Learning-based Decision-making under Stochastic and Adversarial Uncertainties

by

Yili Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Applied and Interdisciplinary Mathematics)
in the University of Michigan
2023

Doctoral Committee:

Assistant Professor Asaf Cohen, Co-Chair
Associate Professor Vijay G Subramanian, Co-Chair
Professor Erhan Bayraktar
Research Assistant Professor Ali Kara
Professor Virginia Young

Yili Zhang

zhyili@umich.edu

ORCID iD: 0000-0002-1380-2083

DEDICATION

To my grandma who never saw the completion of this journey, and to myself three years ago who decided to continue this journey.

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my advisors—Professor Asaf Cohen and Professor Vijay Subramanian—not only for their patience and rigorous mentorship, but also for their constant support and encouragement, especially during the hard times of the pandemic. It is my great pleasure to have them as my advisors, and I cannot imagine the completion of this thesis without their help.

I would also like to extend my deepest gratitude to my committee members and my collaborators: Professor Erhan Bayraktar, Professor Virginia Young, Postdoctoral Professor Ali Kara, and Professor Ibrahim Ekren. I am grateful for my committee members' thorough reviewing of my work. Their insightful feedback has helped to shape this thesis into a better version. My first collaborators Professor Bayraktar and Professor Ekren not only introduced me to the field of machine learning but also exposed me to examples of successful researchers.

I must also thank everyone from the math graduate office for being supportive and helpful during my studying as a Ph.D. student at Umich. Their exceptional day-to-day work has not only ensured that I had a well planned defense, but also provided invaluable support throughout my role as a graduate student instructor.

I would like to have my special thanks to all my family: my parents, grandparents, aunts, uncles, sisters, and brothers. We were not able to be physically close yet we were supportive to each other through instant massaging during the past few years. I wish I were there when we lost our dearest grandmother, mother, and mentor.

I have met a lot of great people here in Ann Arbor, and I would like to say thank you to all of them for being part of my journey and creating unforgettable memories with me: Jiajia

Guo, Ryan Kohl, Kexin Li, Yiqing Lu, Christiana Mavroyiakoumou, Binglin Song, Nancy Wang, Yueqiao Wu, Jingjie Zhang, Xin Zhang and so many more that I would not be able to make it an inclusive list.

I would also like to mention my thanks to Oreo, Pie, and O-tilde. You are my best cats.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

FIGURE

ABSTRACT

This thesis studies two online learning problems in which the efficiency of the proposed strategies is studied in terms of their regret. The first problem deals with designing learning algorithms that optimize the social welfare of a single server queuing system when both the arrival and service rates are unknown and the second one involves finding an asymptotically optimal strategy for the problem of prediction with expert advice.

In Chapter II, we consider a long-term average profit maximizing admission control problem in an M/M/1 queuing system with unknown service and arrival rates. We propose a learning-based dispatching algorithm and characterize its regret with respect to optimal threshold dispatch policies. We show that our algorithm achieves $O(1)$ regret when feasible, and $O(\ln^{1+\epsilon}(N))$ otherwise for every $\epsilon > 0$ where N is the number of customers arrived.

In Chapter III, we consider the problem of prediction with expert advice under the adversarial setting with a geometric stopping time from a zero-sum game point of view. In the literature, it has been shown in this setting that the discounted cost problem yields the solution of the long-term average cost problem as the discount parameter goes to 0, when quantities are appropriately scaled. We will focus on this asymptotic regime, and hence, the long-term average behavior of this game. We show that it is optimal for nature to play the "comb strategies" in the 4 expert game and that the best regret grows as $O(1/\sqrt{2\delta})$, where δ is the parameter of the geometric stopping time.

CHAPTER I

Introduction

Online machine learning refers to a learning paradigm where the model is trained on data that is presented in a sequential fashion. A huge variety of learning problems can be considered online learning problems, and the techniques used to solve these may vary from problem to problem. Yet, links, similarities, and methods inspired by each other may also appear. To measure the performance of the proposed strategies, regret, which is often defined to be the difference between the optimal gain and actual gain, is a popular and natural choice of the objective to analyze; see [Hoi et al. 2021; Hazan 2021] and their references. An online machine learning algorithm can adjust its predictions and decisions at a low computational cost according to the new data that has arrived. This enables the algorithm to adapt and improve the model it learns in real-time. This paradigm is useful in scenarios where data arrives continuously.

This thesis is devoted to two problems in online learning where decisions need to be made repeatedly under uncertainty. The goals and methodologies used for these two problems are quite different. The first problem focuses on designing admission policies in a single server queuing system to optimize the long-term average social welfare; see Chapter II. The second problem deals with finding an asymptotic optimal strategy for the problem of predicting with expert advice under adversarial environments; see Chapter III. Despite the different types of uncertainty considered in the two problems, we aim to learn and analyze an adaptive control that achieves sub-linear regret in both problems. Both problems fall under the topic of model-based adaptive control as discussed in books [Landau et al. 2011; Ioannou and Sun 2012; Benosman 2016]. The trade-off between exploitation and exploration is also a common challenge in the dispatching/scheduling problems in queuing systems and the problem of learning with expert advice, in addition to other online learning problems, for example, multi-armed bandit problems and problems in Markov Decision Processes (MDPs) that are developed in Reinforcement Learning (RL).

Learning in queuing systems naturally falls under the category of online learning since customers usually arrive at the system in sequential order. Various learning problems have been considered with different models, see survey [Walton and Xu 2021], which also points out the link between queuing theoretic results with adversarial learning, and also [Neely 2010]. Multi-armed bandit problems have been thoroughly studied and many algorithms are developed and analyzed, see [Lattimore and Szepesvári 2020b]. Adapting the classical algorithms from this topic, [Yang, Srikant, and Ying 2023] and [Krishnasamy, Sen, et al. 2021] studied the problem of learning the scheduling policy in queuing systems and proposed algorithms that achieve sub-linear regret when model statistics are unknown. Also working on learning a scheduling policy in queuing systems, [Agrawal and R. Jia 2022] and [Salodkar et al. 2008] use an MDP formulation to analyze this learning problem and proposed algorithms that converge to the optimal scheduling policy. An overview of applications of MDPs in queues is provided in [Li et al. 2019]. With applications in routers and network switches, [Kahe and Jahangir 2019], [Xiong et al. 2010] and [H. Zhang et al. 2003] developed self-tuning Active Queue Management algorithms that do not require careful tuning of the parameters to reduce the congestion in the network.

Motivated by the job dispatching problem for online computing demands, in Chapter II, we consider a learning problem of an $M/M/1$ queuing model that was first proposed and analyzed in [Naor 1969]. We design a two-phase algorithm with a skippable explicit exploration phase. Our algorithm achieves $O(1)$ regret on the regret accumulated when feasible and $O(\ln^{1+\epsilon}(N))$ for every $\epsilon > 0$ otherwise, where N is the number of arrivals.

In this model, a positive reward R is obtained after a customer finishes service, and a holding cost of C is collected per unit of time for each customer in the system (including any customer in service). Upon the arrival of a customer, with the goal of optimizing the long-term average social welfare, a dispatcher decides whether to admit this customer to the queue. Having joined the queue, the customer only leaves after service is received; the customer does not return to the system if denied entry. If all the model parameters are known, then the optimal admission policies found in [Naor 1969] is a threshold policy in which arrivals are not admitted whenever the system occupancy (queue-length plus customer in service) reaches a static threshold value.

We consider the learning problem where both the arrival and service rates are unknown but the reward R and per unit time holding cost C are known. Our learning-based algorithm consists of batches with each batch being composed of an optional fixed-length forced exploration phase (phase 1) and an exploitation phase (phase 2) whose length increases with batch index. During the (potential) exploration phase, every arriving customer is admitted to the system. During the exploitation phase, the learning dispatcher uses a fixed threshold that is computed using all samples of the service completions and inter-arrival times collected from all prior exploitation phases as well as exploration phases. In general systems, one would first need to show the stability of the queue under the learning algorithm before characterizing regret performance. However, since our proposed learning algorithm has a fixed-length exploration phase and uses a threshold policy in the exploitation phase, the queue is stable with probability 1 under the proposed learning algorithm. To reduce the regret we omit the exploration phase in the next batch if the threshold used in the exploitation phase of the previous batch (that just ended to start the new batch) is non-zero.

The expected regret is defined as the difference between the accumulated net gain of an optimal dispatcher and that of the learning dispatcher. We show how to translate the regret from a given time horizon to a function of the number of arrivals within said time horizon, and hence, we work with analyzing the queuing system at customer arrivals. Then we show that our learning dispatcher is able to achieve $O(1)$ regret when all optimal thresholds with full information are non-zero, and achieves an $O(\ln^{1+\epsilon}(N))$ regret for $\epsilon > 0$ in the case that an optimal threshold with full information is 0 (i.e., an optimal policy is to reject all arrivals), where N is the number of arrivals. We also perform numerical experiments to examine the finite-time performance of our algorithm. This chapter is based on [Cohen, V. G. Subramanian, and Y. Zhang 2022]. Part of the work has been presented at the 58th Annual Allerton Conference on Communication, Control, and Computing at the University of Illinois at Urbana-Champaign (September, 2022).

The problem of learning with expert advice for an agent has been studied widely from the perspective of designing algorithms (for the agent which can also be referred to as the player) that achieve sub-linear accumulated regret when comparing the gain of the agent with the best-performing expert. This problem is typically regarded as a zero-sum game between the agent and nature, with the optimal min-max regret representing the value of this game. Both upper and lower regret bounds for regret of the agents have been studied under various settings; see [Cesa-Bianchi and Lugosi 2006] for a comprehensive study. It is well known that the optimal regret of $O(\sqrt{T\log(N)})$ can be achieved by the weighted average algorithm [Littlestone and Warmuth 1994; Vovk 1990] where T is the number of the rounds and N is the number of the experts when the loss function is bounded and convex. Recent work has discussed more sophisticated regret bounds, see [Koolen and Erven 2015; Mhammedi, Koolen, and Erven 2019], and proposed algorithms that can outperform the best expert [Hoeven, Zhivotovskiy, and Cesa-Bianchi 2022]. The optimality of an adaptive weighted-average algorithm, when one expert is malicious, is proved in [Bayraktar, Poor, and X. Zhang 2021]. Taking a PDE approach, [Bayraktar, Ekren, and X. Zhang 2022] worked in Wasserstein space and considered the expert problem with partial information where both nature (again the adversary) and the agent update their belief of the distribution on the accumulative rewards of the experts by performing a Bayesian update, and [Bayraktar, Ekren, and X. Zhang 2021] analyzed the situation when both the agent and nature have common knowledge of the probability vector of the gain of experts, and nature can at most affect the gain of one expert.

We focus on finding the asymptotically optimal strategy for nature in a repeatedly played zero-sum game between an agent(or player) and nature in Chapter III. The maturity time of this game is determined by a geometric random variable with parameter $\delta > 0$; note that this leads to a discounted cost formulation as the game stops at this time. The two parties of this game interact through a set of N experts. At each round, with the history of the gains of each expert and the player available to both parties of the game, the player and nature choose subsets of the experts simultaneously. The gain of the expert(s) chosen by nature increases by 1, and so does the reward of the player if this expert is also chosen by the player. The regret of the player at time t is defined to be the difference between the gain of the best-performing expert and the gain of the player. The player aims to minimize the regret while nature aims to maximize it. The lower and upper values of this game are given by max-min and min-max of the expected regret. This problem was first studied in [Cover 1967] for the 2-experts case. Observed that it is a weakly dominant strategy for nature to choose from one of the vertices of the convex polytope of the set of possible distributions over the experts. Optimal strategies of the player and nature were found by solving the dynamic programming equation directly in [Gravin, Peres, and Sivan 2016] for the 3-experts case. A re-scaled game was studied in [Drenska 2017] where a partial differential equation (PDE) is derived for the limiting value function of the re-scaled game.

We explicitly solve the PDE that is derived in [Drenska 2017], and show that an optimal strategy for nature in the 4-expert case is the following. In each round, nature divides the experts into two sets. One set contains the experts with the highest and third-highest gains. The other contains the experts with the second and fourth-highest gains. Ties are broken at random. Nature then chooses between the two sets uniformly at random. This strategy is called the “comb strategy”. It is a weakly dominant strategy for nature and was first introduced in [Gravin, Peres, and Sivan 2016]. We also prove that the regret of the limiting game grows as $O(\pi/\sqrt{32\delta})$ as δ goes to 0. Both the optimal strategy for nature, and the order of the regret were conjectured in [Gravin, Peres, and Sivan 2016], and hence, our work settles the conjecture. This chapter is based on [Bayraktar, Ekren, and Y. Zhang 2020]. Part of the work has been presented in the Financial/Actuarial Mathematics Seminar at the University of Michigan (April, 2019).

CHAPTER II

Learning-based Optimal Admission Control In A Single Server Queuing System

2.1 Introduction

We consider admission control for a first-in-first-out (FIFO) single-class single-server queuing model with Poisson arrivals and exponential service times. Specifically, there is a dispatcher that decides on admitting arrivals with the goal to maximize the long-term average profit – each admitted arrival yields a positive reward R (obtained after a customer finishes service), which is balanced by a holding cost for the (homogeneous) customers waiting in the queue. The buffer capacity of this queue is infinite and the dispatcher may decide upon arrivals to reject any customers joining the queue with the profit objective in mind. When the service and arrival rates are known, this model was studied in [Naor 1969]. In our investigation, we will consider the situation where the dispatcher does not have knowledge of either the arrival rate or the service rate. One potential application is the job dispatching problem for online computing demands, especially when the computing servers are provided by a third-party cloud computing platform: the dispatcher may negotiate the reward and cost with the customers, and thus, have information (via market research) on the arrival rate of the jobs, but since the servers are provided by a third-party platform, the dispatcher may not know the service rate. Despite prior market research, it is, however, plausible that the dispatcher doesn't know the arrival rate accurately.

[Naor 1969] studied two problems: 1) the optimal policy for the self-optimization problem where customers are maximizing their own net (expected) profit so that a selfish Wardrop equilibrium is of interest; as well as 2) the optimal policy for the social welfare maximization problem where a dispatcher is aiming at maximizing the long-term average profit so that a social Wardrop equilibrium is of interest. In both problems, a threshold policy was shown to be optimal: 1) in the self-optimization problem, arrivals do not join the queue if the queue-length upon arrival is high enough; and 2) in the social-welfare maximization problem, the dispatcher doesn't admit arrivals whenever a threshold level is reached. [Naor 1969] showed that the threshold for the social welfare maximization problem is not greater than the threshold for the self-optimization problem. Our investigation and the accompanying algorithm are primarily designed for the social welfare optimization problem where the dispatcher is interested in learning how to perform at the same level of efficiency as if knowing the actual arrival and service rate. Any learning-based algorithm will necessarily need exploration which could violate incentive-compatibility constraints (even *ex-ante* and not only *ex-post*) of individual utility maximizing agents. Hence, we do not consider the self-optimization version of the problem in this manuscript.

In our analysis, we will couple two queuing systems: a *learning system*, whose dispatcher does not know the arrival and service rate *apriori*, and a *genie-aided* system, whose dispatcher has full information of the model parameters. We refer to the corresponding algorithm and dispatcher of the two systems as the *learning algorithm*, *learning dispatcher* and *genie-aided algorithm*, *genie-aided dispatcher*, respectively. Our figure of metric at a given time t will be the difference between the net expected profits of a genie-aided algorithm and the learning algorithm, i.e., the expected regret.

Contributions: We propose a learning-based dispatching algorithm that achieves an $O(1)$ regret when (genie-aided) optimal algorithms use a non-zero threshold, and achieves an

$O(\ln^{1+\epsilon}(N))$ regret for any specified $\epsilon > 0$ when it is optimal to use threshold 0, where N denotes the number of arrivals [1]; see Remark 2.4.3 for a refinement on the achievable regret. Our learning-based algorithm consists of batches with each batch being composed of an optional forced exploration phase (phase 1) and an exploitation phase (phase 2) whose length increases with batch index. The exploration phase is omitted if there are new samples collected from the exploitation phase that just ended. Our learning algorithm uses samples collected from all the exploitation phases as well as from any exploration phases; the former is important if the exploration phase is omitted.

For the system studied in [Naor 1969], not all values of the unknown model parameters result in a unique optimal static threshold policy. For some specific choices of the model parameters, there exist two optimal static thresholds, and therefore all the policies that stochastically alternate between the two static optimal thresholds also achieve the optimal long-term average profit. As mentioned earlier, we are interested in analyzing the regret – defined to be the difference between the expected profit of the learning and genie-aided systems. When the optimal policy is unique, there is no ambiguity in the definition of the regret as there is a fixed optimal policy to compare against. However, when there are multiple policies that are optimal, we need to specify a particular optimal policy that we are comparing against. Among the multiple optimal policies, we compare against a policy with a specific way of randomizing between the two static optimal thresholds, and then we prove that we can achieve similar regret as when there exists a unique optimal policy, which is of order $O(1)$ when both thresholds are positive, and of order $O(\ln^{1+\epsilon}(N))$ for any specified $\epsilon > 0$ when 0 is an optimal threshold and N is the number of customers that have arrived; Remark 2.4.3 applies with non-unique thresholds too.

In our setting, we do not exclude the case where the genie-aided dispatcher uses a static threshold 0, and hence, rejects all customers. This leads to a balancing act for the

[1] We show how to translate the regret from the number of arrivals to a time horizon.

dispatcher: quickly transitioning to reject all customers if the true threshold is 0 versus admitting customers infinitely often otherwise (based on the optimal threshold), and all of this while not being aware of the true optimal admission policy. With this in mind, for learning to not stall, the existence of the exploration phase is crucial when the true threshold is positive. A naive learning scheme that only uses the empirical average service time as an estimate of the unknown parameter may perform poorly: a few extremely long service times at the beginning may mislead the learning dispatcher to think that the service rate is low, and hence, result in it not accepting customers into the queue even when the genie-aided dispatcher uses a non-zero threshold; see plots in Section 2.6.

Related work: On the topic of finding optimal controls vis-a-vis individual and social welfare maximization, there are many models that have studied generalizations of the model introduced in [Naor 1969]. [Knudsen 1972] generalized the model in [Naor 1969] to multiple servers with a non-linear cost for customers waiting in the system. The reward for customers served is constant and customers arrive according to a Poisson process. The service times of the customers are exponentially distributed and are independent of the identity of the currently active server. [Lippman and Stidham 1977] studied a single queue model with Poisson arrivals and non-decreasing, concave service rate with respect to the number of customers in the system. The holding cost per unit of time for each customer is constant and the rewards for the customers entering the system are *i.i.d.* random variables with finite mean. The authors first considered the discounted net profit in the finite horizon case (in terms of the total number of admissions and service completions), and then extended the analysis to the non-discounted and infinite horizon case. [Johansen and Stidham 1980] studied the problem of finding the optimal admission policy of a system with general service and arrival processes. In the problem's setting, the net profit is discounted and the authors considered the finite horizon (in terms of the number of arriving customers) case. The rewards of the customers are *i.i.d.* random variables with finite mean and the non-negative waiting cost is a function of the number of customers in the system as well as the total number of past arrivals. All the works [Knudsen 1972; Lippman and Stidham 1977; Johansen and Stidham 1980] compared the optimal policy for the individual and social welfare maximization problems and showed that the optimal policies for both optimization problems are threshold policies that depend on the rewards of customers. Moreover, they also showed that the optimal threshold for the social welfare maximization problem is no greater than the individual maximization problem. Assuming a random arrival rate, [Y. Chen and Hasenbein 2020] showed that the optimal thresholds for the social welfare maximization problem are no larger than the individual maximization problem when either the queue length is observable or unobservable. They also showed that the optimal threshold for the revenue maximization problem may not coincide with the social welfare maximization problem when the queue is unobservable.

Learning unknown parameters to operate optimally in queuing systems, and analyzing queuing systems with model uncertainly have both been studied under various settings – see the tutorial [Walton and Xu 2021] for a recent overview. Our paper focuses on regret analysis in comparison with an optimal algorithm when the parameters are known. Under this framework, there is growing literature considering different models and various types of regret. [Adler, Moharrami, and V. Subramanian 2022] considered an Erlang-B blocking system with unknown arrival and service rates, where a customer is either blocked or receives service immediately. The authors proposed an algorithm that observes the system upon arrivals and converges to the optimal policy that either admits all customers when there is a free server, or blocks all customers. In our setting, the queue has infinite capacity, customers may wait in the queue, and the dispatcher observes the whole history of the queue-length when making a decision. The reward of admitting a customer in both our paper and [Adler, Moharrami, and V. Subramanian 2022] is only realized in the future as it involves knowledge of service times and (in our case also) waiting times, and the expected net profit requires knowledge of the arrival and service rates; this precludes the direct use of Reinforcement Learning based methods discussed in [Sutton and Barto 2018] and [Bertsekas 2019]. Stability is always assured in [Adler, Moharrami, and V. Subramanian 2022] since the maximum system occupancy is bounded (finite number of servers with no queuing). The queuing system is stable under any optimal policy for the problem we consider. However, under an arbitrary learning dispatcher, the supremum of the queue-lengths may be unbounded when the service rate is unknown. We will discuss the impact of this on our analysis in Section 2.2.3. [Krishnasamy, Arapostathis, et al. 2018] first considered a discrete-time single-server queuing system with multi-class customers and unknown service rates, and then modified and extended their algorithms to parallel multi-server queuing systems, again with multi-class customers. In the model customers of class i have (per unit-time) waiting cost c_i when waiting in the queue and Bernoulli services with the service success probability at server j being $\mu_{i,j}$ for class i (i.e., geometrically distributed service-times). They proposed a $c\mu$-rule-based algorithm that achieves constant regret compared to using the $c\mu$ rule with the true service rates. The $c\mu$ rule prioritizes the service of customers of type i at server j when $c_i\mu_{i,j}$ is higher. Optimality of the $c\mu$ rule has been proved in various settings, especially in the single server case; see [W. E. Smith 1956; Shwartz and Makowski 1986; Buyukkoc, Varaiya, and Walrand 1985] and [Cox and W. L. Smith 1961, Chapter 3]. [Zhong, Birge, and Ward 2022] considered the problem of learning the optimal static scheduling policy in a multi-class many-server queuing system with time-varying Poisson arrivals. Customers of type i have exponentially distributed patience with rate θ_i and exponentially distributed service requirements with rate μ_i. Unlike in [Krishnasamy, Arapostathis, et al. 2018], where stability is not guaranteed for arbitrary scheduling policies, the impatience of the customers helps to stabilize the queue without any extra requirements on the scheduling policy. The authors compared their Learn-Then-Schedule learning algorithm with the $c\mu/\theta$-rule and showed that their learning algorithm achieves a $\Theta(\log(T))$ regret where T is the (finite) time-horizon. For a discrete-time multi-class parallel-server system, when compared to the algorithm which matches a queue to a server for which the success service probability is the highest among all possible matches of this queue to any other server, [Krishnasamy, Sen, et al. 2021] used a multi-armed bandit viewpoint and proposed Q-UCB and Q-Thompson sampling algorithms that achieve $O(\mathrm{poly}(\log(T))/T)$ queue-regret as the time horizon T goes to infinity. [Stahlbuhk, Shrader, and Modiano 2021] focused on a single-server discrete-time queue, and showed the existence of queue-length-based policies that can achieve an $O(1)$ regret. When each server has its own queue, [Choudhury et al. 2021] studied the discrete-time routing problem when service rate and queue-length are not known. Taking a Markov Decision Process (MDP) viewpoint, [Agrawal and R. Jia 2022] considered a discrete-time inventory control problem where orders to be made arrive with delay and the decision-maker observes solely the sales and not the demands. Thereafter, a holding cost is collected for each unit of the good that is in storage. At each time step, the decision-maker needs to make new orders and aims to minimize the total expected holding cost. The authors studied the problem of learning the proper units of orders to be made at each time step when the distribution of the demand is unknown. The algorithm they proposed achieves an $O(\sqrt{T})$ regret (for horizon T) when compared to the best base-stock policy.

With the goal of stabilizing the queues and also minimizing penalties enforced in a discrete-time system, [Neely, Rager, and La Porta 2012] proposed an algorithm that learns a set of Max-Weight functionals that depend on the unknown underlying distribution, and make two-stage decisions (which are shown to correspond to scheduling choices in illustrated examples). The proposed algorithm stabilizes the system considered and achieves at most linear regret in the accumulated penalties when compared to the optimal controller. Considering a scheduling problem with unknown arrival and channel statistics, [Krishnasamy, Akhil, et al. 2018] considered the transformation scheduling problem with activation cost. Under their proposed explore-exploit policy with the exploration probability going to 0 slowly and together with a Max-Weight scheduling policy using learned statistics, the network is shown to be stable and the algorithm achieves at most linear regret in the accumulated switching and activating cost when comparing to the optimal scheduler with the knowledge of the model statistics. The error bound on the long-term average in both works can be made arbitrarily small (when compared to the optimal cost) by changing algorithm parameters. Instead of having explicit exploration, [Yang, Srikant, and Ying 2023] studied a discrete-time multi-server queuing system, and proposed a Max-Weight with discounted Upper Confidence Bound (UCB) scheduling algorithm. Their main result shows the stability of the queuing system under the proposed algorithm.

There is a growing literature that studies online dynamic pricing in service systems using queuing models. We discuss some relevant recent work next. The authors of [X. Chen, Liu, and Hong 2022] considered optimal pricing with congestion in a $GI/GI/1$ queue where there is unit cost depends on the service rate, the arrival rate depends on the service fee, and where customers experience congestion given by the average queue-length of the system. As the cost as a function of the service rate and the dependence of the arrival rate in chosen price is unknown, the authors proposed a gradient-based online learning algorithm that achieves a sub-linear regret when compared with the accumulated profit obtained with the optimal service rate and fee (using steady-state quantities). Also considering an online learning version of finding a proper price amongst a finite set of prices, [H. Jia, Shi, and Shen 2022] considered a multi-server queuing model with Poisson arrivals and exponential services where the dependence of arrival and service rates price chosen is unknown (with the values unknown as well but such that the load for each choice is strictly less than 1). Two online batch processing algorithms based on UCB and Thompson sampling are proposed in [H. Jia, Shi, and Shen 2022]. Both algorithms achieve sub-linear regret (optimal up to logarithmic factors) when compared with the accumulated profit achieved by the optimal price choice.

In our work, we consider a paradigm where there's uncertainty in the model parameters.

A different type of uncertainty, often called Knightian uncertainty, was studied in [Atar, Castiel, and Shadmi 2022], [Cohen 2019a], [Cohen 2019b], and [Cohen and Saha 2021] for multi-class queuing systems in the heavy traffic regime. In these models, the decision-maker is looking for robust control for a class of models. The uncertainty is modeled by including an adversarial player who chooses a worst-case scenario. Hence, the robust control problem is formulated via a stochastic game between the decision maker and the adverse player. Optimality is then characterized by studying Stackelberg equilibria.

