Information Asymmetries in Data-Driven and Sustainable Operations: Stochastic Models and Adaptive Algorithms for Strategic Agents

By

Ilgin Dogan

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Industrial Engineering and Operations Research

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Zuo-Jun Max Shen, Co-chair
Associate Professor Anil Aswani, Co-chair
Professor Alexandre M. Bayen
Professor Philip M. Kaminsky

Spring 2024

Information Asymmetries in Data-Driven and Sustainable Operations:
Stochastic Models and Adaptive Algorithms for Strategic Agents

Abstract

Information Asymmetries in Data-Driven and Sustainable Operations: Stochastic Models and Adaptive Algorithms for Strategic Agents

by

Ilgin Dogan

Doctor of Philosophy in Engineering - Industrial Engineering and Operations Research

University of California, Berkeley

Professor Zuo-Jun Max Shen, Co-chair

Associate Professor Anil Aswani, Co-chair

The modern landscape of operations management (OM) has undergone a profound paradigm shift driven by two surging forces: 1) the integration of expansive real-time data inflow, and 2) the recognition of ambiguity in navigating operational disruptions due to climate crisis. In this transition to data-driven and sustainable operations, a fundamental challenge lies in isolating the lack of transparency in collaboration willingness and misaligned economic motives of strategic agents (i.e., stakeholders) in socio-technical systems.

Motivated by contributing to this breakthrough, this dissertation establishes a foundational theory that leverages data-driven decision-making to proactively mitigate intricate uncertainties, arising from imperfect model insights and information asymmetries, hindering sustainable OM. The dissertation begins by exploring nonlinear and non-stationary control systems under imperfect knowledge of the reward function and system dynamics—a nontrivial scenario common in applications like balancing occupant comfort and energy efficiency in buildings. Expanding on this rigorous control-theoretic learning analysis, the majority of the dissertation is devoted to devising novel, data-driven, and adaptive incentive frameworks to tackle unexplored information disparities in the context of repeated principal-agent games.

Inspired by several real-world applications, such as forest conservation incentives in Payment for Ecosystem Services and renewable energy aggregator contracts for utility grids, this dissertation introduces the "hidden agent rewards" model within a multi-armed bandit framework, where: a principal learns to proactively lead an agent's choices by sequentially offering menus of incentives which contribute to the agent's hidden rewards for a finite set of arms. Designing policies in this setting is challenging, because it entails analyzing dynamic externalities imposed by two separate learning algorithms trained in parallel by strategic parties. To the best of our knowledge, this dissertation presents i) the first generic stochastic sequential model for this widely applicable information imbalance context, and ii) the first methodological framework that contends with the principal's trade-off between consistently learning the agent's rewards and maximizing their own rewards through adaptive incentives. We examine two scenarios: one where the agent has perfect knowledge of their reward model and another where the agent learns their model over time, potentially leading to misleading choices for the principal. In both cases, solid statistical consistency and regret guarantees are proven to persist without restricting the agent's algorithm or reward distributions. Throughout the dissertation, these theoretical results, along with versatile practical insights, outline a prosperous future research landscape to enhance various incentive practices in OM confronting the hidden objectives of incentivized agents.

To my mother, Hülya, for being my irreplaceable ever-bright lodestar.

To my grandmother, Gülhanım, for her exceptional humor to live by.

To my lifemate, Semih, for our odyssey hand in hand.

Contents

List of Figures

List of Tables

Acknowledgments

The successful completion of this dissertation and my journey at UC Berkeley would not have been possible without the support of many people, whom I acknowledge here.

First and foremost, I extend my sincere thanks to my advisors, Zuo-Jun Max Shen and Anil Aswani, by whom it was a great privilege and honor to be guided. Max is a very visionary professor, a worldly-wise person, and a generous advisor. Being advised by him not only gives a chance to benefit from his insightful vision but also entails being a part of a group of bright and dedicated researchers. I am grateful for all the opportunities that have come to me thanks to being a member of his esteemed academic family. Anil is one of the finest intellects in our domain who has consistently amazed me with his wisdom, resourcefulness, and technical skills. I am deeply indebted for his unwavering kindness, in-depth feedback, and invaluable mentorship, including his genuine support and encouragement during my academic job search. Both Max and Anil have wholeheartedly supported my academic and personal growth, and they are the academics whose paths I aim to follow.

Besides my advisors, I owe thanks to the people who have enhanced my research with their valuable contributions. I would like to express my gratitude to my renowned committee members, Philip M. Kaminsky and Alexandre M. Bayen, for their constructive feedback and services in the progression of this dissertation. I am sincerely thankful to my collaborators, Ho-Yin Mak and Yoon Lee, for their generosity and openness to exchange. I would like to extend special recognition to Ho-Yin, to whom I am much obliged for his sincere support in my academic job search. I feel truly fortunate to have had the pleasurable and enlightening opportunity to work with him.

My graduate school experience at UC Berkeley has been positively touched by many individuals. Special thanks to our department chair, Alper Atamtürk, for his valuable support. I am thankful for the department's rewarding trust in me to teach IEOR 151, Service Operations Design & Analysis. I appreciate the faculty of IEOR and College of Computing, Data Science & Society for enriching my skill set with their state-of-the-art courses. Many thanks to past and present IEOR staff for their constant assistance: Anayancy Paz, Diana Salazar, Keith McAleer, Heather Iwata, Ginnie Sadil, Goldie Negelev, and Rebecca Pauling. Warm thanks to my peers for all the shared moments and friendship: Caleb Bugg, Cecilie Greisgaard Petersen, Hansheng Jiang, Jehum Cho, Julie Mulvaney-Kemp, Junyu Cao, Mahan Tajrobehkar, Marie Pelagie Elimbi Moudio, Meng Qi, Mengxin Wang, Mo Liu, Pedro Hespanhol, Ruijie Zhou, Ruojie Zeng, Salar Fattahi, Sheng Liu, Yoon Lee, and Yunduan Lin. Further, I offer my great thanks to my colleagues at Apple and Meta for their valuable insights during my internships.

On my path to UC Berkeley, I owe a debt of gratitude to several distinguished academics. My genuine thanks go to my past master's advisors, Murat Köksalan and Banu Lokman, who have always been there to advise and encourage me. I am truly proud of starting my academic journey under their role modeling. I want to also thank Meral Azizoğlu, for her continuing support and tenderness, and the IE faculty at the Middle East Technical University, for the outstanding education that equipped and inspired my academic pursuits.

Last and most important, I am indebted beyond words to my family for always being my shelter, leverage, and biggest strength. All my accomplishments would be impossible without their unconditional love and lasting faith in my dreams. My deepest and most heartfelt thanks extend to them.

My brother, Çağrı, thank you for forever being kids together. My father, Halil, thank you for always being in my corner. I am most grateful for both of your assurances of unwavering protection under any circumstance.

My lifemate, Semih, thank you for completing my life and walking hand in hand. I cannot thank you enough for your persistent encouragement, love, and support in every way.

My dear grandmother, Gülhanım, thank you for your astonishing touch on our lives with your exceptional sense of humor and unique words of wisdom.

Finally, my mother, Hülya, the person whom I owe everything and hence deserves endless thanks. It was she who cultivated my desire for lifelong education and ensured every educational opportunity was available to me, thanks to her admirable resilience, love, and vision. I dedicate this dissertation to my mother and to the loving memory of my grandmother.

Chapter 1

Introduction

1.1 Motivation and Vision

Socio-technical systems relevant to Operations Management (OM) operate in dynamic environments shaped by several components such as information sharing, environmental and social disruptions, technological advancements, and supply chain complexities. All these factors interact in an intricate way, which introduces significant uncertainties into decision-making processes of agents (i.e., decision makers, stakeholders, controllers) within the system.

Various practical OM settings involve sequential problems where agents make decisions over time by observing stochastic outcomes (i.e., rewards, costs) from the interacted dynamic system. In these settings, uncertainty arises from numerous sources that can be very broadly classified into two categories:

i) *Imperfect model insight:* Imperfect information about the model (rewards and/or dynamics) of the interacted system.

ii) *Information asymmetries:* Unequal distribution of information among agents interacting in the system.

These uncertainties have traditionally been recognized as grand challenges for OM and have been well studied in the literature for diverse contexts. However, in today's era, traditional OM is transforming into *data-driven and sustainable OM*, primarily by two drivers: 1) the growing reliance on vast amounts of data to manage operations, and 2) the acknowledgment of the inevitable need for adopting sustainability practices. Due to this evolving landscape of data-driven and sustainable operations, information asymmetries are now taking on an escalated significance and becoming even more pronounced than in the past.

1.1.1 Information Asymmetries in Data-Driven and Sustainable Operations Management

In the present landscape, businesses are overflown with information from various sources. Further, they face the challenge of integrating sustainability into daily operations while simultaneously engaging with multiple stakeholders, such as governments, local communities, and investors, who may have conflicting goals across three key pillars: social equity, environment, and economy. As a result, it becomes essential to ensure that all stakeholders have access to accurate and timely data. Nevertheless, significant information disparities are triggered by limitations in the data shared among these stakeholders.

One example of this case is about transparency of supply chain emissions data, which is a major complication for companies striving for net-zero commitments. In the retail sector, a substantial portion of greenhouse gas emissions originates from third party suppliers involved across entire product cycle, spanning from manufacturing to packaging. Therefore, major retailers have initiated collaborative programs, such as Walmart's Project Gigaton (Walmart 2023) and Apple's Supplier Responsibility Program (Apple 2023), to encourage their suppliers to disclose emissions data. Yet, this transition to more transparent practices poses challenges to suppliers' traditional competitive advantage. Although retailers may ensure visibility into the specific sustainable input materials or energy-efficient technologies employed by suppliers, they often lack comprehensive insight into the specific expenses incurred by suppliers in implementing these practices (e.g., R&D investments, costs of certifications, integration of emission tracking systems). Addressing this information gap requires designing smart and adaptive contracts to learn suppliers' true objectives and incentivize them for emissions transparency while still maintaining their economic viability in the market.

Another related example is about the information asymmetries between program designers and owners of natural resources involved in Payment for Ecosystem Services (PES) programs (Salzman et al. 2018). PES programs establish a positive economic incentive mechanism by which governments and non-governmental organizations contract local landowners or service providers, such as private forest owners and farmers, to incentivize them to prevent environmental degradation, such as deforestation (Warnes et al. 2023). From the perspective of payment providers, the goal of these incentive mechanisms is to maximize the environmental benefits obtained from the amount of conserved forests while minimizing the required payments. However, a prevalent obstacle in effectiveness of these contracts is the private information of forest owners about their opportunity costs and the amount of deforestation they would choose without any incentives (Engel et al. 2016, Li et al. 2023). One promising solution to bridge this information gap and reveal the true willingness of forest owners is to design data-driven and performance-based incentive policies. These policies should incorporate the effects of learning from the conservation actions of forest owners over time.

1.1.2 Goals of This Dissertation

As evident from the presented examples, a primary challenge in sustainable OM lies in isolating the lack of transparency regarding the willingness to collaborate and the hidden economic drivers of the self-interested strategic agents. *This dissertation envisions to tackle this emerging challenge by devising novel estimation techniques coupled with proactive, intelligent, and learning-based incentive algorithms. The aim is to establish a foundational theory that neatly navigate information asymmetries hindering sustainability within socio-technical and data-driven OM systems.*

This goal necessitates holistic frameworks where traditional OM techniques are integrated with recent advances from statistical learning and data-driven decision-making. Throughout the course of a sequential decision problem, the fundamental trade-off occurs between a) leveraging the limited data at hand to extract latent information from the environment of interest (i.e., *exploration*), and b) making safer choices towards optimizing the ultimate objective (i.e., *exploitation*). As described earlier in this chapter, these latencies may manifest as i) imperfect information about the model, and ii) imperfect information about other agent(s) with whom interactions occur. Studying the exploration/exploitation trade-off requires jointly tackling both sources of uncertainties while proactively accounting for the potential externalities arising from strategic behaviors of the interacted agent(s).

To address this multi-faceted complexity, this dissertation offers solid theoretical foundations to stochastic data-driven models and computational algorithms developed for specific information asymmetric contexts, which have remained under-explored in the related literature. The proposed methodological frameworks revitalizes classical OM theory by drawing from a rich array of techniques in online sequential learning, statistics, optimization, model predictive control, multi-armed bandits, and principal-agent theory.

The distinct information asymmetry settings explored in this dissertation appear in real-world applications across diverse OM domains, including but not limited to healthcare analytics, transportation planning, and inventory management. Despite this wide relevance, the selected application examples attempt to place a particular emphasis on the domain of sustainable operations. In each chapter, these examples outline the practical implications of the proposed analytical frameworks, which are intentionally designed to be generic, explainable, and intuitively interpretable. This design avoids restrictive specialized technical assumptions and proves a rigorous theoretical basis along with functional insights. Building upon this core understanding, this dissertation foresees a trajectory for future research that can facilitate the extension and customization of the developed approaches for various settings.

1.2 Outline and Main Contributions

In light of the fundamental challenges and goals mentioned above, each chapter of this dissertation explores unique classes of sequential decision problems within particular contexts of imperfect model knowledge and asymmetric information. While these scenarios are broadly applicable, existing literature lacks a thorough examination of their technical complexities, leading to a limited understanding of regret limits and dynamic trade-offs for strategic agents.

The dissertation begins by (Chapter 2) investigating a nontrivial learning-based control setting characterized by the uncertainty of a nonlinear model, where both the system dynamics and reward function are partially-unknown. Building on this rigorous technical analysis, the subsequent chapters constitute the major focus of the dissertation. Chapters 3 and 4 address the design of data-driven and adaptive incentive mechanisms under novel information asymmetry settings, layered on top of the uncertainty in the reward model. The dissertation concludes in Chapter 5 with prosperous future research avenues and includes three appendices providing detailed proofs of the technical results presented in each chapter.

Each chapter incorporates material that has either been published or is currently undergoing the review process at peer-reviewed journals. The main contributions of each chapter, along with citation information for the associated papers, are summarized below. As a remark, the first chapter is designed to be self-contained and can be read independently if desired. The content in the last two chapters is suggested to be covered in sequential order.

- **Chapter 2** explores nonlinear and non-stationary systems with imperfect knowledge of both the reward function and system dynamics, motivated by a number of real-world applications such as enhancing the energy-efficiency of Heating, Ventilation, Air-Conditioning (HVAC) systems and optimizing clinical inventory management. For this setting of partial model insight, this chapter addresses the exploration-exploitation trade-off by integrating learning and adaptation into the model predictive control (MPC).

  Though this trade-off has been extensively studied for linear systems; it is less well-studied for learning-based control of nonlinear systems. A major challenge in the nonlinear setting is that, unlike the linear case, there is no explicit characterization of an optimal controller for a given set of model parameters. We propose the use of a finite-horizon oracle controller with perfect knowledge of all parameters as a reference for optimal policy. First, this allows us to propose a new regret notion with respect to this finite-horizon oracle. Second, this allows us to develop non-myopic policies in the context of learning-based MPC and multi-armed bandits (MAB's) that attain low regret (i.e., square-root regret up to a log-squared factor). The conducted statistical and control-theoretic analyses bridge system stability and policy regret, further supported by the numerical results on a HVAC model.

  Related Paper:

  Dogan I, Shen ZJM, and Aswani A (2023) Regret Analysis of Learning-Based MPC With Partially-Unknown Cost Function. *IEEE Transactions on Automatic Control*, doi: 10.1109/TAC.2023.3328827.

- **Chapter 3** sheds light on a nontrivial information asymmetry faced in data-driven and sequential incentive design in the context of repeated principal-agent games. In practice, unlike previous models, the principal can often only observe the actions, not the rewards, of a self-interested agent. This arises in numerous applications, such as routing incentives for sustainable transportation planning and personalized incentives for medical adherence.

  Designing policies in this setting is challenging because existing estimation methods cannot directly learn the agent's rewards. Existing work often overlook the principal's trade-off between consistently learning the agent's rewards (i.e., exploration) and maximizing their own rewards through adaptive incentives (i.e., exploitation). As a result, this scenario, as well as similar ones, remains under-explored. To bridge the gap, we introduce the *hidden rewards* model within a MAB framework, where: the principal gives a different incentive for each bandit arm, the agent picks an arm to maximize its own expected reward plus incentive, and the principal observes the chosen arm and receives a distinct reward (from the agent's). In this chapter, we consider agents with *perfect-knowledge* of their own expected rewards for each arm. First, we design a statistically consistent estimator for the agent's expected rewards. Since our estimator uses as data the sequence of incentives offered and subsequently chosen arms, it can be regarded as an analogy of online inverse optimization in MAB's. Then, we construct a policy that we prove achieves a low regret (i.e., square-root regret up to a log factor) and conclude with experiments demonstrating its applicability to an instance of sustainable route choice problem.

  Related Paper: Dogan I, Shen ZJM, and Aswani A (2023) Repeated principal-agent games with unobserved agent rewards and perfect-knowledge agents. *arXiv preprint arXiv:2304.07407.*

- **Chapter 4** studies the repeated *hidden rewards* setting in an even more challenging environment, where the agent must learn their expected rewards over time by tackling a MAB problem. Relevant to deforestation incentives in PES and renewable energy aggregation contracts, this setup is considerably more complex due to uncertainty in agent choices, potentially misleading the principal. We propose a novel theoretical framework that facilitates the dynamic and sequential externalities between two separate learning algorithms trained in parallel by these two strategic parties. We present a new estimator for the agent's reward expectations in bounded continuous spaces that is formulated as a tractable optimization model leveraging the agent's noisy choices. The estimator is then united with a data-driven incentive policy to address the principal's trade-off while ensuring high-probability incentive compatibility for the agent. We provide rigorous guarantees for the estimator's finite-sample consistency and the policy's regret bound. These theoretical results remain robust without restricting the type of the agent's algorithm and are justified by simulations performed in the context of green energy aggregation contracts.

  Related Paper:

  Dogan I, Shen ZJM, and Aswani A (2023) Estimating and Incentivizing Imperfect-Knowledge Agents with Hidden Rewards. *arXiv preprint arXiv:2308.06717.*

Chapter 2

Regret Analysis of Learning-Based Model Predictive Control with Partially-Unknown Cost Function

2.1 Introduction

Model predictive control (MPC) has been used for a wide range of applications, including: sustainable crop production (Hu and You 2022), carbon tax policies for greenhouse gas emissions (Chu et al. 2012), chemical process controls (Eaton and Rawlings 1992, Arefi et al. 2006), power electronics (Vazquez et al. 2014), aerospace systems (Di Cairano and Kolmanovsky 2018, Eren et al. 2017), and heating, ventilation, and air-conditioning (HVAC) systems (Aswani et al. 2011, Maasoumy and Sangiovanni-Vincentelli 2012). More recent work has studied the design of adaptive or learning-based MPC (LBMPC) schemes that ensure constraint satisfaction in the presence of models that are updated as more system measurements become available (Negenborn et al. 2004, Aswani et al. 2013, Karnchanachari et al. 2020, Gros and Zanon 2020).

In the context of learning-based control, the constraint satisfaction and robustness are often provided by leveraging LBMPC while the performance is optimized by directly utilizing sequential data-driven approaches, such as reinforcement learning (RL) (Hewing et al. 2020). However, the related research that have designed RL algorithms for nonlinear discrete-time control systems are limited (Koller et al. 2018, Wabersich and Zeilinger 2018, Gros and Zanon 2019, Abbasi-Yadkori et al. 2019, Chen et al. 2019, Kakade et al. 2020, Agarwal et al. 2020, Boffi et al. 2021). Some of them are concerned with asymptotic stability of the closed-loop system (known as safety in RL), while the others provide finite-time regret bounds. These bounds quantify the difference between the control performance of the data-driven control policy and that of an oracle control policy with perfect knowledge of the model uncertainty.

In this line of RL research for nonlinear learning-based control, this chapter aims to better connect these two areas by jointly presenting a rigorous regret analysis and proving stability for the proposed adaptive control policies. There are two main novel aspects of our work compared to the existing studies. First, we highlight that comparing finite-horizon policies with different horizon lengths leads to ambiguous regret notions in evaluation of learning-based control policies. For this reason that we will discuss in detail, we propose a new regret notion that compares a finite-horizon learning-based policy with a finite-horizon oracle controller as the benchmark. Second, we bound this regret notion for a class of learning-based control policies for which we prove constraint satisfaction. An important aspect of our regret analysis is that we have to consider the stability of our policy when bounding the regret. In this sense, our analysis draws a connection between stability of the nonlinear control system and regret performance of the learning policy.

2.1.1 Partially-Unknown Cost Function

The classical MPC setup assumes that the system dynamics are exactly known and that a cost function that is to be minimized over a finite horizon is also exactly known. However, there are many real-world applications where the system dynamics or cost function may be partially-unknown, and it is such systems that motivate the study of integrating learning or adaptation into the MPC setting. Because such applications motivate this chapter, we briefly describe some examples.

2.1.1.1 Energy-Efficiency of Heating, Ventilation, Air-Conditioning (HVAC)

Because HVAC systems comprise a substantial portion of overall building energy usage, there has been a substantial body of work on the use of MPC to improve the energy-efficiency of HVAC Aswani et al. (2011), Oldewurtel et al. (2012), Ma et al. (2011), Afram and Janabi-Sharifi (2014), Ostadijafari and Dubey (2019), Bianchini et al. (2019), Chen et al. (2020), Fang et al. (2020), Kumar et al. (2020), Raman et al. (2020). However, these past works typically assume that a cost function that precisely characterizes the trade-off between energy-efficiency and occupant comfort is exactly known. In practice, the quantity of trade-off is different for each occupant and is *a priori* unknown to the controller. It then makes sense from an applications standpoint to try to learn an ideal trade-off from occupant-reported data Aswani et al. (2018) and then adapt the MPC operation in response, which is an example of MPC with a partially-unknown cost function.

2.1.1.2 Clinical-Inventory Management

Inventory management in hospitals involves periodically restocking drugs and medical supplies, and MPC approaches to inventory management Velarde et al. (2014), Schildbach and Morari (2016), Jurado et al. (2016), Maestre et al. (2018), Garcia et al. (2020) are powerful because they naturally capture the dynamics inherent in consuming and purchasing drugs and supplies. However, these past works typically assume that the dynamics of consumption are completely characterized. But in hospitals, the demand for drugs and supplies is difficult to characterize because of unforeseeable events like medical emergencies. Considering the practical implications, it becomes essential to try to learn more information about the demand arising from such events and then adapt the MPC operation in response, which is an example of MPC with learning for the partially-unknown dynamics.

2.1.2 Exploration/Exploitation Trade-Off

An inherent challenge in LBMPC is that of dual-control, which is the problem of jointly optimizing the control to minimize a cost function (of the states and control inputs) and to steer the system in a way that provides more information about any unknown system dynamics or cost function parameters (Mesbah 2018). This challenge is often termed the *exploration/exploitation* trade-off. This trade-off has been formally studied in the setting of stationary or weakly-nonstationary multi-armed bandits (MAB's) (Thompson 1933, Agrawal and Goyal 2013, Mintz et al. 2017), RL for finite Markov chains (Heger 1994, Pashenkova et al. 1996, Gaskett 2003, Jaksch et al. 2010, Moldovan and Abbeel 2012, Biyik et al. 2019, Budd et al. 2020), and RL for linear control systems (Bradtke et al. 1994, Vrabie et al. 2009, Kiumarsi-Khomartash et al. 2014, Klenske and Hennig 2016, Kiumarsi-Khomartash et al. 2017, Cohen et al. 2019, Lale et al. 2020, Simchowitz and Foster 2020).

Extending these ideas to study the exploration/exploitation trade-off for RL of nonlinear and non-stationary control systems is nontrivial. Most work on MAB's assumes stationarity or weak-stationarity. This is partly because computing the optimal policy for non-stationary bandits is PSPACE-hard (Papadimitriou and Tsitsiklis 1999). Similarly, in the setting of RL of control systems, past work on the exploration/exploitation trade-off for nonlinear systems is limited (Koller et al. 2018, Gros and Zanon 2019, Kakade et al. 2020, Wabersich and Zeilinger 2020, Fan and Ming 2020, Boffi et al. 2021). The reason is that the optimal controller for linear systems with a quadratic cost is completely characterized by the Algebraic Ricatti Equation, which allows this body of work to convert the RL problem into simply a parameter estimation problem. However, for nonlinear control systems there is no such simple characterization of the optimal controller, and so alternative approaches are needed. To bridge these gaps, this chapter designs a learning-based controller for nonlinear and non-stationary systems where the policy explores to improve the estimation methodology embedded in the learning mechanism.

2.1.3 Main Contributions and Chapter Outline

Sect. 2.2 begins with technical preliminaries. In Sect. 2.3, we define our setup, formalize our idea of *non-myopic exploitation*, and prove safety properties for a class of control policies. Next, in Sect. 2.4, we introduce an oracle finite-horizon control policy as a reasonable surrogate to the optimal finite-horizon control policy for our non-myopic exploitation problem. We then introduce a new regret notion called as the *N-step dynamic regret* with respect to this oracle finite-horizon controller.

Sect. 2.5 presents our finite sample statistical consistency analysis for the estimation step of our policy for the unknown cost and system parameters. Then, in Sect. 2.6, we develop a novel *non-myopic ϵ-greedy algorithm* that keeps our system stable and safe. We prove Lipschitzian stability of the best input chosen at each exploitation step and provide a rigorous proof for the asymptotic $O(\sqrt{T}(\log T)^2)$ bound on the N-step dynamic regret.

In Sect. 2.7, we support our theoretical results with numerical experiments that we conduct on a model of HVAC systems with partially-unknown cost and system dynamics functions. Our experiments reveal the effectiveness of performing non-myopic exploitation and show that our finite-horizon non-myopic policy achieves the proved regret bound.

2.2 Preliminaries

This section introduces the related mathematical definitions and notation used in this chapter. Let $\mathcal{U}$ be a polytope in $\mathbb{R}^n$. This set can be represented as intersection of a set of half-spaces specified by a set of linear inequalities (Borrelli et al. 2017), i.e., $\mathcal{U} = \{x : P_i x \leq q_i, i = 1, \ldots, d\}$ where $P_i \in \mathbb{R}^{d\times n}$ and $q_i \in \mathbb{R}^d$.

Let $\mathcal{U}, \mathcal{V}$ be two sets. The linear transformation of $\mathcal{U}$ by matrix $\mathcal{R}$ is given by $\mathcal{R}\mathcal{U} = \{\mathcal{R}u : u \in \mathcal{U}\}$. Their Minkowski sum (Schneider 2013) is defined as $\mathcal{U} \oplus \mathcal{V} = \{u + v : u \in \mathcal{U}; v \in \mathcal{V}\}$, and their Pontryagin set difference (Kolmanovsky and Gilbert 1998) is defined as $\mathcal{U} \ominus \mathcal{V} = \{u : u + \mathcal{V} \subseteq \mathcal{U}\}$. Some useful properties of these definitions include: $\mathcal{R}(\mathcal{U} \ominus \mathcal{V}) \subseteq \mathcal{R}\mathcal{U} \ominus \mathcal{R}\mathcal{V}$ and $(\mathcal{U} \ominus \mathcal{V}) \oplus \mathcal{V} \subseteq \mathcal{U}$.

2.3 Problem Formulation

Let $x_t \in \mathbb{R}^n$ be states and $u_t \in \mathbb{R}^q$ be inputs. We assume $x_t \in \mathcal{X}$ and $u_t \in \mathcal{U}$ are constrained by (compact) polytopes $\mathcal{X}, \mathcal{U}$. The true system dynamics are

$$x_{t+1} = f(x_t, u_t, \theta_0) = Ax_t + Bu_t + g(x_t, u_t, \theta_0) \tag{2.1}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times q}$, $\theta_0 \in \Theta$ for some compact set $\Theta \subseteq \mathbb{R}^p$, and the nonlinear function $g(\cdot, \cdot, \theta) : \mathbb{R}^n \times \mathbb{R}^q \to \mathbb{R}^n$ is parameterized by $\theta \in \Theta$. We assume that

$$\big\{g(x, u, \theta_0) : x \in \mathcal{X}, u \in \mathcal{U}\big\} \subseteq \mathcal{W}, \tag{2.2}$$

for some bounded polytope $\mathcal{W}$, and $A, B, g, \mathcal{W}, \Theta$ are known but θ_0 is not known to the controller. We define $w_t = g(x_t, u_t, \theta_0)$, and note $w_t \in \mathcal{W}$ by assumption. The intuition is we have a nominal linear model and a partially-unknown, nonlinear correction.

At each time step t, the controller receives a stochastic reward r_t from distribution $\mathbb{P}_{x_t,u_t,\theta_0}$ with probability density function $p(r|x_t, u_t, \theta_0)$ and expectation $\mathbb{E}\, r_t = h(x_t, u_t, \theta_0)$. We assume h is parametrically unknown (θ_0 is unknown). This setup can handle stochastic costs c_t (as opposed to rewards) by setting $r_t = -c_t$. We standardize our notation for rewards.

The control problem we study in this chapter is to sequentially choose inputs to maximize expected total reward at the end of a finite time horizon $\mathcal{T} = \{0, \dots, T\}$. At time t, the controller has access to all past rewards $r_0, \dots, r_{t-1}$, past control inputs $u_0, \dots, u_{t-1}$, and past and current states $x_0, \dots, x_t$. As a result, any given control policy $u_t = \Lambda_t(\mathcal{F}_t)$ will be a sequence (with respect to t) of functions of this past information

$$\mathcal{F}_t = \{r_0, \dots, r_{t-1}, u_0, \dots, u_{t-1}, x_0, \dots, x_t\}. \tag{2.3}$$

We distinguish between different control policies by using different superscripts for the sequence of functions Λ_t for $t = 0, \dots, T$ characterizing the policy.

2.3.1 Learning-Based MPC Formulation

LBMPC maintains two different models of the control system: a learned model to enhance performance and a nominal model to provide robustness Aswani et al. (2013). Because A, B are known in our setup, the controller uses as its nominal model dynamics

$$\bar{x}_{t+k+1|t} = A\bar{x}_{t+k|t} + Bu_{t+k|t}, \tag{2.4}$$

where $\bar{x} \in \mathbb{R}^n$ is system state of the nominal model. The "$|t$" notation denotes the initial condition is taken to be $\bar{x}_{t|t} = x_t$, where x_t is the true state at time t. Because $g(\cdot, \cdot, \theta)$ is also known, the controller uses as its learned model the dynamics

$$\tilde{x}_{t+k+1|t} = A\tilde{x}_{t+k|t} + Bu_{t+k|t} + g(\tilde{x}_{t+k|t}, u_{t+k|t}, \widehat{\theta}_t), \tag{2.5}$$

where $\tilde{x}$ is the system state of the learned model and $\widehat{\theta}_t$ is the controller's estimate of θ_0 at time t. Here, LBMPC learns the true model dynamics by updating its estimate of the parameters θ_0 as more system state measurements from the system becomes available.

Before presenting the formulation, we must first discuss the terminal set used for the MPC. Assuming that (A, B) is stabilizable, there exists a constant state-feedback matrix $K \in \mathbb{R}^{q \times n}$ such that $(A + BK)$ is Schur stable. We assume $\Omega \subseteq \mathcal{X}$ is a maximal output admissible disturbance invariant set (Kolmanovsky and Gilbert 1998) meaning that for some stabilizing K it satisfies: a) $\Omega \subseteq \{\overline{x} : \overline{x} \in \mathcal{X} : K\overline{x} \in \mathcal{U}\}$ (*constraint satisfaction*) and b) $(A + BK)\Omega \oplus \mathcal{W} \subseteq \Omega$ (*disturbance invariance*). The intuition is that Ω is a set of states satisfying the constraints $\mathcal{X}$ for which there exists a feasible action keeping the true state within Ω despite the uncertainty of the nominal model. Several algorithms (Kolmanovsky and Gilbert 1998, Limon et al. 2009, Rakovic and Baric 2010, Wang et al. 2021) can compute this set, and so we assume Ω is available to the controller.

With the set Ω, we consider an (simplified) LBMPC variant that maximizes the expected N-step reward. Our results can be generalized straightforwardly to the full formulation (Aswani et al. 2013), but we do not consider this as it adds substantial notational complexity that hinders showcasing the stochastic aspects of our setting. The LBMPC formulation of a finite-horizon N is

$$\begin{aligned} V_N(x_t,\theta,t) = &\max \textstyle\sum_{k=0}^{N} h(\tilde{x}_{t+k|t}, u_{t+k|t}, \theta) && \\ \text{s.t. } & \bar{x}_{t+k+1|t} = A\bar{x}_{t+k|t} + Bu_{t+k|t} && k \in \langle N-1\rangle \\ & \tilde{x}_{t+k+1|t} = A\tilde{x}_{t+k|t} + Bu_{t+k|t} + g(\tilde{x}_{t+k|t}, u_{t+k|t}, \theta) && k \in \langle N-1\rangle \\ & \bar{x}_{t+k|t} \in \mathcal{X} && k \in [N] \\ & u_{t+k|t} \in \mathcal{U} && k \in \langle N\rangle \\ & \bar{x}_{t+1|t} \in \Omega \ominus \mathcal{W},\ \bar{x}_{t|t} = \tilde{x}_{t|t} = x_t && \end{aligned} \tag{2.6}$$

where $\langle k\rangle = \{0,\dots,k\}$ and $[k] = \{1,\dots,k\}$. The difference between this simplified variant and the full formulation is that here we apply the invariant set Ω at the first time step, an idea previously used in (Aswani et al. 2012), whereas the full formulation uses a robust tube framework to apply Ω at the N-th time point. Our results apply to the above LBMPC formulation and may generalize to the similar variants, but it is unclear if they would generalize to other LBMPC forms without further study.

2.3.2 Safety of Learning-Based MPC Variant

Because applying the invariant set to the first time point in an MPC formulation is nonstandard, we first formally prove that this LBMPC variant ensures recursive properties of robust constraint satisfaction and robust feasibility.

Theorem 2.1 *Suppose $\{u_{t|t},\dots,u_{t+N|t}\}$ are feasible for $V_N(x_t,\theta,t)$ for any θ. If Ω is a maximal output admissible disturbance invariant set, then choosing $u_t = u_{t|t}$ ensures that we have: a) $x_{t+1} \in \mathcal{X}$ (robust constraint satisfaction) and b) there exist values $\{u_{t+1|t+1},\dots,u_{t+N|t+1}\}$ that are feasible for $V_N(x_{t+1},\theta',t+1)$ for any θ' (robust feasibility).*

The complete proof of this theorem is provided in Appendix A.1.1.

Remark 2.1 *An important feature of the above result is that there is no required relationship between the θ and θ'. Since estimates of the θ are updated through learning, this shows that the safety properties of this LBMPC variant are decoupled from the design of the learning-process.*

2.3.3 Technical Assumptions

Our learning-based control problem is well-posed under certain regularity assumptions. We will design a policy under the assumptions described below.

Assumption 2.1 *The rewards r_t are conditionally independent given θ_0 and x_0, or equivalently, given θ_0 and the complete sequence of $\{u_0,\dots,u_t,x_0,\dots,x_t\}$.*

The above assumption is similar to the assumption of independence of rewards in the stationary MAB model. This assumption implies that $r_t|\{x_t, \theta_0\}$ is independent of $r_{t'}|\{x_{t'}, \theta_0\}$ for any two time points $t \neq t'$.

Assumption 2.2 *The log-likelihood ratio $\ell(r, x, u; \theta, \theta') = \log \frac{p(r|x,u,\theta)}{p(r|x,u,\theta')}$ of $\mathbb{P}_{x,u,\theta}$ is locally $L_{\ell,x}$-Lipschitz continuous with respect to x on the compact set $\mathcal{X}$ for $\theta, \theta' \in \Theta$, $u \in \mathcal{U}$.*

This ensures continuity of the reward distribution with respect to the parameters. If two parameter sets are close to each other in value, then the resulting distributions will also be similar.

Assumption 2.3 *The distribution $\mathbb{P}_{x,u,\theta}$ for all $x \in \mathcal{X}, u \in \mathcal{U}$, and $\theta \in \Theta$ is sub-Gaussian with parameter σ, and either $p(r|x, u, \theta)$ has a finite support or $\ell(r, u; x, \theta, x', \theta')$ is locally $L_{\ell,r}$-Lipschitz with respect to r.*

This assumption ensures sample averages are close to their means and is satisfied by many distributions (e.g., Gaussian with known variance). Our last condition ensures that the system dynamics and the reward expectation function are well-behaved and that the states do not change too rapidly.

Assumption 2.4 *Repeated composition of the true dynamics with itself up to $N-1$ times, $f^{t+k}(x_t, u_{t|t}, \ldots, u_{t+k|t}, \theta)$, is Lipschitz continuous with respect to $x_t \in \mathcal{X}$ and $u_{t+k|t} \in \mathcal{U}$ with constants $L_{f,x}$ and $L_{f,u}$, respectively. Besides, the expectation $h(x_t, u_t, \theta)$, for $u_t = \Lambda_t(\mathcal{F}_t)$ in (2.3), is Lipschitz continuous with respect to $x_t \in \mathcal{X}$ and $u_t \in \mathcal{U}$ with constants $L_{h,x}$ and $L_{h,u}$, respectively, for all $\theta \in \Theta$.*

2.4 The N-Step Dynamic Regret

Our interest is in evaluating the performance of an LBMPC *exploitation* policy for a given $N \leq T$ that is $\Lambda_t^{E,N}(\mathcal{F}_t) = u^*_{t|t}(\widehat{\theta}_t)$ for the corresponding value from the maximizer of $V_N(x_t, \widehat{\theta}_t, t)$ where $\widehat{\theta}_t$ are the control policy's estimates of the unknown θ_0. Data-driven policies are often evaluated by comparing performance to a benchmark policy, and it is typical to benchmark using the optimal policy (Garivier and Moulines 2008, Besbes et al. 2014, Bouneffouf and Féraud 2016). In our setting, the optimal policy is a sequence of functions $\Lambda_t^*(\mathcal{F}_t)_{t=0}^T$ maximizing $\sum_{t=0}^T h(x_t, u_t, \theta_0)$ subject to the knowledge available to the control policy (which does not include θ_0). However, computing optimal policies for the problems we consider is PSPACE-hard (Papadimitriou and Tsitsiklis 1999). Even their structure is not known for our setup, including for the special case of linear dynamics and quadratic cost function with unknown coefficients.

An alternative benchmark is an oracle policy that has perfect knowledge of θ_0. Specifically, we will use the LBMPC *oracle* policy that is $\Lambda_t^{O,N}(\mathcal{F}_t) = u^*_{t|t}(\theta_0)$ for the corresponding value from the maximizer of $V_N(x_t, \theta_0, t)$ as defined in (2.6). However, there are two subtleties that have to be discussed.

The first subtlety is that the horizon length of the LBMPC oracle policy could potentially be different than the horizon length of the LBMPC policy. However, using different control horizon lengths can lead to different sums of expected rewards over the entire control horizon $\mathcal{T}$. Though this behavior is well known within the MPC community, its implication on evaluating learning-based control policies has not been previously appreciated. The implication is that comparing policies with different horizon lengths leads to a poorly-defined regret notion, and that we should compare oracle policies and learning-based policies with the same finite-horizon.

The second subtlety is that the presence of nonlinear dynamics in our setup means the state trajectory of a system always controlled by a benchmark policy can be very different than that of a system always controlled by a learning policy, even if the learning policy converges towards the benchmark policy.

For this reason, we define a regret notion to compare a finite-horizon benchmark policy to a finite-horizon learning-based policy. We consider an ϵ-greedy policy $\Lambda_t^{\epsilon,N}$ that uses the LBMPC policy $\Lambda_t^{E,N}$ at each greedy exploitation step. Let x_t, u_t be the state and input for the system as controlled by the oracle policy $\Lambda_t^{O,N}$, and let x'_t, u'_t be the state and input for the system as controlled by the ϵ-greedy policy $\Lambda_t^{\epsilon,N}$. Then, the expected *N-step dynamic regret* is defined as

$$R_{N,T} = \textstyle\sum_{t=0}^{T} h(x_t, \Lambda_t^{O,N}(\mathcal{F}_t), \theta_0) - h(x'_t, \Lambda_t^{\epsilon,N}(\mathcal{F}'_t), \theta_0) \tag{2.7}$$

where $\mathcal{F}_t$ is as defined in (2.3) and $\mathcal{F}'_t$ is as defined in (2.3) with x', u' replacing x, u. This definition is closely related to the traditional dynamic regret (Zinkevich 2003, Hall and Willett 2013), and the novel aspect of ours is that it compares two N-step finite-horizon policies.

2.5 Parameter Estimation

Let the variables $\{r_i\}_{i=0}^{t-1}$ be the actual observed values of the rewards up to time t. Using Assumption 2.1, the joint likelihood $p(\{r_i\}_{i=0}^{t-1}|x_0, \ldots, x_t, u_0, \ldots, u_{t-1}, \theta)$ can be expressed as $\prod_{i=0}^{t-1} p(r_i|x_i, u_i, \theta)P(x_i|x_{i-1}, \theta)$. Here, the one step transition likelihood $P(x_i|x_{i-1}, \theta)$ is a degenerate distribution with all probability mass at x_i, by perpetuation of the dynamics $f(x_i, u_i, \theta)$ with initial conditions x_{i-1}. Thus, the maximum likelihood estimator (MLE) for θ is

$$\begin{aligned} \widehat{\theta}_t \in \arg\min_{\theta\in\Theta} &- \textstyle\sum_{i=0}^{t-1} \log p(r_i|x_i, u_i, \theta) \\ \text{s.t. } & x_{i+1} = f(x_i, u_i, \theta)\ \forall i \in \{0, \ldots, t-1\} \end{aligned} \tag{2.8}$$

2.5.1 Solving the MLE Problem

The MLE problem (2.8) can be computed using optimization, dynamic programming, or various filtering approaches that have been proposed for different problem structures. Since the

MLE problem has been extensively studied in the literature, we will not study its numerical computation in this chapter, but rather we give some helpful references in this section.

The Kalman Filter (KF) is a recursive estimator for linear-quadratic (LQ) discrete-time control systems Kalman (1960). In more complex systems with non-Gaussian distributions and nonlinear state transitions, the Extended KF is one of the best-known estimators that is based on linearization of the state equations at each time step Jazwinski (2007), Anderson and Moore (2012). Another well-known filtering approach is the Particle Filter (PF), also known as Sequential Monte Carlo (SMC) method, that can deal with non-Gaussian and nonlinear sequential estimation problems by computing the posterior distributions of the states Kitagawa (1996), Liu and Chen (1998), Doucet et al. (2000), Durbin and Koopman (2000), Doucet et al. (2001). Another approach proposed for maximum likelihood parameter estimation is known as the Expectation-Maximization (EM) Algorithm. It was first introduced in Dempster et al. (1977), and has been used extensively for parameter estimation Ghahramani and Hinton (1996), Shumway and Stoffer (1982).

For practical purposes, these efficient approaches motivate the use of MLE in our policy. Further, if the controller did not have perfect state measurements, we could use the noisy state data to estimate the dynamics in the constraints of (2.8) (Amelin and Granichin 2012, Kalmuk et al. 2017), which would also alleviate any potential infeasibility issues of the MLE.

2.5.2 Concentration Bounds

We further analyze the concentration properties of the solution to (2.8) and take an approach to the theoretical analysis that generalizes that of (Mintz et al. 2017). We begin by introducing the notion of trajectory Kullback–Leibler (KL) divergence. Since this problem includes the joint distribution of a trajectory of values, the concentration bound for the parameter estimates is computed with regards to the trajectory KL divergence.

Definition 2.1 *The trajectory Kullback–Leibler (KL) divergence between the parameter trajectories $\theta, \theta' \in \Theta$ with the same input sequence $\Pi_T = \{u_t\}_{t=0}^T$ is*

$$D_{\Pi_T}(\theta||\theta') = \sum_{i=0}^{T} D_{KL}(\mathbb{P}_{f^i(x_0,\Pi_i,\theta),u_i,\theta}||\mathbb{P}_{f^i(x_0,\Pi_i,\theta'),u_i,\theta'}) \tag{2.9}$$

where Π_i is the given sequence of control inputs from time 0 to i, f^i is the repeated composition of the dynamics f with itself i times subject to Π_i, and D_{KL} is the standard KL-Divergence.

We have an *observability* assumption with the implication that the distance between two different parameters $\theta, \theta' \in \Theta$ is bounded proportional to their trajectory KL divergence.

Assumption 2.5 *For a given input sequence Π_T and parameters $\theta \neq \theta'$, if $D_{\Pi_T}(\theta||\theta') \leq \delta$, then $\|\theta - \theta'\| \leq C\delta$ for $C > 0$.*

We next reformulate the MLE problem (2.8) by removing the state dynamics constraints through repeated composition of f, that is

$$\widehat{\theta}_t \in \arg\min_{\theta\in\Theta} \frac{1}{t-1}\sum_{i=0}^{t-1} \log \frac{p(r_i|f^i(x_0,\Pi_i,\theta_0),u_i,\theta_0)}{p(r_i|f^i(x_0,\Pi_i,\theta),u_i,\theta)} \tag{2.10}$$

This reformulation is helpful for our theoretical analysis since for fixed θ, the expected value of the above objective function under $\mathbb{P}_{x_0,\Pi_T,\theta_0}$ is simply $\frac{1}{t-1}D_{\Pi_T}(\theta_0||\theta)$. Hence, we can interpret the MLE problem as minimizing the trajectory KL divergence between the distribution of potential sets of parameters and that of the true parameter set. This interpretation is helpful for us to derive our concentration inequalities. For conciseness of the focus of our analysis in this chapter, we present the final concentration bound for $\widehat{\theta}_t$ and do not include its proof since it largely follows by the theoretical arguments in Mintz et al. (2017).

Theorem 2.2 *For any constant $\zeta > 0$, we have the bound that*

$$P\bigg(\frac{1}{t-1}D_{\Pi_t}(\theta_0||\widehat{\theta}_t) \leq \zeta + \frac{c_f(d_x,d_\theta)}{\sqrt{t-1}}\bigg) \geq 1 - \exp\bigg(-\frac{\zeta^2(t-1)}{2L_{\ell,r}^2\sigma^2}\bigg) \tag{2.11}$$

where the constant

$$c_f(d_x,d_\theta) = 8L_{f,x}L_{\ell,x}\mathrm{diam}(\mathcal{X})\sqrt{\pi} + 48\sqrt{2}(2)^{\frac{1}{d_x+d_\theta}}L_{f,x}L_{\ell,x}\mathrm{diam}(\Theta\times\mathcal{X})\sqrt{\pi(d_x+d_\theta)} \tag{2.12}$$

depends upon d_x and d_θ (dimensionalities of $\mathcal{X}$ and Θ), and $\mathrm{diam}(\mathcal{X}) = \max_{x,y\in\mathcal{X}}\|x-y\|_2$.

We will use this concentration inequality to prove the regret bound of our non-myopic ϵ-greedy policy that we present next.

2.6 Non-Myopic ϵ-Greedy Approach

We develop a *non-myopic ϵ-greedy algorithm* that can achieve effective regret bounds for the non-stationary and nonlinear LBMPC model introduced in Section 2.3. Our choice of algorithm aims to draw a connection between the control and MAB literature. A possible alternative could be adding additive noise to the control inputs which we leave as a future work. When compared with the other well-known MAB strategies, Thompson Sampling (TS) and Upper Confidence Bound (UCB), ϵ-greedy is significantly easier from a computational standpoint for combining with the LBMPC formulation of our non-myopic exploitation problem. TS requires characterization of the posterior distribution which is indeed not possible under the general dynamics considered. Similarly, UCB requires being able to compute the confidence bounds which is not feasible in this framework. Hence, those strategies are not practical for the kinds of applications we are interested in.

The pseudocode of the *non-myopic ϵ-greedy algorithm* can be found in Algorithm 1. Our algorithm explores randomly according to a non-stationary stochastic process. The initial state x_0 is an arbitrary point from the state space $\mathcal{X}$. At each time $t \in \mathcal{T}$, the algorithm samples a Bernoulli variable s_t based on the exploration probability ϵ_t. If $s_t = 1$, it performs pure exploration by choosing only a single control input $u_{t|t}$. To ensure robust constraint satisfaction and feasibility after exploration, this input is uniform randomly chosen from a set $\overline{\mathcal{U}}(x_t) = \{u : Ax_t + Bu \in \Omega \ominus \mathcal{W}, u \in \mathcal{U}\}$. We note that choosing the remaining N inputs $u_{t+k|t}$ for $k \in [N]$ is redundant at an exploration step since they do not affect the system state and parameter estimation part of our policy. If $s_t = 0$, the algorithm performs a greedy exploitation step by first computing the MLE of the model parameters θ_0. Using these estimates, the algorithm then solves the non-myopic exploitation problem $V_N(x_t, \widehat{\theta}_t, t)$ (2.6) to select the sequence of inputs $\{u^*_{t|t}(\widehat{\theta}_t), \ldots, u^*_{t+N|t}(\widehat{\theta}_t)\}$ with the highest MLE estimated N-step reward at time t. At the end of each time step, the algorithm observes the updated state x_{t+1} and realized reward r_t after applying the chosen input $\Lambda_t^{\epsilon,N}(\mathcal{F}_t)$ to the system.

Algorithm 1 Non-Myopic ϵ-Greedy Algorithm

```
1: Set: $c > 0$ and $x_0 \in \mathcal{X}$
2: for $t \in \mathcal{T}$ do
3:     Set: $\epsilon_t = \min\left\{1, c/t\right\}$
4:     Sample: $s_t \sim$ Bernoulli($\epsilon_t$)
5:     if $s_t = 1$ then
6:         Randomly select: $u_{t|t} \in \overline{\mathcal{U}}(x_t)$
7:         Set: $\Lambda_t^{\epsilon,N}(\mathcal{F}_t) = u_{t|t}$
8:     else
9:         Compute: $\widehat{\theta}_t$ from (2.8)
10:        Compute: $u^*_{t|t}(\widehat{\theta}_t)$ from $V_N(x_t, \widehat{\theta}_t, t)$ (2.6)
11:        Set: $\Lambda_t^{\epsilon,N}(\mathcal{F}_t) = u^*_{t|t}(\widehat{\theta}_t)$
12:    Observe: $r_t$ and $x_{t+1}$
```

Remark 2.2 *If $\mathcal{W}, \mathcal{X}, \mathcal{U}$ are all polytopes, then Ω can be approximated by a polytope arbitrarily well. Then, $\Omega \ominus \mathcal{W}$ is also a polytope. As a result, line 6 involves randomly picking an element from a polytope that can be done in a computationally efficient way using standard algorithms.*

For clarity, we consider a randomization at the initial system state, and then assume noise-free transitions for the subsequent states which is common in the line of RL for finite sample analysis (Bertsekas and Yu 2010, Lazaric et al. 2010, Liu and Wei 2014, Fazel et al. 2018). Our analysis here provides a strong ground for generalization of our policy to the setting of imperfect state measurements as an important direction for future work.

We also note that the exploration probability ϵ_t decays over time. This is critical to reduce the cost of exploration by ensuring the algorithm makes fewer unnecessary explorations as more data collected and the estimates of our policy improve.

2.6.1 Lipschitzian Stability of Non-Myopic Exploitation

The notion of Lipschitzian stability was first introduced in Levy et al. (2000) for finite-dimensional parametric optimization problems. The Lipschitzian stability of optimal solutions is characterized by their behaviour with respect to perturbations in the parameter values. When the feasible set is unperturbed (i.e., independent of the parameter values) Proposition 4.32 in (Bonnans and Shapiro 2013) provides two sufficient conditions for Lipschitzian stability of optimal solutions: i) a second order growth condition, and ii) Lipschitzian continuity of the difference of the perturbed and unperturbed objective functions. To prove the Lipschitzian stability of the non-myopic exploitation policy $\Lambda_t^{E,N}(\mathcal{F}_t) = u^*_{t|t}(\widehat{\theta}_t)$ for each time step $t \in \mathcal{T}$, we next present the theoretical results required to show that these two conditions hold for $V_N(x_t, \widehat{\theta}_t, t)$. The complete proofs of all results in this section are given in Appendix A.1.3.

Lemma 2.1 *Suppose $U_{N,t} = \{u_{t|t}, \ldots, u_{t+N|t}\}$ is a feasible input sequence for $V_N(x_t, \widehat{\theta}_t, t)$. Let $J_N(x_t, U_{N,t}, \widehat{\theta}_t, t)$ be the estimated N-step reward of this input sequence at time t, i.e.,*

$$J_N(x_t, U_{N,t}, \widehat{\theta}_t, t) = \textstyle\sum_{k=0}^{N} h(\tilde{x}_{t+k|t}, u_{t+k|t}, \widehat{\theta}_t) \tag{2.13}$$

where $\tilde{x}_{t+k+1|t} = f(\tilde{x}_{t+k|t}, u_{t+k|t}, \widehat{\theta}_t)$ for $k \in \langle N-1 \rangle$ as given in (2.6). Then, $J_N(x_t, U_{N,t}, \widehat{\theta}_t, t)$ is $(L_{f,u} \cdot L_{h,u})$-Lipschitz continuous with respect to $U_{N,t}$ on the compact set $\mathcal{U}^{N+1}$, i.e.,

$$J_N(x_t, U_{N,t}, \widehat{\theta}_t, t) - J_N(x_t, U'_{N,t}, \widehat{\theta}_t, t) \leq L_{f,u} L_{h,u} ||U_{N,t} - U'_{N,t}|| \tag{2.14}$$

for any feasible input sequence $U'_{N,t} = \{u'_{t|t}, \ldots, u'_{t+N|t}\}$.

Lemma 2.1 implies the second order growth condition for $V_N(x_t, \widehat{\theta}_t, t)$ since it shows that J_N increases at least linearly over a compact set. The proof follows by Assumption 2.4 and the properties of Lipschitz continuity. We next present the second condition required for the Lipschitzian stability of the maximizer of $V_N(x_t, \widehat{\theta}_t, t)$.

Assumption 2.6 *Let the input sequences $U^*_{N,t}(\widehat{\theta}_t) = \{u^*_{t|t}(\widehat{\theta}_t), \ldots, u^*_{t+N|t}(\widehat{\theta}_t)\}$ and $U^*_{N,t}(\theta) = \{u^*_{t|t}(\theta), \ldots, u^*_{t+N|t}(\theta)\}$ be maximizers of $V_N(x_t, \widehat{\theta}_t, t)$ and $V_N(x_t, \theta, t)$, respectively. Then, for $\kappa \geq 0$, we have*

$$\Big|[J_N(x_t, U^*_{N,t}(\widehat{\theta}_t), \widehat{\theta}_t, t) - J_N(x_t, U^*_{N,t}(\widehat{\theta}_t), \theta, t)] - [J_N(x_t, U^*_{N,t}(\theta), \widehat{\theta}_t, t) - J_N(x_t, U^*_{N,t}(\theta), \theta, t)]\Big| \\ \leq \kappa \big\|\widehat{\theta}_t - \theta\big\| \cdot \big\|U^*_{N,t}(\widehat{\theta}_t) - U^*_{N,t}(\theta)\big\| \tag{2.15}$$

We now give a sufficient condition for Assumption 2.6.

Proposition 2.1 *If the property*

$$\Big\|\nabla_u J_N(x_t, U^*_{N,t}(\widehat{\theta}_t), \widehat{\theta}_t, t) - \nabla_u J_N(x_t, U^*_{N,t}(\widehat{\theta}_t), \theta, t)\Big\|_\infty \leq L_J \big\|\widehat{\theta}_t - \theta\big\| \tag{2.16}$$

holds for any $\theta \in \Theta$ and real constant $L_J \geq 0$, then Assumption 2.6 is satisfied.

This result is proven mainly by utilizing the Fundamental Theorem of Calculus for Line Integrals and Hölder's inequality.

Lemma 2.2 *If the state dynamics $f(x, u, \theta)$ and the expectation function $h(x, u, \theta)$ are polynomial functions, then the sufficient condition given in Proposition 2.1 holds.*

A specific example where Lemma 2.2 holds is a discrete-time linear time-invariant system with $f(x, u, \theta) = Ax + Bu$ and $h(x, u, \theta) = x^TQx + u^TRu$ where $\theta = [Q, R, A, B]$.

Lemma 2.3 *If Assumption 2.6 and Lemma 2.1 hold, then the Lipschitzian stability property follows by Proposition 4.32 in Bonnans and Shapiro (2013), i.e., $\big|\big|U^*_{N,t}(\widehat{\theta}_t) - U^*_{N,t}(\theta)\big|\big| \leq c_u^{-1}\kappa\|\widehat{\theta}_t - \theta\|$ for a constant $c_u > 0$.*

Since $\big|\big|u^*_{t|t}(\widehat{\theta}_t) - u^*_{t|t}(\theta)\big|\big| \leq \big|\big|U^*_{N,t}(\widehat{\theta}_t) - U^*_{N,t}(\theta)\big|\big|$, we conclude that the non-myopic exploitation policy $\Lambda_t^{E,N}(\mathcal{F}_t) = u^*_{t|t}(\widehat{\theta}_t)$ corresponding from the maximizer of $V_N(x_t, \widehat{\theta}_t, t)$ is $c_u^{-1}\kappa$-Lipschitz continuous with respect to $\hat{\theta}_t \in \Theta$.

2.6.2 Regret Analysis

We next characterize the cumulative regret performance of the non-myopic ϵ-greedy approach with respect to the expected N-step dynamic regret $R_{N,T}$ (2.7) introduced in Sect. 2.4.

By definition, $R_{N,T}$ compares the LBMPC oracle policy $\Lambda_t^{O,N}(\mathcal{F}_t)$ for the system x_t, u_t as controlled by the oracle policy to our non-myopic ϵ-greedy policy $\Lambda_t^{\epsilon,N}(\mathcal{F}'_t)$ for the system x'_t, u'_t as controlled by the learning-policy that uses the LBMPC policy $\Lambda_t^{E,N}(\mathcal{F}'_t)$ at greedy exploitation steps. Before bounding this regret notion, we first provide an upper bound on a weaker comparison of these two policies, $\Lambda_t^{\epsilon,N}(\mathcal{F}'_t)$ and $\Lambda_t^{O,N}(\mathcal{F}_t)$, by comparing the actions chosen under the states x'_t achieved by $\Lambda_t^{\epsilon,N}(\mathcal{F}'_t)$.

Theorem 2.3 *The non-myopic ϵ-greedy policy $\Lambda_t^{\epsilon,N}(\mathcal{F}'_t)$ and LBMPC oracle policy $\Lambda_t^{O,N}(\mathcal{F}'_t)$ satisfy the following result for the system states x'_t that are achieved by $\Lambda_t^{\epsilon,N}(\mathcal{F}'_t)$:*

$$\begin{aligned}
&\sum_{t=0}^{T} h(x'_t, \Lambda^{O,N}(\mathcal{F}'_t), \theta_0) - \sum_{t=0}^{T} h(x'_t, \Lambda^{\epsilon,N}(\mathcal{F}'_t), \theta_0) \\
&\quad \leq \mathcal{M}\exp\left(\frac{c_f^2(d_x, d_\theta)}{2L_{\ell,r}^2\sigma^2}\right)(\mathcal{C} + \log T) + \mathcal{M}c(1 - \log(c+1) + \log T) \\
&\qquad + \frac{L_{h,u}\kappa C\sqrt{4L_{\ell,r}^2\sigma^2}}{c_u}\sqrt{T}\log T
\end{aligned} \tag{2.17}$$

where $C > 0$, $c_f(d_x, d_\theta)$ is the constant in Theorem 2.2, and $\mathcal{C}$ is a bound on the finite summation $\sum_{t=1}^{9}\exp(-(\log t)^2)$.

Using the result of Theorem 2.3, we next show the cumulative regret behaviour of our non-myopic ϵ-greedy policy $\Lambda^{\epsilon,N}(\mathcal{F}'_t)$ by assuming the stability of the LBMPC oracle policy $\Lambda_t^{O,N}(\mathcal{F}_t)$ for $t \in \mathcal{T}$. If the LBMPC from Sect. 2.3 does not provide stability, the full LBMPC formulation (Aswani et al. 2013) can achieve stability. Our results in this chapter generalize to the full formulation but at the expense of substantial notational complexity.

Assumption 2.7 *Let $x_{eq} \in \Omega$ be an equilibrium for the LBMPC system in Sect. 2.3. For $\alpha \in [0, 2/3]$ and $\mathcal{F}_t$ as in (2.3), the LBMPC oracle policy $\Lambda_t^{O,N}(\mathcal{F}_t)$ satisfies $\|Ax_t + B\Lambda_t^{O,N}(\mathcal{F}_t) + g(x_t, \Lambda_t^{O,N}(\mathcal{F}_t), \theta_0) - x_{eq}\| \leq \alpha\|x_t - x_{eq}\|\ \forall t$.*

Exponential stability of the nonlinear LBMPC implied by this assumption can be ensured under certain sufficient conditions established in the literature (Mayne et al. 2000, Pannocchia et al. 2011). Generalizing the results with less restrictive stability notions poses future research.

Theorem 2.4 *For $4 \leq c \leq \sqrt[4]{T}/\sqrt{3}$, the expected N-step dynamic regret $R_{N,T}$ (2.7) for a policy $\Lambda^{\epsilon,N}(\mathcal{F}'_t)$ computed by Algorithm 1 satisfies*

$$\begin{aligned} R_{N,T} \leq{}& 2L_{h,x}\sqrt{T}\mathrm{diam}(\mathcal{X}) + \frac{2L_{h,x}c(3-\alpha)}{1-\alpha}\mathrm{diam}(\mathcal{X})\log T + \frac{4L_{h,x}\overline{C}c^2}{\alpha}\sqrt{T}(\log T)^3 \\ &+ \mathcal{M}\exp(\frac{c_f^2(d_x, d_\theta)}{2L_{\ell,r}^2\sigma^2})(\mathcal{C} + \log T) + \mathcal{M}c(1 - \log(c+1) + \log T) \\ &+ \frac{L_{h,u}\kappa C\sqrt{4L_{\ell,r}^2\sigma^2}}{c_u}\sqrt{T}\log T \end{aligned} \tag{2.18}$$

with probability at least

$$\begin{aligned} \Bigg[1 - (T - 2\sqrt{T})\exp\Bigg(-\frac{4c^2\left(\log\frac{e(2\sqrt{T}+2}{c+1}\right)^2}{2c\log(2\sqrt{T}+1) + \frac{2c^2}{2\sqrt{T}+1} + \frac{4c^2}{3}\log\frac{e(2\sqrt{T}+2)}{c+1}}\Bigg) \\ - \exp\Bigg(-\frac{c^2\big(\log\frac{T}{2\sqrt{T}+1}\big)^2}{(4 + \frac{2}{3}c^2)\log T}\Bigg)\Bigg] \end{aligned} \tag{2.19}$$

where $\overline{C} = c_u^{-1}(\|B\| + L_{f,u})\kappa C\sqrt{4L_{\ell,r}^2\sigma^2}$.

Remark 2.3 *This instantaneous bound on N-step dynamic regret implies an asymptotic cumulative regret of order $O(\sqrt{T}(\log T)^3)$ for the non-myopic ϵ-greedy policy with respect to the expected N-step dynamic regret (2.7).*

2.7 Numerical Experiments

In this section, we conduct numerical experiments to show the effectiveness of our non-myopic ϵ-greedy approach. We use the cumulative expected N-step dynamic regret (2.7) and the cumulative expected reward as comparison metrics. All experiments were run using Python 3.7.4 and Anaconda on a laptop computer with 2.3 GHz 8-Core Intel Core i9 processor and 16GB DDR4 RAM. We use the MOSEK Optimizer API in Python to solve all the optimization problems MOSEK ApS (2019).

We simulate an HVAC system (see Sect. 2.1.1.1), using a discrete time model from Aswani et al. (2011) with 15 minutes sampling interval. The system state is monitored by the interior temperature setpoint of the room at each period and follows the dynamics model inspired by the physics of convective heat transfer

$$x_{t+1} = k_r x_t - k_c u_t + k_v v_t + q_t \tag{2.20}$$

where $x_t \in [20, 24]$ in $^\circ C$, $k_r > 0$ is the time constant of the room, $k_c > 0$ is the temperature change over a 15-minute system delay caused by cooling for an AC duty cycle of $u_t \in [0, 0.5]$, $k_v > 0$ is the time constant for heat transfer from the room to the outside, v_t is the outside temperature in $^\circ C$, and q_t is the heating load of the occupants and equipment within the room over a system delay. We note that the time constants k_r and k_v are dimensionless.

We assume $r_t = -c_t \sim \mathcal{N}\left(h(x_t, u_t, \theta_0), \sigma^2\right)$ for $h(x_t, u_t, \theta_0) = \gamma_1 p_t u_t + (x_t - \gamma_2 - v_t)^2$ where p_t is the electricity price assumed to follow a peak-pricing plan between 12-6 p.m. over an 24 hour day. The $\gamma_1 p_t u_t$ accounts for energy use, and $v_t + \gamma_2$ indicates a setpoint preference that adjusts with outside temperature (ASHRAE 2013). We suppose $\theta_0 = [q_t, \gamma_1, \gamma_2]$ are unknown to the controller, and use $\sigma = 1, k_r = 0.64, k_c = 2.64$, and $k_v = 0.10$ (Aswani et al. 2011). We assume that the non-stationary parameters v_t and q_t are generated from a sinusoidal distribution with a single peak over 24 hours and average values of 6.98 and 17 in $^\circ C$, respectively. The experiments are replicated 1000 times for $N \in \{1, 10\}$ over a time horizon of length $T = 100,000$. All metrics are reported as averages across these replicates.

