_{PNE}^H.

Lemma 20 (***Lower bound for social welfare at PNE***)**.** *Let Γ_N^H be a homogeneous game corresponding to CPR game Γ with each player $i \in \mathcal{N}$ having $\{\mu_i^S\}^H = \mu_i^S$ and $\{\mu_i^R\}^H = \frac{\mu_i^S}{h_N}$. Let Ψ_{PNE}^Γ and Ψ_{PNE}^H be the social welfare functions for Γ and Γ_N^H, respectively, evaluated at their unique PNEs. Then $\Psi_{PNE}^\Gamma \geq \Psi_{PNE}^H \geq \mu_T^S r^S + \frac{\mu_T^S - \overline{x}}{a_N} f_N(\overline{x})$, where $\overline{x}$ is the unique maximizer of f_i, i.e. $\frac{df_i}{dx}(\overline{x}) = 0$.*

Proof. Let $\lambda^* = [\lambda_1^*, \ldots, \lambda_N^*]$ (Corollary 1) and $\lambda^H = [\lambda_H, \ldots, \lambda_H]$ (Lemma 18) be the unique PNEs for the CPR games Γ and Γ_N^H, respectively. Let $x^* = \mu_T^S - \sum_{i=1}^N a_i \lambda_i^*$ and $x^H = \mu_T^S - N a_N \lambda_H$ be their slackness parameters at respective PNEs.

Step 1: *We show that $x^H \geq x^*$ using contradiction.*

Let $x^* > x^H$. Recall that at PNE, $\frac{df_i}{dx} > 0$ (Corollary 1). Using strict concavity of f_i (Lemma 15), we have $\frac{df_i}{dx}(x^H) > \frac{df_i}{dx}(x^*) > 0$. Therefore, $\frac{f_N(x^*)}{\frac{df_N}{dx}(x^*)} > \frac{f_N(x^H)}{\frac{df_N}{dx}(x^H)} = a_N \lambda_N^H$. Recall that $f_1(x) \geq \cdots \geq f_N(x)$ for any x, and $\frac{df_i}{dx} = \frac{dr^R(x)}{dx}(1 - p(x)) - r^R(x)\frac{dp(x)}{dx}$ is independent of i. Hence, $\frac{f_i(x^*)}{\frac{df_i}{dx}(x^*)} > a_N \lambda_N^H$ for any i. Using $h_N \geq h_i$, we get $a_i \mu_i^R = \mu_i^S + \frac{\mu_i^S}{h_i} \geq \mu_i^S + \frac{\mu_i^S}{h_N} = a_N \{\mu_i^R\}^H > a_N \lambda_N^H$. Therefore, $a_i \lambda_i^* = \min\left\{\frac{f_i(x^*)}{\frac{df_i}{dx}(x^*)}, a_i \mu_i^R\right\} > a_N \lambda_N^H$, for any i. Hence, $x^* < x^H$, which is a contradiction. Therefore, $x^H \geq x^*$ (equivalently, $\sum_{i=1}^N a_i \lambda_i^* \geq N a_N \lambda_N^H$) and $\frac{df_i}{dx}(x^*) \geq \frac{df_i}{dx}(x^H) > 0$.

Step 2: *We show that $\sum_i f_i(x^*) \geq N f_N(x^H)$ & $\sum_i \lambda_i^* \geq N \lambda_N^H$.*

Let $d \leq N$ be the support for λ^*. Therefore, $\lambda_i^* = \min\left\{\frac{f_i(x^*)}{a_i \frac{df_i}{dx}(x^*)}, \mu_i^R\right\}$ for every $i \leq d$, , and $\lambda_i^* = 0$ for any $i > d$. Therefore, $\sum_{i=1}^d \frac{f_i(x^*)}{\frac{df_i}{dx}(x^*)} \geq \sum_{i=1}^d a_i \lambda_i^* \geq N a_N \lambda_N^H = N \frac{f_N(x^H)}{\frac{df_N}{dx}(x^H)}$. Using $\frac{df_i}{dx}(x^*) \geq \frac{df_N}{dx}(x^H) > 0$ ($\frac{df_i}{dx}$ is independent of i), we get $\sum_{i=1}^d f_i(x^*) \geq N f_N(x^H)$. Additionally, $\sum_{i=1}^N a_i \lambda_i^* \geq N a_N \lambda_N^H$ implies $\sum_{i=1}^d \lambda_i^* \geq \sum_{i=1}^N \frac{a_i}{a_N} \lambda_i^* \geq N \lambda_N^H$.

Step 3: *We show that $\Psi_{PNE}^\Gamma \geq \Psi_{PNE}^H$.*

Using $h_N \geq h_i$, we have $\mu_i^R = \frac{\mu_i^S}{h_i} \geq \frac{\mu_i^S}{h_N} = \{\mu_i^R\}^H > \lambda_N^H$. Let $d_1 \leq d$ be the largest index of player satisfying $\frac{f_i(x^*)}{a_i \frac{df_i}{dx}} > \lambda_N^H$. Since $\frac{f_N(x^*)}{a_N \frac{df_N}{dx}(x^*)} < \lambda_N^H$, $d_1 < N$. Therefore, λ_i^* satisfies,

$$\lambda_i^* = \begin{cases} \min\left\{\frac{f_i(x^*)}{a_i \frac{df_i}{dx}(x^*)}, \mu_i^R\right\} > \lambda_N^H, & \text{for } i \leq d_1, \\ \frac{f_i(x^*)}{a_i \frac{df_i}{dx}(x^*)} \leq \lambda_N^H, & \text{for } d_1 + 1 \leq i \leq d. \end{cases} \tag{8.39}$$