Outline of paper: In Section 2.2 we introduce the model, propose our learning algorithm and state our main results. In Section 2.3, we state some preliminary results, including the properties of the coupling introduced in Section 2.2. Section 2.4 and 2.5 are devoted to the analysis of our learning algorithm and include the proof of our main results. Section 2.6 provides the finite-time performance of our algorithm via simulations. In section 2.7 we summarize our result.

2.2 The learning problem and the main results

In this section, we introduce the stochastic model and the learning algorithm. Specifically, in Section 2.2.1 we introduce the optimal admission control problem for the queuing system studied in [Naor 1969]. In this model, all the parameters are known. The same model but with unknown service and arrival rate is introduced in Section 2.2.2. We couple the models with known and unknown parameters so that we can characterize the regret of our learning dispatcher. Our learning algorithm is provided in Section 2.2.3. Finally, in Section 2.2.4 we state the main results.

2.2.1 The stochastic model with known parameters

[Naor 1969] studied the self-optimization and social welfare maximization problems for the following model. Homogeneous customers arrive at a single server queue according to a Poisson process with a rate $0 < \lambda < \infty$. When a customer arrives, and only then, the dispatcher decides whether to admit this customer to the queue or not. A customer that is not admitted (i.e., rejected) leaves and does not return. An admitted customer remains in the queue until being served. Upon service completion, the dispatcher receives a reward $R > 0$. Once the service is completed, the customer leaves the queue. The dispatcher suffers from a waiting/holding cost at the rate of $C > 0$ per time unit for each customer in the queue until service completion. The service requirements for the customers are *i.i.d.* EXP(μ) (i.e., exponentially distributed random variables with the rate $0 < \mu < \infty$). The dispatcher's goal is to maximize the social welfare, i.e., to maximize the long-term average profit accrued by serving customers – the ergodic-reward maximization problem. Let $Q(t)$ denote the queue-length of the system at time t, $N_A(t)$ denote the number of customers that arrived at the system until and including time t, then for an admission policy ρ the long-term average profit can be expressed as:

$$\liminf_{T\to\infty} \frac{1}{T} \left(\sum_{i=1}^{N_A(T)} R\mathbb{1}_{\{\text{Policy } \rho \text{ admits customer } i\}} - \int_0^T CQ(t)dt \right), \tag{II.1}$$

where throughout the paper, $\mathbb{1}_A$ is the indicator function of event A: namely, $\mathbb{1}_A = 1$ if A happens and 0 otherwise.

The optimal admission policy of the dispatcher in [Naor 1969] is a static *threshold policy*. That is, there is a threshold that depends on the parameters of the model, such that the dispatcher admits an arriving customer if and only if the queue-length upon arrival is strictly below this threshold. [Naor 1969] studied optimal admission control for the ergodic cost minimization problem by choosing the best threshold value among all possible thresholds. When the dispatcher uses a static threshold policy with a threshold K, the result is an $M/M/1/K$ queueing system. The queue-length process of such a system has a stationary distribution and is also ergodic. Note that the optimal threshold can then be determined by computing the expected reward using the stationary distribution of the $M/M/1/K$ queueing system for all possible values of K. Using this logic [Naor 1969] characterized the optimal threshold via the function $V : \mathbb{N} \times (0, \infty)^2 \to [0, \infty)$, given by:

$$V(K, y, z) = \begin{cases} \frac{K(y-z)-z(1-(z/y)^K)}{(y-z)^2}, & \text{if } y \neq z, \\ \frac{K(K+1)}{2y}, & \text{if } y = z. \end{cases} \tag{II.2}$$

The following proposition states a few properties of this function $V(\cdot, \cdot, \cdot)$.

Proposition 2.2.1. *The following hold:*

1. *For all fixed K, the function $V(K, \cdot, \cdot)$ is continuous in its domain.*

2. *For all fixed (y, z), $V(K, y, z)$ is strictly increasing in K.*

Proof. Note that when $K = 0$, $V(0, y, z) = 0$ for all $(y, z) \in (0, \infty)^2$. Consider any point $(K, y, z) \in \mathbb{N}^+ \times (0, \infty)^2$. In order to prove the continuity of V, it will be easier to rely on an alternative formulation of V based on the stationary distribution which we now provide. Let p_i^K denote the stationary probability of having the queue-length equal to i and let E_K denote the stationary expected queue-length when using the threshold policy with a threshold K. One can show that:

$$V(K, y, z) = \frac{E_{K-1} - E_K}{p_K^K - p_{K-1}^{K-1}} \frac{1}{z}, \qquad \text{where} \qquad p_i^K = \frac{(z/y)^i}{\sum_{i=0}^K (z/y)^i} \qquad \text{and} \qquad E_K = \sum_{i=0}^K i p_i^K.$$

Clearly, when $(y, z) \in (0, \infty)^2$, $1/z$, E_K, E_{K-1}, p_K^K and p_{K-1}^{K-1} are all continuous in (y, z). Moreover, $p_{K-1}^{K-1} \neq p_K^K$ for all $(y, z) \in (0, \infty)^2$.

Now, let us consider the function $V(K, y, z)$ for any fixed $(y, z) \in (0, \infty)^2$. To show the monotonic increasing property, we consider the function $f : [0, \infty) \to [0, \infty)$, $f(K) = V(K, y, z)$ by extending the definition of $V(\cdot, \cdot, \cdot)$ to real-valued K. From (II.2), it follows that when $y = z$, $f(K)$ is strictly increasing. Now, we focus on the case $y \neq z$. Computing the derivative of $f(K)$, we get:

$$f'(K) = \frac{(y - z) + z(z/y)^K \ln(z/y)}{(y - z)^2}.$$

Using the inequality $\ln(x) > 1 - 1/x$ for all $x > 0, x \neq 1$, we get:

$$(y - z) + z(z/y)^K \ln(z/y) > (y - z) + z(z/y)^K (1 - y/z) = (y - z)(1 - (z/y)^K) > 0,$$

for all $y \neq z$. This shows that $f(K)$ is strictly increasing, which implies that $V(K, y, z)$ is strictly increasing in K for all fixed $(y, z) \in (0, \infty)^2$. □

Using these properties [Naor 1969] showed that for every service rate μ and arrival rate λ the following inequalities for integer x

$$V(x, \mu, \lambda) \leq \frac{R}{C} < V(x + 1, \mu, \lambda) \tag{II.3}$$

have a unique solution $x = \bar{K}$, and this $\bar{K}$ is an optimal admittance threshold for the problem considered. Moreover, when $V\left(\bar{K}, \mu, \lambda\right) < R/C$, the optimal threshold is unique. However, when $V\left(\bar{K}, \mu, \lambda\right) = R/C$, both $\bar{K}$ and $\bar{K} - 1$ are optimal thresholds; hence, any policy that randomizes between the two thresholds at each arrival is also optimal[2].

Let $m := 1/\mu$ and $\nu := 1/\lambda$ denote the average service time and the average inter-arrival times respectively. Consider a pair of the true service and arrival rates (μ, λ) for which

[2] We discuss what we mean by “optimal" in Remark 2.5.1 after we specify the strategy to which we compare our learning algorithm in the case that there are multiple optimal thresholds.

there exists a unique optimal threshold and the corresponding $\bar{K}$ satisfying (II.3) with strict inequalities. Proposition 2.2.1 implies that there exist $\delta_1 > 0$ and $\delta_2 > 0$, both depending on μ and λ, such that for all pairs of points $(\hat{m}, \hat{\nu})$, where

$$m - \delta_1 < \hat{m} < m + \delta_1 \quad \text{and} \quad \nu - \delta_2 < \hat{\nu} < \nu + \delta_2, \tag{II.4}$$

we have:

$$V(\bar{K}, 1/\hat{m}, 1/\hat{\nu}) < \frac{R}{C} < V(\bar{K}+1, 1/\hat{m}, 1/\hat{\nu}). \tag{II.5}$$

That is, if one can estimate the average service time and the average inter-arrival time accurately so the inequality (II.4) is satisfied, one can obtain the corresponding $\bar{K}$ by solving (II.3) using $1/\hat{m}$ and $1/\hat{\nu}$ instead of μ and λ.

When equality holds in (II.3), for pairs of the true service and arrival rates (μ, λ) and the corresponding $\bar{K}$ that satisfies $V(\bar{K}, \mu, \lambda) = R/C$, there exist $\tilde{\delta}_1 > 0$ and $\tilde{\delta}_2 > 0$, both depending on μ and λ, such that for all pairs of points $(\hat{m}, \hat{v})$ where

$$m - \tilde{\delta}_1 < \hat{m} < m + \tilde{\delta}_1 \quad \text{and} \quad \nu - \tilde{\delta}_2 < \hat{\nu} < \nu + \tilde{\delta}_2, \tag{II.6}$$

we have:

$$V(\bar{K}-1, 1/\hat{m}, 1/\hat{\nu}) < \frac{R}{C} < V(\bar{K}+1, 1/\hat{m}, 1/\hat{\nu}). \tag{II.7}$$

That is, as long as the estimated average service time and average inter-arrival time are accurate enough to satisfy inequality (II.6), the integer solved from inequality (II.3) using $1/\hat{m}$ and $1/\hat{\nu}$ in place of μ and λ will be in the set of optimal thresholds, that is, $\{\bar{K}-1, \bar{K}\}$.

2.2.2 The learning system and the genie-aided system

We assume that the reward R and the cost per time unit C are known to the learning dispatcher, but neither the service rate μ nor the arrival rate λ. Consider again the potential application of job dispatch for online computing demands. When the computation clusters are provided by a third-party cloud computing platform, the dispatcher of the online computing jobs may not have knowledge about the configuration of the servers and their service rate. The dispatcher may also be unfamiliar with the customers who demand services, and therefore may only possess limited knowledge of the arrival rate. In our model, the dispatcher continuously observes the queue-length and past admission control decisions. Hence, we restrict the dispatcher to admission controls that at the time of a new arrival, admit or reject based on the entire history of the queue-length until the arrival time, and also the past admission control decisions. We call such controls *admissible*. Note that based on the FIFO serving discipline that's used, we can infer the time to enter service for all customers entering service by time t, and also the departure epochs for all the customers departing (after completing service) by t. Therefore, when a new customer arrives, the dispatcher can estimate the mean service time (also the service rate) using the service times of the customers that have departed before the new arrival, and use it for admission control. Further, knowledge of all past admission control decisions enables the dispatcher to obtain information on all past inter-arrival times, which will then be used to compute the statistics for the arrival process, i.e., the arrival rate.

We measure the performance of a policy chosen by the learning dispatcher by the regret it incurs in comparison to an optimal policy. Specifically, we use the difference between the expected net profit under the given learning-based control/policy and the best expected net profit the dispatcher could have obtained had it known the parameters μ and λ. To rigorously define the regret, we introduce some relevant processes for both the genie-aided and the learning systems.

We will use the marker $\bar{}$ to denote processes associated with the *genie-aided system* (dispatcher knows μ). The processes without a marker are associated with the *learning system* (dispatcher does not know μ). We let

- $\bar{Q}(t)$ and $Q(t)$ denote the queue-length at time t;
- $\bar{Q}_i$ and Q_i denote the queue-length right before the arrival of the i^{th} customer;
- $\bar{N}_A(t)$ and $N_A(t)$ denote the number of customers that have arrived at the system until and including time t;
- $\bar{N}_{\text{join}}(t)$ and $N_{\text{join}}(t)$ denote the number of customers that have joined the queue until and including time t;
- $\bar{T}_i^A$ and T_i^A denote the arrival time of the i^{th} customer to the system (i.e., $\bar{T}_i^A = \inf\{t : \bar{N}_A(t) \geq i\}$ and $T_i^A = \inf\{t : N_A(t) \geq i\}$, respectively);
- $\bar{K}_i$ and K_i denote the threshold policy used by the respective dispatchers at the arrival of the i^{th} customer.

2.2.2.1 A coupling between the two systems

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ rich enough to support two independent Poisson processes $(P(t))_{t\geq 0}$ and $(N_A(t))_{t\geq 0}$ with rates μ and λ, respectively. Set $\bar{N}_A = N_A$ so the arrival processes to both systems are the same. Let T_i^{PD} denote the i^{th} jump time of P. The service requirements of the customers that are being served at time t by all systems to be analyzed are determined as follows: the head of the line customer of each system (assuming not empty) completes her service at the time of the next jump of $P(t)$. Note that it may be the case that the services of the currently in-service customers are initiated at different times for the learning and genie-aided systems. Nevertheless, because the exponential distribution is memoryless, this does not change the distribution of the random process corresponding to the two systems, and in particular the distribution of the customer's service times. In other words, the time between the beginning of a service of a customer and the next jump of P is EXP(μ) distributed. Hence, we refer to $P(t)$ as the *potential departure process*, and to $\{T_i^{PD}\}_{i\geq 1}$ as the *potential departure times*, i.e., when there is a jump in P, and the queue-length is larger than 0, there will be a departure of a customer, but when the queue-length is 0, i.e., no customer is being served, this potential departure is *wasted*. Therefore, $\{P(T_i^A) - P(T_{i-1}^A)\}_{i\geq 1}$ is the number of potential services between two consecutive arrivals for both systems.

Now, we will use the underlying processes $\bar{N}_A = N_A$ and P to couple the queue-length processes of both systems assuming that a threshold policy is used in each system. Consider a sequence of random variables $\{K_i\}_{i\geq 0}$ taking vales in $\mathbb{N}$, such that each K_i is measurable with respect to the filtration generated by the queue-length until time T_i^A: since T_i^A is a stopping time for the filtration being used, we can define the σ-algebra $\mathcal{F}_{T_i^A} := \mathcal{F}_i$ (for short) using the original filtration $\mathcal{F}_T = \sigma(Q(t) : t \leq T)$ in the usual way (See [Durrett 2016]). We use $\{K_i\}_{i\geq 0}$ as a sequence of thresholds. Similarly, we use $\{\bar{K}_i\}_{i\geq 0}$ to denote the sequence of thresholds used by the genie-aided dispatcher. We refer to any such $\{K_i\}_{i\geq 0}$ as a threshold policy. For the coupled genie-aided and learning systems, we have the following: for any $i \geq 1$,

$$\begin{aligned} Q_i &= \left(Q_{i-1} + \mathbb{1}_{\{Q_{i-1}<K_{i-1}\}} - (P(T_i^A) - P(T_{i-1}^A))\right)^+, \\ \text{and } \bar{Q}_i &= \left(\bar{Q}_{i-1} + \mathbb{1}_{\{\bar{Q}_{i-1}<\bar{K}_{i-1}\}} - (P(T_i^A) - P(T_{i-1}^A))\right)^+, \end{aligned}$$

where for $x \in \mathbb{R}$, $(x)^+ := \max(x, 0)$. Similarly, we have:

$$Q(t) = \left(Q_n + \mathbb{1}_{\{Q_n<K_n\}} - (P(t) - P(T_n^A))\right)^+, \tag{II.8}$$

$$\text{and } \bar{Q}(t) = \left(\bar{Q}_n + \mathbb{1}_{\{\bar{Q}_n<\bar{K}_n\}} - (P(t) - P(T_n^A))\right)^+, \tag{II.9}$$

where $n := \max\{m : T_m^A < t\}$. Once the initial queue-lengths Q_0 and $\bar{Q}_0$ are specified in $\mathbb{Z}_+$, by induction one can show that the processes $\{Q_i\}_{i\geq 0}$ and $\{\bar{Q}_i\}_{i\geq 0}$ are well-defined, and using these $\{Q_t\}_{t\geq 0}$ and $\{\bar{Q}_t\}_{t\geq 0}$ are also well-defined.

2.2.2.2 The regret.

Let $\mathbb{E}[\cdot]$ be expectation associated with $(\Omega, \mathcal{F}, \mathbb{P})$. Then, the regret is given by

$$G(t) := \mathbb{E}\left[R\,\bar{N}_{\text{join}}(t) - C\int_0^t \bar{Q}(u)du - \left(R\,N_{\text{join}}(t) - C\int_0^t Q(u)du\right)\right].$$

This definition of the regret compares the net reward processes of the learning and genie-aided systems: if the learning-based admission control algorithm achieves the same long-term average profit, then this will allow us to estimate the sub-linear offset. The genie-aided dispatcher uses a static threshold policy that maximizes the long-term average profit described in (II.1). Note that when equality does not hold in (II.3), the genie-aided policy is unique so there is no ambiguity in the definition of the regret. In this case, $\bar{K}_i \equiv \bar{K}$, where $\bar{K}$ uniquely satisfies inequality (II.3). However, when equality holds in (II.3), the genie-aided policy is not unique. We will compare our learning algorithm with a particular optimal genie-aided system that will be specified in Section 2.5.

Consider a threshold policy for the learning system, $\{K_i\}_{i\geq 0}$, and a threshold policy for the genie-aided system, $\{\bar{K}_i\}_{i\geq 0}$, the regret can be estimated as:

$$\begin{aligned} G(t) &= \mathbb{E}\left[R\sum_{i=1}^{N_A(t)}\left(\mathbb{1}_{\{\bar{Q}_i<\bar{K}_i\}} - \mathbb{1}_{\{Q_i<K_i\}}\right)\right] - \mathbb{E}\left[C\int_0^t \Big(\bar{Q}(u) - Q(u)\Big)du\right] \\ &\leq \mathbb{E}\left[R\sum_{i=1}^{N_A(t)}\left|\mathbb{1}_{\{\bar{Q}_i<\bar{K}_i\}} - \mathbb{1}_{\{Q_i<K_i\}}\right|\right] + \mathbb{E}\left[C\int_0^t \Big|\bar{Q}(u) - Q(u)\Big|du\right]. \qquad \text{(II.10)} \end{aligned}$$

From (II.8) and (II.9), we note that

$$\left|\bar{Q}(t) - Q(t)\right| \le \left|\bar{Q}_n + \mathbb{1}_{\{\bar{Q}_n < \bar{K}_n\}} - \left(Q_n + \mathbb{1}_{\{Q_n < K_n\}}\right)\right|.$$

This expression helps us to get an upper bound for the integral $\int_0^t |\bar{Q}(u) - Q(u)| du$ in (II.10) as follows:

$$\int\limits_0^t \left|\bar{Q}(u) - Q(u)\right| du \le \sum_{i=0}^{N_A(t)} \left(T_{i+1}^A - T_i^A\right)\left(|\bar{Q}_i - Q_i| + |\mathbb{1}_{\{\bar{Q}_i < \bar{K}_i\}} - \mathbb{1}_{\{Q_i < K_i\}}|\right).$$

Substituting the above bound in (II.10), we get:

$$\begin{aligned} G(t) \le \mathbb{E}\left[R \sum_{i=1}^{N_A(t)} \left|\mathbb{1}_{\{\bar{Q}_i < \bar{K}_i\}} - \mathbb{1}_{\{Q_i < K_i\}}\right|\right] + \mathbb{E}\left[C \sum_{i=0}^{N_A(t)} (T_{i+1}^A - T_i^A)\left|\bar{Q}_i - Q_i\right|\right] \\ + \mathbb{E}\left[C \sum_{i=0}^{N_A(t)} \left(T_{i+1}^A - T_i^A\right)\left|\mathbb{1}_{\{\bar{Q}_i < \bar{K}_i\}} - \mathbb{1}_{\{Q_i < K_i\}}\right|\right]. \qquad \text{(II.11)} \end{aligned}$$

Note that the (future) inter-arrival time $T_{i+1}^A - T_i^A$ is independent of the queue-length of the learning and genie-aided systems Q_i and $\bar{Q}_i$, respectively, as well as the threshold used at the arrival of the i^{th} customer K_i and $\bar{K}_i$. In particular, $T_{i+1}^A - T_i^A$ is independent of $|\bar{Q}_i - Q_i|$ and $|\mathbb{1}_{\{\bar{Q}_i < \bar{K}_i\}} - \mathbb{1}_{\{Q_i < K_i\}}|$. Then as the increments of the Poisson process are independent, we have:

$$\begin{aligned} \mathbb{E}\left[C \sum_{i=0}^{N_A(t)} (T_{i+1}^A - T_i^A)\left|\bar{Q}_i - Q_i\right|\right] &= \mathbb{E}\left[C \sum_{i=0}^{\infty} (T_{i+1}^A - T_i^A)\left|\bar{Q}_i - Q_i\right| \mathbb{1}_{\{T_i^A \le t\}}\right] \\ &= C \sum_{i=0}^{\infty} \mathbb{E}\left[(T_{i+1}^A - T_i^A)\left|\bar{Q}_i - Q_i\right| \mathbb{1}_{\{T_i^A \le t\}}\right] \quad \text{(MCT)} \\ &= C \sum_{i=0}^{\infty} \mathbb{E}\left[\frac{1}{\lambda}\left|\bar{Q}_i - Q_i\right| \mathbb{1}_{\{T_i^A \le t\}}\right] \quad \text{(By independence)} \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\frac{C}{\lambda}\sum_{i=0}^{\infty}\left|\bar{Q}_i - Q_i\right|\mathbb{1}_{\{T_i^A \le t\}}\right] \quad \text{(MCT)}\\
&= \frac{C}{\lambda}\mathbb{E}\left[\sum_{i=0}^{N_A(t)}\left|\bar{Q}_i - Q_i\right|\right],
\end{aligned}
$$

where MCT stands for the Monotone Convergence Theorem. Similarly, we can also simplify $\mathbb{E}\left[C\sum_{i=0}^{N_A(t)}(T_{i+1}^A - T_i^A)\left|\mathbb{1}_{\{\bar{Q}_i<\bar{K}_i\}} - \mathbb{1}_{\{Q_i<K_i\}}\right|\right]$ to get

$$
\begin{aligned}
G(t) &\le \mathbb{E}\left[\left(R+\frac{C}{\lambda}\right)\sum_{i=1}^{N_A(t)}\left|\mathbb{1}_{\{\bar{Q}_i<\bar{K}_i\}} - \mathbb{1}_{\{Q_i<K_i\}}\right|\right] + \mathbb{E}\left[\frac{C}{\lambda}\sum_{i=0}^{N_A(t)}\left|\bar{Q}_i - Q_i\right|\right]\\
&\le \left(R+\frac{C}{\lambda}\right)\mathbb{E}\left[\sum_{i=1}^{N_A(t)}\left|\mathbb{1}_{\{\bar{Q}_i<\bar{K}_i\}} - \mathbb{1}_{\{Q_i<K_i\}}\right| + \left|\bar{Q}_i - Q_i\right|\right]. \qquad \text{(II.12)}
\end{aligned}
$$

Following this bound, from now on, we analyze the systems at the arrival epochs $\{T_i^A\}_{i\ge 1}$.

With the shift to analyzing the systems at arrival epochs, we will characterize the regret in terms of the total number of arrivals N. We use $\tilde{G}(N) := G(T_N^A)$ to denote the total regret accumulated up to the arrival of the N^{th} customer. Recall that $m = 1/\mu$ denote the average service time and $\nu = 1/\lambda$ denote the inter-arrival time. We assume that $0 < m < \infty$ and $0 < \nu < \infty$: we allow for the average service time to be large, and it is possible to have $\bar{K} = 0$ where the optimal policy for the genie-aided system is to reject any arriving customer. Note that when $\bar{K} = 0$, equality in (II.3) is not possible for $R, C > 0$, therefore the optimal policy is unique, and $\bar{K}_i = \bar{K} = 0$ for all $i \ge 0$. If the genie-aided dispatcher always admits customers when the queue is empty and the learning dispatcher knows this, then the algorithm design would be simpler: there is no need to balance exploration and exploitation explicitly. With this knowledge, a learning dispatcher can achieve constant regret using a policy that always accepts customers when the queue is empty and uses a threshold computed by solving the inequalities (II.3) using the empirical service rate otherwise. The conflicting requirements for a learning algorithm in the two different regimes – $\bar{K} = 0$ (stop admitting customers soon) versus $\bar{K} > 0$ (admit customers infinitely often but at the correct rate via the right choice of the threshold) – are critical to the difficulty of our problem and its analysis.

2.2.3 The learning algorithm

We propose (and study) Algorithm 1 for learning-based social-welfare maximizing dispatch that consists of a sequence of *batches*, where each batch has two phases: phase 1 for exploration and phase 2 for exploitation. For customer i who arrives during phase 1 (assuming that a phase 1 is used), we can assume that $K_i = \infty$ as this customer is admitted in the queue no matter the queue-length at this arrival. However, in our algorithm, we will fix any exploration phase (if used) for all batches to last for exactly l_1 arrivals, and so, the threshold K_i is effectively $K_i = l_1$ for all arrivals in any phase 1. At the beginning of phase 2 of the j^{th} batch $K(j)$ is computed by finding the minimum between $K^*(j)$ and the integer that solves inequalities $V(x, 1/\hat{m}, 1/\hat{\nu}) \leq R/C < V(x+1, 1/\hat{m}, 1/\hat{\nu})$. The computed $K(j)$ will be used for the entire exploitation phase of batch j. That is, for customers i_1 and i_2 who arrive during phase 2 of the j^{th} batch, $K_{i_1} = K_{i_2} = K(j)$ and these customers are admitted to the queue when the queue-length seen at their arrival is strictly less than $K(j)$. For technical reasons, we will insist that at the termination of phase 2, the queue is empty. As the batch number increases, our algorithm will extend the length of the exploitation phase and reduce the occurrences of the exploration phases.