Figures 2.1 and 2.2 show the expected N-step dynamic regret accrued up to time $T = 100,000$ by the policies for $N = 1$ and $N = 10$, respectively. As discussed in Sect. 2.4, the notion of N-step dynamic regret is defined to compare two N-step finite-horizon policies for a given N. Hence, contrasting policies with varying horizon lengths would lead to a poorly-defined comparison. However, we observe that these empirical regret curves are compatible with our theoretical asymptotic regret bound $O(\sqrt{T}(\log T)^3)$ proved in Sect. 3.3.2. We further observe that the N-step dynamic regret for a given finite-time horizon is not expected to be monotonic. The reason of this non-monotonic behavior is the choice of the benchmark policy. We benchmark using the LBMPC oracle policy $\Lambda_t^{O,N}(\mathcal{F}_t)$ that optimizes N-step ahead under the full knowledge of cost and system parameters as a surrogate for the optimal policy $\Lambda_t^*(\mathcal{F}_t)$ that optimizes up to the end of time horizon.

Lastly, we compare the cumulative expected costs of the policies for $N = 1$ and $N = 10$ by subtracting the expected cost of $\Lambda_t^{\epsilon,10}(\mathcal{F}_t')$ from that of $\Lambda_t^{\epsilon,1}(\mathcal{F}_t')$. Figure 2.3 shows that

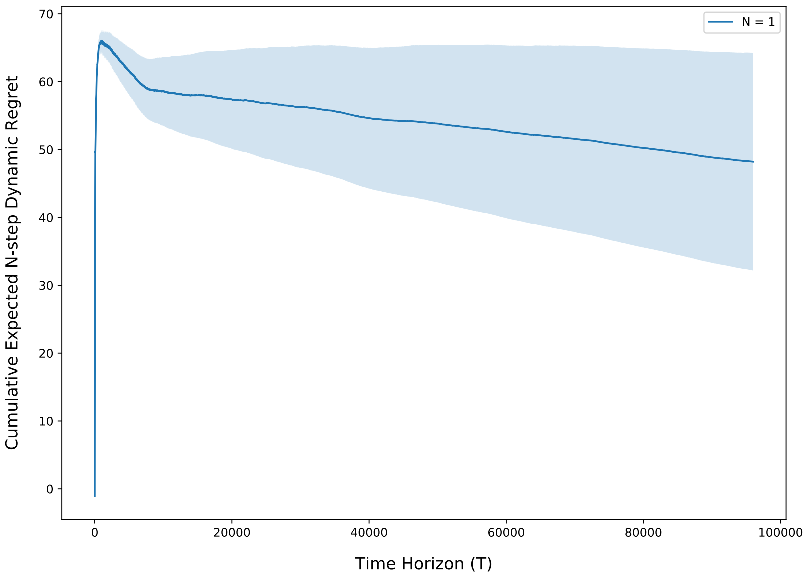


Figure 2.1: Expected 1-step dynamic regret. The shaded region represents the standard error over 1000 replications.

we obtain significantly lower costs with higher N value and that this gain increases roughly linearly as the time horizon gets longer. This implies that non-myopic policies utilizing information on future improvements in cumulative costs provide better approximations to optimal policies over a finite-time horizon.

2.8 Conclusions

This chapter proposes a novel learning-based control framework at the intersection of LBMPC and RL for studying the exploration/exploitation trade-off in nonlinear and non-stationary systems. A critical issue we consider is stability, which is one of the unique (and not previously well-studied) issues that arises with RL for nonlinear systems. The developed class of LBMPC policies embraces a statistically consistent parameter estimation approach and has been proven to attain low regret in settings with partially-unknown cost function and system dynamics.

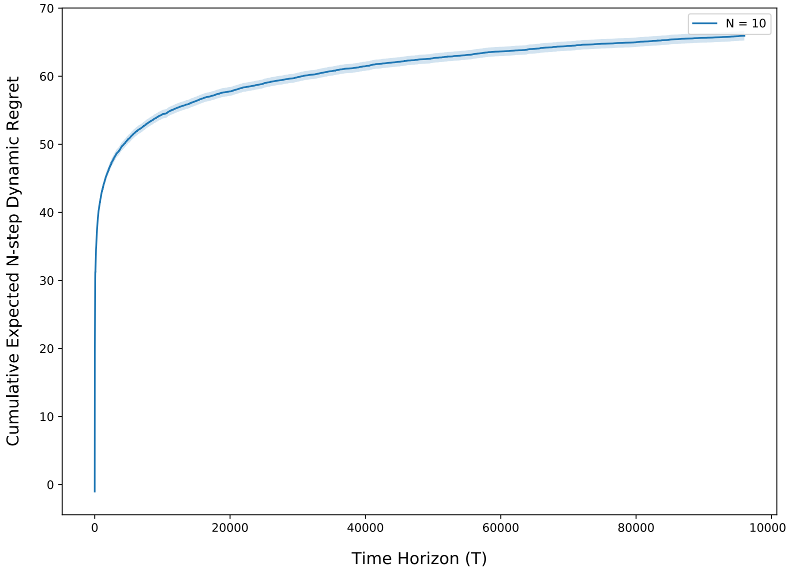


Figure 2.2: Expected 10-step dynamic regret. The shaded region represents the standard error over 1000 replications.

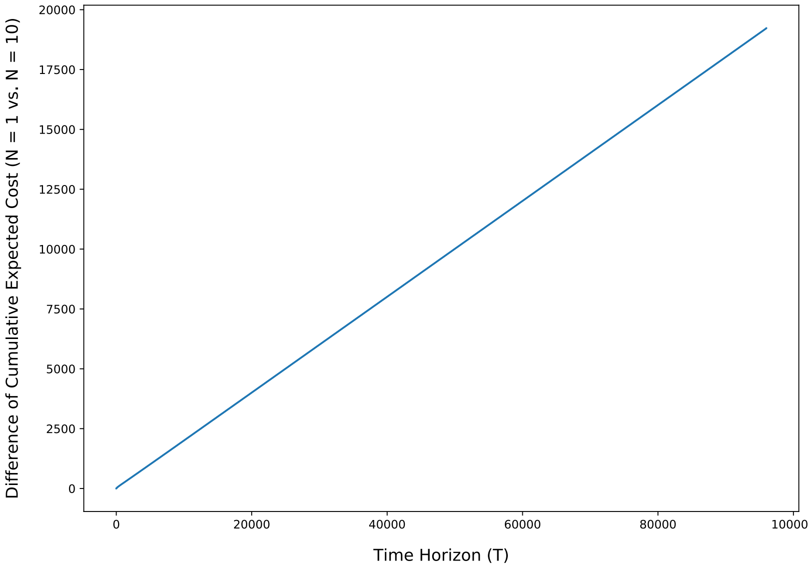


Figure 2.3: Difference of cumulative expected costs for $N = 1, 10$.

Chapter 3

Repeated Principal-Agent Games with Hidden Rewards: Perfect-Knowledge Agents

3.1 Introduction

System designers frequently use the idea of providing incentives to stakeholders as a powerful means of steering the stakeholders for their own benefit. Operations management includes many such examples, such as incentivizing suppliers for emissions transparency, vertical collaboration between shippers and carriers for sustainable route planning, performance-based bonuses offered to ride-hailing drivers, monetary incentives provided to patients for medical adherence, and quality-contingent bonus payments for workers in crowdsourcing platforms.

As discussed in the Introduction chapter, a primary challenge in designing efficient incentives lies in isolating the lack of transparency regarding the true willingness and economic objectives of the self-interested stakeholders. In many real-world settings, the problem of designing efficient incentives can be framed as a repeated principal-agent problem where a primary party (i.e., principal) seeks to optimize their ultimate objective by designing sequential incentive policies to lead a self-interested stakeholder (i.e., agent) with a private decision-making process. This privacy imposes an information asymmetry between the principal and the agent that can appear in the form of either an *adverse selection*, in which the information about the agent's true preferences or rewards are hidden from the principal, or a *moral hazard*, in which the actions chosen by the agent are hidden from the principal (Bolton and Dewatripont 2004). For instance, in the context of employment incentives designed by an employer, the hidden information in an adverse selection setting could be the level of productivity of an employee whereas a hidden action in the moral hazard setting could be the total effort level of the employee. More generally, the hidden information in the adverse selection setting can be seen as an unknown "type" or "preferences" of the agent that directly affects the action chosen by the agent, which in turn determines both the agent's and the principal's rewards. These situations require specification of the agent's private information and the distributional-knowledge that the principal has concerning that information.

Existing literature on repeated principal-agent models mostly studies the moral hazard setting, with a more recent focus on the problem of estimating agent's unknown model parameters under hidden actions (e.g., Ho et al. 2016, Kaynar and Siddiq 2022). On the other hand, the adverse selection setting is mostly studied either for single-period static games (Navabi and Nayyar 2018, Chade and Swinkels 2019, Gottlieb and Moreira 2022) or else for the repeated dynamic games where restrictive assumptions are made on, for example, dimension of the agent's action space, utility function of the agent, and relationship between principal's rewards and agent's unknown type (e.g., Halac et al. 2016, Eső and Szentes 2017, Maheshwari et al. 2022). Furthermore, the statistical estimation and learning problem has not previously been explored for the repeated adverse selection setting. However, system designers in practice require more generic and richer dynamic approaches that leverage data only on past incentives and the agent's past actions without necessarily imposing a specific structure on the reward model or type distribution of the agent.

Our main goal in this chapter is to open a new window to repeated principal-agent models under adverse selection from the perspective of statistical learning theory. In particular, we consider an unexplored setting of adverse selection where the principal can only observe the history of the agent's actions but remains uninformed about the associated rewards of the agent. To enhance the practical relevance of our approach, we design a generic and simple model. We assume that the agent has the perfect knowledge of their reward model and picks the reward-maximizing action based on the incentives provided by the principal at each period. Under this repeated *hidden rewards* setting, we are mainly interested in studying the following two research questions:

1. How to compute a statistically consistent estimator for a non-parametric reward model of the agent?
2. How to design data-driven and adaptive incentives that will attain low regret to the principal?

3.1.1 Motivating Real-Life Applications

3.1.1.1 Sustainable and Collaborative Route Planning with Backhauling

Backhauling is a way of improving the efficiency of shipping vehicles by providing pickup loads for them on their way back to the origin depot. It has been widely applied in logistics operations to reduce both the transportation costs of companies and negative environmental impacts due to fuel consumption and pollutant emissions (Early 2011, Juan et al. 2014, Turkensteen and Hasle 2017). In the context of collaborative transportation in a supply chain network, backhauling is a complex, yet powerful, tool for achieving green closed-loop logistics. Due to the hierarchical relationship between shipper companies and carriers in a transportation network, it is often studied as a form of vertical collaboration in which companies create integrated outbound-inbound routes – instead of dedicated delivery and dedicated pickup routes – and provide incentives (i.e., side payments) to carriers to induce these routes (Ergun et al. 2007, Audy et al. 2012, Marques et al. 2020, Santos et al. 2021).

These existing approaches focus on solving the shipper's single-period static routing and pricing problems by using techniques mostly from optimization theory. However, in practice, shippers face these decisions and interact with carriers dynamically and repeatedly at every shipment request. Therefore, there is clearly a need for designing the vertical collaboration between a shipper and a carrier as a sequential learning and decision-making process. In that regard, this incentive design problem can be formulated as a repeated game between a principal (shipper) and an agent (carrier) under adverse selection. The goal of the shipper is to initiate the use of pre-planned integrated routes for their linehaul and backhaul customers to minimize their total transportation costs, whereas the carrier aims to maximize their total profits from the selected routes. At the end of each shipment request, the shipper observes the set of routes chosen by the carrier after the provided incentives while the total profit obtained by the carrier stands as invisible information to the shipper – which makes it more challenging for the shipper to predict and orient the carrier's future selections. Taking into account all these features, the repeated *hidden rewards* model and the adaptive incentive policy proposed in this chapter can explicitly account the goals and interactions of both parties and yield effective incentive plans by leveraging the available data over a given contract period.

3.1.1.2 Personalized Incentive Design for Medical Adherence

The problem of patients not following medication dosing instructions is recognized as a major and widespread issue around the world. Such lack of adherence to a medication regime leads to not only poor health outcomes but also substantial financial costs Osterberg and Blaschke (2005). According to WHO (2003), medication non-adherence is observed 50% of the time, which may increase up to 80% for relatively asymptomatic diseases such as hypertension (Brown et al. 2016). Research reveals various reasons for this problem including individual-level factors (e.g., medication side effects), social factors (e.g., cultural beliefs), and economic factors (e.g., transportation costs to clinics) (WHO 2003, Bosworth 2010, Long et al. 2011). To overcome some of these concerns, incentive programs that provide financial rewards to the patients are commonly employed and shown to effectively improve medical adherence. There is a related literature in medicine and economics on examining the effects of these monetary incentives using empirical analyses (Lagarde et al. 2007, Gneezy et al. 2011) and in operations management on quantitatively designing financial incentives for different market contexts (Aswani et al. 2018, Ghamat et al. 2018, Guo et al. 2019, Suen et al. 2022).

The design of financial incentives throughout a medication regime with finite length adequately features a repeated principal-agent problem under the *hidden rewards* setting that we introduce in this chapter. Given their personal preferences and characteristics (i.e., type), the patient (i.e., agent) exhibits certain adherence behaviors in order to maximize their total utility which is comprised of benefits obtained through the improvements in their health conditions, costs incurred due adherence, and incentives offered by the healthcare provider. On the other hand, the goal of the healthcare provider (i.e., principal) is to maximize the clearance rate, that is the rate at which the infected patient is recovered, by designing motivating payments to improve their adherence actions. This payment design problem is nontrivial due to scarce clinical resources and the information asymmetry between the provider and the patient. Although the healthcare provider can often fully observe the patient's adherence actions, the realized utilities of the patient often stands as a private information to the provider. Because the data-driven incentive design framework presented in this chapter is based on a generic model without restrictive technical assumptions, we propose that it could be easily leveraged to overcome the problem of medical non-adherence.

3.1.2 Main Contributions and Chapter Outline

We next present the outline and main methodological contributions of this chapter in more detail.

Consistent estimator. In Section 3.2.1, we provide the details of the principal-agent setting introduced above. Then, we introduce a novel estimator for a non-parametric reward model of a reward-maximizing agent with finite set of actions in Section 3.2.2. Our estimator is formulated exactly as a linear optimization model that estimates the expected rewards of all actions without assuming any functional form or any specific distributional property. In accordance with the *hidden rewards* setting, the only input to our estimator is the data on past incentives and past actions chosen by the agent. In Section 3.2.3, we give results proving identifiability and finite-sample statistical consistency of the proposed estimator. Essentially, we prove probability bounds on the diameter of the random polytope defined by the feasible space of our estimator in each time period.

Data-driven and low-regret incentives. Section 3.3.1 describes a practical and computationally efficient ϵ-greedy policy for the principal's adaptive incentives over a finite time horizon of length T. By utilizing the finite-sample concentration bounds derived for our estimator, we compute the regret of the proposed policy with respect to an oracle incentive policy that maximizes the principal's expected net reward at each time step under the perfect knowledge of all reward expectations. Section 3.3.2 presents a rigorous regret bound of order $O(\sqrt{T \log T})$ for the repeated principal-agent models with hidden agent rewards.

Discussion and Numerical results. Our approach assumes that the agent's decisions are consistent with a fixed vector of reward expectations. However, we also consider the case when there is no guarantee that the agent is truthful about their preferences. In some cases where the agent might also be knowledgeable about the principal's model, they can increase the information rent extracted from the principal by pretending their vector of reward expectations are different. In Section 3.4, we provide a discussion from the perspective of the reward-maximizing agent and argue that our incentives are designed in a way that maximizes the principal's expected net reward subject to the agent's information rent. To support our theoretical results and demonstrate our data-driven incentive framework, we conduct simulations on an instance of the sustainable route planning problem outlined earlier. In Section 3.5, we share the details of our experimental setting and numerical results.

Lastly, we conclude in Section 3.6 by discussing future work that might be steered by our analyses in this chapter. We include the proofs for all theoretical results provided in the main text in Appendix B.

3.1.3 Related Literature

3.1.3.1 Repeated Principal-Agent Models

There is a rich and extensive literature on principal-agent models in economics (Holmström 1979, Grossman and Hart 1983, Hart and Holmström 1987) and in operations management (Martimort and Laffont 2009). For repeated models, most existing studies focuses on the moral hazard setting (Radner 1981, Rogerson 1985, Spear and Srivastava 1987, Abreu et al. 1990, Plambeck and Zenios 2000, Conitzer and Garera 2006, Sannikov 2008, 2013). Several of them study the problem of estimating the agent's model when actions are hidden (Vera-Hernandez 2003, Misra et al. 2005, Misra and Nair 2011, Ho et al. 2016, Kaynar and Siddiq 2022). On the other hand, related work on the design of incentives under the adverse selection setting is relatively scarce. In many of them, the agent's type (e.g., level of effort or probability of being successful) is considered as an additional, unknown information on top of a moral hazard setting (Dionne and Lasserre 1985, Sundadam and Banks 1991, Gayle and Miller 2015, Williams 2015, Halac et al. 2016, Eső and Szentes 2017). Only a few of these works study the estimation problem for the hidden type setting, and they use statistical estimation methods such as least squares approximation (Lee and Zenios 2012), minimization of a sum of squared criterion function (Gayle and Miller 2015), and simulation-based maximum likelihood estimation (Aswani et al. 2019, Mintz et al. 2023). However, the adverse selection setting studied in these papers comes with limiting assumptions such as the assumption that the agent's type parameter belongs to a discrete set.

Our work differs from these studies in several ways. Although the *hidden rewards* setting has various application areas, we are not aware of any other work studying this novel and non-trivial dynamic principal-agent model. The estimation problem we consider in this setting involves estimating reward expectation values which belong to a bounded continuous space. Differently from the existing work summarized above, we solve a practical linear program and follow a set-based estimation approach to estimate these continuous mean rewards. Furthermore, regarding the incentive design problem, these past papers do not consider the exploration-exploitation trade-off faced by the principal, and hence, they are not able to provide guarantees on how close to optimal their solutions are. In this chapter, we take a sequential learning approach to compute adaptive and efficient incentives for the principal and perform a regret analysis for the considered repeated adverse selection models.

3.1.3.2 Multi-Armed Bandits for Incentive Design

A related line of research from sequential decision-making includes the use of a multi-armed bandit (MAB) framework for mechanism design. MAB's are widely applied to dynamic auction design problems which are closely related with the incentive design in dynamic principal-agent problems (Nazerzadeh et al. 2008, Devanur and Kakade 2009, Jain et al. 2014, Amin et al. 2014, Biswas et al. 2015, Ho et al. 2016, Braverman et al. 2019, Bhat et al. 2019, Abhishek et al. 2020, Shweta and Sujit 2020, Han et al. 2020, Simchowitz and Slivkins 2021, Wang et al. 2022, Gao et al. 2022).

The principal's problem in our repeated hidden rewards game is directly applicable to the MAB framework. At each iteration of the game, the principal offers a set of incentives corresponding to the set of arms (i.e., actions) in the agent's model and generates a random reward through the arm selected by the agent. As the interaction between these two parties proceeds, the principal faces a trade-off between consistently learning the unknown reward expectations of each agent arm (i.e., *exploring* the agent's arm space through adverse incentives to encourage selection of different arms) and maximizing their own cumulative net reward (i.e., *exploiting* arms estimated to yield the highest expected rewards by motivating the agent to select them with the minimal incentives). From this perspective, the MAB framework is regarded to be useful in effectively navigating the principal's exploration-exploitation trade-off while designing efficient data-driven incentives.

3.1.3.3 Inverse Optimization

Inverse optimization is a framework for inferring parameters of an optimization model from the observed solution data that are typically corrupted by noise (Ahuja and Orlin 2001, Heuberger 2004). More recent work in this area probes into estimating the model of a decision-making agent by formulating the agent's model as a linear or a convex optimization problem in offline settings (where data are available a priori) (Keshavarz et al. 2011, Bertsimas et al. 2015, Esfahani et al. 2018, Aswani et al. 2018, Chan et al. 2019, 2022) or in online settings (where data arrive sequentially) (Bärmann et al. 2018, Dong et al. 2018, Dong and Zeng 2020, Maheshwari et al. 2023). Different from these studies, we do not assume any specific structure of the agent's decision-making problem, but instead we consider a reward-maximizing agent with finite action space. This case of estimating the non-parametric model of a reward-maximizing agent is also addressed by Kaynar and Siddiq (2022), who study the offline static setting of the principal-agent problem under moral hazard. A key distinction is that we study the online dynamic setting of the repeated principal-agent problem under adverse selection. In the hidden rewards setting that we examine, we design an estimator for the expected rewards of the agent's arms, whose only input is the data of reward-maximizing arms in response to the provided incentives in the past. In that respect, the principal's estimation problem can be regarded as an analogy of online inverse optimization in MAB's. Moreover, to prove consistency of the principal's estimator in this setting, we build upon initial ideas of statistics with set-valued functions (Aswani 2019).

3.1.4 Mathematical Notation

We first specify our notational conventions throughout this chapter. All vectors are denoted by boldfaced lowercase letters. A vector $\mathbf{x}$ whose entries are indexed by a set $\mathcal{M} = [1, \ldots, M]$ is defined as $\mathbf{x} = (x_m)_{m \in \mathcal{M}}$. If each entry x_m belongs to a set $\mathcal{X}$, then we have $\mathbf{x} \in \mathcal{X}^M$. The ℓ_∞-norm of the vector $\mathbf{x}$ is defined by $\|\mathbf{x}\|_\infty = \max(|x_1|, \ldots, |x_M|)$. Further, the cardinality of a set $\mathcal{X}$ is denoted by $|\mathcal{X}|$, and $\mathbb{1}(\cdot)$ denotes the indicator function that takes value 1 when its argument $(\cdot)$ is true, and 0 otherwise. Lastly, the notations $\mathbf{0}_n$ and $\mathbf{1}_n$ are used for the all-zeros and all-ones vectors of size n, respectively, and $\mathbb{P}(\cdot)$ is used for probabilities.

3.2 The Repeated Game and Principal's Estimation

We start this section by introducing our repeated principal-agent model under adverse selection and continue by presenting our novel estimator along with the associated statistical results.

3.2.1 The Repeated Game with Hidden Rewards

We consider a repeated play between a principal and an agent over a finite time horizon $\mathcal{T} = [1, \ldots, T]$. At each time step $t \in \mathcal{T}$, the principal offers a vector of incentives $\boldsymbol{\pi}_t = (\pi_{t,a})_{a \in \mathcal{A}}$ corresponding to the set of all possible actions of the agent $\mathcal{A} = \{1, \ldots, n\}$. Then, the agent takes the action $i_t(\boldsymbol{\pi}_t)$ with the maximum expected total reward after the incentives $\boldsymbol{\pi}_t$, i.e.,

$$i_t(\boldsymbol{\pi}_t) := \arg\max\, (\mathbf{r}^0 + \boldsymbol{\pi}_t) = \underset{a \in \mathcal{A}}{\arg\max} \left(r_a^0 + \pi_{t,a} \right) \tag{3.1}$$

where $\mathbf{r}^0 = (r_a^0)_{a \in \mathcal{A}}$ is the true vector of expected rewards of the agent and is only known by the agent. We assume that $r_a^0, \forall a \in \mathcal{A}$ belongs to a compact set $\mathcal{R} = [R_{\min}, R_{\max}] \subset \mathbb{R}$ where $R_{\max} - R_{\min} \geq 1$. Based on the action chosen by the agent, the principal collects a stochastic reward outcome denoted by $\mu_{t, i_t(\boldsymbol{\pi}_t)} \sim \mathbb{F}_{\theta^0_{i_t(\boldsymbol{\pi}_t)}, i_t(\boldsymbol{\pi}_t)}$ with expectation $\theta^0_{i_t(\boldsymbol{\pi}_t)} \in \Theta$ where Θ is a known compact set.

The true mean reward vectors $\mathbf{r}^0$ and $\boldsymbol{\theta}^0 = (\theta_a^0)_{a \in \mathcal{A}}$ are unknown by the principal. The principal can only observe the selected action $i_t(\boldsymbol{\pi}_t)$ and their own net reward realization $\mu_{t, i_t(\boldsymbol{\pi}_t)} - \sum_{a \in \mathcal{A}} \pi_{t,a}$. In this setting, to ensure that our research problems are well-posed, it suffices to assume that the range of incentives that the principal is able to provide to the agent covers the range of the agent's reward expectations.

Assumption 3.1 *The incentives $\pi_{t,a}$, $\forall a \in \mathcal{A}$ belongs to a compact set $\mathcal{C} = [\underline{C}, \overline{C}]$ where $\underline{C} = R_{\min}$ and $\overline{C} = R_{\max} + \gamma$ for some constant $0 < \gamma \leq R_{\max} - R_{\min} - 1$.*

Because the principal's goal is to provide incentives that will drive the agent's decisions, this assumption ensures that the magnitudes of the incentives are large enough to have an effect on the relative order of the actions with respect to their total rewards after adding the incentives.

3.2.2 The Estimator

Due to the information asymmetry in our repeated adverse selection model, the learning process of the principal comprises estimating the agent's expected reward vector $\mathbf{r}^0$ by solely watching the actions maximizing the total reward vector $\mathbf{r}^0 + \boldsymbol{\pi}_\tau$ in the past time periods $\tau \leq t$. Our fundamental observation of this estimation problem is that the differences of pairs of entries of $\mathbf{r}^0$ is crucial for the statistical analysis, not the individual values of the entries. With this observation on hand, we must first discuss an ambiguity in this problem before formulating our estimator.

Suppose the principal offers an incentive vector $\boldsymbol{\pi}_\tau$ at time τ. We can trivially find two different mean reward vectors for which the principal's estimation problem will be ill-defined. To see this, we consider the vectors $\mathbf{r}' \in \mathcal{R}^n$ and $\mathbf{r}'' = \mathbf{r}' + m\mathbf{1}_n$, where m is any constant scalar such that $\mathbf{r}'' \in \mathcal{R}^n$. Then, the key is to notice that $\arg\max\ \mathbf{r}' + \boldsymbol{\pi}_\tau = \arg\max\ \mathbf{r}'' + \boldsymbol{\pi}_\tau$. Since these two vectors $\mathbf{r}'$ and $\mathbf{r}''$ yield the same maximizer arm, the principal will not be able to distinguish them in the considered affine space. To overcome this issue of identifiability, we remove one redundant dimension from the considered estimation problem by setting all the differences of pairs of $\mathbf{r}$'s entries with respect to a reference point 0.

Definition 3.1 *For a mean reward vector $\mathbf{r} = (r_1, r_2, \ldots, r_n) \in \mathcal{R}^n$, we define $\mathbf{s}$ as the normalized mean reward vector that is without loss of generality defined by $\mathbf{s} := \mathbf{r} - r_1\mathbf{1}_n = (0, r_2 - r_1, \ldots, r_n - r_1)$ and belongs to the compact set $\mathcal{S}^n = [R_{\min} - R_{\max}, R_{\max} - R_{\min}]^n$.*

This dimensionality reduction allows us to decrease our degrees of freedom and derive the identifiability of our estimator. We further note that the maximizer $i_\tau(\boldsymbol{\pi}_\tau)$ of the total expected reward vector $\mathbf{r}^0 + \boldsymbol{\pi}_\tau$ is also the maximizer of $\mathbf{s}^0 + \boldsymbol{\pi}_\tau$. Thus, we define our estimator and conduct theoretical analyses with respect to the normalized reward vector $\mathbf{s}$.

Next, we formalize our estimator for $\mathbf{s}^0$. Let $\boldsymbol{\Pi}_t = \{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{t-1}\}$ be the sequence of incentives offered by the principal and $I_t(\boldsymbol{\Pi}_t) = \{i_1(\boldsymbol{\pi}_1), \ldots, i_{t-1}(\boldsymbol{\pi}_{t-1})\}$ be the sequence of actions chosen by the agent up to time t. Then, the principal's estimate $\widehat{\mathbf{s}}_t\left(I_t(\boldsymbol{\Pi}_t), \boldsymbol{\Pi}_t\right)$ at time t for the agent's normalized mean reward vector $\mathbf{s}^0$ is formulated as

$$\widehat{\mathbf{s}}_t\left(I_t(\boldsymbol{\Pi}_t), \boldsymbol{\Pi}_t\right) \in \arg\min\ 0 \tag{3.2}$$

$$\text{s.t. } s_{i_\tau(\boldsymbol{\pi}_\tau)} + \pi_{\tau, i_\tau(\boldsymbol{\pi}_\tau)} \geq s_a + \pi_{\tau,a} \qquad \forall a \in \mathcal{A},\ \tau = 1, \ldots, t-1 \tag{3.3}$$

$$s_1 = 0,\ s_a \in \mathcal{S} \qquad \forall a \in \mathcal{A} \tag{3.4}$$

This optimization problem can be regarded as the feasibility version of the set-membership estimation problem (Schweppe 1967, Hespanhol and Aswani 2020). Further, we can reformulate it by defining the loss function

$$L\left(\mathbf{s}, I_t(\boldsymbol{\Pi}_t), \boldsymbol{\Pi}_t\right) = \sum_{\tau=1}^{t-1} \ell\left(\mathbf{s}, i_\tau(\boldsymbol{\pi}_\tau), \boldsymbol{\pi}_\tau\right) \tag{3.5}$$

which is the sum of $t-1$ extended real-valued functions given by

$$\ell\left(\mathbf{s}, i_\tau(\boldsymbol{\pi}_\tau), \boldsymbol{\pi}_\tau\right) = \left\{ \begin{array}{ll} 0, & \text{if } \; s_{i_\tau(\boldsymbol{\pi}_\tau)} + \pi_{\tau, i_\tau(\boldsymbol{\pi}_\tau)} \geq s_a + \pi_{\tau,a}, \;\; \forall a \in \mathcal{A} \\ +\infty, & \text{otherwise} \end{array} \right\}. \tag{3.6}$$

Now, we reformulate our feasibility estimator as

$$\widehat{\mathbf{s}}_t\left(I_t(\boldsymbol{\Pi}_t), \boldsymbol{\Pi}_t\right) \in \underset{s_1=0,\; s_a\in\mathcal{S},\forall a\in\mathcal{A}}{\arg\min} \;\; L\left(\mathbf{s}, I_t(\boldsymbol{\Pi}_t), \boldsymbol{\Pi}_t\right) \tag{3.7}$$

Note that we may use the simplified notation $\widehat{\mathbf{s}}_t$ throughout the chapter for conciseness. We next present the results of our statistical analysis for the estimator in (3.7).

3.2.3 Identifiability and Consistency

The convergence behavior of the sequence of estimates $\widehat{\mathbf{s}}_t$ depends on a characterization of the loss function known as an *identifiability condition* (Van der Vaart 2000) that ensures the loss function is minimized uniquely by the true vector $\mathbf{s}^0$. We start our consistency analysis by proving the identifiability of our estimator (3.7). The identifiability of our estimation problem requires characterizing the set of incentive vectors that distinguishes between $\mathbf{s}^0$ and an incorrect estimate $\widehat{\mathbf{s}}_t$. We first provide some intermediate results in Propositions 3.1 – 3.3 and then formalize the final identifiability result for our estimator in Proposition 3.4.

Let $\mathcal{N}(\mathbf{s}^0, \beta) \subset \mathcal{S}^n$ be an open neighborhood centered around $\mathbf{s}^0$ with diameter $\beta > 0$ such that $\mathcal{N}(\mathbf{s}^0, \beta) := \{\mathbf{s} : \|\mathbf{s}-\mathbf{s}^0\|_\infty \leq \beta\}$, and consider the compact set $\mathcal{F} := \mathcal{S}^n \setminus \mathcal{N}(\mathbf{s}^0, \beta)$. We define an open ball $\mathcal{B}(\mathbf{s}^j, d) := \{\mathbf{s} : \|\mathbf{s} - \mathbf{s}^j\|_\infty < d\}$ centered around a vector $\mathbf{s}^j$ with diameter $d > 0$. Since $\mathcal{F}$ is compact, for some finite $q > 0$ and $d < \beta$, there is a finite subcover $\{\mathcal{B}(\mathbf{s}^j, d) : \mathbf{s}^j \in \mathcal{F}\}_{j=1}^q$ of a collection of open balls covering $\mathcal{F}$. Given a normalized reward vector $\mathbf{s} \in \mathcal{B}(\mathbf{s}^j, d), j \in \{1, \ldots, q\}$, our arguments in the following propositions will be based on the following indices:

- $K := \arg\max_{a\in\mathcal{A}} s_a$ (the set of indices corresponding to the maximizers of $\mathbf{s}$)
- $K^0 := \arg\max_{a\in\mathcal{A}} s^0_a$ (the set of indices corresponding to the maximizers of $\mathbf{s}^0$)
- $b \in \arg\max_{a\in\mathcal{A}} |s^0_a - s_a|$ (the index of an entry with the highest absolute value in $\mathbf{s}^0 - \mathbf{s}$)

Proposition 3.1 *Suppose that $K^0 \cap K = \emptyset$ for a given vector $\mathbf{s} \in \mathcal{B}(\mathbf{s}^j, d), j \in \{1, \ldots, q\}$, and that the principal chooses each incentive $\pi_{t,a}$ uniformly randomly from the compact set $\mathcal{C}$, that is $\pi_{t,a} \sim U(\underline{C}, \overline{C}), \forall a \in \mathcal{A}$, at time $t \in \mathcal{T}$. Then,*

$$\mathbb{P}\left(\ell\left(\mathbf{s}, i_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) = +\infty\right) \geq \left(\frac{1}{2} - \frac{\left(\overline{C} - \underline{C} - s^0_{\kappa^0} + s^0_{\kappa}\right)^2}{2(\overline{C} - \underline{C})^2}\right)\left(1 - \frac{\gamma + \beta - d}{\overline{C} - \underline{C}}\right)^2 \left(\frac{\gamma}{\overline{C} - \underline{C}}\right)^{n-2} \tag{3.8}$$

for any $\kappa \in K$, $\kappa^0 \in K^0$, and γ as introduced in Assumption 3.1.

Proposition 3.2 *Suppose that $K^0 \cap K \neq \emptyset$, $b \notin K^0 \cap K$ for a given vector $\mathbf{s} \in \mathcal{B}(\mathbf{s}^j, d), j \in \{1, \ldots, q\}$ and that $\pi_{t,a} \sim U(\underline{C}, \overline{C}), \forall a \in \mathcal{A}$ at time $t \in \mathcal{T}$. We define the quantity $\omega = \sup_{\mathbf{s} \in \mathcal{B}(\mathbf{s}^j, d)} \max_{a \in \mathcal{A}} \{|s_a^0|, |s_a|\}$ as the largest absolute value among the entries of $\mathbf{s}^0$ and of all the vectors in $\mathcal{B}(\mathbf{s}^j, d)$. Then,*

$$\mathbb{P}\left(\ell\left(\mathbf{s}, i_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) = +\infty\right) \geq \frac{\beta^2}{(\overline{C} - \underline{C})^2} \left(1 - \frac{\gamma + \omega}{\overline{C} - \underline{C}}\right)^2 \left(\frac{\gamma}{\overline{C} - \underline{C}}\right)^{n-2}. \tag{3.9}$$

Proposition 3.3 *Suppose that $K^0 \cap K \neq \emptyset$, $b \in K^0 \cap K$ for a given vector $\mathbf{s} \in \mathcal{B}(\mathbf{s}^j, d), j \in \{1, \ldots, q\}$, and that $\pi_{t,a} \sim U(\underline{C}, \overline{C}), \forall a \in \mathcal{A}$ at time $t \in \mathcal{T}$. Then,*

$$\mathbb{P}\left(\ell\left(\mathbf{s}, i_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) = +\infty\right) \geq \frac{\beta^2}{(\overline{C} - \underline{C})^2} \left(1 - \frac{\gamma + \beta - d}{\overline{C} - \underline{C}}\right) \left(1 - \frac{\gamma + \omega}{\overline{C} - \underline{C}}\right) \left(\frac{\gamma}{\overline{C} - \underline{C}}\right)^{n-2} \tag{3.10}$$

for the constant ω defined in Proposition 3.2.

Propositions 3.1 – 3.3 analyze three mutually exclusive cases for a given reward vector $\mathbf{s}$ and the true reward vector $\mathbf{s}^0$. In all cases, these results show that as the distance β between the considered vector $\mathbf{s}$ and the true vector $\mathbf{s}^0$ increases, the probability that the estimator (3.7) will be able to differentiate these two vectors is also increasing proportional to β^2, and that this probability of invalidating an incorrect estimate is always strictly positive. In other words, they state that the unknown mean reward vector $\mathbf{s}^0$ can be learned from the input data collected by offering randomly chosen incentives that explore the agent's action space. Proposition 3.4 combines these results to show that our adverse selection model satisfies an identifiability property required for a precise inference on the agent's hidden rewards.

Proposition 3.4 ***(Identifiability)*** *At time $t \in \mathcal{T}$, suppose that $\pi_{t,a} \sim U(\underline{C}, \overline{C}), \forall a \in \mathcal{A}$. Then, for any normalized reward vector $\mathbf{s} \in \mathcal{F}$, we have*

$$\mathbb{P}\left(\ell\left(\mathbf{s}, i_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) = +\infty\right) \geq \alpha\beta^2 \tag{3.11}$$

for some constant $\alpha > 0$.

Theorem 3.1 presents the finite-sample concentration behavior for our estimator with respect to the loss function (3.5). The main sketch of the proof of Theorem 3.1 follows by the existence of the finite subcover $\{\mathcal{B}(\mathbf{s}^j, d) : \mathbf{s}^j \in \mathcal{F}\}_{j=1}^q$ of an open covering of $\mathcal{F}$ and by using the result of Proposition 3.4 for each of the open balls in this subcover. Then, the final inequality is obtained by using volume ratios to bound the covering number q. The complete proof is given in Appendix B.1.1. The intuition behind the upper bound given in (3.12) is that the learning rate of the principal's estimator depends on the number of time periods at which the principal is exploring the action space of the agent.

Theorem 3.1 *Let $\eta(1,t)$ be the number of time steps that the principal chooses each incentive $\pi_{t,a}$ uniformly randomly from the compact set $\mathcal{C}$ up to time t, that is $\eta(1,t) = |\Lambda(1,t)|$ where $\Lambda(1,t) = \{\tau : 1 \leq \tau \leq t-1,\ \pi_{\tau,a} \sim U(\underline{C}, \overline{C}), a \in \mathcal{A}\}$. Then, we have*

$$\mathbb{P}\left(\inf_{\mathbf{s}\in\mathcal{F}} L\left(\mathbf{s}, I_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) < +\infty\right) \leq \exp\left(-\alpha(\eta(1,t)-1)\beta^2 - \log\beta + n\log(R_{\max} - R_{\min})\right) \tag{3.12}$$

where $\mathcal{F} = \{\mathbf{s} \in \mathcal{S}^n : \|\mathbf{s} - \mathbf{s}^0\|_\infty > \beta\}$ as before.

This theorem is useful because it allows us to derive our finite-sample concentration inequality with respect to the distance between our estimates $\widehat{\mathbf{s}}_t$ and the true reward vector $\mathbf{s}^0$. We conclude this section with an alternative statement of Theorem 3.1.

Corollary 3.1 (Finite-Sample Concentration Bound) *The principal's estimator in (3.7) satisfies*

$$\mathbb{P}\left(\|\mathbf{s}^0 - \widehat{\mathbf{s}}_t\|_\infty > \beta\right) \leq \exp\left(-\alpha(\eta(1,t)-1)\beta^2 - \log\beta + n\log(R_{\max} - R_{\min})\right) \tag{3.13}$$

for any $\beta > 0$.

Recall that the radius of a polytope is the maximum distance between any two points in it. Then, because both $\widehat{\mathbf{s}}_t$ and $\mathbf{s}^0$ are feasible solutions to (3.7), this corollary can be also interpreted as a probability bound on the radius of the random polytope defined by the constraints of our estimation problem.

3.3 Principal's Adaptive Incentive Framework

In this section, we develop an adaptive incentive policy that yields an efficient regret bound for the principal in the repeated adverse selection game described in Section 3.2.1. As per the considered model setting, the principal needs to learn their own expected rewards $\boldsymbol{\theta}^0$ in addition to the agent's rewards. Because the principal can fully observe the reward outcomes $\mu_{t,i_t(\boldsymbol{\pi}_t)}$ that they get through the agent's decisions, we consider an unbiased estimator under the following assumption about the principal's reward distribution family.

Assumption 3.2 *The principal's rewards $\mu_{t,a}$'s for an arm $a \in \mathcal{A}$ are independent and follow a sub-Gaussian distribution $\mathbb{F}_{\theta_a,a}$ for all $\theta_a \in \Theta$.*

This assumption states that the rewards $\mu_{t,a}$ and $\mu_{t',a}$ collected by the principal at any two time points t, t' where the agent chooses arm a are independent from each other. Assumption 3.2 is a mild assumption that is commonly encountered in many MAB models.

Let $T(a,t) = |\{\tau \in \mathcal{T} : \tau \leq t-1, i_\tau(\boldsymbol{\pi}_\tau) = a\}|$ be the number of time points that the agent selects arm a up to time t. Then, the principal's estimator for $\theta_a^0, \forall a \in \mathcal{A}$ is given by

$$\widehat{\theta}_{t,a} = \frac{1}{T(a,t)} \sum_{\tau=1}^{t-1} \mu_{\tau,a} \mathbb{1}\left(i_\tau(\boldsymbol{\pi}_\tau) = a\right) \tag{3.14}$$

which is the sample mean of the principal's reward outcomes for agent's arm a up to time t. If the principal's reward distribution $\mathbb{F}_{\theta_a^0,a}$ for any $a \in \mathcal{A}$ is an exponential family distribution where the sufficient statistic is equal to the random variable itself, such as Bernoulli, Poisson, and the multinomial distributions, then $\widehat{\theta}_{t,a}$ corresponds to the maximum likelihood estimator for θ_a^0.

3.3.1 Principal's ϵ-Greedy Algorithm

We develop an ϵ-greedy algorithm that integrates the principal's estimation problem and the incentive design problem in a practical learning framework. The pseudocode of the principal's ϵ-greedy algorithm is given in Algorithm 2.

During the first $n = |\mathcal{A}|$ time periods, the principal makes the agent select each of the n actions once so that the principal will be able to record a reward observation and have an initial estimate of θ_a^0 for all $a \in \mathcal{A}$. To achieve this, the principal offers the maximum possible incentive ($\overline{C}$) for the desired action which is sufficient to make it the agent's reward-maximizer action by Assumption 3.1. After this initialization period, at each time point $t \in [n+1, \ldots, T]$, the algorithm first updates the estimate of θ^0 for the most recently played action $i_{t-1}(\boldsymbol{\pi}_{t-1})$, and then samples a Bernoulli random variable x_t based on the exploration probability ϵ_t. If $x_t = 1$, then the algorithm performs a pure exploration step by simply choosing an incentive vector $\boldsymbol{\pi}_t = (\pi_{t,a})_{a\in\mathcal{A}}$ where each component $\pi_{t,a}$ is selected uniformly randomly from the compact set $\mathcal{C}$. On the other hand, if $x_t = 0$, then the principal performs a greedy exploitation by first updating their estimate $\widehat{\mathbf{s}}_t$ for the unknown mean rewards of the agent by solving the estimation problem (3.7). Next, the principal computes the vector of incentives $\mathbf{c}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t)$ that maximizes their estimated expected net reward at time t. The expected net reward of the principal is computed by subtracting the provided total incentives at that time step from the expected reward that the principal will collect through the action which will be chosen by the agent. However, since the agent's true rewards are unknown, the principal cannot exactly know in advance the action that will be chosen by the agent after the provided incentives. Therefore, the principal tries to incentivize the agent to select the action that is estimated to maximize the principal's expected net reward at that time step. This is ensured by incrementing the calculated incentive by an additional amount to account for the uncertainty in the estimates of the agent's expected rewards.

For that purpose, using $\widehat{\boldsymbol{\theta}}_t$ and $\widehat{\mathbf{s}}_t$, the principal first estimates the minimum incentives (denoted by $(\widetilde{c}_a^j)_{a\in\mathcal{A}}$) required to make the agent pick an action $j \in \mathcal{A}$ and the corresponding expected net reward value (denoted by $\widetilde{V}(j, \widehat{\mathbf{s}}_t; \widehat{\boldsymbol{\theta}}_t)$) that will be observed after action j is

Algorithm 2 Principal's ϵ-Greedy Algorithm

```
Set: $m \geq 4$, $\alpha > 0$
for $t \in [1, \dots, n]$ do
    Set: $\boldsymbol{\pi}_t = (\pi_{t,a})_{a \in \mathcal{A}}$ where $\pi_{t,a} = \overline{C}$ for $a = t$ and $\pi_{t,a} = 0$ for all $a \neq t$
    if $t \geq 2$ then $\widehat{\theta}_{t,i_{t-1}(\boldsymbol{\pi}_{t-1})} = \mu_{t-1,i_{t-1}(\boldsymbol{\pi}_{t-1})}$
    Observe: $i_t(\boldsymbol{\pi}_t) = \arg\max_{a \in \mathcal{A}} (s_a^0 + \pi_{t,a})$ and $\mu_{t,i_t(\boldsymbol{\pi}_t)}$
for $t \in [n+1, \dots, T]$ do
    Compute: $\widehat{\theta}_{t,i_{t-1}(\boldsymbol{\pi}_{t-1})} \in \frac{1}{T(i_{t-1}(\boldsymbol{\pi}_{t-1}),t)} \sum_{\tau=1}^{t-1} \mu_{\tau,i_{t-1}(\boldsymbol{\pi}_{t-1})} \mathbb{1}(i_\tau(\boldsymbol{\pi}_\tau) = i_{t-1}(\boldsymbol{\pi}_{t-1}))$
    Set: $\epsilon_t = \min\left\{1, {}^{m}\!/_{t}\right\}$
    Sample: $x_t \sim \text{Bernoulli}(\epsilon_t)$
    if $x_t = 1$ then
        Sample: $\pi_{t,a} \sim \mathcal{U}\left(\underline{C}, \overline{C}\right)$ for all $a \in \mathcal{A}$
        Set: $\boldsymbol{\pi}_t = (\pi_{t,a})_{a \in \mathcal{A}}$
    else
        Compute: $\beta_t = \sqrt{\frac{\log(\eta(1,t)-1)}{\alpha(\eta(1,t)-1)}}$ where $\eta(1,t) = \left|\{\tau : x_\tau = 1, n+1 \leq \tau \leq t-1\}\right|$
        Compute: $\widehat{\mathbf{s}}_t \in \arg\min\left\{L\left(\mathbf{s}, I_t(\boldsymbol{\Pi}_t), \boldsymbol{\Pi}_t\right) \middle| s_1 = 0, s_a \in \mathcal{S}, \forall a \in \mathcal{A}\right\}$
        for $j \in \mathcal{A}$ do
            Compute: $\widetilde{V}(j, \widehat{\mathbf{s}}_t; \widehat{\boldsymbol{\theta}}_t) = \widehat{\theta}_{t,j} - \left(\max_{a \in \mathcal{A}} \widehat{s}_{t,a}\right) + \widehat{s}_{t,j} - 2\beta_t$
        Compute: $j_t^* = \arg\max_{j \in \mathcal{A}} \widetilde{V}(j, \widehat{\mathbf{s}}_t; \widehat{\boldsymbol{\theta}}_t)$
        Set: $c_{j_t^*}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t) = \left(\max_{a \in \mathcal{A}} \widehat{s}_{t,a}\right) - \widehat{s}_{t,j_t^*} + 2\beta_t$ and $c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t) = 0$ for all $a \neq j_t^*$
        Set: $\boldsymbol{\pi}_t = (c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t))_{a \in \mathcal{A}}$
    Observe: $i_t(\boldsymbol{\pi}_t) = \arg\max_{a \in \mathcal{A}} (s_a^0 + \pi_{t,a})$ and $\mu_{t,i_t(\boldsymbol{\pi}_t)}$
```

taken by the agent:

$$\widetilde{c}_j^j = \left(\max_{a \in \mathcal{A}} \widehat{s}_{t,a}\right) - \widehat{s}_{t,j} + 2\beta_t \tag{3.15}$$

$$\widetilde{c}_a^j = 0, \quad \forall a \in \mathcal{A},\ a \neq j \tag{3.16}$$

$$\widetilde{V}(j, \widehat{\mathbf{s}}_t; \widehat{\boldsymbol{\theta}}_t) = \widehat{\theta}_{t,j} - \sum_{a \in \mathcal{A}} \widetilde{c}_a^j = \widehat{\theta}_{t,j} - \left(\max_{a \in \mathcal{A}} \widehat{s}_{t,a}\right) + \widehat{s}_{t,j} - 2\beta_t \tag{3.17}$$

where $\beta_t > 0, \forall t$. After computing these values for every action $j \in \mathcal{A}$, the principal chooses the set of incentives corresponding to action j_t^* that brings the highest $\widetilde{V}(j, \widehat{\mathbf{s}}_t; \widehat{\boldsymbol{\theta}}_t)$ value. The chosen vector of incentives is denoted by $\mathbf{c}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t) = (c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t))_{a \in \mathcal{A}}$ such that $c_{j_t^*}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t) =$

$(\max_{a\in\mathcal{A}} \widehat{s}_{t,a}) - \widehat{s}_{t,j_t^*} + 2\beta_t$ and $c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t) = 0, \forall a \neq j_t^*$ where $j_t^* \in \arg\max_{j\in\mathcal{A}} \widetilde{V}(j, \widehat{\mathbf{s}}_t; \widehat{\boldsymbol{\theta}}_t)$. We show that the design of these exploitation incentives are purposeful in the sense that they drive the agent's reward-maximizer action to be j_t^* with high probability. We formalize this property in Proposition 3.5 in the next subsection.

At the end of each time period, the principal provides the selected incentives $\boldsymbol{\pi}_t$ to the agent and observes the reward-maximizer arm $i_t(\boldsymbol{\pi}_t)$ chosen by the agent. As a result, the principal receives a net reward of $\mu_{t,i_t(\boldsymbol{\pi}_t)} - \sum_{a\in\mathcal{A}} \pi_{t,a}$, and the agent collects a total reward of $s^0_{i_t(\boldsymbol{\pi}_t)} + \pi_{t,i_t(\boldsymbol{\pi}_t)}$. We reiterate that the principal does not observe the agent's reward associated with the chosen action.

Remark 3.1 *The arithmetic operations performed to compute the exploitation incentives in lines 16-19 of Algorithm 2 have a complexity of $O(n)$ where $n = |\mathcal{A}|$. This implies that the computational complexity of the principal's bandit algorithm is linear in the dimension of the agent's model.*

3.3.2 Regret Bound

We compute the regret of a policy $\Pi_{\epsilon,T} = \{\boldsymbol{\pi}_t\}_{t\in\mathcal{T}}$ generated by Algorithm 2 by comparing it with an *oracle* incentive policy with respect to the cumulative expected net reward obtained by the principal. An oracle incentive policy is defined as the policy with perfect knowledge of all the reward expectations $\boldsymbol{\theta}^0$ and $\mathbf{s}^0$. Let $\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0)$ be the constant oracle incentives that maximize the principal's expected net reward at each time step over the time horizon $\mathcal{T}$. The oracle incentives are computed in a similar way to the computation of the exploitation incentives in Algorithm 2. We first solve for the minimum incentives required to make an action $j \in \mathcal{A}$ the reward-maximizer action of the agent and compute the associated expected net reward value $\widetilde{V}(j, \mathbf{s}^0; \boldsymbol{\theta}^0)$ as follows:

$$\widetilde{c}^j_j = \left(\max_{a\in\mathcal{A}} s^0_a\right) - s^0_j \tag{3.18}$$

$$\widetilde{c}^j_a = 0, \quad \forall a \in \mathcal{A},\ a \neq j \tag{3.19}$$

$$\widetilde{V}(j, \mathbf{s}^0; \boldsymbol{\theta}^0) = \theta^0_j - \sum_{a\in\mathcal{A}} \widetilde{c}^j_a = \theta^0_j - \left(\max_{a\in\mathcal{A}} s^0_a\right) + s^0_j \tag{3.20}$$

Then, the oracle policy chooses the set of incentives corresponding to the agent action $j^{*,0}$ that has the highest $\widetilde{V}(j, \mathbf{s}^0; \boldsymbol{\theta}^0)$ value, that is $j^{*,0} := \arg\max_{j\in\mathcal{A}} \widetilde{V}(j, \mathbf{s}^0; \boldsymbol{\theta}^0)$. We note that by construction of the oracle incentives, this action is same as the action that maximizes the agent's total reward after the incentives, i.e., $j^{*,0} = i(\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0)) = \arg\max_{a\in\mathcal{A}} s^0_a + c_a(\boldsymbol{\theta}^0, \mathbf{s}^0)$ where

$$c_{i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))}(\boldsymbol{\theta}^0, \mathbf{s}^0) = \max_{a\in\mathcal{A}} s^0_a - s^0_{i(\mathbf{c}^0(\boldsymbol{\theta}^0,\mathbf{s}^0))} + \varsigma \tag{3.21}$$

$$c_a(\boldsymbol{\theta}^0, \mathbf{s}^0) = 0, \quad \forall a \neq i(\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0)). \tag{3.22}$$

for a sufficiently small constant $\varsigma > 0$ which helps avoiding the occurrence of multiple maximizer actions for the agent. Then, the principal's expected net reward at any time step under the oracle policy is given as

$$V(\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0); \boldsymbol{\theta}^0) = \theta^0_{i(\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0))} - \max_{a \in \mathcal{A}} s^0_a + s^0_{i(\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0))} - \varsigma. \tag{3.23}$$

Similarly, we compute $V_t(\boldsymbol{\pi}_t; \boldsymbol{\theta}^0)$ as the expected net reward of the principal at time t under the incentives generated by Algorithm 2 as

$$V_t(\boldsymbol{\pi}_t; \boldsymbol{\theta}^0) = \theta^0_{i_t(\boldsymbol{\pi}_t)} - \sum_{a \in \mathcal{A}} \pi_{t,a} \tag{3.24}$$

where $i_t(\boldsymbol{\pi}_t)$ is as given in line (21) of Algorithm 2. Lastly, we define the regret of a policy $\Pi_{\epsilon,T} = \{\boldsymbol{\pi}_t\}_{t \in \mathcal{T}}$ with respect to the cumulative expected net reward obtained by the principal.

$$\text{Regret}\,(\Pi_{\epsilon,T}) = \sum_{t \in \mathcal{T}} V(\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0); \boldsymbol{\theta}^0) - V_t(\boldsymbol{\pi}_t; \boldsymbol{\theta}^0) \tag{3.25}$$

We provide a rigorous regret bound for the principal's ϵ-greedy algorithm in Theorem 3.2. We next present several intermediate theoretical results that will be used to prove our regret bound.

Lemma 3.1 *Let $\mathcal{T}^{\text{xplore}} \in \mathcal{T}$ and $\mathcal{T}^{\text{xploit}} \in \mathcal{T}$ be the set of random time steps that Algorithm 2 performs exploration (lines 11-12) and exploitation (lines 15-20), respectively. Then, the following probability bound holds at any $t \in \mathcal{T}^{\text{xploit}}$:*

$$\mathbb{P}\left(\max_{a \in \mathcal{A}} \widehat{s}_{t,a} - \widehat{s}_{t,j^*_t} + 2\beta_t \geq \max_{a \in \mathcal{A}} s^0_a - s^0_{j^*_t}\right) \\ > 1 - \exp\left(-\alpha(\eta(1,t) - 1)\beta_t^2 - \log \beta_t + n \log(R_{\max} - R_{\min})\right) \tag{3.26}$$

where $\eta(1,t) = |\{\tau : 1 \leq \tau \leq t-1, \tau \in \mathcal{T}^{\text{xplore}}\}|$ as introduced in Theorem 3.1.

The main observation required for the proof of this lemma is that the desired event is implied by the event $\|\mathbf{s}^0 - \widehat{\mathbf{s}}_t\|_\infty \leq \beta_t$. Hence, the lower bound on the probability that the desired event holds is directly obtained by using the result of Corollary 3.1.

Proposition 3.5 *At any time $t \in \mathcal{T}^{\text{xploit}}$, the probability that the agent will pick arm j^*_t after the exploitation incentives $\mathbf{c}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t)$ is bounded by*

$$\mathbb{P}\left(j^*_t = i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t))\right) > 1 - \exp\left(-\alpha(\eta(1,t) - 1)\beta_t^2 - \log \beta_t + n \log(R_{\max} - R_{\min})\right). \tag{3.27}$$

We recall that the principal estimates that the action j^*_t will yield the highest expected net reward to themselves, and hence desires that j^*_t will be chosen by the agent after observing the exploitation incentives. From this perspective, the implication of the last result is that the exploitation incentives are successful in making j^*_t the total reward maximizer action for the agent with high probability. This result is proved in a straightforward way by using the definition of our exploitation incentives and the result of Lemma 3.1.

Proposition 3.6 *Suppose $\beta_t = \sqrt{\frac{\log(\eta(1,t)-1)}{\alpha(\eta(1,t)-1)}}$ for all $t \in \mathcal{T}$. Then, we have*

$$\mathbb{P}\left(i(\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0)) \neq i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t))\right) \leq \frac{4n}{\eta(1,t)-1} + \frac{2n(R_{\max} - R_{\min})^n\sqrt{\alpha}}{\sqrt{(\eta(1,t)-1)\log(\eta(1,t)-1)}}. \tag{3.28}$$

This result shows a decreasing (over time) upper bound on the probability that the action selected by the agent under the exploitation incentives $\mathbf{c}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t)$ will not be the true reward-maximizer action that would be selected by the agent under the oracle incentives $\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0)$. The proof follows by mainly using the finite-sample concentration bounds for the principal's estimates $\widehat{\boldsymbol{\theta}}_t$ and $\widehat{\mathbf{s}}_t$ and the result of Proposition 3.5.

Theorem 3.2 (Finite-Sample Regret Bound) *The regret of a policy $\Pi_{\epsilon,T}$ computed by the Principal's ϵ-Greedy Algorithm (2) is bounded by*

$$\begin{aligned}\text{Regret}\left(\Pi_{\epsilon,T}\right) \leq\ & \frac{8}{\sqrt{\alpha}}\sqrt{T\log T} + 8n\left(\overline{C} - \underline{C} + \text{diam}(\Theta)\right)(R_{\max} - R_{\min})^n\sqrt{\alpha}\sqrt{T} \\ & + \left(n(\overline{C} - \underline{C})(m+8) + \text{diam}(\Theta)m\right)\log T \\ & + m\left(n(\overline{C} - \underline{C}) + \text{diam}(\Theta)\right) + B_1 + B_2\end{aligned} \tag{3.29}$$

where $B_1 + B_2 = \frac{4}{\sqrt{\alpha}}\sqrt{\frac{\log(m-1)}{m-1}} + \frac{2n\left(2(\overline{C}-\underline{C})+\text{diam}(\Theta)\right)(R_{\max}-R_{\min})^n\sqrt{\alpha}}{\sqrt{m-2}} + \frac{4n\left(\overline{C}-\underline{C}+\text{diam}(\Theta)\right)}{m-1}$ *and* $\text{diam}(\Theta) = \max_{a,a' \in \mathcal{A}} \theta^0_a - \theta^0_{a'}$ *are finite and strictly positive constants.*

Remark 3.2 *This finite-sample regret bound corresponds to an asymptotic regret at a rate of order $O(\sqrt{T\log T})$ for the proposed adaptive incentive design framework.*

The proof details for all the results in this section can be found in Appendix B.1.2.

3.4 Agent's Information Rent

In this section, we present a discussion of our repeated principal-agent model from the agent's perspective. According to the information structure that we study in this chapter, the only observable information to the principal are the actions taken by the reward-maximizing agent. The principal needs to estimate the agent's true preferences and rewards under this information asymmetry. Our data-driven framework assumes that the agent acts truthfully, so that the sequence of their actions is selected in a consistent way with their true expected reward vector $\mathbf{s}^0$. In spite of that, there exists an unavoidable information rent given to the agent due to the information asymmetry in our model as in every other adverse selection model. This strictly positive information rent always presents and is an inherent part of our hidden rewards setting. However, the principal's goal is to minimize the amount they pay to the agent on top of this minimal amount of information rent. The way we design the principal's exploitation incentives given in (3.15)-(3.17) allows the principal to achieve this goal. Assuming that the agent picks their actions with respect to a fixed expected reward vector (that is only known by the agent), we implicitly induce incentive compatibility when we optimize the principal's incentives such that they will make the agent pick the arm that the principal wants them to pick. However, the agent could just pretend that their true rewards $\mathbf{s}^0$ are different from the beginning of the sequential game, and pick all their actions in accordance with these "pretended" rewards in order to extract a higher information rent from the principal and maximize their total rewards. Under the hidden rewards setting, there is no way for the principal to prohibit the agent from this misbehavior which allows them to maximize the information rent they collect from the principal as we show in this section. We also note that avoiding this extra information rent could be possible in other principal-agent designs where more information about the agent's reward model is accessible by the principal. For instance, the principal could know in advance the discrete set of the agent's mean reward values without necessarily knowing which value belongs to which action. Analyzing such settings in which the principal would be able to offer incentives that get the agent to reveal their true preferences is beyond the scope of this chapter, yet it stands as an interesting future research direction.

From the standpoint of the reward-maximizer agent, we can formalize the agent's problem as an optimization model that maximizes the information rent they are extracting from the principal. The main observation here is that the maximum possible value of the agent's information rent is finite and can be achieved by a sophisticated agent who is also knowledgeable about the principal's rewards. Recall that the principal offers the incentives that will induce the agent to pick the action which would yield the highest net expected reward to the principal. Assuming that the agent is informed about $\boldsymbol{\theta}^0$ and $\mathbf{s}^0$, they could demand extra payment from the principal by taking their actions with respect to a fixed "pretended" mean reward vector $\check{\mathbf{s}}(\mathbf{s}^0, \boldsymbol{\theta}^0)$ throughout the entire time horizon. We next formalize this idea in the following optimization problem.

$$\begin{aligned} \check{\mathbf{s}}(\mathbf{s}^0, \boldsymbol{\theta}^0) \;\; \in \;\; & \arg\max_{\mathbf{s},\boldsymbol{\pi}} \;\; s_a^0 + \pi_a \\ & \text{s.t. } a = \arg\max_{a' \in \mathcal{A}} \theta_{a'}^0 - \pi_{a'} \\ & \quad \pi_a > 0,\ \pi_a \in \mathcal{C},\ \pi_{a'} = 0\ \forall a' \in \mathcal{A} \setminus \{a\} \\ & \quad b = \arg\max_{a' \in \mathcal{A}} s_{a'} + \pi_{a'} \\ & \quad a = b \end{aligned} \tag{3.30}$$

The objective function of this optimization problem maximizes the agent's true expected reward (after the incentives) obtained from selecting action a which is further specified by the constraints. The first constraint implies that action a maximizes the principal's expected net reward when the incentives $\boldsymbol{\pi}$ are selected as given in the second set of constraints. Then, the third and last constraints ensure that the incentives are designed in such a way that action a is also the reward-maximizer for the agent who pretend their rewards as $\check{\mathbf{s}}$.

Proposition 3.7 *The agent's optimization problem (3.30) is feasible, and the agent can maximize their information rent by choosing its solution* $\check{\mathbf{s}}(\mathbf{s}^0, \boldsymbol{\theta}^0)$ *as their "pretended" fixed mean reward vector during the course of their repeated play with the principal.*

The complete proof of this proposition is provided in Appendix B.1.3. Recall that in Section 3.3.2, we show that when the agent plays truthfully in accordance with their true mean reward vector $\mathbf{s}^0$ and the principal follows the oracle incentive policy $\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0)$, then the agent gets their minimum possible expected total reward. We start the proof by showing that this solution is feasible to the problem (3.30), yet it yields the worst-case result for the agent. We continue by proving the existence of other feasible solutions which use mean reward vectors that are different than $\mathbf{s}^0$ and return higher information rents to the agent. These feasible solutions are proposed for two mutually exclusive cases based on whether the maximizer actions of the principal's and the agent's mean rewards, $\boldsymbol{\theta}^0$ and $\mathbf{s}^0$, are the same with each other or not. We next present two numerical examples that illustrate the feasible solutions proposed in the proof for each of these two cases.

Example 3.1 *Consider a model with three actions* $\mathcal{A} = \{1, 2, 3\}$*. Let the agent's true mean reward vector be* $\mathbf{s}^0 = (s_1^0, s_2^0, s_3^0) = (0, 4, 3)$ *and the principal's true mean reward vector be* $\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2^0, \theta_3^0) = (1, 8, 2)$*. Notice that the principal does not need to incentivize the agent in this case because the reward-maximizer actions for both parties are the same with each other. The principal can just offer the incentives* $\boldsymbol{\pi} = (0, 0, 0)$ *that yield the highest possible expected net reward to them (which is 8) and the worst-case expected total reward to the agent (which is 4). Now, suppose that the agent is untruthful and playing according to the rewards* $\mathbf{s} = (0, 4, 9.5)$*. In that case, if the principal offers the same incentives* $\boldsymbol{\pi}$*, then the agent will pick the third action and the principal's expected net reward will be 2. However, the principal can obtain a relatively higher expected net reward by offering a different set of incentives that will get the agent to pick the second action. Suppose that the principal gives the incentives* $\widetilde{\boldsymbol{\pi}} = (0, 5.9, 0)$*, which is a feasible solution to the agent's optimization problem together with the chosen* $\mathbf{s}$*. Then, the expected net reward of the principal becomes* $\theta_2^0 - \widetilde{\pi}_2 = 2.5$ *whereas the agent's expected total reward jumps to* $s_2^0 + \widetilde{\pi}_2 = 9.5$*. As a result, the agent collects an extra information rent of* $9.5 - 4 = 5.5$ *which is the difference between their expected total rewards when they are truthfully playing with* $\mathbf{s}^0$ *and when they are pretending their rewards are* $\mathbf{s}$*.*

Example 3.2 *Consider a model with four actions* $\mathcal{A} = \{1, 2, 3, 4\}$*. Let the agent's true mean reward vector be* $\mathbf{s}^0 = (s_1^0, s_2^0, s_3^0, s_4^0) = (0, 4, 3, 6)$ *and the principal's true mean reward vector be* $\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2^0, \theta_3^0, \theta_4^0) = (1, 8, 7, 2)$*. If the agent plays in accordance with their true rewards, then* $\mathbf{s} = \mathbf{s}^0$ *and* $\boldsymbol{\pi} = (0, 2.1, 0, 0)$ *will yield a feasible solution to (3.30) with* $a = b = 2$*. With this solution, the principal's expected net reward will be* $\theta_2^0 - \pi_2 = 8 - 2.1 = 5.9$ *and the agent's expected total reward will be* $s_2^0 + \pi_2 = 4 + 2.1 = 6.1$*. On the other hand, consider the rewards* $\mathbf{s} = (0, 4, 3, 7.8)$ *and the incentives* $\widetilde{\boldsymbol{\pi}} = (0, 3.9, 0, 0)$*. These vectors result in another feasible solution in which the principal's expected net reward decreases to* $\theta_2^0 - \widetilde{\pi}_2 = 8 - 3.9 = 4.1$

whereas the agent's expected total reward rises to $s_2^0 + \widetilde{\pi}_2 = 4 + 3.9 = 7.9$*. As can be seen, the agent gains a higher information rent in this case by pretending their rewards are* $\mathbf{s}$ *and capturing an extra amount of* 1.8 *from the principal's expected profits.*

As stated before, achieving the maximum information rent would require a significant amount of sophistication from the agent, which may not be the case in practice. As the agent is less knowledgeable about the principal's model, they will get less information rent. However, regardless of the knowledge level, the agent's behavior needs to be based on a fixed vector of mean rewards. Whether it is the true vector or a "pretended" vector, the taken actions will be essentially consistent with the same reward vector throughout the entire time horizon — aligning with the underlying assumption in our repeated principal-agent model. Therefore, we highlight that our framework is designed to maximize the principal's expected net reward subject to the information rent that the agent takes.