Hence,

$$\begin{aligned} \Psi_{PNE}^\Gamma &= \mu_T^S r^S + \sum_{i=1}^d \lambda_i^* f_i(x^*) \\ &\overset{(1^*)}{=} \mu_T^S r^S + \lambda_N^H \sum_{i=1}^d f_i(x^*) + \sum_{i=1}^{d_1} (\lambda_i^* - \lambda_N^H) f_i(x^*) \end{aligned}$$

$$
\begin{aligned}
&+\sum_{i=d_1+1}^{d}(\lambda_i^* - \lambda_N^H)f_i(x^*)\\
&\overset{(2^*)}{\geq} \mu_T^S r^S + \lambda_N^H N f_N(x^H) + \sum_{i=1}^{d_1}(\lambda_i^* - \lambda_N^H) f_{d_1+1}(x^*)\\
&+\sum_{i=d_1+1}^{d}(\lambda_i^* - \lambda_N^H)f_{d_1+1}(x^*)
\end{aligned}
$$

$$
\begin{aligned}
&= \mu_T^S r^S + \lambda_N^H N f_N(x^H) + f_{d_1+1}(x^*)\sum_{i=1}^{d}(\lambda_i^* - \lambda_N^H)\\
&\overset{(3^*)}{\geq} \mu_T^S r^S + \lambda_N^H N f_N(x^H) = \Psi_{PNE}^H,
\end{aligned}
$$

where (1^*) follows by adding and subtracting $\lambda_N^H \sum_{i=1}^d f_i(x^*)$. (2^*) follows from $\sum_{i=1}^d f_i(x^*) \geq N f_N(x^H)$ (Step 2), (8.39), and the fact that $f_1(x^*) \geq \cdots \geq f_N(x^*)$. (3^*) follows by recalling that $\sum_{i=1}^d \lambda_i^* \geq N\lambda_N^H$ (Step 2).

Step 4: *We show that* $\Psi_{PNE}^\Gamma \geq \Psi_{PNE}^H \geq \mu_T^S r^S + \frac{\mu_T^S - \overline{x}}{a_N} f_N(\overline{x})$.

Let $\overline{x}$ be the unique maximizer of f_i. From Lemma 19, $\Psi_{PNE}^H = \Psi_{SW}^H = \max\{\Psi^H\} \geq \mu_T^S r^S + \lambda_T^R f_N(\mu_T^S - a_N\lambda_T^R)$ for any λ_T^R. Choosing $\lambda_T^R = \frac{\mu_T^S - \overline{x}}{a_N}$, we obtain the desired bounds. □

Proof of Theorem 8: The global optimum of social welfare function is upper bounded by:

$$
\begin{aligned}
\Psi_{SW}^\Gamma &= \mu_T^S r^S + \max_{\lambda_i^R}\left\{\sum_{i=1}^N \lambda_i^R f_i(x)\right\}\\
&\leq \mu_T^S r^S + \max_{\lambda_i^R}\left\{\lambda_T^R\right\}\max_x\{f_i(x)\}\\
&\overset{(1^*)}{\leq} \mu_T^S r^S + \mu_T^S f_i(\overline{x})\\
&\leq \mu_T^S(r^S + r^R(\overline{x})(1-p(\overline{x}))), \qquad (8.40)
\end{aligned}
$$

where (1^*) is obtained using the system constraint $x > 0$ which implies $\mu_T^S \geq \sum_{i=1}^N a_i\lambda_i^R \geq \lambda_T^R$.

From Lemma 20, Ψ_{PNE}^Γ is lower bounded by:

$$
\begin{aligned}
\Psi_{PNE}^\Gamma &\geq \mu_T^S r^S + \frac{\mu_T^S - \overline{x}}{a_N} f_N(\overline{x})\\
&= \frac{1}{a_N}\left(\overline{x} a_N r^S + (\mu_T^S - \overline{x})(r^S + r^R(\overline{x})(1-p(\overline{x})))\right)
\end{aligned}
$$

$$\geq \frac{1}{a_N}(\mu_T^S - \overline{x})(r^S + r^R(\overline{x})(1 - p(\overline{x}))). \tag{8.41}$$

Using (8.40) and (8.41), we get, $PoA = \frac{\Psi_{SW}^{\Gamma}}{\Psi_{PNE}^{\Gamma}} \leq \frac{\mu_T^S a_N}{\mu_T^S - \overline{x}}$.

Now we establish the bounds on η_{TRI} and η_{LI}. Let x_{PNE} and x_{SW} be the slackness parameter corresponding to the PNE and social welfare, respectively. Recall that $\frac{df_i}{dx} > 0$ (Corollary 1), for $x \in \{x_{PNE}, x_{SW}\}$. Hence, using strict concavity of f_i, we have $x_{PNE}, x_{SW} \in (0,\ \overline{x})$. Therefore, $\mu_T^S - \overline{x} < \sum_{i=1}^N a_i\{\lambda_i^R\}_{PNE} < \mu_T^S$, and $\mu_T^S - \overline{x} < \sum_{i=1}^N a_i\{\lambda_i^R\}_{SW} < \mu_T^S$. Hence, η_{TRI} and η_{LI} are upper bounded by:

$$\eta_{TRI} = \frac{(\lambda_T^R)_{SW}}{(\lambda_T^R)_{PNE}} < \frac{\mu_T^S a_N}{(\mu_T^S - \overline{x})a_1}, \text{and}$$

$$\eta_{LI} = \frac{(\sum_{i=1}^N a_i \lambda_i^R)_{PNE}}{(\sum_{i=1}^N a_i \lambda_i^R)_{SW}} < \frac{\mu_T^S}{\mu_T^S - \overline{x}}.$$

□