Here is some notation that we use in the algorithm:

- l_1: A positive integer representing the length of phase 1, $l_1 > 1$;
- l_2: A positive integer representing the initial minimum length of phase 2, $l_2 \geq l_1$;
- i: A positive integer which is the index of the arriving customer from the very beginning. It is used to update the belief of the average arrival rate;

Algorithm 1: Learning-based customer dispatch, with unknown service and arrival rate.

```
$i = 0$; $j = 0$; $\alpha_j$ grows at polynomial rate in $j$ ; $s = 0$;
  $K^*(j) = \max\{\lfloor \ln(j) \rfloor, 0\} + l_1 + Q_0$.
while $i \leq N$ do
    $j = j + 1$;
    % If the phase $1$ of the $j^{th}$ batch happens, it sees $l_1$ customers.
    if $j == 1$ or $(K(j-1) == 0$ and $B^j == 1)$ then
        for the next $l_1$ customers do
            $i = i + 1$;
            % we update the belief of the average arrival time when there is a
              new arrival.
            $\hat{\nu} = \hat{\nu} + \frac{\text{inter-arrival time observed } - \hat{\nu}}{i}$;
            Exploration phase: customers always join the queue, $K_i = l_1$.
        end
        if there are $\mathrm{S}_{\text{cnt}} > 0$ new services completed during this phase 1 then
            for $\text{cnt} = 1$ to $\mathrm{S}_{\text{cnt}}$ do
                $s = s + 1$;
                $\hat{m} = \hat{m} + \frac{\text{service time of the s}^{\text{th}} \text{ customer that completed service} - \hat{m}}{s}$;
            end
        end
    end
    Compute integer $K$, which satisfies
      $V(K, 1/\hat{m}, 1/\hat{\nu}) \leq R/C < V(K+1, 1/\hat{m}, 1/\hat{\nu})$;
    Set $K(j) = \min\{K^*(j), K\}$;
    count = 0 ;
    % The phase $2$ of the $j^{th}$ batch sees at least $\alpha_j l_2$ customers. The
      queue-length is $0$ when phase $2$ ends.
    while $count < \alpha_j l_2$ or $Q_i > 0$ do
        count = count +1;
        $i = i + 1$;
        $\hat{\nu} = \hat{\nu} + \frac{\text{inter-arrival time observed } - \hat{\nu}}{i}$;
        Customers join the queue if and only if the queue-length is smaller than $K(j)$,
          and so $K_i = K(j)$.
    end
    if there are $\mathrm{S}_{\text{cnt}} > 0$ new services completed during this phase 2 then
        for $\text{cnt} = 1$ to $\mathrm{S}_{\text{cnt}}$ do
            $s = s + 1$;
            $\hat{m} = \hat{m} + \frac{\text{service time of the s}^{\text{th}} \text{ customer that completed service} - \hat{m}}{s}$;
        end
    end
end
```

- j: A positive integer that indices the batch number;
- $\alpha_j \geq 1$: Growth factor for the length of phase 2 in the j^{th} batch which ensures that the phase 2 duration lasts for at least $\lceil \alpha_j l_2 \rceil$ arrivals;
- B^j: A Bernoulli random variable that is independent of everything else, where $\mathbb{P}\left[B^j = 1\right] = 1$ for $j = 1$, and $\mathbb{P}\left[B^j = 1\right] = \ln^{\epsilon}(j)/j$ for $j > 1$ and fixed $\epsilon > 0$. If the threshold used in the previous batch (the $(j-1)^{\text{th}}$ batch) is 0, the random variable B^j will be used to determine if phase 1 will happen;
- $K(j)$: the threshold used by the learning dispatcher during phase 2 of the j^{th} batch;
- $K^*(j)$: the upper bound of the threshold used by the learning algorithm. This parameter slowly increases to infinity, and is chosen to be larger than the initial queue-length, Q_0, and the length of phase 1, i.e., l_1;
- S_{cnt}: A counter which counts for the number of completed services in each phase. This counter is used to update the belief of the average service rate after each phase.

Note that Algorithm 1 enforces an exploration phase only for the first batch, and then utilizes one in a probabilistic manner when the learned threshold in the previous batch is 0. When the genie-aided system uses a non-zero threshold, as the number of services experienced by the customers admitted by the dispatcher increases, the threshold learned by the algorithm will quickly become non-zero for phase 2. In this scenario, the exploration phase can potentially be eschewed, and, in fact, should be used more and more infrequently as time progresses so that the regret is not large. In fact, in our algorithm we completely eliminate a phase 1 for a batch if in the previous batch the threshold of its phase 2 is positive: some customers will be admitted in a phase 2 with a positive threshold so new service time estimates will obtain, and on the contrary, a phase 2 with a 0 threshold will not admit any customers. However, allowing for an exploration phase is necessary. When the genie-aided system uses a non-zero threshold, it is possible that the learning system sees the first few service times being long enough so that the learned threshold is 0. Then, without the exploration phase, the learning system will stop admitting any customers to the queue, and therefore, will not get any more samples to update its false belief. Although this is a low-probability event, the probability of this happening is non-negligible for any fixed length l_1 of the exploration.

The frequency of the exploration phase in our algorithm is controlled by the distribution of B^j. Our theoretical regret analysis uses $\mathbb{P}\left[B^j = 1\right] = \ln(j)/j$. When the genie-aided system uses the threshold 0, the exploration phase should not happen too often. This is because every time the learning system admits a customer into the queue, the regret increases. Hence, this regime demands that phase 1 be eschewed as fast as possible. However, as the algorithm is unaware of the parameter regime (even whether the optimal threshold is zero or non-zero), we necessarily need enough phase 1s when the threshold from the previous batch is 0. Hence, to combat the regret accumulation from phase 1s when the optimal policy is not to admit any arrivals, we increase the length of phase 2 (the exploitation phase) as the batch count increases. The control of the length of phase 2 of the j^{th} batch is achieved using parameter α_j: phase 2 of the j^{th} batch will last for at least $\lceil \alpha_j l_2 \rceil$ arrivals. Whereas we do require that α_j grows to infinity, we do not want it to grow too fast as this could lead to poor performance: when the thresholds used by the learning and genie-aided systems do not match in a batch, there may be too much regret accumulated during that batch if there is a large value of α_j for small j (when the probability of an error is higher).

Note that $K^*(j) = \max\{\lfloor \ln(j) \rfloor, 0\} + l_1 + Q_0$ is a deterministic function, with $K^*(j)$ no smaller than l_1 and the initial queue-length of the learning system Q_0 (when Q_0 is chosen in a deterministic manner). We also note that $\lim_{j\to\infty} K^*(j) = \infty$. This ensures that as the number of batches increases, eventually, the (true) optimal thresholds will be smaller than this upper bound. Note that for all $j \geq \lceil e^{\bar{K}} \rceil$ batches, $K^*(j) \geq \bar{K}$. Therefore, if the estimations on the service and arrival rates are accurate during batch j for $j \geq \lceil e^{\bar{K}} \rceil$, then the learning dispatcher will be using $\bar{K}$ during phase 2. Although $\lceil e^{\bar{K}} \rceil$ can be a large number, it is a fixed constant (fixing μ and λ), and the total expected regret accumulated during the first $\lfloor e^{\bar{K}} \rfloor$ batches will also be a constant (see Remark 2.4.1). Therefore, in our analysis we focus our analysis on the regret accumulated when $j \geq \lceil e^{\bar{K}} \rceil$.

2.2.4 Main results: Regret bounds for Algorithm 1

Theorem 2.2.1. *Assume that the initial queue-length for the learning and genie-aided systems are the same, and* 0 *is not in the set of optimal thresholds used by the genie-aided system. Then, Algorithm 1 achieves* $O(1)$ *regret as* $N \to \infty$*, where* N *is the total number of arrivals.*

Theorem 2.2.2. *Assume that the initial queue-length for the learning and genie-aided systems are the same, and* 0 *is in the set of optimal thresholds used by the genie-aided system. Then, Algorithm 1 achieves* $O(\ln^{1+\epsilon}(N))$ *regret for any specified* $\epsilon > 0$ *as* $N \to \infty$*, where* N *is the total number of arriving customers.*

When the learning and genie-aided systems have different initial queue-lengths, as stated in Remark 2.3.1 below, the regret characterization still holds. This is done by introducing another genie-aided system that has the same initial queue-length as the learning system. Thereafter, we will use Proposition 2.3.1 (discussed in the following section), which shows that if two coupled systems use the same threshold policy, then the ordering of their queue-lengths is preserved. We end this section by pointing out that the regret characterization in Theorem 2.2.2 can be changed to $O(\log^{1+\epsilon}(N))$ for all $\epsilon > 0$ as $N \to \infty$; see the discussion in Remark 2.4.3.

2.3 Preliminary results

We will use a few coupled systems to prove the main results. Besides the coupling between the learning and the genie-aided systems mentioned before, we will also compare the queue-length process of the learning system with systems using the same threshold policy but with different initial queue-lengths. The following results are proved for systems coupled by having the same arrival process and with the service time of the customers in the queue of both systems begin determined by the same Poisson process from $t = 0$.

The next proposition states that the order of the queue-lengths of two coupled systems is preserved over time if their threshold policies satisfy certain conditions. This is a core preliminary result that is used in different ways, and helps us establish our main results in considerable generality. Consider two systems G and L coupled through process $\{N_A(t)\}_{t\geq 0}$ and $\{P(t)\}_{t\geq 0}$ as described in Section 2.2.2.1, but with possibly different initial queue-lengths and (threshold) admission policies. Let $Q^G(t)$ and $Q^L(t)$ denote the queue-length at time t of the two systems, respectively. Let $\{K_i^G\}_{i\geq 0}$ and $\{K_i^L\}_{i\geq 0}$ denote the threshold policies of the two systems, respectively.

Proposition 2.3.1.

1. *If the dispatchers for the two coupled systems G and L use the same threshold admission policy for all arrivals, i.e., $K_i^G = K_i^L$ for all i, then with probability 1, the order of their queue-lengths is preserved for all time, that is,*

$$Q^G(0) \geq Q^L(0) \implies Q^G(t) \geq Q^L(t), \qquad \forall t \geq 0. \tag{II.13}$$

2. *Assume that both systems have the same initial queue-length $q := Q^G(0) = Q^L(0)$. Let $D^G(t)$ and $D^L(t)$ denote the number of departures up to time t for the systems G and L, respectively. If $K_i^G \geq K_i^L$ for all i, then with probability 1,*

$$Q^G(t) \geq Q^L(t) \textit{ and } D^G(t) \geq D^L(t), \qquad \forall t \geq 0. \tag{II.14}$$

Moreover, every customer that joins the queue in the system L necessarily joins the queue *in the system G when static thresholds $K^G \geq K^L$ are used in the two systems, respectively, and $q \leq K^L$.*

Before proving the proposition, we state a useful corollary.

Corollary 2.3.1. *Assume that phase 1 of the j^{th} batch did not happen and the queue-length processes of the learning and genie-aided systems are coupled. If the two systems use the same threshold during the phase 2 of the j^{th} batch and if the queue-length of the genie-aided system hits 0 during this phase 2, then the queue-lengths of both systems are 0 at the end of this phase 2.*

Proof of Corollary 2.3.1. Recall that under the proposed algorithm, the queue-length of the learning system is 0 at the end of each phase 2. Hence, the result follows immediately by Proposition 2.3.1. □

Proof of Proposition 2.3.1. Let us start by proving the first part of Proposition 2.3.1. Since the queue-length process is a jump process, it is sufficient to show that after each jump, the queue-lengths of the two systems satisfy (II.13). Note that the set of potential jump times is the union of the arrival times (jumps times in the arrival process) and the jump times in the Poisson process that determines the service process. Let $\{t_l\}_{l\geq 0} = \{T_i^A\}_{i\geq 0} \cup \{T_i^{PD}\}_{i\geq 0}$ denote the ordered countable set of potential jump times of the queue-length process, where $t_{l-1} < t_l$. By the superposition property of independent Poisson processes, with probability 1, $\{T_i^A\}_i \cap \{T_i^{PD}\}_i = \emptyset$, so that at any time instant t_l, either there is an arrival, or there is a potential departure. Let Q_l^G and Q_l^L denote the queue-lengths immediately before the l^{th} potential jump of the system G and L, respectively. Also, let Q_0^G and Q_0^L, respectively, denote the initial queue-length of the two systems.

The proof follows by induction. Fix $n > 0$ and assume $Q_l^G \geq Q_l^L$ holds for all $l \leq n$. Immediately after time t_n, one of the following can happen:

- If $Q_n^G = Q_n^L$: In case the jump at time t_n is due to a service completion or a service wasted, $Q_{n+1}^G = Q_{n+1}^L$. If the jump is due to a new arriving customer, the dispatcher will make the same choice in both systems, and $Q_{n+1}^G = Q_{n+1}^L$ holds.

- If $Q_n^G > Q_n^L \geq 0$: In case the jump at time t_n is due to a service completion or a service wasted, $Q_{n+1}^G \geq Q_{n+1}^L$. Otherwise, the jump is due to an arriving customer. We have $Q_{n+1}^G \geq Q_n^G \geq Q_n^L + 1 \geq Q_{n+1}^L$.

Now, let us consider the second part of Proposition 2.3.1. First, we show that $Q^G(t) \geq Q^L(t)$ holds for all t. Again, it is sufficient to show $Q_l^G \geq Q_l^L$ for every $l > 0$, the proof of which follows by induction. Fix $n > 0$ and assume that $Q_l^G \geq Q_l^L$ for all $l \leq n$. Immediately after t_n, one of the following can happen:

- If $Q_n^G = Q_n^L$: In case the jump at time t_n is due to a service completion or a service wasted, then $Q_{n+1}^G = Q_{n+1}^L$. Otherwise, the jump is due to an arriving customer. Since $K_i^G \geq K_i^L$ for all i, this customer is admitted in system L only if also admitted in system G and we have $Q_{n+1}^G \geq Q_{n+1}^L$.

- If $Q_n^G > Q_n^L \geq 0$: As before, either both processes jump in the same direction at time t_n or only one of them jumps (which would be the L system). In either case, $Q_{n+1}^G \geq Q_{n+1}^L$.

Since $Q^G(t) \geq Q^L(t)$ holds for all t it follows that whenever there is a service completion in system L then there is one also in G. Therefore, $D^G(t) \geq D^L(t)$.

Now assume that the static thresholds K^G and K^L are used in the systems G and L, respectively. To show that every customer who joins the queue in system L also joins the queue in system G, we will show first that $Q^G(t) - Q^L(t) \leq K^G - K^L$. Fix a $n > 0$ and assume that $Q_l^G - Q_l^L \leq K^G - K^L$ holds for all $l \leq n$. One of the following can happen immediately after time t_n:

- If $Q_n^G - Q_n^L = K^G - K^L$: Under this case, either we have $\{Q_n^G = K^G, Q_n^L = K^L\}$, or $\{Q_n^L \leq Q_n^G < K^G, Q_n^L < K^L\}$. Then, only when $Q_n^G = K^G - K^L$, $Q_n^L = 0$, and the jump is due to a service completion or service being wasted, the queue-length processes of the two systems evolve differently: system G has a service completion but not L. However, $Q_{n+1}^G - Q_{n+1}^L \leq K^G - K^L$ still holds.

- If $Q_n^G - Q_n^L < K^G - K^L$: Either we have $\{Q_n^L \leq Q_n^G < K^G, Q_n^L = K^L\}$, or $\{Q_n^L \leq Q_n^G < K^G, Q_n^L < K^L\}$. When $\{Q_n^L \leq Q_n^G < K^G, Q_n^L = K^L\}$, if the jump is due to an arriving customer, the dispatcher in the system G will assign this customer to the queue but not the dispatcher in the system L. Otherwise, both systems have a service completion. Then, $Q_{n+1}^G - Q_{n+1}^L \leq K^G - K^L$ holds in either case. When $\{Q_n^L \leq Q_n^G < K^G, Q_n^L < K^L\}$, if the jump is due to a new arriving customer, the dispatchers in both systems admit the customers to the queue. Otherwise, the jump is due to a service completion or service being wasted, where it is possible that only in system G there is a service completion. Again, $Q_{n+1}^G - Q_{n+1}^L \leq K^G - K^L$ holds in either cases.

At the time T_l^A, which corresponds to the arrival of the l^{th} customer, assume that this customer is admitted to the queue in the system L but not in G. We must have $Q_l^L < K^L$ and $Q_l^G = K^G$, i.e., $Q_l^G - Q_l^L > K^G - K^L$. This is a contradiction. Therefore, for any arriving customer, either the dispatchers in both systems G and L make the same admission decision, or only the dispatcher in the system G admits this customer. As a result, any customer who joins the queue in the system L necessarily joins the queue in the system G. □

Remark 2.3.1. *In case the genie-aided system and the learning system have different initial queue-lengths, we can introduce a second genie-aided system that has the same initial queue-length as the learning system and is also coupled with the two systems using the procedure from Section 2.2.2.1. Let Q_i' denote the queue-length of this new system right before the i^{th} arrival customer, $G'(N)$ denote the regret of the learning algorithm with respect to the second* *genie-aided system. Using the triangle inequality and equation* (II.12)*, we get:*

$$\tilde{G}(N) \leq \left(R + \frac{C}{\lambda}\right)\mathbb{E}\left[\sum_{i=1}^{N}\left|\mathbb{1}_{\{\bar{Q}_i<\bar{K}_i\}} - \mathbb{1}_{\{Q'_i<\bar{K}_i\}}\right| + \left|\bar{Q}_i - Q'_i\right|\right] + G'(N).$$

Theorems 2.2.1 and 2.2.2 provide regret bounds for $G'(N)$. By Proposition 2.3.1, the orders of Q'_i and $\bar{Q}_i$ are preserved, thus after both queue-length processes hit 0, Q'_i and $\bar{Q}_i$ will evolve together. Since the expected time of both queue-length processes to hit 0 *simultaneously is finite, the regret characterization in Theorems 2.2.1 and 2.2.2 still holds.*

2.4 Unique admittance threshold case

In this section, we analyze the case where (II.3) holds with strict inequality. In this case, the genie-aided dispatcher uses a unique optimal threshold $\bar{K}$, and the resulting queue-length process has a stationary distribution.

In section 2.4.1, we start by providing an estimate for the number of samples of completed service times that the learning algorithm uses in order to estimate the average service time, and then to update the threshold policy for each phase 2: see Proposition 2.4.1. We use it to estimate the probability that the learning system can obtain an accurate estimate of the average service time: see Proposition 2.4.2. Combining the above estimate with the probability that the learning system can obtain an accurate estimation on the arrival rate, see Proposition 2.4.4, we can bound the probability of the learning system using the same threshold as the genie-aided system; see Corollary 2.4.1. In section 2.4.2, we estimate the regret of the learning algorithm because of having phase 1 (if used) and using incorrect thresholds in phase 2 separately. Proposition 2.4.6 we consider “bad" events where there will be regret accumulated during phase 2 because of using the wrong threshold. In addition, we will use an upper bound on the difference between the queue-length processes of the learning and genie-aided system to bound the regret accumulated because of the existence of phase 1 (if used) in Lemma 2.4.1

and because of using the wrong threshold during phase 2 in Lemma 2.4.2. The proof of Theorem 2.2.1 and 2.2.2 are stated in section 2.4.3 and 2.4.4 respectively.

2.4.1 Sample estimation

First, we state and prove some results on the number of samples the learning dispatcher gets on the inter-arrival times and completed service times, and the resulting implications on the estimates of the arrival and service rates.

In the following proposition, we show that with high probability, the number of samples of completed service times that the learning algorithm can observe is sufficiently large at the beginning of the phase 2 of the j^{th} batch. For this, we use the fact that (by design) each phase 2 is longer than phase 1.

Proposition 2.4.1. *Let D_j denote the number of observed service times up to the beginning of phase 2 of the j^{th} batch. Then,*

$$\mathbb{P}\left[D_j \leq \frac{l_1 \ln^{1+\epsilon}(j)\mu}{4(1+\epsilon)(\lambda+\mu)}\right] \leq \exp\left(-\frac{l_1 \ln^{1+\epsilon}(j)\mu}{16(1+\epsilon)(\lambda+\mu)}\right) + \exp\left(-\frac{C_0(\epsilon)}{8} - \frac{\ln^{1+\epsilon}(j)}{8(1+\epsilon)}\right),$$

where $C_0(\epsilon) := 1 + \sum_{i=2}^{\lfloor e^\epsilon \rfloor} \frac{\ln^\epsilon(i)}{i} - \frac{\ln^{1+\epsilon}(\lceil e^\epsilon \rceil)}{1+\epsilon}$ is a constant depending on the choice of ϵ.

Proof. Consider the epoch which is the beginning of phase 2 of the j^{th} batch. Let $\hat{X}^j$ denote the total number of arrivals that the learning dispatcher sees during the past batches and the potential phase 1 of the j^{th} batch . Note that $\hat{X}^j$ counts for the arrivals in phase 1's (when they occur), and all past phase 2's using a threshold ≥ 1.

The following inequality holds when $\alpha_j l_2 \geq l_1$ for all j:

$$\hat{X}^j \geq l_1 + \sum_{i=1}^{j-1} \left(\mathbb{1}_{\{K(i)>0\}}\alpha_i l_2 + \mathbb{1}_{\{K(i)=0\}} B^{i+1} l_1\right) \geq l_1 \sum_{i=1}^{j} B^i.$$

Observing the function $\ln^\epsilon(x)/x$ is decreasing when $x \geq e^\epsilon$, when $j \geq \lceil e^\epsilon \rceil$, we have:

$$\frac{\ln^{1+\epsilon}(j)}{1+\epsilon} - \frac{\ln^{1+\epsilon}(\lceil e^{\epsilon} \rceil)}{1+\epsilon} = \int_{\lceil e^{\epsilon} \rceil}^{j} \frac{\ln^{\epsilon}(x)}{x} dx \leq \sum_{i=\lceil e^{\epsilon} \rceil}^{j} \frac{\ln^{\epsilon}(i)}{i},$$

$$\sum_{i=\lceil e^{\epsilon} \rceil}^{j} \frac{\ln^{\epsilon}(i)}{i} \leq \frac{\ln^{\epsilon}(\lceil e^{\epsilon} \rceil)}{\lceil e^{\epsilon} \rceil} + \int_{e^{\epsilon}}^{j} \frac{\ln^{\epsilon}(x)}{x} dx = \frac{\ln^{\epsilon}(\lceil e^{\epsilon} \rceil)}{\lceil e^{\epsilon} \rceil} + \frac{\ln^{1+\epsilon}(j)}{1+\epsilon} - \frac{\ln^{1+\epsilon}(e^{\epsilon})}{1+\epsilon}.$$

Set

$$C_0(\epsilon) := 1 + \sum_{i=2}^{\lfloor e^{\epsilon} \rfloor} \frac{\ln^{\epsilon}(i)}{i} - \frac{\ln^{1+\epsilon}(\lceil e^{\epsilon} \rceil)}{1+\epsilon} \qquad \text{and} \qquad \tilde{C_0}(\epsilon) := 1 + \sum_{i=2}^{\lceil e^{\epsilon} \rceil} \frac{\ln^{\epsilon}(i)}{i} - \frac{\ln^{1+\epsilon}(e^{\epsilon})}{1+\epsilon},$$

we get:

$$C_0(\epsilon) + \frac{\ln^{1+\epsilon}(j)}{1+\epsilon} \leq \mathbb{E}\left[\sum_{i=1}^{j} B^i\right] \leq \tilde{C_0}(\epsilon) + \frac{\ln^{1+\epsilon}(j)}{1+\epsilon}.$$

Using the multiplicative Chernoff bound for independent Bernoulli random variables, the inequalities above, and $\tilde{C_0}(\epsilon) \geq 0$ for all $\epsilon > 0$, we get the following upper bound on the probability of $\hat{X}^j$ being small:

$$\mathbb{P}\left[\hat{X}^j < \frac{l_1 \ln^{1+\epsilon}(j)}{2(1+\epsilon)}\right] \leq \mathbb{P}\left[l_1 \sum_{i=1}^{j} B^j < \frac{l_1 \ln^{1+\epsilon}(j)}{2(1+\epsilon)}\right] \leq \exp\left(-\frac{C_0(\epsilon)}{8} - \frac{\ln^{1+\epsilon}(j)}{8(1+\epsilon)}\right).$$

Let ζ_i be a Bernoulli random variable such that $\zeta_i = 1$ when there is at least one potential service completion between the arrival time of the i^{th} and $(i+1)^{th}$ customer. The random variables $\{\zeta_i\}_i$ are *i.i.d.* and $\mathbb{P}[\zeta_i = 1] = \mu/(\lambda+\mu)$. When the threshold used is at least 1, if the i^{th} customer is rejected, the queue-length at the arrival of this customer is non-zero; obviously, when the i^{th} customer is admitted to the queue, the queue-length right after the arrival of this customer is non-zero. In either case, if there are any potential services during the inter-arrival times between the i^{th} and $(i+1)^{th}$ customers, at least one of the completed services is observed by the learning dispatcher. This implies that $\sum_{i \text{ counted in } \hat{X}_j} \zeta_i = \sum_{n=0}^{\hat{X}_j} \zeta_{cnt_n} \leq D_j$, where cnt_n is a sub-sequence of i and cnt_n is the index from the beginning of the n^{th} arrival customer that is counted in $\hat{X}_j$. Then we have:

$$\mathbb{P}\left[D_j \leq \frac{l_1 \ln^{1+\epsilon}(j)\mu}{4(1+\epsilon)(\lambda+\mu)} \middle| \hat{X}^j \geq \frac{l_1 \ln^{1+\epsilon}(j)}{2(1+\epsilon)}\right] \leq \mathbb{P}\left[\sum_{n=1}^{\lceil l_1 \ln^{1+\epsilon}(j)/2(1+\epsilon)\rceil} \zeta_{cnt_n} \leq \frac{l_1 \ln^{1+\epsilon}(j)\mu}{4(1+\epsilon)(\lambda+\mu)}\right]$$
$$\leq \exp\left(-\frac{l_1 \ln^{1+\epsilon}(j)\mu}{16(1+\epsilon)(\lambda+\mu)}\right).$$

We dropped the conditioning in the first inequality using $\sum_{i=1}^{\hat{X}^j} \zeta_i \leq D_j$, and $\mathbb{P}[\sum_{i=1}^{n+1} \zeta_i \leq c] \leq \mathbb{P}[\sum_{i=1}^{n} \zeta_i \leq c]$ for all $n, c \in \mathbb{Z}^+$, and the second inequality follows from multiplicative Chernoff bound for independent Bernoulli random variables. Combining the results above, we obtain:

$$\mathbb{P}\left[D_j \leq \frac{l_1 \ln^{1+\epsilon}(j)\mu}{4(1+\epsilon)(\lambda+\mu)}\right] = \mathbb{P}\left[D_n \leq \frac{l_1 \ln^{1+\epsilon}(j)\mu}{4(1+\epsilon)(\lambda+\mu)} \middle| \hat{X}^j \geq \frac{l_1 \ln^{1+\epsilon}(j)}{2(1+\epsilon)}\right] \mathbb{P}\left[\hat{X}^j \geq \frac{l_1 \ln^{1+\epsilon}(j)}{2(1+\epsilon)}\right]$$
$$+ \mathbb{P}\left[D_j \leq \frac{l_1 \ln^{1+\epsilon}(j)\mu}{4(1+\epsilon)(\lambda+\mu)} \middle| \hat{X}^j < \frac{l_1 \ln^{1+\epsilon}(j)}{2(1+\epsilon)}\right] \mathbb{P}\left[\hat{X}^j < \frac{l_1 \ln^{1+\epsilon}(j)}{2(1+\epsilon)}\right]$$
$$\leq \exp\left(-\frac{l_1 \ln^{1+\epsilon}(j)\mu}{16(1+\epsilon)(\lambda+\mu)}\right) + \exp\left(-\frac{C_0(\epsilon)}{8} - \frac{\ln^{1+\epsilon}(j)}{8(1+\epsilon)}\right).$$

This completes the proof. □

Using Proposition 2.4.1 above, in the next proposition we will establish that with high probability, the learning dispatcher will have an accurate estimate of the average service time, and therefore the service rate.