3.5 Numerical Experiments

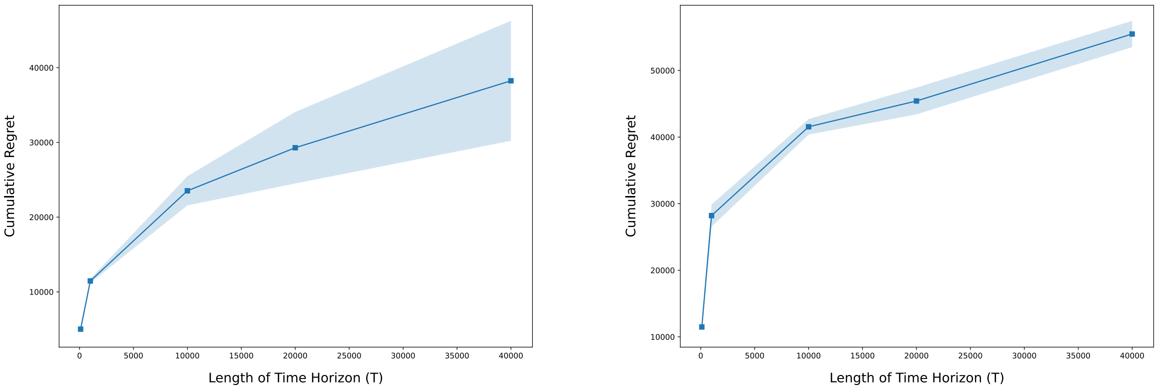
We aim to support our theoretical results for the repeated principal-agent models with hidden agent rewards by conducting simulation experiments in which the proposed data-driven incentives are compared with the derived oracle incentives. Our experimental setting is based on an instance of the sustainable and collaborative transportation planning problem introduced in Section 3.1.1.1.

Consider a transportation network composed of the linehaul and backhaul customers of a shipper who acknowledges that their total carbon emissions and cost of logistics operations can be reduced by the use of pre-planned integrated outbound-inbound routes. Let $\mathcal{A} = \{1, \dots, n\}$ be the discrete set of all possible pure inbound routes, pure outbound routes, and the offered outbound-inbound routes for the given network. Each route $a \in \mathcal{A}$ brings a stochastic cost to the shipper with an expectation ζ_a^0. Note that our setup can handle stochastic costs (as opposed to rewards) by setting the expected reward as the negative of the expected cost, i.e., $\theta_a^0 = -\zeta^0$. Thus, we will continue using our standard notation. Suppose the shipper works with a carrier who wants to maximize their total expected profit (note that the s_a^0's are invisible to the shipper) and may be also serving to other shippers. The goal of the shipper is to motivate the carrier to collaborate with them and perform the most efficient (for the shipper) outbound-inbound routes over a sequence of shipment requests, $\{1, \dots, T\}$.

We run our experiments for multiple combinations of the parameters $n \in \{5, 10\}$ (i.e., total number of alternative routes) and $T \in \{10^2, 10^3, 10^4, 2 \cdot 10^4, 4 \cdot 10^4\}$ (i.e., total number of shipment requests). Each setting is replicated five times, and the average and standard deviation of our regret metric (3.25) are reported across these replicates. We assume that the feasible range of incentives is given by $\mathcal{C} = [-20, 60]$, and the principal's stochastic costs for each route $a \in \mathcal{A}$ follow a Gaussian distribution $\mathcal{N}(\theta_a^0, 5)$. The input parameter m for Algorithm 2 is chosen as $m = 30$ in all settings which implies that the principal explores during the first 30 periods of the considered contract horizon after the initialization period (see lines 2-4). The values selected for the vectors $\boldsymbol{\theta}^0$ and $\mathbf{r}^0$ are presented in Table B.1 in Appendix B.2.

Figure 3.1 shows the cumulative regret accrued by the principal's ϵ-greedy algorithm for different values of n and T. As expected, our approach achieves a sublinear regret that matches with the asymptotic order $O(\sqrt{T}\log T)$ proven by our theoretical analyses.

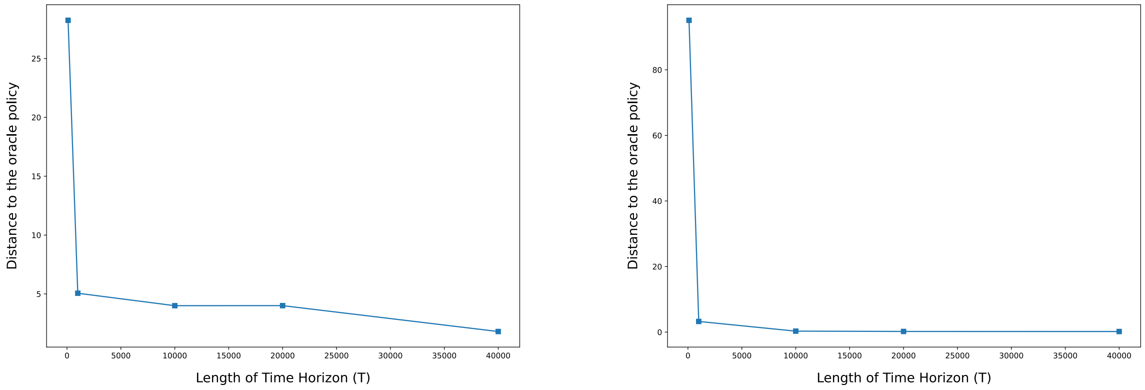
A significant theoretical challenge in the shipper's problem is that they need to compute an incentive amount for each and every route as accurately as possible in order to optimize their own profits. The shipper has to estimate the expected profits consistently not only for the desired integrated outbound-inbound routes but also for all the separate outbound and inbound routes. Thus, the difficulty level of the shipper's problem increases as the size of the carrier's action space increases. To highlight this challenge, we present a more direct measure of how close the menu of incentives designed by Algorithm 2 gets to the oracle menu of incentives at the end of a finite contract horizon. As highlighted, because every alternative route matters the same, we measure the distance between the two sets of incentives by using the ℓ_1 norm – in which all the entries of the vectors are weighted equally. As can be seen in Figure 3.2, the proposed incentive design mechanism is able to consistently converge to the oracle incentive policy, and it achieves a better convergence as the length of the contract horizon gets longer. Further, a comparison of Figures 3.2a and 3.2b reveals that our data-driven framework is able to achieve the same accuracy even when the size of action space is doubled.



(a) Regret for $n = 5$

(b) Regret for $n = 10$

Figure 3.1: The cumulative regret of the policies generated by Algorithm 2. The shaded regions represent the standard error over all replications.



(a) $\|\boldsymbol{\pi}_T - \mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0)\|_1$ for $n = 5$

(b) $\|\boldsymbol{\pi}_T - \mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0)\|_1$ for $n = 10$

Figure 3.2: The ℓ_1 distance between the oracle incentives and the incentives reached by Algorithm 2 at the end of the time horizon. For any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^K$, the ℓ_1 distance is defined by $\|\mathbf{x} - \mathbf{y}\|_1 = \sum_{k=1}^{K} |x_k - y_k|$.

3.6 Conclusions

We conclude by summarizing this chapter's primary contributions to the principal-agent theory and data-driven contract design literature. In this chapter, we introduce a novel repeated principal-agent setting which has not been explored in earlier studies even though it is applicable to many real-life problems. In particular, we analyze an adverse selection model where the principal can solely observe the agent's decisions while the agent's true preferences and rewards stay hidden from the principal. To enhance the practical relevance of our theoretical studies, we keep our model as generic as possible.

The two integrated dimensions of the considered research problem are: i) estimation of the agent's unknown reward model, and ii) design of adaptive incentives that will maximize the principal's cumulative net rewards over a finite time horizon. We first introduce our novel estimator and prove its identifiability and finite-sample concentration bound. Then, we formalize the principal's data-driven incentives and unite them with our estimator in an ϵ-greedy bandit algorithm. We conduct a rigorous regret analysis for this algorithm and support our theoretical results by demonstrating the performance of our approach in the simulations of a sustainable route planning model.

In the *hidden rewards* model herein, we assume that the reward-maximizing agent has full knowledge of their reward model and is able to take the true reward-maximizer action at every period. A more challenging model would consider an agent with imperfect knowledge of their model. In that case, our analyses will also need to involve the learning process of the agent who trains their algorithm on top of the principal's learning process. As can be expected, the dynamic interactions between these two learning parties will add substantial complexity both to the estimation and the incentive design problems. In the next chapter, we will extend our model and analyses to address this intricate scenario, and we will also highlight possible future directions for further work building upon our contributions.

Chapter 4

Repeated Principal-Agent Games with Hidden Rewards: Learning Agents

4.1 Introduction

In practice, incentive providers (i.e., principals) often cannot observe the reward realizations of incentivized agents, which is in contrast to many principal-agent models that have been previously studied. This information asymmetry challenges the principal to consistently estimate the agent's hidden rewards by solely watching the agent's decisions, which becomes even more challenging when the agent also has to learn its own rewards. This intricate setting offers not only wide practical relevance but also poses interesting open theoretical questions, which are intended to be addressed by this chapter.

This chapter analyzes the *hidden rewards* game introduced in Chapter 3 in a more complex environment, where: there is a learning agent that tackles a multi-armed bandit (MAB) problem to acquire the knowledge of their true reward-maximizers under the offered incentives. In this repeated principal-agent game, the main theoretical challenge is sourced from the dynamic and sequential interactions taking place between the two strategic decision-makers. In each play of the game, first the principal offers a menu of incentives to the agent, and then the learning agent makes a choice from a finite set of actions, which in turn determines the rewards collected by both players. In other words, there is a two-sided sequential externality in this setting, whereby the agent's imperfect knowledge imposes additional costs on the principal and the principal's incentives impose a more challenging decision-making environment for the agent with imperfect knowledge. This chapter considers that both the principal and the agent observe stochastic rewards with unknown (to both) expectations, and that both parties aim to maximize their own cumulative expected rewards at the end of the game.

For this complex setting, we jointly tackle the two coupled facets of the principal's problem:

i) learning the agent's hidden reward expectations by training a consistent estimator,

ii) designing an incentive mechanism to lead the agent's learning algorithm in favor of the principal.

The statistical and regret analyses in this chapter show that this setup yields substantially higher complexity and necessitates a distinct theoretical analysis compared to the "perfect agent" setting in Chapter 3. The implication is that the principal's learning algorithm is trained on top of the agent's learning process, and the major complication arises from the uncertainty in the agent's choices. There is now no guarantee that these choices are the true maximizers of the agent's rewards. Both the estimation errors and cost of explorations incurred by the agent directly contribute to the cumulative regret of the principal over the considered time horizon. By marrying classical principal-agent theory with statistics and online learning, we offer a robust data-driven incentive design framework without necessarily restricting the type of the agent's MAB algorithm and prove that the proposed policy attains sublinear regret for the principal.

4.1.1 Motivating Real-Life Applications

4.1.1.1 Adaptive Contracts for Sustainable Energy Aggregation

Among many attempts to address the devastating and growing impact of the climate crisis, transition to clean green energy stands as one of the most effective and widespread processes. Today, there is a rising awareness on the value of renewable and large-scale distributed energy storage for the clean energy transition. With this motivation, the so-called "aggregators" have started to play a key role in electricity markets. An aggregator is a company operating a grid-scale virtual power plant that pools energy supply available in their distributed battery systems and sells this capacity in electricity markets during peak demand or emergency periods (International Renewable Energy Agency 2019, Kennedy 2021, Berntzen et al. 2021). This aggregation operation provides substantial advantages to the market stakeholders, the utility companies, and the independent aggregator firms on the supply side and the residential and commercial customers on the demand side. The primary benefits can be summarized as reducing costs for utilities and communities, decreasing carbon emissions, and improving power reliability. To achieve these benefits, the aggregator has to motivate the customers to be flexible and allow the aggregator to use the backup power available in their electric vehicles or solar energy storage systems (Bindra and Revankar 2018, Biggins et al. 2022). For this purpose, the aggregator offers benefits to the participating customers such as a compensation for their contribution to the energy supply in the grid. On the higher level, it falls to the utility company to encourage the aggregator to initiate flexibility in their storage capacity through increasing their investments and managing voluntary customer participation in the aggregation program.

The sequential game between the utility company (i.e., principal) and independent aggregator company (i.e., agent) can be effectively modeled as a repeated unobserved rewards setting with two strategic and learning players. In this model, the actions of the aggregator is defined as the amount of storage capacity (MW/h) (sourced from electric vehicles or household heat pumps) reserved for the use of utility grid, and the payments of the utility company are defined as the service fees offered for purchased storage (MW/h) in an aggregation contract (which is organized typically on an hourly basis). The realized profits of the utility and the aggregator observed as a result of these contracts depend on several sources of uncertainty including variations in renewable energy generation (e.g., wind, solar, hydro, etc.), electricity demand, and market prices. Such contracts are utilized in recent applications such as the Emergency Load Reduction program launched by Tesla and PG&E (McCarthy 2022) and the Resilient Home program initiated by Sunrun and East Bay Community Energy in California (East Bay Community Energy 2020). Taking into account the conflicting objectives of multiple stakeholders and unknown stochastic components in these systems, we believe that the adaptive incentive policies proposed in this chapter can help the utility company to offer smart contracts that explicitly considers their sequential interactions with the aggregator and expands the participation of the aggregator.

4.1.1.2 Deforestation Incentives for Payment for Ecosystem Services

According to the UN Food and Agriculture Organization, deforestation since 1990 is estimated to have reached 420 million hectares worldwide, with an annual rate of around 10 million hectares every year since then, the majority of which belongs to tropical forests (Food and FAO). However, the rate of forest loss is reported to have declined substantially in the last five years due to various interventions, such as Payment for Ecosystem Services (PES) (Busch and Ferretti-Gallon 2017, Seymour and Harris 2019). PES programs establish a positive economic incentive mechanism by which governments and non-governmental organizations contract owners of natural resources (i.e., local landowners or service providers, such as private forest owners and farmers) to incentivize them to prevent environmental degradation, such as deforestation (Warnes et al. 2023).

From the perspective of payment providers, the goal of these incentive mechanisms is to maximize the environmental benefits obtained from the amount of conserved forests with the minimal amount of payments. However, a prevalent obstacle in effectiveness of these contracts involves the information disparities between program designers and the forest owners (Salzman et al. 2018). Typically, the forest owners hold private information about their opportunity costs and the amount of deforestation they would choose without any incentives (Engel et al. 2016). Only a few papers in the literature have studied the contract design problem in the PES context by using a principal-agent framework in which the agent has a privately-known baseline conservation level for a known initial forest area (Mason and Plantinga 2013, Li et al. 2023). These studies characterize the optimal contracts for static one-shot games between providers and forest owners, while leaving the analysis open for dynamic performance-based incentive policies. To bridge this gap, we suggest that PES incentives should incorporate the effects of learning from forest owners' conservation actions over time. Our *hidden rewards* model can be effectively extended to formalize this information asymmetric setup, and the proposed data-driven incentive design framework can help reveal the true willingness of forest owners for their land use.

4.1.2 Main Contributions and Chapter Outline

We next present an outline of this chapter by featuring our main contributions to the theory and applications in the related literature areas, which are summarized in Section 3.1.3 of Chapter 3.

Consistent estimator fed by the agent's MAB. We start by introducing our repeated principal-agent game with hidden rewards of a learning agent in Section 4.2.1. In accordance with this model, in Section 4.2.2, we propose a novel estimator for the agent's expected reward for each bandit arm – which uses as data the sequence of incentives offered and subsequently chosen arms by the agent's MAB algorithm. Our estimator is formulated exactly as a linear optimization model without assuming any functional form or any specific distributional property. Using this formulation, we next prove an identifiability property and a finite-sample concentration bound of the proposed estimator under a mild assumption on the probability that the agent's MAB algorithm does not select the true reward-maximizer arms in response to the offered incentives. We present these statistical results in Sections 4.2.3 and 4.2.4.

Robust, data-driven incentive policy. Section 4.3.1 embeds our consistent estimator into a MAB framework and presents the principal's adaptive and data-driven incentive policy using a practical and computationally efficient ϵ-greedy approach. By utilizing the finite-sample consistency results we derived for our estimator, we compute the regret of the proposed policy with respect to an oracle incentive policy that maximizes the principal's expected net reward at each time step under perfect knowledge of all reward expectations. Section 4.3.2 presents a rigorous sublinear regret bound for the principal under the sequential uncertainty imposed through the agent's choices.

Agent's behavior. In Section 4.4, we present a theoretical analysis and a discussion from the perspective of the selfish learning agent. We highlight that our statistical consistency and regret bound results for the principal are proven without restricting the type or structure of the agent's MAB algorithm. As mentioned above, our only assumption about the agent's behavior is a probability bound associated with the inaccuracy of the arms chosen by their algorithm. In Section 4.4.1, we show the mildness of our assumption by proving that it is satisfied when the agent uses a naive ϵ-greedy algorithm to make their decisions throughout the sequential game. Furthermore, in Section 4.4.2, we discuss that the learning framework proposed for the principal considers a self-interested agent with particular sophistication where they are not knowledgeable about or attempting to learn the principal's model. In that respect, our approach can be regarded as the worst case bound for the circumstances in which the agent has enough sophistication to maximize their information rent. However, regardless of the level of agent's complexity, our data-driven incentive mechanism is designed in a way that maximizes the principal's cumulative expected net reward subject to the existence of the agent's information rent.

Numerical results. Lastly, we conduct simulation experiments in the context of sequential aggregator contracts for the green energy storage operations described above. In Section 4.5, we share details of our experimental setting and numerical results supporting our finite-sample bounds on concentration of the proposed estimator and regret of the proposed incentive policy. We demonstrate the applicability and efficiency of our framework in enhancing a renewable, reliable, and smart utility grid for communities.

We conclude this chapter in Section 4.6 by discussing promising future directions. The proofs of all theoretical results provided in the main text are included in Appendix C.

4.2 Principal's Consistent Estimator

This section presents the studied repeated adverse selection model, our novel estimator, and the associated statistical consistency results. We note that this chapter follows the same conventions for mathematical notation as those given in Section 3.1.4 of Chapter 3.

4.2.1 The Repeated Game with Hidden Rewards of Learning Agents

We consider a sequential game between an incentive-provider principal and an incentivized agent over a finite time horizon $\mathcal{T} = [1, \ldots, T]$. The agent solves a MAB problem with a discrete set of arms (or actions) $\mathcal{A} = \{1, \ldots, n\}$. At each time step $t \in \mathcal{T}$, the principal first chooses an incentive amount $\pi_{t,a}$ for each bandit arm of the agent and offers the vector of incentives $\boldsymbol{\pi}_t = (\pi_{t,a})_{a \in \mathcal{A}}$. Then, the agent's MAB algorithm selects the arm $\upsilon_t(\boldsymbol{\pi}_t)$ which brings *i)* a stochastic reward outcome to the principal denoted by $\mu_{t,\upsilon_t(\boldsymbol{\pi}_t)}$ that follows a distribution $\mathbb{F}^{\text{pr}}_{\theta^0_{\upsilon_t(\boldsymbol{\pi}_t)},\upsilon_t(\boldsymbol{\pi}_t)}$ associated with the arm $\upsilon_t(\boldsymbol{\pi}_t)$ with expectation $\theta^0_{\upsilon_t(\boldsymbol{\pi}_t)} \in \Theta$ where Θ is a known compact set, and *ii)* a stochastic reward outcome to the agent denoted by $\rho_{t,\upsilon_t(\boldsymbol{\pi}_t)}$ that follows a distribution $\mathbb{F}^{\text{ag}}_{r^0_{\upsilon_t(\boldsymbol{\pi}_t)},\upsilon_t(\boldsymbol{\pi}_t)}$ associated with the arm $\upsilon_t(\boldsymbol{\pi}_t)$ with expectation $r^0_{\upsilon_t(\boldsymbol{\pi}_t)} \in \mathcal{R}$ where $\mathcal{R} = [R_{\min}, R_{\max}] \subset \mathbb{R}$ is a known compact set such that $R_{\max} - R_{\min} \geq 1$. We highlight that the principal can only observe the selected arm $\upsilon_t(\boldsymbol{\pi}_t)$ and their own net reward realization $\mu_{t,\upsilon_t(\boldsymbol{\pi}_t)} - \sum_{a \in \mathcal{A}} \pi_{t,a}$ at the end of each period.

In this setting, the ground truth mean reward vectors $\mathbf{r}^0 = (r^0_a)_{a \in \mathcal{A}}$ and $\boldsymbol{\theta}^0 = (\theta^0_a)_{a \in \mathcal{A}}$ are unknown both to the agent and to the principal. To ensure that our research problems are well-posed, it suffices to assume that the feasible range of the principal's incentives subsumes the range of the agent's reward expectations.

Assumption 4.1 *The incentives $\pi_{t,a}$ for all $a \in \mathcal{A}$ belongs to a compact set $\mathcal{C} = [\underline{C}, \overline{C}]$ where $\underline{C} = R_{\min}$ and $\overline{C} = R_{\max} + \gamma$ for some constant $0 < \gamma \leq R_{\max} - R_{\min} - 1$.*

Similar to Assumption 3.1 in the previous chapter, Assumption 4.1 ensures that the magnitudes of the principal's incentives can be chosen sufficiently large to change the relative ordering of the bandit arms with respect to their expected rewards after adding the incentives. This assures that the principal is able to provide incentives that will steer the agent's decisions into the desirable ones.

4.2.2 The Consistent Estimator

The problem of designing adaptive incentives throughout the described repeated principal-agent play involves a fundamental challenge for the principal: estimating the agent's mean rewards $\mathbf{r}^0$ by only using the data of past incentives offered and chosen arms in response to these incentives. Because we consider an imperfect-knowledge agent, the agent also trains their own estimator to predict $\mathbf{r}^0$ and chooses their arms by following a sequential learning algorithm that aims to maximize their estimated total mean reward after incentives. This suggests that the principal indeed tries to estimate a certain reward vector, such that the sequence of chosen arms are the maximizers of that vector plus the incentive vectors offered at the corresponding time periods. Because of this structure, the principal's estimation in this setting entails the same ambiguity discussed in Section 3.2.2.

To resolve this issue and ensure that our estimator satisfies an identifiability property, we apply a dimensionality reduction to the agent's model and define the *normalized* mean reward vector $\mathbf{s}$ as before.

Definition 4.1 *For a mean reward vector* $\mathbf{r} = (r_1, r_2, \ldots, r_n) \in \mathcal{R}^n$*, we define* $\mathbf{s}$ *as the normalized mean reward vector that is without loss of generality defined by* $\mathbf{s} := \mathbf{r} - r_1\mathbf{1}_n = (0, r_2 - r_1, \ldots, r_n - r_1)$ *and belongs to the compact set* $\mathcal{S}^n = [R_{\min} - R_{\max}, R_{\max} - R_{\min}]^n$*.*

We recall that this normalization does not change the accuracy of our estimation because the maximizer entries of $\mathbf{r}^0 + \boldsymbol{\pi}_\tau$ and $\mathbf{s}^0 + \boldsymbol{\pi}_\tau$ are the same with each other. It essential to observe that what matters for the consistency of the principal's estimation is the differences of pairs of entries of $\mathbf{r}^0$, rather than its individual entries. This observation will allow as to derive the identifiability result for our estimator in the next subsection.

As highlighted before, the principal's estimator has only two sequences of inputs: $\boldsymbol{\Pi}_t = \{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_{t-1}\}$ is the sequence of the incentives chosen by the principal up to time t and $\Upsilon_t(\boldsymbol{\Pi}_t) = \{\upsilon_1(\boldsymbol{\pi}_1), \upsilon_2(\boldsymbol{\pi}_2), \ldots, \upsilon_{t-1}(\boldsymbol{\pi}_{t-1})\}$ is the sequence of the arms chosen by the agent up to time t. As we stated above, these chosen arms are based on the agent's own estimator which is trained in parallel to the principal's estimator, and hence, there is no guarantee that these arms are the true maximizers for the agent under the offered incentives. The implication of this case is that there is an additional source of uncertainty carried to the principal's estimator through the arms chosen by the agent's learning algorithm. Taking into account this uncertainty in the behaviors of the agent, we formalize the principal's estimate $\widehat{\mathbf{s}}_t^{\mathrm{pr}}\left(\Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)$ at time t for the agent's true normalized mean reward vector $\mathbf{s}^0$:

$$\begin{aligned}
\widehat{\mathbf{s}}_t^{\mathrm{pr}}\left(\Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) \in \arg\min \ & \sum_{\tau=1}^{t-1} y_\tau && (4.1)\\
\text{s.t. } \ & s_{\upsilon_\tau(\boldsymbol{\pi}_\tau)} + \pi_{\tau,\upsilon_\tau(\boldsymbol{\pi}_\tau)} + y_\tau \geq s_a + \pi_{\tau,a} \qquad \forall a \in \mathcal{A},\ \tau = 1,\dots,t-1 && (4.2)\\
& y_\tau \in \mathbb{R} \qquad \tau = 1,\dots,t-1 && (4.3)\\
& s_1 = 0,\ s_a \in \mathcal{S} \qquad \forall a \in \mathcal{A} && (4.4)
\end{aligned}$$

where y_τ's are the slack variables used to contend with the agent's unknown behavior. By introducing the loss function

$$L\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) = \sum_{\tau=1}^{t-1} \ell\left(\mathbf{s}, \upsilon_\tau(\boldsymbol{\pi}_\tau), \boldsymbol{\pi}_\tau\right) \tag{4.5}$$

where $\ell\left(\mathbf{s}, \upsilon_\tau(\boldsymbol{\pi}_\tau), \boldsymbol{\pi}_\tau\right) = \max_{a\in\mathcal{A}}\left(s_a + \pi_{\tau,a} - s_{\upsilon_\tau(\boldsymbol{\pi}_\tau)} - \pi_{\tau,\upsilon_\tau(\boldsymbol{\pi}_\tau)}\right)$, we can reformulate the linear optimization problem above as

$$\widehat{\mathbf{s}}_t^{\mathrm{pr}}\left(\Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) \in \underset{s_1=0,\ s_a\in\mathcal{S},\forall a\in\mathcal{A}}{\arg\min} L\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) \tag{4.6}$$

For notational simplicity, we use the simplified notation $\widehat{\mathbf{s}}_t^{\mathrm{pr}}$ throughout the rest of this chapter.

4.2.3 Identifiability

The first step of our finite-sample convergence analysis for the principal's estimator (4.6) is to prove that our estimator satisfies an identifiability property that ensures the loss function (4.5) is minimized uniquely by the true reward vector $\mathbf{s}^0$ (Van der Vaart 2000). The implication of this condition is that our estimator should be able to distinguish between $\mathbf{s}^0$ and an incorrect estimate $\widehat{\mathbf{s}}_t^{\mathrm{pr}}$ for a given set of incentives. As pointed out in the previous subsection, the characterization of such incentives is based on the differences of pairs of entries of $\mathbf{s}^0$.

With this observation on hand, we give our identifiability result in Proposition 4.4 that is proven using intermediate results in Propositions 4.1 – 4.3. Though these intermediate results may superficially look similar to the ones in Section 3.2.3, the results here differ in a fundamental way because we must take into account the unknown decision rule of the "imperfect-knowledge agent". Before presenting these results, we clarify that our statistical analysis employs the sets $\left(\mathcal{N}(\mathbf{s}^0,\beta), \mathcal{F}, \{\mathcal{B}(\mathbf{s}^j,d)\}_{j=1}^q\right)$ and the indices (K, K^0, b) exactly as defined in the previous chapter.

Proposition 4.1 *Suppose that $K^0 \cap K = \emptyset$ for a given vector $\mathbf{s} \in \mathcal{B}(\mathbf{s}^j, d), j \in \{1, \ldots, q\}$. Consider the incentives chosen uniformly randomly from the compact set $\mathcal{C}$, i.e., $\pi_{t,a} \sim U(\underline{C}, \overline{C}), \forall a \in \mathcal{A}$. Then, we bound the following probability conditioned on the case that the agent chooses the true maximizer arm at time $t \in \mathcal{T}$.*

$$\mathbb{P}\left(\ell\left(\mathbf{s}, \upsilon_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) \geq o \middle| \upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a \in \mathcal{A}} \left(s_a^0 + \pi_{t,a}\right)\right) \geq \left(\frac{(s_{\kappa^0}^0 - s_\kappa^0)^2 - o^2}{2(\overline{C} - \underline{C})^2}\right)\left(1 - \frac{\gamma + \beta - d}{\overline{C} - \underline{C}}\right)^2 \left(\frac{\gamma}{\overline{C} - \underline{C}}\right)^{n-2} \tag{4.7}$$

which holds for any two indices $\kappa^0 \in K^0$, $\kappa \in K$, and any constant $o \in (0, \delta)$ where $\delta := \max_{a \in \mathcal{A}} s_a^0 - \max_{a \in \mathcal{A} \setminus \{K^0\}} s_a^0$ is the difference between the largest and second largest entries of $\mathbf{s}^0$.

Proposition 4.2 *Suppose that $K^0 \cap K \neq \emptyset$, $b \notin K^0 \cap K$ for a given vector $\mathbf{s} \in \mathcal{B}(\mathbf{s}^j, d), j \in \{1, \ldots, q\}$. Consider that $\pi_{t,a} \sim U(\underline{C}, \overline{C}), \forall a \in \mathcal{A}$ and the agent chooses the true maximizer arm at time $t \in \mathcal{T}$. Then,*

$$\mathbb{P}\left(\ell\left(\mathbf{s}, \upsilon_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) \geq o \middle| \upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a \in \mathcal{A}} \left(s_a^0 + \pi_{t,a}\right)\right) \geq \frac{(\beta - o)^2}{(\overline{C} - \underline{C})^2}\left(1 - \frac{\gamma + \omega}{\overline{C} - \underline{C}}\right)^2 \left(\frac{\gamma}{\overline{C} - \underline{C}}\right)^{n-2} \tag{4.8}$$

for any constant $o \in (0, \beta)$ where $\omega = \sup_{\mathbf{s} \in \mathcal{B}(\mathbf{s}^j, d)} \max_{a \in \mathcal{A}}(|s_a^0|, |s_a|)$ is the largest absolute value observed among the entries of $\mathbf{s}^0$ and of all the vectors in $\mathcal{B}(\mathbf{s}^j, d)$.

Proposition 4.3 *Suppose that $K^0 \cap K \neq \emptyset$, $b \in K^0 \cap K$ for a given vector $\mathbf{s} \in \mathcal{B}(\mathbf{s}^j, d), j \in \{1, \ldots, q\}$. Consider that $\pi_{t,a} \sim U(\underline{C}, \overline{C}), \forall a \in \mathcal{A}$ and the agent chooses the true maximizer arm at time $t \in \mathcal{T}$. Then,*

$$\mathbb{P}\left(\ell\left(\mathbf{s}, \upsilon_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) \geq o \middle| \upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a \in \mathcal{A}} \left(s_a^0 + \pi_{t,a}\right)\right) \geq \frac{(\beta - o)^2}{(\overline{C} - \underline{C})^2}\left(1 - \frac{\gamma + \beta - d}{\overline{C} - \underline{C}}\right)\left(1 - \frac{\gamma + \omega}{\overline{C} - \underline{C}}\right)\left(\frac{\gamma}{\overline{C} - \underline{C}}\right)^{n-2} \tag{4.9}$$

for any constant $o \in (0, \beta)$.

Our goal in these propositions is to show that the principal's estimator (4.6) is able to differentiate between the true mean reward vector $\mathbf{s}^0$ and a different reward vector $\mathbf{s}$ that is at least β away from $\mathbf{s}^0$ in terms of the ℓ_∞ norm. Due to the structure of our model, we prove this property separately for three mutually exclusive cases defined based on the sets of indices K and K^0 and the index b introduced earlier. However, each of these results are proven for the event that the agent's algorithm picks the true maximizer arm in response to the given random incentives. As we aim to offer a generic approach without limiting the type of the algorithm used by the agent, we need to consider a mild assumption on the learning behavior of the agent. In particular, we need to assume that after a transient period of learning, the agent's algorithm will choose an incorrect (i.e., different than the true maximizer) arm at a decreasing rate as the game move forwards over the considered time horizon. We now specify this rate in the following statement.

Assumption 4.2 *Let* $p_t := \mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t) \neq \arg\max_{a\in\mathcal{A}}\left(s_a^0 + \pi_{t,a}\right)\right)$ *be the probability that the agent does not select the true reward-maximizer arm at time* t*. There exists a constant* $k \geq 1$ *such that* $p_t \leq k\dfrac{\sqrt{\log 2t}}{\sqrt{t}}$ *at any time step* $t \in [\widetilde{k}, T]$ *where* $\widetilde{k} \geq 2$ *is the minimum value satisfying* $k\sqrt{\log 2\widetilde{k}} < \sqrt{\widetilde{k}}$*.*

We later provide a validation of this assumption by showing that it is satisfied when the agent uses a classical MAB algorithm and the principal uses the proposed data-driven incentive policy presented in the next section. Now, we unite the results in Propositions 4.1 – 4.3 and obtain our final identifiability statement.

Proposition 4.4 *At time* $t \in [\widetilde{k}, T]$*, suppose that* $\pi_{t,a} \sim U(\underline{C}, \overline{C}), \forall a \in \mathcal{A}$*. Then, for any normalized reward vector* $\mathbf{s} \in \mathcal{F}$*, we have*

$$\mathbb{P}\left(\ell\left(\mathbf{s}, \upsilon_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) \geq o\right) \geq \alpha(\beta - o)^2 \left(1 - k\frac{\sqrt{\log 2t}}{\sqrt{t}}\right) \tag{4.10}$$

for any constant $o \in (0, \beta)$ *and* $\alpha > 0$*.*

This result proves that our estimator (4.6) can identify and refute an incorrect estimate $\mathbf{s} \in \mathcal{F}$ with a strictly positive probability whose rate is proportional to the square of the estimation error β. Therefore, as the game progresses and the agent learns its own reward model, the principal is also capable of effectively learning $\mathbf{s}^0$ from the arms chosen by the agent in response to the random incentives offered to explore the agent's bandit model.

4.2.4 Finite-Sample Concentration Bound

We derive our statistical consistency result for the principal's estimator (4.6) in several intermediate steps. In these steps, we prove finite-sample concentration inequalities with respect to the behavior of the loss function (4.5) evaluated at any incorrect reward vector $\mathbf{s} \in \mathcal{F}$ and the loss function evaluated at the true reward vector $\mathbf{s}^0$.

Proposition 4.5 *Let* $\eta(\widetilde{k}, t)$ *be the number of time steps that the principal chooses each incentive* $\pi_{t,a}$ *uniformly randomly from the compact set* $\mathcal{C}$ *within the time interval* $[\widetilde{k}, t-1]$*, i.e.,*

$$\eta(\widetilde{k}, t) = \left|\Lambda(\widetilde{k}, t)\right| \quad \text{where} \quad \Lambda(\widetilde{k}, t) := \{\tau : \widetilde{k} \leq \tau \leq t-1,\ \pi_{\tau,a} \sim U(\underline{C}, \overline{C}), \forall a \in \mathcal{A}\} \tag{4.11}$$

For the given sequences of incentives $\mathbf{\Pi}_t$ *and chosen arms* $\Upsilon_t(\mathbf{\Pi}_t)$*, the total estimation loss over these time steps is defined as*

$$L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) = \sum_{\tau\in\Lambda(\widetilde{k},t)} \ell\left(\mathbf{s}, \upsilon_\tau(\boldsymbol{\pi}_\tau), \boldsymbol{\pi}_\tau\right) \tag{4.12}$$

and satisfies

$$\mathbb{P}\left(\left|L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) - \mathbb{E}L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)\right| \geq \nu\right) \\ \leq 2\exp\left(-\frac{2\nu^2}{(\eta(\widetilde{k},t)-1)n\left(6R_{\max} - 6R_{\min} + 2\gamma\right)^2}\right) \tag{4.13}$$

for any constant $\nu > 0$ *and mean reward vector* $\mathbf{s} \in \mathcal{S}$*.*

The proof of this result follows by using the bounded differences inequality (i.e., McDiarmid's inequality) (Boucheron et al. 2013) and the definition of our single-step loss function $\ell\left(\mathbf{s}, \upsilon_\tau(\boldsymbol{\pi}_\tau), \boldsymbol{\pi}_\tau\right)$.

Proposition 4.6 *For the given sequences of incentives* $\mathbf{\Pi}_t$ *and chosen arms* $\Upsilon_t(\mathbf{\Pi}_t)$*, the concentration of the loss function* $L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)$ *in (4.12) within the the compact set* $\mathcal{F} = \{\mathbf{s} \in \mathcal{S}^n : \|\mathbf{s} - \mathbf{s}^0\|_\infty > \beta\}$ *is given as*

$$\mathbb{P}\left(\sup_{\mathbf{s}\in\mathcal{F}}\left|L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) - \mathbb{E}L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)\right| \geq \nu\right) \\ \leq 2\exp\left(-\frac{2\nu^2}{(\eta(\widetilde{k},t)-1)n(6R_{\max} - 6R_{\min} + 2\gamma)^2} - \log\beta + n\log(R_{\max} - R_{\min})\right) \tag{4.14}$$

for any constant $\nu > 0$*.*

This proposition is proven by using the result of Proposition 4.5 and bounding the covering number q for $\mathcal{F}$ by volume ratios. We now compute a lower bound for the minimum possible expected loss over $\Lambda(\widetilde{k},t)$ achieved within $\mathcal{F}$, i.e., the set of feasible reward vectors that are at least β away from $\mathbf{s}^0$.

Lemma 4.1 *We define the minimizer of the loss (4.12) within the compact set* $\mathcal{F}$ *as* $\mathbf{s}_t^{\mathcal{F}} := \arg\inf_{\mathbf{s}\in\mathcal{F}} L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)$*. Then, given the sequences of incentives* $\mathbf{\Pi}_t$ *and chosen arms* $\Upsilon_t(\mathbf{\Pi}_t)$*, we have*

$$\mathbb{E}L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}_t^{\mathcal{F}}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) \geq \frac{4\alpha\left(1 - k\sqrt{\log 2\widetilde{k}}/\sqrt{\widetilde{k}}\right)^2}{27}\beta^3\mathbb{E}\eta(\widetilde{k},t) \tag{4.15}$$

for any $t \in [\widetilde{k}, T]$*.*

We prove this result by using Assumption 4.2 and our identifiability result in Proposition 4.4. We continue by deriving an upper bound for the expected total loss up to time t evaluated at the true mean reward vector $\mathbf{s}^0$.

Lemma 4.2 *The expectation of the total loss of the principal's estimator (4.6) computed for the true mean reward vector* $\mathbf{s}^0$ *is bounded by*

$$\mathbb{E}L\left(\mathbf{s}^0, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) \leq 3k\left(3(R_{\max} - R_{\min}) + \gamma\right)\sqrt{t\log(2t)} \tag{4.16}$$

In the last lemma of this section, we present the concentration inequality for $\mathbf{s}^0$.

Lemma 4.3 *For the given sequences of incentives* $\mathbf{\Pi}_t$ *and chosen arms* $\Upsilon_t(\mathbf{\Pi}_t)$*, the concentration of the total loss of the principal's estimator evaluated at* $\mathbf{s}^0$ *is given as*

$$\mathbb{P}\left(L\left(\mathbf{s}^0, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) - \mathbb{E}L\left(\mathbf{s}^0, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) \geq \nu\right) \leq \exp\left(-\frac{2\nu^2}{(t-1)\left(3R_{\max} - 3R_{\min} + \gamma\right)^2}\right) \tag{4.17}$$

for any $\nu > 0$.

This lemma is proven by first observing that the loss for $\mathbf{s}^0$ at any time step becomes 0 when the agent selects the true maximizer arm, and thus, it suffices to only consider the time steps where the agent's algorithm makes an inaccurate decision. Then, the proof follows by using Hoeffding's Inequality (Boucheron et al. 2013).

The last step is to combine Proposition 4.6 and Lemmas 4.1-4.3 to obtain the final finite-sample concentration bound for the principal's estimator with respect to the total loss function (4.5).

Theorem 4.1 *We introduce the quantity*

$$\lambda_t = \frac{4\alpha\left(1 - k\sqrt{\log 2\widetilde{k}}/\sqrt{\widetilde{k}}\right)^2}{27}\beta^3\mathbb{E}\eta(\widetilde{k}, t) - 3k\left(3(R_{\max} - R_{\min}) + \gamma\right)\sqrt{t\log(2t)}. \tag{4.18}$$

Given the sequences of incentives $\mathbf{\Pi}_t$ *and agent's choices* $\Upsilon_t(\mathbf{\Pi}_t)$*, we show that*

$$\begin{aligned}\mathbb{P}&\left(\inf_{\mathbf{s}\in\mathcal{F}} L\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) \leq L\left(\mathbf{s}^0, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)\right) \\ &\leq 2\exp\left(-\frac{2\lambda_t^2}{(t-1)16n(6R_{\max} - 6R_{\min} + 2\gamma)^2} - \log\beta + n\log(2R_{\max} - 2R_{\min})\right)\end{aligned} \tag{4.19}$$

for $t \in [\widetilde{k}, T]$.

Alternatively, we can reinterpret Theorem 4.1 and obtain a concentration with respect to the ℓ_∞ distance between our estimates $\widehat{\mathbf{s}}_t$ and the true reward vector $\mathbf{s}^0$.

Corollary 4.1 (Finite-Sample Concentration Bound) *The principal's estimator in (4.6) satisfies*

$$\begin{aligned}&\mathbb{P}\left(\|\mathbf{s}^0 - \widehat{\mathbf{s}}_t^{\text{pr}}\|_\infty > \beta\right)\\ &\quad \le 2\exp\left(-\frac{2\lambda_t^2}{(t-1)16n(6R_{\max} - 6R_{\min} + 2\gamma)^2} - \log\beta + n\log(2R_{\max} - 2R_{\min})\right) \qquad (4.20)\end{aligned}$$

for any $\beta > 0$ and $t \in [\widetilde{k}, T]$ where λ_t is as defined in (4.18).

It is important to note that in this bound the learning rate of the principal directly depends on $\eta(\widetilde{k}, t)$, which is the number of periods at which they offer random incentives from an offset time $\widetilde{k}$ and beyond to explore the arm space of the agent. More importantly, the existence of this offset point shows that the rate that the principal explores the agent's bandit model must be greater than the agent's exploration rate. Recall that we consider a selfish agent who is only interested in learning their own reward model whereas the principal needs to learn both their own rewards and the agent's rewards to be able to design effective incentives. Because the principal's learning process is fed by the agent's decisions over the considered repeated game, the principal's learning will be possible if and only if the agent's learning is successful. To restate, the principal has to wait for a sufficient time after which the agent will start playing the right arm the most fraction of the time. That is why our finite-sample concentration bound holds for the time periods $t \ge \widetilde{k}$ where $\widetilde{k}$ is an increasing function of the agent's parameter k as given in Assumption 4.2. (We note that the offset point of time $\widetilde{k}$ can be computed numerically when k is known.) The implication of this result is that for higher k values, it will take a longer time for the agent to start playing correctly in a consistent way and for the principal's estimator to converge to $\mathbf{s}^0$. We conclude this section by emphasizing this discussion in the following remark.

Remark 4.1 *The rationale behind the condition $t \ge \widetilde{k}$ in Corollary 4.1 relies on an essential dynamic of the repeated adverse selection games we study in this chapter. It reflects the fact that the finite-sample consistency of the principal's estimator is attainable only after a transient learning period of the agent's algorithm.*

This fundamental observation once again brings us back to the main theoretical challenge we highlighted in the introduction of the chapter. The adverse impact of the agent's learning process on the principal's inferences immensely accumulates the costs of the principal, as we will show in our regret analysis in the next section.

4.3 The Robust Data-Driven Incentive Policy

Our novel and consistent estimator allows the principal to design an adaptive and easy to compute menu of incentives for the agent's bandit arms. We propose a MAB framework for the principal within which we unify the principal's estimator and the data-driven incentives we design in this section.

Before we present the details of the proposed learning framework, we recall that the principal's problem involves estimating their own reward expectations $\boldsymbol{\theta}^0$ in addition to the agent's rewards. However, the former estimation is a more manageable problem than the latter because the principal can fully observe their own reward outcomes $\mu_{t,\upsilon_t(\boldsymbol{\pi}_t)}$ realized through the arms chosen by the agent.

Assumption 4.3 *For each agent arm $a \in \mathcal{A}$, the principal observes independent reward realizations $\mu_{t,a}, t \in \mathcal{T}$ from a sub-Gaussian distribution $\mathbb{F}^{\mathrm{pr}}_{\theta^0_a,a}$ for all $\theta^0_a \in \Theta$. Similarly, the agent's rewards $\rho_{t,a}, t \in \mathcal{T}$ are independent from each other and follow a sub-Gaussian distribution for all $r^0_a \in \mathcal{R}$.*

This assumption about the reward distribution families of the principal and the agent is a mild and common condition encountered in many MAB models. Under this assumption, we define the quantity $T(a,t) = |\{\tau : 1 \leq \tau \leq t-1, \upsilon_\tau(\boldsymbol{\pi}_\tau) = a\}|$ as the number of time points when the agent selects arm a up to time t. Then, we consider the sample mean of the principal's reward outcomes up to time t as the principal's estimate for $\theta^0_a, \forall a \in \mathcal{A}$.

$$\widehat{\theta}_{t,a} = \frac{1}{T(a,t)} \sum_{\tau=1}^{t-1} \mu_{\tau,a} \mathbb{1}\left(\upsilon_\tau(\boldsymbol{\pi}_\tau) = a\right) \tag{4.21}$$

We note that $\widehat{\theta}_{t,a}$ is an unbiased estimator and is the same as the maximum likelihood estimator for many common exponential family distributions where the sufficient statistic is equal to the random variable itself, including the Gaussian, Bernoulli, Poisson, and multinomial distributions.

4.3.1 Principal's ϵ-Greedy Policy

We next present a sequential learning framework that utilizes the principal's estimators $\widehat{\mathbf{s}}^{\mathrm{pr}}_t$ and $\widehat{\boldsymbol{\theta}}_t$ to compute an adaptive and efficient incentive policy. Keeping in mind practicality, we develop an ϵ-greedy algorithm for which we provide pseudocode in Algorithm 3.

To initiate the principal's and agent's learning processes, we consider an initialization period over the first $n = |\mathcal{A}|$ time periods throughout which the agent is induced to pick each of the n arms once to observe a reward realization and compute a starting estimate of the associated reward expectation. To achieve this, the principal offers the maximum possible incentive $(\overline{C})$ for the desired arm in each step that would be sufficient to make that arm reward-maximizer for the agent by Assumption 4.1.

At each time step $t \in [n+1, \dots, T]$ after the initialization period, the principal first updates their estimate for their own mean reward associated with the most recently chosen arm $\theta^0_{v_{t-1}(\boldsymbol{\pi}_{t-1})}$. Then, the principal samples a Bernoulli random variable x^{pr}_t with success probability ϵ^{pr}_t which corresponds to the principal's exploration probability. In accordance with our observation highlighted in Remark 4.1 following our statistical analysis in Section 4.2.4, here we clarify that the principal's rate of the exploration is designed to be greater than the agent's learning rate in accordance with Assumption 4.2.

If $x^{\text{pr}}_t = 1$, then the principal performs an exploration step where they offer incentives $\boldsymbol{\pi}_t = (\pi_{t,a})_{a\in\mathcal{A}}$ that are uniformly randomly selected from the feasible range of incentives $\mathcal{C} = [\underline{C}, \overline{C}]$. Otherwise, for $x^{\text{pr}}_t = 0$ the principal prefers a less-risky, greedy exploitation play to maximize their expected net reward in that period. They update their estimate $\widehat{\mathbf{s}}^{\text{pr}}_t$ for the agent's mean rewards by solving (4.6) and use it to compute the incentives $\mathbf{c}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}^{\text{pr}}_t)$ maximizing their estimated expected net reward at time t. Recall that the principal's expected net reward in a period is equal to their fixed expected reward for the chosen arm minus the sum of incentives offered for each bandit arm. Therefore, our goal in an exploitation period is to compute the minimum vector of incentives that will steer the agent's choice into the desired arm (that is estimated to yield the highest expected net reward to the principal) in that period while inducing the agent's incentive compatibility. Using the most recent estimates $\widehat{\boldsymbol{\theta}}_t$ and $\widehat{\mathbf{s}}^{\text{pr}}_t$, for each arm $j \in \mathcal{A}$, we compute the minimum amount of incentives $(\widetilde{c}^j_a)_{a\in\mathcal{A}}$ that make j the maximizer arm for the agent and the corresponding expected net reward $\widetilde{V}(j, \widehat{\mathbf{s}}^{\text{pr}}_t; \widehat{\boldsymbol{\theta}}_t)$ of the principal in case j is actually chosen by the agent in response to these incentives. Further, to contend with the principal's estimation error in $\widehat{\mathbf{s}}^{\text{pr}}_t$ and the unknown behavior of the learning agent, we add a buffer amount to the computed minimum incentives and obtain

$$\widetilde{c}^j_j = \left(\max_{a\in\mathcal{A}} \widehat{s}^{\text{pr}}_{t,a}\right) - \widehat{s}^{\text{pr}}_{t,j} + 2\beta_t \tag{4.22}$$

$$\widetilde{c}^j_a = 0, \quad \forall a \in \mathcal{A},\ a \neq j \tag{4.23}$$

$$\widetilde{V}(j, \widehat{\mathbf{s}}^{\text{pr}}_t; \widehat{\boldsymbol{\theta}}_t) = \widehat{\theta}_{t,j} - \sum_{a\in\mathcal{A}} \widetilde{c}^j_a = \widehat{\theta}_{t,j} - \left(\max_{a\in\mathcal{A}} \widehat{s}^{\text{pr}}_{t,a}\right) + \widehat{s}^{\text{pr}}_{t,j} - 2\beta_t \tag{4.24}$$

where $\beta_t = B\frac{\sqrt{\log 2t}}{t^{w/3}}$ with $B = \frac{3k(3(R_{\max}-R_{\min})+\gamma)^n \sqrt[6]{32n}}{1-k\sqrt{\log 2\widetilde{k}}/\sqrt{\widetilde{k}}}$. After computing these values for each arm $j \in \mathcal{A}$, the principal chooses the set of incentives corresponding to the arm j^*_t that is estimated to yield the highest $\widetilde{V}(j, \widehat{\mathbf{s}}^{\text{pr}}_t; \widehat{\boldsymbol{\theta}}_t)$ to the principal. To reiterate, the exploitation incentives are purposefully designed to make the desired arm j^*_t reward-maximizer for the agent with high probability by leveraging the statistically consistent estimator proposed in Section 4.2. Lastly, we clarify that the principal only observes the arm $v_t(\boldsymbol{\pi}_t)$ chosen by the agent in response to the offered incentives and their own net reward realization $\mu_{t,v_t(\boldsymbol{\pi}_t)} - \sum_{a\in\mathcal{A}} \pi_{t,a}$ at the end of each period t. The total reward collected by the agent $\rho_{t,v_t(\boldsymbol{\pi}_t)} + \pi_{t,v_t(\boldsymbol{\pi}_t)}$ remains as private knowledge in the considered hidden rewards setting.

Remark 4.2 *Computation of the exploitation incentives through lines 16–19 in Algorithm 3 includes arithmetic operations with a linear computational complexity $O(n)$ in terms of the dimension of the agent's bandit model $n = |\mathcal{A}|$.*

Algorithm 3 Principal's ϵ-Greedy Algorithm

```
 1: Set: $m^{\mathrm{pr}} \geq 1$ and $w \in (0, 1/4)$
 2: for $t \in [1, \dots, n]$ do
 3:     Set: $\boldsymbol{\pi}_t = (\pi_{t,a})_{a \in \mathcal{A}}$ where $\pi_{t,a} = \overline{C}$ for $a = t$ and $\pi_{t,a} = 0$ for all $a \neq t$
 4:     Observe: $\upsilon_t(\boldsymbol{\pi}_t) = t$ and $\mu_{t,\upsilon_t(\boldsymbol{\pi}_t)}$
 5:     if $t \geq 2$ then $\widehat{\theta}_{t,\upsilon_{t-1}(\boldsymbol{\pi}_{t-1})} = \mu_{t-1,\upsilon_{t-1}(\boldsymbol{\pi}_{t-1})}$
 6: for $t \in [n+1, \dots, T]$ do
 7:     Compute: $\widehat{\theta}_{t,\upsilon_{t-1}(\boldsymbol{\pi}_{t-1})} \in \frac{1}{T(\upsilon_{t-1}(\boldsymbol{\pi}_{t-1}), t)} \sum_{\tau=1}^{t-1} \mu_{\tau,\upsilon_{t-1}(\boldsymbol{\pi}_{t-1})} \mathbb{1}(\upsilon_\tau(\boldsymbol{\pi}_\tau) = \upsilon_{t-1}(\boldsymbol{\pi}_{t-1}))$
 8:     Set: $\epsilon_t^{\mathrm{pr}} = \min\left\{1, \frac{m^{\mathrm{pr}}}{t^{1/2-w}}\right\}$
 9:     Sample: $x_t^{\mathrm{pr}} \sim \mathrm{Bernoulli}(\epsilon_t^{\mathrm{pr}})$
10:     if $x_t^{\mathrm{pr}} = 1$ then
11:         Sample: $\pi_{t,a} \sim \mathcal{U}\left(\underline{C}, \overline{C}\right)$ for all $a \in \mathcal{A}$
12:         Set: $\boldsymbol{\pi}_t = (\pi_{t,a})_{a \in \mathcal{A}}$
13:     else
14:         Compute: $\beta_t = B \frac{\sqrt{\log 2t}}{t^{w/3}}$
15:         Compute: $\widehat{\mathbf{s}}_t^{\mathrm{pr}} \in \arg\min \left\{ L\left(\mathbf{s}, \Upsilon_t(\boldsymbol{\Pi}_t), \boldsymbol{\Pi}_t\right) \middle| s_1 = 0,\ s_a \in \mathcal{S}\ \forall a \in \mathcal{A} \right\}$
16:         for $j \in \mathcal{A}$ do
17:             Compute: $\widetilde{V}(j, \widehat{\mathbf{s}}_t^{\mathrm{pr}}; \widehat{\boldsymbol{\theta}}_t) = \widehat{\theta}_{t,j} - \left(\max_{a \in \mathcal{A}} \widehat{s}_{t,a}^{\mathrm{pr}}\right) + \widehat{s}_{t,j}^{\mathrm{pr}} - 2\beta_t$
18:         Compute: $j_t^* = \arg\max_{j \in \mathcal{A}} \widetilde{V}(j, \widehat{\mathbf{s}}_t^{\mathrm{pr}}; \widehat{\boldsymbol{\theta}}_t)$
19:         Set: $c_{j_t^*}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) = \left(\max_{a \in \mathcal{A}} \widehat{s}_{t,a}^{\mathrm{pr}}\right) - \widehat{s}_{t,j_t^*}^{\mathrm{pr}} + 2\beta_t$ and $c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) = 0$ for all $a \neq j_t^*$
20:         Set: $\boldsymbol{\pi}_t = (c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}))_{a \in \mathcal{A}}$
21:     Observe: $\upsilon_t(\boldsymbol{\pi}_t)$ and $\mu_{t,\upsilon_t(\boldsymbol{\pi}_t)}$
```

4.3.2 Principal's Finite-Sample Regret

To show the efficiency and effectiveness of the designed incentive mechanism, we conduct a rigorous regret analysis where regret is defined in terms of the expected net reward of the principal accumulated over a finite time horizon.

4.3.2.1 Oracle Incentive Policy

As a benchmark to the proposed incentive policy, we define an oracle incentive policy that maximizes the principal's expected net reward under full knowledge of all reward expectations $\mathbf{s}^0$ and $\boldsymbol{\theta}^0$. In other words, the oracle policy computes the exploitation incentives by using the true mean reward values and assuming a perfect-knowledge agent, as used in Chapter 3. Similar to the procedure described above, for each arm $j \in \mathcal{A}$, we solve for the minimum incentives $(\widetilde{c}^j_a)_{a\in\mathcal{A}}$ required to make $j \in \mathcal{A}$ maximizer of the agent's total reward and compute the corresponding expected net reward of the principal $\widetilde{V}(j, \mathbf{s}^0; \boldsymbol{\theta}^0)$ as follows.

$$\widetilde{c}^j_j = \left(\max_{a\in\mathcal{A}} s^0_a\right) - s^0_j \tag{4.25}$$

$$\widetilde{c}^j_a = 0, \quad \forall a \in \mathcal{A},\ a \neq j \tag{4.26}$$

$$\widetilde{V}(j, \mathbf{s}^0; \boldsymbol{\theta}^0) = \theta^0_j - \sum_{a\in\mathcal{A}} \widetilde{c}^j_a = \theta^0_j - \left(\max_{a\in\mathcal{A}} s^0_a\right) + s^0_j \tag{4.27}$$

Then, the oracle policy chooses the set of incentives corresponding to desired arm $j^{*,0} := \arg\max_{j\in\mathcal{A}} \widetilde{V}(j, \mathbf{s}^0; \boldsymbol{\theta}^0)$. The computed oracle incentives are denoted by $\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0)$ and given as

$$c_{j^{*,0}}(\boldsymbol{\theta}^0, \mathbf{s}^0) = \left(\max_{a\in\mathcal{A}} s^0_a\right) - s^0_{j^{*,0}} + \varsigma \tag{4.28}$$

$$c_a(\boldsymbol{\theta}^0, \mathbf{s}^0) = 0, \quad \forall a \neq j^{*,0} \tag{4.29}$$

where $\varsigma > 0$ is an arbitrarily small constant that helps avoiding the occurrence of multiple maximizer arms for the agent. Because the oracle policy fully knows the agent's expected rewards and hence the true reward-maximizer arm, the main difference between oracle incentives and the exploitation incentives computed by Algorithm 3 is the absence of the buffer amount β_t. Further, we observe that the desired arm $j^{*,0}$ of the oracle policy is equal to the agent's true maximizer arm in response to the oracle incentives, i.e., $j^{*,0} = \upsilon(\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0)) = \arg\max_{a\in\mathcal{A}} s^0_a + c_a(\boldsymbol{\theta}^0, \mathbf{s}^0)$.

4.3.2.2 Regret Bound

At any time period $t \in \mathcal{T}$, the expected net reward of the principal under the oracle incentive policy is given as

$$V(\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0); \boldsymbol{\theta}^0) = \theta^0_{j^{*,0}} - \max_{a\in\mathcal{A}} s^0_a + s^0_{j^{*,0}} - \varsigma \tag{4.30}$$

and under the proposed incentive policy generated by Algorithm 3 is given as

$$V_t(\boldsymbol{\pi}_t; \boldsymbol{\theta}^0) = \theta^0_{\upsilon_t(\boldsymbol{\pi}_t)} - \sum_{a\in\mathcal{A}} \pi_{t,a}. \tag{4.31}$$

Accordingly, we define the regret of the proposed ϵ-greedy incentive policy $\Pi_{\epsilon,T} = \{\boldsymbol{\pi}_t\}_{t\in\mathcal{T}}$ with respect to the oracle incentive policy over a finite time horizon $\mathcal{T}$ as

$$\text{Regret}\left(\Pi_{\epsilon,T}\right) = \sum_{t\in\mathcal{T}} V(\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0); \boldsymbol{\theta}^0) - V_t(\boldsymbol{\pi}_t; \boldsymbol{\theta}^0). \tag{4.32}$$

Before we present our upper bound for this regret notion, we first prove a useful theoretical result showing a probability bound on the accuracy of the arm j_t^* that is estimated by Algorithm 3 to yield the maximum expected net reward to the principal at time t.

Proposition 4.7 *The probability that the estimated best arm (j_t^*) for the principal in an exploitation step of Algorithm 3 (see lines 16–19) is the same as the true best arm ($j^{*,0}$) chosen by the oracle policy is bounded from above by*

$$\mathbb{P}\left(j_t^* \neq j^{*,0}\right) \leq \frac{n}{t} + \frac{n^{5/6}2^{n+1}}{3^{n+1}k\sqrt[6]{32}}\frac{1}{\sqrt{t\log 2t}} \tag{4.33}$$

for $t \in [\widetilde{k}, T]$ where $\widetilde{k}$ is as introduced in Assumption 4.2.

This probability bound on the accuracy of our estimate for the principal's desired arm at each period has two main components corresponding to the estimation gap of $\widehat{\mathbf{s}}_t^{\text{pr}}$ and the error in $\widehat{\boldsymbol{\theta}}_t$. Therefore, the proof follows by mainly using our finite-sample concentration bound for $\widehat{\mathbf{s}}_t^{\text{pr}}$ given in Corollary 4.1 and Hoeffding's Inequality (Boucheron et al. 2013) for the consistency of $\widehat{\boldsymbol{\theta}}_t$.

Lastly, we combine this result with the results of our statistical analysis and prove a rigorous regret bound for the proposed incentive design framework of the principal.

Theorem 4.2 (Finite-Sample Regret Bound) *The finite-sample regret bound of a policy $\Pi_{\epsilon,T}$ computed by the Principal's ϵ-Greedy Algorithm (3) is proven to be*

$$\begin{aligned}\text{Regret}\left(\Pi_{\epsilon,T}\right) \leq\ & \frac{12B}{3-w}T^{1-w/3}\sqrt{\log 2T} + 2k\Theta^{\max}\sqrt{T\log 2T} + n^2\left(\overline{C} - \underline{C} + \Theta^{\max}\right)\log T \\ & + \Theta^{\max}\widetilde{k} + \frac{2^{n+1}\left(\Theta^{\max}(2n^{11/6} + 1/\sqrt[6]{n}) + n^{5/6}(\overline{C} - \underline{C})\left(1+2n\right)\right)}{3^{n+1}k\sqrt[6]{32}}\sqrt{T} \\ & + m^{\text{pr}}\left(n(\overline{C} - \underline{C}) + \Theta^{\max}\right)\left(\frac{2}{2w+1}T^{w+1/2} + \frac{2w-1}{2w+1}\right)\end{aligned} \tag{4.34}$$

where $\Theta^{\max}$ is defined as the upper bound on θ_a^0's.

Remark 4.3 *The finite-sample regret bound proved for the proposed data-driven and adaptive incentive policy implies an asymptotic regret bound of order $O\left(T^{11/12+\sigma}\sqrt{\log T}\right)$ where σ can be made arbitrarily close to 0.*

We emphasize that the principal's regret also comprises the agent's regret through the uncertainty in the arms chosen over the course of the repeated *hidden rewards* game. Our regret bound reflects the substantial complexity and adverse impact resulting from the two parallel learning algorithms that are dynamically interacting with each other throughout the considered time horizon. For this challenging setting, our generic and practically relevant incentive mechanism is able to achieve a sublinear regret performance, that remain robust without restricting the type of the learning algorithm used by the incentivized agent.

4.4 The Learning Agent's Behavior

Although our goal in this chapter is to mainly address the estimation and incentive design problems faced by the principal in the considered repeated hidden rewards game, this section provides a further theoretical analysis of the agent's learning behavior and a discussion on the information rent gained by the self-interested agent.

4.4.1 Validation of Assumption 4.2

As discussed earlier and highlighted in Remark 4.1, convergence of the principal's learning policy is only attainable after convergence of the agent's learning policy. Therefore, our theoretical analysis for statistical consistency of the proposed estimator in Section 4.2 and regret bound of the proposed incentive policy in Section 4.3 needed to assume a probability bound on the learning behavior of the agent. We specified this bound in Assumption 4.2, and here we show the mildness of this assumption by proving that it is satisfied by a naive ϵ-greedy algorithm that could be used by the agent to make their decisions throughout the sequential game.

We consider the ϵ-greedy algorithm presented in Algorithm 4. The first $n = |\mathcal{A}|$ time steps constitute the initialization period where the agent selects each of their n arms to obtain a random reward realization and have an initial estimate for their associated mean reward. Because the agent can fully observe their own rewards, we consider the sample mean of their random reward outcomes as their estimate $\widehat{\boldsymbol{s}}_t^{\mathrm{ag}}$ for the mean rewards $\mathbf{s}^0$. We note that this corresponds to an unbiased estimator under Assumption 4.3.