Proposition 2.4.2. *Let $\hat{m}(j)$ denote the empirical service time estimated by the learning dispatcher at the beginning of phase 2 of the j^{th} batch. For the proposed algorithm,*

$$\mathbb{P}\left[|\hat{m}(j) - m| > \Delta_1\right] \leq C_1 \exp(-C_2 \ln^{1+\epsilon}(j)), \tag{II.15}$$

where

$$\begin{aligned} C_1 &:= \max\left\{\exp\left(-\frac{C_0(\epsilon)}{8}\right),\ \frac{2\exp\left(\Delta_1^2/(8m^2)\right)}{\exp\left(\Delta_1^2/(8m^2)\right)-1},\ 1\right\}, \\ C_2 &:= \min\left\{\frac{l_1\mu}{16(1+\epsilon)(\lambda+\mu)},\ \frac{1}{8(1+\epsilon)},\ \frac{l_1\mu\Delta_1^2}{32(1+\epsilon)m(\lambda m+1)}\right\}, \end{aligned} \tag{II.16}$$

with $\Delta_1 := \min\{\delta_1, 2m\}$*, and* δ_1 *is the constant from inequality* (II.4) *which is one part of the condition needed for the conclusion in* (II.5).

The proof of the proposition relies upon tail concentration bounds for sub-exponential random variables. We follow the definition and concentration bounds as in [Wainwright 2019, Section 2.1].

Definition 2.4.1. *A random variable* X *with mean* μ *is called sub-exponential if there are non-negative parameters* (α^2, β) *such that* $\mathbb{E}[e^{\gamma(X-\mu)}] \le e^{\frac{\alpha^2\gamma^2}{2}}$ *for all* $|\gamma| < \frac{1}{\beta}$.

Proposition 2.4.3. *Suppose that* X *is sub-exponential with parameters* (α^2, β)*. Then:*

$$\mathbb{P}[X \ge \mu + t] \le \begin{cases} e^{-\frac{t^2}{2\alpha^2}}, & 0 \le t \le \frac{\alpha^2}{\beta}, \\ e^{-\frac{t}{2\beta}}, & t \ge \frac{\alpha^2}{\beta}, \end{cases} = \max\left\{e^{-\frac{t^2}{2\alpha^2}}, e^{-\frac{t}{2\beta}}\right\}.$$

Proof of Proposition 2.4.2. Let S_i denote the service time of the i^{th} service completion. Since S_i are *i.i.d.* with distribution EXP$(1/m)$, which is a $(4m^2, 2m)$ sub-exponential random variable, $\sum_{i=1}^n S_i$ is a $(4m^2n, 2m)$ sub-exponential random variable; see [Vershynin 2018, Section 2.8]. Observe that $0 \le k\Delta_1 \le 2mk$. Using the sub-exponential concentration bounds above, we get:

$$\begin{aligned} \mathbb{P}\left[|\hat{m}(j) - m| > \Delta_1 | D_j > n\right] &\le \sum_{k=n+1}^{\infty} \mathbb{P}\left[\left|\sum_{i=1}^{k} S_i - km\right| \ge k\Delta_1\right] \\ &\le \sum_{k=n+1}^{\infty} 2\exp\left(-\frac{k\Delta_1^2}{8m^2}\right) \le \frac{2\exp\left(\Delta_1^2/(8m^2)\right)}{\exp\left(\Delta_1^2/(8m^2)\right)-1}\exp\left(-\frac{(n+1)\Delta_1^2}{8m^2}\right). \end{aligned}$$

The third inequality follows by the geometric sum formula.

Then, substituting $n = \lfloor l_1 \ln^{1+\epsilon}(j)\mu/(4(1+\epsilon)(\lambda+\mu))\rfloor$, we get:

$$
\begin{aligned}
\mathbb{P}\left[|\hat{m}(j) - m| > \Delta_1 \middle| D_j > \frac{l_1 \ln^{1+\epsilon}(j)\mu}{4(1+\epsilon)(\lambda+\mu)}\right] &= \mathbb{P}\left[|\hat{m}(j) - m| > \Delta_1 \middle| D_j > \left\lfloor \frac{l_1 \ln^{1+\epsilon}(j)\mu}{4(1+\epsilon)(\lambda+\mu)} \right\rfloor\right] \\
&\leq \frac{2\exp\left(\Delta_1^2/(8m^2)\right)}{\exp\left(\Delta_1^2/(8m^2)\right) - 1} \exp\left(-\left(\left\lfloor \frac{l_1 \ln^{1+\epsilon}(j)\mu}{4(1+\epsilon)(\lambda+\mu)} \right\rfloor + 1\right)\frac{\Delta_1^2}{8m^2}\right) \\
&\leq \frac{2\exp\left(\Delta_1^2/(8m^2)\right)}{\exp\left(\Delta_1^2/(8m^2)\right) - 1} \exp\left(-\frac{l_1\mu \ln^{1+\epsilon}(j)\Delta_1^2}{32(1+\epsilon)m^2(\lambda+\mu)}\right).
\end{aligned}
$$

Using the last upper bound and Proposition 2.4.1, we find:

$$
\begin{aligned}
\mathbb{P}\left[|\hat{m}(j) - m| > \Delta_1\right] &= \mathbb{P}\left[|\hat{m}(j) - m| > \Delta_1 \middle| D_j \leq \frac{l_1 \ln^{1+\epsilon}(j)\mu}{4(1+\epsilon)(\lambda+\mu)}\right] \mathbb{P}\left[D_j \leq \frac{l_1 \ln^{1+\epsilon}(j)\mu}{4(1+\epsilon)(\lambda+\mu)}\right] \\
&\quad + \mathbb{P}\left[|\hat{m}(j) = m| > \Delta_1 \middle| D_j > \frac{l_1 \ln^{1+\epsilon}(j)\mu}{4(1+\epsilon)(\lambda+\mu)}\right] \mathbb{P}\left[D_j > \frac{l_1 \ln^{1+\epsilon}(j)\mu}{4(1+\epsilon)(\lambda+\mu)}\right] \\
&\leq \exp\left(-\frac{l_1 \ln^{1+\epsilon}(j)\mu}{16(1+\epsilon)(\lambda+\mu)}\right) + \exp\left(-\frac{C_0(\epsilon)}{8} - \frac{\ln^{1+\epsilon}(j)}{8(1+\epsilon)}\right) \\
&\quad + \frac{2\exp\left(\Delta_1^2/(8m^2)\right)}{\exp\left(\Delta_1^2/(8m^2)\right) - 1} \exp\left(-\frac{l_1\mu \ln^{1+\epsilon}(j)\Delta_1^2}{32(1+\epsilon)m^2(\lambda+\mu)}\right) \\
&\leq C_1 \exp(-C_2 \ln^{1+\epsilon}(j)),
\end{aligned}
$$

where C_1 and C_2 are given by (II.16).

□

Proposition 2.4.4. *Let $\nu(j)$ denote the empirical inter-arrival time estimated by the learning dispatcher at the beginning of phase 2 of the j^{th} batch. For the proposed algorithm,*

$$
\mathbb{P}\left[|\nu - \hat{\nu}(j)| > \Delta_2\right] \leq C_3 \exp(-C_4 \beta_j),
$$

where

$$
C_3 := \frac{2\exp(\Delta_2^2/(8\nu^2))}{\exp(\Delta_1^2/(8\nu^2)) - 1}, \quad C_4 := \frac{l_1 \Delta_2^2}{8\nu^2}, \quad \text{and} \quad \beta_j := 1 + \sum_{i=1}^{j-1} \alpha_i, \tag{II.17}
$$

with $\Delta_2 := \min\{\delta_2, 2\nu\}$, and δ_2 is the constant from inequality (II.4) *which is the second part of the condition needed for the conclusion in* (II.5).

Proof. Note that no matter whether customers are admitted to the queue or not, the learning dispatcher is able to observe all arrivals. We always have the first phase 1, and that the number of customers who arrived during the j^{th} phase 2 is at least $\alpha_j l_2$. Note that we also have $l_2 > l_1$. Let $\beta_j = 1 + \sum_{i=1}^{j-1} \alpha_i$. Right before the j^{th} phase 2, there are at least $l_1 + \sum_{n=1}^{j-1} \alpha_n l_2 \geq \beta_j l_1$ customers that have arrived at the system, and the learning dispatcher would have observed all the inter-arrival times. Following a similar logic as in the proof of Proposition 2.4.2, let A_i denote the inter-arrival time of consecutive customers. A_i are *i.i.d.* with distribution $\mathrm{EXP}(1/\nu)$, which is a $(4\nu^2, 2\nu)$ sub-exponential random variable. Using the concentration result detailed in Proposition 2.4.3 for sub-exponential random variables, we have:

$$\begin{aligned}\mathbb{P}\left[|\nu - \hat{\nu}(j)| > \Delta_2\right] &\leq \sum_{k=\beta_j}^{\infty} \mathbb{P}\left[\left|\sum_{i=1}^{k} A_i - k\nu\right| > k\Delta_2\right] \\ &\leq \sum_{k=\beta_j}^{\infty} 2\exp\left(-\frac{k\Delta_2^2}{8\nu^2}\right) \leq \frac{2\exp(\Delta_2^2/(8\nu^2))}{\exp(\Delta_1^2/(8\nu^2)) - 1} \exp\left(-\frac{\beta_j l_1 \Delta_2^2}{8\nu^2}\right),\end{aligned}$$

which establishes the result. □

Note that since $\alpha_j \geq 1$ for all j, $\beta_j \geq j$. Therefore, as the number of batches, j, increases, the probability of not having a correct estimate of the average arrival rate decreases faster than the probability of not having a correct estimate of the average service time. In the following corollary, we will combine Propositions 2.4.2 and 2.4.4 to get a bound on the probability of the learning dispatcher not using (an optimal) threshold $\bar{K}$ when j is large.

Corollary 2.4.1. *For the proposed algorithm, when* $j \geq \lceil e^{\bar{K}} \rceil$,

$$\mathbb{P}\left[K(j) \neq \bar{K}\right] \leq C_1 \exp(-C_2 \ln^{1+\epsilon}(j)) + C_3 \exp(-C_4 \beta_j), \tag{II.18}$$

where C_1 *and* C_2 *are defined in* (II.16)*;* C_3 *and* C_4 *are defined in* (II.17).

Proof. Recall that for the true arrival and service rates λ and μ, we have

$$\hat{V}(\bar{K}, \mu, \lambda) < \frac{R}{C} < \hat{V}(\bar{K} + 1, \mu, \lambda).$$

Proposition 2.2.1 says that if $\hat{m}$ and $\hat{\nu}$ satisfy inequality (II.4), then the learning dispatcher would be able to solve for the desired threshold $\bar{K}$. Moreover, since $j > e^{\bar{K}}$, $K^*(j) \geq \bar{K}$, i.e., the learning dispatcher would be able to use $\bar{K}$ in the j^{th} phase 2. Using Proposition 2.4.2 and Proposition 2.4.4, we have:

$$\begin{aligned} \mathbb{P}\left[K(j) \neq \bar{K}\right] &\leq \mathbb{P}\left[|m - \hat{m}(j)| > \Delta_1\right] + \mathbb{P}\left[|\nu - \hat{\nu}(j)| > \Delta_2\right] \\ &\leq C_1 \exp(-C_2 \ln^{1+\epsilon}(j)) + C_3 \exp(-C_4 \beta_j), \end{aligned}$$

which concludes the proof. □

When the learning dispatcher has knowledge of either μ or λ, one can obtain an inequality similar to that in Corollary 2.4.1 by setting the corresponding bound from Propositions 2.4.2 and 2.4.4 to 0. When the service rate is known and the arrival rate is not known, then a better characterization of the regret obtains; see Remark 2.4.2.

2.4.2 Regret accumulated in each phase

We now analyze the regret. Let G_1^j denote the expected regret accumulated during the period starting with the (potential) phase 1 and ending at the first time the queue is emptied in the immediate phase 2 for the j^{th} batch that follows. Let G_2^j denote the expected regret accumulated in the remainder of phase 2 of the j^{th} batch. Whenever phase 1 of the j^{th} batch does not happen, there is no regret to be grouped to G_1^j, and the regret accumulated in phase 2 is entirely in G_2^j; in this case, the regret accumulated during the entire j^{th} batch is also solely in G_2^j. Both G_1^j and G_2^j count for the regret accumulated because of not having accurate estimates of the service rate as well as not estimating the arrival rate accurately. Intuitively, G_1^j takes into consideration the regret accumulated because of the existence of a phase 1, and G_2^j considers the regret accumulated because of the learning system using an incorrect threshold. Despite the subtleties, for easier recall, we refer to G_i^j as the regret accumulated in phase $i \in \{1, 2\}$ of batch j.

Let N denote the number of arrivals as a function of which we will determine the regret. Then, we have:

$$\tilde{G}(N) \leq \mathbb{E}\left[\sum_{j=1}^{J}(G_1^j + G_2^j)\right] \leq \sum_{j=1}^{\lceil N/l_2 \rceil}(G_1^j + G_2^j), \tag{II.19}$$

where $J := J(N)$ is the total number of batches until N arrivals including the batch in progress or initiated by the N^{th} arrival. The last inequality follows by the observation:

$$N \geq \sum_{i=1}^{J} \alpha_i l_2 \geq \beta_J l_2 \geq J l_2,$$

which implies $J \leq N/l_2$ a.s. When one uses α_j that grows like j^α, for some $\alpha > 0$, we obtain that J is of order of $O(N^{1/(a+1)})$. This adjustment would not affect the order of the regret but only the constants: see Sections 2.4.3 and 2.4.4.

For each j, we will analyze G_1^j and G_2^j separately. Let $\mathcal{E}_1^j$ denote the event that phase 1 of the j^{th} batch happens. Since in the proposed algorithm, we always have the first phase 1, we have $\mathbb{P}[\mathcal{E}_1^1] = 1$. Phase 1 is omitted when the threshold used in the previous phase 2 is non-zero. By the independence of B^j and $K(j)$, for $j > 1$ we have:

$$\begin{aligned}\mathbb{P}\left[\mathcal{E}_1^j\right] &= \mathbb{P}\left[\mathcal{E}_1^j \middle| K(j-1) = 0\right] \mathbb{P}\left[K(j-1) = 0\right] + \mathbb{P}\left[\mathcal{E}_1^j \middle| K(j-1) \neq 0\right] \mathbb{P}\left[K(j-1) \neq 0\right] \\ &= \mathbb{P}\left[B^j = 1\right] \mathbb{P}\left[K(j-1) = 0\right]. \end{aligned} \tag{II.20}$$

Let $\mathcal{E}_2^j$ denote the event that $K(j) = \bar{K}$, and $\mathcal{E}_3^j$ denote the event that the queue-lengths of the two systems are the same at the beginning of the j^{th} batch, i.e.,

$$\mathcal{E}_2^j := \{K(j) = \bar{K}\} \qquad \text{and} \qquad \mathcal{E}_3^j := \{Q_{n^j} = \bar{Q}_{n^j}\}.$$

Also, denote by $\tau^{K,l}$ the number of arrivals during a busy period of an $M/M/1/K$ queue with initial queue-length l. The proof of lemmas 2.4.1 and 2.4.2 rely on an upper bound of $\mathbb{E}\left[\tau^{K,l}\right]$ which is stated in the following proposition.

Proposition 2.4.5. *Consider an $M/M/1/K$ queue with arrival rate λ, service rate μ and intial queue length $0 < l \leq K$.*

$$\mathbb{E}\left[\tau^{K,l}\right] \leq g(l;K), \tag{II.21}$$

where

$$g(1;K) = \begin{cases} \frac{\lambda/\mu+1}{\lambda/\mu-1}\left(\left(\frac{\lambda}{\mu}\right)^K - 1\right), & \lambda \neq \mu, \\ 2K, & \lambda = \mu, \end{cases}$$

and for all $1 < l \leq K$,

$$g(l;K) = \begin{cases} \frac{\lambda/\mu+1}{(\lambda/\mu-1)^2}\left(\left(1-\left(\frac{\lambda}{\mu}\right)^l\right)\left(\left(\frac{\lambda}{\mu}\right)^{K+1} - \frac{\lambda}{\mu} + 1\right) + (l-1)\left(1-\frac{\lambda}{\mu}\right)\right), & \lambda \neq \mu, \\ l(2K-l+1), & \lambda = \mu. \end{cases}$$

In particular, $\mathbb{E}\left[\tau^{K,l}\right]$ is of order $O((\lambda/\mu)^K + K^2)$.

Proof. Consider a finite state Markov chain with state space $\{0, 1, ..., .K\}$, and with the following transition matrix:

$$
\begin{aligned}
&p(0,0) = 1; \\
&p(l,l+1) = \frac{\lambda}{\lambda+\mu}, \quad p(l,l-1) = \frac{\mu}{\lambda+\mu}, \qquad \text{when } l \in \{1, ..., K-1\}; \\
&p(K,K) = \frac{\lambda}{\lambda+\mu}, \quad p(K,K-1) = \frac{\mu}{\lambda+\mu};
\end{aligned}
$$

Let $g(l;K)$ denote the expected number of jumps of this Markov chain until it hits 0 for the first time when the initial state is l and the threshold is K. Conditional on the first jump, we obtain the following relationship for $g(l:K)$,

$$
\begin{aligned}
&g(l;K) = \frac{\lambda}{\lambda+\mu} g(l+1;K) + \frac{\mu}{\lambda+\mu} g(l-1;K) + 1, \qquad \text{when } l \in \{1, ..., K-1\}; \\
&g(K;K) = \frac{\lambda}{\lambda+\mu} g(K;K) + \frac{\mu}{\lambda+\mu} g(K-1;K) + 1;
\end{aligned}
$$

together with the condition $g(0;K) = 0$, we can solve for $g(l;K)$, and obtain:

$$
g(1;K) = \begin{cases} \frac{\lambda/\mu+1}{\lambda/\mu-1} \left(\left(\frac{\lambda}{\mu}\right)^K - 1 \right), & \lambda \neq \mu, \\ 2K, & \lambda = \mu, \end{cases}
$$

and for all $1 < l \leq K$,

$$
g(l;K) = \begin{cases} \frac{\lambda/\mu+1}{(\lambda/\mu-1)^2} \left(\left(1 - \left(\frac{\lambda}{\mu}\right)^l\right) \left(\left(\frac{\lambda}{\mu}\right)^{K+1} - \frac{\lambda}{\mu} + 1 \right) + (l-1)\left(1 - \frac{\lambda}{\mu}\right) \right), & \lambda \neq \mu, \\ l(2K-l+1), & \lambda = \mu. \end{cases}
$$

From the transition probabilities of the Markov chain, $g(n:K)$ is also the expected number of services and arrivals of the corresponding $M/M/1/K$ queue with arrival rate $\lambda > 0$, service rate $\mu > 0$ and initial queue length l during the busy period which is initiated with

n customers in the queue. Since each arrival must also be served when the Markov chain hits 0, $\mathbb{E}\left[\tau^{K,l}\right] \leq g(l;K) \leq 2\mathbb{E}\left[\tau^{K,l}\right] + K$. Therefore, $g(l;K)$ serves as an upper bound on $\mathbb{E}\left[\tau^{K,l}\right]$. This upper-bound is tight in the sense that $g(l;K)$ is at most $2\mathbb{E}\left[\tau^{K,l}\right] + K$. □

Lemma 2.4.1. *For $j > e^{\bar{K}}$, we have the following:*

1. When $\bar{K} > 0$,

$$G_1^j \leq \left(R + \frac{C}{\lambda}\right)\left(l_1^2 + \left(\bar{K}+1\right)l_1 + \left(1 + K^*(j)\right)g(l_1;K^*(j))\right)\mathbb{P}\left[\mathcal{E}_1^j\right];$$

2. When $\bar{K} = 0$,

$$G_1^j \leq \left(R + \frac{C}{\lambda}\right)\left(l_1^2 + l_1 + C_5\right)\mathbb{P}\left[\mathcal{E}_1^j\right] + \left(R + \frac{C}{\lambda}\right)(1 + K^*(j))g(l_1;K^*(j))\mathbb{P}\left[\left(\mathcal{E}_2^j\right)^c\right];$$

where

$$C_5 := (1 + l_1)\,\frac{l_1\lambda}{\mu}.$$

The function $g(l;K)$ is defined in Proposition 2.4.5, and is $O((\lambda/\mu)^K + K^2)$ for all $l \leq K$.

Proof. Let n^j denote the total number of customers that arrived until the beginning of the j^{th} batch, and $L_1^j := \min\{n \mid Q_{n^j+l_1+n} = 0\}$. Recall that $\mathcal{E}_1^j$ denotes the event that phase 1 happens during the j^{th} batch. Using (II.12) and observing that regret accumulates in G_1^j only when $\mathcal{E}_1^j$ happens, we have:

$$\begin{aligned}
G_1^j \leq & \left(R+\frac{C}{\lambda}\right)\mathbb{E}\left[\sum_{i=n^j+1}^{n^j+l_1}\left|\mathbb{1}_{\{\bar{Q}_i<\bar{K}_i\}}-\mathbb{1}_{\{Q_i<K_i\}}\right|+\left|\bar{Q}_i-Q_i\right|\middle|\mathcal{E}_1^j\right]\mathbb{P}\left[\mathcal{E}_1^j\right] \\
& +\left(R+\frac{C}{\lambda}\right)\mathbb{E}\left[\sum_{i=n^j+l_1+1}^{n^j+l_1+L_1^j}\left|\mathbb{1}_{\{\bar{Q}_i<\bar{K}_i\}}-\mathbb{1}_{\{Q_i<K_i\}}\right|+\left|\bar{Q}_i-Q_i\right|\middle|\mathcal{E}_1^j\right]\mathbb{P}\left[\mathcal{E}_1^j\right] \\
& =: (I)+(II).
\end{aligned}$$

Note that I is a bound on the regret accumulated during phase 1 of the j^{th} batch (when it occurs), and II is a bound on the regret accumulated in phase 2 of the j^{th} batch until the queue is emptied for the first time in this phase 2. When $\bar{K}>0$ or $\bar{K}=0$, we can follow the same logic to bound I, i.e. the regret accumulated during phase 1 for $j>e^{\bar{K}}$:

$$(I)\leq\left(R+\frac{C}{\lambda}\right)\mathbb{E}\left[\sum_{i=n^j}^{n^j+l_1}\left(1+\bar{K}+l_1\right)\right]\mathbb{P}\left[\mathcal{E}_1^j\right]\leq\left(R+\frac{C}{\lambda}\right)\left(l_1^2+\left(\bar{K}+1\right)l_1\right)\mathbb{P}\left[\mathcal{E}_1^j\right].$$

Now, we bound II in the case $\bar{K}>0$. We use $K^*(j)$ to obtain a bound on the queue-length difference of the two systems as well as the expectation of L_1^j. The queue-length of the learning system at the beginning of each phase 2 is at most l_1 since the queue-length of the learning system is 0 at the end of the previous phase 2. Moreover, the threshold used by the learning dispatcher in the j^{th} batch is bounded above by $K^*(j)\geq l_1$. Hence the queue-length of the learning system is bounded by $K^*(j)$ during phase 2. Consider a system S_2 that uses the admission policy with threshold $K^*(j)$ and which is coupled with the learning system according to Section 2.2.2.1. Assume that the initial queue-length of S_2 is the same as the queue-length of the learning system at the beginning of the j^{th} phase 2 which is at most l_1. Note that the threshold used in the learning system is less or equal to the one used in S_2. Let τ denote the total number of arrivals during the first busy period of the system S_2. Using Proposition 2.3.1, we get $Q_i\leq Q_i^{S_2}$ for $n^j+l_1+1\leq i\leq n^j+l_1+L_1^j$, and

$\mathbb{E}[L_1^j|\mathcal{E}_1^j] \leq \mathbb{E}\left[\tau^{K^*(j),l_1}\right]$. Using Proposition 2.4.5, and together with the upper bound $K^*(j)$ of the queue-length of the learning system, we get :

$$(II) \leq \left(R + \frac{C}{\lambda}\right)(1 + K^*(j))\,\mathbb{E}\left[L_1^j|\mathcal{E}_1^j\right]\mathbb{P}\left[\mathcal{E}_1^j\right] \leq \left(R + \frac{C}{\lambda}\right)(1 + K^*(j))\,g(l_1; K^*(j))\mathbb{P}\left[\mathcal{E}_1^j\right],$$

where $F_1(j)$ is defined in the statement of Lemma 2.4.1.

Together, we have the following bound for G_1^j when $\bar{K} > 0$:

$$G_1^j \leq \left(R + \frac{C}{\lambda}\right)\Big(l_1^2 + \left(\bar{K} + 1\right) l_1 + \left(1 + K^*(j)\right) g(l_1; K^*(j))\Big)\mathbb{P}\left[\mathcal{E}_1^j\right].$$

In the case of $\bar{K} = 0$, we take a slightly different path of analyzing II: we consider the threshold used in the j^{th} phase 2 to get a better regret bound compared to using the same argument as in the case $\bar{K} > 0$. We have:

$$\begin{aligned}
(II) &= \left(R + \frac{C}{\lambda}\right)\mathbb{E}\left[\sum_{i=n^j+l_1+1}^{n^j+l_1+L_1^j} \left|\mathbb{1}_{\{\bar{Q}_i < \bar{K}_i\}} - \mathbb{1}_{\{Q_i < K_i\}}\right| + \left|\bar{Q}_i - Q_i\right| \middle| \mathcal{E}_1^j \cap \mathcal{E}_2^j\right]\mathbb{P}\left[\mathcal{E}_1^j \cap \mathcal{E}_2^j\right] \\
&\quad + \left(R + \frac{C}{\lambda}\right)\mathbb{E}\left[\sum_{i=n^j+l_1+1}^{n^j+l_1+L_1^j} \left|\mathbb{1}_{\{\bar{Q}_i < \bar{K}_i\}} - \mathbb{1}_{\{Q_i < K_i\}}\right| + \left|\bar{Q}_i - Q_i\right| \middle| \mathcal{E}_1^j \cap (\mathcal{E}_2^j)^c\right]\mathbb{P}\left[\mathcal{E}_1^j \cap (\mathcal{E}_2^j)^c\right] \\
&\leq \left(R + \frac{C}{\lambda}\right)(1 + l_1)\,\mathbb{E}\left[L_1^j|\mathcal{E}_1^j \cap \mathcal{E}_2^j\right]\mathbb{P}\left[\mathcal{E}_1^j \cap \mathcal{E}_2^j\right] \\
&\quad + \left(R + \frac{C}{\lambda}\right)(1 + K^*(j))\,\mathbb{E}\left[L_1^j|\mathcal{E}_1^j \cap (\mathcal{E}_2^j)^c\right]\mathbb{P}\left[\mathcal{E}_1^j \cap (\mathcal{E}_2^j)^c\right] \\
&\leq \left(R + \frac{C}{\lambda}\right)(1 + l_1)\,\frac{l_1\lambda}{\mu}\mathbb{P}\left[\mathcal{E}_1^j\right] + \left(R + \frac{C}{\lambda}\right)(1 + K^*(j))\,g(l_1; K^*(j))\mathbb{P}\left[(\mathcal{E}_2^j)^c\right].
\end{aligned}$$

The first follows since the total number of customers admitted in phase 1 is l_1, and since in the case $\bar{K} = 0$ and under $\mathcal{E}_2^j$, the threshold used in phase 2 is 0. Under $\mathcal{E}_1^j \cap \mathcal{E}_2^j$, the learning system does not accept any new customers to the queue, and $\mathbb{E}[\mathcal{E}_1^j \cap \mathcal{E}_2^j]$ is the number of arrivals during the period of serving all the remaining customers in the queue. Observe the queue-length of the learning system at the beginning of phase 2 is at most l_1, conditioning on the time used to serve l_1 customers, we get the desired bound on $\mathbb{E}[\mathcal{E}_1^j \cap \mathcal{E}_2^j]$. The bound on

$\mathbb{E}[L_1^j|\mathcal{E}_1^j \cap (\mathcal{E}_2^j)^c]$ follows the same logic as the bound of $\mathbb{E}[L_1^j|\mathcal{E}_1^j]$. Combined with the bound for I, we get the desired result. $\square$

We observe that under the event $\mathcal{E}_2^j \cap \mathcal{E}_3^j$, there will be no regret accumulated in G_2^j: indeed, under the event $(\mathcal{E}_1^j)^c \cap \mathcal{E}_2^j \cap \mathcal{E}_3^j$, the dispatcher of the learning system and the dispatcher of the genie-aided system will make the same decision on every arrival customer in phase 2 of the j^{th} batch. As a result, their queue-lengths will be matched and there will be no regret accumulated during this exploitation phase, thus also no regret accumulated in G_2^j. The threshold used in phase 1 can be considered as the maximum allowed value, namely $K^*(j)(\geq l_1)$, since all the arriving customers during phase 1 are admitted. Under the event $\mathcal{E}_2^j$, the threshold used in the j^{th} phase 2 is the same as the genie-aided system. Therefore, under the event $\mathcal{E}_1^j \cap \mathcal{E}_2^j \cap \mathcal{E}_3^j$, although phase 1 of the j^{th} batch happens, the queue-length at the beginning of the j^{th} batch is the same for both systems and the thresholds used in the learning system is no smaller than the threshold used in the genie-aided system. The coupling between the learning and genie-aided system preserves the order between the queue-lengths of the two systems as proved in Proposition $2.3.1$: when the queue-length of the learning system hits 0 the first time after phase 1, the queue-length of the genie-aided system is also 0. Therefore, under event $\mathcal{E}_1^j \cap \mathcal{E}_2^j \cap \mathcal{E}_3^j$, after the queue-length of the learning system hits 0 after phase 1, the queue-lengths of the learning and genie-aided system are matched, and no regret is accumulated in G_2^j.

The next proposition shows that the probability of the event $\mathcal{E}_2^j \cap \mathcal{E}_3^j$ is high. We use De Morgan's law to get an upper bound on the probability of this event by using already characterized bounds on the probabilities of a few events.

Proposition 2.4.6. *Fix $j \geq \lceil e^{\bar{K}} \rceil$. Then, we have the following:*

1. In the case $\bar{K} > 0$,

$$\mathbb{P}\left[\left(\mathcal{E}_2^j \cap \mathcal{E}_3^j\right)^c\right] \leq C_1 \exp\Big(-C_2 \ln^{1+\epsilon}(j)\Big) + C_1 \exp\Big(-C_2 \ln^{1+\epsilon}(j-1)\Big) + C_3 \exp\Big(-C_4 \beta_j\Big) + C_3 \exp\Big(-C_4 \beta_{j-1}\Big) + \big(c_{\bar{K}}\big)^{\alpha_{j-1} l_2}.$$

2. In the case $\bar{K} = 0$,

$$\mathbb{P}\left[\left(\mathcal{E}_2^j \cap \mathcal{E}_3^j\right)^c\right] \leq C_1 \exp(-C_2 \ln^{1+\epsilon}(j)) + C_3 \exp(-C_4 \beta_j).$$

The constants C_1, C_2, C_3 and C_4 are defined in (II.16) *and* (II.17)*, and*

$$c_{\bar{K}} := 1 - \left(\frac{\mu}{\lambda + \mu}\right)^{\bar{K}} \in (0,1).$$

Proof. We first consider the case $\bar{K} > 0$. Let $\mathcal{E}_4^j$ denote the event that the queue-length of the genie-aided system hits 0 during phase 2 of the j^{th} batch. The probability that at least $\bar{K}$ potential services occur between two consecutive inter-arrivals is $1 - c_{\bar{K}}$. Since the genie-aided system is an $M/M/1/\bar{K}$ queue, there are at most $\bar{K}$ customers in the queue. Since the total number of arrivals during the phase 2 of the j^{th} batch is at least $\alpha_j l_2$, we get:

$$\mathbb{P}[(\mathcal{E}_4^j)^c] \leq (c_{\bar{K}})^{\alpha_j l_2}.$$

By Corollary 2.3.1, we have:

$$\mathbb{P}\left[\left(\mathcal{E}_3^j\right)^c \mid \mathcal{E}_2^{j-1}\right] \leq \mathbb{P}\left[\left(\mathcal{E}_4^{j-1}\right)^c \mid \mathcal{E}_2^{j-1}\right] \leq (c_{\bar{K}})^{\alpha_{j-1} l_2}.$$

Using De Morgan's laws we can re-write the event $(\mathcal{E}_2^j \cap \mathcal{E}_3^j)^c$ as $(\mathcal{E}_2^j)^c \cup (\mathcal{E}_3^j)^c$ and by using

Corollary 2.4.1 for $j > e^{\bar{K}}$ we obtain:

$$\begin{aligned}\mathbb{P}\left[\left(\mathcal{E}_2^j \cap \mathcal{E}_3^j\right)^c\right] &\leq \mathbb{P}\left[\left(\mathcal{E}_2^j\right)^c\right] + \mathbb{P}\left[\left(\mathcal{E}_2^{j-1}\right)^c\right] + \mathbb{P}\left[\left(\mathcal{E}_3^j\right)^c \,\middle|\, \mathcal{E}_2^{j-1}\right] \\ &\leq C_1 \exp\Big(- C_2 \ln^{1+\epsilon}(j)\Big) + C_1 \exp\Big(- C_2 \ln^{1+\epsilon}(j-1)\Big) \\ &\qquad + C_3 \exp\Big(- C_4 \beta_j\Big) + C_3 \exp\Big(- C_4 \beta_{j-1}\Big) + (c_{\bar{K}})^{\alpha_{j-1} l_2}.\end{aligned}$$

In case that $\bar{K} = 0$, the queue-length of the genie-aided system is always 0, and $\mathcal{E}_3^j$ happens with probability 1. Hence,

$$\mathbb{P}\left[\left(\mathcal{E}_2^j \cap \mathcal{E}_3^j\right)^c\right] = \mathbb{P}\left[(\mathcal{E}_2^j)^c\right] \leq C_1 \exp(-C_2 \ln^{1+\epsilon}(j)) + C_3 \exp(-C_4 \beta_j).$$

This completes the proof. □

Next, we will estimate G_2^j, which considers the regret accumulated during the j^{th} batch after the first time the queue-length of the learning system hit 0 during the j^{th} phase 2 if there is a phase 1, and considers the regret accumulated during phase 2 if phase 1 did not happen. As we mentioned before, only under the event $\left(\mathcal{E}_2^j \cap \mathcal{E}_3^j\right)^c$, regret is accumulated to G_2^j.

Lemma 2.4.2. *For $j > e^{\bar{K}}$,*

$$G_2^j \leq \left(R + \frac{C}{\lambda}\right) \Big((1 + K^*(j))\, \alpha_j l_2 + \big(1 + K^*(j)\big) g(K^*(j); K^*(j))\Big) \mathbb{P}\left[\left(\mathcal{E}_2^j \cap \mathcal{E}_3^j\right)^c\right],$$

with $g(l; K)$ defined in Proposition 2.4.5.

Proof. Let $\tilde{n}^j$ denote the total number of customers that arrived until the beginning of phase 2 of the j^{th} batch. Note that when phase 1 did not happen in the j^{th} batch, $\tilde{n}^j = n^j$, and when phase 1 happened, $\tilde{n}^j = n^j + l_1$. However, since we are analyzing the regret accumulated in phase 2 because of using an incorrect threshold and not conditional on having a phase 1 or no, using $\tilde{n}^j$ would give simpler expressions during the analysis. By its definition, G_2^j takes into consideration only part of the regret that is accumulated in phase 2. Since we are interested in finding an upper bound, we will "double-count" parts of the regret that are already considered in G_1^j in the case that there is a phase 1 and compute the regret accumulated during phase 2. Set $L_2^j := \min\{n \mid Q_{\tilde{n}^j+\alpha_j l_2+n} = 0\}$. This is the total number of arriving customers beyond the first $\alpha_j l_2$ ones during the exploitation phase for the j^{th} batch. Using (II.12) and $(\mathcal{E}_2^j \cap \mathcal{E}_3^j)^c$, we get:

$$\begin{aligned}
G_2^j &\leq \left(R + \frac{C}{\lambda}\right) \mathbb{E}\left[\sum_{i=\tilde{n}^j+1}^{\tilde{n}^j+\alpha_j l_2+L_2^j} \left|\mathbb{1}_{\left\{\bar{Q}_i<\bar{K}\right\}} - \mathbb{1}_{\{Q_i<K(j)\}}\right| \mathbb{1}_{\left\{\left(\mathcal{E}_2^j\cap\mathcal{E}_3^j\right)^c\right\}}\right] \\
&\quad + \left(R + \frac{C}{\lambda}\right) \mathbb{E}\left[\sum_{i=\tilde{n}^j}^{\tilde{n}^j+\alpha_j l_2+L_2^j} \left|\bar{Q}_i - Q_i\right| \mathbb{1}_{\left\{\left(\mathcal{E}_2^j\cap\mathcal{E}_3^j\right)^c\right\}}\right] \\
&=: \left(R + \frac{C}{\lambda}\right) \Big((III) + (IV)\Big).
\end{aligned}$$

In what follows we bound the two expectations on the RHS. For the first expectation, since $|\mathbb{1}_{\{\bar{Q}_i<\bar{K}\}} - \mathbb{1}_{\{Q_i<K(j)\}}| \leq 1$, after splitting phase 2 into two parts, we get:

$$\begin{aligned}
(III) &\leq \mathbb{E}\left[\sum_{i=\tilde{n}^j+1}^{\tilde{n}^j+\alpha_j l_2} \mathbb{1}_{\left\{\left(\mathcal{E}_2^j\cap\mathcal{E}_3^j\right)^c\right\}}\right] + \mathbb{E}\left[\sum_{i=\tilde{n}^j+\alpha_j l_2+1}^{\tilde{n}^j+\alpha_j l_2+L_2^j} \mathbb{1}_{\left\{\left(\mathcal{E}_2^j\cap\mathcal{E}_3^j\right)^c\right\}}\right] \\
&= \mathbb{E}\left[\alpha_j l_2 \mathbb{1}_{\left\{\left(\mathcal{E}_2^j\cap\mathcal{E}_3^j\right)^c\right\}}\right] + \mathbb{E}\left[L_2^j \mathbb{1}_{\left\{\left(\mathcal{E}_2^j\cap\mathcal{E}_3^j\right)^c\right\}}\right] \\
&= \alpha_j l_2 \mathbb{P}\left[\left(\mathcal{E}_2^j \cap \mathcal{E}_3^j\right)^c\right] + \mathbb{E}\left[L_2^j \middle| \left(\mathcal{E}_2^j \cap \mathcal{E}_3^j\right)^c\right] \mathbb{P}\left[\left(\mathcal{E}_2^j \cap \mathcal{E}_3^j\right)^c\right].
\end{aligned}$$

Using a similar way of analyzing L_1^j in the proof of Lemma 2.4.1 but comparing with a coupled system that uses threshold $K^*(j)$ and having initial queue-length $K^*(j)$, we get:

$$\mathbb{E}\left[L_2^j \middle| \left(\mathcal{E}_2^j \cap \mathcal{E}_3^j\right)^c\right] \leq \mathbb{E}\left[\tau^{K^*(j),K^*(j)}\right] \leq g(K^*(j); K^*(j)).$$

Together with the inequalities above, we get a bound for (III):

$$(III) \leq (\alpha_j l_2 + g(K^*(j); K^*(j))) \, \mathbb{P}\left[\left(\mathcal{E}_2^j \cap \mathcal{E}_3^j\right)^c\right].$$

We can split (IV) in a similar manner as above, and then, together with $Q_i \leq K^*(j)$, we have:

$$\begin{aligned}(IV) &\leq K^*(j) \left(\mathbb{E}\left[\sum_{i=\tilde{n}^j}^{\tilde{n}^j+\alpha_j l_2} \mathbb{1}_{\left\{\left(\mathcal{E}_2^j \cap \mathcal{E}_3^j\right)^c\right\}} \right] + \mathbb{E}\left[\sum_{i=\tilde{n}^j \alpha_j l_2}^{\tilde{n}^j+\alpha_j l_2+L_2^j} \mathbb{1}_{\left\{\left(\mathcal{E}_2^j \cap \mathcal{E}_3^j\right)^c\right\}} \right] \right) \\ &\leq K^*(j) \left(\alpha_j l_2 + g(K^*(j); K^*(j))\right) \mathbb{P}\left[\left(\mathcal{E}_2^j \cap \mathcal{E}_3^j\right)^c\right].\end{aligned}$$

Combining the bounds for (III) and (IV) , we get:

$$G_2^j \leq \left(R + \frac{C}{\lambda}\right) \Big(\left(1 + K^*(j)\right) \alpha_j l_2 + \big(1 + K^*(j)\big) g(K^*(j), K^*(j)) \Big) \mathbb{P}\left[\left(\mathcal{E}_2^j \cap \mathcal{E}_3^j\right)^c\right].$$

with $g(l; K)$ defined in Proposition 2.4.5. □

Before proving the regret bound for Algorithm 1, the following remark gives an upper bound on the regret accumulated during the first $\lfloor e^{\bar{K}} \rfloor$ batches where the upper-bound of the threshold used in the phase 2 of the learning systems may be smaller than $\bar{K}$.

Remark 2.4.1. *Recall that the queue-length of each batch does not exceed $K^*(j)$ in the j^{th} batch. Following the definition of $K^*(j)$, when $j \geq \lceil e^{\bar{K}} \rceil$, $K^*(j) \geq \bar{K} + l_1 + Q_0 \geq \bar{K}$. The regret accumulated during the first $\lfloor e^{\bar{K}} \rfloor$ batches is at the most*

$$G_0 := \left(R + \frac{C}{\lambda}\right) \sum_{j=1}^{\lceil e^{\bar{K}} \rceil} \left(K^*(j) + \bar{K} + 1\right) \left(l_1 + \alpha_j l_2 + g(K^*(j); K^*(j))\right),$$

where $g(l; K)$ is defined in Proposition 2.4.5. This bound is loose since it assumes that phase 1 happens at each batch and a worst-case assumption of regret being accumulated at all times is enforced. Note that the bound is a finite function of the system parameters.

2.4.3 Proof of theorem 2.2.1

In the case that $\bar{K} > 0$, using inequality (II.19), Lemma 2.4.1, and Lemma 2.4.2, we have:

$$\begin{aligned}\sum_{j=\lceil e^{\bar{K}}\rceil}^{\lceil N/l_2\rceil} G_1^j + G_2^j \leq \left(R + \frac{C}{\lambda}\right) \sum_{j=\lceil e^{\bar{K}}\rceil}^{\lceil N/l_2\rceil} \Big(l_1^2 + \left(\bar{K}+1\right) l_1 + \left(1+K^*(j)\right) g(l_1; K^*(j))\Big) \mathbb{P}\left[\mathcal{E}_1^j\right] \\ + \left(R + \frac{C}{\lambda}\right) \sum_{j=\lceil e^{\bar{K}}\rceil}^{\lceil N/l_2\rceil} \Big((1+K^*(j))\alpha_j l_2 + \left(1+K^*(j)\right) g(K^*(j); K^*(j))\Big) \mathbb{P}\left[\left(\mathcal{E}_2^j \cap \mathcal{E}_3^j\right)^c\right].\end{aligned}$$

Substituting values/bounds for $\mathbb{P}[\mathcal{E}_1^j]$ and $\mathbb{P}[\left(\mathcal{E}_2^j \cap \mathcal{E}_3^j\right)^c]$ from Corollary 2.4.1 and Proposition 2.4.6, we get:

$$\begin{aligned}&\sum_{j=\lceil e^{\bar{K}}\rceil}^{\lceil N/l_2\rceil} G_1^j + G_2^j \\ &\leq \sum_{j=\lceil e^{\bar{K}}\rceil}^{\lceil N/l_2\rceil} \left(R + \frac{C}{\lambda}\right) \Big(l_1^2 + (\bar{K}+1)\, l_1 + (1+K^*(j)) g(l_1, K^*(j))\Big) \frac{\ln^\epsilon(j)}{j} \Big(C_1 \exp(-C_2 \ln^{1+\epsilon}(j)) + C_3 e^{-C_4 \beta_j}\Big) \\ &\quad + \sum_{j=\lceil e^{\bar{K}}\rceil}^{\lceil N/l_2\rceil} \left(R + \frac{C}{\lambda}\right) (1+K^*(j)) \Big(\alpha_j l_2 + g(K^*(j), K^*(j))\Big) \\ &\qquad \times \Big(C_1 \exp\Big(-C_2 \ln^{1+\epsilon}(j-1)\Big) + C_1 \exp\Big(-C_2 \ln^{1+\epsilon}(j)\Big) + C_3 e^{-C_4 \beta_j} + C_3 e^{-C_4 \beta_{j-1}} + \left(c_{\bar{K}}\right)^{\alpha_{j-1} l_2}\Big),\end{aligned}$$

where $g(l; K)$ is defined in Proposition 2.4.5 and is of order $O((\lambda/\mu)^K + K^2)$. Recall that $\beta_j \geq j$. All terms involved are partial sums of convergent series when α_j increases to infinity as a function bounded by polynomial in j. Therefore $\lim_{N\to\infty} G(N)$ is bounded, and the proposed algorithm achieves $O(1)$ regret in the case that $\bar{K} > 0$.

2.4.4 Proof of Theorem 2.2.2

Similarly to the proof of Theorem 2.2.1, using inequality (II.19), Lemma 2.4.1, Lemma 2.4.2, Corollary 2.4.1 and Proposition 2.4.6, we have:

$$\begin{aligned}\sum_{j=\lceil e^{\bar{K}}\rceil}^{\lceil N/l_2\rceil} G_1^j + G_2^j \le & \sum_{j=\lceil e^{\bar{K}}\rceil}^{\lceil N/l_2\rceil} \left(R+\frac{C}{\lambda}\right)(l_1^2+l_1+C_5)\frac{\ln^{\epsilon}(j)}{j} \\ & + \sum_{j=\lceil e^{\bar{K}}\rceil}^{\lceil N/l_2\rceil} \left(R+\frac{C}{\lambda}\right)(1+K^*(j))g(l_1,K^*(j))\Big(C_1\exp(-C_2\ln^{1+\epsilon}(j))+C_3\exp(-C_4\beta_j)\Big) \\ & + \sum_{j=\lceil e^{\bar{K}}\rceil}^{\lceil N/l_2\rceil} \left(R+\frac{C}{\lambda}\right)(1+K^*(j))\Big(\alpha_j l_2+g(K^*(j);K^*(j))\Big)\Big(C_1\exp(-C_2\ln^{1+\epsilon}(j))+C_3\exp(-C_4\beta_j)\Big).\end{aligned}$$

The dominant term on the RHS above is

$$\sum_{j=\lceil e^{\bar{K}}\rceil}^{\lceil N/l_2\rceil} \left(R+\frac{C}{\lambda}\right)\left(l_1^2+l_1+C_5\right)\frac{\ln^{\epsilon}(j)}{j}.$$

When N is large, we have

$$\sum_{j=2}^{\lceil N/l_2\rceil} \frac{\ln^{\epsilon}(j)}{j} = O(\ln^{1+\epsilon}(N)).$$

Hence the regret for $\bar{K}=0$ is of order $O(\ln^{1+\epsilon}(N))$.

Remark 2.4.2. *We mentioned earlier that one can adapt the analysis to the case when only the service rate is unknown or only the arrival rate is unknown by adjusting the probability of the learning system using the optimal thresholds in phase 2 and receiving similar regret bounds. As shown in the prof above, in the case when the optimal threshold is 0, the reason why the regret is* $O(\ln^{1+\epsilon}(N))$ *is that phase 1 is likely to happen infinitely often so that enough samples of the service rate can be obtained. This explicit exploration phase is necessary when the service rate is unknown. However, when only the arrival rate is unknown, the learning system would always obtain free samples for the arrival rates whether accepting customers to the queue or not. In this case, it would be unnecessary to explore explicitly, so that an* $O(1)$ *regret results similar to the case where the optimal threshold is non-zero when one always* *omits phase 1 and only the arrival rate is unknown.*

Remark 2.4.3. *The regret analysis above showed that we can obtain constant regret for the case where the optimal thresholds are non-zeros, and an $O(\ln^{1+\epsilon}(N))$ regret when 0 is an optimal threshold for any fixed $\epsilon > 0$. From the proof of Theorem 2.2.2, the order of the regret is a result of explicit exploration as it is the dominant term. One natural question is the following: can we further reduce the order of the regret in the case that 0 is an optimal threshold while preserving the constant regret in the case that the optimal threshold is non-zero, if we reduce $\mathbb{P}\left[B^j = 1\right]$, the probability of having phase 1 when the previous phase 2 uses threshold 0? Following the steps of our proof we can show that having $\mathbb{P}\left[B^j = 1\right] = \ln(\ln(j))/j$ would result in regret accumulating slower than $O(\ln^{1+\epsilon}(N))$ for any $\epsilon > 0$ in the case that 0 is a optimal threshold, and constant regret in the case that the optimal threshold is non-zero. However, this result would hold for large enough N, as the finite time performance of using $\mathbb{P}\left[B^j = 1\right] = \ln(\ln(j))/j$ may not out-perform our discussed choices for $\mathbb{P}\left[B^j = 1\right]$ as it would require j to be extremely large (but still finite) to show improved performance.*

Remark 2.4.4. *We believe that the dramatically different behaviors for our algorithm between cases when 0 is an optimal threshold, and when it is not, is fundamental to our problem owing to completely different demands in two parameter regimes: in one case, no customers should be dispatched at all, versus the other case where asymptotically a positive fraction of customers are dispatched. Hence, we conjecture that for any given learning-based dispatching algorithm the regret accumulated would grow at least at $\Omega(\ln(N))$ when the parameters are chosen in an adversarial manner. Note that our algorithm satisfies this conjecture. We will argue later on in Section 2.6 that an Upper-Confidence Bound (UCB) scheme will have a worst-case regret over parameter choices of $\Omega(\ln(N))$.*

2.5 Non-unique admittance threshold case

When the dispatcher uses a static threshold policy, the queue-length process is Markovian and ergodic. [Naor 1969] showed that the social welfare (long-term average profit in (II.1)) is maximized when using the static threshold $\bar{K}$ that uniquely satisfies (II.3) by analyzing the stationary distributions of the queue-length process for all possible static threshold policies. When (II.3) holds with equality and $\bar{K} \geq 1$, static thresholds $\bar{K}$ and $\bar{K}-1$ are both optimal, and furthermore, policies that (stochastically) alternate between the thresholds $\bar{K}$ and $\bar{K}-1$ with a fixed probability yield the same long-term average profit, i.e., are optimal for the ergodic reward maximization problem. This complicates our regret analysis as we will need to pick a specific ergodic reward-maximizing policy for our regret analysis.

In Section 2.5.1 we analyze the learned threshold; in Section 2.5.2, we introduce the specific ergodic reward maximizing genie-aided dispatcher that we will compare to, which we will label the alternating genie-aided dispatcher; and finally, Section 2.5.3 is devoted to the analysis of the regret of the learning algorithm compared to the specific genie-aided dispatcher introduced earlier.

2.5.1 Threshold used by the learning dispatcher in phase 2.

Following Algorithm 1, the threshold used by the learning dispatcher in the j^{th} phase 2 is $K(j) = \min(K^*(j), K)$, where K is the unique integer that satisfies the inequality $V(K, 1/\hat{m}, \hat{\nu}) \leq R/C < V(K+1, 1/\hat{m}, 1/\hat{\nu})$, where $\hat{m}$ is the empirical average service time, and $\hat{\nu}$ is the empirical inter-arrival time, computed using all completed services and observed arrivals before each phase 2. As mentioned earlier, the threshold is fixed throughout each phase 2. Proposition 2.2.1 implies that as long as the estimations are accurate so that inequalities (II.6) are satisfied, and when $j \geq \lceil e^{\bar{K}} \rceil$, the learning dispatcher would use a threshold in $\{\bar{K}, \bar{K}-1\}$ during the j^{th} phase 2. Proposition 2.4.1 still holds when equality holds in (II.3). Unlike in the previous case where we showed that eventually, the learning dispatcher uses the same threshold $\bar{K}$ as the genie-aided dispatcher in phase 2, we now show that as the number of batches goes to infinity, the learning algorithm will (eventually) stochastically alternate only between the thresholds $\bar{K}$ or $\bar{K}-1$. We first state the analogous of Proposition 2.4.2, Proposition 2.4.4 and Corollary 2.4.1.