After that period, at each time step $t \in [n+1, \ldots, T]$, the agent first observes the menu of incentives $\boldsymbol{\pi}_t$ offered by the principal and updates their estimate for their own mean reward associated with the most recently chosen arm $\widehat{s}_{t,\upsilon_{t-1}(\boldsymbol{\pi}_{t-1})}^{\mathrm{ag}}$. Then, the algorithm samples a Bernoulli random variable x_t^{ag} based on the agent's exploration probability ϵ_t^{ag}. If $x_t^{\mathrm{ag}} = 1$, the agent explores their MAB model by randomly selecting an arm for the current time step. Otherwise, the agent performs a greedy exploitation by picking the arm that maximizes their estimated expected reward plus the offered incentive. At the end of each play, the self-interested agent only observes the stochastic reward they collect from the chosen arm.

Algorithm 4 Agent's ϵ-Greedy Algorithm

```
Set: $m^{\mathrm{ag}} \geq 1$
for $t \in [1, \dots, n]$ do
    Observe: $\boldsymbol{\pi}_t$
    Set: $\upsilon_t(\boldsymbol{\pi}_t) = t$
    Observe: $\rho_{t,\upsilon_t(\boldsymbol{\pi}_t)}$
    if $t \geq 2$ then $\widehat{s}^{\mathrm{ag}}_{t,\upsilon_{t-1}(\boldsymbol{\pi}_{t-1})} = \rho_{t-1,\upsilon_{t-1}(\boldsymbol{\pi}_{t-1})}$
for $t \in [n+1, \dots, T]$ do
    Observe: $\boldsymbol{\pi}_t$
    Compute: $\widehat{s}^{\mathrm{ag}}_{t,\upsilon_{t-1}(\boldsymbol{\pi}_{t-1})} \in \frac{1}{T(\upsilon_{t-1}(\boldsymbol{\pi}_{t-1}), t)} \sum_{\tau=1}^{t-1} \rho_{\tau,\upsilon_{t-1}(\boldsymbol{\pi}_{t-1})} \mathbb{1}(\upsilon_\tau(\boldsymbol{\pi}_\tau) = \upsilon_{t-1}(\boldsymbol{\pi}_{t-1}))$
    Set: $\epsilon^{\mathrm{ag}}_t = \min\left\{1, \frac{m^{\mathrm{ag}}}{\sqrt{t}}\right\}$
    Sample: $x^{\mathrm{ag}}_t \sim \mathrm{Bernoulli}(\epsilon^{\mathrm{ag}}_t)$
    if $x^{\mathrm{ag}}_t = 1$ then Sample: $\upsilon_t(\boldsymbol{\pi}_t) \in \mathcal{A}$
    else Set: $\upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a \in \mathcal{A}} \widehat{s}^{\mathrm{ag}}_{t,a} + \pi_{t,a}$
    Observe: $\rho_{t,\upsilon_t(\boldsymbol{\pi}_t)}$
```

We show that Algorithm 4 is consistent with Assumption 4.2 by first proving two useful lemmas and then combining them in Proposition 4.8. For notational convenience, we let $\mathcal{T}^{\mathrm{pr-xplore}} \in \mathcal{T}$ and $\mathcal{T}^{\mathrm{pr-xploit}} \in \mathcal{T}$ be the set of random time steps that the principal's algorithm (3) performs exploration (lines 11–12) and exploitation (lines 14–20), respectively. Similarly, we define $\mathcal{T}^{\mathrm{ag-xplore}}, \mathcal{T}^{\mathrm{ag-xploit}} \in \mathcal{T}$ as the set of random time steps that the agent's algorithm (4) performs exploration and exploitation, respectively.

Lemma 4.4 *Consider a time step $t \in [\widetilde{k}, T]$ where the principal offers the exploitation incentives computed according to Algorithm 3 and the agent performs a greedy exploitation by picking the reward-maximizer arm according to Algorithm 4. Then, the probability that the agent does not select the true reward-maximizer arm at time t is bounded by*

$$\begin{aligned}
&\mathbb{P}\Big(\upsilon_t(\boldsymbol{\pi}_t) \neq \arg\max_{a \in \mathcal{A}} s^0_a + \pi_{t,a} \Big| t \in \mathcal{T}^{\mathrm{ag-xploit}} \cap \mathcal{T}^{\mathrm{pr-xploit}}\Big) \\
&\quad \leq \frac{2n^3 \left(\exp\left(\frac{2(m^{\mathrm{ag}})^2 B^2}{4n(R_{\max} - R_{\min})^2}\right) + 1\right)}{t-1} \\
&\qquad + 8n^2 \exp\left(-\frac{2\lambda_t^2}{(t-1)16n(6R_{\max} - 6R_{\min} + 2\gamma)^2} - \log\frac{\beta_t}{2} + n\log(2R_{\max} - 2R_{\min})\right)
\end{aligned} \tag{4.35}$$

where λ_t is as defined in (4.18) and β_t is as specified in (14).

Lemma 4.5 *Consider a time step $t \in \mathcal{T}$ where the principal offers random exploration incentives according to Algorithm 3 and the agent performs a greedy exploitation by picking the reward-maximizer arm according to Algorithm 4. Then, the probability that the agent does not select the true reward-maximizer arm at time t is bounded by*

$$\mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t) \neq \arg\max_{a\in\mathcal{A}} s_a^0 + \pi_{t,a} \middle| t \in \mathcal{T}^{\text{ag}-\text{xploit}} \cap \mathcal{T}^{\text{pr}-\text{xplore}}\right) \\ \leq \frac{4n^2\sqrt{n}\sqrt{\log 2t}}{\sqrt{m^{\text{ag}}}\sqrt[4]{t}} + \frac{2n^2(\exp(2m^{\text{ag}})+1)}{t-1} \quad (4.36)$$

In our analysis for the agent's learning process, we need to take into account the externality imposed by the principal's incentives in each period which requires considering the principal's exploration/exploitation plays separately. Lemmas 4.4 and 4.5 correspond to our results for each of these cases, and their detailed proofs are provided in Appendix C.1.3. Next, we combine these intermediate results and obtain our final proposition.

Proposition 4.8 *The probability that the agent's algorithm (4) does not select the true reward-maximizer arm at time $t \in [\widetilde{k}, T]$ satisfies*

$$\mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t) \neq \arg\max_{a\in\mathcal{A}} s_a^0 + \pi_{t,a}\right) = O\left(\frac{\sqrt{\log 2t}}{\sqrt{t}}\right) \quad (4.37)$$

Proposition 4.8 is proven by using the mathematical induction technique. Our theoretical analysis in this section shows that the probability bound associated with the inaccuracy of the arms chosen by the agent directly depends on: i) exploration rates of the agent and the principal, ii) the agent's estimation gap between $\widehat{\mathbf{s}}_t^{\text{ag}}$ and $\mathbf{s}^0$, and iii) the principal's estimation gap between $\widehat{\mathbf{s}}_t^{\text{pr}}$ and $\mathbf{s}^0$. The complete proof is provided in Appendix C.1.3.

Remark 4.4 *Proposition 4.8 demonstrates the validity and feasibility of Assumption 4.2 by establishing its satisfaction with a classical MAB algorithm.*

4.4.2 Agent's Information Rent

As in every adverse selection model, there is an unavoidable information rent that the agent extracts from the principal due to the information asymmetry inherent in the considered hidden rewards setting. Our data-driven incentive design framework aims to minimize the extra amount principal the pays to the agent on top of this minimal amount of information rent by inducing the agent's incentive compatibility. We implicitly ensure this when we optimize the principal's exploitation incentives (16)–(19) which are designed to drive the agent pick the arm that the principal wants them to pick. However, our analysis assumes that the incentivized agent acts truthfully in the sense that their choices are made in a consistent way with their estimated expected reward vector $\widehat{\mathbf{s}}_t^{\text{ag}}$ – which may not be the case in practice.

From the standpoint of the self-interested agent, even though the agent is able to consistently learn their true rewards throughout the sequential game, we observe that they could just pretend that their learned rewards are different and make their choices in accordance with these "pretended" rewards to extract a higher information rent from the principal and maximize their total rewards. In that regard, as also discussed in Section 3.4, the agent's problem can be formalized as an optimization model that maximizes their information rent. The optimum objective value can be achieved when the agent is knowledgeable about the principal's rewards as well, enabling them to misinform the principal and demand extra payment. It is essential to notice that there is no way for the principal to prohibit the agent from this misbehavior in the hidden rewards settings studied in the previous and current chapters, whereas it could be possible in other adverse selection models where the principal has access to more information about the agent's reward model. Analyzing such models remains an interesting future research direction.

Our framework considers a self-interested agent with particular sophistication who is not attempting to learn the principal's rewards. Therefore, our solution can be regarded as yielding the worst-case result for an agent with higher sophistication. On the other hand, in most of the real-life applications that we consider, the incentivized agents do not have enough elaboration on learning the principal's model. Further, as their knowledge level decreases, they can only get smaller amounts of information rent from the principal. In any case, our approach is designed to minimize the principal's payments and maximize their cumulative expected net reward subject to any amount of information rent that the agent takes regardless of their level of complexity.

4.5 Numerical Experiments

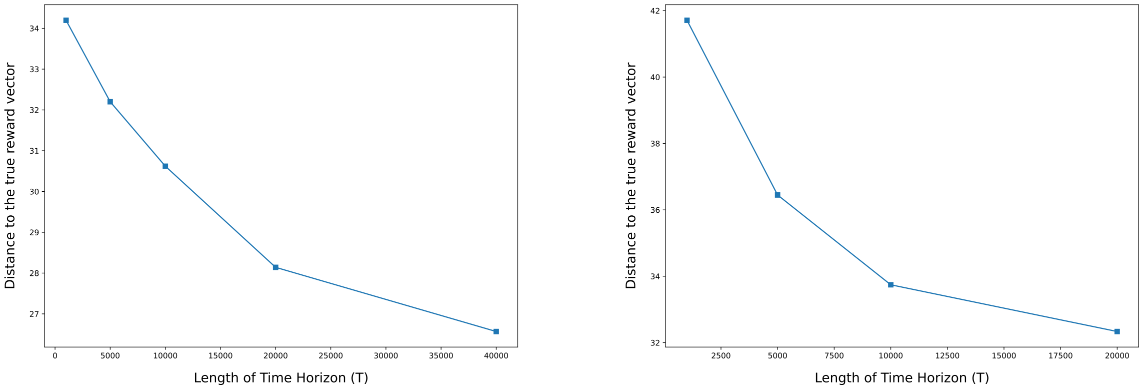
This section presents the numerical results supporting our theoretical bounds on the finite-sample concentration of the proposed estimator and on the convergence of the proposed incentive policy. We conduct simulation experiments on an instance of the contract design problem described in Section 4.1.1.1 for grid-scale distributed storage and aggregation of renewable energy.

Consider an independent aggregator company that operates distributed rechargeable battery storage systems to offer an affordable and environmentally sustainable alternative to the utility grid, and a utility company that acknowledges that it can reduce its costs and carbon footprint, as well as improve the reliability of its service, by contracting the extra supply capacity available in the aggregator's battery systems during peak demand or extreme weather emergency periods. These contracts are done by participation of the aggregator in the day-ahead energy market that allows the aggregator generate income by sharing their battery capacities with the utility. In this context, the utility needs to consistently estimate the aggregator's profits to design smart and efficient payments that will expand the aggregator's participation in the market.

We model this problem by leveraging the repeated adverse selection model given in Section 4.2.1. The day-ahead market operates on an hourly basis where the utility company (i.e., principal) offers a price for energy delivered (dollar/MW) by the aggregator (i.e., agent) for each hour of the next day. Consistent with the related literature (e.g. Bessa et al. 2011, Zhao et al. 2015, Rahimiyan and Baringo 2015), the aggregator is assumed to be a price-taker that only decides the shared energy quantities. Their decisions are assumed to have no effect on the marginal electricity price, given that most electricity markets are still dominated by large conventional generators. Each time step t in our sequential model corresponds to an hourly session over the entire period $\mathcal{T}$ of contract between the aggregator and the utility. The aggregator's problem is formalized as a MAB model where each arm $a \in \mathcal{A}$ corresponds to a different range of power capacity (MW) they provide for the use of the utility. The supply capacity is bounded by the maximum total amount of energy that can be consumed by the aggregator's entire battery systems. For example, if an electric vehicle is consuming 2 kW in an hour and the aggregator owns one thousand electric vehicles connected in residential charging points of the clients who allow the aggregator to control their charging processes, then the aggregator can provide a capacity of up to 2 MW per hour (Bessa et al. 2011). In this scenario, we can divide this feasible interval of shared capacity $[0, 2]$ into a discrete number of subintervals to define the aggregator's arms, for example $\mathcal{A} = \{[0, 0.5], [0.5, 1], [1, 1.5], [1.5, 2]\}$. In each session, the aggregator chooses an arm $\upsilon_t(\boldsymbol{\pi}_t)$ in response to the payment scheme $\boldsymbol{\pi}_t$ offered by the utility company. This choice, along with several sources of uncertainty such as variations in renewable energy generation, observed electricity demand in the community, and market prices, is uncontrollable by the utility. These factors collectively determine the realized profits of the aggregator $(\rho_{t,\upsilon_t(\boldsymbol{\pi}_t)})$ and the utility $(\mu_{t,\upsilon_t(\boldsymbol{\pi}_t)})$ in each session.

We run our experiments for multiple combinations of the dimension of the aggregator's MAB model $n \in \{5, 10\}$ and the length of the contract period $T \in \{10^3, 5 \cdot 10^3, 10 \cdot 10^3, 20 \cdot 10^3, 40 \cdot 10^3\}$ in hours. Each setting is replicated five times, and the average and standard deviation of our performance metrics are reported across these replicates. The realizations of the profits of the aggregator and those of the utility company are assumed to follow Gaussian distributions, $\mathcal{N}(\mathbf{r}^0, 10)$ and $\mathcal{N}(\boldsymbol{\theta}^0, 10)$, respectively, where the values used for the expectations $\mathbf{r}^0$ and $\boldsymbol{\theta}^0$ are given in Table B.1 in Appendix C.2. Further, we take the compact set to which the aggregator's expected profits belong as $\mathcal{R} = [-20, 50]$ and the feasible range of the utility's payments as $\mathcal{C} = [-20, 60]$ (which makes $\gamma = 10$ as introduced in Assumption 4.1). The input parameters for the utility's algorithm (3) are chosen as $m^{\text{pr}} = 5$ and $w = 1/5$ in all settings, implying that the utility explores with probability one during the first $214 - n$ sessions after the initialization period (lines 2-5). To simulate the aggregator's choices, we use Algorithm (4) with parameter $m^{\text{ag}} = 10$, which implies that the aggregator explores with probability one during the first $100 - n$ sessions after the initialization period (lines 2-6).

We consider three main metrics to demonstrate the performance of our approach in the considered experimental setting: 1) concentration of our estimator towards the true expected profit vector $\mathbf{s}^0$, 2) convergence of our payment policy to the oracle payment policy introduced in Section 4.3.2.1, and 3) cumulative regret of the proposed united data-driven framework.



(a) $\|\mathbf{s}^0 - \widehat{\mathbf{s}}_T^{\mathrm{pr}}\|_\infty$ for $n = 5$

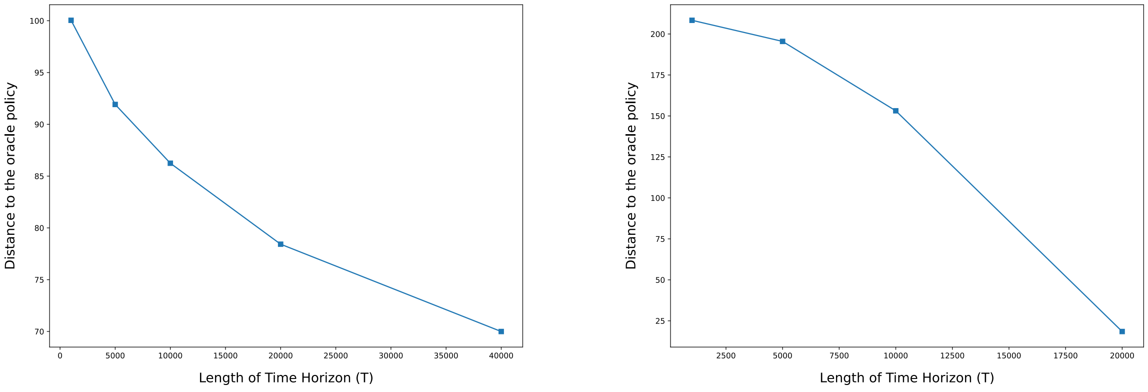
(b) $\|\mathbf{s}^0 - \widehat{\mathbf{s}}_T^{\mathrm{pr}}\|_\infty$ for $n = 10$

Figure 4.1: Estimator concentration measured in terms of the ℓ_∞ distance between the true mean profit vector $\mathbf{s}^0$ and the final mean profit vector $\widehat{\mathbf{s}}_T^{\mathrm{pr}}$ obtained by the proposed estimator.

First, we provide a direct measure of the accuracy of the proposed estimator (4.6) for the aggregator's mean profits to support our finite-sample concentration results proven in Section 4.2. In Figure 4.1, we present the ℓ_∞-distance between the final estimated mean profit vector $\widehat{\mathbf{s}}_T^{\mathrm{pr}}$ and the true mean profit vector $\mathbf{s}^0$. These results display the consistency of our novel estimator and the achieved accuracy, even when the size of the MAB model is doubled.

Second, we measure the convergence of the payment policy generated by Algorithm 3 to the oracle payment policy. A significant challenge in the utility's contract design problem is that the inherent learning problem gets harder as the number of arms (i.e., the supply capacity) of the aggregator gets larger. That is because they need to compute a payment amount for each possible arm (not only for the desired one) as accurately as possible in order to steer the aggregator's participation in the sequential day-ahead sessions. Because every alternative arm matters the same, we measure the distance between the proposed payment vector and the oracle payment vector in terms of the ℓ_1 norm metric which weights all the entries of the vectors equally. As seen in Figure 4.2, the proposed payment mechanism is able to consistently converge to the oracle mechanism, and it achieves a better convergence as the length of the contract horizon gets longer. Further, a comparison of Figures 4.2a and 4.2b reveals the highlighted challenge regarding the relation between the difficulty of the contract design problem and the dimension of the considered MAB model.

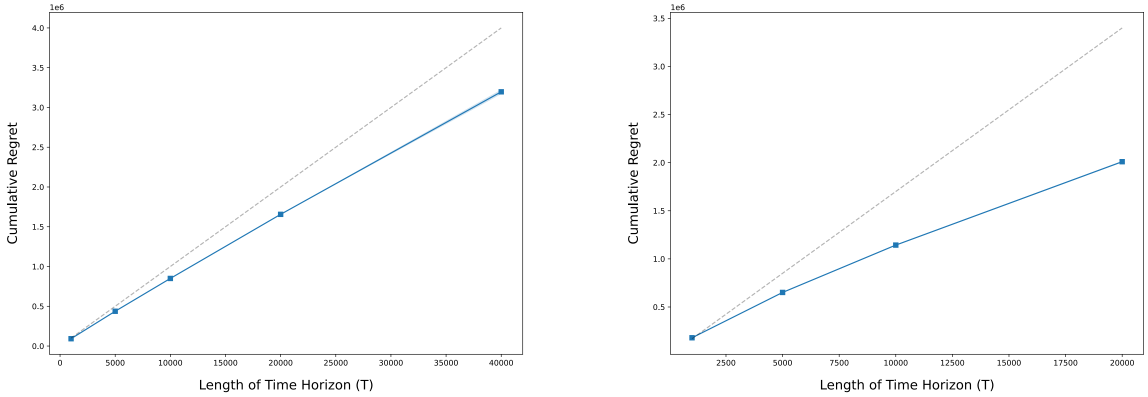
Last, we present the overall performance of our adaptive incentive framework where we unite our consistent estimator with the proposed policy. Figure 4.3 shows the cumulative regret accrued by the utility company for different n and T values. As expected, our approach achieves a sublinear regret that matches with the proven asymptotic order in Remark 4.3.



(a) $\|\boldsymbol{\pi}_T - \mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0)\|_1$ for $n = 5$

(b) $\|\boldsymbol{\pi}_T - \mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0)\|_1$ for $n = 10$

Figure 4.2: Policy convergence measured in terms of the ℓ_1 distance between the oracle payments $\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0)$ and the payments $\boldsymbol{\pi}_T$ reached by Algorithm 3 at the end of the contract horizon. $\big($For any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^M$, the ℓ_1 distance is defined by $\|\mathbf{x}-\mathbf{y}\|_1 = \sum_{m=1}^{M} |x_m - y_m|$.$\big)$



(a) Regret for $n = 5$

(b) Regret for $n = 10$

Figure 4.3: The average cumulative regret of the policies generated by Algorithm 3.

4.6 Conclusions and Future Directions

Motivated by a variety of unexplored incentive design applications, the analyses in this chapter aim to bridge the well-established principal-agent theory with recent advances in online sequential learning theory. We model a generic and practically relevant repeated adverse selection game within a MAB framework where a principal incentivizes a self-interested learning agent by only watching the agent's decisions in response to the offered incentives over a finite time horizon. To effectively lead the agent and ensure incentive compatibility, the principal must consistently adapt their incentives by learning the agent's true reward expectations (which are also unknown to the agent), without observing the agent's random reward realizations in each play of the game.

Because we consider an imperfect-knowledge agent, we engage with two algorithms trained in parallel by the principal and the agent through a two-way dynamic interaction over the course of the game. Within this complex scenario, we address both the estimation/learning and the sequential incentive design problems that the principal faces. We primarily focus on their exploration/exploitation trade-off and provide rigorous convergence guarantees that are independent of the type of the agent's algorithm.We also provide insights from the perspective of the selfish agent, who seeks to maximize their rewards by extracting higher information rents from the principal. Lastly, we reinforce our theoretical findings with numerical experiments, justifying the effectiveness and efficiency of the proposed framework in managing aggregated renewable energy supply for smart and reliable community grids.

The *hidden agent rewards* setting remains mostly unexplored in the literature due to the analytical intricacy of jointly studying the learning and incentive design problems. However, we believe our contributions in this dissertation can channel promising future work for data-driven contract design within the sustainable OM literature. One interesting direction would be to consider a multi-agent setting where a principal gets to collaborate with multiple reward-maximizer selfish agents. We suppose that our model and integrated framework is applicable to scenarios where the agents collectively work as a team and the principal provides team incentives based on the observed team-level decisions. On the other hand, studying a scenario where the incentives need to be designed separately for each individual selfish-agent (who might be also communicating with other agents) would require a completely different approach and theoretical analysis. Furthermore, the repeated adverse selection models that we introduce in Chapters 3 and 4 are designed purposefully to be as generic and simple as possible to improve practical relevance for various domains. Our models can be further extended and specialized to accommodate features of a particular incentive application that may attract the OM practitioners.

Chapter 5

Concluding Remarks and Foresight

It is evident that a seismic and unprecedented paradigm shift has emerged in the management of operations within the contemporary socio-technical systems. This shift is propelled by the dual forces of an increasing dependence on extensive data flow and the challenge of surviving unpredictable disruptions stemming from the climate crisis. Motivated by contributing to this breakthrough, this dissertation introduces novel approaches that leverage flowing data to derive sustainable and collaborative operational solutions, with a focus on mitigating externalities associated with incomplete information about the system model and stakeholder strategies.

An essential aspect of this mitigation is to strike a balance in the trade-off of allocating resources between exploring potential better alternatives for long-term profits and exploiting current knowledge to maximize short-term gains. Each chapter of this dissertation examines this exploration/exploitation trade-off under distinct externality settings with model uncertainty and information asymmetry, which have not been thoroughly explored in the literature due to their analytical complexities. The major portion is devoted to designing data-driven and adaptive incentive mechanisms that assist a strategic learner consistently guide another self-interested agent to extract latent information about their true rewards while proactively mitigating their adversarial actions. The developed frameworks unify tools from classical and advanced methodologies to analyze the trade-offs faced by both parties concurrently over a finite-horizon game. This generic context caters to different OM applications, ranging from renewable energy aggregation contracts for utility grids to forest conservation incentives in PES programs.

The methodological contributions of this dissertation prove a rigorous theoretical basis for bridging information gaps in data-driven and sustainable decision-making across diverse OM domains. This theory serves as the essence for future research, which could be developed to enhance its applicability and confront various business constraints. To further advance this prosperous research landscape, one can address an additional layer of information imbalance through environmental data monitoring mechanisms. In many sustainability practices, asymmetry arises not only from the unknown true rewards of incentivized parties but also from a significant lacuna in observing their actual actions. In this regard, an open research question emerges: How can modern monitoring technologies and novel forms of environmental data be incorporated to inform adaptive incentive structures? For instance, the incentive policies employed in forest conservation projects could be reinforced through the utilization of high-resolution canopy height maps generated by artificial intelligence and remote sensing technologies for the incentivized land areas (e.g. Lang et al. 2023). Similarly, emission transparency contracts between retailers and suppliers should be shaped by data gathered from continuous/predictive emissions monitoring systems (CEMS/PEMS) and other pertinent technologies. Towards this direction, aligning these new environmental data types with the practical implications of this dissertation holds substantial potential to yield intelligent incentive frameworks. We foresee that these data-driven advanced frameworks will promote responsive, profitable, and sustainable OM practices in our ever-changing and disruptive world.

Bibliography

Abbasi-Yadkori Y, Lazic N, Szepesvari C (2019) Model-free linear quadratic control via reduction to expert prediction. *AISTATS*.

Abhishek K, Jain S, Gujar S (2020) Designing truthful contextual multi-armed bandits based sponsored search auctions. *arXiv preprint arXiv:2002.11349* .

Abreu D, Pearce D, Stacchetti E (1990) Toward a theory of discounted repeated games with imperfect monitoring. *Econometrica: Journal of the Econometric Society* 1041–1063.

Afram A, Janabi-Sharifi F (2014) Theory and applications of HVAC control systems–A review of model predictive control (MPC). *Building and Environment* 72:343–355.

Agarwal N, Brukhim N, Hazan E, Lu Z (2020) Boosting for Control of Dynamical Systems. *ICML*, 96–103.

Agrawal S, Goyal N (2013) Thompson sampling for contextual bandits with linear payoffs. *ICML*, 127–135.

Ahuja RK, Orlin JB (2001) Inverse optimization. *Operations Research* 49(5):771–783.

Amelin KS, Granichin ON (2012) Randomized controls for linear plants and confidence regions for parameters under external arbitrary noise. *2012 American Control Conference (ACC)* .

Amin K, Rostamizadeh A, Syed U (2014) Repeated Contextual Auctions with Strategic Buyers. *Advances in Neural Information Processing Systems*, volume 27.

Anderson BD, Moore JB (2012) *Optimal filtering* (Dover).

Apple (2023) Supplier Responsibility. Accessed February 1, 2024, https://www.apple.com/supplier-responsibility/.

Arefi M, Montazeri A, Poshtan J, Jahed-Motlagh M (2006) Nonlinear Model Predictive Control of Chemical Processes with a Wiener Identification Approach. *2006 IEEE International Conference on Industrial Technology*, 1735–1740, URL http://dx.doi.org/10.1109/ICIT.2006.372470.

ASHRAE (2013) Ansi/ashrae standard 55-2013: Thermal environmental conditions for human occupancy.

Aswani A (2019) Statistics with set-valued functions: applications to inverse approximate optimization. *Mathematical Programming* 174(1-2):225–251.

Aswani A, Bouffard P, Tomlin C (2012) Extensions of learning-based model predictive control for real-time application to a quadrotor helicopter. *2012 American Control Conference (ACC)*, 4661–4666.

Aswani A, Gonzalez H, Sastry SS, Tomlin C (2013) Provably safe and robust learning-based model predictive control. *Automatica* 49.

Aswani A, Master N, Taneja J, Culler D, Tomlin C (2011) Reducing transient and steady state electricity consumption in HVAC using learning-based model-predictive control. *Proc. IEEE* 100.

Aswani A, Shen ZJ, Siddiq A (2018) Inverse optimization with noisy data. *Operations Research* 66(3):870–892.

Aswani A, Shen ZJM, Siddiq A (2019) Data-driven incentive design in the medicare shared savings program. *Operations Research* 67(4):1002–1026.

Audy JF, Lehoux N, D'Amours S, Rönnqvist M (2012) A framework for an efficient implementation of logistics collaborations. *International Transactions in Operational Research* 19(5):633–657.

Bärmann A, Martin A, Pokutta S, Schneider O (2018) An online-learning approach to inverse optimization. *arXiv preprint arXiv:1810.12997* .

Berntzen L, Meng Q, Johannessen MR, Vesin B, Brekke T, Laur I (2021) The aggregator as a storage provider. *2021 11th International Conference on Power and Energy Systems (ICPES)*, 190–195 (IEEE).

Bertsekas DP, Yu H (2010) Distributed asynchronous policy iteration in dynamic programming. *Allerton*, 1368–1375.

Bertsimas D, Gupta V, Paschalidis IC (2015) Data-driven estimation in equilibrium using inverse optimization. *Mathematical Programming* 153(2):595–633.

Besbes O, Gur Y, Zeevi A (2014) Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 199–207.

Bessa RJ, Matos MA, Soares FJ, Lopes JAP (2011) Optimized bidding of a ev aggregation agent in the electricity market. *IEEE Transactions on Smart Grid* 3(1):443–452.

Bhat S, Jain S, Gujar S, Narahari Y (2019) An optimal bidimensional multi-armed bandit auction for multi-unit procurement. *Annals of Mathematics and Artificial Intelligence* 85(1):1–19.

Bianchini G, Casini M, Pepe D, Vicino A, Zanvettor GG (2019) An integrated model predictive control approach for optimal hvac and energy storage operation in large-scale buildings. *Applied Energy* 240:327–340.

Biggins F, Ejeh JO, Brown S (2022) Going, going, gone: Optimising the bidding strategy for an energy storage aggregator and its value in supporting community energy storage. *Energy Reports* 8.

Bindra H, Revankar S (2018) *Storage and Hybridization of Nuclear Energy: Techno-Economic Integration of Renewable and Nuclear Energy* (Academic Press).

Biswas A, Jain S, Mandal D, Narahari Y (2015) A truthful budget feasible multi-armed bandit mechanism for crowdsourcing time critical tasks. *AAMAS*, 1101–1109.

Biyik E, Margoliash J, Alimo SR, Sadigh D (2019) Efficient and safe exploration in deterministic markov decision processes with unknown transition models. *ACC*, 1792–1799.

Boffi NM, Tu S, Slotine JJE (2021) Regret bounds for adaptive nonlinear control. *Learning for Dynamics and Control*.

Bolton P, Dewatripont M (2004) *Contract theory* (MIT press).

Bonnans JF, Shapiro A (2013) *Perturbation analysis of optimization problems* (Springer Science & Business Media).

Borrelli F, Bemporad A, Morari M (2017) *Predictive Control for Linear and Hybrid Systems* (Cambridge University Press).

Bosworth HB (2010) Medication adherence. *Improving patient treatment adherence*, 68–94 (Springer).

Boucheron S, Lugosi G, Massart P (2013) *Concentration inequalities: A nonasymptotic theory of independence* (Oxford university press).

Bouneffouf D, Féraud R (2016) Multi-armed bandit problem with known trend. *Neurocomputing* 205:16–21.

Bradtke SJ, Ydstie BE, Barto AG (1994) Adaptive linear quadratic control using policy iteration. *ACC*, volume 3, 3475–3479.

Braverman M, Mao J, Schneider J, Weinberg SM (2019) Multi-armed bandit problems with strategic arms. *Conference on Learning Theory*, 383–416 (PMLR).

Brown MT, Bussell J, Dutta S, Davis K, Strong S, Mathew S (2016) Medication adherence: truth and consequences. *The American journal of the medical sciences* 351(4):387–399.

Budd M, Lacerda B, Duckworth P, West A, Lennox B, Hawes N (2020) Markov decision processes with unknown state feature values for safe exploration using gaussian processes. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).*

Busch J, Ferretti-Gallon K (2017) What drives deforestation and what stops it? a meta-analysis. *Review of Environmental Economics and Policy* .

Chade H, Swinkels J (2019) Disentangling moral hazard and adverse selection. Technical report, Working Paper, Arizona State University.

Chan TC, Eberg M, Forster K, Holloway C, Ieraci L, Shalaby Y, Yousefi N (2022) An inverse optimization approach to measuring clinical pathway concordance. *Management Science* 68(3):1882–1903.

Chan TC, Lee T, Terekhov D (2019) Inverse optimization: Closed-form solutions, geometry, and goodness of fit. *Management Science* 65(3):1115–1135.

Chen C, Modares H, Xie K, Lewis FL, Wan Y, Xie S (2019) Reinforcement learning-based adaptive optimal exponential tracking control of linear systems with unknown dynamics. *IEEE Transactions on Automatic Control* 64(11):4423–4438.

Chen Y, Tong Z, Zheng Y, Samuelson H, Norford L (2020) Transfer learning with deep neural networks for model predictive control of hvac and natural ventilation in smart buildings. *Journal of Cleaner Production* 254:119866.

Chu B, Duncan S, Papachristodoulou A, Hepburn C (2012) Using economic model predictive control to design sustainable policies for mitigating climate change. *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, 406–411.

Cohen A, Koren T, Mansour Y (2019) Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. *Proceedings of the 36th International Conference on Machine Learning*, volume 97.

Conitzer V, Garera N (2006) Learning algorithms for online principal-agent problems (and selling goods online). *Proceedings of the 23rd International Conference on Machine Learning*, 209–216, ICML '06.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39(1):1–22.

Devanur NR, Kakade SM (2009) The price of truthfulness for pay-per-click auctions. *EC '09*.

Di Cairano S, Kolmanovsky IV (2018) Real-time optimization and model predictive control for aerospace and automotive applications. *2018 Annual American Control Conference (ACC)*, URL `http://dx.doi.org/10.23919/ACC.2018.8431585`.

Dionne G, Lasserre P (1985) Adverse Selection, Repeated Insurance Contracts and Announcement Strategy. *The Review of Economic Studies* 52(4):719–723.

Dong C, Chen Y, Zeng B (2018) Generalized inverse optimization through online learning. *Advances in Neural Information Processing Systems* 31.

Dong C, Zeng B (2020) Inverse multiobjective optimization through online learning. *arXiv preprint arXiv:2010.06140* .

Doucet A, De Freitas N, Gordon N (2001) An introduction to sequential monte carlo methods. *Sequential Monte Carlo methods in practice*, 3–14 (Springer).

Doucet A, Godsill S, Andrieu C (2000) On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing* 10.

Durbin J, Koopman SJ (2000) Time series analysis of non-gaussian observations based on state space models from both classical and bayesian perspectives. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62(1):3–56.

Early C (2011) Delivering greener logistics. URL `https://www.iema.net/articles/delivering-greener-logistics`.

East Bay Community Energy (2020) EBCE Launches First-of-its-kind Home Battery Backup Program. `https://ebce.org/news-and-events/ebce-launches-first-of-its-kind-home-battery-backup-program/`.

Eaton JW, Rawlings JB (1992) Model-predictive control of chemical processes. *Chemical Engineering Science* 47(4):705–720.

Engel S, et al. (2016) The devil in the detail: a practical guide on designing payments for environmental services. *International Review of Environmental and Resource Economics* 9(1–2):131–177.

Eren U, Prach A, Koçer BB, Raković SV, Kayacan E, Açıkmeşe B (2017) Model predictive control in aerospace systems: Current state and opportunities. *Journal of Guidance, Control, and Dynamics* 40(7):1541–1566.

Ergun O, Kuyzu G, Savelsbergh M (2007) Reducing truckload transportation costs through collaboration. *Transportation science* 41(2):206–221.

Esfahani PM, Shafieezadeh-Abadeh S, Hanasusanto GA, Kuhn D (2018) Data-driven inverse optimization with imperfect information. *Mathematical Programming* 167(1):191–234.

Eső P, Szentes B (2017) Dynamic contracting: An irrelevance theorem. *Theoretical Economics* 12(1):109–139.

Estep D (2010) *Practical Analysis in One Variable* (Springer).

Fan Y, Ming Y (2020) Efficient exploration for model-based reinforcement learning with continuous states and actions. *arXiv preprint arXiv:2012.09613* .

Fang J, Ma R, Deng Y (2020) Identification of the optimal control strategies for the energy-efficient ventilation under the model predictive control. *Sustainable Cities and Society* 53.

Fazel M, Ge R, Kakade S, Mesbahi M (2018) Global convergence of policy gradient methods for the linear quadratic regulator. *International Conference on Machine Learning*, 1467–1476 (PMLR).

Food TUN, (FAO) AO (2020) Global forest resources assessment 2020. URL `https://www.fao.org/forest-resources-assessment/2020/en/`.

Gao G, Huang S, Huang H, Xiao M, Wu J, Sun YE, Zhang S (2022) Combination of auction theory and multi-armed bandits: Model, algorithm, and application. *IEEE Transactions on Mobile Computing* .

Garcia IF, Chanfreut P, Jurado I, Maestre JM (2020) A data-based model predictive decision support system for inventory management in hospitals. *IEEE Journal of Biomedical and Health Informatics* .

Garivier A, Moulines E (2008) On Upper-Confidence Bound Policies for Non-Stationary Bandit Problems. *arXiv preprint arXiv:0805.3415* .

Gaskett C (2003) Reinforcement learning under circumstances beyond its control. `https://eprint.iacr.org/1990/001`.

Gayle GL, Miller RA (2015) Identifying and testing models of managerial compensation. *The Review of Economic Studies* 82(3):1074–1118.

Ghahramani Z, Hinton GE (1996) Parameter estimation for linear dynamical systems. Technical report, Technical Report CRG-TR-96-2, University of Toronto, Dept. of Computer Science.

Ghamat S, Zaric GS, Pun H (2018) Contracts to promote optimal use of optional diagnostic tests in cancer treatment. *Production and Operations Management* 27(12):2184–2200.

Gneezy U, Meier S, Rey-Biel P (2011) When and why incentives (don't) work to modify behavior. *Journal of economic perspectives* 25(4):191–210.

Gottlieb D, Moreira H (2022) Simple contracts with adverse selection and moral hazard. *Theoretical Economics* 17(3):1357–1401.

Gros S, Zanon M (2019) Data-driven economic nmpc using reinforcement learning. *IEEE TAC* 65(2):636–648.

Gros S, Zanon M (2020) Reinforcement learning for mixed-integer problems based on mpc. *ArXiv* abs/2004.01430.

Grossman S, Hart O (1983) An analysis of the principal-agent problem. *Econometrica* 51(1):7–45.

Guo P, Tang CS, Wang Y, Zhao M (2019) The impact of reimbursement policy on social welfare, revisit rate, and waiting time in a public healthcare system: Fee-for-service versus bundled payment. *Manufacturing & Service Operations Management* 21(1):154–170.

Halac M, Kartik N, Liu Q (2016) Optimal contracts for experimentation. *The Review of Economic Studies* 83(3):1040–1091.

Hall EC, Willett RM (2013) Dynamical models and tracking regret in online convex programming. *arXiv preprint arXiv:1301.1254* .

Han Y, Zhou Z, Flores A, Ordentlich E, Weissman T (2020) Learning to bid optimally and efficiently in adversarial first-price auctions. *arXiv preprint arXiv:2007.04568* .

Hart O, Holmström B (1987) The theory of contracts. *Advances in economic theory: Fifth world congress*, volume 71, 155 (Cambridge).

Heger M (1994) Consideration of risk in reinforcement learning. *Machine Learning Proceedings 1994*, 105–111 (Elsevier).

Hespanhol P, Aswani A (2020) Statistical consistency of set-membership estimator for linear systems. *IEEE Control Systems Letters* 4(3):668–673.

Heuberger C (2004) Inverse combinatorial optimization: A survey on problems, methods, and results. *Journal of combinatorial optimization* 8(3):329–361.

Hewing L, Wabersich KP, Menner M, Zeilinger MN (2020) Learning-based model predictive control: Toward safe learning in control. *Annual Review of Control, Robotics, and Autonomous Systems* 3.

Ho CJ, Slivkins A, Vaughan J (2016) Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. *Journal of Artificial Intelligence Research* 55:317–359.

Holmström B (1979) Moral hazard and observability. *The Bell journal of economics* 74–91.

Hu G, You F (2022) Renewable energy-powered semi-closed greenhouse for sustainable crop production using model predictive control and machine learning for energy management. *Renewable and Sustainable Energy Reviews* 168:112790.

International Renewable Energy Agency (2019) Innovation landscape brief: Aggregators.

Jain S, Narayanaswamy B, Narahari Y (2014) A multiarmed bandit incentive mechanism for crowdsourcing demand response in smart grids. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

Jaksch T, Ortner R, Auer P (2010) Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research* 11(4).

Jazwinski AH (2007) *Stochastic processes and filtering theory* (Courier Corporation).

Juan AA, Faulin J, Pérez-Bernabeu E, Jozefowiez N (2014) Horizontal cooperation in vehicle routing problems with backhauling and environmental criteria. *Procedia - Social and Behavioral Sciences* 111:1133–1141.

Jurado I, Maestre JM, Velarde P, Ocampo-Martínez C, Fernández I, Tejera BI, del Prado JR (2016) Stock management in hospital pharmacy using chance-constrained model predictive control. *Computers in biology and medicine* 72:248–255.

Kakade S, Krishnamurthy A, Lowrey K, Ohnishi M, Sun W (2020) Information theoretic regret bounds for online nonlinear control. *arXiv preprint arXiv:2006.12466* .

Kalman RE (1960) A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* .

Kalmuk A, Tyushev K, Granichin ON, Yuchi M (2017) Online parameter estimation for MPC model uncertainties based on LSCR approach. *IEEE CCTA* .

Karnchanachari N, Valls MI, Hoeller D, Hutter M (2020) Practical Reinforcement Learning For MPC: Learning from sparse objectives in under an hour on a real robot. *L4DC*.

Kaynar N, Siddiq A (2022) Estimating Effects of Incentive Contracts in Online Labor Platforms. *Management Science* 69(4):2106–2126.

Kennedy R (2021) Energy storage aggregation unlocks benefits for homeowners, grid operators, and installers. `https://pv-magazine-usa.com/2021/07/13/energy-storage-aggregation-unlocks-benefits-for-homeowners-grid-operators-and-installers/`.

Keshavarz A, Wang Y, Boyd S (2011) Imputing a convex objective function. *2011 IEEE international symposium on intelligent control*, 613–619 (IEEE).

Kitagawa G (1996) Monte carlo filter and smoother for non-gaussian nonlinear state space models. *J Comput Graph Stat* 5(1).

Kiumarsi-Khomartash B, Lewis F, Jiang Z (2017) H∞ control of linear discrete-time systems: Off-policy reinforcement learning. *Autom.* 78:144–152.

Kiumarsi-Khomartash B, Lewis F, Modares H, Karimpour A, Naghibi-Sistani MB (2014) Reinforcement q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics. *Autom.* 50.

Klenske ED, Hennig P (2016) Dual control for approximate bayesian reinforcement learning. *ArXiv* abs/1510.03591.

Koller T, Berkenkamp F, Turchetta M, Krause A (2018) Learning-based model predictive control for safe exploration. *2018 IEEE conference on decision and control (CDC)*, 6059–6066 (IEEE).

Kolmanovsky I, Gilbert EG (1998) Theory and computation of disturbance invariant sets for discrete-time linear systems. *Mathematical Problems in Engineering* 4:317–367.

Kumar R, Wenzel MJ, ElBsat MN, Risbeck MJ, Drees KH, Zavala VM (2020) Stochastic model predictive control for central hvac plants. *Journal of Process Control* 90:1–17.

Lagarde M, Haines A, Palmer N (2007) Conditional cash transfers for improving uptake of health interventions in low-and middle-income countries: a systematic review. *Jama* 298(16):1900–1910.

Lale S, Azizzadenesheli K, Hassibi B, Anandkumar A (2020) Regret bound of adaptive control in linear quadratic gaussian (lqg) systems. *ArXiv* abs/2003.05999.

Lang N, Jetz W, Schindler K, Wegner JD (2023) A high-resolution canopy height model of the earth. *Nature Ecology & Evolution* 7(11):1778–1789.

Lazaric A, Ghavamzadeh M, Munos R (2010) Analysis of a classification-based policy iteration algorithm. *ICML*.

Lee DKK, Zenios SA (2012) An evidence-based incentive system for medicare's end-stage renal disease program. *Management Science* 58(6):1092–1105.

Levy AB, Poliquin RA, Rockafellar RT (2000) Stability of locally optimal solutions. *SIAM Journal on Optimization* 10(2):580–604.

Li WD, Ashlagi I, Lo I (2023) Simple and approximately optimal contracts for payment for ecosystem services. *Management Science* 69(12):7821–7837.

Limon D, Alamo T, Raimondo DM, De La Peña DM, Bravo JM, Ferramosca A, Camacho EF (2009) Input-to-state stability: a unifying framework for robust model predictive control. *Nonlinear Model Predictive Control: Towards New Challenging Applications* 1–26.

Liu D, Wei Q (2014) Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems. *IEEE Transactions on Neural Networks and Learning Systems* 25.

Liu JS, Chen R (1998) Sequential monte carlo methods for dynamic systems. *Journal of the American statistical association* 93(443):1032–1044.

Long Q, Smith H, Zhang T, Tang S, Garner P (2011) Patient medical costs for tuberculosis treatment and impact on adherence in china: a systematic review. *BMC public health* 11(1):1–9.

Ma Y, Borrelli F, Hencey B, Coffey B, Bengea S, Haves P (2011) Model predictive control for the operation of building cooling systems. *IEEE Transactions on control systems technology* 20(3):796–803.

Maasoumy M, Sangiovanni-Vincentelli A (2012) Total and peak energy consumption minimization of building hvac systems using model predictive control. *IEEE Design & Test of Computers* 29.

Maestre J, Fernández M, Jurado I (2018) An application of economic model predictive control to inventory management in hospitals. *Control Engineering Practice* 71:120–128.

Maheshwari C, Kulkarni K, Wu M, Sastry SS (2022) Inducing social optimality in games via adaptive incentive design. *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2864–2869 (IEEE).

Maheshwari C, Sasty SS, Ratliff L, Mazumdar E (2023) Convergent First-Order Methods for Bilevel Optimization and Stackelberg Games. *arXiv preprint arXiv:2302.01421* .

Marques A, Soares R, Santos MJ, Amorim P (2020) Integrated planning of inbound and outbound logistics with a rich vehicle routing problem with backhauls. *Omega* 92:102172.

Martimort D, Laffont JJ (2009) *The Theory of Incentives: The Principal-Agent Model* (Princeton University Press).

Mason CF, Plantinga AJ (2013) The additionality problem with offsets: Optimal contracts for carbon sequestration in forests. *Journal of Environmental Economics and Management* 66(1):1–14.

Mayne DQ, Rawlings JB, Rao CV, Scokaert PO (2000) Constrained model predictive control: Stability and optimality. *Automatica* 36(6):789–814.

McCarthy E (2022) PG&E, Tesla launch program to use customers' Powerwall batteries to tackle California reliability concerns. `https://www.utilitydive.com/news/pge-tesla-launch-program-to-use-customers-powerwall-batteries-to-tackle/626297/`.

Mesbah A (2018) Stochastic model predictive control with active uncertainty learning: A survey on dual control. *Annu. Rev. Control.* 45:107–117.

Mintz Y, Aswani A, Kaminsky P, Flowers E, Fukuoka Y (2017) Non-stationary bandits with habituation and recovery dynamics. *Operations Research* 68.

Mintz Y, Aswani A, Kaminsky P, Flowers E, Fukuoka Y (2023) Behavioral analytics for myopic agents. *European Journal of Operational Research* 310(2):793–811.

Misra S, Coughlan AT, Narasimhan C (2005) Salesforce compensation: An analytical and empirical examination of the agency theoretic approach. *Quantitative Marketing and Economics* 3(1):5–39.

Misra S, Nair HS (2011) A structural model of sales-force compensation dynamics: Estimation and field implementation. *Quantitative Marketing and Economics* 9(3):211–257.

Moldovan TM, Abbeel P (2012) Safe exploration in markov decision processes. *arXiv preprint arXiv:1205.4810* .

MOSEK ApS (2019) *MOSEK Optimizer API for Python 9.2.37*. URL `https://docs.mosek.com/9.2/pythonapi/index.html`.

Navabi S, Nayyar A (2018) Optimal auction design for flexible consumers. *IEEE Transactions on Control of Network Systems* 6(1):138–150.

Nazerzadeh H, Saberi A, Vohra RV (2008) Dynamic cost-per-action mechanisms and applications to online advertising. *WWW*.

Negenborn R, De Schutter B, Wiering M, Hellendoorn J (2004) Experience-based model predictive control using reinforcement learning. *Proceedings of the 8th TRAIL Congress*.

Oldewurtel F, Parisio A, Jones CN, Gyalistras D, Gwerder M, Stauch V, Lehmann B, Morari M (2012) Use of model predictive control and weather forecasts for energy efficient building climate control. *Energy and Buildings* 45:15–27.

Ostadijafari M, Dubey A (2019) Linear model-predictive controller (LMPC) for building's heating ventilation and air conditioning (HVAC) system. *IEEE CCTA*, 617–623 (IEEE).

Osterberg L, Blaschke T (2005) Adherence to medication. *New England journal of medicine* 353(5):487–497.

Pannocchia G, Rawlings JB, Wright SJ (2011) Conditions under which suboptimal nonlinear mpc is inherently robust. *Systems & Control Letters* 60(9):747–755.

Papadimitriou CH, Tsitsiklis JN (1999) The complexity of optimal queuing network control. *Mathematics of Operations Research* .

Pashenkova E, Rish I, Dechter R (1996) Value iteration and policy iteration algorithms for markov decision problem. *AAAI'96: Workshop on Structural Issues in Planning and Temporal Reasoning* (Citeseer).

Plambeck EL, Zenios SA (2000) Performance-based incentives in a dynamic principal-agent model. *Manufacturing & service operations management* 2(3):240–263.

Radner R (1981) Monitoring cooperative agreements in a repeated principal-agent relationship. *Econometrica: Journal of the Econometric Society* 1127–1148.

Rahimiyan M, Baringo L (2015) Strategic bidding for a virtual power plant in the day-ahead and real-time markets: A price-taker robust optimization approach. *IEEE Transactions on Power Systems* 31(4):2676–2687.

Rakovic SV, Baric M (2010) Parameterized robust control invariant sets for linear systems: Theoretical advances and computational remarks. *IEEE Transactions on Automatic Control* 55.

Raman NS, Devaprasad K, Chen B, Ingley HA, Barooah P (2020) Model predictive control for energy-efficient hvac operation with humidity and latent heat considerations. *Applied Energy* 279:115765.

Rogerson WP (1985) Repeated moral hazard. *Econometrica: Journal of the Econometric Society* 69–76.

Salzman J, Bennett G, Carroll N, Goldstein A, Jenkins M (2018) The global status and trends of payments for ecosystem services. *Nature Sustainability* 1(3):136–144.

Sannikov Y (2008) A continuous- time version of the principal: Agent problem. *The Review of Economic Studies* 75(3):957–984.

Sannikov Y (2013) Contracts: The theory of dynamic principal—agent relationships and the continuous-time approach. *Advances in Economics and Econometrics: Volume 1, Economic Theory: Tenth World Congress*, volume 49, 89.

Santos MJ, Curcio E, Amorim P, Carvalho M, Marques A (2021) A bilevel approach for the collaborative transportation planning problem. *International Journal of Production Economics* 233:108004.

Schildbach G, Morari M (2016) Scenario-based model predictive control for multi-echelon supply chain management. *European Journal of Operational Research* 252(2):540–549.

Schneider R (2013) *Convex bodies: the Brunn–Minkowski theory*. Encyclopedia of Mathematics and its Applications (Cambridge University Press), 2 edition.

Schweppe FC (1967) Recursive state estimation: Unknown but bounded errors and system inputs. *Sixth Symposium on Adaptive Processes*, 102–107.

Seymour F, Harris NL (2019) Reducing tropical deforestation. *Science* 365(6455):756–757.

Shumway RH, Stoffer DS (1982) An approach to time series smoothing and forecasting using the em algorithm. *Journal of time series analysis* 3(4):253–264.

Shweta J, Sujit G (2020) A multiarmed bandit based incentive mechanism for a subset selection of customers for demand response in smart grids. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2046–2053.

Simchowitz M, Foster D (2020) Naive exploration is optimal for online LQR. *ICML*, volume 119.

Simchowitz M, Slivkins A (2021) Exploration and incentives in reinforcement learning. *arXiv preprint arXiv:2103.00360* .

Spear SE, Srivastava S (1987) On repeated moral hazard with discounting. *The Review of Economic Studies* 54(4):599–617.

Suen Sc, Negoescu D, Goh J (2022) Design of incentive programs for optimal medication adherence in the presence of observable consumption. *Operations Research* 70(3):1691–1716.

Sundadam RK, Banks JS (1991) Adverse Selection and Moral hazard in a Repeated Elections Models. Working Papers, University of Rochester - Center for Economic Research (RCER).

Thompson WR (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4):285–294.

Turkensteen M, Hasle G (2017) Combining pickups and deliveries in vehicle routing–an assessment of carbon emission effects. *Transportation Research Part C: Emerging Technologies* 80:117–132.

Van der Vaart AW (2000) *Asymptotic Statistics*, volume 3 (Cambridge University Press).

Vazquez S, Leon JI, Franquelo LG, Rodriguez J, Young HA, Marquez A, Zanchetta P (2014) Model predictive control: A review of its applications in power electronics. *IEEE industrial electronics magazine* 8(1):16–31.

Velarde P, Maestre J, Jurado I, Fernandez I, Tejera BI, del Prado J (2014) Application of robust model predictive control to inventory management in hospitalary pharmacy. *IEEE ETFA*.

Vera-Hernandez M (2003) Structural estimation of a principal-agent model: moral hazard in medical insurance. *RAND Journal of Economics* 670–693.

Vrabie D, Pastravanu OC, Abu-Khalaf M, Lewis F (2009) Adaptive optimal control for continuous-time linear systems based on policy iteration. *Autom.* 45.

Wabersich KP, Zeilinger MN (2018) Safe exploration of nonlinear dynamical systems: A predictive safety filter for reinforcement learning. *arXiv preprint arXiv:1812.05506* .

Wabersich KP, Zeilinger MN (2020) Performance and safety of bayesian model predictive control: Scalable model-based rl with guarantees. *arXiv preprint arXiv:2006.03483* .

Walmart (2023) Project Gigaton. Accessed February 1, 2024, `https://www.walmartsustainabilityhub.com/project-gigaton/`.

Wang Z, Gao L, Huang J (2022) Socially-optimal mechanism design for incentivized online learning. *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, 1828–1837.

Wang Z, Jungers RM, Ong CJ (2021) Computation of the maximal invariant set of discrete-time linear systems subject to a class of non-convex constraints. *Automatica* 125.

Warnes X, de Zegher JF, Iancu D, Plambeck E (2023) Area conditions and positive incentives: Engaging local communities to protect forests. *Available at SSRN 4609761* .

WHO (2003) *Adherence to long-term therapies: evidence for action* (World Health Organization).

Williams N (2015) A solvable continuous time dynamic principal–agent model. *Journal of Economic Theory* 159:989–1015.

Zhao Q, Shen Y, Li M (2015) Control and bidding strategy for virtual power plants with renewable generation and inelastic demand in electricity markets. *IEEE Transactions on Sustainable Energy* 7(2):562–575.

Zinkevich M (2003) Online convex programming and generalized infinitesimal gradient ascent. *ICML*, 928–936.

Appendices

Appendix A

Appendix for Chapter 2

A.1 Proofs of Theoretical Results

A.1.1 Results in Section 2.3

Proof of Theorem 2.1. We first prove the robust constraint satisfaction property. Since $\{u_{t|t}, \ldots, u_{t+N|t}\}$ are feasible for $V_N(x_t, \theta, t)$, then we have $\overline{x}_{t+1|t} = Ax_t + Bu_{t|t} \in \Omega \ominus \mathcal{W}$ by (2.6). By relating the true dynamics to the nominal model, the true next state is $x_{t+1} = \overline{x}_{t+1|t} + w_t$ for some $w_t \in \mathcal{W}$. This means $x_{t+1} \in (\Omega \ominus \mathcal{W}) \oplus \mathcal{W} \subseteq \Omega \subseteq \mathcal{X}$ where the last set inclusion follows from the constraint satisfaction property in the definition of Ω.

We next prove the robust feasibility property. By the definition (2.6) of $V_N(x_{t+1}, \theta', t+1)$, we have that $\bar{x}_{t+1|t+1} = x_{t+1}$. However, we just showed that $x_{t+1} \in \Omega$. Hence $\bar{x}_{t+1|t+1} \in \Omega$. Now set $u_{t+1|t+1} = Kx_{t+1}$, and note that the constraint satisfaction property of Ω means $u_{t+1|t+1} \in \mathcal{U}$. Since $\overline{x}_{t+2|t+1} = A\overline{x}_{t+1|t+1} + Bu_{t+1|t+1} = (A + BK)\overline{x}_{t+1|t+1}$, we have

$$\overline{x}_{t+2|t+1} \in (A + BK)\Omega \subseteq ((A + BK)\Omega \oplus \mathcal{W}) \ominus \mathcal{W} \subseteq \Omega \ominus \mathcal{W} \tag{A.1}$$

where the last set inclusion follows by the disturbance invariance property of Ω. So $\bar{x}_{t+2|t+1} \in \Omega \ominus \mathcal{W} \subseteq \Omega \subseteq \mathcal{X}$ by the constraint satisfaction property of Ω. We can sequentially repeat this argument with $u_{t+k+1|t+1} = Kx_{t+k+1|t+1}$ to show this choice results in $u_{t+k+1|t+1} \in \mathcal{U}$ and $x_{t+k+1|t+1} \in \mathcal{X}$ for $k \in [N-1]$. Thus $\{u_{t+1|t+1}, \ldots, u_{t+N|t+1}\}$ are feasible for $V_N(x_{t+1}, \theta', t+1)$. □

A.1.2 Results in Section 2.5

Proof of Theorem 2.2. Omitted. Refer to Section 3 of Mintz et al. (2017). □

A.1.3 Results in Section 2.6

Proof of Lemma 2.1. By Assumption 2.4, $f^{t+k}(x_t, u_{t|t}, \ldots, u_{t+k|t}, \widehat{\theta})$ is $L_{f,u}$-Lipschitz continuous and $h(\tilde{x}_{t+k|t}, u_{t+k|t}, \widehat{\theta}_t)$ is $L_{h,u}$-Lipschitz continuous with respect to $u_{t+k|t} \in \mathcal{U}$.

Then, by preservation of Lipschitz continuity across functional compositions and addition, we have the desired condition. □

Proof of Proposition 2.1. Let $s(\tau) = U^*_{N,t}(\widehat{\theta}_t) + \tau \cdot (U^*_{N,t}(\theta) - U^*_{N,t}(\widehat{\theta}_t))$. This implies $s(0) = U^*_{N,t}(\widehat{\theta}_t)$ and $s(1) = U^*_{N,t}(\theta)$. Then,

$$\begin{aligned}
&[J_N(x_t, U^*_{N,t}(\widehat{\theta}_t), \widehat{\theta}_t, t) - J_N(x_t, U^*_{N,t}(\widehat{\theta}_t), \theta, t)] - [J_N(x_t, U^*_{N,t}(\theta), \widehat{\theta}_t, t) - J_N(x_t, U^*_{N,t}(\theta), \theta, t)] \\
&= \int_0^1 \nabla_U J(x_t, s(\tau), \widehat{\theta}_t, t)^T (U^*_{N,t}(\theta) - U^*_{N,t}(\widehat{\theta}_t)) d\tau \\
&\quad - \int_0^1 \nabla_U J(x_t, s(\tau), \theta, t)^T (U^*_{N,t}(\theta) - U^*_{N,t}(\widehat{\theta}_t)) d\tau \qquad \text{(A.2)}
\end{aligned}$$

where the last equality follows by the Fundamental Theorem of Calculus for Line Integrals. Then, we continue as

$$\begin{aligned}
&= \left| \int_0^1 [\nabla_U J(x_t, s(\tau), \widehat{\theta}_t, t) - \nabla_U J(x_t, s(\tau), \theta, t)]^T (U^*_{N,t}(\theta) - U^*_{N,t}(\widehat{\theta}_t)) d\tau \right| \qquad \text{(A.3)} \\
&\leq \int_0^1 \left\| \nabla_U J(x_t, s(\tau), \widehat{\theta}_t, t) - \nabla_U J(x_t, s(\tau), \theta, t) \right\|_\infty \left\| U^*_{N,t}(\theta) - U^*_{N,t}(\widehat{\theta}_t) \right\|_1 d\tau \qquad \text{(A.4)} \\
&\leq L_J \left\| \widehat{\theta}_t - \theta \right\| \cdot \left\| U^*_{N,t}(\theta) - U^*_{N,t}(\widehat{\theta}_t) \right\|_1 \qquad \text{(A.5)} \\
&\leq \sqrt{N} L_J \left\| \widehat{\theta}_t - \theta \right\| \cdot \left\| U^*_{N,t}(\theta) - U^*_{N,t}(\widehat{\theta}_t) \right\|_2 \qquad \text{(A.6)}
\end{aligned}$$

where (A.4) follows by Hölder's inequality, and (A.5) follows by the assumed property in Proposition 2.1. This gives us the desired result in Assumption 2.6 by setting $\kappa = \sqrt{N} L_J$. □

Proof of Lemma 2.2. Since (2.13) is the average of compositions of two polynomials f and h, it is polynomial. Then, $\nabla_u J_N(x, U, \theta, t)$ is polynomial on the bounded domain $\mathcal{X} \times \mathcal{U}^{N+1} \times \Theta$.