Proposition 2.5.1. *Let $\hat{m}(j)$ denote the empirical service time estimated by the learning dispatcher at the beginning of phase 2 of the j^{th} batch. For the proposed algorithm, in case that $V(\bar{K},\mu,\lambda) = R/C$, we have,*

$$\mathbb{P}\left[|\hat{m}(j) = m| > \tilde{\Delta}_1\right] \leq \tilde{C}_1 \exp(-\tilde{C}_2 \ln^{1+\epsilon}(j)), \tag{II.22}$$

where

$$\begin{aligned} \tilde{C}_1 &:= \max\left\{\exp\left(-\frac{C_0(\epsilon)}{8(1+\epsilon)}\right), \frac{2\exp\left(\tilde{\Delta}_1^2/(8m^2)\right)}{\exp\left(\tilde{\Delta}_1^2/(8m^2)\right)-1}, 1\right\}, \\ \tilde{C}_2 &:= \min\left\{\frac{l_1\mu}{16(1+\epsilon)(\lambda+\mu)}, \frac{1}{8(1+\epsilon)}, \frac{l_1\mu\tilde{\Delta}_1^2}{32(1+\epsilon)m(\lambda m+1)}\right\}, \end{aligned} \tag{II.23}$$

with $\tilde{\Delta}_1 := \min\{\tilde{\delta}_1, 2m\}$, and $\tilde{\delta}_1$ is a constant for the first inequality in (II.6) *which is one part of the condition needed to reach the conclusion in* (II.7).

Proof. The proof is the same as the proof of Proposition 2.4.2, but with different constants. □

Proposition 2.5.2. *Let $\nu(j)$ denote the empirical inter-arrival time estimated by the learning dispatcher at the beginning of phase 2 of the j^{th} batch. For the proposed algorithm, in case that $V(\bar{K},\mu,\lambda) = R/C$, we have,*

$$\mathbb{P}\left[|\nu - \hat{\nu}(j)| > \tilde{\Delta}_2\right] \leq \tilde{C}_3 \exp(-\tilde{C}_4 \beta_j),$$

where

$$\tilde{C}_3 := \frac{2\exp(\tilde{\Delta}_2^2/(8\nu^2))}{\exp(\tilde{\Delta}_1^2/(8\nu^2)) - 1} \qquad \textit{and} \qquad \tilde{C}_4 := \frac{l_1\tilde{\Delta}_2^2}{8\nu^2}, \tag{II.24}$$

where β_j is defined in Proposition 2.4.4, and $\tilde{\Delta}_2 := \min\{\tilde{\delta}_2, 2\nu\}$, where $\tilde{\delta}_2$ is the constant in the second inequality in (II.6) *that is the second part needed to reach the conclusion in* (II.7).

Proof. The proof is the same as the proof of Proposition 2.4.4, but with different constants.

□

Corollary 2.5.1. *For the proposed algorithm, when $j \geq \lceil e^{\bar{K}} \rceil$, in case that $V(\bar{K}, \mu, \lambda) = R/C$,*

$$\mathbb{P}\left[\{K(j) \neq \bar{K}\} \cap \{K(j) \neq \bar{K} - 1\}\right] \leq \tilde{C}_1 \exp(-\tilde{C}_2 \ln^{1+\epsilon}(j)) + \tilde{C}_3 \exp(-\tilde{C}_4\beta_j), \tag{II.25}$$

where $\tilde{C}_1$ and $\tilde{C}_2$ are defined in (II.23) *and $\tilde{C}_3$ and C_4 are defined in* (II.24).

Proof. The proof for this proposition follows the same logic as the proof of Corollary 2.4.1, but with different constants. □

Corollary 2.5.2. *In case that $V(\bar{K}, \mu, \lambda) = R/C$, there exists a random index $\mathcal{J}$ that is finite with probability* 1*, where the learning algorithm would use threshold $\bar{K}$ or $\bar{K} - 1$ after the $\mathcal{J}^{th}$ batch.*

Proof. We show that the learning algorithm uses thresholds that are not $\bar{K}$ nor $\bar{K} - 1$ only finitely many times with probability 1. From Corollary 2.5.1, when $\bar{K} > 1$, we have:

$$\sum_{j=1}^{\infty} \mathbb{P}\left[\Big(\{K(j) = \bar{K}\} \cup \{K(j) = \bar{K} - 1\}\Big)^c\right] \leq \sum_{j=1}^{\infty} \tilde{C}_1 \exp(-\tilde{C}_2 \ln^{1+\epsilon}(j)) + \tilde{C}_3 \exp(-\tilde{C}_4 j^2)$$
$$< \infty.$$

By the Borel–Cantelli lemma (See [Durrett 2016]), we have

$$\mathbb{P}\left[\limsup_{j\to\infty}\left(\{K(j)=\bar{K}\}\cup\{K(j)=\bar{K}-1\}\right)^c\right]=0,$$

that is, with probability 1, the learning algorithm uses thresholds not in $\{\bar{K},\bar{K}-1\}$ only a finite number of times. Thus, almost surely the learning algorithm uses the optimal thresholds $\bar{K}$ and $\bar{K}-1$ after a finite random time. When $\bar{K}=1$, a similar proof holds. □

2.5.2 An alternating genie-aided dispatcher coupled with the learning dispatcher that maximizes the long-term average profit

If we compare our learning algorithm with a genie-aided system that uses a static threshold $\bar{K}$ (or alternatively $\bar{K}-1$), the regret will not be constant even when $\bar{K}>1$. The reason is that the learning dispatcher may switch between the thresholds $\bar{K}$ and $\bar{K}-1$ in different phase 2s even when $\hat{m}\in(m-\epsilon,m+\epsilon)$, where ϵ is sufficiently small. However, we can compare the queue-length process under the learning dispatcher with an optimal genie-aided dispatcher to which we refer to as the *alternating genie-aided dispatcher*: a dispatcher who may change the threshold used between $\bar{K}$ and $\bar{K}-1$ at the beginning of any busy cycle (a busy period plus an immediately following idle period). We will ensure that the threshold-changing policy of this alternating genie-aided dispatcher is adapted to the filtration generated by the queue-lengths of the two systems and the random variable B^j, with the threshold remaining unchanged during each busy cycle. It is worth mentioning that although the learning dispatcher may compute and change the threshold at the beginning of each phase 2 (which may involve multiple busy cycles), only the genie-aided dispatcher may change the threshold at the beginning of a busy cycle. This alternating genie-aided dispatcher is aware of the fact that the learning dispatcher follows Algorithm 1 and can compute the threshold learned by the learning dispatcher. This alternating genie-aided dispatcher is coupled with the learning dispatcher under the coupling described in Section 2.2.2.1. Moreover, when a customer arrives, having seen the realization of B^j, this genie-aided dispatcher is aware of whether this customer arrives during a phase 1 or 2 of the learning system, and would pick the proper threshold to use when this customer initiates a busy cycle.

Recall that K_i denotes the threshold used by the learning system at the arrival of the i^{th} customer. Following similar notation as in Section 2.2 for the alternating genie-aided dispatcher, let $\tilde{K}_i$ denote the threshold policy used at the arrival of the i^{th} customer, $\tilde{Q}_i$ denote the queue-length right before the arrival of the i^{th} customer, $\tilde{Q}(t)$ denote the queue-length at time t, τ_n^B denote the time of the beginning of the n^{th} busy cycle, $\tilde{N}_A(\tau_n^B)$ denote the index of the arrival customer who arrives at the beginning of the n^{th} busy cycle, $\tilde{N}(t)$ denote the total number of completed busy cycles up to time t, and $\tilde{K}^n$ denote the threshold used during the n^{th} busy cycle; note that $\tau_1^B = 0$. At the beginning of each busy cycle, the alternating genie-aided dispatcher then chooses a threshold $\tilde{K}^n \in \{\bar{K}, \bar{K}-1\}$, where we have

$$\tilde{K}^n = \begin{cases} \bar{K}-1, & \text{if } n = 1, \\ \bar{K}-1, & \text{if } n > 1 \text{ and } \{K_{\tilde{N}_A(\tau_n^B)} \leq \bar{K}-1 \text{ OR customer } \tilde{N}_A(\tau_n^B) \text{ arrives during phase } 1\}, \\ \bar{K}, & \text{if } n > 1 \text{ and } \{K_{\tilde{N}_A(\tau_n^B)} \geq \bar{K} \text{ AND customer } \tilde{N}_A(\tau_n^B) \text{ arrives during phase } 2\}. \end{cases} \tag{II.26}$$

That is, when the customer who initiates a busy cycle in the genie-aided system arrives during phase 1 of the learning system, the genie-aided dispatcher uses threshold $\bar{K}-1$ in the initiated busy cycle. When the customer arrives during phase 2, in the initiated busy cycle, the genie-aided dispatcher uses a threshold from $\{\bar{K}, \bar{K}-1\}$ that is closer to the threshold used by the learning system. This threshold choice would help to preserve the queue-lengths ordering under desired events, as explained in subsection 2.5.3. In other words, for customers i_1 and i_2 who arrive during the n^{th} busy cycle, i.e., $\tilde{N}_A(\tau_n^B) \leq i_1 < i_2 < \tilde{N}_A(\tau_{n+1}^B)$, we have $\tilde{K}_{i_1} = \tilde{K}_{i_2} = \tilde{K}^n$. This switching policy is adapted to the filtration generated by the queue-lengths of the genie-aided and learning systems. Since the learning algorithm always has the first exploration phase, we set $\tilde{K}^1 = \bar{K} - 1$.

The following proposition shows the optimality of the alternating genie-aided dispatcher described above using the strong law of large numbers for martingales.

Proposition 2.5.3. *Consider a dispatcher who uses a static threshold policy, either $\bar{K}$ or $\bar{K}-1$, during a busy cycle, and may switch between these two thresholds only at the beginning of a busy cycle following the switching rule described in* (II.26). *The long-term average profit of the system under this dispatcher is the same as a dispatcher using either one of the static thresholds $\bar{K}$ or $\bar{K}-1$.*

Proof. Assume the initial queue-length is some $a \in \{0, 1, \dots, \bar{K}\}$, where the particular value doesn't impact the asymptotic results. We are interested in finding:

$$\begin{aligned}
&\liminf_{t\to\infty} \frac{1}{t}\left(aR + \sum_{i=1}^{\tilde{N}_A(t)} R\mathbb{1}_{\{\tilde{Q}_i \le \tilde{K}_i\}} - \int_0^t C\tilde{Q}(u)du\right)\\
&= \liminf_{t\to\infty} \frac{1}{t}\left(aR + \sum_{i=1}^{\tilde{N}_A(\tau_2^B)-1} R\mathbb{1}_{\{\tilde{Q}_i \le \tilde{K}^1\}} - \int_0^{\tau_2^B} C\tilde{Q}(u)du\right)\\
&\qquad + \liminf_{t\to\infty} \frac{1}{t}\left(\sum_{n=2}^{\tilde{N}(t)}\left(\sum_{i=\tilde{N}_A(\tau_n^B)}^{\tilde{N}_A(\tau_{n+1}^B)-1} R\mathbb{1}_{\{\tilde{Q}_i \le \tilde{K}^n\}} - \int_{\tau_n^B}^{\tau_{n+1}^B} C\tilde{Q}(u)du\right)\right)\\
&\qquad + \liminf_{t\to\infty} \frac{1}{t}\left(\sum_{i=\tilde{N}_A\left(\tau^B_{\tilde{N}(t)+1}\right)}^{\tilde{N}_A(t)} R\mathbb{1}_{\{\tilde{Q}_i \le \tilde{K}^{\tilde{N}(t)+1}\}} - \int_{\tau^B_{\tilde{N}(t)+1}}^{t} C\tilde{Q}(u)du\right). \qquad \text{(II.27)}
\end{aligned}$$

Let the tuple $(X_n, \mathcal{B}_n)$ denote the total net profit and duration of the n^{th} busy cycle under this dispatcher. For the first busy cycle, we have:

$$X_1 := aR + \sum_{i=1}^{\tilde{N}_A(\tau_2^B)-1} R\mathbb{1}_{\{\tilde{Q}_i \le \tilde{K}^1\}} - \int_0^{\tau_2^B} C\tilde{Q}(u)du, \text{ and } \mathcal{B}_1 := \tau_2^B.$$

For $n \geq 2$, we have:

$$X_n := \sum_{i=\tilde{N}_A(\tau_n)}^{\tilde{N}_A(\tau_{n+1}^B)-1} R\mathbb{1}_{\{\tilde{Q}_i \leq \tilde{K}^n\}} - \int_{\tau_n^B}^{\tau_{n+1}^B} C\tilde{Q}(u)du, \text{ and } \mathcal{B}_n := \tau_{n+1}^B - \tau_n^B.$$

We can rewrite (II.27) as:

$$\liminf_{t\to\infty} \frac{1}{t} \sum_{n=2}^{\tilde{N}(t)} X_n + \liminf_{t\to\infty} \frac{1}{t} \left(X_1 + \sum_{i=\tilde{N}_A(\tau_{\tilde{N}(t)+1}^B)}^{\tilde{N}_A(t)} R\mathbb{1}_{\{\tilde{Q}_i \leq \tilde{K}^{\tilde{N}(t)+1}\}} - \int_{\tau_{\tilde{N}(t)+1}^B}^{t} C\tilde{Q}(u)du \right).$$

When the initial queue-length is finite, $\mathbb{E}[\mathcal{B}_1]$ and $\mathbb{E}[(\mathcal{B}_1)^2]$ are finite; see [Takagi and Tarabia 2009].

Let $(Y_n^{\bar{K}}, \mathcal{B}_n^{\bar{K}})$ denote the total net profit and the duration of the n^{th} busy cycle of a dispatcher that uses static threshold $\bar{K}$ and with initial queue-length 1, and let $\mathcal{Y}^{\bar{K}}(t)$ denote the accumulated total net profit of this dispatcher up to time t. Setting the initial queue-length to 1 is owing to a generic busy cycle starting as such. The random variables $(Y_n^{\bar{K}}, \mathcal{B}_n^{\bar{K}})$ are *i.i.d.*, and $\mathcal{Y}^{\bar{K}}(t)$ is a renewal reward process: see [Durrett 2016, Section 3.1]. Similarly, we can define $(Y_n^{\bar{K}-1}, \mathcal{B}_n^{\bar{K}-1})$ and $\mathcal{Y}^{\bar{K}-1}(t)$ for a dispatcher that uses static threshold $\bar{K}-1$. [Naor 1969] showed that there exists a constant $\mathcal{O}$ denoting the optimal long-term average profit of the dispatcher, where with probability 1,

$$\lim_{t\to\infty} \frac{1}{t}\mathcal{Y}^{\bar{K}}(t) = \lim_{t\to\infty} \frac{1}{t}\mathcal{Y}^{\bar{K}-1}(t) = \mathcal{O}.$$

By the renewal-reward theorem, [Durrett 2016, Section 3.1], we have:

$$\mathbb{E}\left[Y_1^{\bar{K}}\right] = \mathbb{E}\left[\mathcal{B}_1^{\bar{K}}\right]\mathcal{O}, \quad \text{and} \quad \mathbb{E}\left[Y_1^{\bar{K}-1}\right] = \mathbb{E}\left[\mathcal{B}_1^{\bar{K}-1}\right]\mathcal{O}.$$

Let $\tilde{\mathcal{F}}_{n-1} := \tilde{\mathcal{F}}_{\tau_n}$ denote the sigma-algebra generated by the queue-length process of the coupled learning dispatcher and the dispatcher described in Proposition 2.5.3 up to time τ_n^B (the end of the $(n-1)^{th}$ busy cycle of the dispatcher described in Proposition 2.5.3). By the independence of the Poisson arrival and Poisson potential service process, the distribution of $(X_n, \mathcal{B}_n)$ conditioned on $\tilde{\mathcal{F}}_{n-1}$ is the same as the distribution of $(X_n, \mathcal{B}_n)$ conditioned on the filtration generated by $\tilde{K}^n$. Moreover, for $n \geq 2$, $(X_n, \mathcal{B}_n)$ conditioned on the event $\{\tilde{K}^n = \bar{K}\}$ has the same distribution as $(Y_1^{\bar{K}}, \mathcal{B}_1^{\bar{K}})$ and $(X_n, \mathcal{B}_n)$ conditional on the event $\{\tilde{K}^n = \bar{K}-1\}$ has the same distribution as $(Y_1^{\bar{K}-1}, \mathcal{B}_1^{\bar{K}-1})$. Using these, for $i \geq 2$, we have:

$$\begin{aligned}\mathbb{E}\left[\mathcal{B}_n\right] &= \mathbb{E}\left[\mathcal{B}_n \middle| \tilde{K}^n = \bar{K}\right] \mathbb{P}\left[\tilde{K}^n = \bar{K}\right] + \mathbb{E}\left[\mathcal{B}_n \middle| \tilde{K}^n = \bar{K}-1\right] \mathbb{P}\left[\tilde{K}^n = \bar{K}-1\right] \\ &= \mathbb{E}\left[\mathcal{B}_1^{\bar{K}}\right] \mathbb{P}\left[\tilde{K}^n = \bar{K}\right] + \mathbb{E}\left[\mathcal{B}_1^{\bar{K}-1}\right] \mathbb{P}\left[\tilde{K}^n = \bar{K}-1\right],\end{aligned}$$

and similarly,

$$\begin{aligned}\mathbb{E}\left[(\mathcal{B}_n)^2\right] &= \mathbb{E}\left[(\mathcal{B}_n)^2 \middle| \tilde{K}^n = \bar{K}\right] \mathbb{P}\left[\tilde{K}^n = \bar{K}\right] + \mathbb{E}\left[(\mathcal{B}_n)^2 \middle| \tilde{K}^n = \bar{K}-1\right] \mathbb{P}\left[\tilde{K}^n = \bar{K}-1\right] \\ &= \mathbb{E}\left[(\mathcal{B}_1^{\bar{K}})^2\right] \mathbb{P}\left[\tilde{K}^n = \bar{K}\right] + \mathbb{E}\left[(\mathcal{B}_1^{\bar{K}-1})^2\right] \mathbb{P}\left[\tilde{K}^n = \bar{K}-1\right].\end{aligned}$$

Both $\mathcal{B}_1^{\bar{K}}$ and $\mathcal{B}_1^{\bar{K}-1}$ have finite first and second moments, [Takagi and Tarabia 2009], and thus, so does $\mathcal{B}_i$.

Let $\tilde{N}_{\text{join}}^n$ denote the number of the customers joining the queue during the n^{th} busy cycle under the dispatching policy described in Proposition 2.5.3. Observe that the total number of arrivals joining the queue and services are equal during a busy cycle except for the first one for which there are exactly a more service completions than the number of customers joining the queue during the first busy cycle. When there are at least $\bar{K}$ potential services between two consecutive arrivals, the queue-length under the dispatcher described in Proposition 2.5.3

hits 0 and a busy period ends. Therefore, for any integer M, we have:

$$\mathbb{P}\left[\tilde{N}^n_{\text{join}} > M\right] \leq \left(1 - \left(\frac{\mu}{\lambda+\mu}\right)^{\bar{K}}\right)^M,$$

which then implies that the random variable $\tilde{N}^i_J$ has finite first and second moments.

Since $|X_n| \leq R\tilde{N}^n_{\text{join}} + C\bar{K}\mathcal{B}_n$, a.s., for all $n \geq 2$, and $|X_1| \leq R\tilde{N}^1_{\text{join}} + aR + C\bar{K}\mathcal{B}_1$ a.s., we can conclude that X_n also has finite first and second moments, and it is clear that with probability 1,

$$\liminf_{t\to\infty} \frac{1}{t}\left(X_1 + \sum_{n=\tilde{N}_A\left(\tau^B_{\tilde{N}(t)+1}\right)}^{\tilde{N}_A(t)} R\mathbb{1}_{\{\tilde{Q}_i \leq \tilde{K}^{\tilde{N}(t)+1}\}} - \int_{\tau^B_{\tilde{N}(t)+1}}^t C\tilde{Q}(u)du\right) = 0.$$

For almost every sample path, there exists t^* such that $\tilde{N}(t) > 1$ for all $t \geq t^*$, and we have the following upper and lower bounds with probability 1:

$$\liminf_{t\to\infty} \frac{1}{\sum_{n=1}^{\tilde{N}(t)+1} \mathcal{B}_i} \sum_{n=2}^{\tilde{N}(t)} X_n \leq \liminf_{t\to\infty} \frac{1}{t} \sum_{n=2}^{\tilde{N}(t)} X_n \leq \liminf_{t\to\infty} \frac{1}{\sum_{n=2}^{\tilde{N}(t)} \mathcal{B}_n} \sum_{i=2}^{\tilde{N}(t)} X_n.$$

We show $\liminf_{t\to\infty}(1/t)\sum_{n=2}^{\tilde{N}(t)} X_n = \mathcal{O}$ a.s. by showing that with probability 1, both

$$\liminf_{t\to\infty} \frac{1}{\sum_{n=1}^{\tilde{N}(t)+1} \mathcal{B}_n} \sum_{n=2}^{\tilde{N}(t)} X_n = \mathcal{O}, \text{ and} \tag{II.28}$$

$$\liminf_{t\to\infty} \frac{1}{\sum_{n=2}^{\tilde{N}(t)} \mathcal{B}_n} \sum_{n=2}^{\tilde{N}(t)} X_n = \mathcal{O}. \tag{II.29}$$

Note that we have:

$$\liminf_{t\to\infty} \frac{1}{\sum_{n=1}^{\tilde{N}(t)+1} \mathcal{B}_n} \sum_{n=2}^{\tilde{N}(t)} X_n = \liminf_{t\to\infty} \frac{\sum_{n=2}^{\tilde{N}(t)} \mathcal{B}_n}{\sum_{n=1}^{\tilde{N}(t)+1} \mathcal{B}_n} \frac{1}{\sum_{n=2}^{\tilde{N}(t)} \mathcal{B}_n} \sum_{n=2}^{\tilde{N}(t)} X_n$$
$$= \liminf_{t\to\infty} \frac{\tilde{N}(t)+1}{\sum_{n=1}^{\tilde{N}(t)+1} \mathcal{B}_n} \times \frac{\sum_{n=2}^{\tilde{N}(t)} \mathcal{B}_n}{\tilde{N}(t)-1} \times \frac{\tilde{N}(t)-1}{\tilde{N}(t)+1} \times \frac{1}{\sum_{n=2}^{\tilde{N}(t)} \mathcal{B}_n} \sum_{n=2}^{\tilde{N}(t)} X_n.$$

We can also rewrite (II.29) as

$$\liminf_{n\to\infty} \frac{\tilde{N}(t)-1}{\sum_{n=2}^{\tilde{N}(t)} \mathcal{B}_n} \frac{1}{\tilde{N}(t)-1} \sum_{n=2}^{\tilde{N}(t)} (X_n - \mathcal{B}_n \mathcal{O}) = 0.$$

Note that $\lim_{t\to\infty} \tilde{N}(t) = \infty$ and $\lim_{t\to\infty} \sum_{n=2}^{\tilde{N}(t)} \mathcal{B}_n = \infty$ a.s., which in turn imply that a.s. we have:

$$\liminf_{t\to\infty} \frac{\tilde{N}(t)+1}{\sum_{n=1}^{\tilde{N}(t)+1} \mathcal{B}_n} = \liminf_{k\to\infty} \frac{k}{\sum_{n=1}^{k} \mathcal{B}_n} = \liminf_{t\to\infty} \frac{\tilde{N}(t)-1}{\sum_{n=2}^{\tilde{N}(t)} \mathcal{B}_n} \text{ and } \lim_{t\to\infty} \frac{\tilde{N}(t)-1}{\tilde{N}(t)+1} = \lim_{k\to\infty} \frac{k-1}{k+1} = 1.$$

Then, in order to establish (II.28) and (II.29), it is sufficient to show that with probability 1,

$$\liminf_{k\to\infty} \frac{1}{k-1} \sum_{n=2}^{k} (X_n - \mathcal{B}_n \mathcal{O}) = 0, \text{ and} \tag{II.30}$$

$$0 < \liminf_{k\to\infty} \frac{k}{\sum_{n=1}^{k} \mathcal{B}_n} \le \limsup_{k\to\infty} \frac{k}{\sum_{n=1}^{k} \mathcal{B}_n} < \infty. \tag{II.31}$$

We will prove (II.30) by using the strong law of large numbers for martingales [Csörgő 1968, Theorem 1]. Let $M_k = \sum_{n=2}^{k} (X_n - \mathcal{B}_n \mathcal{O})$ for $k \ge 2$, $M_1 = 0$. Clearly $\mathbb{E}[|M_k|] < \infty$ for all k. Also,

$$\begin{aligned}\mathbb{E}\left[M_{k+1} - M_k \middle| \tilde{\mathcal{F}}_k\right] &= \mathbb{E}\left[X_{k+1} - \mathcal{B}_{k+1}\mathcal{O} \middle| \tilde{\mathcal{F}}_k\right] \\ &= \mathbb{E}\left[X_{k+1} - \mathcal{B}_{k+1}\mathcal{O} \middle| \tilde{K}^k\right] \\ &= \mathbb{1}_{\{\tilde{K}^{k+1} = \bar{K}\}} \mathbb{E}\left[Y_1^{\bar{K}} - \mathcal{B}_1^{\bar{K}} \mathcal{O}\right] + \mathbb{1}_{\{\tilde{K}^{k+1} = \bar{K}-1\}} \mathbb{E}\left[Y_1^{\bar{K}-1} - \mathcal{B}_1^{\bar{K}-1} \mathcal{O}\right] = 0.\end{aligned} \tag{II.32}$$

The second equality follows since the distribution of $(X_n, \mathcal{B}_n)$ conditioned on $\tilde{\mathcal{F}}_{n-1}$ is the same as the distribution of $(X_n, \mathcal{B}_n)$ conditioned on the filtration generated by $\tilde{K}^n$ for all $n \geq 2$. Therefore, we have shown that M_k is a martingale with respect to filtration $\{\tilde{\mathcal{F}}_k\}_{k\geq 1}$ with martingale difference sequence $X_k - \mathcal{B}_k\mathcal{O}$ for $k \geq 2$.

Next, we will show that $\sum_{k=2}^{\infty} k^{-2}\mathbb{E}\left[(X_k - \mathcal{B}_k\mathcal{O})^2\right]$ is finite. For $k \geq 2$, we have:

$$\begin{aligned}
\mathbb{E}\left[(X_k - \mathcal{B}_k\mathcal{O})^2\right] &= \mathbb{E}\left[\left(\sum_{i=\tilde{N}_A(\tau_k^B)}^{\tilde{N}_A(\tau_{k+1}^B)-1} R\mathbb{1}_{\{\tilde{Q}_i \leq \tilde{K}^k\}} - \int_{\tau_k^B}^{\tau_{k+1}^B} C\tilde{Q}(u)du - \mathcal{B}_n\mathcal{O}\right)^2\right] \\
&\leq \mathbb{E}\left[\left(\sum_{i=\tilde{N}_A(\tau_k^B)}^{\tilde{N}_A(\tau_{k+1}^B)-1} R\mathbb{1}_{\{\tilde{Q}_i \leq \tilde{K}^k\}}\right)^2 + \left(\int_{\tau_k^B}^{\tau_{k+1}^B} C\tilde{Q}(u)du + \mathcal{B}_n O\right)^2\right] \\
&\leq \mathbb{E}\left[R^2(\tilde{N}^k_{\text{join}})^2 + (\mathcal{B}_k)^2(\mathcal{O} + C\bar{K})^2\right],
\end{aligned}$$

where we recall that $\tilde{N}^k_{\text{join}}$ denotes the customers joining the queue during the k^{th} busy cycle, and $\mathcal{B}_k = \tau_{k+1}^B - \tau_k^B$ is the duration of the k^{th} busy cycle. When $k \geq 2$, both $\tilde{N}^k_{\text{join}}$ and $\mathcal{B}_k$ have finite second moments that do not depend on k, so that $\sum_{k=2}^{\infty} k^{-2}\mathbb{E}\left[(X_k - \mathcal{B}_k\mathcal{O})^2\right] < \infty$. Therefore, by the strong law of large numbers for martingales [Csörgő 1968, Theorem 1], (II.30) holds.

Next, we prove (II.31). Consider a dispatcher that uses the static threshold policy $\bar{K}$, which is coupled with the dispatcher described in Proposition 2.5.3, and also has initial queue-length a. The duration of the n^{th} busy cycles of this dispatcher is denoted $\tilde{\mathcal{B}}_n^{\bar{K}}$. The random variables $\tilde{\mathcal{B}}_n^{\bar{K}}$s are *i.i.d.* for all $n \geq 2$. Although having a different distribution, $\tilde{\mathcal{B}}_1^{\bar{K}}$ is independent of $\tilde{\mathcal{B}}_n^{\bar{K}}$ for all $n \geq 2$.

Using Proposition 2.3.1, observe that on any sample path, when the dispatcher that uses the static threshold $\bar{K}$ has experienced k busy periods, the dispatcher described in Proposition 2.5.3 would have experienced more than k busy periods. Thus, we can conclude that, with probability 1,

$$\sum_{n=1}^{k} \tilde{\mathcal{B}}_i^{\bar{K}} \geq \sum_{n=1}^{k} \mathcal{B}_k,$$

for all k. Moreover, since $\mathcal{B}_n^{\bar{K}}$s have finite first moments, [Takagi and Tarabia 2009], and are non-negative, they are finite a.s. Therefore, $\lim_{k\to\infty} k/\sum_{n=1}^{k} \tilde{\mathcal{B}}_n^{\bar{K}} = 1/\mathbb{E}[\mathcal{B}_2^{\bar{K}}]$ exists a.s. and is strictly positive. Therefore, with probability 1, we have:

$$\liminf_{k\to\infty} \frac{k}{\sum_{n=1}^{k} \mathcal{B}_n} \geq \lim_{k\to\infty} \frac{k}{\sum_{n=1}^{k} \tilde{\mathcal{B}}_n^{\bar{K}}} = \frac{1}{\mathbb{E}[\mathcal{B}_2^{\bar{K}}]} > 0.$$

Similarly, comparing with the dispatcher using static threshold policy $\bar{K}-1$ that is coupled with the genie-aided dispatcher described in Proposition 2.5.3, with probability 1, we have:

$$\limsup_{k\to\infty} \frac{k}{\sum_{n=1}^{k} \mathcal{B}_n} \leq \lim_{k\to\infty} \frac{k}{\sum_{n=1}^{k} \tilde{\mathcal{B}}_n^{\bar{K}-1}} = \frac{1}{\mathbb{E}[\mathcal{B}_2^{\bar{K}-1}]} < \infty.$$

The last two results imply (II.31). Then, (II.31) and (II.30) prove the desired result. □

Remark 2.5.1. *When there exists a unique optimal threshold policy, the definition of regret is straightforward and without any ambiguity. However, in the case where there are multiple optimal threshold policies, we need to define the regret with respect to one of the optimal policies. Proposition 2.5.3 shows that the alternating genie-aided system is asymptotically optimal for almost all sample paths in the sense that it achieves the same long-term average profit as the system that uses either static threshold $\bar{K}$ or $\bar{K}-1$ starting from the beginning. The* total *net profit achieved by this alternating genie-aided system up to time T is not necessarily equal to the total net profit achieved by the genie-aided system using static threshold $\bar{K}$ or $\bar{K}-1$. These three policies (including the two static policies) do not necessarily achieve the same net profit up to time T on given sample paths of the arrival and service processes. Note that by Propositoin 2.3.1, the net profit process of the alternating genie-aided system during any busy cycle is either the same as the gain of one of the systems* *using static thresholds $\bar{K}$ and $\bar{K} - 1$ or the net profit during the busy cycle is no smaller than the gain in the system using the static threshold $\bar{K}$: consider the case that the alternating system switches from using threshold $\bar{K} - 1$ to $\bar{K}$, and the queue-length hits $\bar{K}$ during the current busy cycle. This is the only case where the behavior of the alternating genie-aided system may be different from the two systems using a static threshold. However, during the time between the switch and the time that the queue length of the alternating system hits $\bar{K}$ in the current busy cycle, the queue-length of the system using threshold $\bar{K}$ is greater than or equal to the queue-length of the alternating system. Moreover, the number of customers being served is the same for these two systems (in the current busy cycle). A similar but opposite comparison can be made with the system using static threshold $\bar{K} - 1$. In fact, the total net profit achieved (as a function of time) by the two systems using the static thresholds $\bar{K}$ and $\bar{K} - 1$, respectively, are not necessarily equal on given sample paths of the arrival and service processes either. We expect that the difference between the net profit of pairs of such systems obeys a Central-Limit Theorem behavior (including a functional form of the Central-Limit Theorem) when appropriately normalized and scaled (in time).*

Take as a concrete example the situation where $\bar{K} = 1$ and $\bar{K} - 1 = 0$ are both optimal thresholds and assume that the initial queue length is 0 for both systems. Using the inequalities in II.3, we get that these two optimal thresholds only occur when $C/\mu = R$. The system that uses the static threshold 0 does not admit any customers into the system, and clearly achieves a total net profit equal to 0 for any time T. The system that uses the static threshold 1 admits a customer in the queue if and only if the system is empty when this customer arrives. The busy periods of this system using the static threshold 1 are exactly the periods when a single customer is served, and the expected net profit during any busy period of this system is $R - C/\mu = 0$. However, this does not imply that the total net profit up to time T of the system using threshold 1 is 0. In fact, the difference of the total net profit between these two systems over the busy periods of the system using threshold 1 is a sum of mean-zero random variables *(with each random variable being $R - C \times S$ where $S \sim \mathrm{EXP}(\mu)$ is the service time of the customer-in-service), which, intuitively, will lead to the claimed Central-Limit Theorem behavior. Furthermore, by the (finite-time) Law of the Iterated Logarithm [Balsubramani 2015], along (almost all) sample paths the difference of the total net profit of the two systems may grow at most as $O(\sqrt{T\ln(\ln(T))})$ (with high probability).*

For this example, we can also carry out an explicit analysis of $\mathbb{E}[\mathcal{G}(t)]$, the expected total net profit up to any time t of the system using static threshold 1. With the assumption that the initial queue length is 0, it is easier to consider the busy cycle as the idle period together with the consecutive busy period. Let $(Y_n^1, \mathcal{B}_n^1)$ denote the total net profit and the duration of the n^{th} busy cycle of the dispatcher that uses threshold 1. As mentioned in the previous paragraph, $\mathbb{E}[Y_n^1] = 0$ for all n. The random variables $\mathcal{B}_n^1$ are i.i.d. and have the same distribution as $A + S$, where A is an $\mathrm{EXP}(\lambda)$ random variable and S is an $\mathrm{EXP}(\mu)$ random variable independent of A. Let $N(t)$ denote the number of completed busy cycles until time t, $n(t) = \mathbb{E}\left[N(t)\right]$ denote the expected number of completed busy cycles up to time t, $\sigma_s(t)$ denote the residual service time of the current busy cycle at time t, and $\tau_t = \sum_{n=1}^{N(t)+1} \mathcal{B}_n^1$ denote the end-time of the current busy cycle. Recalling that the reward R is given to the dispatcher at each service completion, we have:

$$\mathbb{E}\left[\mathcal{G}(t)\right] = \mathbb{E}\left[\mathcal{G}(\tau_t)\right] - R + C\mathbb{E}[\sigma_s(t)].$$

Note that $n(t)$ is the renewal function of the associated (alternating) renewal process with renewal interval distributed the same as $A + S$. By standard renewal theory arguments, $n(t)$ is finite for all t, and $N(t) + 1$ is a stopping time of the sequence $(Y_n^1, \mathbb{B}_n^1)$. Applying Wald's *equality, we get*

$$\mathbb{E}[\mathcal{G}(\tau_t)] = \mathbb{E}\left[\sum_{i=1}^{N(t)+1} Y_i^1\right] = \mathbb{E}\left[N(t)+1\right]\mathbb{E}\left[Y_1^1\right] = 0.$$

Note that the distribution of $\sigma_s(t)$ follows $\mathrm{EXP}(\mu)$: if at time t the busy period has not started yet, clearly the residual service time is an $\mathrm{EXP}(\mu)$ random variable. If there is a customer being served at time t, the busy cycle ends at the completion of this service. Using the memory-less property of exponential random variable, the residual service time is again an $\mathrm{EXP}(\mu)$ random variable.

Then, using $\mathbb{E}[\mathcal{G}(\tau_t)] = 0$, we get:

$$\mathbb{E}\left[\mathcal{G}(t)\right] = \mathbb{E}\left[\mathcal{G}(\tau_t)\right] - R + C\mathbb{E}[\sigma_s(t)] = 0 - R + C/\mu = 0.$$

Despite admitting a customer when the queue is empty, the expected net profit at any time is exactly 0 for the dispatcher using static threshold 1 when both $\bar{K} = 1$ and $\bar{K} - 1 = 0$ are optimal thresholds. We expect that a similar but more complicated computation using renewal theory (as the memory-less argument no longer holds for the busy period, which is now a phase-type distribution, plus we need to determine the remaining workload to be served) can be carried out for systems using threshold $\bar{K} > 1$ and $\bar{K} - 1 > 0$, when both are optimal thresholds. We expect that as $t \to \infty$, the expected total net profit of the two systems using static thresholds differ by at most a constant, and so is the difference of the expected total net profit of the alternating system and the two systems using a static threshold. These questions are outside the scope of the paper and are left for future research.

2.5.3 Regret analysis with respect to the alternating genie-aided dispatcher.

In Proposition 2.5.3 we proved that the alternating genie-aided dispatcher described in Section 2.5.2 that uses $\bar{K}$ and $\bar{K}-1$ “in favor" of the learning algorithm is optimal for (II.1). Next, we bound the regret of the learning dispatcher when compared with this genie-aided dispatcher.

Recall from Section 2.5.2 that $\tilde{K}_i$ denotes the threshold used by the alternating genie-aided dispatcher at the arrival of the i^{th} arriving customer. Following (II.12), we have:

$$G(t) \leq \left(R + \frac{C}{\lambda}\right)\mathbb{E}\left[\sum_{i=1}^{N_A(t)} \left|\mathbb{1}_{\{\tilde{Q}_i < \tilde{K}_i\}} - \mathbb{1}_{\{Q_i < K_i\}}\right| + |\tilde{Q}_i - Q_i|\right].$$

Similar to the earlier analysis, assuming that both systems start with the same initial queue-length, we use $\tilde{G}_1^j$ to denote the expected regret accumulated during the (potential) phase 1 and the first time the queue is emptied in the consecutive phase 2 for the j^{th} batch. Again, we use $\tilde{G}_2^j$ to denote the expected regret accumulated in the remainder of (the phase 2 of the) j^{th} batch.

Set $\tilde{\mathcal{E}}_2^j := \{K(j) = \bar{K}\} \cup \{K(j) = \bar{K} - 1\}$. We will reuse the events $\mathcal{E}_1^j$ and $\mathcal{E}_3^j$ that were first introduced in Section 2.4. Recall that $\mathcal{E}_1^j$ denotes the event that phase 1 of the j^{th} batch happens, and $\mathcal{E}_3^j = \{Q_{n^j} = \tilde{Q}_{n^j}\}$ denotes the event that at the beginning of the j^{th} phase 2 of the learning system, the queue-length of the two systems are the same.

Only under the event $\mathcal{E}_1^j$ there is a regret contribution to $\tilde{G}_1^j$ (since otherwise phase 1 of the j^{th} batch is omitted, and the queue-length at the beginning of phase 2 is 0). Under the event $(\mathcal{E}_1^j)^c \cap \tilde{\mathcal{E}}_2^j \cap \mathcal{E}_3^j$, there is no regret contribution to $\tilde{G}_2^j$: indeed, for this batch of customers, $\tilde{\mathcal{E}}_2^j$ ensures the learned threshold is either $\bar{K}$ or $\bar{K}-1$. The event $(\mathcal{E}_1^j)^c$ ensures that phase 1 is omitted, so the queue-length at the beginning of this phase 2 of the learning system is 0. Moreover, $\mathcal{E}_3^j$ ensures that the queue-length of the alternating genie-aided system is also 0 at this time, which means that the arrival of the first customer of this phase 2 initiates a busy cycle for both systems. In this case, the alternating genie-aided system would pick the same threshold used as the learning system for all the busy cycles in this phase 2. Both systems would make the same choices of admitting each arrival in this phase 2, and the queue-length processes of the two systems would also coincide for the entire phase 2. Under the event $\mathcal{E}_1^j \cap \tilde{\mathcal{E}}_2^j \cap \mathcal{E}_3^j$, although phase 1 happens, Proposition 2.3.1 tells us that the queue-length of the learning system at the end of phase 1 is no smaller than the queue-length of the genie-aided system. The event $\tilde{\mathcal{E}}_2^j$ ensures that the threshold used by the learning system during the entire phase 2 is no smaller than the threshold used by the genie-aided system (since the genie-aided system would be either using the same threshold as the learning system when a busy cycle is initiated by a customer who arrives during phase 2 or using threshold $\bar{K}-1$ when a busy cycle is initiated by a customer who arrives during phase 1), when the queue-length of the learning system hits 0 for the first time after phase 1, the queue-length of the genie-aided system also hits 0. The next proposition gives a bound that holds in the current setting for the probability of $\left(\tilde{\mathcal{E}}_2^j \cap \mathcal{E}_3^j\right)^c$.

Proposition 2.5.4. *Fix $j \geq \lceil e^{\bar{K}} \rceil$. In case that $V(\bar{K}, \mu, \lambda) = R/C$, we have the following:*

$$\mathbb{P}\left[\left(\tilde{\mathcal{E}}_2^j \cap \mathcal{E}_3^j\right)^c\right] \leq \tilde{C}_1 \exp\Big(-\tilde{C}_2 \mathrm{ln}^{1+\epsilon}(j)\Big) + \tilde{C}_1 \exp\Big(-\tilde{C}_2 \ln^{1+\epsilon}(j-1)\Big) \\ + \tilde{C}_3 \exp\Big(-\tilde{C}_4 \beta_j\Big) + \tilde{C}_3 \exp\Big(-\tilde{C}_4 \beta_{j-1}\Big) + \big(c_{\bar{K}}\big)^{\alpha_{j-1} l_2}.$$

$\tilde{C}_1$, $\tilde{C}_2$, $\tilde{C}_3$ and $\tilde{C}_4$ are defined in (II.23) *and* (II.24)*, and*

$$c_{\bar{K}} := 1 - \left(\frac{\mu}{\lambda + \mu}\right)^{\bar{K}} \in (0, 1).$$

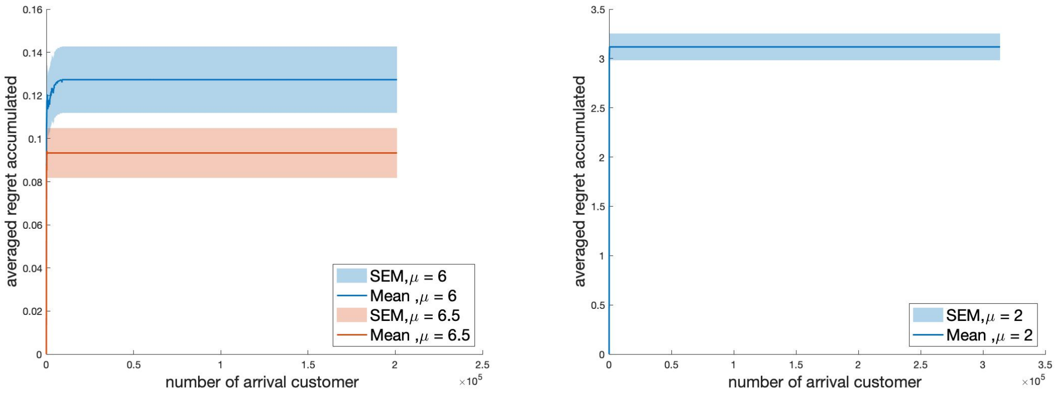
Proof. The proof for both cases $\bar{K} > 1$ and $\bar{K} = 1$ follows the same logic as in the case $\bar{K} > 0$ in Proposition 2.4.6. □

Since we are using l_1, $K^*(j)$ and $\bar{K}$ to bound the queue-length in the proof of Lemmas 2.4.1 and 2.4.2, these two lemmas still hold when the optimal threshold is not unique. It should be now clear that Theorem 2.2.1 and Theorem 2.2.2 also hold when equality holds in (II.3).

2.6 Simulation-based numerical results

In this section, we demonstrate the performance of our proposed Algorithm 1 using simulations. To compute the regret we compare our algorithm to the genie-aided system that has the knowledge of the arrival and service rates and uses the optimal strategy proposed by [Naor 1969]. For the simulations, we set the initial queue-length to be 0 for both the genie-aided and learning systems. For all numerical experiments, unless specified otherwise, we use the following set of parameters: $l_2 = 10$, $C = R = 1$, $\mathbb{E}[B^j] = \ln(j)/j$, $\alpha_j = j$ where recall that l_2 is the minimum length of phase 2, C is the cost per unit time, R is the reward granted to the dispatcher when each service completes, B^j is the random variable which controls the probability of having phase 1 when the threshold used in the previous phase 2 is 0, and α_j is the rate at which the minimum length of phase 2 increases. Note that, unless specified otherwise, we use $\epsilon = 1$ in $\mathbb{E}[B^j] = \ln^{\epsilon}(j)/j$. We vary μ and λ for different experiments, and explore zero and non-zero optimal threshold cases, as well as the cases where the optimal threshold is unique and when it is not unique. To show the pattern of the regret within a reasonable number of arriving customers, when the largest optimal threshold is 0, we use $l_1 = 1$ and when the largest optimal threshold is positive, we use $l_1 = 3$, where l_1 is the length of phase 1 (when used), and stays unchanged for all batches. Our theoretical analysis holds for arbitrary choices of the constants $l_1 \geq 1$. However, when l_1 is large and the service rate is small, it will take a long time for the queue to empty during phase 2, and therefore, will require more arrivals to show the correct asymptotic behavior of the regret.

The finite-time performance of the simulated results agrees qualitatively with our upper bound: when an optimal strategy is to use threshold 0, the learning system achieves an expected regret that grows in a sub-linear manner; and when all optimal strategies use a non-zero threshold, the learning system achieves an $O(1)$ expected regret.



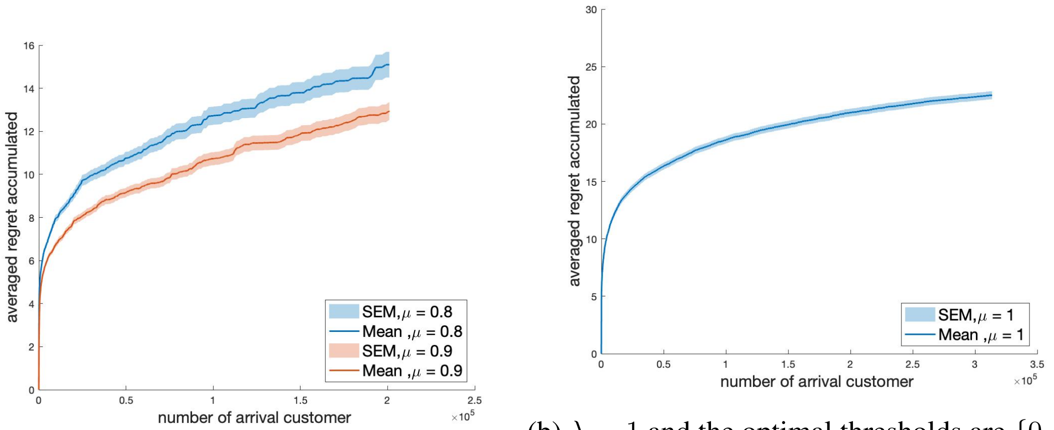
(a) $\lambda = 1$, $R = 1$, and the optimal threshold is $\bar{K} = 5$.

(b) $\lambda = 1$, $R = \frac{129}{32}$, and the optimal thresholds $\{4, 5\}$ ($\bar{K} = 5$).

Figure II.1: Regret of the learning system when all optimal thresholds are positive. We set $C = 1$, $\mathbb{E}[B^j] = \ln(j)/j$, $K^*(j) \sim \ln(j)$, and $\alpha_j = j$.

Expected regret with non-zero optimal thresholds: Figure II.1a shows the variation of the (expected) regret with respect to the number of arrivals for $\mu = 6$ and $\mu = 6.5$ when $l_1 = 3$ and $\lambda = 1$. The regret is averaged over 1000 simulations and there are more than $2 * 10^5$ customers arrivals to the system. The optimal threshold is unique, and the genie-aided dispatcher uses the threshold $\bar{K} = 5$ in both cases that are plotted in Figure II.1a. The initial upper bound is $K^*(1) = l_1$, which is smaller than the optimal threshold but increases slowly so that eventually $\bar{K} < K^*(j)$ for large j. As shown in the analysis and the numerical experiments, the regret is $O(1)$. Figure II.1b shows the regret plot with respect to the number of arrivals for $\mu = 2$, $\lambda = 1$ and $R = 129/32$ with $l_1 = 3$. The regret is averaged over 2000 simulations and there are more than $2 * 10^5$ customers arrivals to the system. In this case, the optimal threshold is not unique: both $\bar{K} - 1 = 4$ and $\bar{K} = 5$ are optimal thresholds. The alternating genie-aided algorithm uses the policy that is described in Proposition 2.5.3

and only changes the threshold used between busy cycles. Similarly, as in Figure II.1a, the learning algorithm will not be able to use $\bar{K}$ in the first few batches because of the truncation. The plots indicate that constant regret is accumulated, which is consistent with our analytical results; interestingly, in all cases, convergence to the constant regret value happens rapidly.



(a) $\lambda = 1$ and the optimal threshold is $\bar{K} = 0$.

(b) $\lambda = 1$ and the optimal thresholds are $\{0, 1\}$; $\bar{K} = 1$.

Figure II.2: Regret of the learning system when an optimal threshold is zero. We set $C = R = 1, \mathbb{E}[B^j] = \ln(j)/j$, $K^*(j) \sim \ln(j)$ and $\alpha_j = j$.

Expected regret with zero being an optimal threshold: Figure II.2a shows how the regret changes with respect to the number of arrivals for $\mu = 0.8$ and $\mu = 0.9$ when $l_1 = 1$ and $\lambda = 1$. The regret is averaged over 2000 simulations and there are more than 10^5 customers arrived in the system. In both cases shown in Figure II.2a, the genie-aided dispatcher uses threshold $\bar{K} = 0$. Figure II.2b shows the regret plot with respect to the number of customers for $\mu = 1$ and $\lambda = 1$ when $l_1 = 3$. The regret is averaged over 2000 simulations and there are more than $2 * 10^5$ customers arrived in the system. In this case, the optimal threshold is not unique: both $\bar{K} - 1 = 0$ and $\bar{K} = 1$ are optimal thresholds. The alternating genie-aided dispatcher uses the policy that is described in Proposition 2.5.3 and only changes the threshold between busy cycles. The plots indicate that sub-linear regret is accumulated in all cases. Here, when the learning dispatcher uses threshold 0 in phase 2 of a given batch, the existence of the forced exploration phase in the next batch results in regret being accumulated. Note that for all plots shown in Figure II.2, the optimal thresholds can be used by the learning dispatcher in phase 2 right from the first batch.

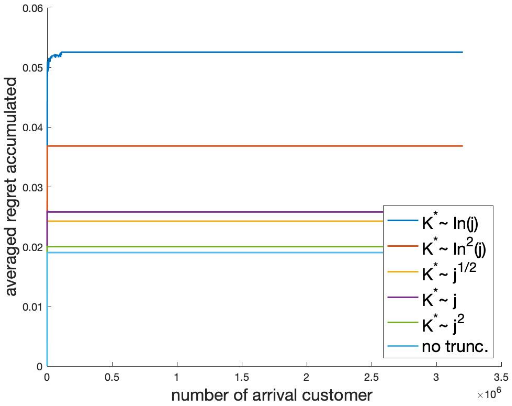
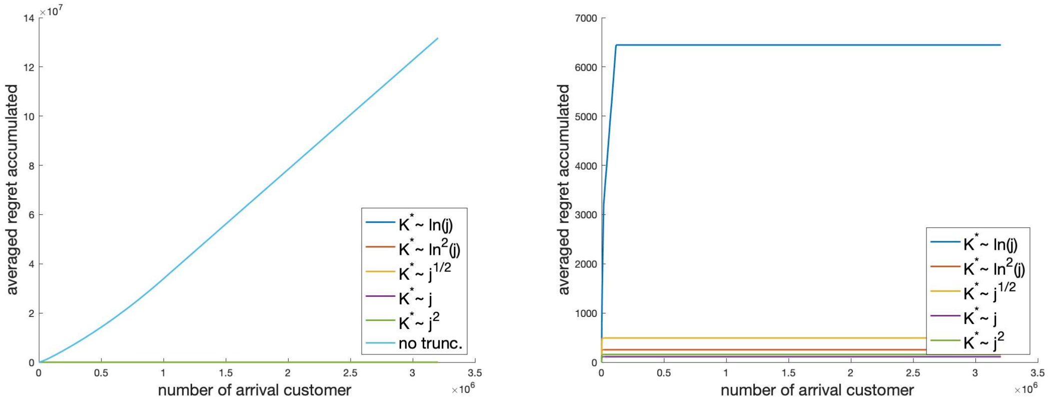


Figure II.3: Regret of the learning system when $\mu = 9$, $\lambda = 1$, $R = C = 1$ and the optimal threshold is 8 using and not using the truncation for the threshold used in phase 2. We set $\mathbb{E}[B^j] = \ln(j)/j$ and $\alpha_j = j$.