Hence, by Corollary 8.2 in Estep (2010), $\nabla_u J_N(x, U, \theta, t)$ is locally Lipschitz with respect to $\theta \in \Theta$ for any $x \in \mathcal{X}, U \in \mathcal{U}^{N+1}, t \in \mathcal{T}$. □

Proof of Theorem 2.3. For notational convenience, let

$$\mathbb{E}[M_t] = h(x'_t, \Lambda^{O,N}(\mathcal{F}'_t), \theta_0) - h(x'_t, \Lambda^{\epsilon,N}(\mathcal{F}'_t), \theta_0) \qquad \text{(A.7)}$$

Let $\mathcal{T}^{\text{xit}} \in \mathcal{T}$ and $\mathcal{T}^{\text{xre}} \in \mathcal{T}$ be the set of random time points that Algorithm (1) performs exploitation and exploration, respectively. Noticing that the cardinalities $\#\mathcal{T}^{\text{xit}}$, $\#\mathcal{T}^{\text{xre}}$ are random variables, we have

$$\sum_{t=0}^{T} \mathbb{E}[M_t] = \textstyle\sum_{t\in\mathcal{T}^{\text{xit}}} h(x'_t, u^*_{t|t}(\theta_0), \theta_0) - h(x'_t, u^*_{t|t}(\widehat{\theta}_t), \theta_0) \\ + \sum_{t\in\mathcal{T}^{\text{xre}}} h(x'_t, u^*_{t|t}(\theta_0), \theta_0) - h(x'_t, u_{t|t}, \theta_0) \quad \text{(A.8)}$$

We note that $\mathbb{E}[M_t]$ is a bounded value since $\mathcal{X}, \Theta, \mathcal{U}$ are all compact sets and $h(x, u, \theta)$ is a bounded continuous function on this domain. Then, assuming $\mathbb{E}[M_t] \leq \mathcal{M}$, we obtain

$$\left[\sum_{t=0}^{T} \mathbb{E}[M_t] \middle| \mathcal{T}^{\text{xit}}\right] \leq \mathcal{M}\mathbb{E}[\#\mathcal{T}^{\text{xre}}] + \sum_{t\in\mathcal{T}^{\text{xit}}} h(x'_t, u^*_{t|t}(\theta_0), \theta_0) - h(x'_t, u^*_{t|t}(\widehat{\theta}_t), \theta_0) \quad \text{(A.9)}$$

We can rewrite each term inside the summation above as

$$\begin{aligned} & h(x'_t, u^*_{t|t}(\theta_0), \theta_0) - h(x'_t, u^*_{t|t}(\widehat{\theta}_t), \theta_0) \\ &= \mathbb{E}[M_t | D_{\Pi_t}(\theta_0 || \widehat{\theta}_t) \leq \delta_{\widehat{\theta}_t}, x_t, \theta_0, \widehat{\theta}_t] P(D_{\Pi_t}(\theta_0 || \widehat{\theta}_t) \leq \delta_{\widehat{\theta}_t}) \\ &\quad + \mathbb{E}[M_t | D_{\Pi_t}(\theta_0 || \widehat{\theta}_t) \geq \delta_{\widehat{\theta}_t}, x_t, \theta_0, \widehat{\theta}_t] P(D_{\Pi_t}(\theta_0 || \widehat{\theta}_t) \geq \delta_{\widehat{\theta}_t}) \\ &= (13, a) + (13, b) \end{aligned} \quad \text{(A.10)}$$

Let $\varepsilon(\delta_{\widehat{\theta}_t}) = \max\{\|\theta_0 - \widehat{\theta}_t\| : D_{\Pi_t}(\theta_0 || \widehat{\theta}_t) \leq \delta_{\widehat{\theta}_t}\}, \forall t \in \mathcal{T}^{\text{xit}}$.

$$\begin{aligned} \sum_{t\in\mathcal{T}^{\text{xit}}} (13, a) &= \sum_{t\in\mathcal{T}^{\text{xit}}} h(x'_t, u^*_{t|t}(\theta_0), \theta_0) - h(x'_t, u^*_{t|t}(\widehat{\theta}_t), \theta_0) & \text{(A.11)} \\ &\leq \sum_{t\in\mathcal{T}^{\text{xit}}} L_{h,u} \left\| u^*_{t|t}(\theta_0) - u^*_{t|t}(\widehat{\theta}_t) \right\| & \text{(A.12)} \\ &\leq \sum_{t\in\mathcal{T}^{\text{xit}}} L_{h,u} \left\| U^*_{N|t}(\theta_0) - U^*_{N|t}(\widehat{\theta}_t) \right\| & \text{(A.13)} \\ &\leq \frac{L_{h,u}\kappa}{c_u} \sum_{t\in\mathcal{T}^{\text{xit}}} \left\| \theta_0 - \widehat{\theta}_t \right\| & \text{(A.14)} \\ &\leq \frac{L_{h,u}\kappa}{c_u} \sum_{t\in\mathcal{T}^{\text{xit}}} \varepsilon(\delta_{\widehat{\theta}_t}) & \text{(A.15)} \end{aligned}$$

where (A.12) follows by Assumption 2.4 and (A.14) follows by Lemma 2.3. Now, we have $\varepsilon(\delta_{\widehat{\theta}_t}) = C\delta_{\widehat{\theta}_t}$ for a constant $C > 0$ by Assumption 2.5 and let $\eta(t) = |\{s \in \mathcal{T}^{\text{xit}} : s \leq t\}|$. Then, for $\delta_{\widehat{\theta}_t} = \sqrt{4L_{\ell,r}^2\sigma^2}\log\eta(t)/\sqrt{\eta(t)}$, we obtain

$$\leq \frac{L_{h,u}\kappa C\sqrt{4L_{\ell,r}^2\sigma^2}}{c_u}\sqrt{\#\mathcal{T}^{\text{xit}}}\log\#\mathcal{T}^{\text{xit}} \tag{A.16}$$

$$\leq \frac{L_{h,u}\kappa C\sqrt{4L_{\ell,r}^2\sigma^2}}{c_u}\sqrt{T}\log T \tag{A.17}$$

To bound the second term in (A.10), recall $\mathbb{E}[M_t] \leq \mathcal{M}$. Then,

$$\sum_{t\in\mathcal{T}^{\text{xit}}}(13,b) \leq \mathcal{M}\sum_{t\in\mathcal{T}^{\text{xit}}}\exp\left(\frac{-(\delta_{\widehat{\theta}_t}\sqrt{t-1}-c_f(d_x,d_\theta))^2}{2L_{\ell,r}^2\sigma^2}\right) \tag{A.18}$$

$$\leq \mathcal{M}\sum_{t\in\mathcal{T}^{\text{xit}}}\exp\left(\frac{-\delta_{\widehat{\theta}_t}^2(t-1)/2+c_f^2(d_x,d_\theta)}{2L_{\ell,r}^2\sigma^2}\right) \tag{A.19}$$

$$\leq \mathcal{M}\exp\left(\frac{c_f^2(d_x,d_\theta)}{2L_{\ell,r}^2\sigma^2}\right)\left(\sum_{t=1}^{9}\exp(-(\log t)^2\right)+\sum_{t\in\mathcal{T}^{\text{xit}},t\geq 10}\exp(-\log t)) \tag{A.20}$$

$$\leq \mathcal{M}\exp\left(\frac{c_f^2(d_x,d_\theta)}{2L_{\ell,r}^2\sigma^2}\right)(\mathcal{C}+\log T) \tag{A.21}$$

where (A.18) follows by Theorem 2.2 and $\mathcal{C}$ can be approximated as 2.2232. Lastly, we bound the first term in (A.9): $\mathcal{M}\mathbb{E}[\#\mathcal{T}^{\text{xre}}] = \mathcal{M}\sum_{t=0}^{T}\min\{1,\frac{c}{t}\} \leq \mathcal{M}(c+\sum_{t=c+1}^{T}\frac{c}{t}) \leq \mathcal{M}c(1-\log(c+1)+\log T)$. Substituting these into (A.9):

$$\left[\sum_{t=0}^{T}\mathbb{E}[M_t]\middle|\mathcal{T}^{\text{xit}}\right] \leq \mathcal{M}\exp\left(\frac{c_f^2(d_x,d_\theta)}{2L_{\ell,r}^2\sigma^2}\right)(\mathcal{C}+\log T)$$
$$+\mathcal{M}c(1-\log(c+1)+\log T)+\frac{L_{h,u}\kappa C\sqrt{4L_{\ell,r}^2\sigma^2}}{c_u}\sqrt{T}\log T$$

and taking the expectation gives us the desired result. □

Proof of Theorem 2.4. By Assumption 2.4 and the upper bound in (2.17),

$$R_{N,T} = \sum_{t=0}^{T}h(x_t,\Lambda_t^{O,N}(\mathcal{F}_t),\theta_0)-h(x_t',\Lambda_t^{O,N}(\mathcal{F}_t'),\theta_0)$$
$$+\sum_{t=0}^{T}h(x_t',\Lambda_t^{O,N}(\mathcal{F}_t'),\theta_0)-h(x_t',\Lambda_t^{\epsilon,N}(\mathcal{F}_t'),\theta_0)$$

$$\leq L_{h,x} \sum_{t=0}^{T} \|x_t - x_t'\| \; + \; (2.17) \tag{A.22}$$

Algorithm (1) performs exploration at random times according to a non-stationary stochastic process over $\mathcal{T}$. We divide $\mathcal{T}$ into "inter-explore intervals" composed of an exploration and the subsequent exploitations until the next one is reached. Let $I_k = [\underline{I}_k, \overline{I}_k]$ be the k^{th} subinterval such that $I_{-1} = [0, 2\lceil\sqrt{T}\rceil]$, $I_0 = [2\lceil\sqrt{T}\rceil + 1, t_1^{\text{xre}} - 1]$, $I_k = [t_k^{\text{xre}}, t_{k+1}^{\text{xre}} - 1]$ for $k \in [1, K-1]$ where t_k^{xre} is the k^{th} exploration step after time $2\lceil\sqrt{T}\rceil$, and $I_K = [t_K^{\text{xre}}, T]$ where $K = \sum_{t=2\lceil\sqrt{T}\rceil+1}^{T} s_t$ and $s_t \sim \text{Bernoulli}\left(\min\left\{1, {}^{c}\!/_{t}\right\}\right)$. Then, $\sum_{t=0}^{T} \|x_t - x_t'\| = \sum_{k=-1}^{K} \sum_{t \in I_k} \|x_t - x_t'\|$. The key idea is that regret over each $I_k, k \in [0, K]$ is bounded above by the regret over $S_k = [\underline{S}_k, \overline{S}_k] = [\underline{I}_k, T]$ that includes a single exploration at time $\underline{I}_k$ followed by exploitation steps thereafter up to T.

Suppose Algorithm 1 uses $\Lambda_t^{O,N}(\mathcal{F}_t)$ at all greedy exploitation steps of $S_k, k \in [0, K]$. Since $x_t \in \mathcal{X}$ for $t \in \mathcal{T}^{\text{xre}}$ and $\mathcal{X}$ is compact, $\|x_t - x_{eq}\| \leq \text{diam}(\mathcal{X}),\ t \in \mathcal{T}^{\text{xre}}$. Then, by Assumption 2.7,

$$\sum_{t \in I_k} \|x_t - x_{eq}\| \leq \sum_{t \in S_k} \|x_t - x_{eq}\| = \|x_{t_k^{\text{xre}}} - x_{eq}\| + \sum_{t=\underline{S}_k+1}^{\overline{S}_k} \|x_t - x_{eq}\| \tag{A.23}$$

$$\leq \text{diam}(\mathcal{X}) + \sum_{t=\underline{S}_k+1}^{\overline{S}_k} \alpha^{t-\underline{S}_k} \text{diam}(\mathcal{X}) \leq \frac{\text{diam}(\mathcal{X})}{(1-\alpha)} \tag{A.24}$$

Next, suppose instead $\Lambda_t^{E,N}(\mathcal{F}_t')$ is used at all greedy exploitation steps of $S_k, k \in [0, K]$. Observe the convergence of $\Lambda^{E,N}(\mathcal{F}_t')$:

$$\begin{aligned}
&\left\| Ax_t' + B\Lambda_t^{E,N}(\mathcal{F}_t') + g(x_t', \Lambda_t^{E,N}(\mathcal{F}_t'), \theta_0) - x_{eq} \right\| \\
&\leq \left\| Ax_t' + B\Lambda_t^{O,N}(\mathcal{F}_t') + g(x_t', \Lambda_t^{O,N}(\mathcal{F}_t'), \theta^o) - x_{eq} \right\| \\
&\quad + \left\| B\Lambda_t^{E,N}(\mathcal{F}_t') + g(x_t', \Lambda_t^{E,N}(\mathcal{F}_t'), \theta_0) - B\Lambda_t^{O,N}(\mathcal{F}_t') - g(x_t', \Lambda_t^{O,N}(\mathcal{F}_t'), \theta_0) \right\| && \text{(A.25)} \\
&\leq \alpha\|x_t' - x_{eq}\| + \left\| B\Lambda_t^{E,N}(\mathcal{F}_t') + g(x_t', \Lambda_t^{E,N}(\mathcal{F}_t'), \theta_0) - B\Lambda_t^{O,N}(\mathcal{F}_t') - g(x_t', \Lambda_t^{O,N}(\mathcal{F}_t'), \theta_0) \right\| && \text{(A.26)}
\end{aligned}$$

where (A.25) follows by the triangle inequality and (A.26) follows by Assumption 2.7. Recall that $\|\Lambda_t^{E,N}(\mathcal{F}_t') - \Lambda_t^{O,N}(\mathcal{F}_t')\| \leq \frac{\kappa C\sqrt{4L_{\ell,r}^2\sigma^2}}{c_u} \frac{\log\eta(t)}{\sqrt{\eta(t)}}$ as followed from (A.12) to (A.16). By Assumption 2.4, we have $(A.26) \leq \alpha\|x_t' - x_{eq}\| + \overline{C}\frac{\log\eta(t)}{\sqrt{\eta(t)}}$, where $\overline{C} = \frac{(\|B\|+L_{f,u})\kappa C\sqrt{4L_{\ell,r}^2\sigma^2}}{c_u}$. Then, for $k \in [0, K]$,

$$\sum_{t \in I_k} \|x_t' - x_{eq}\| \leq \sum_{t \in S_k} \|x_t' - x_{eq}\| \tag{A.27}$$

$$= \|x'_{t_k^{\text{xre}}} - x_{eq}\| + \sum_{t=\underline{S}_k+1}^{\overline{S}_k} \|x'_t - x_{eq}\| \tag{A.28}$$

$$\leq \text{diam}(\mathcal{X}) + \sum_{t=t_k^{\text{xre}}+1}^{T} \left[\alpha^{t-t_k^{\text{xre}}} \text{diam}(\mathcal{X}) + \overline{C} \sum_{i=t_k^{\text{xre}}+1}^{t-1} \frac{\alpha^{t-1-i} \log \eta(i)}{\sqrt{\eta(i)}}\right] \tag{A.29}$$

$$\leq \frac{2-\alpha}{1-\alpha} \text{diam}(\mathcal{X}) + \overline{C} \sum_{t=t_k^{\text{xre}}+1}^{T} \sum_{i=t_k^{\text{xre}}+1}^{t-1} \alpha^{t-1-i} \frac{\log \eta(i)}{\sqrt{\eta(i)}} \tag{A.30}$$

Recall $\eta(i) = i - \sum_{j=1}^{i} s_j$ where $s_j \sim \text{Bernoulli}(\min\{1, c/j\})$ and $\mathbb{E}\sum_{j=1}^{i} s_j \leq c + \int_{j=c}^{i} \frac{c}{j} dj = c \log \frac{ei}{c}$. By conditioning on the event $\mathcal{E}_i = \{\sum_{j=1}^{i} s_j \leq 3\mathbb{E}\sum_{j=1}^{i} s_j\}$, we get $\eta(i) \geq i - 3\mathbb{E}\sum_{j=1}^{i} s_j \geq i - 3c \log \frac{ei}{c} \geq \frac{i}{c^2}$ where the last inequality holds for all $i \geq 2\lceil\sqrt{T}\rceil + 1 \geq 6c^2 + 1$. Then, for $\alpha \in [0, 2/3]$,

$$\leq \frac{2-\alpha}{1-\alpha} \text{diam}(\mathcal{X}) + \overline{C}c \sum_{t=t_k^{\text{xre}}+1}^{T} \sum_{i=t_k^{\text{xre}}+1}^{t-1} \alpha^{t-i-1} \frac{\log i}{\sqrt{i}} \tag{A.31}$$

$$\leq \frac{2-\alpha}{1-\alpha} \text{diam}(\mathcal{X}) + \frac{\overline{C}c}{\alpha} \sum_{t=t_k^{\text{xre}}+1}^{T} \sum_{i=t_k^{\text{xre}}+1}^{t-1} \frac{\log i}{(t-i)\sqrt{i}} \tag{A.32}$$

$$\leq \frac{2-\alpha}{1-\alpha} \text{diam}(\mathcal{X}) + \frac{2\overline{C}c}{\alpha} \sum_{t=t_k^{\text{xre}}+1}^{T} \frac{(\log t)^2}{\sqrt{t}} \tag{A.33}$$

$$\leq \frac{2-\alpha}{1-\alpha} \text{diam}(\mathcal{X}) + \frac{2\overline{C}c}{\alpha} \sqrt{T}(\log T)^2 \tag{A.34}$$

Note $\mathbb{E}\sum_{j=1}^{i} s_j \geq c \log \frac{e(i+1)}{c+1}$ and $\text{Var}(\sum_{j=1}^{i} s_j) = \sum_{j=1}^{i} \frac{c}{j} \cdot \frac{j-c}{j} \leq c \log i + \frac{c^2}{i}$. Then, by Bernstein's inequality (Corollary 2.11 in Boucheron et al. (2013)), it follows that (A.34) holds with probability

$$\mathbb{P}\Big(\bigcap_{i=t_k^{\text{xre}}+1}^{T} \mathcal{E}_i\Big) \geq 1 - \sum_{i=t_k^{\text{xre}}+1}^{T} P(\overline{\mathcal{E}}_i) \tag{A.35}$$

$$\geq 1 - \sum_{i=t_k^{\text{xre}}+1}^{T} \exp\left(-\frac{4c^2\big(\log\frac{e(i+1)}{c+1}\big)^2}{2c\log i + \frac{2c^2}{i} + \frac{4c^2}{3}\log\frac{e(i+1)}{c+1}}\right) \tag{A.36}$$

$$\geq 1 - (T - 2\sqrt{T}) \exp\left(-\frac{4c^2\big(\log\frac{e(2\sqrt{T}+2)}{c+1}\big)^2}{2c\log(2\sqrt{T}+1) + \frac{2c^2}{2\sqrt{T}+1} + \frac{4c^2}{3}\log\frac{e(2\sqrt{T}+2)}{c+1}}\right) \tag{A.37}$$

The above bounds for $\Lambda^{O,N}(\mathcal{F}_t)$ and (A.34) for $\Lambda^{\epsilon,N}(\mathcal{F}'_t)$ allow us to bound the deviation of the system trajectory under the learning policy from the one under the oracle policy over I_{-1} as $\sum_{t\in I_{-1}} \|x_t - x'_t\| \leq 2\sqrt{T}\mathrm{diam}(\mathcal{X})$ and over $I_k, k \geq 0$ as

$$\sum_{t\in I_k} \|x_t - x'_t\| \leq \sum_{t\in S_k} \|x_t - x_{eq}\| + \sum_{t\in S_k} \|x'_t - x_{eq}\| \leq \frac{3-\alpha}{1-\alpha}\mathrm{diam}(\mathcal{X}) + \frac{2\overline{C}c}{\alpha}\sqrt{T}(\log T)^2 \tag{A.38}$$

Combining this with (A.22), we obtain

$$R_{N,T} \leq 2L_{h,x}\sqrt{T}\mathrm{diam}(\mathcal{X}) + L_{h,x}K\Big(\frac{3-\alpha}{1-\alpha}\mathrm{diam}(\mathcal{X}) + \frac{2\overline{C}\sqrt{c}}{\alpha}\sqrt{T}(\log T)^2\Big) + (2.17) \tag{A.39}$$

and it remains to bound K. Note $\mathbb{E}K \geq c\log\frac{T+1}{2\sqrt{T}+1}$, $\mathrm{Var}(K) \leq 2\log T$, and Bernstein's inequality yields $\mathbb{P}(K \leq 2\mathbb{E}K) \geq 1 - \exp(-\frac{c^2(\log\frac{T}{2\sqrt{T}+1})^2}{(4+\frac{2}{3}c^2)\log T})$. Bounding K by $2\mathbb{E}K \leq 2c\log T$ gives the desired result. □

Appendix B

Appendix for Chapter 3

B.1 Proofs of Theoretical Results

B.1.1 Results in Section 3.2

Proof of Proposition 3.1. We first note that $\ell\left(\mathbf{s}, i_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) = +\infty$ is obtained when the action selected by the agent (the maximizer of $\mathbf{s}^0 + \boldsymbol{\pi}_t$) is not the same as the maximizer of $\mathbf{s} + \boldsymbol{\pi}_t$. Since now we consider the case that $K^0 \cap K = \emptyset$, we already observe different indices for the largest entries of the true normalized rewards $\mathbf{s}^0$ and the considered normalized rewards $\mathbf{s}$ before adding the incentives. Hence, we can observe the desired event ($\ell\left(\mathbf{s}, i_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) = +\infty$) by simply choosing the incentive amounts in such a way that the new maximizers after adding the incentives will still belong to the sets K and K^0. Suppose we have

$$\pi_{t,a} < R_{\min} + \gamma + \beta - d \text{ for all } a \in \mathcal{A} \setminus \{\kappa, \kappa^0\} \tag{B.1}$$

$$\pi_{t,a} \geq R_{\min} + \gamma + \beta - d \text{ for } a \in \{\kappa, \kappa^0\} \tag{B.2}$$

Note that (B.1) and (B.2) are valid conditions according to Assumption 3.1. Now, recall that a vector $\mathbf{s} \in \mathcal{B}(\mathbf{s}^j, d)$ satisfies $\|\mathbf{s}^0 - \mathbf{s}\|_\infty > \beta$ by definition. We define $\widetilde{\mathbf{s}}^j := \arg\inf_{\mathbf{s} \in \mathcal{B}(\mathbf{s}^j, d)} \|\mathbf{s}^0 - \mathbf{s}\|_\infty$ as the closest vector (with respect to the ℓ_∞-norm) in ball $\mathcal{B}(\mathbf{s}^j, d)$ to the true reward vector $\mathbf{s}^0$. Then, we have $\|\mathbf{s}^0 - \widetilde{\mathbf{s}}^j\|_\infty \geq \beta - d$ by construction, and it follows that

$$\begin{aligned}
&\mathbb{P}\left(\ell\left(\mathbf{s}, i_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) = +\infty\right) \\
&\geq \mathbb{P}\left(\bigcup_{x \in \mathcal{A}, y \in \mathcal{A}, y \neq x} x = \arg\max_{a \in \mathcal{A}} \left(s_a^0 + \pi_{t,a}\right), y = \arg\max_{a \in \mathcal{A}} \left(s_a + \pi_{t,a}\right)\right) && \text{(B.3)} \\
&\geq \mathbb{P}\left(\kappa = \arg\max_{a \in \mathcal{A}}(s_a + \pi_{t,a}), \kappa^0 = \arg\max_{a \in \mathcal{A}}(s_a^0 + \pi_{t,a})\right) \text{ for any } \kappa \in K,\ \kappa^0 \in K^0 && \text{(B.4)} \\
&\geq \mathbb{P}\left(\kappa = \arg\max_{a \in \mathcal{A}}(s_a + \pi_{t,a}), \kappa^0 = \arg\max_{a \in \mathcal{A}}(s_a^0 + \pi_{t,a}) \middle| (B.1), (B.2)\right) \mathbb{P}\left((B.1), (B.2)\right) && \text{(B.5)}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{P}\left(s_{\kappa^0} - s_{\kappa} < \pi_{t,\kappa} - \pi_{t,\kappa^0} < s^0_{\kappa^0} - s^0_{\kappa}\right) \cdot \prod_{a \in \{\kappa, \kappa^0\}} \mathbb{P}\left(\pi_{t,a} \geq R_{\min} + \gamma + \beta - d\right) \\
&\qquad \cdot \prod_{a \in \mathcal{A} \setminus \{\kappa, \kappa^0\}} \mathbb{P}\left(\pi_{t,a} < R_{\min} + \gamma + \beta - d\right) && \text{(B.6)} \\
&\geq \mathbb{P}\left(s_{\kappa^0} - s_{\kappa} < \pi_{t,\kappa} - \pi_{t,\kappa^0} < s^0_{\kappa^0} - s^0_{\kappa}\right) \cdot \prod_{a \in \{\kappa, \kappa^0\}} \mathbb{P}\left(\pi_{t,a} \geq R_{\min} + \gamma + \beta - d\right) \\
&\qquad \cdot \prod_{a \in \mathcal{A} \setminus \{\kappa, \kappa^0\}} \mathbb{P}\left(\pi_{t,a} \leq R_{\min} + \gamma\right) && \text{(B.7)} \\
&= \mathbb{P}\left(s_{\kappa^0} - s_{\kappa} < \pi_{t,\kappa} - \pi_{t,\kappa^0} < s^0_{\kappa^0} - s^0_{\kappa}\right) \cdot \prod_{a \in \{\kappa, \kappa^0\}} \left(1 - \frac{R_{\min} + \gamma + \beta - d - \underline{C}}{\overline{C} - \underline{C}}\right) \\
&\qquad \cdot \prod_{a \in \mathcal{A} \setminus \{\kappa, \kappa^0\}} \frac{R_{\min} + \gamma - \underline{C}}{\overline{C} - \underline{C}} && \text{(B.8)} \\
&= \mathbb{P}\left(s_{\kappa^0} - s_{\kappa} < \pi_{t,\kappa} - \pi_{t,\kappa^0} < s^0_{\kappa^0} - s^0_{\kappa}\right) \prod_{a \in \{\kappa, \kappa^0\}} \left(1 - \frac{\gamma + \beta - d}{\overline{C} - \underline{C}}\right) \prod_{a \in \mathcal{A} \setminus \{\kappa, \kappa^0\}} \frac{\gamma}{\overline{C} - \underline{C}} && \text{(B.9)}
\end{aligned}
$$

where (B.6) follows since $\pi_{t,a}$'s are considered to be independent random variables, (B.8) follows since $\pi_{t,a} \sim \mathcal{U}(\underline{C}, \overline{C}), \forall a \in \mathcal{A}$, and (B.9) follows since $\underline{C} = R_{\min}$ by Assumption 3.1. For the first term in (B.9), notice that the case that $s_{\kappa^0} - s_{\kappa} = s^0_{\kappa^0} - s^0_{\kappa} = 0$ cannot occur. This can only happen if $\kappa^0 \in K$ and $\kappa \in K^0$ which contradicts with the condition $K^0 \cap K_t = \emptyset$. Similarly, $\mathbf{s}^0$ cannot be the all-zeros vector under the given condition $K^0 \cap K_t = \emptyset$. Thus, the following always holds under the given condition: $s^0_{\kappa^0} - s^0_{\kappa} > 0$, $s_{\kappa^0} - s_{\kappa} < 0$, and $\mathbf{s}^0 \neq \mathbf{0}_n$. Then, we obtain

$$
(B.9) \geq \mathbb{P}\left(0 \leq \pi_{t,\kappa} - \pi_{t,\kappa^0} < s^0_{\kappa^0} - s^0_{\kappa}\right) \left(1 - \frac{\gamma + \beta - d}{\overline{C} - \underline{C}}\right)^2 \left(\frac{\gamma}{\overline{C} - \underline{C}}\right)^{n-2} \tag{B.10}
$$

The probability term in the last inequality can be computed by using the cumulative distribution function (cdf) of $\pi_{t,a} - \pi_{t,a'}$ – which is the difference of two identically and independently distributed (iid) Uniform random variables. The difference $\pi_{t,a} - \pi_{t,a'}$ follows a triangular distribution whose cdf can be explicitly computed as follows.

$$
\mathbb{P}\left(\pi_{t,a} - \pi_{t,a'} \leq \Delta\right) = \left\{
\begin{array}{ll}
0, & \text{for } \Delta < \underline{C} - \overline{C} \\
\int\limits_{\underline{C}}^{\overline{C}+\Delta} \int\limits_{\pi_{t,a}-\Delta}^{\overline{C}} \frac{1}{(\overline{C}-\underline{C})^2} d\pi_{t,a} d\pi_{t,a'}, & \text{for } \underline{C} - \overline{C} \leq \Delta < 0 \\
1 - \int\limits_{\underline{C}+\Delta}^{\overline{C}} \int\limits_{\underline{C}}^{\pi_{t,a}-\Delta} \frac{1}{(\overline{C}-\underline{C})^2} d\pi_{t,a} d\pi_{t,a'}, & \text{for } 0 \leq \Delta \leq \overline{C} - \underline{C} \\
1, & \text{for } \Delta \geq \overline{C} - \underline{C}
\end{array}
\right\} \tag{B.11}
$$

$$= \left\{ \begin{array}{ll} 0, & \text{for } \Delta < \underline{C} - \overline{C} \\ \frac{(\Delta + \overline{C} - \underline{C})^2}{2(\overline{C} - \underline{C})^2}, & \text{for } \underline{C} - \overline{C} \leq \Delta < 0 \\ 1 - \frac{\left(\Delta + \underline{C} - \overline{C}\right)^2}{2(\overline{C} - \underline{C})^2}, & \text{for } 0 \leq \Delta \leq \overline{C} - \underline{C} \\ 1, & \text{for } \Delta \geq \overline{C} - \underline{C} \end{array} \right\} \tag{B.12}$$

Since by construction we have $\mathcal{R} \subseteq [\underline{C}, \overline{C}]$, we know that $0 < s^0_{\kappa^0} - s^0_{\kappa} \leq \overline{C} - \underline{C}$ holds. Thus, we have

$$\mathbb{P}\left(0 \leq \pi_{t,\kappa} - \pi_{t,\kappa^0} < s^0_{\kappa^0} - s^0_{\kappa}\right) = 1 - \frac{\left(s^0_{\kappa^0} - s^0_{\kappa} + \underline{C} - \overline{C}\right)^2}{2(\overline{C} - \underline{C})^2} - \frac{1}{2} \tag{B.13}$$

$$= \frac{1}{2} - \frac{\left(\overline{C} - \underline{C} - s^0_{\kappa^0} + s^0_{\kappa}\right)^2}{2(\overline{C} - \underline{C})^2} > 0 \tag{B.14}$$

Combining this last result with (B.10), we obtain the desired result and conclude. □

Proof of Proposition 3.2. Recall that the event $\ell\left(\mathbf{s}, i_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) = +\infty$ is observed when the maximizer entries of the total reward vectors $\mathbf{s}^0 + \boldsymbol{\pi}_t$ and $\mathbf{s} + \boldsymbol{\pi}_t$ are different from each other. Hence, to prove the lower bound in (3.9), we will consider the case when $\arg\max_{a \in \mathcal{A}}\left(s_a + \pi_{t,a}\right) = 1$ and $\arg\max_{a \in \mathcal{A}}\left(s^0_a + \pi_{t,a}\right) = b$ because we know that $b \neq 1$. As we have $s_1 = s^0_1 = 0$ by construction, having $b = 1$ would imply that $\mathbf{s}^0 = \mathbf{s} = \mathbf{0}_n$. However, this contradicts with the fact that $\mathbf{s} \in \mathcal{B}(\mathbf{s}^j, d) \in \mathcal{F}$ which means $\|\mathbf{s}^0 - \mathbf{s}\|_\infty = |s^0_b - s_b| > \beta$ must be satisfied.

With this consideration, let $\omega = \sup_{\mathbf{s} \in \mathcal{B}(\mathbf{s}^j, d)} \max_{a \in \mathcal{A}}\{|s^0_a|, |s_a|\}$ be the largest absolute value observed among the entries of $\mathbf{s}^0$ and of all vectors in $\mathcal{B}(\mathbf{s}^j, d)$. Then, suppose we have

$$\pi_{t,a} < R_{\min} + \gamma + \beta - d \text{ for all } a \in \mathcal{A} \setminus \{1, b\} \tag{B.15}$$

$$\pi_{t,a} \geq R_{\min} + \gamma + \omega \text{ for } a \in \{1, b\}. \tag{B.16}$$

Note that (B.15) and (B.16) are consistent with Assumption 3.1. Further, they imply that the indices in the sets K^0 and K are no more maximizers after adding the incentives in (B.15)-(B.16). To restate, we now have $s^0_{\kappa^0} + \pi_{t,\kappa^0} < s^0_a + \pi_{t,a}$ and $s_\kappa + \pi_{t,\kappa} < s_a + \pi_{t,a}$ for any $\kappa^0 \in K^0$, $\kappa \in K$, $a \in \{1, b\}$. Further, if the events $s^0_1 + \pi_{t,1} < s^0_b + \pi_{t,b}$ and $s_b + \pi_{t,b} < s_1 + \pi_{t,1}$ also hold, then we will obtain the desired case (that is $\arg\max_{a \in \mathcal{A}}(s_a + \pi_{t,a}) = 1$ and $\arg\max_{a \in \mathcal{A}}(s^0_a + \pi_{t,a}) = b$). Our proof will be based on this observation.

Since $|s^0_b - s_b| > \beta$ by definition, we know that $|s^0_b - s_b| > |s^0_1 - s_1| = 0$. Suppose that without loss of generality, we have $s^0_b - s_b > s^0_1 - s_1 = 0$ and $s^0_b - s_b > \beta$. Then, we get

$$\mathbb{P}\left(\ell\left(\mathbf{s}, i_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) = +\infty\right)$$

$$\geq \mathbb{P}\left(\bigcup_{x \in \mathcal{A}, y \in \mathcal{A}, y \neq x} x = \arg\max_{a \in \mathcal{A}}\left(s^0_a + \pi_{t,a}\right), y = \arg\max_{a \in \mathcal{A}}\left(s_a + \pi_{t,a}\right)\right) \tag{B.17}$$

$$\geq \mathbb{P}\left(1 = \arg\max_{a\in\mathcal{A}}(s_a + \pi_{t,a}), b = \arg\max_{a\in\mathcal{A}}(s_a^0 + \pi_{t,a})\right) \tag{B.18}$$

$$\geq \mathbb{P}\left(1 = \arg\max_{a\in\mathcal{A}}(s_a + \pi_{t,a}), b = \arg\max_{a\in\mathcal{A}}(s_a^0 + \pi_{t,a}) \middle| (B.15), (B.16)\right) \mathbb{P}\left((B.15), (B.16)\right) \tag{B.19}$$

$$= \mathbb{P}\left(s_b - s_1 < \pi_{t,1} - \pi_{t,b} < s_b^0 - s_1^0\right) \cdot \prod_{a\in\{1,b\}} \mathbb{P}\left(\pi_{t,a} \geq R_{\min} + \gamma + \omega\right) \cdot \prod_{a\in\mathcal{A}\setminus\{1,b\}} \mathbb{P}\left(\pi_{t,a} < R_{\min} + \gamma + \beta - d\right) \tag{B.20}$$

$$\geq \mathbb{P}\left(s_b - s_1 < \pi_{t,1} - \pi_{t,b} < s_b^0 - s_1^0\right) \cdot \prod_{a\in\{1,b\}} \mathbb{P}\left(\pi_{t,a} \geq R_{\min} + \gamma + \omega\right) \cdot \prod_{a\in\mathcal{A}\setminus\{1,b\}} \mathbb{P}\left(\pi_{t,a} \leq R_{\min} + \gamma\right) \tag{B.21}$$

$$= \mathbb{P}\left(s_b - s_1 < \pi_{t,1} - \pi_{t,b} < s_b^0 - s_1^0\right) \prod_{a\in\{1,b\}} \left(1 - \frac{\gamma + \omega}{\overline{C} - \underline{C}}\right) \prod_{a\in\mathcal{A}\setminus\{1,b\}} \frac{\gamma}{\overline{C} - \underline{C}} \tag{B.22}$$

where (B.21) and (B.22) follow since $\underline{C} = R_{\min}$ by Assumption 3.1 and $\pi_{t,a}$'s are independent random variables with $\pi_{t,a} \sim \mathcal{U}(\underline{C}, \overline{C}), \forall a \in \mathcal{A}$.

We next compute a lower bound for the first term in (B.22).

$$\mathbb{P}\left(s_b - s_1 < \pi_{t,1} - \pi_{t,b} < s_b^0 - s_1^0\right) = \mathbb{P}\left(s_b - s_1 + s_b^0 - s_b^0 < \pi_{t,1} - \pi_{t,b} < s_b^0 - s_1^0\right) \tag{B.23}$$

$$\geq \mathbb{P}\left(s_b^0 - \beta - s_1 < \pi_{t,1} - \pi_{t,b} < s_b^0 - s_1^0\right) \tag{B.24}$$

$$= \mathbb{P}\left(s_b^0 - \beta < \pi_{t,1} - \pi_{t,b} < s_b^0\right) \tag{B.25}$$

We can compute the probability in the last line above by using the cdf derived in (B.12). Since the cdf is a piecewise function, we need to consider the two disjoint cases given as:

- *Case 1:* $\underline{C} - \overline{C} \leq s_b^0 < 0$
- *Case 2:* $0 \leq s_b^0 \leq \overline{C} - \underline{C}$

We also consider the following subcases to derive the probability bounds for the two cases above.

- *Subcase 1:* $\underline{C} - \overline{C} \leq s_b^0 < 0$ and $\underline{C} - \overline{C} \leq s_b^0 - \beta < 0$
- *Subcase 2:* $0 \leq s_b^0 \leq \overline{C} - \underline{C}$ and $0 \leq s_b^0 - \beta \leq \overline{C} - \underline{C}$

We can bound $\mathbb{P}\left(s_b^0 - \beta < \pi_{t,1} - \pi_{t,b} < s_b^0\right)$ from below under *Subcase 1* and *Subcase 2* as follows.

$$\mathbb{P}\left(s_b^0 - \beta \leq \pi_{t,1} - \pi_{t,b} < s_b^0, \textit{Subcase 1}\right) = \frac{(s_b^0 + \overline{C} - \underline{C})^2}{2(\overline{C} - \underline{C})^2} - \frac{(s_b^0 - \beta + \overline{C} - \underline{C})^2}{2(\overline{C} - \underline{C})^2} \tag{B.26}$$

$$= \frac{(s_b^0)^2 - (s_b^0 - \beta)^2 + 2(s_b^0 - (s_b^0 - \beta))(\overline{C} - \underline{C})}{2(\overline{C} - \underline{C})^2} \tag{B.27}$$

$$= \frac{(s_b^0)^2 - (s_b^0 - \beta)^2 + 2\beta(\overline{C} - \underline{C})}{2(\overline{C} - \underline{C})^2} \tag{B.28}$$

$$= \frac{-\beta^2 + 2\beta(s_b^0 + \overline{C} - \underline{C})}{2(\overline{C} - \underline{C})^2} \tag{B.29}$$

$$\geq \frac{-\beta^2 + 2\beta^2}{2(\overline{C} - \underline{C})^2} \tag{B.30}$$

$$= \frac{\beta^2}{2(\overline{C} - \underline{C})^2} \tag{B.31}$$

where second to the last line follows since we have $0 < \beta \leq s_b^0 + \overline{C} - \underline{C}$ in this subcase.

$$\mathbb{P}\left(s_b^0 - \beta \leq \pi_{t,1} - \pi_{t,b} < s_b^0, \textit{Subcase 2}\right) = 1 - \frac{(s_b^0 + \underline{C} - \overline{C})^2}{2(\overline{C} - \underline{C})^2} - 1 + \frac{(s_b^0 - \beta + \underline{C} - \overline{C})^2}{2(\overline{C} - \underline{C})^2} \tag{B.32}$$

$$= \frac{(\overline{C} - \underline{C} - s_b^0 + \beta)^2}{2(\overline{C} - \underline{C})^2} - \frac{(\overline{C} - \underline{C} - s_b^0)^2}{2(\overline{C} - \underline{C})^2} \tag{B.33}$$

$$= \frac{\beta^2 + 2\beta(\overline{C} - \underline{C} - s_b^0)}{2(\overline{C} - \underline{C})^2} \tag{B.34}$$

$$\geq \frac{\beta^2}{2(\overline{C} - \underline{C})^2} \tag{B.35}$$

where the last inequality follows since we have $\overline{C} - \underline{C} - s_b^0 \geq 0$ and $\beta > 0$ by definition. Now, since *Case 1* and *Case 2* are mutually exclusive events, we combine everything and obtain

$$\mathbb{P}\left(s_b^0 - \beta \leq \pi_{t,1} - \pi_{t,b} < s_b^0\right)$$

$$= \mathbb{P}\left(s_b^0 - \beta \leq \pi_{t,1} - \pi_{t,b} < s_b^0, \textit{Case 1}\right) + \mathbb{P}\left(s_b^0 - \beta \leq \pi_{t,1} - \pi_{t,b} < s_b^0, \textit{Case 2}\right) \tag{B.36}$$

$$\geq \mathbb{P}\left(s_b^0 - \beta \leq \pi_{t,1} - \pi_{t,b} < s_b^0, \textit{Subcase 1}\right) + \mathbb{P}\left(s_b^0 - \beta \leq \pi_{t,1} - \pi_{t,b} < s_b^0, \textit{Subcase 2}\right) \tag{B.37}$$

$$\geq \frac{\beta^2}{(\overline{C} - \underline{C})^2} \tag{B.38}$$

Combining this last result with (B.22), we obtain

$$\mathbb{P}\left(\ell\left(\mathbf{s}, i_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) = +\infty\right) \geq \frac{\beta^2}{(\overline{C} - \underline{C})^2}\left(1 - \frac{\gamma + \omega}{\overline{C} - \underline{C}}\right)^2 \left(\frac{\gamma}{\overline{C} - \underline{C}}\right)^{n-2} \tag{B.39}$$

□

Proof of Proposition 3.3. We follow a mainly similar argument as in the proof of Proposition 3.2. Recall that we know $b \neq 1$ since $\|\mathbf{s}^0 - \mathbf{s}\|_\infty > \beta$ by construction as explained in the previous proof, and that either $s_b > 0$ or $s_b^0 > 0$ holds. We also have $\omega = \sup_{\mathbf{s} \in \mathcal{B}(\mathbf{s}^j, d)} \max_{a \in \mathcal{A}} \{|s_a^0|, |s_a|\}$ as before. Now, consider the following conditions on the incentives

$$\pi_{t,a} < R_{\min} + \gamma + \beta - d \text{ for all } a \in \mathcal{A} \setminus \{1, b\} \tag{B.40}$$

$$\pi_{t,b} \geq R_{\min} + \gamma + \beta - d \tag{B.41}$$

$$\pi_{t,1} \geq R_{\min} + \gamma + \omega \tag{B.42}$$

which are compatible with Assumption 3.1. Now, since $|s_b^0 - s_b| > \beta$ by definition of $\mathbf{s}$, we know that $|s_b^0 - s_b| > |s_1^0 - s_1| = 0$. Suppose that without loss of generality, we have $s_b^0 - s_b > s_1^0 - s_1 = 0$ and $s_b^0 - s_b > \beta$. Then, we obtain

$$\mathbb{P}\left(\ell\left(\mathbf{s}, i_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) = +\infty\right)$$

$$\geq \mathbb{P}\left(\bigcup_{x \in \mathcal{A}, y \in \mathcal{A}, y \neq x} x = \arg\max_{a \in \mathcal{A}} \left(s_a^0 + \pi_{t,a}\right), y = \arg\max_{a \in \mathcal{A}} \left(s_a + \pi_{t,a}\right) \right) \tag{B.43}$$

$$\geq \mathbb{P}\left(1 = \arg\max_{a \in \mathcal{A}}(s_a + \pi_{t,a}), b = \arg\max_{a \in \mathcal{A}}(s_a^0 + \pi_{t,a})\right) \tag{B.44}$$

$$\geq \mathbb{P}\left(1 = \arg\max_{a \in \mathcal{A}}(s_a + \pi_{t,a}), b = \arg\max_{a \in \mathcal{A}}(s_a^0 + \pi_{t,a}) \middle| (B.40) - (B.42)\right) \mathbb{P}\left((B.40) - (B.42)\right) \tag{B.45}$$

$$= \mathbb{P}\left(s_b - s_1 < \pi_{t,1} - \pi_{t,b} < s_b^0 - s_1^0\right) \mathbb{P}\left(B.41\right) \mathbb{P}\left(B.42\right) \prod_{a \in \mathcal{A} \setminus \{1,b\}} \mathbb{P}\left(\pi_{t,a} < R_{\min} + \gamma + \beta - d\right) \tag{B.46}$$

$$\geq \mathbb{P}\left(s_b - s_1 < \pi_{t,1} - \pi_{t,b} < s_b^0 - s_1^0\right) \mathbb{P}\left(B.41\right) \mathbb{P}\left(B.42\right) \prod_{a \in \mathcal{A} \setminus \{1,b\}} \mathbb{P}\left(\pi_{t,a} \leq R_{\min} + \gamma\right) \tag{B.47}$$

$$= \mathbb{P}\left(s_b - s_1 < \pi_{t,1} - \pi_{t,b} < s_b^0 - s_1^0\right) \left(1 - \frac{\gamma + \beta - d}{\overline{C} - \underline{C}}\right) \left(1 - \frac{\gamma + \omega}{\overline{C} - \underline{C}}\right) \prod_{a \in \mathcal{A} \setminus \{1,b\}} \frac{\gamma}{\overline{C} - \underline{C}} \tag{B.48}$$

where (B.46) follows as $\pi_{t,a}$'s are independent random variables and (B.48) follows since $\underline{C} = R_{\min}$ by Assumption 3.1 and $\pi_{t,a} \sim \mathcal{U}(\underline{C}, \overline{C}), \forall a \in \mathcal{A}$. Then, we obtain the following lower bound for the first term in (B.48)

$$\mathbb{P}\left(s_b - s_1 < \pi_{t,1} - \pi_{t,b} < s_b^0 - s_1^0\right) \geq \mathbb{P}\left(s_b^0 - \beta < c_{t,1} - c_{t,b} < s_b^0\right) \geq \frac{\beta^2}{(\overline{C} - \underline{C})^2} \tag{B.49}$$

by using similar arguments as in (B.23)-(B.38) from the proof of Proposition 3.2. Lastly, combining this result with (B.48), we obtain

$$\mathbb{P}\left(\ell\left(\mathbf{s}, i_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) = +\infty\right) \geq \frac{\beta^2}{(\overline{C}-\underline{C})^2}\left(1-\frac{\gamma+\beta-d}{\overline{C}-\underline{C}}\right)\left(1-\frac{\gamma+\omega}{\overline{C}-\underline{C}}\right)\left(\frac{\gamma}{\overline{C}-\underline{C}}\right)^{n-2} \tag{B.50}$$

□

Proof of Proposition 3.4. Notice that the following three conditions are mutually exclusive events:

i. $K^0 \cap K = \emptyset$

ii. $K^0 \cap K \neq \emptyset$ and $b \notin K^0 \cap K$

iii. $K^0 \cap K \neq \emptyset$ and $b \in K^0 \cap K$

Hence, we can unite the results of Propositions 3.1, 3.2, and 3.3 and obtain

$$\begin{aligned}\mathbb{P}\left(\ell\left(\mathbf{s}, i_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) = +\infty\right) &= \sum_{j\in\{i,ii,iii\}} \mathbb{P}\left(\ell\left(\mathbf{s}, i_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) = +\infty, j\right) &\text{(B.51)}\\ &\geq \alpha\beta^2 &\text{(B.52)}\end{aligned}$$

for some constant $\alpha > 0$. □

Proof of Theorem 3.1. Recall that we define an open ball $\mathcal{B}(\mathbf{s}^j, d) := \{\mathbf{s} : \|\mathbf{s}-\mathbf{s}^j\|_\infty < d\}$ centered around a vector $\mathbf{s}^j$ with diameter $d > 0$. Since $\mathcal{F} = \{\mathbf{s} \in \mathcal{S}^n : \|\mathbf{s}-\mathbf{s}^0\|_\infty > \beta\}$ is compact, there is a finite subcover $\{\mathcal{B}(\mathbf{s}^j, d) : \mathbf{s}^j \in \mathcal{F}\}_{j=1}^q$ of a collection of open balls covering $\mathcal{F}$ where $d < \beta$. Further, we define $\overline{\mathbf{s}}_t^j := \arg\inf_{\mathbf{s}\in\mathcal{B}(\mathbf{s}^j,d)} L\left(\mathbf{s}, I_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)$. Now, since $\mathcal{F} \subseteq \bigcup_{j=1}^q \mathcal{B}(\mathbf{s}^j, d)$, we have

$$\begin{aligned}\inf_{\mathbf{s}\in\mathcal{F}} L\left(\mathbf{s}, I_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) = \inf_{\mathbf{s}\in\mathcal{F}} \sum_{\tau=1}^{t-1} \ell\left(\mathbf{s}, i_\tau(\boldsymbol{\pi}_\tau), \boldsymbol{\pi}_\tau\right) &\geq \min_{j\in[q]} \inf_{\mathbf{s}\in\mathcal{B}(\mathbf{s}^j,d)} \sum_{\tau=1}^{t-1} \ell\left(\mathbf{s}, i_\tau(\boldsymbol{\pi}_\tau), \boldsymbol{\pi}_\tau\right) &\text{(B.53)}\\ &\geq \min_{j\in[q]} \sum_{\tau=1}^{t-1} \ell\left(\overline{\mathbf{s}}_t^j, i_\tau(\boldsymbol{\pi}_\tau), \boldsymbol{\pi}_\tau\right) &\text{(B.54)}\\ &\geq \min_{j\in[q]} \sum_{\tau\in\Lambda(1,t)} \ell\left(\overline{\mathbf{s}}_t^j, i_\tau(\boldsymbol{\pi}_\tau), \boldsymbol{\pi}_\tau\right) &\text{(B.55)}\end{aligned}$$

where $[q] = \{1, \ldots, q\}$. We then follow by

$$\mathbb{P}\left(\inf_{\mathbf{s}\in\mathcal{F}} L\left(\mathbf{s}, I_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) < +\infty\right) \leq \mathbb{P}\left(\min_{j\in[q]} \sum_{\tau\in\Lambda(1,t)} \ell\left(\overline{\mathbf{s}}_t^j, i_\tau(\boldsymbol{\pi}_\tau), \boldsymbol{\pi}_\tau\right) < +\infty\right) \tag{B.56}$$

$$\leq \mathbb{P}\left(\bigcup_{j\in[q]} \sum_{\tau\in\Lambda(1,t)} \ell\left(\overline{\mathbf{s}}_t^j, i_\tau(\boldsymbol{\pi}_\tau), \boldsymbol{\pi}_\tau\right) < +\infty\right) \tag{B.57}$$

$$\leq \sum_{j\in[q]} \mathbb{P}\left(\sum_{\tau\in\Lambda(1,t)} \ell\left(\overline{\mathbf{s}}_t^j, i_\tau(\boldsymbol{\pi}_\tau), \boldsymbol{\pi}_\tau\right) < +\infty\right) \tag{B.58}$$

$$= \sum_{j\in[q]} \mathbb{P}\left(\ell\left(\overline{\mathbf{s}}_t^j, i_\tau(\boldsymbol{\pi}_\tau), \boldsymbol{\pi}_\tau\right) < +\infty,\ \forall\tau\in\Lambda(1,t)\right) \tag{B.59}$$

$$= \sum_{j\in[q]} \prod_{\tau\in\Lambda(1,t)} \mathbb{P}\left(\ell\left(\overline{\mathbf{s}}_t^j, i_\tau(\boldsymbol{\pi}_\tau), \boldsymbol{\pi}_\tau\right) < +\infty\right) \tag{B.60}$$

$$= \sum_{j\in[q]} \prod_{\tau\in\Lambda(1,t)} \left[1 - \mathbb{P}\left(\ell\left(\overline{\mathbf{s}}_t^j, i_\tau(\boldsymbol{\pi}_\tau), \boldsymbol{\pi}_\tau\right) = +\infty\right)\right] \tag{B.61}$$

$$\leq \sum_{j\in[q]} \prod_{\tau\in\Lambda(1,t)} \left(1 - \alpha\beta^2\right) \tag{B.62}$$

where the first inequality follows by (B.55), (B.58) follows by the Boole's inequality (a.k.a. union bound), (B.60) follows by the assumption of independence of the time steps, and (B.62) follows by the identifiability condition provided in Proposition 3.4. Note that we prove Proposition 3.4 for any vector $\mathbf{s} \in \mathcal{B}(\mathbf{s}^j, d)$, and hence, it also holds for $\overline{\mathbf{s}}_t^j$. We continue as

$$(B.62) = \sum_{j\in[q]} \left(1 - \alpha\beta^2\right)^{\eta(1,t)-1} \tag{B.63}$$

$$= \sum_{j\in[q]} \exp\left((\eta(1,t) - 1)\log\left(1 - \alpha\beta^2\right)\right) \tag{B.64}$$

$$\leq \sum_{j\in[q]} \exp\left(-\alpha(\eta(1,t) - 1)\beta^2\right) \tag{B.65}$$

$$= q\exp\left(-\alpha(\eta(1,t) - 1)\beta^2\right) \tag{B.66}$$

where (B.65) follows by an upper bound on natural logarithm: $\log x \leq x - 1$ for $x > 0$, which can be proven by the Mean Value Theorem and works by selecting $x = 1 - \alpha\beta^2$ in our case.

Next, we provide an upper bound for the covering number q by using the volume ratios. Recall that $\mathcal{S} = [R_{\min} - R_{\max}, R_{\max} - R_{\min}]$ by definition, and hence,

$$q = \mathcal{N}(d, \mathcal{F}, \|\cdot\|) \leq \frac{\mathrm{vol}(\mathcal{F})}{\mathrm{vol}(\mathcal{B}(\mathbf{s}^j, d))} \leq \frac{\mathrm{vol}(\mathcal{S}^n)}{\mathrm{vol}(\mathcal{B}(\mathbf{s}^j, d))} \leq \frac{(R_{\max} - R_{\min})^n}{d^n} \tag{B.67}$$

Suppose we have $d = \sqrt[n]{\beta}$. Then, combining everything, we obtain

$$\mathbb{P}\left(\inf_{\mathbf{s}\in\mathcal{F}} L\left(\mathbf{s}, I_t(\boldsymbol{\Pi}_t), \boldsymbol{\Pi}_t\right) < +\infty\right) \leq \frac{(R_{\max} - R_{\min})^n}{\beta} \exp\left(-\alpha(\eta(1,t) - 1)\beta^2\right) \tag{B.68}$$

$$= \exp\left(-\alpha(\eta(1,t)-1)\beta^2 - \log\beta + n\log(R_{\max} - R_{\min})\right) \tag{B.69}$$

□

Proof of Corollary 3.1. We first highlight that the result of Theorem 3.1 is proven for any normalized reward vector $\mathbf{s} \in \mathcal{F} \subset \mathcal{S}^n$ that satisfies $\|\mathbf{s}^0 - \mathbf{s}\|_\infty > \beta$ by definition. Also, recall that the principal's estimator $\widehat{\mathbf{s}}_t \in \mathcal{S}^n$ is defined in (3.7) such that it satisfies $L\left(\widehat{\mathbf{s}}_t, I_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) < +\infty$. Then, we have the following implication

$$\left\{\|\mathbf{s}^0 - \widehat{\mathbf{s}}_t\|_\infty > \beta\right\} \subseteq \left\{\exists \mathbf{s} : \|\mathbf{s}^0 - \mathbf{s}\|_\infty > \beta \text{ and } L\left(\mathbf{s}, I_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) < +\infty\right\} \tag{B.70}$$

which gives us the desired bound as

$$\mathbb{P}\left(\|\mathbf{s}^0 - \widehat{\mathbf{s}}_t\|_\infty > \beta\right) \leq \mathbb{P}\left(\inf_{\mathbf{s}\in\mathcal{F}} L\left(\mathbf{s}, I_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) < +\infty\right) \tag{B.71}$$

$$\leq \exp\left(-\alpha(\eta(1,t)-1)\beta^2 - \log\beta + n\log(R_{\max} - R_{\min})\right) \tag{B.72}$$

where the last inequality follows by Theorem 3.1. □

B.1.2 Results in Section 3.3

Proof of Lemma 3.1. We start by defining the indices $\kappa_t \in \arg\max_{a\in\mathcal{A}} \widehat{s}_{t,a}$ and $\kappa^0 \in \arg\max_{a\in\mathcal{A}} s^0_a$ for notational convenience. Then, we can rewrite the given probability as

$$\mathbb{P}\left(\max_{a\in\mathcal{A}} \widehat{s}_{t,a} - \widehat{s}_{t,j^*_t} + 2\beta_t \geq \max_{a\in\mathcal{A}} s^0_a - s^0_{j^*_t}\right)$$

$$= \mathbb{P}\left(\widehat{s}_{t,\kappa_t} - \widehat{s}_{t,j^*_t} + 2\beta_t \geq s^0_{\kappa^0} - s^0_{j^*_t}\right) \tag{B.73}$$

$$= \mathbb{P}\left(2\beta_t \geq s^0_{\kappa^0} - \widehat{s}_{t,\kappa_t} + \widehat{s}_{t,j^*_t} - s^0_{j^*_t}\right) \tag{B.74}$$

$$= \mathbb{P}\left(2\beta_t \geq s^0_{\kappa^0} - \widehat{s}_{t,\kappa_t} + \widehat{s}_{t,j^*_t} - s^0_{j^*_t} + \widehat{s}_{t,\kappa^0} - \widehat{s}_{t,\kappa^0}\right) \tag{B.75}$$

$$= \mathbb{P}\left(2\beta_t \geq \left(s^0_{\kappa^0} - \widehat{s}_{t,\kappa^0}\right) + \left(\widehat{s}_{t,\kappa^0} - \widehat{s}_{t,\kappa_t}\right) + \left(\widehat{s}_{t,j^*_t} - s^0_{j^*_t}\right)\right) \tag{B.76}$$

where the second term inside the parenthesis satisfies

$$0 \geq \widehat{s}_{t,\kappa^0} - \widehat{s}_{t,\kappa_t} \tag{B.77}$$

Further, if $\|\mathbf{s}^0 - \widehat{\mathbf{s}}_t\|_\infty \leq \beta_t$, then we have

$$\beta_t \geq s^0_{\kappa^0} - \widehat{s}_{t,\kappa^0} \tag{B.78}$$

$$\beta_t \geq \widehat{s}_{t,j^*_t} - s^0_{j^*_t} \tag{B.79}$$

Hence, we have

$$(B.77) - (B.79) \Longrightarrow 2\beta_t \geq \left(s^0_{\kappa^0} - \widehat{s}_{t,\kappa^0}\right) + \left(\widehat{s}_{t,\kappa^0} - \widehat{s}_{t,\kappa_t}\right) + \left(\widehat{s}_{t,j^*_t} - s^0_{j^*_t}\right) \tag{B.80}$$

when $\|\mathbf{s}^0 - \widehat{\mathbf{s}}_t\|_\infty \leq \beta_t$ holds. Combining this result with Corollary 3.1, we conclude the proof.

$$\begin{aligned}
&\mathbb{P}\left(2\beta_t \geq \left(s^0_{\kappa^0} - \widehat{s}_{t,\kappa^0}\right) + \left(\widehat{s}_{t,\kappa^0} - \widehat{s}_{t,\kappa_t}\right) + \left(\widehat{s}_{t,j^*_t} - s^0_{j^*_t}\right)\right) && \text{(B.81)}\\
&\geq \mathbb{P}\left(\|\mathbf{s}^0 - \widehat{\mathbf{s}}_t\|_\infty \leq \beta_t\right) && \text{(B.82)}\\
&> 1 - \exp\left(-\alpha(\eta(1,t) - 1)\beta_t^2 - \log\beta_t + n\log(R_{\max} - R_{\min})\right) && \text{(B.83)}
\end{aligned}$$

□

Proof of Proposition 3.5. By construction of Algorithm 2, the arm that the agent picks at time $t \in \mathcal{T}^{\text{xploit}}$ is defined as $i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t)) = \arg\max_{a\in\mathcal{A}} s^0_a + c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t)$. This implies

$$s^0_{i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))} + c_{i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t) > s^0_a + c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t) \quad \forall a \in \mathcal{A} \setminus \left\{i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t)\right\} \tag{B.84}$$

Then, the probability that the agent picks arm j^*_t at time t is bounded by

$$\begin{aligned}
\mathbb{P}\left(j^*_t = i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t))\right) &= \mathbb{P}\left(s^0_{j^*_t} + c_{j^*_t}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t) > s^0_a + c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t),\ \forall a \in \mathcal{A}\setminus\{j^*_t\}\right) && \text{(B.85)}\\
&= \mathbb{P}\left(s^0_{j^*_t} + \left(\max_{a\in\mathcal{A}} \widehat{s}_{t,a}\right) - \widehat{s}_{t,j^*_t} + 2\beta_t > s^0_a,\ \forall a \in \mathcal{A}\setminus\{j^*_t\}\right) && \text{(B.86)}\\
&\geq \mathbb{P}\left(s^0_{j^*_t} + \left(\max_{a\in\mathcal{A}} \widehat{s}_{t,a}\right) - \widehat{s}_{t,j^*_t} + 2\beta_t \geq \max_{a\in\mathcal{A}} s^0_a\right) && \text{(B.87)}\\
&> 1 - \exp\left(-\alpha(\eta(1,t) - 1)\beta_t^2 - \log\beta_t + n\log(R_{\max} - R_{\min})\right) && \text{(B.88)}
\end{aligned}$$

where the last inequality follows by Lemma 3.1. □

Proof of Proposition 3.6. First, recall that we define the true reward-maximizer action under the oracle incentives in Section 3.3.2 as

$$i(\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0)) = \arg\max_{j\in\mathcal{A}} \widetilde{V}(j, \mathbf{s}^0; \boldsymbol{\theta}^0) = \arg\max_{j\in\mathcal{A}} \theta^0_j - \left(\max_{a\in\mathcal{A}} s^0_a\right) + s^0_j \tag{B.89}$$

Then, we introduce the set $\mathcal{A}_t = \mathcal{A} \setminus \{i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t))\}$ for notational convenience and obtain

$$\begin{aligned}
\mathbb{P}\left(i(\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0)) \neq i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t))\right) &\leq \sum_{a\in\mathcal{A}_t} \mathbb{P}\left(\widetilde{V}(i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t)), \mathbf{s}^0; \boldsymbol{\theta}^0) < \widetilde{V}(a, \mathbf{s}^0; \boldsymbol{\theta}^0)\right) && \text{(B.90)}\\
&= \sum_{a\in\mathcal{A}_t} \mathbb{P}\left(\theta^0_{i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))} - \theta^0_a < s^0_a - s^0_{i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))}\right) && \text{(B.91)}
\end{aligned}$$

We continue by conditioning on whether the action picked by the agent under the exploitation incentives is same as the action with the highest estimated net reward to the principal (j_t^*).

$$\begin{aligned}(B.91) = \sum_{a\in\mathcal{A}_t} & \mathbb{P}\left(\theta^0_{i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))} - \theta^0_a < s^0_a - s^0_{i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))} \middle| j^*_t = i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))\right)\mathbb{P}\left(j^*_t = i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))\right)\\ &+ \mathbb{P}\left(\theta^0_{i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))} - \theta^0_a < s^0_a - s^0_{i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))} \middle| j^*_t \neq i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))\right)\mathbb{P}\left(j^*_t \neq i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))\right) \qquad \text{(B.92)}\\ \leq \sum_{a\in\mathcal{A}_t} & \mathbb{P}\left(\theta^0_{i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))} - \theta^0_a < s^0_a - s^0_{i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))} \middle| j^*_t = i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))\right) + \mathbb{P}\left(j^*_t \neq i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))\right) \qquad \text{(B.93)}\\ \leq \sum_{a\in\mathcal{A}_t} & \mathbb{P}\left(\theta^0_{j^*_t} - \theta^0_a < s^0_a - s^0_{j^*_t}\right)\\ &+ \exp\left(-\alpha(\eta(1,t)-1)\beta_t^2 - \log\beta_t + n\log(R_{\max}-R_{\min})\right) \qquad \text{(B.94)}\end{aligned}$$

where the last inequality follows by Proposition 3.5. Now, by definition of j_t^*, we have $\widetilde{V}(a,\widehat{\mathbf{s}}_t;\widehat{\boldsymbol{\theta}}) < \widetilde{V}(j_t^*,\widehat{\mathbf{s}}_t;\widehat{\boldsymbol{\theta}})$, $\forall a\in\mathcal{A}, a\neq j_t^*$ which implies $\widehat{\theta}_{t,a} - \widehat{\theta}_{t,j_t^*} < \widehat{s}_{t,j_t^*} - \widehat{s}_{t,a}, \forall a \in\mathcal{A}, a\neq j_t^*$. Combining this inequality with the first term in (B.94), we have

$$\begin{aligned}&\mathbb{P}\left(\theta^0_{j^*_t} - \theta^0_a < s^0_a - s^0_{j^*_t}\right)\\ &= \mathbb{P}\left((\theta^0_{j^*_t} - \widehat{\theta}_{t,j^*_t}) + (\widehat{\theta}_{t,a} - \theta^0_a) < (s^0_a - \widehat{s}_{t,a}) + (\widehat{s}_{t,j^*_t} - s^0_{j^*_t})\right) \qquad \text{(B.95)}\\ &= \mathbb{P}\left((\theta^0_{j^*_t} - \widehat{\theta}_{t,j^*_t}) + (\widehat{\theta}_{t,a} - \theta^0_a) < (s^0_a - \widehat{s}_{t,a}) + (\widehat{s}_{t,j^*_t} - s^0_{j^*_t}) \middle| \|\mathbf{s}^0 - \widehat{\mathbf{s}}_t\|_\infty \leq \beta_t\right)\\ &\qquad\qquad\qquad\qquad \cdot\mathbb{P}\left(\|\mathbf{s}^0 - \widehat{\mathbf{s}}_t\|_\infty \leq \beta_t\right)\\ &\quad + \mathbb{P}\left((\theta^0_{j^*_t} - \widehat{\theta}_{t,j^*_t}) + (\widehat{\theta}_{t,a} - \theta^0_a) < (s^0_a - \widehat{s}_{t,a}) + (\widehat{s}_{t,j^*_t} - s^0_{j^*_t}) \middle| \|\mathbf{s}^0 - \widehat{\mathbf{s}}_t\|_\infty > \beta_t\right)\\ &\qquad\qquad\qquad\qquad \cdot\mathbb{P}\left(\|\mathbf{s}^0 - \widehat{\mathbf{s}}_t\|_\infty > \beta_t\right) \qquad \text{(B.96)}\\ &\leq \mathbb{P}\left((\theta^0_{j^*_t} - \widehat{\theta}_{t,j^*_t}) + (\widehat{\theta}_{t,a} - \theta^0_a) < (s^0_a - \widehat{s}_{t,a}) + (\widehat{s}_{t,j^*_t} - s^0_{j^*_t}) \middle| \|\mathbf{s}^0 - \widehat{\mathbf{s}}_t\|_\infty \leq \beta_t\right)\\ &\quad + \mathbb{P}\left(\|\mathbf{s}^0 - \widehat{\mathbf{s}}_t\|_\infty > \beta_t\right) \qquad \text{(B.97)}\\ &\leq \mathbb{P}\left((\theta^0_{j^*_t} - \widehat{\theta}_{t,j^*_t}) + (\widehat{\theta}_{t,a} - \theta^0_a) < 2\beta_t\right)\\ &\quad + \exp\left(-\alpha(\eta(1,t)-1)\beta_t^2 - \log\beta_t + n\log(R_{\max}-R_{\min})\right) \qquad \text{(B.98)}\end{aligned}$$

where the last line follows by the finite-sample concentration bound in Corollary 3.1. Next, we bound the first term above as follows.

$$\begin{aligned}&\mathbb{P}\left(\theta^0_{j^*_t} - \widehat{\theta}_{t,j^*_t} < 2\beta_t - (\widehat{\theta}_{t,a} - \theta^0_a)\right)\\ &= \mathbb{P}\left(\theta^0_{j^*_t} - \widehat{\theta}_{t,j^*_t} < 2\beta_t - (\widehat{\theta}_{t,a} - \theta^0_a) \middle| \widehat{\theta}_{t,a} - \theta^0_a < 3\beta_t\right)\mathbb{P}\left(\widehat{\theta}_{t,a} - \theta^0_a < 3\beta_t\right)\\ &\quad + \mathbb{P}\left(\theta^0_{j^*_t} - \widehat{\theta}_{t,j^*_t} < 2\beta_t - (\widehat{\theta}_{t,a} - \theta^0_a) \middle| \widehat{\theta}_{t,a} - \theta^0_a \geq 3\beta_t\right)\mathbb{P}\left(\widehat{\theta}_{t,a} - \theta^0_a \geq 3\beta_t\right) \qquad \text{(B.99)}\end{aligned}$$

$$\leq \mathbb{P}\left(\widehat{\theta}_{t,j_t^*} - \theta^0_{j_t^*} > \beta_t\right) + \mathbb{P}\left(\widehat{\theta}_{t,a} - \theta^0_a \geq 3\beta_t\right) \tag{B.100}$$

$$\leq \mathbb{P}\left(\widehat{\theta}_{t,j_t^*} - \theta^0_{j_t^*} > \beta_t\right) + \mathbb{P}\left(\widehat{\theta}_{t,a} - \theta^0_a \geq \beta_t\right) \tag{B.101}$$

Notice that we bound the two probability terms in the last line in the same way by definition of $\widehat{\theta}_{t,a}$'s (3.14). For any $a \in \mathcal{A}$, let $\overline{T}(a,t) = |\{\tau \in \Lambda(1,t) : i_\tau(\boldsymbol{\pi}_\tau) = a\}|$ be the number of exploration steps up to time t at which the agent's reward-maximizer arm is action a. Thus, $\overline{T}(a,t)$ is the sum of $\eta(1,t)$ independent Bernoulli random variables with success probabilities $\mathbb{P}\left(a = \arg\max_{a' \in \mathcal{A}} s^0_{a'} + \pi_{\tau,a'}\right)$. Then,

$$\mathbb{E}\overline{T}(a,t) = \sum_{\tau \in \Lambda(1,t)} \mathbb{P}\left(a = \arg\max_{a' \in \mathcal{A}} s^0_{a'} + \pi_{\tau,a'}\right) \geq \sum_{\tau \in \Lambda(1,t)} \left[\sum_{a' \in \mathcal{A}} \frac{(s^0_a - s^0_{a'} + \overline{C} - \underline{C})^2}{2(\overline{C} - \underline{C})^2}\right] \tag{B.102}$$

$$= n\frac{(s^0_a + \gamma)^2}{2(\overline{C} - \underline{C})^2}\eta(1,t) \tag{B.103}$$

where the term in the square brackets in (B.102) follows by using the cdf derived in (B.12). Since the cdf is defined as a piecewise function, it suffices to only consider the case when $\underline{C} - \overline{C} \leq s^0_a - s^0_{a'} < 0$ holds to find a lower bound on the probability $\mathbb{P}\left(a = \arg\max_{a' \in \mathcal{A}} s^0_{a'} + \pi_{\tau,a'}\right)$. Further, (B.103) follows since by definition we know that $s^0_{a'} \leq R_{\max} - R_{\min} = \overline{C} - \underline{C} - \gamma$ for all $a' \in \mathcal{A}$.

Now, observing that $\overline{T}(a,t) \leq T(a,t)$, we can use Hoeffding's Inequality (Boucheron et al. 2013) to proceed. For any $a \in \mathcal{A}$,

$$\begin{aligned}
&\mathbb{P}\left(\widehat{\theta}_{t,a} - \theta^0_a > \beta_t\right) \\
&= \mathbb{P}\left(\widehat{\theta}_{t,a} - \theta^0_a > \beta_t \middle| \overline{T}(a,t) > \mathbb{E}\overline{T}(a,t)\frac{4\alpha(R_{\max} - R_{\min})^2(\overline{C} - \underline{C})^2}{n(s^0_a + \gamma)^2}\right) \\
&\qquad \cdot \mathbb{P}\left(\overline{T}(a,t) > \mathbb{E}\overline{T}(a,t)\frac{4\alpha(R_{\max} - R_{\min})^2(\overline{C} - \underline{C})^2}{n(s^0_a + \gamma)^2}\right) \\
&\quad + \mathbb{P}\left(\widehat{\theta}_{t,a} - \theta^0_a > \beta_t \middle| \overline{T}(a,t) \leq \mathbb{E}\overline{T}(a,t)\frac{4\alpha(R_{\max} - R_{\min})^2(\overline{C} - \underline{C})^2}{n(s^0_a + \gamma)^2}\right) \\
&\qquad \cdot \mathbb{P}\left(\overline{T}(a,t) \leq \mathbb{E}\overline{T}(a,t)\frac{4\alpha(R_{\max} - R_{\min})^2(\overline{C} - \underline{C})^2}{n(s^0_a + \gamma)^2}\right)
\end{aligned} \tag{B.104}$$

$$\begin{aligned}
&\leq \mathbb{P}\left(\widehat{\theta}_{t,a} - \theta^0_a > \beta_t \middle| \overline{T}(a,t) > \mathbb{E}\overline{T}(a,t)\frac{4\alpha(R_{\max} - R_{\min})^2(\overline{C} - \underline{C})^2}{n(s^0_a + \gamma)^2}\right) \\
&\quad + \mathbb{P}\left(\overline{T}(a,t) \leq \mathbb{E}\overline{T}(a,t)\frac{4\alpha(R_{\max} - R_{\min})^2(\overline{C} - \underline{C})^2}{n(s^0_a + \gamma)^2}\right)
\end{aligned} \tag{B.105}$$

$$\leq \exp\left(- \frac{2\frac{4\alpha(R_{\max} - R_{\min})^2(\overline{C} - \underline{C})^2}{n(s^0_a + \gamma)^2}\mathbb{E}\overline{T}(a,t)\beta_t^2}{4\alpha(R_{\max} - R_{\min})^2}\right)$$

$$+\exp\left(-2\eta(1,t)\Big(\frac{4\alpha(R_{\max}-R_{\min})^2(\overline{C}-\underline{C})^2}{n(s_a^0+\gamma)^2}\Big)^2\mathbb{E}\overline{T}(a,t)^2\right) \tag{B.106}$$

$$\leq \exp\left(-\eta(1,t)\frac{\log(\eta(1,t)-1)}{\eta(1,t)-1}\right)+\exp\big(-\eta(1,t)^3\big) \tag{B.107}$$

$$\leq \frac{1}{\eta(1,t)-1}+\frac{1}{\eta(1,t)^3} \tag{B.108}$$

$$\leq \frac{2}{\eta(1,t)-1} \tag{B.109}$$

where (B.107) follows by substituting the lower bound in (B.103) and $\beta_t = \sqrt{\frac{\log(\eta(1,t)-1)}{\alpha(\eta(1,t)-1)}}$. Lastly, we combine this result with (B.94), (B.98), and (B.101) and conclude our proof.

$$\mathbb{P}\left(i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\neq i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))\right)$$

$$\leq \frac{4n}{\eta(1,t)-1}+2n\exp\big(-\alpha(\eta(1,t)-1)\beta_t^2-\log\beta_t+n\log(R_{\max}-R_{\min})\big) \tag{B.110}$$

$$=\frac{4n}{\eta(1,t)-1}+2n\exp\left(-\log(\eta(1,t)-1)-\log\sqrt{\frac{\log(\eta(1,t)-1)}{\alpha(\eta(1,t)-1)}}+n\log(R_{\max}-R_{\min})\right) \tag{B.111}$$

$$=\frac{4n}{\eta(1,t)-1}+\frac{2n(R_{\max}-R_{\min})^n\sqrt{\alpha}}{\sqrt{(\eta(1,t)-1)\log(\eta(1,t)-1)}} \tag{B.112}$$

□

Proof of Theorem 3.2. The expected net reward of the principal defined in (3.24) has two main components: cost incurred due to the offered incentives and mean reward collected through the arm chosen by the agent. Accordingly, we decompose our regret notion (3.25) into two main parts as follows.

$$\text{Regret}\,(\Pi_{\epsilon,T}) = \sum_{t\in\mathcal{T}} V(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0);\boldsymbol{\theta}^0)-V_t(\boldsymbol{\pi}_t;\boldsymbol{\theta}^0) \tag{B.113}$$

$$=\sum_{t\in\mathcal{T}}\sum_{a\in\mathcal{A}}\big[\pi_{t,a}-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\big]+\sum_{t\in\mathcal{T}}\Big[\theta^0_{i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))}-\theta^0_{i_t(\boldsymbol{\pi}_t)}\Big] \tag{B.114}$$

First, we provide an upper bound for the first part of (B.114).

$$\sum_{t\in\mathcal{T}}\sum_{a\in\mathcal{A}}\big[\pi_{t,a}-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\big]\leq\sum_{t\in\mathcal{T}^{\text{xplore}}}\sum_{a\in\mathcal{A}}\big[\pi_{t,a}-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\big]+\sum_{t\in\mathcal{T}^{\text{xploit}}}\sum_{a\in\mathcal{A}}\Big[c_a(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t)-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\Big] \tag{B.115}$$

Notice that the cardinalities $|\mathcal{T}^{\text{xplore}}|$ and $|\mathcal{T}^{\text{xploit}}|$ are random variables. Then,

$$\mathbb{E}\Big[\sum_{t\in\mathcal{T}^{\text{xplore}}}\sum_{a\in\mathcal{A}}\big[\pi_{t,a}-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\big]\,\Big|\mathcal{T}^{\text{xplore}}\Big]\leq n(\overline{C}-\underline{C})|\mathcal{T}^{\text{xplore}}| \tag{B.116}$$

Taking the expectation of both sides of the last inequality, we obtain the following upper bound for the first summation term in (B.115).

$$\begin{aligned}
\sum_{t\in\mathcal{T}^{\text{xplore}}}\sum_{a\in\mathcal{A}}\left[\pi_{t,a}-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\right] \le n(\overline{C}-\underline{C})\mathbb{E}|\mathcal{T}^{\text{xplore}}| &= n(\overline{C}-\underline{C})\sum_{t=1}^{T}\min\left\{1,\frac{m}{t}\right\} && \text{(B.117)}\\
&\le n(\overline{C}-\underline{C})\sum_{t=1}^{T}\frac{m}{t} && \text{(B.118)}\\
&\le n(\overline{C}-\underline{C})\left(m+\int_{t=1}^{T}\frac{m}{t}\right) && \text{(B.119)}\\
&= nm(\overline{C}-\underline{C})(1+\log T) && \text{(B.120)}
\end{aligned}$$

where the last equality follows by the finite sum formula of the harmonic series. Next, we bound the second part of (B.115) as follows.

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t\in\mathcal{T}^{\text{xploit}}}\sum_{a\in\mathcal{A}}\left[c_a(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t)-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\right]\Big|\mathcal{T}^{\text{xploit}}\right]\\
&=\sum_{t\in\mathcal{T}^{\text{xploit}}}\sum_{a\in\mathcal{A}}\mathbb{E}\left[c_a(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t)-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\Big|\|\mathbf{s}^0-\widehat{\mathbf{s}}_t\|_\infty\le\beta_t\right]\mathbb{P}\left(\|\mathbf{s}^0-\widehat{\mathbf{s}}_t\|_\infty\le\beta_t\right)\\
&\quad+\sum_{t\in\mathcal{T}^{\text{xploit}}}\sum_{a\in\mathcal{A}}\mathbb{E}\left[c_a(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t)-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\Big|\|\mathbf{s}^0-\widehat{\mathbf{s}}_t\|_\infty>\beta_t\right]\mathbb{P}\left(\|\mathbf{s}^0-\widehat{\mathbf{s}}_t\|_\infty>\beta_t\right) && \text{(B.121)}\\
&\le\sum_{t\in\mathcal{T}^{\text{xploit}}}\sum_{a\in\mathcal{A}}\mathbb{E}\left[c_a(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t)-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\Big|\|\mathbf{s}^0-\widehat{\mathbf{s}}_t\|_\infty\le\beta_t\right]\\
&\quad+\sum_{t\in\mathcal{T}^{\text{xploit}}}\left(\overline{C}-\underline{C}\right)\exp\left(-\alpha(\eta(t)-1)\beta_t^2-\log\beta_t+n\log R_{\max}\right) && \text{(B.122)}
\end{aligned}$$

where the last line follows by Corollary 3.1. To compute an upper bound for the first term in the last inequality, we proceed as

$$\begin{aligned}
&\sum_{t\in\mathcal{T}^{\text{xploit}}}\sum_{a\in\mathcal{A}}\mathbb{E}\left[c_a(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t)-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\Big|\|\mathbf{s}^0-\widehat{\mathbf{s}}_t\|_\infty\le\beta_t\right]\\
&=\sum_{t\in\mathcal{T}^{\text{xploit}}}\sum_{a\in\mathcal{A}}\mathbb{E}\left[c_a(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t)-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\Big|\|\mathbf{s}^0-\widehat{\mathbf{s}}_t\|_\infty\le\beta_t,j_t^*=i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\right]\mathbb{P}\left(j_t^*=i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\right)\\
&\qquad+\mathbb{E}\left[c_a(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t)-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\Big|\|\mathbf{s}^0-\widehat{\mathbf{s}}_t\|_\infty\le\beta_t,j_t^*\ne i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\right]\mathbb{P}\left(j_t^*\ne i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\right)\\
& && \text{(B.123)}\\
&\le\sum_{t\in\mathcal{T}^{\text{xploit}}}\left(\max_{a\in\mathcal{A}}\widehat{s}_{t,a}-\widehat{s}_{t,j_t^*}+2\beta_t-\max_{a\in\mathcal{A}}s_a^0+s^0_{i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))}\Big|\|\mathbf{s}^0-\widehat{\mathbf{s}}_t\|_\infty\le\beta_t,j_t^*=i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\right)\\
&\qquad+\left(\overline{C}-\underline{C}\right)\mathbb{P}\left(j_t^*\ne i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\right) && \text{(B.124)}
\end{aligned}$$

As before, we use the indices $\kappa_t \in \arg\max_{a\in\mathcal{A}} \widehat{s}_{t,a}$ and $\kappa^0 \in \arg\max_{a\in\mathcal{A}} s^0_a$ for notational convenience.

$$\begin{aligned} &= \sum_{t\in\mathcal{T}^{\text{xploit}}} \left(\widehat{s}_{t,\kappa_t} - \widehat{s}_{t,j_t^*} + 2\beta_t - s^0_{\kappa^0} + s^0_{i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))} \middle| \|\mathbf{s}^0 - \widehat{\mathbf{s}}_t\|_\infty \le \beta_t, j_t^* = i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\right) \\ &\qquad + \left(\overline{C} - \underline{C}\right) \mathbb{P}\left(j_t^* \neq i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\right) \end{aligned} \tag{B.125}$$

$$\begin{aligned} &= \sum_{t\in\mathcal{T}^{\text{xploit}}} \Big((\widehat{s}_{t,\kappa_t} - s^0_{\kappa_t}) + (s^0_{\kappa_t} - s^0_{\kappa^0}) + (s^0_{i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))} - \widehat{s}_{t,j_t^*}) + 2\beta_t \\ &\qquad\qquad\qquad \Big| \|\mathbf{s}^0 - \widehat{\mathbf{s}}_t\|_\infty \le \beta_t, j_t^* = i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\Big) \\ &\qquad + \left(\overline{C} - \underline{C}\right) \mathbb{P}\left(j_t^* \neq i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\right) \end{aligned} \tag{B.126}$$

$$\le \sum_{t\in\mathcal{T}^{\text{xploit}}} 4\beta_t + \left(\overline{C} - \underline{C}\right) \mathbb{P}\left(j_t^* \neq i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\right) \tag{B.127}$$

At this step, we continue by observing that

$$\mathbb{P}\left(j_t^* = i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\right) \ge \mathbb{P}\left(j_t^* = i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t)),\ i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t)) = i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\right) \tag{B.128}$$

which implies

$$\begin{aligned} \mathbb{P}\left(j_t^* \neq i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\right) &\le 1 - \mathbb{P}\left(j_t^* = i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t)),\ i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t)) = i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\right) && \text{(B.129)} \\ &= 1 - \left[1 - \mathbb{P}\left(j_t^* \neq i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t)) \bigcup i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t)) \neq i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\right)\right] && \text{(B.130)} \\ &\le \mathbb{P}\left(j_t^* \neq i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))\right) + \mathbb{P}\left(i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t)) \neq i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\right) && \text{(B.131)} \\ &\le \exp\left(-\alpha(\eta(1,t)-1)\beta_t^2 - \log\beta_t + n\log(R_{\max} - R_{\min})\right) \\ &\quad + \frac{4n}{\eta(1,t)-1} + \frac{2n(R_{\max}-R_{\min})^n\sqrt{\alpha}}{\sqrt{(\eta(1,t)-1)\log(\eta(1,t)-1)}} && \text{(B.132)} \end{aligned}$$

where (B.130) follows by the fact that $\mathbb{P}(\cap_i A_i) = 1 - \mathbb{P}(\cup_i \overline{A}_i)$ for a set of events A_i's, (B.131) follows by the Boole's inequality (a.k.a. union bound), and the last inequality follows by Propositions 3.5 and 3.6.

Combining the last result with (B.122) and (B.127) for $\beta_t = \sqrt{\frac{\log(\eta(1,t)-1)}{\alpha(\eta(1,t)-1)}}$, we obtain

$$\begin{aligned} &\mathbb{E}\left[\sum_{t\in\mathcal{T}^{\text{xploit}}} \sum_{a\in\mathcal{A}} \left[c_a(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t) - c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\right] \middle| \mathcal{T}^{\text{xploit}}\right] \\ &\le \sum_{t\in\mathcal{T}^{\text{xploit}}} 4\beta_t + \left(\overline{C} - \underline{C}\right) \Bigg(2\exp\left(-\alpha(\eta(1,t)-1)\beta_t^2 - \log\beta_t + n\log(R_{\max} - R_{\min})\right) \end{aligned}$$

$$+\frac{4n}{\eta(1,t)-1}+\frac{2n(R_{\max}-R_{\min})^n\sqrt{\alpha}}{\sqrt{(\eta(1,t)-1)\log(\eta(1,t)-1)}}\Bigg) \tag{B.133}$$

$$=\sum_{t\in\mathcal{T}^{\text{xploit}}} 4\sqrt{\frac{\log(\eta(1,t)-1)}{\alpha(\eta(1,t)-1)}}+\sum_{t\in\mathcal{T}^{\text{xploit}}}\frac{4n\left(\overline{C}-\underline{C}\right)(R_{\max}-R_{\min})^n\sqrt{\alpha}}{\sqrt{(\eta(1,t)-1)\log(\eta(1,t)-1)}}+\sum_{t\in\mathcal{T}^{\text{xploit}}}\frac{4n\left(\overline{C}-\underline{C}\right)}{\eta(1,t)-1} \tag{B.134}$$

$$\leq\sum_{t\in\mathcal{T}^{\text{xploit}}} 4\sqrt{\frac{\log(\eta(1,t)-1)}{\alpha(\eta(1,t)-1)}}+\sum_{t\in\mathcal{T}^{\text{xploit}}}\frac{4n\left(\overline{C}-\underline{C}\right)(R_{\max}-R_{\min})^n\sqrt{\alpha}}{\sqrt{\eta(1,t)-2}}+\sum_{t\in\mathcal{T}^{\text{xploit}}}\frac{4n\left(\overline{C}-\underline{C}\right)}{\eta(1,t)-1} \tag{B.135}$$

where the second term in (B.135) follows by the following bound on the natural logarithm: $1-1/x\leq\log x$ for $x>0$. Now, recall that the principal's ϵ-Greedy Algorithm (2) performs pure exploration over the first m steps of the finite time horizon $\mathcal{T}$. This implies $\eta(1,t)\geq m$ and $t\geq m+1$ for any $t\in\mathcal{T}^{\text{xploit}}$. Then, because the terms of the three summations in (B.135) are monotone decreasing functions of $\eta(1,t)$ for $m\geq 4$, we can bound these finite summations with the corresponding definite integrals plus the first terms of these series.

$$\begin{aligned}&(B.135)\\&\leq\frac{4}{\sqrt{\alpha}}\int_{x=m}^{|\mathcal{T}^{\text{xploit}}|}\sqrt{\frac{\log(x-1)}{x-1}}dx+\int_{t=m}^{|\mathcal{T}^{\text{xploit}}|}\frac{4n\left(\overline{C}-\underline{C}\right)(R_{\max}-R_{\min})^n\sqrt{\alpha}}{\sqrt{x-2}}dx\\&\quad+\int_{t=m}^{|\mathcal{T}^{\text{xploit}}|}\frac{4n\left(\overline{C}-\underline{C}\right)}{x-1}dx+B_1\end{aligned} \tag{B.136}$$

where $B_1=\frac{4}{\sqrt{\alpha}}\sqrt{\frac{\log(m-1)}{m-1}}+\frac{4n\left(\overline{C}-\underline{C}\right)(R_{\max}-R_{\min})^n\sqrt{\alpha}}{\sqrt{m-2}}+\frac{4n\left(\overline{C}-\underline{C}\right)}{m-1}$,

$$\begin{aligned}&\leq\frac{8}{\sqrt{\alpha}}\sqrt{(|\mathcal{T}^{\text{xploit}}|-1)\log(|\mathcal{T}^{\text{xploit}}|-1)}\\&\quad+8n\left(\overline{C}-\underline{C}\right)(R_{\max}-R_{\min})^n\sqrt{\alpha}\sqrt{|\mathcal{T}^{\text{xploit}}|-2}+4n\left(\overline{C}-\underline{C}\right)\log(|\mathcal{T}^{\text{xploit}}|-1)+B_1\end{aligned} \tag{B.137}$$

$$\leq\frac{8}{\sqrt{\alpha}}\sqrt{T\log T}+8n\left(\overline{C}-\underline{C}\right)(R_{\max}-R_{\min})^n\sqrt{\alpha}\sqrt{T}+4n\left(\overline{C}-\underline{C}\right)\log T+B_1 \tag{B.138}$$

By taking the expectation of the last result, we have

$$\begin{aligned}\sum_{t\in\mathcal{T}^{\text{xploit}}}\sum_{a\in\mathcal{A}}\left[c_a(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t)-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\right]&\leq\frac{8}{\sqrt{\alpha}}\sqrt{T\log T}+8n\left(\overline{C}-\underline{C}\right)(R_{\max}-R_{\min})^n\sqrt{\alpha}\sqrt{T}\\&\quad+4n\left(\overline{C}-\underline{C}\right)\log T+B_1\end{aligned} \tag{B.139}$$

Combining the results in (B.120) and (B.139) with (B.115), we obtain the following upper bound for the first part of our regret bound in (B.114).

$$\begin{aligned}\sum_{t\in\mathcal{T}}\sum_{a\in\mathcal{A}}\left[\pi_{t,a}-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\right] &\le nm(\overline{C}-\underline{C})(1+\log T)+\frac{8}{\sqrt{\alpha}}\sqrt{T\log T}\\ &\quad+8n\left(\overline{C}-\underline{C}\right)(R_{\max}-R_{\min})^n\sqrt{\alpha}\sqrt{T}+4n\left(\overline{C}-\underline{C}\right)\log T+B_1\end{aligned}\tag{B.140}$$

Next, we consider the second part of our regret bound in (B.114).

$$\sum_{t\in\mathcal{T}}\left[\theta^0_{i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))}-\theta^0_{i_t(\boldsymbol{\pi}_t)}\right]=\sum_{t\in\mathcal{T}^{\text{xplore}}}\left[\theta^0_{i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))}-\theta^0_{i_t(\boldsymbol{\pi}_t)}\right]+\sum_{t\in\mathcal{T}^{\text{xploit}}}\left[\theta^0_{i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))}-\theta^0_{i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))}\right]\tag{B.141}$$

We recall that the principal's reward expectations θ^0_a belong to a known compact set Θ and define $\operatorname{diam}(\Theta):=\max_{a,a'\in\mathcal{A}}\theta^0_a-\theta^0_{a'}$. As earlier, we consider that $|\mathcal{T}^{\text{xplore}}|$ and $|\mathcal{T}^{\text{xploit}}|$ are random variables, and bound the first term in (B.141) by following a similar argument as in (B.117)-(B.120).

$$\sum_{t\in\mathcal{T}^{\text{xplore}}}\left[\theta^0_{i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))}-\theta^0_{i_t(\boldsymbol{\pi}_t)}\right]\le\operatorname{diam}(\Theta)\mathbb{E}|\mathcal{T}^{\text{xplore}}|\le\operatorname{diam}(\Theta)m(1+\log T)\tag{B.142}$$

We continue by deriving the upper bound for the second term in (B.141).

$$\begin{aligned}&\mathbb{E}\left[\sum_{t\in\mathcal{T}^{\text{xploit}}}\left[\theta^0_{i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))}-\theta^0_{i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))}\right]\middle|\mathcal{T}^{\text{xploit}}\right]\\ &=\sum_{t\in\mathcal{T}^{\text{xploit}}}\mathbb{E}\left[\mu_{t,i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))}-\mu_{t,i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))}\middle|i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\neq i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))\right]\\ &\qquad\qquad\cdot\mathbb{P}\left(i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\neq i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))\right)\end{aligned}\tag{B.143}$$

$$\le\operatorname{diam}(\Theta)\sum_{t\in\mathcal{T}^{\text{xploit}}}\mathbb{P}\left(i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\neq i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))\right)\tag{B.144}$$

$$\le\operatorname{diam}(\Theta)\sum_{t\in\mathcal{T}^{\text{xploit}}}\frac{4n}{\eta(1,t)-1}+\frac{2n(R_{\max}-R_{\min})^n\sqrt{\alpha}}{\sqrt{(\eta(1,t)-1)\log(\eta(1,t)-1)}}\tag{B.145}$$

which follows by Proposition 3.6. By following similar arguments as in (B.133) - (B.138), we obtain

$$\le 8n\operatorname{diam}(\Theta)(R_{\max}-R_{\min})^n\sqrt{\alpha}\sqrt{T}+4n\left(\overline{C}-\underline{C}\right)\log T+B_2\tag{B.146}$$

where $B_2 = \frac{2n\text{diam}(\Theta)(R_{\max}-R_{\min})^n\sqrt{\alpha}}{\sqrt{m-2}} + \frac{4n\text{diam}(\Theta)}{m-1}$. We then take the expectation of this result and get

$$\begin{aligned}\sum_{t\in\mathcal{T}^{\text{xploit}}} &\left[\theta^0_{i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))} - \theta^0_{i_t(\mathbf{c}(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t))}\right] \\ &\leq 8n\text{diam}(\Theta)(R_{\max}-R_{\min})^n\sqrt{\alpha}\sqrt{T} + 4n\left(\overline{C}-\underline{C}\right)\log T + B_2 \quad \text{(B.147)}\end{aligned}$$

Together (B.142) and (B.147) gives the following upper bound for the second part of our regret.

$$\begin{aligned}\sum_{t\in\mathcal{T}} \left[\theta^0_{i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))} - \theta^0_{i_t(\boldsymbol{\pi}_t)}\right] &\leq \text{diam}(\Theta)m(1+\log T) + 8n\text{diam}(\Theta)(R_{\max}-R_{\min})^n\sqrt{\alpha}\sqrt{T} \\ &\quad + 4n\left(\overline{C}-\underline{C}\right)\log T + B_2 \quad \text{(B.148)}\end{aligned}$$

Finally, we join the upper bounds in (B.140) and (B.148) to achieve the regret bound presented in Theorem 3.2.

$$\begin{aligned}\text{Regret}\left(\Pi_{\epsilon,T}\right) \leq\ &\frac{8}{\sqrt{\alpha}}\sqrt{T\log T} + 8n\left(\overline{C}-\underline{C}+\text{diam}(\Theta)\right)(R_{\max}-R_{\min})^n\sqrt{\alpha}\sqrt{T} \\ &+ \left(n(\overline{C}-\underline{C})(m+8) + \text{diam}(\Theta)m\right)\log T \\ &+ m\left(n(\overline{C}-\underline{C}) + \text{diam}(\Theta)\right) + B_1 + B_2 \quad \text{(B.149)}\end{aligned}$$

□

B.1.3 Results in Section 3.4

Proof of Proposition 3.7. First, recall that in Section 3.3.2, we show that if the agent behaves truthfully in accordance with their true mean reward vector $\mathbf{s}^0$ and the principal follows the oracle incentive policy $\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0)$, then the agent gets their minimum possible expected total reward (which is equal to $\max_{a'\in\mathcal{A}} s^0_{a'} + \varsigma$ for a sufficiently small constant $\varsigma > 0$).

In this proof, we start by demonstrating this result again by using the agent's optimization problem (3.30). To recall, the oracle incentive policy first computes the maximum net expected reward that the principal can get from the selection of each action $j \in \mathcal{A}$. This amount was computed as: $\widetilde{V}(j,\mathbf{s}^0;\boldsymbol{\theta}^0) =$ (principal's expected reward from j) $-$ (the minimum total incentives to make j agent's reward-maximizer action) $= \theta^0_j - \left(\max_{a'\in\mathcal{A}} s^0_{a'} - s^0_j\right)$. Then, we denoted the action corresponding to the highest of these values as $j^{*,0} = \arg\max_{j\in\mathcal{A}} \widetilde{V}(j,\mathbf{s}^0;\boldsymbol{\theta}^0)$ and the agent's true reward maximizer action as $i(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0)) = \arg\max_{j\in\mathcal{A}}\left(s^0_j + c_j(\boldsymbol{\theta}^0,\mathbf{s}^0)\right)$.

Now, in the optimization problem (3.30), we let $\mathbf{s} = \mathbf{s}^0$ and $\boldsymbol{\pi} = \mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0)$ where $\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0)$ is as given in (3.21)-(3.22). Then, we have $a = j^{*,0}$ satisfying the first and second constraints and $b = i(\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0))$ satisfying the third constraint. As discussed in Section 3.3.2, the oracle incentives are designed such that $j^{*,0} = i(\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0))$, and thus they also satisfy the last constraint of (3.30). This shows that $\mathbf{s}^0$ and the oracle incentive policy $\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0)$ together yield a feasible solution to the agent's optimization problem. Under this feasible solution, the principal's expected net reward is $\theta^0_{j^{*,0}} - c_{j^{*,0}}(\boldsymbol{\theta}^0, \mathbf{s}^0) = \theta^0_{j^{*,0}} - \max_{a' \in \mathcal{A}} s^0_{a'} + s^0_{j^{*,0}} - \varsigma$, and the agent's expected total reward (i.e., the value of the objective function) is $s^0_{j^{*,0}} + c_j^{*,0}(\boldsymbol{\theta}^0, \mathbf{s}^0) = \max_{a' \in \mathcal{A}} s^0_{a'} + \varsigma$.

Second, we show that there exists a different feasible solution to the agent's optimization problem (3.30) and that this solution yields a higher profit to the agent than the truthful (and worst-case) solution above. To show this, we need to consider two mutually exclusive cases based on the maximizer actions of the principal and the agent: $\kappa^0 := \arg\max_{a' \in \mathcal{A}} s^0_{a'}$ and $q^0 := \arg\max_{a' \in \mathcal{A}} \theta^0_{a'}$.

<u>*Case 1:*</u> $\kappa^0 = q^0$. In this case, notice that the principal does not need to incentivize the agent at all to get them pick the desired action, and thus we have $j^{*,0} = q^0$. However, the agent can pretend that they have a different reward vector whose reward-maximizer action is different than $j^{*,0}$. This way, the agent can oblige the principal to offer them positive incentives for selecting $j^{*,0}$. Let $\overline{q}^0 = \arg\max_{a' \in \mathcal{A} \setminus \{q^0\}} \theta^0_{a'}$ be the action associated with the second highest true mean reward of the principal. We define the quantity $Q_1 := \theta^0_{q^0} - \theta^0_{\overline{q}^0}$. Then, we consider the solution $(\mathbf{s}, \boldsymbol{\pi})$ where $\mathbf{s}$ is such that $s_{\overline{q}^0} = s^0_{q^0} + Q_1 - 2\varsigma$ and $s_j = s^0_j$, $\forall j \neq \overline{q}^0$ and $\boldsymbol{\pi}$ is such that $\pi_{q^0} = Q_1 - \varsigma$ and $\pi_j = 0$, $\forall j \neq q^0$, for a sufficiently small constant $\varsigma > 0$. For $a = b = q^0$, this solution is feasible to the agent's problem (3.30) and yields an expected net reward $\theta^0_{q^0} - \pi_{q^0} = \theta^0_{q^0} - Q_1 + \varsigma$ to the principal and an expected total reward $s^0_{q^0} + \pi_{q^0} = \max_{a' \in \mathcal{A}} s^0_{a'} + Q_1 - \varsigma$ to the agent (which is the value of the objective function for this solution). This result shows that the agent can increase their expected total reward by using the considered reward vector $\mathbf{s}$ and extracting an extra amount of $Q_1 - 2\varsigma$ from the principal.

<u>*Case 2:*</u> $\kappa^0 \neq q^0$. Now, we define a new quantity corresponding to the difference between the highest and second highest $\widetilde{V}(j, \mathbf{s}^0; \boldsymbol{\theta}^0)$. Let this quantity be $Q_2 := \widetilde{V}(j^{*,0}, \mathbf{s}^0; \boldsymbol{\theta}^0) - \max_{j \in \mathcal{A} \setminus \{j^{*,0}\}} \widetilde{V}(j, \mathbf{s}^0; \boldsymbol{\theta}^0)$. Then, we consider the solution $(\mathbf{s}, \boldsymbol{\pi})$ where $\mathbf{s}$ is such that $s_{\kappa^0} = s_{\kappa^0} + Q_2 - 2\varsigma$ and $s_a = s^0_a$, $\forall a \neq \kappa^0$ and $\boldsymbol{\pi}$ is such that $\pi_{j^{*,0}} = s^0_{\kappa^0} - s^0_{j^{*,0}} + Q_2 - \varsigma$ and $\pi_a = 0$, $\forall a \neq j^{*,0}$ for a sufficiently small constant $\varsigma > 0$. Note that this solution is feasible to the agent's problem (3.30) for $a = b = j^{*,0}$. Then, the principal's expected net reward becomes $\theta^0_{j^{*,0}} - \pi_{j^{*,0}} = \theta^0_{j^{*,0}} - \max_{a \in \mathcal{A}} s^0_a + s^0_{j^{*,0}} - Q_2 + \varsigma$ and the agent's expected total reward (i.e., the value of the objective function) becomes $s^0_{j^{*,0}} + \pi_{j^{*,0}} = \max_{a \in \mathcal{A}} s^0_a + Q_2 - \varsigma$. In other words, there is a feasible solution of (3.30) that increases the agent's expected reward (and decreases the principal's expected net reward) by $Q_2 - 2\varsigma$ as compared to the worst-case solution above.

These example solutions prove that the optimization problem given in (3.30) is feasible and designed to maximize the agent's information rent by the use of an untrue mean reward vector. □

B.2 Parameters for Numerical Experiments

In our simulations, we demonstrate the performance of our data-driven approach for different values of n (i.e., the cardinality of the carrier's action space). The parameter intervals are set to $\Theta = [0, 100]$ and $\mathcal{R} = [-20, 50]$, and the entries of the vectors $\boldsymbol{\theta}^0$ and $\mathbf{r}^0$ are randomly generated from these sets as reported below.

n	$\boldsymbol{\theta}^0$	$\mathbf{r}^0$
5	(29, 1, 14, 26, 15)	(14, -24, -4, 19, 29)
10	(0, 44, 51, 65, 9, 35, 69, 91, 51, 44)	(-4, 8, 22, -12, -2, 46, -8, 16, 38, 14)

Table B.1: Experimental parameters for different numbers of alternative carrier routes

Appendix C

Appendix for Chapter 4

C.1 Proofs of Theoretical Results

C.1.1 Results in Section 4.2

Proof of Proposition 4.1. Given that the agent selects the true reward-maximizer arm in the considered period t, i.e., $\upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a\in\mathcal{A}} (s_a^0 + \pi_{t,a})$, the construction of our estimator (4.6) implies that the one-stage loss function $\ell(\mathbf{s}, \upsilon_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t)$ can be strictly positive if and only if the selected arm satisfies $\upsilon_t(\boldsymbol{\pi}_t) \neq \arg\max_{a\in\mathcal{A}} (s_a + \pi_{t,a})$ for the given vector $\mathbf{s} \in \mathcal{B}(\mathbf{s}^j, d), j \in \{1, \ldots, q\}$.

Because we consider the scenario when $K^0 \cap K = \emptyset$, we already observe different maximizer indices for the true rewards $\mathbf{s}^0$ and the considered rewards $\mathbf{s}$ before adding the incentives. Therefore, we can simply choose a set of incentives such that the new maximizers after adding the incentives will still belong to the sets K and K^0. For that purpose, we consider the incentives

$$\pi_{t,a} < R_{\min} + \gamma + \beta - d \text{ for all } a \in \mathcal{A} \setminus \{\kappa, \kappa^0\} \tag{C.1}$$

$$\pi_{t,a} \geq R_{\min} + \gamma + \beta - d \text{ for } a \in \{\kappa, \kappa^0\} \tag{C.2}$$

for any $\kappa \in K,\ \kappa^0 \in K^0$ where γ is as introduced in Assumption 4.1. We note that (C.1) and (C.2) are valid conditions based on Assumption 4.1.

For the rest of our analysis, we define $\widetilde{\mathbf{s}}^j := \arg\inf_{\mathbf{s}\in\mathcal{B}(\mathbf{s}^j,d)} \|\mathbf{s}^0 - \mathbf{s}\|_\infty$ as the closest vector (with respect to the ℓ_∞-norm) in ball $\mathcal{B}(\mathbf{s}^j, d)$ to the true reward vector $\mathbf{s}^0$. Then, we have $\|\mathbf{s}^0 - \widetilde{\mathbf{s}}^j\|_\infty \geq \beta - d$ by construction, and we obtain

$$\begin{aligned}
&\mathbb{P}\left(\ell(\mathbf{s}, \upsilon_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t) \geq o \middle| \upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a\in\mathcal{A}} \left(s_a^0 + \pi_{t,a}\right)\right) \\
&= \mathbb{P}\left(\max_{a\in\mathcal{A}} (s_a + \pi_{t,a}) - s_{\upsilon_t(\boldsymbol{\pi}_t)} - \pi_{t,\upsilon_t(\boldsymbol{\pi}_t)} \geq o \middle| \upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a\in\mathcal{A}} \left(s_a^0 + \pi_{t,a}\right)\right) \qquad \text{(C.3)}
\end{aligned}$$

$$\geq \mathbb{P}\Bigg(\bigcup_{x\in\mathcal{A},y\in\mathcal{A},y\neq x} x = \arg\max_{a\in\mathcal{A}}\left(s^0_a + \pi_{t,a}\right), y = \arg\max_{a\in\mathcal{A}}\left(s_a + \pi_{t,a}\right), s_y + \pi_{t,y} - s_x - \pi_{t,x} \geq o \Bigg) \tag{C.4}$$

$$\geq \mathbb{P}\left(\kappa = \arg\max_{a\in\mathcal{A}}(s_a + \pi_{t,a}), \kappa^0 = \arg\max_{a\in\mathcal{A}}(s^0_a + \pi_{t,a}), s_\kappa + \pi_{t,\kappa} - s_{\kappa^0} - \pi_{t,\kappa^0} \geq o\right) \tag{C.5}$$

$$\geq \mathbb{P}\Bigg(\kappa = \arg\max_{a\in\mathcal{A}}(s_a + \pi_{t,a}), \kappa^0 = \arg\max_{a\in\mathcal{A}}(s^0_a + \pi_{t,a}), s_\kappa + \pi_{t,\kappa} - s_{\kappa^0} - \pi_{t,\kappa^0} \geq o \Bigg| (C.1), (C.2)\Bigg) \cdot \mathbb{P}\left((C.1), (C.2)\right) \tag{C.6}$$

$$= \mathbb{P}\left(s_{\kappa^0} - s_\kappa + o \leq \pi_{t,\kappa} - \pi_{t,\kappa^0} < s^0_{\kappa^0} - s^0_\kappa\right) \cdot \prod_{a\in\{\kappa,\kappa^0\}} \mathbb{P}\left(\pi_{t,a} \geq R_{\min} + \gamma + \beta - d\right) \cdot \prod_{a\in\mathcal{A}\setminus\{\kappa,\kappa^0\}} \mathbb{P}\left(\pi_{t,a} < R_{\min} + \gamma + \beta - d\right) \tag{C.7}$$

$$\geq \mathbb{P}\left(s_{\kappa^0} - s_\kappa + o \leq \pi_{t,\kappa} - \pi_{t,\kappa^0} < s^0_{\kappa^0} - s^0_\kappa\right) \cdot \prod_{a\in\{\kappa,\kappa^0\}} \mathbb{P}\left(\pi_{t,a} \geq R_{\min} + \gamma + \beta - d\right) \cdot \prod_{a\in\mathcal{A}\setminus\{\kappa,\kappa^0\}} \mathbb{P}\left(\pi_{t,a} \leq R_{\min} + \gamma\right) \tag{C.8}$$

$$= \mathbb{P}\left(s_{\kappa^0} - s_\kappa + o \leq \pi_{t,\kappa} - \pi_{t,\kappa^0} < s^0_{\kappa^0} - s^0_\kappa\right) \cdot \prod_{a\in\{\kappa,\kappa^0\}} \left(1 - \frac{R_{\min} + \gamma + \beta - d - \underline{C}}{\overline{C} - \underline{C}}\right) \cdot \prod_{a\in\mathcal{A}\setminus\{\kappa,\kappa^0\}} \frac{R_{\min} + \gamma - \underline{C}}{\overline{C} - \underline{C}} \tag{C.9}$$

$$= \mathbb{P}\left(s_{\kappa^0} - s_\kappa + o \leq \pi_{t,\kappa} - \pi_{t,\kappa^0} < s^0_{\kappa^0} - s^0_\kappa\right)] \cdot \prod_{a\in\{\kappa,\kappa^0\}} \left(1 - \frac{\gamma + \beta - d}{\overline{C} - \underline{C}}\right) \cdot \prod_{a\in\mathcal{A}\setminus\{\kappa,\kappa^0\}} \frac{\gamma}{\overline{C} - \underline{C}} \tag{C.10}$$

where (C.7) follows since $\pi_{t,a}$'s are considered to be independent random variables, (C.9) follows since $\pi_{t,a} \sim \mathcal{U}(\underline{C}, \overline{C}), \forall a \in \mathcal{A}$, and (C.10) follows since $\underline{C} = R_{\min}$ by Assumption 4.1.

For the first term in (C.10), notice that the case that $s_{\kappa^0} - s_\kappa = s^0_{\kappa^0} - s^0_\kappa = 0$ cannot occur. This can only happen if $\kappa^0 \in K$ and $\kappa \in K^0$ which contradicts with the condition $K^0 \cap K_t = \emptyset$. Similarly, $\mathbf{s}^0$ cannot be the all-zeros vector under the given condition $K^0 \cap K_t = \emptyset$. Further, since $0 < o < \delta = s^0_{\kappa^0} - \max_{a\in\mathcal{A}\setminus\{K^0\}} s^0_a$ by definition, we know that $s_{\kappa^0} - s_\kappa + o < s^0_{\kappa^0} - s^0_\kappa$ holds. This implies that the first probability term in (C.10) has a nonzero value and can be bounded by using the cumulative distribution function (cdf) of $\pi_{t,a} - \pi_{t,a'}$ which is the difference of two identically and independently distributed (iid) Uniform random variables. The difference $\pi_{t,a} - \pi_{t,a'}$ follows a triangular distribution whose cdf can be explicitly computed as in B.12 from Appendix B.1.1. Using this cdf, we derive a strictly positive lower bound for the first term in (C.10).

$$
\begin{aligned}
&\mathbb{P}\left(s_{\kappa^0} - s_\kappa + o \le \pi_{t,\kappa} - \pi_{t,\kappa^0} < s^0_{\kappa^0} - s^0_\kappa\right) \\
&\ge \mathbb{P}\left(s_{\kappa^0} - s_\kappa + o \le \pi_{t,\kappa} - \pi_{t,\kappa^0} < s^0_{\kappa^0} - s^0_\kappa,\ s_{\kappa^0} - s_\kappa + o < 0\right) && \text{(C.11)} \\
&= 1 - \frac{(s^0_{\kappa^0} - s^0_\kappa + \underline{C} - \overline{C})^2}{2(\overline{C} - \underline{C})^2} - \frac{(s_{\kappa^0} - s_\kappa + o + \overline{C} - \underline{C})^2}{2(\overline{C} - \underline{C})^2} && \text{(C.12)} \\
&= 1 - \frac{(s^0_{\kappa^0} - s^0_\kappa)^2 + (s_{\kappa^0} - s_\kappa + o)^2 + 2(\overline{C} - \underline{C})^2 + 2(\overline{C} - \underline{C})(s_{\kappa^0} - s_\kappa + o - s^0_{\kappa^0} + s^0_\kappa)}{2(\overline{C} - \underline{C})^2} \\
& && \text{(C.13)} \\
&= \frac{-(s^0_{\kappa^0} - s^0_\kappa)^2 - (s_{\kappa^0} - s_\kappa + o)^2 + 2(\overline{C} - \underline{C})(s_\kappa - s_{\kappa^0} - o + s^0_{\kappa^0} - s^0_\kappa)}{2(\overline{C} - \underline{C})^2} && \text{(C.14)} \\
&\ge \frac{-(s^0_{\kappa^0} - s^0_\kappa)^2 - (s_{\kappa^0} - s_\kappa + o)^2 + 2(\overline{C} - \underline{C})(s^0_{\kappa^0} - s^0_\kappa)}{2(\overline{C} - \underline{C})^2} && \text{(C.15)}
\end{aligned}
$$

where the last line follows since we consider the case $s_{\kappa^0} - s_\kappa + o < 0$ for this lower bound. Then,

$$\ge \frac{(s^0_{\kappa^0} - s^0_\kappa)^2 - o^2}{2(\overline{C} - \underline{C})^2} > 0 \qquad \text{(C.16)}$$

which follows since $s^0_{\kappa^0} - s^0_\kappa \le \overline{C} - \underline{C}$ by definition and $s_{\kappa^0} - s_\kappa < 0$ based on the considered case. Combining this last result with (C.10), we obtain the following nonzero lower bound for the desired probability.

$$
\begin{aligned}
&\mathbb{P}\left(\ell\left(\mathbf{s}, \upsilon_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) \ge o \,\middle|\, \upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a\in\mathcal{A}} \left(s^0_a + \pi_{t,a}\right)\right) \\
&\qquad\qquad \ge \left(\frac{(s^0_{\kappa^0} - s^0_\kappa)^2 - o^2}{2(\overline{C} - \underline{C})^2}\right)\left(1 - \frac{\gamma + \beta - d}{\overline{C} - \underline{C}}\right)^2 \left(\frac{\gamma}{\overline{C} - \underline{C}}\right)^{n-2} \qquad \text{(C.17)}
\end{aligned}
$$

□

Proof of Proposition 4.2. Similar to our observation in the proof of Proposition 4.1, the construction of our estimator (4.6) implies that the one-stage loss function $\ell\left(\mathbf{s}, \upsilon_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right)$ can be strictly positive if and only if we have $\upsilon_t(\boldsymbol{\pi}_t) \ne \arg\max_{a\in\mathcal{A}}\left(s_a + \pi_{t,a}\right)$ for the given vector $\mathbf{s} \in \mathcal{B}(\mathbf{s}^j, d), j \in \{1, \ldots, q\}$ under the condition of $\upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a\in\mathcal{A}}\left(s^0_a + \pi_{t,a}\right)$. Therefore, to prove the lower bound in (4.8), we will consider the event where $\arg\max_{a\in\mathcal{A}}(s_a + \pi_{t,a}) = 1$ and $\arg\max_{a\in\mathcal{A}}(s^0_a + \pi_{t,a}) = b$ because of the fact that $b \ne 1$. As we have $s_1 = s^0_1 = 0$ by construction, having $b = 1$ would imply that $\mathbf{s}^0 = \mathbf{s} = \mathbf{0}_n$, and that would contradict with the definition of $\mathbf{s}$ which implies $\|\mathbf{s}^0 - \mathbf{s}\|_\infty = |s^0_b - s_b| > \beta$.

Let $\omega = \sup_{\mathbf{s}\in\mathcal{B}(\mathbf{s}^j,d)} \max_{a\in\mathcal{A}}\{|s^0_a|, |s_a|\}$ be the largest absolute value observed among the entries of $\mathbf{s}^0$ and of all vectors in $\mathcal{B}(\mathbf{s}^j, d)$. Then, we suppose

$$\pi_{t,a} < R_{\min} + \gamma + \beta - d \text{ for all } a \in \mathcal{A} \setminus \{1, b\} \tag{C.18}$$

$$\pi_{t,a} \geq R_{\min} + \gamma + \omega \text{ for } a \in \{1, b\}. \tag{C.19}$$

which are consistent with Assumption 4.1. These conditions imply that $s^0_{\kappa^0} + \pi_{t,\kappa^0} < s^0_a + \pi_{t,a}$ and $s_\kappa + \pi_{t,\kappa} < s_a + \pi_{t,a}$ hold for any $\kappa^0 \in K^0$, $\kappa \in K$, $a \in \{1, b\}$, and that the indices in the sets K^0 and K are no more maximizers after adding the incentives in (C.18)-(C.19). Thus, we will obtain the desired case (that is $\arg\max_{a\in\mathcal{A}}(s_a + \pi_{t,a}) = 1$ and $\arg\max_{a\in\mathcal{A}}(s^0_a + \pi_{t,a}) = b$) if the events $s^0_1 + \pi_{t,1} < s^0_b + \pi_{t,b}$ and $s_b + \pi_{t,b} < s_1 + \pi_{t,1}$ hold.

Further, because $|s^0_b - s_b| > \beta$ by definition, we know that $|s^0_b - s_b| > |s^0_1 - s_1| = 0$. Suppose that without loss of generality, we have $s^0_b - s_b > s^0_1 - s_1 = 0$ and $s^0_b - s_b > \beta$. Then, our proof follows as

$$\mathbb{P}\left(\ell\left(\mathbf{s}, \upsilon_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) \geq o \middle| \upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a\in\mathcal{A}}\left(s^0_a + \pi_{t,a}\right)\right)$$

$$= \mathbb{P}\left(\max_{a\in\mathcal{A}}\left(s_a + \pi_{t,a}\right) - s_{\upsilon_t(\boldsymbol{\pi}_t)} - \pi_{t,\upsilon_t(\boldsymbol{\pi}_t)} \geq o \middle| \upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a\in\mathcal{A}}\{s^0_a + \pi_{t,a}\}\right) \tag{C.20}$$

$$\geq \mathbb{P}\Bigg(\bigcup_{x\in\mathcal{A}, y\in\mathcal{A}, y\neq x} x = \arg\max_{a\in\mathcal{A}}\left(s^0_a + \pi_{t,a}\right), y = \arg\max_{a\in\mathcal{A}}\left(s_a + \pi_{t,a}\right), s_y + \pi_{t,y} - s_x - \pi_{t,x} \geq o\Bigg) \tag{C.21}$$

$$\geq \mathbb{P}\left(1 = \arg\max_{a\in\mathcal{A}}(s_a + \pi_{t,a}), b = \arg\max_{a\in\mathcal{A}}(s^0_a + \pi_{t,a}), s_1 + \pi_{t,1} - s_b - \pi_{t,b} \geq o\right) \tag{C.22}$$

$$\geq \mathbb{P}\Bigg(1 = \arg\max_{a\in\mathcal{A}}(s_a + \pi_{t,a}), b = \arg\max_{a\in\mathcal{A}}(s^0_a + \pi_{t,a}), s_1 + \pi_{t,1} - s_b - \pi_{t,b} \geq o \,\Bigg|\, (C.18), (C.19)\Bigg)\mathbb{P}\left((C.18), (C.19)\right) \tag{C.23}$$

$$= \mathbb{P}\left(s_b - s_1 + o \leq \pi_{t,1} - \pi_{t,b} < s^0_b - s^0_1\right) \cdot \prod_{a\in\{1,b\}} \mathbb{P}\left(\pi_{t,a} \geq R_{\min} + \gamma + \omega\right) \cdot \prod_{a\in\mathcal{A}\setminus\{1,b\}} \mathbb{P}\left(\pi_{t,a} < R_{\min} + \gamma + \beta - d\right) \tag{C.24}$$

$$\geq \mathbb{P}\left(s_b - s_1 + o \leq \pi_{t,1} - \pi_{t,b} < s^0_b - s^0_1\right) \cdot \prod_{a\in\{1,b\}} \mathbb{P}\left(\pi_{t,a} \geq R_{\min} + \gamma + \omega\right) \cdot \prod_{a\in\mathcal{A}\setminus\{1,b\}} \mathbb{P}\left(\pi_{t,a} \leq R_{\min} + \gamma\right) \tag{C.25}$$

$$= \mathbb{P}\left(s_b - s_1 + o \leq \pi_{t,1} - \pi_{t,b} < s^0_b - s^0_1\right) \cdot \prod_{a\in\{1,b\}} \left(1 - \frac{\gamma + \omega}{\overline{C} - \underline{C}}\right) \cdot \prod_{a\in\mathcal{A}\setminus\{1,b\}} \frac{\gamma}{\overline{C} - \underline{C}} \tag{C.26}$$

where (C.25) and (C.26) follow since $\underline{C} = R_{\min}$ by Assumption 4.1 and $\pi_{t,a}$'s are independent random variables with $\pi_{t,a} \sim \mathcal{U}(\underline{C}, \overline{C}), \forall a \in \mathcal{A}$. Next, we bound the first term in (C.26).

$$\begin{aligned}
\mathbb{P}\left(s_b - s_1 + o \leq \pi_{t,1} - \pi_{t,b} < s_b^0 - s_1^0\right) &= \mathbb{P}\left(s_b - s_1 + s_b^0 - s_b^0 + o \leq \pi_{t,1} - \pi_{t,b} < s_b^0 - s_1^0\right) && \text{(C.27)}\\
&\geq \mathbb{P}\left(s_b^0 - \beta - s_1 + o \leq \pi_{t,1} - \pi_{t,b} < s_b^0 - s_1^0\right) && \text{(C.28)}\\
&= \mathbb{P}\left(s_b^0 - \beta + o \leq \pi_{t,1} - \pi_{t,b} < s_b^0\right) && \text{(C.29)}
\end{aligned}$$

We can further bound the last probability by using the cdf (B.12) that is defined as a piecewise function. For that purpose, we consider the following two mutually exclusive cases.

- *Case 1:* $\underline{C} - \overline{C} \leq s_b^0 < 0$
- *Case 2:* $0 \leq s_b^0 \leq \overline{C} - \underline{C}$

We also consider the following subcases to bound the two cases above.

- *Subcase 1:* $\underline{C} - \overline{C} \leq s_b^0 < 0$ and $\underline{C} - \overline{C} \leq s_b^0 - \beta + o < 0$
- *Subcase 2:* $0 \leq s_b^0 \leq \overline{C} - \underline{C}$ and $0 \leq s_b^0 - \beta + o \leq \overline{C} - \underline{C}$

and compute the lower bounds for these subcases as follows.

$$\begin{aligned}
&\mathbb{P}\left(s_b^0 - \beta + o \leq \pi_{t,1} - \pi_{t,b} < s_b^0, \textit{Subcase 1}\right)\\
&= \frac{(s_b^0 + \overline{C} - \underline{C})^2}{2(\overline{C} - \underline{C})^2} - \frac{(s_b^0 - \beta + o + \overline{C} - \underline{C})^2}{2(\overline{C} - \underline{C})^2} && \text{(C.30)}\\
&= \frac{(s_b^0)^2 - (s_b^0 - \beta + o)^2 + 2(s_b^0 - (s_b^0 - \beta + o))(\overline{C} - \underline{C})}{2(\overline{C} - \underline{C})^2} && \text{(C.31)}\\
&= \frac{(s_b^0)^2 - (s_b^0 - \beta + o)^2 + 2(\beta - o)(\overline{C} - \underline{C})}{2(\overline{C} - \underline{C})^2} && \text{(C.32)}\\
&= \frac{(s_b^0)^2 - (s_b^0)^2 - (\beta - o)^2 + 2(\beta - o)s_b^0 + 2(\beta - o)(\overline{C} - \underline{C})}{2(\overline{C} - \underline{C})^2} && \text{(C.33)}\\
&= \frac{-(\beta - o)^2 + 2(\beta - o)(s_b^0 + \overline{C} - \underline{C})}{2(\overline{C} - \underline{C})^2} && \text{(C.34)}\\
&\geq \frac{-(\beta - o)^2 + 2(\beta - o)^2}{2(\overline{C} - \underline{C})^2} && \text{(C.35)}\\
&= \frac{(\beta - o)^2}{2(\overline{C} - \underline{C})^2} && \text{(C.36)}
\end{aligned}$$

where second to the last line follows since we have $0 < \beta - o \leq s_b^0 + \overline{C} - \underline{C}$ in this subcase.

$$\mathbb{P}\left(s_b^0 - \beta + o \leq \pi_{t,1} - \pi_{t,b} < s_b^0, \textit{Subcase 2}\right)$$

$$= 1 - \frac{(s_b^0 + \underline{C} - \overline{C})^2}{2(\overline{C} - \underline{C})^2} - 1 + \frac{(s_b^0 - \beta + o + \underline{C} - \overline{C})^2}{2(\overline{C} - \underline{C})^2} \tag{C.37}$$

$$= \frac{(\overline{C} - \underline{C} - s_b^0 + \beta - o)^2}{2(\overline{C} - \underline{C})^2} - \frac{(\overline{C} - \underline{C} - s_b^0)^2}{2(\overline{C} - \underline{C})^2} \tag{C.38}$$

$$= \frac{(\beta - o)^2 + 2(\beta - o)(\overline{C} - \underline{C} - s_b^0)}{2(\overline{C} - \underline{C})^2} \tag{C.39}$$

$$\geq \frac{(\beta - o)^2}{2(\overline{C} - \underline{C})^2} \tag{C.40}$$

where the last inequality follows since we have $\overline{C} - \underline{C} - s_b^0 \geq 0$ and $\beta - o > 0$ by definition. Next, we use the lower bounds for *Subcase 1* and *Subcase 2* to bound *Case 1* and *Case 2.*

$$\mathbb{P}\left(s_b^0 - \beta + o \leq \pi_{t,1} - \pi_{t,b} < s_b^0\right)$$

$$= \mathbb{P}\left(s_b^0 - \beta + o \leq \pi_{t,1} - \pi_{t,b} < s_b^0, \textit{Case 1}\right) + \mathbb{P}\left(s_b^0 - \beta + o \leq \pi_{t,1} - \pi_{t,b} < s_b^0, \textit{Case 2}\right) \tag{C.41}$$

$$\geq \mathbb{P}\left(s_b^0 - \beta + o \leq \pi_{t,1} - \pi_{t,b} < s_b^0, \textit{Subcase 1}\right) + \mathbb{P}\left(s_b^0 - \beta + o \leq \pi_{t,1} - \pi_{t,b} < s_b^0, \textit{Subcase 2}\right) \tag{C.42}$$

$$\geq \frac{(\beta - o)^2}{(\overline{C} - \underline{C})^2} \tag{C.43}$$

We conclude our proof by combining the last result with (C.26) which yields

$$\mathbb{P}\left(\ell\left(\mathbf{s}, \upsilon_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) \geq o \middle| \upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a \in \mathcal{A}} \left(s_a^0 + \pi_{t,a}\right)\right)$$

$$\geq \frac{(\beta - o)^2}{(\overline{C} - \underline{C})^2} \left(1 - \frac{\gamma + \omega}{\overline{C} - \underline{C}}\right)^2 \left(\frac{\gamma}{\overline{C} - \underline{C}}\right)^{n-2} \tag{C.44}$$

□

Proof of Proposition 4.3. Our proof mainly follows arguments similar to those in the proof of Proposition 4.2. Recall that $b \neq 1$ (because $\mathbf{s}$ is defined such that $\|\mathbf{s}^0 - \mathbf{s}\|_\infty > \beta$) and that either $s_b > 0$ or $s_b^0 > 0$ holds. We use the definition of $\omega = \sup_{\mathbf{s} \in \mathcal{B}(\mathbf{s}^j, d)} \max_{a \in \mathcal{A}} \{|s_a^0|, |s_a|\}$ and consider the following set of incentives.

$$\pi_{t,a} < R_{\min} + \gamma + \beta - d \text{ for all } a \in \mathcal{A} \setminus \{1, b\} \tag{C.45}$$

$$\pi_{t,b} \geq R_{\min} + \gamma + \beta - d \tag{C.46}$$

$$\pi_{t,1} \geq R_{\min} + \gamma + \omega \tag{C.47}$$

which are compatible with Assumption 4.1.

By construction of $\mathbf{s}$, we know that $|s_b - s_b^0| > |s_1 - s_1^0| = 0$. Then, without loss of generality, we suppose that $s_b^0 - s_b > s_1^0 - s_1 = 0$ and $s_b^0 - s_b > \beta$.

$$\mathbb{P}\left(\ell\left(\mathbf{s}, \upsilon_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) \geq o \middle| \upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a\in\mathcal{A}}\left(s_a^0 + \pi_{t,a}\right)\right)$$

$$= \mathbb{P}\left(\max_{a\in\mathcal{A}}\left(s_a + \pi_{t,a}\right) - s_{\upsilon_t(\boldsymbol{\pi}_t)} - \pi_{t,\upsilon_t(\boldsymbol{\pi}_t)} \geq o \middle| \upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a\in\mathcal{A}}\{s_a^0 + \pi_{t,a}\}\right) \tag{C.48}$$

$$\geq \mathbb{P}\Bigg(\bigcup_{x\in\mathcal{A}, y\in\mathcal{A}, y\neq x} x = \arg\max_{a\in\mathcal{A}}\left(s_a^0 + \pi_{t,a}\right), y = \arg\max_{a\in\mathcal{A}}\left(s_a + \pi_{t,a}\right), s_y + \pi_{t,y} - s_x - \pi_{t,x} \geq o\Bigg) \tag{C.49}$$

$$\geq \mathbb{P}\left(1 = \arg\max_{a\in\mathcal{A}}(s_a + \pi_{t,a}), b = \arg\max_{a\in\mathcal{A}}(s_a^0 + \pi_{t,a}), s_1 + \pi_{t,1} - s_b - \pi_{t,b} \geq o\right) \tag{C.50}$$

$$\geq \mathbb{P}\Bigg(1 = \arg\max_{a\in\mathcal{A}}(s_a + \pi_{t,a}), b = \arg\max_{a\in\mathcal{A}}(s_a^0 + \pi_{t,a}), s_1 + \pi_{t,1} - s_b - \pi_{t,b} \geq o \Bigg| (C.45) - (C.47)\Bigg) \cdot \mathbb{P}\left((C.45) - (C.47)\right) \tag{C.51}$$

$$= \mathbb{P}\left(s_b - s_1 + o \leq \pi_{t,1} - \pi_{t,b} < s_b^0 - s_1^0\right) \cdot \mathbb{P}\left(C.46\right) \cdot \mathbb{P}\left(C.47\right) \cdot \prod_{a\in\mathcal{A}\setminus\{1,b\}} \mathbb{P}\left(\pi_{t,a} < R_{\min} + \gamma + \beta - d\right) \tag{C.52}$$

$$\geq \mathbb{P}\left(s_b - s_1 + o \leq \pi_{t,1} - \pi_{t,b} < s_b^0 - s_1^0\right) \cdot \mathbb{P}\left(C.46\right) \cdot \mathbb{P}\left(C.47\right) \cdot \prod_{a\in\mathcal{A}\setminus\{1,b\}} \mathbb{P}\left(\pi_{t,a} \leq R_{\min} + \gamma\right) \tag{C.53}$$

$$= \mathbb{P}\left(s_b - s_1 + o \leq \pi_{t,1} - \pi_{t,b} < s_b^0 - s_1^0\right)\left(1 - \frac{\gamma + \beta - d}{\overline{C} - \underline{C}}\right)\left(1 - \frac{\gamma + \omega}{\overline{C} - \underline{C}}\right) \cdot \prod_{a\in\mathcal{A}\setminus\{1,b\}} \frac{\gamma}{\overline{C} - \underline{C}} \tag{C.54}$$

where (C.52) follows as $\pi_{t,a}$'s are independent random variables and (C.54) follows since we assume that $\underline{C} = R_{\min}$ by Assumption 4.1 and $\pi_{t,a} \sim \mathcal{U}(\underline{C}, \overline{C}), \forall a \in \mathcal{A}$.

Then, by using similar arguments as in (C.27)-(C.43) from the proof of Proposition 4.2, we get the following lower bound for the first term in (C.54)

$$\mathbb{P}\left(s_b - s_1 + o \leq \pi_{t,1} - \pi_{t,b} < s_b^0 - s_1^0\right) \geq \mathbb{P}\left(s_b^0 - \beta + o \leq \pi_{t,1} - \pi_{t,b} < s_b^0\right) \geq \frac{(\beta - o)^2}{(\overline{C} - \underline{C})^2} \tag{C.55}$$

and obtain the desired result.

$$\mathbb{P}\left(\ell\left(\mathbf{s}, \upsilon_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) \geq o \middle| \upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a\in\mathcal{A}}\left(s_a^0 + \pi_{t,a}\right)\right)$$
$$\geq \frac{(\beta-o)^2}{(\overline{C}-\underline{C})^2}\left(1-\frac{\gamma+\beta-d}{\overline{C}-\underline{C}}\right)\left(1-\frac{\gamma+\omega}{\overline{C}-\underline{C}}\right)\left(\frac{\gamma}{\overline{C}-\underline{C}}\right)^{n-2} \tag{C.56}$$

□

Proof of Proposition 4.4. We derive the desired lower bound by conditioning on the case when the imperfect-knowledge agent selects the true maximizer arm at time $t \in \mathcal{T}$.

$$\mathbb{P}\left(\ell\left(\mathbf{s}, \upsilon_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) \geq o\right)$$
$$\geq \mathbb{P}\left(\ell\left(\mathbf{s}, \upsilon_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) \geq o \middle| \upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a\in\mathcal{A}}(s_a^0 + \pi_{t,a})\right)\mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a\in\mathcal{A}}(s_a^0 + \pi_{t,a})\right) \tag{C.57}$$
$$= \mathbb{P}\left(\ell\left(\mathbf{s}, \upsilon_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) \geq o \middle| \upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a\in\mathcal{A}}(s_a^0 + \pi_{t,a})\right)(1-p_t) \tag{C.58}$$
$$\geq \mathbb{P}\left(\ell\left(\mathbf{s}, \upsilon_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) \geq o \middle| \upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a\in\mathcal{A}}(s_a^0 + \pi_{t,a})\right)\left(1-k\frac{\sqrt{\log 2t}}{\sqrt{t}}\right) \tag{C.59}$$

where the last inequality follows by Assumption 4.2. Then, we observe that the following three conditions are mutually exclusive events

i. $K^0 \cap K = \emptyset$

ii. $K^0 \cap K \neq \emptyset$ and $b \notin K^0 \cap K$

iii. $K^0 \cap K \neq \emptyset$ and $b \in K^0 \cap K$

which allows us to combine the results of Propositions 4.1-4.3 and obtain

$$(C.59) = \sum_{j\in\{i,ii,iii\}}\mathbb{P}\left(\ell\left(\mathbf{s}, \upsilon_t(\boldsymbol{\pi}_t), \boldsymbol{\pi}_t\right) \geq o,\ j \middle| \upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a\in\mathcal{A}}(s_a^0 + \pi_{t,a})\right)\left(1-k\frac{\sqrt{\log 2t}}{\sqrt{t}}\right) \tag{C.60}$$
$$\geq \alpha(\beta-o)^2\left(1-k\frac{\sqrt{\log 2t}}{\sqrt{t}}\right) \tag{C.61}$$

for some constant $\alpha > 0$. □

Proof of Proposition 4.5. We derive the given concentration bound by using the bounded differences inequality (i.e., McDiarmid's inequality) (Boucheron et al. 2013). So, we begin by showing that the loss function $L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)$ has the *bounded differences property.*

We first note that the single-step loss function

$$\ell\left(\mathbf{s}, \upsilon_\tau(\boldsymbol{\pi}_\tau), \boldsymbol{\pi}_\tau\right) = \max_{a\in\mathcal{A}} \left(s_a + \pi_{\tau,a} - s_{\upsilon_\tau(\boldsymbol{\pi}_\tau)} - \pi_{\tau,\upsilon_\tau(\boldsymbol{\pi}_\tau)}\right) \tag{C.62}$$

is bounded from below and above as

$$\begin{aligned}(2R_{\min} - R_{\max}) - (2R_{\max} + \gamma - R_{\min}) &\leq \ell\left(\mathbf{s}, \upsilon_\tau(\boldsymbol{\pi}_\tau), \boldsymbol{\pi}_\tau\right)\\ &\leq (2R_{\max} + \gamma - R_{\min}) - (2R_{\min} - R_{\max})\end{aligned} \tag{C.63}$$

$$3R_{\min} - 3R_{\max} - \gamma \leq \ell\left(\mathbf{s}, \upsilon_\tau(\boldsymbol{\pi}_\tau), \boldsymbol{\pi}_\tau\right) \leq 3R_{\max} - 3R_{\min} + \gamma \tag{C.64}$$

since $s_a \in \mathcal{S} = [R_{\min} - R_{\max}, R_{\max} - R_{\min}]$ and $\pi_{\tau,a} \in \mathcal{C} = [R_{\min}, R_{\max} + \gamma]$ for all $a \in \mathcal{A}, \tau \in \mathcal{T}$ by definition.

Now, recall that the sequence of incentives $\mathbf{\Pi}_t = \{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{\widetilde{\tau}}, \ldots, \boldsymbol{\pi}_{t-1}\} \in \mathcal{C}^{n\times(t-1)}$ includes $(t-1)$ vectors of dimension $n = |\mathcal{A}|$ by definition. For a particular $\widetilde{\tau} \in \{1, \ldots, t-1\}$, we define $\boldsymbol{\pi}'_{\widetilde{\tau}} = \{\pi_{\widetilde{\tau},1}, \ldots, \pi'_{\widetilde{\tau},p}, \ldots, \pi_{\widetilde{\tau},n}\}$ such that it has the same components with $\boldsymbol{\pi}_{\widetilde{\tau}}$ except the value at index p. Then, for $\mathbf{\Pi}'_t = \{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}'_{\widetilde{\tau}}, \ldots, \boldsymbol{\pi}_{t-1}\}$, we have

$$\begin{aligned}&\left|L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) - L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}'_t), \mathbf{\Pi}'_t\right)\right|\\ &= \left|\sum_{\tau\in\Lambda(\widetilde{k},t)} \ell\left(\mathbf{s}, \upsilon_\tau(\boldsymbol{\pi}_\tau), \boldsymbol{\pi}_\tau\right) - \ell\left(\mathbf{s}, \upsilon_\tau(\boldsymbol{\pi}'_\tau), \boldsymbol{\pi}'_\tau\right)\right| &\text{(C.65)}\\ &= \left|\ell\left(\mathbf{s}, \upsilon_{\widetilde{\tau}}(\boldsymbol{\pi}_{\widetilde{\tau}}), \boldsymbol{\pi}_{\widetilde{\tau}}\right) - \ell\left(\mathbf{s}, \upsilon_{\widetilde{\tau}}(\boldsymbol{\pi}'_{\widetilde{\tau}}), \boldsymbol{\pi}'_{\widetilde{\tau}}\right)\right| &\text{(C.66)}\\ &\leq 6R_{\max} - 6R_{\min} + 2\gamma &\text{(C.67)}\end{aligned}$$

where the last result follows by (C.64). Because this result holds for any $\widetilde{\tau} \in \{1, \ldots, t-1\}$ and any index $p \in \{1, \ldots, n\}$, it shows that the bounded differences property holds for $L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)$. Therefore, we can directly use the bounded differences inequality (Boucheron et al. 2013) to obtain

$$\begin{aligned}\mathbb{P}\Big(L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) - \mathbb{E}L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) \geq \nu\Big)\\ \leq \exp\left(-\frac{2\nu^2}{(\eta(\widetilde{k},t)-1)n\left(6R_{\max} - 6R_{\min} + 2\gamma\right)^2}\right)\end{aligned} \tag{C.68}$$

for any $\nu > 0$. Because the bounded differences property $(C.67)$ is symmetric, the loss $L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)$ also satisfies the lower-tail inequality

$$\begin{aligned}\mathbb{P}\Big(L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) - \mathbb{E}L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) \leq -\nu\Big)\\ \leq \exp\left(-\frac{2\nu^2}{(\eta(\widetilde{k},t)-1)n\left(6R_{\max} - 6R_{\min} + 2\gamma\right)^2}\right)\end{aligned} \tag{C.69}$$

for any $\nu > 0$. Lastly, combining the last two inequalities gives us desired the concentration bound and completes our proof. □

Proof of Proposition 4.6. First, we recall the finite subcover $\{\mathcal{B}(\mathbf{s}^j, d) : \mathbf{s}^j \in \mathcal{F}\}_{j=1}^q$ of a collection of open balls covering $\mathcal{F}$ for finite $q > 0$ and $d < \beta$. Now, we also define the vector $\bar{\mathbf{s}}_t^j = \arg\sup_{\mathbf{s}\in\mathcal{B}(\mathbf{s}^j,d)} \left| L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) - \mathbb{E}L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)\right|$. Then, we have

$$\begin{aligned}
&\sup_{\mathbf{s}\in\mathcal{F}} \left| L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) - \mathbb{E}L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)\right| \\
&\leq \max_{j\in[q]} \sup_{\mathbf{s}\in\mathcal{B}(\mathbf{s}^j,d)} \left| L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) - \mathbb{E}L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)\right| &\text{(C.70)}\\
&= \max_{j\in[q]} \left| L^{\Lambda(\widetilde{k},t)}\left(\bar{\mathbf{s}}_t^j, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) - \mathbb{E}L^{\Lambda(\widetilde{k},t)}\left(\bar{\mathbf{s}}_t^j, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)\right| &\text{(C.71)}
\end{aligned}$$

where $[q] = \{1, \ldots, q\}$, and the first inequality follows because $\mathcal{F} \subseteq \bigcup_{j=1}^q \mathcal{B}(\mathbf{s}^j, d)$ by construction. We then follow by

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\mathbf{s}\in\mathcal{F}} \left| L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) - \mathbb{E}L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)\right| \geq \nu\right) \\
&\leq \mathbb{P}\left(\max_{j\in[q]} \left| L^{\Lambda(\widetilde{k},t)}\left(\bar{\mathbf{s}}_t^j, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) - \mathbb{E}L^{\Lambda(\widetilde{k},t)}\left(\bar{\mathbf{s}}_t^j, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)\right| \geq \nu\right) &\text{(C.72)}\\
&\leq \mathbb{P}\Bigg(\bigcup_{j\in[q]} \left| L^{\Lambda(\widetilde{k},t)}\left(\bar{\mathbf{s}}_t^j, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) - \mathbb{E}L^{\Lambda(\widetilde{k},t)}\left(\bar{\mathbf{s}}_t^j, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)\right| \geq \nu\Bigg) &\text{(C.73)}\\
&\leq \sum_{j\in[q]} \mathbb{P}\left(\left| L^{\Lambda(\widetilde{k},t)}\left(\bar{\mathbf{s}}_t^j, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) - \mathbb{E}L^{\Lambda(\widetilde{k},t)}\left(\bar{\mathbf{s}}_t^j, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)\right| \geq \nu\right) &\text{(C.74)}\\
&\leq \sum_{j\in[q]} 2\exp\left(-\frac{2\nu^2}{(\eta(\widetilde{k},t)-1)n(6R_{\max} - 6R_{\min} + 2\gamma)^2}\right) &\text{(C.75)}\\
&= 2q\exp\left(-\frac{2\nu^2}{(\eta(\widetilde{k},t)-1)n(6R_{\max} - 6R_{\min} + 2\gamma)^2}\right) &\text{(C.76)}
\end{aligned}$$

where (C.74) follows by Boole's inequality (i.e., union bound) and (C.75) follows by Proposition 4.5. Note that Proposition 4.5 holds for any vector $\mathbf{s} \in \mathcal{S}$, and hence, it also holds for $\bar{\mathbf{s}}_t^j$.

Lastly, it remains to provide an upper bound for the covering number q. We compute this bound by using the volume ratios and definition of the set $\mathcal{S} = [R_{\min} - R_{\max}, R_{\max} - R_{\min}]$.

$$q = \mathcal{N}(d, \mathcal{F}, \|\cdot\|) \leq \frac{\mathrm{vol}(\mathcal{F})}{\mathrm{vol}(\mathcal{B}(\mathbf{s}^j, d))} \leq \frac{\mathrm{vol}(\mathcal{S}^n)}{\mathrm{vol}(\mathcal{B}(\mathbf{s}^j, d))} \leq \frac{(R_{\max} - R_{\min})^n}{d^n} \tag{C.77}$$

Suppose that we have $d = \sqrt[n]{\beta}$. Then, combining (C.76) and (C.77) gives us the desired result.

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\mathbf{s}\in\mathcal{F}}\left|L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s},\Upsilon_t(\mathbf{\Pi}_t),\mathbf{\Pi}_t\right)-\mathbb{E}L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s},\Upsilon_t(\mathbf{\Pi}_t),\mathbf{\Pi}_t\right)\right|\geq\nu\right)\\
&\leq 2\frac{(R_{\max}-R_{\min})^n}{\beta}\exp\left(-\frac{2\nu^2}{(\eta(\widetilde{k},t)-1)n(6R_{\max}-6R_{\min}+2\gamma)^2}\right) && \text{(C.78)}\\
&=2\exp\left(-\frac{2\nu^2}{(\eta(\widetilde{k},t)-1)n(6R_{\max}-6R_{\min}+2\gamma)^2}-\log\beta+n\log(R_{\max}-R_{\min})\right) && \text{(C.79)}
\end{aligned}$$

□

Proof of Lemma 4.1. The given lower bound can be derived by considering the time steps $\tau\in\Lambda(\widetilde{k},t)$ where the agent selects a reward-maximizer arm $\upsilon_\tau(\boldsymbol{\pi}_\tau)=\arg\max_{a\in\mathcal{A}} s_a^0+\pi_{\tau,a}$. Further, recall that the set $\Lambda(\widetilde{k},t)$ consists of random time points which makes $\eta(\widetilde{k},t)$ a random variable. Thus, we start by computing a lower bound for the conditional expected loss given the set $\Lambda(\widetilde{k},t)$.

$$\begin{aligned}
&\mathbb{E}\left[L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}_t^{\mathcal{F}},\Upsilon_t(\mathbf{\Pi}_t),\mathbf{\Pi}_t\right)\middle|\Lambda(\widetilde{k},t)\right] && \text{(C.80)}\\
&=\mathbb{E}\Bigg[\sum_{\tau\in\Lambda(\widetilde{k},t)}\ell\left(\mathbf{s}_t^{\mathcal{F}},\upsilon_\tau(\boldsymbol{\pi}_\tau),\boldsymbol{\pi}_\tau\right)\Big|\Lambda(\widetilde{k},t)\Bigg] && \text{(C.81)}\\
&\geq\mathbb{E}\Bigg[\sum_{\tau\in\Lambda(\widetilde{k},t)}\ell\left(\mathbf{s}_t^{\mathcal{F}},\upsilon_\tau(\boldsymbol{\pi}_\tau),\boldsymbol{\pi}_\tau\right)\cdot\mathbf{1}\Big\{\upsilon_\tau(\boldsymbol{\pi}_\tau)=\arg\max_{a\in\mathcal{A}}s_a^0+\pi_{\tau,a}\Big\}\Big|\Lambda(\widetilde{k},t)\Bigg] && \text{(C.82)}\\
&\geq\sum_{\tau\in\Lambda(\widetilde{k},t)}o\mathbb{P}\left(\ell\left(\mathbf{s}_t^{\mathcal{F}},\upsilon_\tau(\boldsymbol{\pi}_\tau),\boldsymbol{\pi}_\tau\right)\geq o\right)(1-p_\tau) && \text{(C.83)}\\
&\geq\alpha o(\beta-o)^2\sum_{\tau\in\Lambda(\widetilde{k},t)}\left(1-k\frac{\sqrt{\log 2\tau}}{\sqrt{\tau}}\right)^2 && \text{(C.84)}\\
&\geq\alpha o(\beta-o)^2\sum_{\tau\in\Lambda(\widetilde{k},t)}\left(1-k\sqrt{\log 2\widetilde{k}}/\sqrt{\widetilde{k}}\right)^2 && \text{(C.85)}\\
&=\alpha\left(1-k\sqrt{\log 2\widetilde{k}}/\sqrt{\widetilde{k}}\right)^2 o(\beta-o)^2\eta(\widetilde{k},t) && \text{(C.86)}
\end{aligned}$$

where (C.83) follows by considering whether the single-step loss function $\ell\left(\mathbf{s}_t^{\mathcal{F}},\upsilon_\tau(\boldsymbol{\pi}_\tau),\boldsymbol{\pi}_\tau\right)$ is zero or strictly positive, and (C.84) follows by Assumption 4.2 and Proposition 4.4 for any $o\in(0,\beta)$.

Next, we take the expectation of both sides of the last inequality and get

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}_t^{\mathcal{F}},\Upsilon_t(\mathbf{\Pi}_t),\mathbf{\Pi}_t\right)\middle|\Lambda(\widetilde{k},t)\right]\right] &= \mathbb{E}L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}_t^{\mathcal{F}},\Upsilon_t(\mathbf{\Pi}_t),\mathbf{\Pi}_t\right)\\ &\geq \alpha\left(1-k\sqrt{\log 2\widetilde{k}}/\sqrt{\widetilde{k}}\right)^2 o(\beta-o)^2\mathbb{E}\eta(\widetilde{k},t)\end{aligned} \tag{C.87}$$

Lastly, we substitute $o=\beta/3$ to maximize the lower bound that we have in the last line above and conclude our proof by

$$\mathbb{E}L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}_t^{\mathcal{F}},\Upsilon_t(\mathbf{\Pi}_t),\mathbf{\Pi}_t\right) \geq \frac{4\alpha\left(1-k\sqrt{\log 2\widetilde{k}}/\sqrt{\widetilde{k}}\right)^2}{27}\beta^3\mathbb{E}\eta(\widetilde{k},t) \tag{C.88}$$

□

Proof of Lemma 4.2. First, we recall that the single-step loss $\ell\left(\mathbf{s}^0,\upsilon_\tau(\boldsymbol{\pi}_\tau),\boldsymbol{\pi}_\tau\right)$ becomes 0 when the agent selects the true maximizer arm (i.e., $\upsilon_\tau(\boldsymbol{\pi}_\tau)=\arg\max_{a\in\mathcal{A}} s_a^0+\pi_{\tau,a}$). Thus, in this proof, it suffices to consider only the time steps where the agent selects an arbitrary non-maximizer arm $\upsilon_\tau(\boldsymbol{\pi}_\tau)\in\mathcal{A}$. Because $\upsilon_\tau(\boldsymbol{\pi}_\tau)$ is not the true maximizer arm, we have $s^0_{\upsilon_\tau(\boldsymbol{\pi}_\tau)}+\pi_{\tau,\upsilon_\tau(\boldsymbol{\pi}_\tau)}<s_a^0+\pi_{\tau,a}$ for some $a\in\mathcal{A}\setminus\{\upsilon_\tau(\boldsymbol{\pi}_\tau)\}$. Further, we know that by definition

$$\ell\left(\mathbf{s}^0,\upsilon_\tau(\boldsymbol{\pi}_\tau),\boldsymbol{\pi}_\tau\right)=\max_{a\in\mathcal{A}}\left(s_a^0+\pi_{\tau,a}\right)-s^0_{\upsilon_\tau(\boldsymbol{\pi}_\tau)}-\pi_{\tau,\upsilon_\tau(\boldsymbol{\pi}_\tau)}\leq 3(R_{\max}-R_{\min})+\gamma \tag{C.89}$$

since $s_a\in\mathcal{S}=[R_{\min}-R_{\max},R_{\max}-R_{\min}]$ and $\pi_{\tau,a}\in\mathcal{C}=[R_{\min},R_{\max}+\gamma]$ for all a,τ. Then, we have

$$\begin{aligned}\mathbb{E}L\left(\mathbf{s}^0,\Upsilon_t(\mathbf{\Pi}_t),\mathbf{\Pi}_t\right) &= \sum_{\tau=1}^{t-1}\mathbb{E}\ell\left(\mathbf{s}^0,\upsilon_\tau(\boldsymbol{\pi}_\tau),\boldsymbol{\pi}_\tau\right) &\text{(C.90)}\\ &= \sum_{\tau=1}^{t-1}\mathbb{E}\ell\left(\mathbf{s}^0,\upsilon_\tau(\boldsymbol{\pi}_\tau),\boldsymbol{\pi}_\tau\right)\mathbf{1}\Big\{\upsilon_\tau(\boldsymbol{\pi}_\tau)\neq\arg\max_{a\in\mathcal{A}}s_a^0+\pi_{\tau,a}\Big\} &\text{(C.91)}\\ &\leq (3(R_{\max}-R_{\min})+\gamma)\sum_{\tau=1}^{t-1}p_\tau &\text{(C.92)}\\ &\leq (3(R_{\max}-R_{\min})+\gamma)\sum_{\tau=1}^{t-1}k\frac{\sqrt{\log 2\tau}}{\sqrt{\tau}} &\text{(C.93)}\\ &\leq k\left(3(R_{\max}-R_{\min})+\gamma\right)\left(\sqrt{\log 2}+\int_{\tau=1}^{t-1}\frac{\sqrt{\log 2\tau}}{\sqrt{\tau}}d\tau\right) &\text{(C.94)}\\ &= k\left(3(R_{\max}-R_{\min})+\gamma\right)\left(\sqrt{\log 2}+\sqrt{2}\sqrt{(2t-2)\log(2t-2)}\right) &\text{(C.95)}\end{aligned}$$

$$\leq k\left(3(R_{\max}-R_{\min})+\gamma\right)\left(\sqrt{\log 2}+2\sqrt{t\log(2t)}\right) \tag{C.96}$$

$$\leq 3k\left(3(R_{\max}-R_{\min})+\gamma\right)\sqrt{t\log(2t)} \tag{C.97}$$

where (C.93) follows by Assumption 4.2. □

Proof of Lemma 4.3. Similar to our argument in the proof of Lemma 4.2, we note that the single-step loss $\ell\left(\mathbf{s}^0, \upsilon_\tau(\boldsymbol{\pi}_\tau), \boldsymbol{\pi}_\tau\right)$ becomes 0 when the agent selects the true reward-maximizer arm, $\upsilon_\tau(\boldsymbol{\pi}_\tau) = \arg\max_{a\in\mathcal{A}} s_a^0 + \pi_{\tau,a}$. Therefore, we only need to bound the single-step loss at the time steps where the agent selects an arbitrary non-maximizer arm $\upsilon_\tau(\boldsymbol{\pi}_\tau) \in \mathcal{A}$.

$$L\left(\mathbf{s}^0, \Upsilon_t(\boldsymbol{\Pi}_t), \boldsymbol{\Pi}_t\right)$$

$$= \sum_{\tau=1}^{t-1} \ell\left(\mathbf{s}^0, \upsilon_\tau(\boldsymbol{\pi}_\tau), \boldsymbol{\pi}_\tau\right) \mathbf{1}\Big\{\upsilon_\tau(\boldsymbol{\pi}_\tau) \neq \arg\max_{a\in\mathcal{A}} s_a^0 + \pi_{\tau,a}\Big\} \tag{C.98}$$

$$= \sum_{\tau=1}^{t-1} \max_{a\in\mathcal{A}} \left\{s_a^0 + \pi_{\tau,a} - s^0_{\upsilon_\tau(\boldsymbol{\pi}_\tau)} - \pi_{\tau,\upsilon_\tau(\boldsymbol{\pi}_\tau)}\right\} \mathbf{1}\Big\{\upsilon_\tau(\boldsymbol{\pi}_\tau) \neq \arg\max_{a\in\mathcal{A}} s_a^0 + \pi_{\tau,a}\Big\} \tag{C.99}$$

We next compute the concentration inequality for the sum of identity functions in the last expression above which are independent Bernoulli variables.

$$\mathbb{P}\left(\sum_{\tau=1}^{t-1} \mathbf{1}\Big\{\upsilon_\tau(\boldsymbol{\pi}_\tau) \neq \arg\max_{a\in\mathcal{A}} s_a^0 + \pi_{\tau,a}\Big\} - \mathbb{E}\sum_{\tau=1}^{t-1} \mathbf{1}\Big\{\upsilon_\tau(\boldsymbol{\pi}_\tau) \neq \arg\max_{a\in\mathcal{A}} s_a^0 + \pi_{\tau,a}\Big\} \geq \overline{\nu}\right)$$
$$\leq \exp\left(-\frac{2\overline{\nu}^2}{t-1}\right) \tag{C.100}$$

where the last inequality follows by Hoeffding's Inequality (Boucheron et al. 2013) for any $\overline{\nu} > 0$. Here, we note that $\max_{a\in\mathcal{A}}\Big\{s_a^0 + \pi_{\tau,a} - s^0_{\upsilon_\tau(\boldsymbol{\pi}_\tau)} - \pi_{\tau,\upsilon_\tau(\boldsymbol{\pi}_\tau)}\Big\}$ for all $\tau \leq t-1$ is a constant value for a given sequence of incentives $\boldsymbol{\Pi}_t = \{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_{t-1}\}$. Further, we know that $\max_{a\in\mathcal{A}}\Big\{s_a^0 + \pi_{\tau,a} - s^0_{\upsilon_\tau(\boldsymbol{\pi}_\tau)} - \pi_{\tau,\upsilon_\tau(\boldsymbol{\pi}_\tau)}\Big\} \leq 3R_{\max} - 3R_{\min} + \gamma$ since $s_a \in \mathcal{S} = [R_{\min} - R_{\max}, R_{\max} - R_{\min}]$ and $\pi_{\tau,a} \in \mathcal{C} = [R_{\min}, R_{\max} + \gamma]$ for all $a \in \mathcal{A}, \tau \in \mathcal{T}$. Then, the last result implies

$$\mathbb{P}\left(L\left(\mathbf{s}^0, \Upsilon_t(\boldsymbol{\Pi}_t), \boldsymbol{\Pi}_t\right) - \mathbb{E}L\left(\mathbf{s}^0, \Upsilon_t(\boldsymbol{\Pi}_t), \boldsymbol{\Pi}_t\right) \geq \left(3R_{\max} - 3R_{\min} + \gamma\right)\overline{\nu}\right) \leq \exp\left(-\frac{2\overline{\nu}^2}{t-1}\right) \tag{C.101}$$

Last, replacing $\nu = \left(3R_{\max} - 3R_{\min} + \gamma\right)\overline{\nu}$ for any $\overline{\nu} > 0$, we obtain the concentration inequality for $L\left(\mathbf{s}^0, \Upsilon_t(\boldsymbol{\Pi}_t), \boldsymbol{\Pi}_t\right)$ as

$$\mathbb{P}\left(L\left(\mathbf{s}^0, \Upsilon_t(\boldsymbol{\Pi}_t), \boldsymbol{\Pi}_t\right) - \mathbb{E}L\left(\mathbf{s}^0, \Upsilon_t(\boldsymbol{\Pi}_t), \boldsymbol{\Pi}_t\right) \geq \nu\right) \leq \exp\left(-\frac{2\nu^2}{(t-1)\left(3R_{\max} - 3R_{\min} + \gamma\right)^2}\right) \tag{C.102}$$

for any $\nu > 0$. □

Proof of Theorem 4.1. We let $\mathbf{s}_t^{\mathcal{F}} = \arg\inf_{\mathbf{s}\in\mathcal{F}} L(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t)$ for notational convenience. Then, we begin by considering any constant $\nu > 0$ such that

$$3\nu < \mathbb{E}L^{\Lambda(\widetilde{k},t)}(\mathbf{s}_t^{\mathcal{F}}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t) - \mathbb{E}L(\mathbf{s}^0, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t) \tag{C.103}$$

where $L^{\Lambda(\widetilde{k},t)}(\mathbf{s}_t^{\mathcal{F}}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t)$ is as introduced in (4.12). For the ν being considered,

$$\begin{aligned} &\text{If } L(\mathbf{s}^0, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t) < \mathbb{E}L(\mathbf{s}^0, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t) + \nu \\ &\text{and } L^{\Lambda(\widetilde{k},t)}(\mathbf{s}_t^{\mathcal{F}}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t) > \mathbb{E}L^{\Lambda(\widetilde{k},t)}(\mathbf{s}_t^{\mathcal{F}}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t) - \nu \end{aligned}$$

then it follows that $L(\mathbf{s}^0, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t) < L^{\Lambda(\widetilde{k},t)}(\mathbf{s}_t^{\mathcal{F}}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t)$. Now, we take the contrapositive of this 'if' statement:
If we have $L^{\Lambda(\widetilde{k},t)}(\mathbf{s}_t^{\mathcal{F}}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t) \leq L\left(\mathbf{s}^0, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)$, then either $L(\mathbf{s}^0, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t) \geq \mathbb{E}L(\mathbf{s}^0, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t) + \nu$ or $L^{\Lambda(\widetilde{k},t)}(\mathbf{s}_t^{\mathcal{F}}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t) \leq \mathbb{E}L^{\Lambda(\widetilde{k},t)}(\mathbf{s}_t^{\mathcal{F}}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t) - \nu$ will hold for the ν being considered.

Further, we know that $L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) \leq L\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)$ for any $\mathbf{s} \in \mathcal{S}$ and all $t \in \mathcal{T}$ by definition. Thus, we have

$$\begin{align} &\mathbb{P}\left(L(\mathbf{s}_t^{\mathcal{F}}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t) \leq L(\mathbf{s}^0, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t)\right) \notag \\ &\quad \leq \mathbb{P}\left(L^{\Lambda(\widetilde{k},t)}(\mathbf{s}_t^{\mathcal{F}}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t) \leq L(\mathbf{s}^0, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t)\right) \tag{C.104} \\ &\quad \leq \mathbb{P}\left(L(\mathbf{s}^0, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t) \geq \mathbb{E}L(\mathbf{s}^0, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t) + \nu\right) \notag \\ &\qquad + \mathbb{P}\left(L^{\Lambda(\widetilde{k},t)}(\mathbf{s}_t^{\mathcal{F}}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t) \leq \mathbb{E}L^{\Lambda(\widetilde{k},t)}(\mathbf{s}_t^{\mathcal{F}}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t) - \nu\right) \tag{C.105} \\ &\quad \leq \exp\left(-\frac{2\nu^2}{(t-1)\left(3R_{\max} - 3R_{\min} + \gamma\right)^2}\right) \notag \\ &\qquad + \frac{(R_{\max} - R_{\min})^n}{\beta} \exp\left(-\frac{2\nu^2}{(\eta(\widetilde{k},t) - 1)n(6R_{\max} - 6R_{\min} + 2\gamma)^2}\right) \tag{C.106} \\ &\quad \leq \exp\left(-\frac{2\nu^2}{(t-1)n\left(6R_{\max} - 6R_{\min} + 2\gamma\right)^2}\right) \notag \\ &\qquad + \frac{(R_{\max} - R_{\min})^n}{\beta} \exp\left(-\frac{2\nu^2}{(t-1)n(6R_{\max} - 6R_{\min} + 2\gamma)^2}\right) \tag{C.107} \\ &\quad = \exp\left(-\frac{2\nu^2}{(t-1)n\left(6R_{\max} - 6R_{\min} + 2\gamma\right)^2}\right)\left(1 + \frac{(R_{\max} - R_{\min})^n}{\beta}\right) \tag{C.108} \\ &\quad \leq \exp\left(-\frac{2\nu^2}{(t-1)n\left(6R_{\max} - 6R_{\min} + 2\gamma\right)^2}\right)\left(1 + \frac{(2R_{\max} - 2R_{\min})^n}{\beta}\right) \tag{C.109} \end{align}$$

where (C.106) follows by Lemma 4.3 and Proposition 4.6. Notice that $\beta < 2R_{\max} - 2R_{\min}$ by construction.

$$\leq 2\exp\left(-\frac{2\nu^2}{(t-1)n(6R_{\max}-6R_{\min}+2\gamma)^2}\right)\frac{(2R_{\max}-2R_{\min})^n}{\beta} \tag{C.110}$$

$$= 2\exp\left(-\frac{2\nu^2}{(t-1)n(6R_{\max}-6R_{\min}+2\gamma)^2} - \log\beta + n\log(2R_{\max}-2R_{\min})\right) \tag{C.111}$$

for the ν being considered. Suppose $\nu = (\mathbb{E}L^{\Lambda(\widetilde{k},t)}(\mathbf{s}_t^{\mathcal{F}}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t) - \mathbb{E}L(\mathbf{s}^0, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t))/4$. Then,

$$= 2\exp\Bigg(-\frac{2(\mathbb{E}L^{\Lambda(\widetilde{k},t)}(\mathbf{s}_t^{\mathcal{F}}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t) - \mathbb{E}L(\mathbf{s}^0, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t))^2}{(t-1)16n(6R_{\max}-6R_{\min}+2\gamma)^2} - \log\beta + n\log(2R_{\max}-2R_{\min})\Bigg) \tag{C.112}$$

$$\leq 2\exp\left(-\frac{2\lambda_t^2}{(t-1)16n(6R_{\max}-6R_{\min}+2\gamma)^2} - \log\beta + n\log(2R_{\max}-2R_{\min})\right) \tag{C.113}$$

where the last inequality follows since $\mathbb{E}L^{\Lambda(\widetilde{k},t)}\left(\mathbf{s}_t^{\mathcal{F}}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) - \mathbb{E}L\left(\mathbf{s}^0, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) \geq \lambda_t > 0$ by Lemma 4.1 and Lemma 4.2 for $t \in [\widetilde{k}, T]$. □

Proof of Corollary 4.1. We first recall that the principal's estimator $\widehat{\mathbf{s}}_t^{\text{pr}}$ (4.6) is defined such that it satisfies $L\left(\widehat{\mathbf{s}}_t^{\text{pr}}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) \leq L\left(\mathbf{s}^0, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)$. Then, we have the following implication

$$\left\{\|\mathbf{s}^0 - \widehat{\mathbf{s}}_t^{\text{pr}}\|_\infty > \beta\right\} \subseteq \left\{\exists \mathbf{s} : \|\mathbf{s}^0 - \mathbf{s}\|_\infty > \beta \text{ and } L\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) \leq L\left(\mathbf{s}^0, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)\right\} \tag{C.114}$$

$$\subseteq \left\{\inf_{\mathbf{s}\in\mathcal{F}} L\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) \leq L\left(\mathbf{s}^0, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)\right\} \tag{C.115}$$

which gives us the desired bound as follows.

$$\mathbb{P}\left(\|\mathbf{s}^0 - \widehat{\mathbf{s}}_t^{\text{pr}}\|_\infty > \beta\right)$$

$$\leq \mathbb{P}\left(\inf_{\mathbf{s}\in\mathcal{F}} L\left(\mathbf{s}, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right) \leq L\left(\mathbf{s}^0, \Upsilon_t(\mathbf{\Pi}_t), \mathbf{\Pi}_t\right)\right) \tag{C.116}$$

$$\leq 2\exp\left(-\frac{2\lambda_t^2}{(t-1)16n(6R_{\max}-6R_{\min}+2\gamma)^2} - \log\beta + n\log(2R_{\max}-2R_{\min})\right) \tag{C.117}$$

where the last inequality follows by Theorem 4.1. □

C.1.2 Results in Section 4.3

Proof of Proposition 4.7. We start by recalling the definition of j_t^* in Algorithm 3 (line 18) which implies

$$\begin{aligned}
\mathbb{P}\left(j_t^* \neq j^{*,0}\right) &\leq \mathbb{P}\Bigg(\bigcup_{j_t^*\in\mathcal{A}} \widetilde{V}(j^{*,0},\widehat{\mathbf{s}}_t^{\mathrm{pr}};\widehat{\boldsymbol{\theta}}_t) < \widetilde{V}(j_t^*,\widehat{\mathbf{s}}_t^{\mathrm{pr}};\widehat{\boldsymbol{\theta}}_t)\Bigg) && \text{(C.118)}\\
&\leq \sum_{j_t^*\in\mathcal{A}} \mathbb{P}\left(\widetilde{V}(j^{*,0},\widehat{\mathbf{s}}_t^{\mathrm{pr}};\widehat{\boldsymbol{\theta}}_t) < \widetilde{V}(j_t^*,\widehat{\mathbf{s}}_t^{\mathrm{pr}};\widehat{\boldsymbol{\theta}}_t)\right) && \text{(C.119)}\\
&= \sum_{j_t^*\in\mathcal{A}} \mathbb{P}\left(\widehat{\theta}_{t,j^{*,0}} - \widehat{\theta}_{t,j_t^*} < \widehat{s}^{\mathrm{pr}}_{t,j_t^*} - \widehat{s}^{\mathrm{pr}}_{t,j^{*,0}}\right) && \text{(C.120)}
\end{aligned}$$

Observe that by definition of $j^{*,0} = \arg\max_{j\in\mathcal{A}} \widetilde{V}(j,\mathbf{s}^0;\boldsymbol{\theta}^0) = \arg\max_{j\in\mathcal{A}} \theta_j^0 - (\max_{a\in\mathcal{A}} s_a^0) + s_j^0$, it follows that $\theta^0_{j_t^*} - \theta^0_{j^{*,0}} \leq s^0_{j^{*,0}} - s^0_{j_t^*}$. Then,

$$\begin{aligned}
&(C.120)\\
&= \sum_{j_t^*\in\mathcal{A}} \mathbb{P}\left((\widehat{\theta}_{t,j^{*,0}} - \theta^0_{j^{*,0}}) + (\theta^0_{j_t^*} - \widehat{\theta}_{t,j_t^*}) < (\widehat{s}^{\mathrm{pr}}_{t,j_t^*} - s^0_{j_t^*}) + (s^0_{j^{*,0}} - \widehat{s}^{\mathrm{pr}}_{t,j^{*,0}})\right) && \text{(C.121)}\\
&= \sum_{j_t^*\in\mathcal{A}} \mathbb{P}\left((\widehat{\theta}_{t,j^{*,0}} - \theta^0_{j^{*,0}}) + (\theta^0_{j_t^*} - \widehat{\theta}_{t,j_t^*}) < (\widehat{s}^{\mathrm{pr}}_{t,j_t^*} - s^0_{j_t^*}) + (s^0_{j^{*,0}} - \widehat{s}^{\mathrm{pr}}_{t,j^{*,0}})\Big| \|\mathbf{s}^0 - \widehat{\mathbf{s}}_t^{\mathrm{pr}}\|_\infty \leq \beta_t\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \cdot\, \mathbb{P}\left(\|\mathbf{s}^0 - \widehat{\mathbf{s}}_t^{\mathrm{pr}}\|_\infty \leq \beta_t\right)\\
&\qquad + \mathbb{P}\left((\widehat{\theta}_{t,j^{*,0}} - \theta^0_{j^{*,0}}) + (\theta^0_{j_t^*} - \widehat{\theta}_{t,j_t^*}) < (\widehat{s}^{\mathrm{pr}}_{t,j_t^*} - s^0_{j_t^*}) + (s^0_{j^{*,0}} - \widehat{s}^{\mathrm{pr}}_{t,j^{*,0}})\Big| \|\mathbf{s}^0 - \widehat{\mathbf{s}}_t^{\mathrm{pr}}\|_\infty > \beta_t\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \cdot\, \mathbb{P}\left(\|\mathbf{s}^0 - \widehat{\mathbf{s}}_t^{\mathrm{pr}}\|_\infty > \beta_t\right)\\
&\leq \sum_{j_t^*\in\mathcal{A}} \mathbb{P}\left((\widehat{\theta}_{t,j^{*,0}} - \theta^0_{j^{*,0}}) + (\theta^0_{j_t^*} - \widehat{\theta}_{t,j_t^*}) \leq 2\beta_t\right) + n\mathbb{P}\left(\|\mathbf{s}^0 - \widehat{\mathbf{s}}_t^{\mathrm{pr}}\|_\infty > \beta_t\right) && \text{(C.122)}\\
&\leq \sum_{j_t^*\in\mathcal{A}} \mathbb{P}\left((\widehat{\theta}_{t,j^{*,0}} - \theta^0_{j^{*,0}}) + (\theta^0_{j_t^*} - \widehat{\theta}_{t,j_t^*}) \leq 2\beta_t\right)\\
&\qquad + 2n\exp\left(-\frac{2\lambda_t^2}{(t-1)16n(6R_{\max} - 6R_{\min} + 2\gamma)^2} - \log\beta_t + n\log(2R_{\max} - 2R_{\min})\right) && \text{(C.123)}
\end{aligned}$$

where the last inequality follows by Theorem 4.1. Then, we can bound the first term in the last result as follows.

$$\begin{aligned}
&\sum_{j_t^*\in\mathcal{A}} \mathbb{P}\left((\widehat{\theta}_{t,j^{*,0}} - \theta^0_{j^{*,0}}) \leq 2\beta_t - (\theta^0_{j_t^*} - \widehat{\theta}_{t,j_t^*})\right)\\
&= \sum_{j_t^*\in\mathcal{A}} \mathbb{P}\left((\widehat{\theta}_{t,j^{*,0}} - \theta^0_{j^{*,0}}) \leq 2\beta_t - (\theta^0_{j_t^*} - \widehat{\theta}_{t,j_t^*})\Big| \theta^0_{j_t^*} - \widehat{\theta}_{t,j_t^*} < 3\beta_t\right) \mathbb{P}\left(\theta^0_{j_t^*} - \widehat{\theta}_{t,j_t^*} < 3\beta_t\right)
\end{aligned}$$

$$+ \mathbb{P}\left((\widehat{\theta}_{t,j^*,0} - \theta^0_{j^*,0}) \leq 2\beta_t - (\theta^0_{j^*_t} - \widehat{\theta}_{t,j^*_t}) \middle| \theta^0_{j^*_t} - \widehat{\theta}_{t,j^*_t} \geq 3\beta_t\right) \mathbb{P}\left(\theta^0_{j^*_t} - \widehat{\theta}_{t,j^*_t} \geq 3\beta_t\right) \tag{C.124}$$

$$\leq \sum_{j^*_t \in \mathcal{A}} \mathbb{P}\left(\widehat{\theta}_{t,j^*,0} - \theta^0_{j^*,0} \leq -\beta_t\right) + \mathbb{P}\left(\widehat{\theta}_{t,j^*_t} - \theta^0_{j^*_t} \leq -3\beta_t\right) \tag{C.125}$$

$$\leq \sum_{j^*_t \in \mathcal{A}} \mathbb{P}\left(\widehat{\theta}_{t,j^*,0} - \theta^0_{j^*,0} \leq -\beta_t\right) + \mathbb{P}\left(\widehat{\theta}_{t,j^*_t} - \theta^0_{j^*_t} \leq -\beta_t\right) \tag{C.126}$$

Note that bounding the two probability terms in the right-hand side of the last inequality follows a very similar approach to each other due to the definition of $\widehat{\theta}_{t,a}$'s given in (4.21). For any $a \in \mathcal{A}$, let $\overline{T}(a,t)$ be the number of time steps $\tau \in [\widetilde{k}, t-1]$ where: the agent picks the true reward-maximizer arm, the principal explores, and arm a is the true reward-maximizer arm for the agent. Then, $\overline{T}(a,t)$ is the sum of $t - \widetilde{k} - 1$ independent indicator variables that are independent Bernoulli random variables with success probabilities given as $(1-p_\tau) \cdot \epsilon^{\text{pr}}_\tau \cdot \mathbb{P}\left(a = \arg\max_{a' \in \mathcal{A}} s^0_{a'} + \pi_{\tau,a'}\right)$. Further, we note that $T(a,t) \geq \overline{T}(a,t)$. Then, for any $a \in \mathcal{A}$, we have

$$\mathbb{P}\left(\widehat{\theta}_{t,a} - \theta^0_a \leq -\beta_t\right)$$

$$\leq \mathbb{P}\left(\widehat{\theta}_{t,a} - \theta^0_a \leq -\beta_t \middle| \overline{T}(a,t) > \mathbb{E}\overline{T}(a,t)/2\right) \mathbb{P}\left(\overline{T}(a,t) > \mathbb{E}\overline{T}(a,t)/2\right)$$

$$+ \mathbb{P}\left(\widehat{\theta}_{t,a} - \theta^0_a \leq -\beta_t \middle| \overline{T}(a,t) \leq \mathbb{E}\overline{T}(a,t)/2\right) \mathbb{P}\left(\overline{T}(a,t) \leq \mathbb{E}\overline{T}(a,t)/2\right) \tag{C.127}$$

$$\leq \mathbb{P}\left(\widehat{\theta}_{t,a} - \theta^0_a \leq -\beta_t \middle| \overline{T}(a,t) > \mathbb{E}\overline{T}(a,t)/2\right) + \mathbb{P}\left(\overline{T}(a,t) \leq \mathbb{E}\overline{T}(a,t)/2\right) \tag{C.128}$$

Next, we use Hoeffding's Inequality (Boucheron et al. 2013) and obtain.

$$\leq \exp\left(-\frac{\mathbb{E}\overline{T}(a,t)\beta_t^2}{(\overline{C} - \underline{C})^2}\right) + \exp\left(-t\frac{\left(\mathbb{E}\overline{T}(a,t)\right)^2}{4}\right) \tag{C.129}$$

We recall that our research problems are well-posed by Assumption 4.1 which ensures that the principal is able to provide incentives whose magnitudes are sufficiently large to steer the agent's decisions. This further implies that the rewards of the principal should be large enough to compensate these incentives. Therefore, in the denominator of the first term above, we take the range of $\mu_{\tau,a}$ as $[\underline{C}, \overline{C}]$ in in accordance with Assumption 4.1.

The next step is to derive a lower bound for $\mathbb{E}\overline{T}(a,t)$ by using the definition of $\overline{T}(a,t)$ and Assumption 4.2.

$$\mathbb{E}\overline{T}(a,t) = \sum_{\tau=\widetilde{k}}^{t-1} (1-p_\tau) \cdot \epsilon^{\text{pr}}_\tau \cdot \mathbb{P}\left(a = \arg\max_{a' \in \mathcal{A}} s^0_{a'} + \pi_{\tau,a'}\right) \tag{C.130}$$

$$\geq \sum_{\tau=\widetilde{k}}^{t-1} \left(1 - k\frac{\sqrt{\log 2\tau}}{\sqrt{\tau}}\right) \frac{m^{\text{pr}}}{\tau^{(1/2-w)}} \mathbb{P}\left(a = \arg\max_{a' \in \mathcal{A}} s^0_{a'} + \pi_{\tau,a'}\right) \tag{C.131}$$

We compute a lower bound on the probability $\mathbb{P}\left(a = \arg\max_{a' \in \mathcal{A}} s^0_{a'} + \pi_{\tau,a'}\right)$ by using the cdf derived in (B.12). Since the cdf is a piecewise function, we can consider the case $\underline{C} - \overline{C} \leq s^0_a - s^0_{a'} < 0$ to derive a lower bound.

$$\begin{aligned}
\mathbb{P}\left(a = \arg\max_{a' \in \mathcal{A}} s^0_{a'} + \pi_{\tau,a'}\right) &= \mathbb{P}\left(s^0_{a'} + \pi_{\tau,a'} < s^0_a + \pi_{\tau,a},\ \forall a' \in \mathcal{A} \setminus \{a\}\right) && \text{(C.132)}\\
&= \prod_{a' \in \mathcal{A}\setminus\{a\}} \mathbb{P}\left(\pi_{\tau,a'} - \pi_{\tau,a} < s^0_a - s^0_{a'}\right) && \text{(C.133)}\\
&\geq \prod_{a' \in \mathcal{A}\setminus\{a\}} \mathbb{P}\left(\pi_{\tau,a'} - \pi_{\tau,a} < s^0_a - s^0_{a'},\ \underline{C} - \overline{C} \leq s^0_a - s^0_{a'} < 0\right) && \text{(C.134)}\\
&= \prod_{a' \in \mathcal{A}\setminus\{a\}} \frac{(s^0_a - s^0_{a'} + \overline{C} - \underline{C})^2}{2(\overline{C} - \underline{C})^2} && \text{(C.135)}\\
&\geq \prod_{a' \in \mathcal{A}\setminus\{a\}} \frac{(s^0_a + \gamma)^2}{2(\overline{C} - \underline{C})^2} && \text{(C.136)}\\
&= \frac{(s^0_a + \gamma)^{2n-2}}{2^{n-1}(\overline{C} - \underline{C})^{2n-2}} && \text{(C.137)}
\end{aligned}$$

where the last inequality follows since $s^0_{a'} \leq R_{\max} - R_{\min}$ for all $a' \in \mathcal{A}$ and $\overline{C} - \underline{C} = R_{\max} - R_{\min} + \gamma$ by definition. Then, we have

$$\begin{aligned}
\mathbb{E}\overline{T}(a,t) &\geq \frac{(s^0_a + \gamma)^{2n-2}}{2^{n-1}(\overline{C} - \underline{C})^{2n-2}} \sum_{\tau=\widetilde{k}}^{t-1} \left(1 - k\frac{\sqrt{\log 2\tau}}{\sqrt{\tau}}\right) \frac{m^{\text{pr}}}{\tau^{(1/2-w)}} && \text{(C.138)}\\
&\geq \frac{(s^0_a + \gamma)^{2n-2}}{2^{n-1}(\overline{C} - \underline{C})^{2n-2}} \left(1 - k\frac{\sqrt{\log 2\widetilde{k}}}{\sqrt{\widetilde{k}}}\right) m^{\text{pr}} \sum_{\tau=\widetilde{k}}^{t-1} \frac{1}{\tau^{(1/2-w)}} && \text{(C.139)}\\
&\geq \frac{(s^0_a + \gamma)^{2n-2}}{2^{n-1}(\overline{C} - \underline{C})^{2n-2}} \left(1 - k\frac{\sqrt{\log 2\widetilde{k}}}{\sqrt{\widetilde{k}}}\right) m^{\text{pr}} \int_{\tau=\widetilde{k}}^{t} \frac{1}{\tau^{(1/2-w)}} d\tau && \text{(C.140)}\\
&= \frac{(s^0_a + \gamma)^{2n-2}}{2^{n-1}(\overline{C} - \underline{C})^{2n-2}} \left(1 - k\frac{\sqrt{\log 2\widetilde{k}}}{\sqrt{\widetilde{k}}}\right) m^{\text{pr}} \frac{2\left(t^{w+1/2} - \widetilde{k}^{w+1/2}\right)}{2w+1} && \text{(C.141)}\\
&\geq \frac{(s^0_a + \gamma)^{2n-2}}{2^{n-1}(\overline{C} - \underline{C})^{2n-2}} \left(1 - k\frac{\sqrt{\log 2\widetilde{k}}}{\sqrt{\widetilde{k}}}\right) m^{\text{pr}} \frac{2 - 2w - 1}{2w+1} t^{w+1/2} && \text{(C.142)}
\end{aligned}$$

where the last inequality always holds for $t \geq \widetilde{k}$ since $0 < w < 1/4$ by definition. Combining this last result with (C.129), we get

$$\mathbb{P}\left(\widehat{\theta}_{t,a} - \theta_a^0 \leq -\beta_t\right)$$

$$\leq \exp\left(-\frac{\frac{(s_a^0+\gamma)^{2n-2}}{2^{n-1}(\overline{C}-\underline{C})^{2n-2}}\left(1-k\frac{\sqrt{\log 2\widetilde{k}}}{\sqrt{\widetilde{k}}}\right) m^{\mathrm{pr}}\frac{2-2w-1}{2w+1}t^{w+1/2}\beta_t^2}{(\overline{C}-\underline{C})^2}\right) + \frac{1}{t^{2w+2}} \tag{C.143}$$

We now substitute $\beta_t = B\frac{\sqrt{\log 2t}}{t^{w/3}}$ as given in Algorithm 3.

$$\leq \exp\left(-\frac{\frac{(s_a^0+\gamma)^{2n-2}}{2^{n-1}(\overline{C}-\underline{C})^{2n-2}}\left(1-k\frac{\sqrt{\log 2\widetilde{k}}}{\sqrt{\widetilde{k}}}\right) m^{\mathrm{pr}}\frac{2-2w-1}{2w+1}t^{w+1/2}\frac{9k^2(R_{\max}-R_{\min}+\gamma)^{2n}\sqrt[3]{32n}}{\left(1-k\sqrt{\log 2\widetilde{k}}/\sqrt{\widetilde{k}}\right)^2}t^{w+1/2}\frac{\log 2t}{t^{2w/3}}}{(\overline{C}-\underline{C})^2}\right) \tag{C.144}$$

$$\leq \exp\left(-\frac{(s_a^0+\gamma)^{2n-2}}{2^{n-1}}m^{\mathrm{pr}}\frac{2-2w-1}{2w+1}\frac{9k^2\sqrt[3]{32n}}{\left(1-k\sqrt{\log 2\widetilde{k}}/\sqrt{\widetilde{k}}\right)}t^{w+1/2}\frac{\log 2t}{t^{2w/3}}\right) \tag{C.145}$$

where the last inequality follows since $\overline{C}-\underline{C} = R_{\max} - R_{\min} + \gamma$ with $\gamma > 0$ by definition.

Now, notice that we can bound each of the constant terms in the last line from below by 1 since we have $n \geq 2$, $m^{\mathrm{pr}} \geq 1$, $0 < \gamma \leq R_{\max} - R_{\min} - 1$, $s_a^0 \geq R_{\min} - R_{\max}$, $0 < w < 1/4$, $k \geq 1$, and $k\sqrt{\log 2\widetilde{k}} < \sqrt{\widetilde{k}}$ by definition. Then,

$$\mathbb{P}\left(\widehat{\theta}_{t,a} - \theta_a^0 \leq -\beta_t\right) \leq \exp\left(-t^{w+1/2}\frac{\log 2t}{t^{2w/3}}\right) \leq \exp(-\log 2t) = \frac{1}{2t} \tag{C.146}$$

Substituting this upper bound in (C.126), we obtain

$$\sum_{j_t^* \in \mathcal{A}} \mathbb{P}\left((\widehat{\theta}_{t,j^*,0} - \theta_{j^*,0}^0) + (\theta_{j_t^*}^0 - \widehat{\theta}_{t,j_t^*}) \leq 2\beta_t\right) \leq \frac{n}{t} \tag{C.147}$$

Next, we compute an upper bound for the second part of (C.123). We recall the definition of λ_t as given in (4.18).

$$\lambda_t = \frac{4\alpha\left(1-k\sqrt{\log 2\widetilde{k}}/\sqrt{\widetilde{k}}\right)^2}{27}\beta_t^3\mathbb{E}\eta(\widetilde{k},t) - 3k\left(3(R_{\max}-R_{\min})+\gamma\right)\sqrt{t\log(2t)} \tag{C.148}$$

where $\eta(\widetilde{k}, t)$ is defined in (4.11) as the number of time steps within the time interval $[\widetilde{k}, t-1]$ where the principal chooses each incentive $\pi_{t,a}$ uniformly randomly from the compact set $\mathcal{C}$. By this definition, we can compute a lower bound for $\mathbb{E}\eta(\widetilde{k}, t)$ as follows.

$$\mathbb{E}\eta(\widetilde{k}, t) = \sum_{\tau=\widetilde{k}}^{t-1} \min\left\{1, m^{\text{pr}} \frac{1}{\tau^{(1/2-w)}}\right\} \geq m^{\text{pr}} \sum_{\tau=\widetilde{m}}^{t-1} \frac{1}{\tau^{(1/2-w)}} \geq m^{\text{pr}} \int_{\tau=\widetilde{m}}^{t} \frac{1}{\tau^{(1/2-w)}} d\tau \tag{C.149}$$

$$= \frac{2m^{\text{pr}}}{2w+1}\left(t^{w+1/2} - \widetilde{m}^{w+1/2}\right) \tag{C.150}$$

$$\geq \frac{2m^{\text{pr}} - 2w - 1}{2w+1} t^{w+1/2} \tag{C.151}$$

where $\widetilde{m} \geq \widetilde{k}$ is the minimum value satisfying $m^{\text{pr}} \leq \widetilde{m}^{(1/2-w)}$. Then, the last inequality always holds for $t \geq \widetilde{m}$, $0 < w < 1/4$, and $m^{\text{pr}} \geq 1$. Using this lower bound for $\mathbb{E}\eta(\widetilde{k}, t)$, we obtain

$$\lambda_t \geq \sqrt{t}\left(\frac{4\alpha\left(1 - k\sqrt{\log 2\widetilde{k}}/\sqrt{\widetilde{k}}\right)^2 (2m^{\text{pr}} - 2w - 1)}{27(2w+1)} \beta_t^3 t^w - 3k\left(3(R_{\max} - R_{\min}) + \gamma\right)\sqrt{\log 2t}\right) \tag{C.152}$$

For the specified choice of β_t, we further have

$$\geq 3k\left(3(R_{\max} - R_{\min}) + \gamma\right)\sqrt{t \log 2t}\left(\frac{4\alpha(2m^{\text{pr}} - 2w - 1)k^2(3(R_{\max} - R_{\min}) + \gamma)^{3n-1}\sqrt{32n}}{3(2w+1)\left(1 - k\sqrt{\log 2\widetilde{k}}/\sqrt{\widetilde{k}}\right)} \log(2t) - 1\right) \tag{C.153}$$

$$\geq 3k\left(3(R_{\max} - R_{\min}) + \gamma\right)\sqrt{t \log 2t}\left(\sqrt{32n}\log(2t) - 1\right) \tag{C.154}$$

$$\geq 3k\left(3(R_{\max} - R_{\min}) + \gamma\right)\sqrt{t \log 2t}\left(\sqrt{32n}\log(2t) - \frac{\sqrt{32n}}{3}\log(2t)\right) \tag{C.155}$$

$$\geq 2\sqrt{32n}k\left(3(R_{\max} - R_{\min}) + \gamma\right)\sqrt{t \log 2t}\log(2t) \tag{C.156}$$

where (C.154) follows by the fact that we can bound each constant term that appears in the coefficient of $\log(2t)$ from below by 1 since we have $n \geq 2$, $k \geq 1$, $0 < w < 1/4$, $m^{\text{pr}} \geq 1$, $k\sqrt{\log 2\widetilde{k}} < \sqrt{\widetilde{k}}$, and $\alpha = \text{constant}/\left(R_{\max} - R_{\min} + \gamma\right)^n$ for some constant > 0 as introduced in Proposition 4.4. Also, second to the last inequality above follows for all $t \geq 1$ since $n \geq 2$ by definition. Then,

$$\exp\left(-\frac{2\lambda_t^2}{16n(6R_{\max} - 6R_{\min} + 2\gamma)^2(t-1)} - \log\beta_t + n\log(2R_{\max} - 2R_{\min})\right)$$

$$\leq \exp\left(-\frac{\lambda_t^2}{32n(3R_{\max} - 3R_{\min} + \gamma)^2 t} - \log\beta_t + n\log(2R_{\max} - 2R_{\min})\right) \tag{C.157}$$

$$\leq \exp\Bigg(- 4k^2(\log 2t)^3 - \log\Bigg(\frac{3k\,(3(R_{\max}-R_{\min})+\gamma)^n \sqrt[6]{32n}}{1-k\sqrt{\log 2\widetilde{k}}/\sqrt{\widetilde{k}}}\frac{\sqrt{\log 2t}}{t^{w/3}}\Bigg) + n\log(2R_{\max}-2R_{\min})\Bigg) \tag{C.158}$$

$$\leq \exp\Bigg(- (\log 2t)^3 - \log\Bigg(\frac{3k\,(3(R_{\max}-R_{\min})+\gamma)^n \sqrt[6]{32n}}{1-k\sqrt{\log 2\widetilde{k}}/\sqrt{\widetilde{k}}}\frac{\sqrt{\log 2t}}{t^{w/3}}\Bigg) + n\log(2R_{\max}-2R_{\min})\Bigg) \tag{C.159}$$

$$\leq \frac{(2R_{\max}-2R_{\min})^n\left(1-k\sqrt{\log 2\widetilde{k}}/\sqrt{\widetilde{k}}\right)}{3k\,(3(R_{\max}-R_{\min})+\gamma)^n \sqrt[6]{32n}}\frac{1}{t^{w/3+1/2}}\frac{t^{w/3}}{\sqrt{\log 2t}} \tag{C.160}$$

$$= \frac{2^n}{3^{n+1}k\sqrt[6]{32n}}\frac{1}{\sqrt{t\log 2t}} \tag{C.161}$$

where second to the last inequality holds since for all $t \geq 1$ and $0 < w < 1/4$ it holds that $\exp(-(\log 2t)^3) \leq \exp(-(\log 2t)) = \frac{1}{t} \leq \frac{1}{t^{w/3+1/2}}$.

Lastly, substituting the upper bounds in (C.147) and (C.161) into (C.123), we get

$$\mathbb{P}\left(j_t^* \neq j^{*,0}\right) \leq \frac{n}{t} + \frac{n^{5/6}2^{n+1}}{3^{n+1}k\sqrt[6]{32}}\frac{1}{\sqrt{t\log 2t}} \tag{C.162}$$

□

Proof of Theorem 4.2. We start by recalling the definition of our regret notion given in (4.32) and decompose it into two main parts: 1) total costs incurred due to the offered incentives, 2) total expected rewards collected through the arms chosen by the agent.

$$\text{Regret}\,(\Pi_{\epsilon,T}) = \sum_{t\in\mathcal{T}} V(\mathbf{c}(\boldsymbol{\theta}^0,\mathrm{s}^0);\boldsymbol{\theta}^0) - V_t(\boldsymbol{\pi}_t;\boldsymbol{\theta}^0) \tag{C.163}$$

$$= \sum_{t\in\mathcal{T}}\sum_{a\in\mathcal{A}}\left[\pi_{t,a} - c_a(\boldsymbol{\theta}^0,\mathrm{s}^0)\right] + \sum_{t\in\mathcal{T}}\left[\theta^0_{\upsilon(\mathbf{c}(\boldsymbol{\theta}^0,\mathrm{s}^0))} - \theta^0_{\upsilon_t(\boldsymbol{\pi}_t)}\right] \tag{C.164}$$

We let $\mathcal{T}^{\mathrm{pr-xplore}} \in \mathcal{T}$ and $\mathcal{T}^{\mathrm{pr-xploit}} \in \mathcal{T}$ be the set of random time steps that the principal's algorithm (3) performs exploration (lines 11-12) and exploitation (lines 14-20), respectively. First, we bound the first part of (C.164) as follows.

$$\sum_{t\in\mathcal{T}}\sum_{a\in\mathcal{A}}\left[\pi_{t,a} - c_a(\boldsymbol{\theta}^0,\mathrm{s}^0)\right] = \sum_{t\in\mathcal{T}^{\mathrm{pr-xplore}}}\sum_{a\in\mathcal{A}}\left[\pi_{t,a} - c_a(\boldsymbol{\theta}^0,\mathrm{s}^0)\right] + \sum_{t\in\mathcal{T}^{\mathrm{pr-xploit}}}\sum_{a\in\mathcal{A}}\left[c_a(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathrm{s}}_t^{\mathrm{pr}}) - c_a(\boldsymbol{\theta}^0,\mathrm{s}^0)\right] \tag{C.165}$$

Notice that the cardinalities of the defined random sets, $|\mathcal{T}^{\mathrm{pr-xplore}}|$ and $|\mathcal{T}^{\mathrm{pr-xploit}}|$, are random variables. Then, the first part of (C.165) is bounded by considering the following conditional expectation.

$$\mathbb{E}\Bigg[\sum_{t\in\mathcal{T}^{\mathrm{pr-xplore}}}\sum_{a\in\mathcal{A}}\pi_{t,a}-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\Big|\mathcal{T}^{\mathrm{pr-xplore}}\Bigg]\leq n(\overline{C}-\underline{C})|\mathcal{T}^{\mathrm{pr-xplore}}| \tag{C.166}$$

Taking the expectation of both sides in the last inequality, we obtain

$$\sum_{t\in\mathcal{T}^{\mathrm{pr-xplore}}}\sum_{a\in\mathcal{A}}\pi_{t,a}-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\leq n(\overline{C}-\underline{C})\mathbb{E}|\mathcal{T}^{\mathrm{pr-xplore}}|$$

$$=n(\overline{C}-\underline{C})\sum_{t=1}^{T}\epsilon_t^{\mathrm{pr}} \tag{C.167}$$

$$=n(\overline{C}-\underline{C})\sum_{t=1}^{T}\min\left\{1,\frac{m^{\mathrm{pr}}}{t^{(1/2-w)}}\right\} \tag{C.168}$$

$$\leq nm^{\mathrm{pr}}(\overline{C}-\underline{C})\sum_{t=1}^{T}\frac{1}{t^{1/2-w}} \tag{C.169}$$

$$\leq m^{\mathrm{pr}}n(\overline{C}-\underline{C})\left(1+\int_{t=1}^{T}\frac{1}{t^{1/2-w}}dt\right) \tag{C.170}$$

$$=m^{\mathrm{pr}}n(\overline{C}-\underline{C})\left(\frac{2}{2w+1}T^{w+1/2}+\frac{2w-1}{2w+1}\right) \tag{C.171}$$

Next, we bound the second part of (C.165) again by considering a conditional expectation.

$$\begin{aligned}&\mathbb{E}\Bigg[\sum_{t\in\mathcal{T}^{\mathrm{pr-xploit}}}\sum_{a\in\mathcal{A}}c_a(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t^{\mathrm{pr}})-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\Big|\mathcal{T}^{\mathrm{pr-xploit}}\Bigg]\\
&=\sum_{t\in\mathcal{T}^{\mathrm{pr-xploit}}}\sum_{a\in\mathcal{A}}\mathbb{E}\left[c_a(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t^{\mathrm{pr}})-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\Big|\mathcal{T}^{\mathrm{pr-xploit}},\|\mathbf{s}^0-\widehat{\mathbf{s}}_t^{\mathrm{pr}}\|_\infty\leq\beta_t\right]\mathbb{P}\left(\|\mathbf{s}^0-\widehat{\mathbf{s}}_t^{\mathrm{pr}}\|_\infty\leq\beta_t\right)\\
&\quad+\sum_{t\in\mathcal{T}^{\mathrm{pr-xploit}}}\sum_{a\in\mathcal{A}}\mathbb{E}\left[c_a(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t^{\mathrm{pr}})-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\Big|\mathcal{T}^{\mathrm{pr-xploit}},\|\mathbf{s}^0-\widehat{\mathbf{s}}_t^{\mathrm{pr}}\|_\infty>\beta_t\right]\mathbb{P}\left(\|\mathbf{s}^0-\widehat{\mathbf{s}}_t^{\mathrm{pr}}\|_\infty>\beta_t\right)\end{aligned} \tag{C.172}$$

$$\begin{aligned}&\leq\sum_{t\in\mathcal{T}^{\mathrm{pr-xploit}}}\sum_{a\in\mathcal{A}}\mathbb{E}\left[c_a(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t^{\mathrm{pr}})-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\Big|\mathcal{T}^{\mathrm{pr-xploit}},\|\mathbf{s}^0-\widehat{\mathbf{s}}_t^{\mathrm{pr}}\|_\infty\leq\beta_t\right]\\
&\quad+2n(\overline{C}-\underline{C})\sum_{t\in\mathcal{T}^{\mathrm{pr-xploit}}}\exp\Bigg(-\frac{2\lambda_t^2}{(t-1)16n(6R_{\max}-6R_{\min}+2\gamma)^2}\\
&\qquad\qquad-\log\beta_t+n\log(2R_{\max}-2R_{\min})\Bigg)\end{aligned} \tag{C.173}$$

where the last inequality follows by Corollary 4.1. For the first term in (C.173), we have

$$
\begin{aligned}
&\sum_{t\in\mathcal{T}^{\text{pr}-\text{xploit}}}\sum_{a\in\mathcal{A}}\mathbb{E}\left[c_a(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t^{\text{pr}})-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\middle|\mathcal{T}^{\text{pr}-\text{xploit}},\|\mathbf{s}^0-\widehat{\mathbf{s}}_t^{\text{pr}}\|_\infty\le\beta_t\right]\\
&=\sum_{t\in\mathcal{T}^{\text{pr}-\text{xploit}}}\sum_{a\in\mathcal{A}}\mathbb{E}\left[c_a(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t^{\text{pr}})-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\middle|\mathcal{T}^{\text{pr}-\text{xploit}},\|\mathbf{s}^0-\widehat{\mathbf{s}}_t^{\text{pr}}\|_\infty\le\beta_t,\ j_t^*=j^{*,0}\right]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\cdot\mathbb{P}\left(j_t^*=j^{*,0}\right)\\
&\qquad\qquad+\mathbb{E}\left[c_a(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t^{\text{pr}})-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\middle|\mathcal{T}^{\text{pr}-\text{xploit}},\|\mathbf{s}^0-\widehat{\mathbf{s}}_t^{\text{pr}}\|_\infty\le\beta_t,\ j_t^*\ne j^{*,0}\right]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\cdot\mathbb{P}\left(j_t^*\ne j^{*,0}\right) \qquad (\text{C.174})\\
&\le\sum_{t\in\mathcal{T}^{\text{pr}-\text{xploit}}}\mathbb{E}\Big[\max_{a\in\mathcal{A}}\widehat{s}_{t,a}^{\text{pr}}-\widehat{s}_{t,j_t^*}^{\text{pr}}+2\beta_t-\max_{a\in\mathcal{A}}s_a^0+s_{j^{*,0}}^0\\
&\qquad\qquad\qquad\qquad\qquad\qquad\Big|\mathcal{T}^{\text{pr}-\text{xploit}},\|\mathbf{s}^0-\widehat{\mathbf{s}}_t^{\text{pr}}\|_\infty\le\beta_t,\ j_t^*=j^{*,0}\Big]\\
&\qquad\qquad+n(\overline{C}-\underline{C})\mathbb{P}\left(j_t^*\ne j^{*,0}\right) \qquad (\text{C.175})
\end{aligned}
$$

For convenience, we continue with the notation $\kappa_t\in\arg\max_{a\in\mathcal{A}}\widehat{s}_{t,a}^{\text{pr}}$ and $\kappa^0\in\arg\max_{a\in\mathcal{A}}s_a^0$.

$$
\begin{aligned}
&=\sum_{t\in\mathcal{T}^{\text{pr}-\text{xploit}}}\mathbb{E}\left[\widehat{s}_{t,\kappa_t}^{\text{pr}}-\widehat{s}_{t,j_t^*}^{\text{pr}}+2\beta_t-s_{\kappa^0}^0+s_{j^{*,0}}^0\middle|\mathcal{T}^{\text{pr}-\text{xploit}},\|\mathbf{s}^0-\widehat{\mathbf{s}}_t^{\text{pr}}\|_\infty\le\beta_t,\ j_t^*=j^{*,0}\right]\\
&\qquad\qquad+n(\overline{C}-\underline{C})\mathbb{P}\left(j_t^*\ne j^{*,0}\right) \qquad (\text{C.176})\\
&=\sum_{t\in\mathcal{T}^{\text{pr}-\text{xploit}}}\mathbb{E}\Big[(\widehat{s}_{t,\kappa_t}^{\text{pr}}-s_{\kappa_t}^0)+(s_{\kappa_t}^0-s_{\kappa^0}^0)+(s_{j^{*,0}}^0-\widehat{s}_{t,j_t^*}^{\text{pr}})+2\beta_t\\
&\qquad\qquad\qquad\qquad\qquad\qquad\Big|\mathcal{T}^{\text{pr}-\text{xploit}},\|\mathbf{s}^0-\widehat{\mathbf{s}}_t^{\text{pr}}\|_\infty\le\beta_t,\ j_t^*=j^{*,0}\Big]\\
&\qquad\qquad+n(\overline{C}-\underline{C})\mathbb{P}\left(j_t^*\ne j^{*,0}\right) \qquad (\text{C.177})\\
&\le\sum_{t\in\mathcal{T}^{\text{pr}-\text{xploit}}}4\beta_t+n(\overline{C}-\underline{C})\mathbb{P}\left(j_t^*\ne j^{*,0}\right) \qquad (\text{C.178})\\
&\le\sum_{t\in\mathcal{T}^{\text{pr}-\text{xploit}}}4\beta_t+\frac{n^2(\overline{C}-\underline{C})}{t}+\frac{n^{11/6}2^{n+1}(\overline{C}-\underline{C})}{3^{n+1}k\sqrt[6]{32}}\frac{1}{\sqrt{t\log 2t}} \qquad (\text{C.179})
\end{aligned}
$$

where the last inequality follows by Proposition 4.7. Substituting $\beta_t=B\dfrac{\sqrt{\log 2t}}{t^{w/3}}$, we have

$$
\begin{aligned}
&=\sum_{t\in\mathcal{T}^{\text{pr}-\text{xploit}}}4B\frac{\sqrt{\log 2t}}{t^{w/3}}+\frac{n^2(\overline{C}-\underline{C})}{t}+\frac{n^{11/6}2^{n+1}(\overline{C}-\underline{C})}{3^{n+1}k\sqrt[6]{32}}\frac{1}{\sqrt{t\log 2t}} \qquad (\text{C.180})\\
&=\frac{12B}{3-w}B|\mathcal{T}^{\text{pr}-\text{xploit}}|^{(1-w/3)}\sqrt{\log 2|\mathcal{T}^{\text{pr}-\text{xploit}}|}+n^2(\overline{C}-\underline{C})\log|\mathcal{T}^{\text{pr}-\text{xploit}}|
\end{aligned}
$$

$$+\frac{n^{11/6}2^{n+2}(\overline{C}-\underline{C})}{3^{n+1}k\sqrt[6]{32}}\sqrt{|\mathcal{T}^{\text{pr}-\text{xploit}}|} \tag{C.181}$$

$$\leq \frac{12B}{3-w}T^{1-w/3}\sqrt{\log 2T}+n^2(\overline{C}-\underline{C})\log T+\frac{n^{11/6}2^{n+2}(\overline{C}-\underline{C})}{3^{n+1}k\sqrt[6]{32}}\sqrt{T} \tag{C.182}$$

Now, for the same choice of β_t, we can bound the second term in (C.173) by following the same arguments as in (C.148)-(C.161) and obtain

$$2n(\overline{C}-\underline{C})\sum_{t\in\mathcal{T}^{\text{pr}-\text{xploit}}}\exp\Bigg(-\frac{2\lambda_t^2}{(t-1)16n(6R_{\max}-6R_{\min}+2\gamma)^2} \\ -\log\beta_t+n\log(2R_{\max}-2R_{\min})\Bigg)$$

$$\leq \frac{2^{n+1}n^{5/6}(\overline{C}-\underline{C})}{3^{n+1}k\sqrt[6]{32}}\sum_{t\in\mathcal{T}^{\text{pr}-\text{xploit}}}\frac{1}{\sqrt{t\log 2t}} \tag{C.183}$$

$$\leq \frac{2^{n+1}n^{5/6}(\overline{C}-\underline{C})}{3^{n+1}k\sqrt[6]{32}}\sqrt{|\mathcal{T}^{\text{pr}-\text{xploit}}|} \tag{C.184}$$

$$\leq \frac{2^{n+1}n^{5/6}(\overline{C}-\underline{C})}{3^{n+1}k\sqrt[6]{32}}\sqrt{T} \tag{C.185}$$

Combining these upper bounds with (C.173), we get

$$\mathbb{E}\Bigg[\sum_{t\in\mathcal{T}^{\text{pr}-\text{xploit}}}\sum_{a\in\mathcal{A}}c_a(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t^{\text{pr}})-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\Big|\mathcal{T}^{\text{pr}-\text{xploit}}\Bigg] \\ \leq \frac{12B}{3-w}T^{1-w/3}\sqrt{\log 2T}+n^2(\overline{C}-\underline{C})\log T+\frac{2^{n+1}n^{5/6}(\overline{C}-\underline{C})\,(1+2n)}{3^{n+1}k\sqrt[6]{32}}\sqrt{T} \tag{C.186}$$

Taking the expectation of both sides in the last inequality, we obtain

$$\sum_{t\in\mathcal{T}^{\text{pr}-\text{xploit}}}\sum_{a\in\mathcal{A}}c_a(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}_t^{\text{pr}})-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0) \\ \leq \frac{12B}{3-w}T^{1-w/3}\sqrt{\log 2T}+n^2(\overline{C}-\underline{C})\log T+\frac{2^{n+1}n^{5/6}(\overline{C}-\underline{C})\,(1+2n)}{3^{n+1}k\sqrt[6]{32}}\sqrt{T} \tag{C.187}$$

Substituting this last result and the result in (C.171) into (C.165), we provide the upper bound for the first part of our regret notion as follows.

$$\sum_{t\in\mathcal{T}}\sum_{a\in\mathcal{A}}\left[\pi_{t,a}-c_a(\boldsymbol{\theta}^0,\mathbf{s}^0)\right]\leq m^{\text{pr}}n(\overline{C}-\underline{C})\left(\frac{2}{2w+1}T^{w+1/2}+\frac{2w-1}{2w+1}\right) \\ +\frac{12B}{3-w}T^{1-w/3}\sqrt{\log 2T}+n^2(\overline{C}-\underline{C})\log T$$

$$+\frac{2^{n+1}n^{5/6}(\overline{C}-\underline{C})\,(1+2n)}{3^{n+1}k\sqrt[6]{32}}\sqrt{T} \tag{C.188}$$

Next, we compute an upper bound for the second part of the regret decomposed in (C.164).

$$\sum_{t\in\mathcal{T}}\left[\theta^0_{\upsilon(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))}-\theta^0_{\upsilon_t(\boldsymbol{\pi}_t)}\right]=\sum_{t\in\mathcal{T}^{\text{pr}-\text{xplore}}}\left[\theta^0_{\upsilon(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))}-\theta^0_{\upsilon_t(\boldsymbol{\pi}_t)}\right]+\sum_{t\in\mathcal{T}^{\text{pr}-\text{xploit}}}\left[\theta^0_{\upsilon(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))}-\theta^0_{\upsilon_t(\boldsymbol{\pi}_t)}\right] \tag{C.189}$$

Because the principal's expected rewards belong to a known compact set Θ, we let $\Theta^{\max}$ to be the upper bound on θ^0_a's. Then, we can bound the first term in the last inequality above by following a similar argument as in (C.166)-(C.171).

$$\sum_{t\in\mathcal{T}^{\text{pr}-\text{xplore}}}\left[\theta^0_{\upsilon(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))}-\theta^0_{\upsilon_t(\boldsymbol{\pi}_t)}\right]\leq\Theta^{\max}\mathbb{E}|\mathcal{T}^{\text{pr}-\text{xplore}}| \tag{C.190}$$

$$\leq\Theta^{\max}m^{\text{pr}}\left(\frac{2}{2w+1}T^{w+1/2}+\frac{2w-1}{2w+1}\right) \tag{C.191}$$

To bound the second term in (C.189), we again start by considering a conditional expectation.

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}^{\text{pr}-\text{xploit}}}\theta^0_{\upsilon(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))}-\theta^0_{\upsilon_t(\boldsymbol{\pi}_t)}\middle|\mathcal{T}^{\text{pr}-\text{xploit}}\right]$$

$$=\mathbb{E}\left[\sum_{t\in\mathcal{T}^{\text{pr}-\text{xploit}}}\theta^0_{\upsilon(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))}-\theta^0_{\upsilon_t(\boldsymbol{\pi}_t)}\middle|\mathcal{T}^{\text{pr}-\text{xploit}}\right]\mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t)\neq\upsilon(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\right) \tag{C.192}$$

$$\leq\Theta^{\max}\sum_{t\in\mathcal{T}^{\text{pr}-\text{xploit}}}\mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t)\neq\upsilon(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\right) \tag{C.193}$$

Now, we need to derive an upper bound for the probability that the arm chosen by the agent in response to the principal's exploitation incentives is different than the arm chosen by the perfect-knowledge agent in response to the oracle incentives. Recall that the perfect-knowledge agent always picks the true reward-maximizer arm, i.e., $\upsilon(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))=\arg\max_{a\in\mathcal{A}}s^0_a+c^0_a(\boldsymbol{\theta}^0,\mathbf{s}^0)$ by construction. On the other hand, for an imperfect-knowledge learning agent, we need to take into account whether the agent picks the true reward-maximizer arm or not at a certain time step.

$$\sum_{t\in\mathcal{T}^{\text{pr}-\text{xploit}}}\mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t)\neq\upsilon(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\right)$$

$$=\sum_{t\in\mathcal{T}^{\text{pr}-\text{xploit}}}\mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t)\neq\upsilon(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\middle|\upsilon_t(\boldsymbol{\pi}_t)\neq\arg\max_{a\in\mathcal{A}}s^0_a+c_a(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}^{\text{pr}}_t)\right)p_t$$

$$+\,\mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t)\neq\upsilon(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))\middle|\upsilon_t(\boldsymbol{\pi}_t)=\arg\max_{a\in\mathcal{A}}s^0_a+c_a(\widehat{\boldsymbol{\theta}}_t,\widehat{\mathbf{s}}^{\text{pr}}_t)\right)(1-p_t) \tag{C.194}$$

$$\leq \sum_{t \in \mathcal{T}^{\mathrm{pr-xploit}}} p_t + \sum_{t \in \mathcal{T}^{\mathrm{pr-xploit}}} \mathbb{P}\left(\upsilon(\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0)) \middle| \upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a \in \mathcal{A}} s_a^0 + c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) \right) \tag{C.195}$$

$$\leq \widetilde{k} + \sum_{t \in \mathcal{T}^{\mathrm{pr-xploit}}, t \geq \widetilde{k}} k \frac{\sqrt{\log 2t}}{\sqrt{t}} + \sum_{t \in \mathcal{T}^{\mathrm{pr-xploit}}} \mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t) \neq \upsilon(\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0)) \middle| \upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a \in \mathcal{A}} s_a^0 + c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) \right) \tag{C.196}$$

where the last inequality follows by Assumption 4.2.

$$\leq \widetilde{k} + 2k\sqrt{|\mathcal{T}^{\mathrm{pr-xploit}}| \log 2|\mathcal{T}^{\mathrm{pr-xploit}}|} + \sum_{t \in \mathcal{T}^{\mathrm{pr-xploit}}} \mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t) \neq \upsilon(\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0)) \middle| \upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a \in \mathcal{A}} s_a^0 + c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) \right) \tag{C.197}$$

$$\leq \widetilde{k} + 2k\sqrt{T \log 2T} + \sum_{t \in \mathcal{T}^{\mathrm{pr-xploit}}} \mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t) \neq \upsilon(\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0)) \middle| \upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a \in \mathcal{A}} s_a^0 + c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) \right) \tag{C.198}$$

We now compute an upper bound for the summation in the last line above by recalling that $\upsilon(\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0)) = j^{*,0}$ as introduced in Section 4.3.2.1.

$$\sum_{t \in \mathcal{T}^{\mathrm{pr-xploit}}} \mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t) \neq \upsilon(\mathbf{c}(\boldsymbol{\theta}^0, \mathbf{s}^0)) \middle| \upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a \in \mathcal{A}} s_a^0 + c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) \right)$$

$$= \sum_{t \in \mathcal{T}^{\mathrm{pr-xploit}}} \mathbb{P}\left(\arg\max_{a \in \mathcal{A}} s_a^0 + c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) \neq j^{*,0} \right) \tag{C.199}$$

$$= \sum_{t \in \mathcal{T}^{\mathrm{pr-xploit}}} \mathbb{P}\left(\bigcup_{a \in \mathcal{A} \setminus \{j^{*,0}\}} s_a^0 + c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) > s_{j^{*,0}}^0 + c_{j^{*,0}}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) \right) \tag{C.200}$$

$$= \sum_{t \in \mathcal{T}^{\mathrm{pr-xploit}}} \mathbb{P}\left(\bigcup_{a \in \mathcal{A} \setminus \{j^{*,0}\}} s_a^0 + c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) > s_{j^{*,0}}^0 + c_{j^{*,0}}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) \middle| j_t^* = j^{*,0} \right) \mathbb{P}\left(j_t^* = j^{*,0} \right) + \mathbb{P}\left(\bigcup_{a \in \mathcal{A} \setminus \{j^{*,0}\}} s_a^0 + c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) > s_{j^{*,0}}^0 + c_{j^{*,0}}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) \middle| j_t^* \neq j^{*,0} \right) \mathbb{P}\left(j_t^* \neq j^{*,0} \right) \tag{C.201}$$

$$
\begin{aligned}
&\leq \sum_{t\in\mathcal{T}^{\mathrm{pr-xploit}}} \mathbb{P}\left(\bigcup_{a\in\mathcal{A}\setminus\{j^{*,0}\}} s_a^0 + c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) > s^0_{j^{*,0}} + c_{j^{*,0}}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) \Big| j_t^* = j^{*,0} \right) \\
&\qquad + \sum_{t\in\mathcal{T}^{\mathrm{pr-xploit}}} \sum_{a\in\mathcal{A}\setminus\{j^{*,0}\}} \frac{n}{t} + \frac{n^{5/6}2^{n+1}}{3^{n+1}k\sqrt[6]{32}} \frac{1}{\sqrt{t\log 2t}} \qquad \text{(C.202)} \\
&\leq \sum_{t\in\mathcal{T}^{\mathrm{pr-xploit}}} \mathbb{P}\left(\bigcup_{a\in\mathcal{A}\setminus\{j^{*,0}\}} s_a^0 + c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) > s^0_{j^{*,0}} + c_{j^{*,0}}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) \Big| j_t^* = j^{*,0} \right) \\
&\qquad + n^2 \log T + \frac{n^{11/6}2^{n+2}}{3^{n+1}k\sqrt[6]{32}} \sqrt{T} \qquad \text{(C.203)}
\end{aligned}
$$

where second to the last line follows by Proposition 4.7 and the last line follows by repeating the same arguments we had in lines (C.180)-(C.182) at the first part of this proof. To bound the first term of the last inequality, we use the definition of the exploitation incentives $c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}})$ introduced in Algorithm 3.

$$
\begin{aligned}
&\sum_{t\in\mathcal{T}^{\mathrm{pr-xploit}}} \mathbb{P}\left(\bigcup_{a\in\mathcal{A}\setminus\{j^{*,0}\}} s_a^0 + c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) > s^0_{j^{*,0}} + c_{j^{*,0}}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) \Big| j_t^* = j^{*,0} \right) \\
&= \sum_{t\in\mathcal{T}^{\mathrm{pr-xploit}}} \mathbb{P}\left(\bigcup_{a\in\mathcal{A}\setminus\{j_t^*\}} s_a^0 + c_a(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) > s^0_{j_t^*} + c_{j_t^*}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) \right) \qquad \text{(C.204)} \\
&= \sum_{t\in\mathcal{T}^{\mathrm{pr-xploit}}} \mathbb{P}\left(\bigcup_{a\in\mathcal{A}\setminus\{j_t^*\}} s_a^0 > s^0_{j_t^*} + \max_{a'\in\mathcal{A}} \widehat{s}^{\mathrm{pr}}_{t,a} - \widehat{s}^{\mathrm{pr}}_{t,j_t^*} + 2\beta_t \right) \qquad \text{(C.205)}
\end{aligned}
$$

Recall the indices $\kappa_t \in \arg\max_{a\in\mathcal{A}} \widehat{s}^{\mathrm{pr}}_{t,a}$, $\kappa^0 \in \arg\max_{a\in\mathcal{A}} s_a^0$ defined earlier for notational convenience.

$$
\begin{aligned}
&\leq \sum_{t\in\mathcal{T}^{\mathrm{pr-xploit}}} \mathbb{P}\left(s^0_{\kappa^0} > s^0_{j_t^*} + \widehat{s}^{\mathrm{pr}}_{t,\kappa_t} - \widehat{s}^{\mathrm{pr}}_{t,j_t^*} + 2\beta_t \right) \qquad \text{(C.206)} \\
&= \sum_{t\in\mathcal{T}^{\mathrm{pr-xploit}}} \mathbb{P}\left((s^0_{\kappa^0} - \widehat{s}^{\mathrm{pr}}_{t,\kappa^0}) + (\widehat{s}^{\mathrm{pr}}_{t,\kappa^0} - \widehat{s}^{\mathrm{pr}}_{t,\kappa_t}) + (\widehat{s}^{\mathrm{pr}}_{t,j_t^*} - s^0_{j_t^*}) > 2\beta_t \right) \qquad \text{(C.207)}
\end{aligned}
$$

Notice that $\widehat{s}^{\mathrm{pr}}_{t,\kappa^0} - \widehat{s}^{\mathrm{pr}}_{t,\kappa_t} \leq 0$ by definition. Then,

$$
\begin{aligned}
&\leq \sum_{t\in\mathcal{T}^{\mathrm{pr-xploit}}} \mathbb{P}\left(2\max_{a\in\mathcal{A}} |s_a^0 - \widehat{s}^{\mathrm{pr}}_{t,a}| > 2\beta_t \right) \qquad \text{(C.208)} \\
&= \sum_{t\in\mathcal{T}^{\mathrm{pr-xploit}}} \mathbb{P}\left(\|\mathbf{s}^0 - \widehat{\mathbf{s}}_t^{\mathrm{pr}}\|_\infty > \beta_t \right) \qquad \text{(C.209)}
\end{aligned}
$$

$$\le 2 \sum_{t\in\mathcal{T}^{\mathrm{pr-xploit}}} \exp\left(-\frac{2\lambda_t^2}{(t-1)16n(6R_{\max}-6R_{\min}+2\gamma)^2} - \log\beta_t + n\log(2R_{\max}-2R_{\min})\right) \tag{C.210}$$

which follows by Corollary 4.1. Then, we follow the same arguments as in (C.148)-(C.161) and continue as

$$\le \frac{2^{n+1}}{3^{n+1}k\sqrt[6]{32n}} \sum_{t\in\mathcal{T}^{\mathrm{pr-xploit}}} \frac{1}{\sqrt{t\log 2t}} \tag{C.211}$$

$$\le \frac{2^{n+1}}{3^{n+1}k\sqrt[6]{32n}} \sqrt{|\mathcal{T}^{\mathrm{pr-xploit}}|} \tag{C.212}$$

$$\le \frac{2^{n+1}}{3^{n+1}k\sqrt[6]{32n}} \sqrt{T} \tag{C.213}$$

We substitute this upper bound into first (C.203), then (C.198), and lastly (C.193). Then, taking the expectation of both sides of the obtained inequality in (C.193) gives us

$$\sum_{t\in\mathcal{T}^{\mathrm{pr-xploit}}} \theta^0_{\upsilon(\mathbf{c}(\boldsymbol{\theta}^0,\mathbf{s}^0))} - \theta^0_{\upsilon_t(\boldsymbol{\pi}_t)}$$

$$\le \Theta^{\max}\left(\widetilde{k} + 2k\sqrt{T\log 2T} + \frac{2^{n+1}}{3^{n+1}k\sqrt[6]{32n}}\sqrt{T} + n^2\log T + \frac{n^{11/6}2^{n+2}}{3^{n+1}k\sqrt[6]{32}}\sqrt{T}\right) \tag{C.214}$$

$$= \Theta^{\max}\left(\widetilde{k} + 2k\sqrt{T\log 2T} + \frac{2^{n+1}\left(2n^{11/6} + 1/\sqrt[6]{n}\right)}{3^{n+1}k\sqrt[6]{32}}\sqrt{T} + n^2\log T\right) \tag{C.215}$$

This result together with (C.191) gives us the upper bound for the second part of the regret. We conclude by combining everything with (C.164) and get the desired regret bound.

$$\mathrm{Regret}\left(\Pi_{\epsilon,T}\right)$$

$$\le \frac{12B}{3-w}T^{1-w/3}\sqrt{\log 2T} + m^{\mathrm{pr}}\left(n(\overline{C}-\underline{C}) + \Theta^{\max}\right)\left(\frac{2}{2w+1}T^{w+1/2} + \frac{2w-1}{2w+1}\right) \tag{C.216}$$

$$+ 2k\Theta^{\max}\sqrt{T\log 2T} + \frac{2^{n+1}\left(\Theta^{\max}(2n^{11/6} + 1/\sqrt[6]{n}) + n^{5/6}(\overline{C}-\underline{C})\left(1+2n\right)\right)}{3^{n+1}k\sqrt[6]{32}}\sqrt{T} \tag{C.217}$$

$$+ n^2\left(\overline{C} - \underline{C} + \Theta^{\max}\right)\log T + \Theta^{\max}\widetilde{k} \tag{C.218}$$

□

C.1.3 Results in Section 4.4

Proof of Lemma 4.4. According to the proposed algorithms (3) and (4), the arm chosen by the agent at time $t \in \mathcal{T}^{\text{ag}-\text{xploit}}$ is defined as $\upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a\in\mathcal{A}} \widehat{s}^{\text{ag}}_{t,a} + \pi_{t,a}$ and the incentives provided by the principal at time $t \in \mathcal{T}^{\text{pr}-\text{xploit}}$ is given by $\boldsymbol{\pi}_t = \mathbf{c}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}^{\text{pr}}_t)$. We start our proof by using these definitions.

$$\begin{aligned}
&\mathbb{P}\Big(\upsilon_t(\boldsymbol{\pi}_t) \neq \arg\max_{a\in\mathcal{A}} s^0_a + \pi_{t,a}\Big| t \in \mathcal{T}^{\text{ag}-\text{xploit}} \cap \mathcal{T}^{\text{pr}-\text{xploit}}\Big) \\
&= \mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t) \neq \arg\max_{a\in\mathcal{A}} s^0_a + \pi_{t,a}\middle|\upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a\in\mathcal{A}} \widehat{s}^{\text{ag}}_{t,a} + \pi_{t,a},\ \boldsymbol{\pi}_t = \mathbf{c}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}^{\text{pr}}_t)\right) \qquad \text{(C.219)} \\
&= \mathbb{P}\left(\bigcup_{a'\in\mathcal{A}}\bigcup_{a\in\mathcal{A}} s^0_a + c_{t,a}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}^{\text{pr}}_t) > s^0_{a'} + c_{t,a'}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}^{\text{pr}}_t)\middle| a' = \arg\max_{a''\in\mathcal{A}} \widehat{s}^{\text{ag}}_{t,a''} + c_{t,a''}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}^{\text{pr}}_t)\right) \\
&\qquad \text{(C.220)} \\
&\leq \sum_{a'\in\mathcal{A}}\sum_{a\in\mathcal{A}} \mathbb{P}\left(s^0_a + c_{t,a}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}^{\text{pr}}_t) > s^0_{a'} + c_{t,a'}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}^{\text{pr}}_t)\middle| a' = \arg\max_{a''\in\mathcal{A}} \widehat{s}^{\text{ag}}_{t,a''} + c_{t,a''}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}^{\text{pr}}_t)\right) \\
&\qquad \text{(C.221)}
\end{aligned}$$

where (C.221) follows by the Boole's inequality (a.k.a. union bound). We further condition on whether the chosen arm $\upsilon_t(\boldsymbol{\pi}_t)$ is same as the desired arm by the principal (j^*_t) or not.

$$\begin{aligned}
&= \sum_{a'\in\mathcal{A}}\sum_{a\in\mathcal{A}} \mathbb{P}\Big(c_{t,a}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}^{\text{pr}}_t) - c_{t,a'}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}^{\text{pr}}_t) > s^0_{a'} - s^0_a \Big| j^*_t = a', a' = \arg\max_{a''\in\mathcal{A}} \widehat{s}^{\text{ag}}_{t,a''} + c_{t,a''}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}^{\text{pr}}_t)\Big) \\
&\qquad \cdot \mathbb{P}\Big(j^*_t = a' \Big| a' = \arg\max_{a''\in\mathcal{A}} \widehat{s}^{\text{ag}}_{t,a''} + c_{t,a''}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}^{\text{pr}}_t)\Big) \\
&\qquad + \mathbb{P}\Big(c_{t,a}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}^{\text{pr}}_t) - c_{t,a'}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}^{\text{pr}}_t) > s^0_{a'} - s^0_a \Big| j^*_t \neq a', a' = \arg\max_{a''\in\mathcal{A}} \widehat{s}^{\text{ag}}_{t,a''} + c_{t,a''}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}^{\text{pr}}_t)\Big) \\
&\qquad \cdot \mathbb{P}\Big(j^*_t \neq a' \Big| a' = \arg\max_{a''\in\mathcal{A}} \widehat{s}^{\text{ag}}_{t,a''} + c_{t,a''}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}^{\text{pr}}_t)\Big) \qquad \text{(C.222)} \\
&\leq \sum_{a'\in\mathcal{A}}\sum_{a\in\mathcal{A}} \mathbb{P}\big(c_{t,a}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}^{\text{pr}}_t) - c_{t,j^*_t}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}^{\text{pr}}_t) > s^0_{j^*_t} - s^0_a\big) \\
&\qquad + \mathbb{P}\Big(j^*_t \neq a' \Big| a' = \arg\max_{a''\in\mathcal{A}} \widehat{s}^{\text{ag}}_{t,a''} + c_{t,a''}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}^{\text{pr}}_t)\Big) \qquad \text{(C.223)}
\end{aligned}$$

We can bound the first term in (C.223) from above by using the definition of the principal's exploitation incentives $\mathbf{c}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}})$ introduced in Algorithm 3.

$$\begin{aligned}
&\sum_{a'\in\mathcal{A}}\sum_{a\in\mathcal{A}}\mathbb{P}\big(c_{t,a}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) - c_{t,j_t^*}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) > s_{j_t^*}^0 - s_a^0\big)\\
&= \sum_{a'\in\mathcal{A}}\sum_{a\in\mathcal{A}}\mathbb{P}\big(\widehat{s}_{t,j_t^*}^{\mathrm{pr}} - s_{j_t^*}^0 + s_a^0 - 2\beta_t > \max_{a''\in\mathcal{A}}\widehat{s}_{t,a''}^{\mathrm{pr}}\big) && \text{(C.224)}\\
&\le \sum_{a'\in\mathcal{A}}\sum_{a\in\mathcal{A}}\mathbb{P}\big(\widehat{s}_{t,j_t^*}^{\mathrm{pr}} - s_{j_t^*}^0 + s_a^0 - 2\beta_t > \widehat{s}_{t,a}^{\mathrm{pr}}\big) && \text{(C.225)}\\
&= n\sum_{a\in\mathcal{A}}\mathbb{P}\big(\widehat{s}_{t,j_t^*}^{\mathrm{pr}} - s_{j_t^*}^0 + s_a^0 - \widehat{s}_{t,a}^{\mathrm{pr}} > 2\beta_t\big) && \text{(C.226)}
\end{aligned}$$

We continue by conditioning the last probability on the intersection of the two events $A = \{\widehat{s}_{t,j_t^*}^{\mathrm{pr}} - s_{j_t^*}^0 \le \beta_t\}$ and $B = \{s_a^0 - \widehat{s}_{t,a}^{\mathrm{pr}} \le \beta_t\}$. Recall that the complement of $A \cap B$ is equal to the union of their complements $\overline{A} \cup \overline{B}$.

$$\begin{aligned}
(C.226) &= n\sum_{a\in\mathcal{A}}\mathbb{P}\big(\widehat{s}_{t,j_t^*}^{\mathrm{pr}} - s_{j_t^*}^0 + s_a^0 - \widehat{s}_{t,a}^{\mathrm{pr}} > 2\beta_t \big| A\cap B\big)\mathbb{P}\big(A\cap B\big)\\
&\quad + n\sum_{a\in\mathcal{A}}\mathbb{P}\big(\widehat{s}_{t,j_t^*}^{\mathrm{pr}} - s_{j_t^*}^0 + s_a^0 - \widehat{s}_{t,a}^{\mathrm{pr}} > 2\beta_t \big| \overline{A}\cup\overline{B}\big)\mathbb{P}\big(\overline{A}\cup\overline{B}\big) && \text{(C.227)}\\
&\le n\sum_{a\in\mathcal{A}}\mathbb{P}\big(\overline{A}\cup\overline{B}\big) && \text{(C.228)}\\
&\le n\sum_{a\in\mathcal{A}}\mathbb{P}\big(\widehat{s}_{t,j_t^*}^{\mathrm{pr}} - s_{j_t^*}^0 > \beta_t\big) + \mathbb{P}\big(\widehat{s}_{t,a}^{\mathrm{pr}} - s_a^0 < -\beta_t\big) && \text{(C.229)}\\
&\le 2n^2\mathbb{P}\big(\|\widehat{\mathbf{s}}_t^{\mathrm{pr}} - \mathbf{s}^0\|_\infty > \beta_t\big) && \text{(C.230)}\\
&\le 4n^2\exp\left(-\frac{2\lambda_t^2}{(t-1)16n(6R_{\max} - 6R_{\min} + 2\gamma)^2} - \log\beta_t + n\log(2R_{\max} - 2R_{\min})\right) && \text{(C.231)}
\end{aligned}$$

where the last result follows by Corollary 4.1. Next, we bound the second term in (C.223).

$$\begin{aligned}
&\sum_{a'\in\mathcal{A}}\sum_{a\in\mathcal{A}}\mathbb{P}\Big(j_t^* \neq a' \Big| a' = \arg\max_{a''\in\mathcal{A}}\widehat{s}_{t,a''}^{\mathrm{ag}} + c_{t,a''}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}})\Big)\\
&\le n^2\mathbb{P}\Big(j_t^* \neq \arg\max_{a''\in\mathcal{A}}\widehat{s}_{t,a''}^{\mathrm{ag}} + c_{t,a''}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}})\Big) && \text{(C.232)}\\
&= n^2\mathbb{P}\Big(\bigcup_{a\in\mathcal{A}}\widehat{s}_{t,a}^{\mathrm{ag}} + c_{t,a}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) > \widehat{s}_{t,j_t^*}^{\mathrm{ag}} + c_{t,j_t^*}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}})\Big) && \text{(C.233)}\\
&\le n^2\sum_{a\in\mathcal{A}}\mathbb{P}\Big(\widehat{s}_{t,a}^{\mathrm{ag}} + c_{t,a}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}}) > \widehat{s}_{t,j_t^*}^{\mathrm{ag}} + c_{t,j_t^*}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}_t^{\mathrm{pr}})\Big) && \text{(C.234)}\\
&= n^2\sum_{a\in\mathcal{A}}\mathbb{P}\Big(\widehat{s}_{t,j_t^*}^{\mathrm{pr}} + \widehat{s}_{t,a}^{\mathrm{ag}} - \widehat{s}_{t,j_t^*}^{\mathrm{ag}} - 2\beta_t > \max_{a'\in\mathcal{A}}\widehat{s}_{t,a'}^{\mathrm{pr}}\Big) && \text{(C.235)}
\end{aligned}$$

$$\leq n^2 \sum_{a\in\mathcal{A}} \mathbb{P}\Big(\widehat{s}^{\mathrm{pr}}_{t,j_t^*} + \widehat{s}^{\mathrm{ag}}_{t,a} - \widehat{s}^{\mathrm{ag}}_{t,j_t^*} - 2\beta_t > \widehat{s}^{\mathrm{pr}}_{t,a}\Big) \tag{C.236}$$

$$= n^2 \sum_{a\in\mathcal{A}} \mathbb{P}\Big(\widehat{s}^{\mathrm{pr}}_{t,j_t^*} + \widehat{s}^{\mathrm{ag}}_{t,a} - \widehat{s}^{\mathrm{ag}}_{t,j_t^*} - \widehat{s}^{\mathrm{pr}}_{t,a} + s^0_a - s^0_a + s^0_{j_t^*} - s^0_{j_t^*} > 2\beta_t\Big) \tag{C.237}$$

$$= n^2 \sum_{a\in\mathcal{A}} \mathbb{P}\Big(\big(\widehat{s}^{\mathrm{pr}}_{t,j_t^*} - s^0_{j_t^*}\big) + \big(\widehat{s}^{\mathrm{ag}}_{t,a} - s^0_a\big) + \big(s^0_{j_t^*} - \widehat{s}^{\mathrm{ag}}_{t,j_t^*}\big) + \big(s^0_a - \widehat{s}^{\mathrm{pr}}_{t,a}\big) > 2\beta_t\Big) \tag{C.238}$$

where (C.233) follows by the Boole's inequality (a.k.a. union bound) and (C.235) follows by definition of $\mathbf{c}(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{s}}^{\mathrm{pr}}_t)$ as in Algorithm 3. We next condition the last probability on the intersection of the two events $C = \{\widehat{s}^{\mathrm{pr}}_{t,j_t^*} - s^0_{j_t^*} \leq \beta_t/2,\ \widehat{s}^{\mathrm{ag}}_{t,a} - s^0_a \leq \beta_t/2\}$ and $D = \{s^0_{j_t^*} - \widehat{s}^{\mathrm{ag}}_{t,j_t^*} \leq \beta_t/2,\ s^0_a - \widehat{s}^{\mathrm{pr}}_{t,a} \leq \beta_t/2\}$.

$$\begin{aligned}
&(C.238)\\
&= n^2 \sum_{a\in\mathcal{A}} \mathbb{P}\Big(\big(\widehat{s}^{\mathrm{pr}}_{t,j_t^*} - s^0_{j_t^*}\big) + \big(\widehat{s}^{\mathrm{ag}}_{t,a} - s^0_a\big) + \big(s^0_{j_t^*} - \widehat{s}^{\mathrm{ag}}_{t,j_t^*}\big) + \big(s^0_a - \widehat{s}^{\mathrm{pr}}_{t,a}\big) > 2\beta_t \Big| C \cap D\Big)\mathbb{P}\Big(C \cap D\Big)\\
&\quad + \mathbb{P}\Big(\big(\widehat{s}^{\mathrm{pr}}_{t,j_t^*} - s^0_{j_t^*}\big) + \big(\widehat{s}^{\mathrm{ag}}_{t,a} - s^0_a\big) + \big(s^0_{j_t^*} - \widehat{s}^{\mathrm{ag}}_{t,j_t^*}\big) + \big(s^0_a - \widehat{s}^{\mathrm{pr}}_{t,a}\big) > 2\beta_t \Big| \overline{C} \cup \overline{D}\Big)\mathbb{P}\Big(\overline{C} \cup \overline{D}\Big)
\end{aligned} \tag{C.239}$$

$$\leq n^2 \sum_{a\in\mathcal{A}} \mathbb{P}\big(\overline{C} \cup \overline{D}\big) \tag{C.240}$$

$$\begin{aligned}
&\leq n^2 \sum_{a\in\mathcal{A}} \mathbb{P}\big(\widehat{s}^{\mathrm{pr}}_{t,j_t^*} - s^0_{j_t^*} > \beta_t/2\big) + \mathbb{P}\big(\widehat{s}^{\mathrm{pr}}_{t,a} - s^0_a < -\beta_t/2\big)\\
&\quad + n^2 \sum_{a\in\mathcal{A}} \mathbb{P}\big(\widehat{s}^{\mathrm{ag}}_{t,a} - s^0_a > \beta_t/2\big) + \mathbb{P}\big(\widehat{s}^{\mathrm{ag}}_{t,j_t^*} - s^0_{j_t^*} < -\beta_t/2\big)
\end{aligned} \tag{C.241}$$

$$\leq 2n^2 \mathbb{P}\big(\|\widehat{\mathbf{s}}^{\mathrm{pr}}_t - \mathbf{s}^0\|_\infty > \beta_t/2\big) + n^2 \sum_{a\in\mathcal{A}} \mathbb{P}\big(\widehat{s}^{\mathrm{ag}}_{t,a} - s^0_a > \beta_t/2\big) + \mathbb{P}\big(\widehat{s}^{\mathrm{ag}}_{t,j_t^*} - s^0_{j_t^*} < -\beta_t/2\big) \tag{C.242}$$

$$\begin{aligned}
&\leq 4n^2 \exp\left(-\frac{2\lambda_t^2}{(t-1)16n(6R_{\max} - 6R_{\min} + 2\gamma)^2} - \log\frac{\beta_t}{2} + n\log(2R_{\max} - 2R_{\min})\right)\\
&\quad + n^2 \sum_{a\in\mathcal{A}} \mathbb{P}\big(\widehat{s}^{\mathrm{ag}}_{t,a} - s^0_a > \beta_t/2\big) + \mathbb{P}\big(\widehat{s}^{\mathrm{ag}}_{t,j_t^*} - s^0_{j_t^*} < -\beta_t/2\big)
\end{aligned} \tag{C.243}$$

where the last inequality follows by Corollary 4.1. We can bound the first probability term in (C.243) by recalling that $s^0_a = \mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a}$ by definition as given in (9). Further, we observe that $T(a,t) \geq T^{\mathrm{ag-xplore}}(a,t)$ for any $a \in \mathcal{A}$, and obtain

$$\begin{aligned}
&n^2 \sum_{a\in\mathcal{A}} \mathbb{P}\big(\widehat{s}^{\mathrm{ag}}_{t,a} - s^0_a > \beta_t/2\big)\\
&= n^2 \sum_{a\in\mathcal{A}} \mathbb{P}\left(\widehat{s}^{\mathrm{ag}}_{t,a} - s^0_a > \beta_t/2 \middle| T^{\mathrm{ag-xplore}}(a,t) > \mathbb{E}T^{\mathrm{ag-xplore}}(a,t) - \frac{2(m^{\mathrm{ag}})^2}{n}\right)
\end{aligned}$$

$$
\begin{aligned}
&\cdot \mathbb{P}\left(T^{\mathrm{ag-xplore}}(a,t) > \mathbb{E}T^{\mathrm{ag-xplore}}(a,t) - \frac{2(m^{\mathrm{ag}})^2}{n}\right)\\
&+ n^2 \sum_{a\in\mathcal{A}} \mathbb{P}\left(\widehat{s}^{\mathrm{ag}}_{t,a} - s^0_a > \beta_t/2 \middle| T^{\mathrm{ag-xplore}}(a,t) \leq \mathbb{E}T^{\mathrm{ag-xplore}}(a,t) - \frac{2(m^{\mathrm{ag}})^2}{n}\right)\\
&\cdot \mathbb{P}\left(T^{\mathrm{ag-xplore}}(a,t) \leq \mathbb{E}T^{\mathrm{ag-xplore}}(a,t) - \frac{2(m^{\mathrm{ag}})^2}{n}\right) \qquad (C.244)\\
\leq\ & n^2 \sum_{a\in\mathcal{A}} \mathbb{P}\left(\widehat{s}^{\mathrm{ag}}_{t,a} - s^0_a > \beta_t/2 \middle| T^{\mathrm{ag-xplore}}(a,t) > \mathbb{E}T^{\mathrm{ag-xplore}}(a,t) - \frac{2(m^{\mathrm{ag}})^2}{n}\right)\\
&+ n^2 \sum_{a\in\mathcal{A}} \mathbb{P}\left(T^{\mathrm{ag-xplore}}(a,t) \leq \mathbb{E}T^{\mathrm{ag-xplore}}(a,t) - \frac{2(m^{\mathrm{ag}})^2}{n}\right) \qquad (C.245)
\end{aligned}
$$

Now, we can use Hoeffding's Inequality (Boucheron et al. 2013) to bound the the first probability term in the last inequality above. For the second probability term, we notice that by construction of Algorithm 4, we have $T^{\mathrm{ag-xplore}}(a,t) = \sum_{\tau=1}^{t-1} \mathbf{1}(\upsilon_\tau(\boldsymbol{\pi}_\tau) = a)$ where indicator variables $\mathbf{1}(\upsilon_\tau(\boldsymbol{\pi}_\tau) = a)$'s are defined as independent Bernoulli random variables with success probabilities $\epsilon^{\mathrm{ag}}_\tau/n$. Hence, Hoeffding's Inequality can also be used for the second term.

$$
\begin{aligned}
(C.245) &\leq n^2 \sum_{a\in\mathcal{A}} \exp\left(-\frac{\left(\mathbb{E}T^{\mathrm{ag-xplore}}(a,t) - 2(m^{\mathrm{ag}})^2/n\right)\beta_t^2}{8(R_{\max}-R_{\min})^2}\right)\\
&\quad + n^2 \sum_{a\in\mathcal{A}} \exp\left(-\frac{4(m^{\mathrm{ag}})^4}{n^2}(t-1)\right) \qquad (C.246)\\
&\leq n^2 \sum_{a\in\mathcal{A}} \exp\left(-\frac{\left(\mathbb{E}T^{\mathrm{ag-xplore}}(a,t) - 2(m^{\mathrm{ag}})^2/n\right)\beta_t^2}{8(R_{\max}-R_{\min})^2}\right) + \frac{n^3}{t-1} \qquad (C.247)
\end{aligned}
$$

Now, we need to compute a lower bound for $\mathbb{E}T^{\mathrm{ag-xplore}}(a,t)$.

$$
\begin{aligned}
\mathbb{E}T^{\mathrm{ag-xplore}}(a,t) = \sum_{\tau=1}^{t-1}\frac{\epsilon^{\mathrm{ag}}_\tau}{n} = \frac{1}{n}\sum_{\tau=1}^{t-1}\min\left\{1, \frac{m^{\mathrm{ag}}}{\sqrt{\tau}}\right\} &\geq \frac{m^{\mathrm{ag}}}{n}\sum_{\tau=(m^{\mathrm{ag}})^2}^{t-1}\frac{1}{\sqrt{\tau}} \qquad (C.248)\\
&\geq \frac{m^{\mathrm{ag}}}{n}\int_{(m^{\mathrm{ag}})^2}^{t}\frac{1}{\sqrt{\tau}}d\tau \qquad (C.249)\\
&= \frac{2m^{\mathrm{ag}}}{n}(\sqrt{t} - m^{\mathrm{ag}}) \qquad (C.250)
\end{aligned}
$$

Using this lower bound on $\mathbb{E}T^{\mathrm{ag-xplore}}(a,t)$, we obtain

$$
(C.247) \leq n^3 \exp\left(\frac{-\frac{m^{\mathrm{ag}}}{n}\sqrt{t}\beta_t^2 + \frac{2(m^{\mathrm{ag}})^2}{n}\beta_t^2}{4(R_{\max}-R_{\min})^2}\right) + \frac{n^3}{t-1} \qquad (C.251)
$$

Recall $\beta_t = B\frac{\sqrt{\log 2t}}{t^{w/3}}$ with $B = \frac{3k(3(R_{\max}-R_{\min})+\gamma)^n \sqrt[6]{32n}}{1-k\sqrt{\log 2\widetilde{k}}/\sqrt{\widetilde{k}}}$. Then,

$$= n^3 \exp\left(\frac{-\frac{m^{\mathrm{ag}}}{n}B^2\sqrt{t}\frac{\log 2t}{t^{2w/3}} + \frac{2(m^{\mathrm{ag}})^2}{n}B^2\frac{\log 2t}{t^{2w/3}}}{4(R_{\max}-R_{\min})^2} \right) + \frac{n^3}{t-1} \tag{C.252}$$

$$\leq n^3 \exp\left(-\sqrt{t}\frac{\log 2t}{t^{2w/3}} + \frac{2(m^{\mathrm{ag}})^2 B^2}{4n(R_{\max}-R_{\min})^2} \right) + \frac{n^3}{t-1} \tag{C.253}$$

$$\leq \frac{n^3\left(\exp\left(\frac{2(m^{\mathrm{ag}})^2B^2}{4n(R_{\max}-R_{\min})^2}\right)+1\right)}{t-1} \tag{C.254}$$

where the second to the last inequality holds for all $t \geq 2$. This gives us

$$n^2\sum_{a\in\mathcal{A}}\mathbb{P}\big(\widehat{s}^{\mathrm{ag}}_{t,a} - s^0_a > \beta_t/2\big) \leq \frac{n^3\left(\exp\left(\frac{2(m^{\mathrm{ag}})^2B^2}{4n(R_{\max}-R_{\min})^2}\right)+1\right)}{t-1} \tag{C.255}$$

Similarly, we have

$$n^2\sum_{a\in\mathcal{A}}\mathbb{P}\big(\widehat{s}^{\mathrm{ag}}_{t,j^*_t} - s^0_{j^*_t} < -\beta_t/2\big) \leq \frac{n^3\left(\exp\left(\frac{2(m^{\mathrm{ag}})^2B^2}{4n(R_{\max}-R_{\min})^2}\right)+1\right)}{t-1} \tag{C.256}$$

Substituting these upper bounds into (C.243) and combining everything, we conclude as

$$\begin{aligned} &\mathbb{P}\Big(\upsilon_t(\boldsymbol{\pi}_t) \neq \arg\max_{a\in\mathcal{A}} s^0_a + \pi_{t,a}\Big| t \in \mathcal{T}^{\mathrm{ag-xploit}} \cap \mathcal{T}^{\mathrm{pr-xploit}}\Big) \\ &\leq \frac{2n^3\left(\exp\left(\frac{2(m^{\mathrm{ag}})^2B^2}{4n(R_{\max}-R_{\min})^2}\right)+1\right)}{t-1} \\ &\quad + 8n^2\exp\left(-\frac{2\lambda_t^2}{(t-1)16n(6R_{\max}-6R_{\min}+2\gamma)^2} - \log\frac{\beta_t}{2} + n\log(2R_{\max}-2R_{\min})\right) \end{aligned} \tag{C.257}$$

□

Proof of Lemma 4.5. By construction of the principal's algorithm (3) and agent's algorithm (4), at any time point $t \in \mathcal{T}^{\mathrm{ag-xploit}} \cap \mathcal{T}^{\mathrm{pr-xplore}}$, we know that the arm chosen by the agent is given as $\upsilon_t(\boldsymbol{\pi}_t) = \arg\max_{a\in\mathcal{A}} \widehat{s}^{\mathrm{ag}}_{t,a} + \pi_{t,a}$ and the incentives provided by the principal are uniformly randomly selected from set $\mathcal{C}$. Then, we start with

$$\begin{aligned} &\mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t) \neq \arg\max_{a\in\mathcal{A}} s^0_a + \pi_{t,a}\Big| t \in \mathcal{T}^{\mathrm{ag-xploit}} \cap \mathcal{T}^{\mathrm{pr-xplore}}\right) \\ &= \mathbb{P}\Big(\upsilon_t(\boldsymbol{\pi}_t) \neq \arg\max_{a\in\mathcal{A}} s^0_a + \pi_{t,a} \end{aligned}$$

$$\Bigg| \upsilon_t(\boldsymbol{\pi}_t) = \operatorname*{arg\,max}_{a\in\mathcal{A}} \widehat{s}^{\mathrm{ag}}_{t,a} + \pi_{t,a},\ \ \pi_{t,a} \sim \mathcal{U}(\underline{C},\overline{C})\ \forall a \in \mathcal{A}\Bigg) \tag{C.258}$$

$$= \mathbb{P}\Bigg(\bigcup_{a''\in\mathcal{A}} \bigcup_{a'\in\mathcal{A}} s^0_{a''} + \pi_{t,a''} < s^0_{a'} + \pi_{t,a'} \\ \Bigg| \widehat{s}^{\mathrm{ag}}_{t,a} + \pi_{t,a} \leq \widehat{s}^{\mathrm{ag}}_{t,a''} + \pi_{t,a''}\ \forall a \in \mathcal{A},\ \ \pi_{t,a} \sim \mathcal{U}(\underline{C},\overline{C})\ \forall a \in \mathcal{A}\Bigg) \tag{C.259}$$

$$\leq \sum_{a''\in\mathcal{A}} \sum_{a'\in\mathcal{A}} \mathbb{P}\Big(s^0_{a''} + \pi_{t,a''} < s^0_{a'} + \pi_{t,a'} \\ \Big| \widehat{s}^{\mathrm{ag}}_{t,a} + \pi_{t,a} \leq \widehat{s}^{\mathrm{ag}}_{t,a''} + \pi_{t,a''}\ \forall a \in \mathcal{A},\ \ \pi_{t,a} \sim \mathcal{U}(\underline{C},\overline{C})\ \forall a \in \mathcal{A}\Big) \tag{C.260}$$

$$\leq \sum_{a''\in\mathcal{A}} \sum_{a'\in\mathcal{A}} \mathbb{P}\left(\widehat{s}^{\mathrm{ag}}_{t,a'} - \widehat{s}^{\mathrm{ag}}_{t,a''} \leq \pi_{t,a''} - \pi_{t,a'} < s^0_{a'} - s^0_{a''} \middle| \pi_{t,a} \sim \mathcal{U}(\underline{C},\overline{C}), \forall a \in \mathcal{A}\right) \tag{C.261}$$

where (C.260) follows by the Boole's inequality (a.k.a. union bound).

Now, we recall that $s_a^0 = \mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a}$ for any $a \in \mathcal{A}$ by definition given in (9) and continue our derivation by conditioning on the intersection of the two events $E = \{\widehat{s}^{\mathrm{ag}}_{t,a''} - \mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a''} < \varphi_t\}$ and $F = \{\mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a'} - \widehat{s}^{\mathrm{ag}}_{t,a'} < \varphi_t\}$ for some $\varphi_t > 0$.

$$
\begin{aligned}
&= \sum_{a'' \in \mathcal{A}} \sum_{a' \in \mathcal{A}} \mathbb{P}\left(\widehat{s}^{\mathrm{ag}}_{t,a'} - \widehat{s}^{\mathrm{ag}}_{t,a''} \le \pi_{t,a''} - \pi_{t,a'} < \mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a'} - \mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a''} \middle| \pi_{t,a} \sim \mathcal{U}(\underline{C}, \overline{C}), \forall a \in \mathcal{A}, E \cap F\right) \\
&\qquad \cdot \mathbb{P}(E \cap F) \\
&\quad + \sum_{a'' \in \mathcal{A}} \sum_{a' \in \mathcal{A}} \mathbb{P}\left(\widehat{s}^{\mathrm{ag}}_{t,a'} - \widehat{s}^{\mathrm{ag}}_{t,a''} \le \pi_{t,a''} - \pi_{t,a'} < \mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a'} - \mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a''} \middle| \pi_{t,a} \sim \mathcal{U}(\underline{C}, \overline{C}), \forall a \in \mathcal{A}, \overline{E} \cup \overline{F}\right) \\
&\qquad \cdot \mathbb{P}(\overline{E} \cup \overline{F})
\end{aligned} \tag{C.262}
$$

$$
\begin{aligned}
&\le \sum_{a'' \in \mathcal{A}} \sum_{a' \in \mathcal{A}} \mathbb{P}\left(\widehat{s}^{\mathrm{ag}}_{t,a'} - \widehat{s}^{\mathrm{ag}}_{t,a''} \le \pi_{t,a''} - \pi_{t,a'} < \mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a'} - \mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a''} \middle| \pi_{t,a} \sim \mathcal{U}(\underline{C}, \overline{C}), \forall a \in \mathcal{A}, E \cap F\right) \\
&\quad + \sum_{a'' \in \mathcal{A}} \sum_{a' \in \mathcal{A}} \mathbb{P}(\overline{E}) + \mathbb{P}(\overline{F})
\end{aligned} \tag{C.263}
$$

For the first term in (C.263), we recall that the difference of the two Uniform random variables, $\pi_{t,a''} - \pi_{t,a'}$, follows a triangular distribution whose cdf is derived in (B.12). Because the cdf is in the form of a piecewise function, we can compute an upper bound to the first term above by considering the case with the highest probability, that is $\widehat{s}^{\mathrm{ag}}_{t,a'} - \widehat{s}^{\mathrm{ag}}_{t,a''} < 0$ and $\mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a'} - \mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a''} \ge 0$ for any $a', a'' \in \mathcal{A}$. Then, it follows that

$$
\sum_{a'' \in \mathcal{A}} \sum_{a' \in \mathcal{A}} \mathbb{P}\left(\widehat{s}^{\mathrm{ag}}_{t,a'} - \widehat{s}^{\mathrm{ag}}_{t,a''} \le \pi_{t,a''} - \pi_{t,a'} < \mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a'} - \mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a''} \middle| \pi_{t,a} \sim \mathcal{U}(\underline{C}, \overline{C}), \forall a \in \mathcal{A}, E \cap F\right)
$$

$$
\le \sum_{a'' \in \mathcal{A}} \sum_{a' \in \mathcal{A}} \left[1 - \frac{(\mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a'} - \mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a''} + \underline{C} - \overline{C})^2}{2(\overline{C} - \underline{C})^2} - \frac{(\widehat{s}^{\mathrm{ag}}_{t,a'} - \widehat{s}^{\mathrm{ag}}_{t,a''} + \overline{C} - \underline{C})^2}{2(\overline{C} - \underline{C})^2} \middle| E \cap F\right] \tag{C.264}
$$

$$
= \sum_{a'' \in \mathcal{A}} \sum_{a' \in \mathcal{A}} \left[\frac{-(\mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a'} - \mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a''})^2 - (\widehat{s}^{\mathrm{ag}}_{t,a'} - \widehat{s}^{\mathrm{ag}}_{t,a''})^2 + 2(\overline{C} - \underline{C})(\mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a'} - \mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a''} - \widehat{s}^{\mathrm{ag}}_{t,a'} + \widehat{s}^{\mathrm{ag}}_{t,a''})}{2(\overline{C} - \underline{C})^2} \middle| E \cap F\right] \tag{C.265}
$$

$$
\le \sum_{a'' \in \mathcal{A}} \sum_{a' \in \mathcal{A}} \left[\frac{\mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a'} - \mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a''} - \widehat{s}^{\mathrm{ag}}_{t,a'} + \widehat{s}^{\mathrm{ag}}_{t,a''}}{\overline{C} - \underline{C}} \middle| E \cap F\right] \tag{C.266}
$$

$$
\le \sum_{a'' \in \mathcal{A}} \sum_{a' \in \mathcal{A}} \frac{2\varphi_t}{\overline{C} - \underline{C}} \tag{C.267}
$$

$$
= \frac{2n^2 \varphi_t}{\overline{C} - \underline{C}} \tag{C.268}
$$

For the second and third terms in (C.263) (which are essentially same with each other), we follow the same arguments as in between (C.244)-(C.251) from the proof of Lemma 4.4.

Then, we get

$$\sum_{a''\in\mathcal{A}}\sum_{a'\in\mathcal{A}}\mathbb{P}(\overline{E})+\mathbb{P}(\overline{F})=\sum_{a''\in\mathcal{A}}\sum_{a'\in\mathcal{A}}\mathbb{P}\left(\widehat{s}^{\mathrm{ag}}_{t,a''}-\mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a''}\geq\varphi_t\right)+\mathbb{P}\left(\mathbb{E}\widehat{s}^{\mathrm{ag}}_{t,a'}-\widehat{s}^{\mathrm{ag}}_{t,a'}\geq\varphi_t\right) \tag{C.269}$$

$$\leq 2n^2\exp\left(\frac{-\frac{m^{\mathrm{ag}}}{n}\sqrt{t}\varphi_t^2+\frac{2(m^{\mathrm{ag}})^2}{n}\varphi_t^2}{4(R_{\max}-R_{\min})^2}\right)+\frac{2n^2}{t-1} \tag{C.270}$$

Suppose $\varphi_t=\dfrac{2\sqrt{n}(R_{\max}-R_{\min})\sqrt{\log 2t}}{\sqrt{m^{\mathrm{ag}}}\sqrt[4]{t}}$. Then,

$$\leq 2n^2\exp\Big(-\log 2t+2m^{\mathrm{ag}}\frac{\log 2t}{\sqrt{t}}\Big)+\frac{2n^2}{t-1} \tag{C.271}$$

$$\leq 2n^2\exp\big(-\log 2t+2m^{\mathrm{ag}}\big)+\frac{2n^2}{t-1} \tag{C.272}$$

$$=2n^2\frac{(\exp(2m^{\mathrm{ag}})+1)}{t-1} \tag{C.273}$$

For the same choice of φ_t above, we combine the last result with (C.268) and obtain

$$\mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t)\neq\arg\max_{a\in\mathcal{A}}s_a^0+\pi_{t,a}\middle|t\in\mathcal{T}^{\mathrm{ag-xploit}}\cap\mathcal{T}^{\mathrm{pr-xplore}}\right)$$

$$\leq\frac{4n^2\sqrt{n}(R_{\max}-R_{\min})\sqrt{\log 2t}}{\sqrt{m^{\mathrm{ag}}}(\overline{C}-\underline{C})\sqrt[4]{t}}+\frac{2n^2(\exp(2m^{\mathrm{ag}})+1)}{t-1} \tag{C.274}$$

$$\leq\frac{4n^2\sqrt{n}\sqrt{\log 2t}}{\sqrt{m^{\mathrm{ag}}}\sqrt[4]{t}}+\frac{2n^2(\exp(2m^{\mathrm{ag}})+1)}{t-1} \tag{C.275}$$

where the last inequality follows since $R_{\max}-R_{\min}<\overline{C}-\underline{C}$ by Assumption 4.1. □

Proof of Proposition 4.8. We prove this result by using the induction technique.

Base Case: Consider $t=1$. We have $1\leq(\text{constant})\sqrt{\log 2}$ for any (constant) ≥ 1.2. Then, since $p_1\leq 1$ by definition, it trivially satisfies Proposition 4.8.

Induction Step: Consider any $t\in[\widetilde{k},T]$. Suppose that Proposition 4.8 holds for all p_τ such that $\tau\in[1,t-1]$. Then, we have

$$p_t=\mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t)\neq\arg\max_{a\in\mathcal{A}}s_a^0+\pi_{t,a}\right) \tag{C.276}$$

$$=\mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t)\neq\arg\max_{a\in\mathcal{A}}s_a^0+\pi_{t,a}\middle|t\in\mathcal{T}^{\mathrm{ag-xploit}}\cap\mathcal{T}^{\mathrm{pr-xploit}}\right)\mathbb{P}\left(t\in\mathcal{T}^{\mathrm{ag-xploit}}\cap\mathcal{T}^{\mathrm{pr-xploit}}\right)$$

$$+\mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t)\neq\arg\max_{a\in\mathcal{A}}s_a^0+\pi_{t,a}\middle|t\in\mathcal{T}^{\mathrm{ag-xploit}}\cap\mathcal{T}^{\mathrm{pr-xplore}}\right)\mathbb{P}\left(t\in\mathcal{T}^{\mathrm{ag-xploit}}\cap\mathcal{T}^{\mathrm{pr-xplore}}\right)$$

$$
\begin{aligned}
&+ \mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t) \neq \arg\max_{a\in\mathcal{A}} s_a^0 + \pi_{t,a} \middle| t \in \mathcal{T}^{\text{ag}-\text{xplore}} \right) \mathbb{P}\left(t \in \mathcal{T}^{\text{ag}-\text{xplore}} \right) && \text{(C.277)} \\
&\leq \mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t) \neq \arg\max_{a\in\mathcal{A}} s_a^0 + \pi_{t,a} \middle| t \in \mathcal{T}^{\text{ag}-\text{xploit}} \cap \mathcal{T}^{\text{pr}-\text{xploit}} \right) \\
&+ \mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t) \neq \arg\max_{a\in\mathcal{A}} s_a^0 + \pi_{t,a} \middle| t \in \mathcal{T}^{\text{ag}-\text{xploit}} \cap \mathcal{T}^{\text{pr}-\text{xplore}} \right) \mathbb{P}\left(t \in \mathcal{T}^{\text{pr}-\text{xplore}} \right) \\
&+ \mathbb{P}\left(t \in \mathcal{T}^{\text{ag}-\text{xplore}} \right) && \text{(C.278)} \\
&\leq \frac{2n^3 \left(\exp\left(\frac{2(m^{\text{ag}})^2 B^2}{4n(R_{\max}-R_{\min})^2} \right) + 1 \right)}{t-1} \\
&+ 8n^2 \exp\left(-\frac{2\lambda_t^2}{(t-1)16n(6R_{\max} - 6R_{\min} + 2\gamma)^2} - \log\frac{\beta_t}{2} + n\log(2R_{\max} - 2R_{\min}) \right) \\
&+ \left(\frac{4n^2\sqrt{n}\sqrt{\log 2t}}{\sqrt{m^{\text{ag}}}\sqrt[4]{t}} + \frac{2n^2(\exp(2m^{\text{ag}}) + 1)}{t-1} \right) \frac{m^{\text{pr}}}{t^{1/2-w}} + \frac{m^{\text{ag}}}{\sqrt{t}} && \text{(C.279)}
\end{aligned}
$$

where the last inequality follows by the results of Lemma 4.4 and Lemma 4.5 and by the constructions of Algorithm 3 and Algorithm 4.

Further, we can bound the second term in the right-hand side of the last inequality by following the same arguments as in between (C.148)-(C.161) and obtain

$$
\begin{aligned}
&\exp\left(-\frac{2\lambda_t^2}{(t-1)16n(6R_{\max} - 6R_{\min} + 2\gamma)^2} - \log\frac{\beta_t}{2} + n\log(2R_{\max} - 2R_{\min}) \right) \\
&\qquad \leq \frac{2^{n+1}}{3^{n+1}k\sqrt[6]{32n}} \frac{1}{\sqrt{t\log 2t}} && \text{(C.280)}
\end{aligned}
$$

Combining this upper bound with (C.279), we obtain the following upper bound on p_t.

$$
\begin{aligned}
p_t &= \mathbb{P}\left(\upsilon_t(\boldsymbol{\pi}_t) \neq \arg\max_{a\in\mathcal{A}} s_a^0 + \pi_{t,a} \right) && \text{(C.281)} \\
&\leq \frac{2n^3 \left(\exp\left(\frac{2(m^{\text{ag}})^2 B^2}{4n(R_{\max}-R_{\min})^2} \right) + 1 \right)}{t-1} + \frac{2^{n+4}n^{11/6}}{3^{n+1}k\sqrt[6]{32}} \frac{1}{\sqrt{t\log 2t}} \\
&+ \left(\frac{4n^2\sqrt{n}\sqrt{\log 2t}}{\sqrt{m^{\text{ag}}}\sqrt[4]{t}} + \frac{2n^2(\exp(2m^{\text{ag}}) + 1)}{t-1} \right) \frac{m^{\text{pr}}}{t^{1/2-w}} + \frac{m^{\text{ag}}}{\sqrt{t}} && \text{(C.282)}
\end{aligned}
$$

which implies $p_t = O\Big(\frac{\sqrt{\log 2t}}{\sqrt{t}}\Big)$.

Conclusion: As both the base case and the inductive step have been proven as true, Proposition 4.8 is established by mathematical induction. □

C.2 Parameters for Numerical Experiments

In our numerical experiments, we test the performance of our data-driven incentive policy for different values of the cardinality of the aggregator's MAB model ($n = |\mathcal{A}|$). The closed intervals that the expected profits of the aggregator and the utility company are set to $\Theta = [0, 100]$ and $\mathcal{R} = [-20, 50]$, respectively, and the entries of the expected profit vectors $\mathbf{r}^0$ and $\boldsymbol{\theta}^0$ are uniformly randomly generated from these intervals as reported below.

n	$\boldsymbol{\theta}^0$	$\mathbf{r}^0$
5	(29, 1, 14, 26, 15)	(14, -24, -4, 19, 29)
10	(0, 44, 51, 65, 9, 35, 69, 91, 51, 44)	(-4, 8, 22, -12, -2, 46, -8, 16, 38, 14)

Table C.1: Expected profits in different settings of the size of the aggregator's MAB model