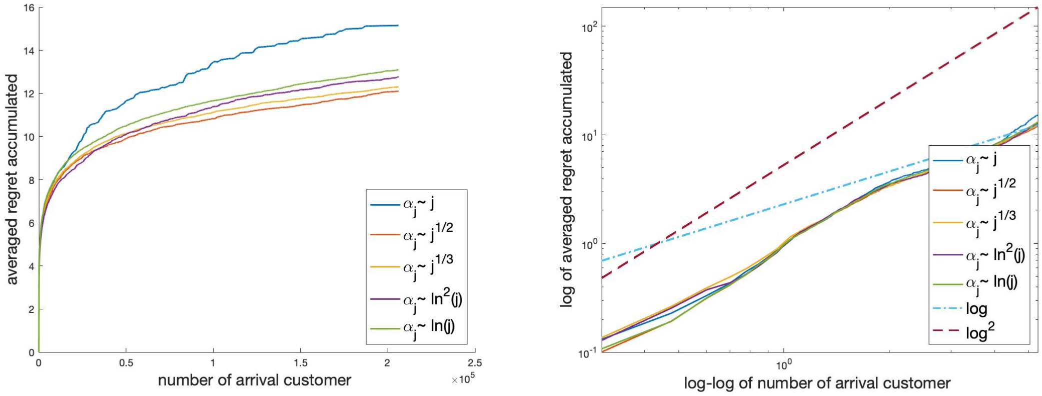


(a) With the no-truncation option included.

(b) Exclude the no-truncation option.

Figure II.4: Regret of the learning system when $\mu = 3$, $\lambda = 3.5$, $R = 21$ and the optimal threshold is 8 using and not using the truncation for the threshold used in phase 2. We set $C = 1$, $\mathbb{E}[B^j] = \ln(j)/j$ and $\alpha_j = j$.

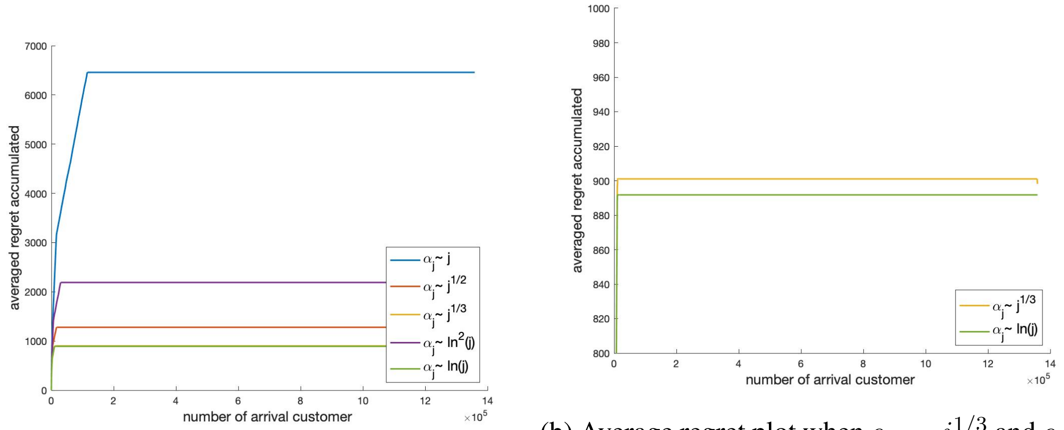
Expected regret with different choices of $K^*(j)$: We introduced truncation with the parameter $K^*(j)$ in our analysis since we needed a bound on the worst-case queue length for the learning system. We obtained a particular order of the regret with the choice of $K^*(j) = \max\{\lfloor \ln(j) \rfloor, 0\} + l_1 + Q_0$. Next, we explore the impact of different choices of $K^*(j)$ in Figure II.3 and Figure II.4. We use $\sim$ to indicate the order at which $K^*(j)$ increases: specifically, $K^*(j) \sim f(j)$ means $K^*(j) = \max\{\lfloor f(j) \rfloor, 0\} + l_1 + Q_0$. The regret values are averaged over 2000 simulations, and there are more than $3 * 10^5$ arrival customers that arrive in more than 700 batches. In Figure II.3, we use $\mu = 9$, $\lambda = 1$, and $R = 1$. The optimal threshold is $\bar{K} = 8$. This Figure indicates that under the chosen model parameters, with or without truncation, a constant regret can be achieved. However, this is not always the case. In Figure II.4, we use $\mu = 3$, $\lambda = 3.5$, and $R = 21$. The optimal threshold is $\bar{K} = 8$. The $M/M/1$ queue with $\mu = 3$ and $\lambda = 3.5$ is not stable. Despite this, Figure II.4b suggests that constant regret is achieved for various truncation choices. However, when no truncation is enforced, the regret accumulated seems to grow linearly with respect to the number of arrivals, see Figure II.4a. This suggests that the truncation is necessary and it helps to ensure a lower regret yet one may use a $K^*(j)$ that grows faster than $\ln(j)$. Confirming this through analysis is a topic to explore in future research.



(a) Average regret plot.

(b) Log versus log-log regret plot.

Figure II.5: Regret accumulated when $\mu = 0.8$, $\lambda = 1$ and $R = 1$ for different choice of α_j. Optimal threshold is $\bar{K} = 0$. We set $C = 1$, $\mathbb{E}[B^j] = \ln(j)/j$ and $K^*(j) \sim \ln(j)$.



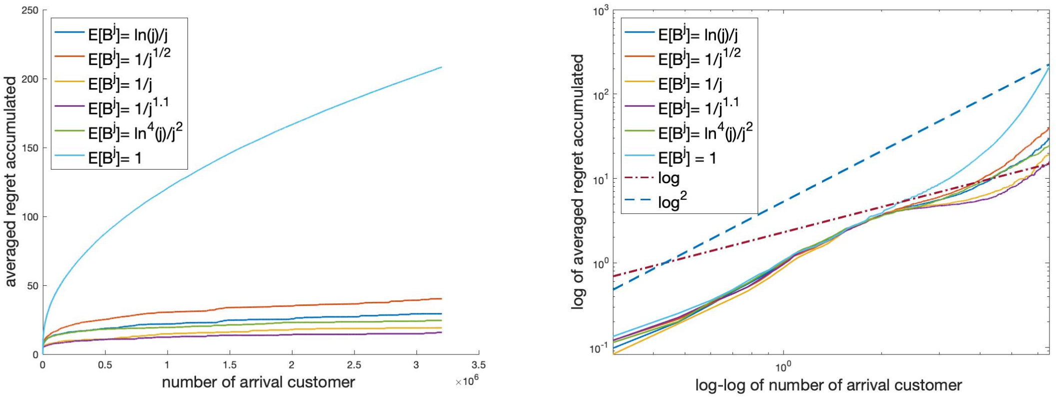
(a) Average regret plot.

(b) Average regret plot when $\alpha_j \sim j^{1/3}$ and $\alpha_j \sim \ln(j)$.

Figure II.6: Regret accumulated when $\mu = 3$, $\lambda = 3.5$, and $R = 21$ for different choices of α_j. Optimal threshold is $\bar{K} = 8$. We set $C = 1$, $\mathbb{E}[B^j] = \ln(j)/j$ and $K^*(j) \sim \ln(j)$.

Expected regret with different choices of α_j: We introduced $\alpha_j l_2$ to be the minimum length of phase 2 for the j^{th} batch. Figure II.5 and II.6 plots the average regret accumulated with different choices of α_j's. In particular, Figure II.5 includes the plot of the average regret accumulated as well as the log versus log-log plot when $\mu = 0.8$, $\lambda = 1$ with more than $2 * 10^5$ arrival customers, and Figure II.6 plots the average regret accumulated when $\mu = 3$, $\lambda = 3.5$ with more than $10 * 10^5$ arrival customers. We use $\alpha_j \sim f(j)$ to denote $\alpha_j = \max\{\lfloor f(j) \rfloor, 1\}$. The regret is averaged over 2000 simulations in both plots. Figure II.5 and II.6 suggests that for all these choices of α_j, a sub-linear regret is accumulated, and having an α_j that grows slower may still be able to achieve the regret bounds proved for $\alpha_j = j$.

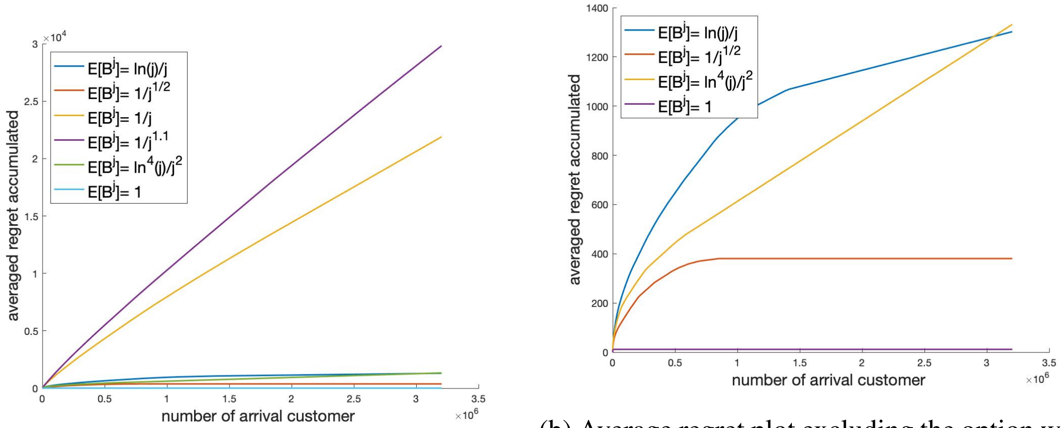
Expected regret with different choices of $\mathbb{E}[B^j]$: We also examined difference choices of $\mathbb{E}[B^j]$, which controls the probability of having a phase 1 when the threshold used in the previous phase 2 is 0. Figure II.7 and II.8 shows the plots of various choices of $\mathbb{E}[B^j]$. From these finite-time experiments, it seems that having a high enough chance to explore during the first few batches the learning dispatcher observes helps to reduce the regret accumulated. However, comparing the plots of $\mathbb{E}[B^j] = \ln^4(j)/j^2$ and $\mathbb{E}[B^j] = \ln(j)/j$ in Figure II.8b, it



(a) Average regret plot.

(b) Log versus log-log regret plot.

Figure II.7: Regret accumulated when $\mu = 0.8$, $\lambda = 1$ and the choices of $\mathbb{E}[B^j]$ vary. The optimal threshold is $\bar{K} = 0$. We set $C = R = 1$, $K^*(j) \sim \ln(j)$ and $\alpha_j = j$.

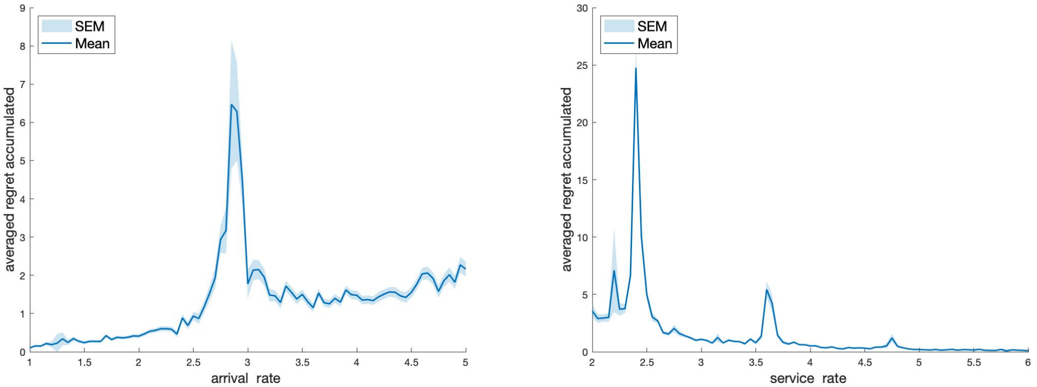


(b) Average regret plot excluding the option when $\mathbb{E}[B^j] = 1/j$ and $\mathbb{E}[B^j] = 1/j^{1.1}$.

(a) Average regret plot.

Figure II.8: Regret accumulated when $\mu = 1.3$, $\lambda = 1$ and the choices of $\mathbb{E}[B^j]$ vary. The optimal threshold is $\bar{K} = 1$ We set $C = R = 1$, $K^*(j) \sim \ln(j)$ and $\alpha_j = j$.

seems that only having a high probability of exploration for the first few batches is not be enough to achieve $O(1)$ regret since the slope of the plot for $\mathbb{E}[B^j] = \ln(j)/j$ decreases a lot faster than the plot of $\mathbb{E}[B^j] = \ln^4(j)/j^2$. Although all the choices of $\mathbb{E}[B^j]$ seem to achieve sub-linear regret for the case $\bar{K} = 0$, always having the exploration phase when the threshold used in the previous phase 2 is 0 accumulates a higher regret with a different scaling behavior.

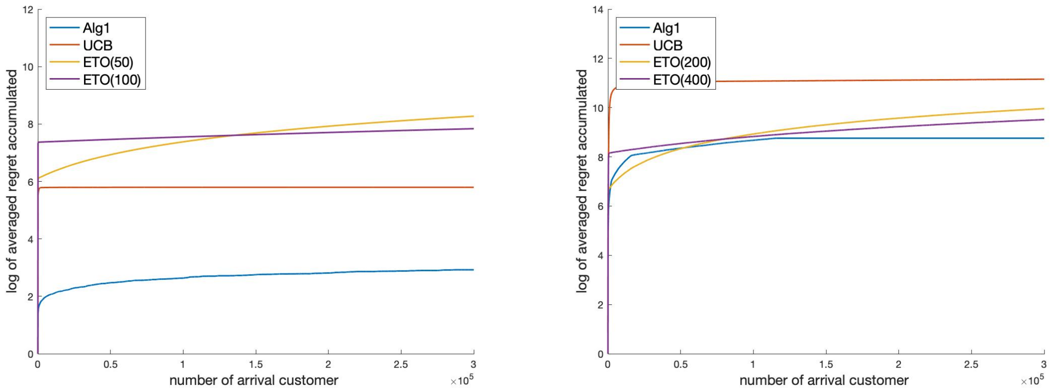


(a) Average regret plot versus various λ's when $\mu = 6$.

(b) Average regret plot versus various μ's when $\lambda = 1$

Figure II.9: Regret plot for various arrival and service rates. We set $C = R = 1$, $\mathbb{E}[B^j] = \ln(j)/j$, $K^*(j) \sim \ln(j)$ and $\alpha_j = j$.

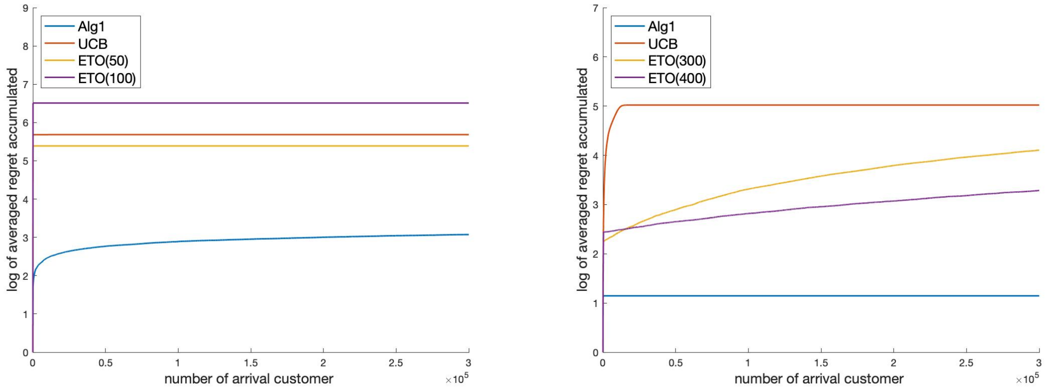
Expected regret with different values of μ and λ: Figure II.9 plots the average regret accumulated when seeing more than $3 * 10^5$ arriving customers when fixing one of the pair of arrival and service rates while varying the other. The regret values are averaged over 600 simulations. From the plot, we observe that when the arrival rate is fixed, as the service rate increases, in general, the regret decreases. However, the decrease is not strict and instead is non-monotonic, where the large cusps are usually around the parameter choices that have non-unique optimal thresholds. When the service rate is fixed, as the arrival rate increases, the regret follows a similar increasing/decreasing trend.



(a) Log average regret plot when $\mu = 0.8$, $\lambda = 1$, $R = 1$ and $\bar{K} = 0$.

(b) Log average regret plot when $\mu = 3$, $\lambda = 3.5$, $R = 21$ and $\bar{K} = 8$.

Figure II.10: Log of regret accumulated when using different algorithms when the optimal threshold is unique. Alg1 is the learning algorithm proposed in Algorithm 1. We set $C = 1$, $\mathbb{E}[B^j] = \ln(j)/j$, $K^*(j) \sim \ln(j)$ and $\alpha_j = j$. ETO(M) is the Estimate-Then-Optimize algorithm that always accepts the first M customers. UCB is the Upper Confidence Bound algorithm.



(a) Log averaged regret plot when $\mu = 1$ and $\lambda = 1$. Both $\bar{K} = 1$ and $\bar{K} - 1 = 0$ are optimal thresholds.

(b) Log of average regret plot when $\mu = 2$, $\lambda = 1$ and $R = 129/32$. Both $\bar{K} = 5$ and $\bar{K} - 1 = 4$ are optimal thresholds.

Figure II.11: Log of regret accumulated when using different algorithms when the optimal thresholds are not unique. Alg1 is the learning algorithm proposed in Algorithm 1. We set $C = 1$, $\mathbb{E}[B^j] = \ln(j)/j$, $K^*(j) \sim \ln(j)$ and $\alpha_j = j$. ETO(M) is the Estimate-Then-Optimize algorithm that always accepts the first M customers. UCB is the Upper Confidence Bound algorithm.

Comparison with benchmark algorithms: We also compared the finite time performance of our proposed Algorithm 1 with a few benchmark algorithms. In Figure II.10 and Figure II.11 we compared Algorithm 1 with the Estimate-Then-Optimize (ETO) algorithm and the Upper Confidence Bound (UCB) algorithm when there are more than $3 * 10^5$ arrival customers and the regrets are averaged over 2000 simulations. We use ETO(M) to denote the ETO algorithm which always accepts the first M customers. We use the UCB algorithm described in [Lattimore and Szepesvári 2020a, Section 7.1] but with UCB bias subtracted from the estimated average service time. Figure II.10a plots the log of average regret for the case when $\mu = 0.8$, $\lambda = 1$ and the optimal threshold is 0. Figure II.10b plots the log of average regret for the case when $\mu = 3$, $\lambda = 3.5$ and the optimal threshold is 8. For the parameters used in these two plots, the optimal threshold is unique. Figure II.11a plots the average regret for the case when $\mu = 1$, $\lambda = 1$ and the optimal thresholds are $\{1, 0\}$. Figure II.11b plots the average regret for the case when $\mu = 2$, $\lambda = 1$ and the optimal thresholds are $\{5, 4\}$. For the parameter choices in Figure II.11, the optimal threshold is not unique. The regret values in these two plots are computed with respect to the alternating genie-aided system which would change the threshold used between $\{\bar{K}, \bar{K} - 1\}$ according to the threshold used by Algorithm1, ETO or UCB.

The order of the regret accumulated by Algorithm 1 and UCB are similar in Figure II.10b and II.11b. However, in Figure II.10a and II.11a where 0 is an optimal threshold, UCB achieves constant regret yet Algorithm 1 achieves a sublinear regret. It is likely that the regret accumulated by Algorithm 1 would slowly increase as the number of arrivals increases and eventually becomes larger than the regret of the UCB algorithm. Our algorithm may choose to use threshold 0 and then a phase 1 may be enforced and regret accumulates because of this. In Figure II.12, we compared the finite time performance of our proposed algorithm with UCB when $\mu = 1.1$ and $\lambda = 1$ with 2000 simulations and more than 10^6 arrival customers. In this case, 1 is the unique optimal threshold. As we can observe from Figure II.12, the regret

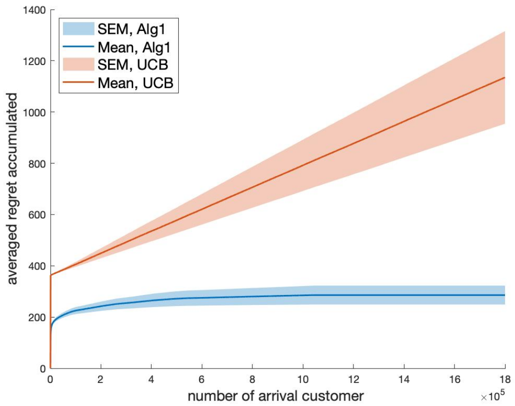


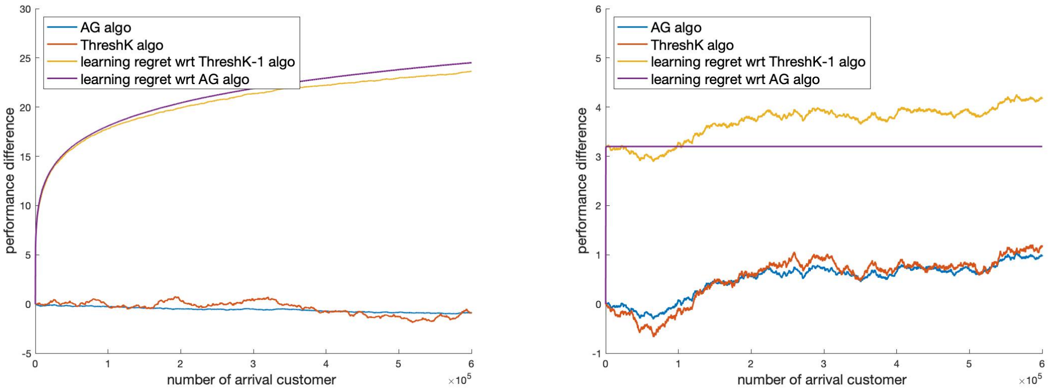
Figure II.12: Regret accumulated when $\mu = 1.1$, $\lambda = 1$, $C = R = 1$ and $\bar{K} = 1$. Alg1 is the learning algorithm proposed in Algorithm 1. We set $l_1 = l_2 = 30$, $B^j = \ln(j)/j$, $K^*(j) \sim \ln(j)$ and $\alpha_j = j$. UCB is the Upper Confidence Bound algorithm.

of UCB increases in a (approximately) linear fashion, while our proposed algorithm is able to achieve constant regret. In fact, we can argue the following for UCB-based dispatching (under the simpler setting of the arrival rate being known):

1. When the optimal threshold(s) is positive, then some bad initial service time samples can result in the estimated threshold being 0. This bad event happens with positive probability for all $\mu > \frac{C}{R}$ (the probability decreases to 0 as $\mu \to \infty$). Whenever this bad event occurs, then the UCB-based dispatching algorithm stops dispatching customers, obtains no new service time samples, and incurs linear regret.

2. When 0 is an optimal threshold, then the corresponding bad event of estimating the threshold as positive is more benign. This holds as dispatching more customers only results in more service time samples, which then help to correct inaccurate estimates. Hence, we expect to achieve a constant or slowly growing (sub-linear) regret.

Note that the explanation above supports the conjecture in Remark 2.4.4 since the worst-case (over parameters) regret of UCB is expected to be linear in N. Moreover, since UCB needs to compute the estimated threshold at every arrival, it requires more computation when compared to Algorithm 1.



(a) $\mu = \lambda = C = R = 1$. Both $\bar{K} = 1$ and $\bar{K} - 1 = 0$ are optimal thresholds.

(b) $\mu = 2$, $\lambda = 1$ and $R = 129/32$. Both $\bar{K} = 5$ and $\bar{K} - 1 = 4$ are optimal thresholds.

Figure II.13: Performance difference between the alternating genie, the genie algorithm using threshold $\bar{K}$ and the genie algorithm using threshold $\bar{K} - 1$. The accumulated net gain of the genie algorithm using threshold $\bar{K} - 1$ is scaled to be 0.

Comparison of different genie-aided algorithms: Figure II.13 compares the accumulated net gain between the alternating genie-aided algorithm (“AG algo" in the legend) coupled with Algorithm 1 and the genie-aided algorithms using threshold $\bar{K}$ (“ThreshK algo" in the legend) or $\bar{K} - 1$ (“ThreshK-1 algo" in the legend) when optimal thresholds are not unique; the accumulated net gain of the genie-aided algorithm using threshold $\bar{K} - 1$ are scaled to be 0. Figure II.13 plots the difference between the net gain obtained by the alternating genie-aided system and the genie-aided system using static threshold $\bar{K} - 1$, and the difference of the net gain between two genie-aided systems using static threshold $\bar{K}$ and $\bar{K} - 1$ over two sets of parameters. We also include the regret accumulated by the learning algorithm compared with the genie-aided algorithm using threshold K-1. The performances of the algorithms are averaged over 18000 simulations. As we can observe from the plots, the regret accumulated by the learning algorithm (with respect to either the alternating genie-aided system or the genie-aided system using threshold $K - 1$) dominates the performance difference between the alternating genie-aided system and the genie-aided system using threshold $K-1$, and the performing difference between the genie-aided system using threshold K and the genie-aided system using threshold $K-1$. This is more evidence in favor of Remark 2.5.1.

2.7 Conclusions

In this paper, we considered a social welfare maximizing problem, which was first proposed and studied in [Naor 1969]. We studied the learning problem of finding the proper threshold admission policy when the service and arrival rates are unknown. We proposed a learning algorithm that consists of batches where each batch has an optional exploration phase with a fixed length and an exploitation phase. When the optimal policy is unique, we showed that our learning algorithm achieves an $O(1)$ regret whenever the optimal threshold is non-zero, and achieves an $O(\ln^{1+\epsilon}(N))$ regret when the optimal threshold is zero, where N denotes the total number of arrival customers to the systems. When the optimal policy is not unique, we specified a particular optimal policy to compare with, and proved that similar regret bounds hold for our learning algorithm.

CHAPTER III

On The Asymptotic Optimality Of The Comb Strategy For Prediction with Expert Advice

3.1 Introduction

In this chapter we use PDE tools to analyze one of the classical problems in machine learning, namely prediction with expert advice. In this framework, a game is played between a player and nature (also called the adversary in the learning literature). At each time step, given past information, the player has to choose an expert among $N > 0$ experts. Simultaneously nature chooses a set of winning experts. Then, both choices are announced. If the player chooses an expert belonging to the set of winning experts, the player also wins. The objective of the player is to minimize his regret with respect to the best performing expert, i.e., minimize

$$R_T = \max_i G_T^i - G_T$$

where G_T^i is the total gain of the expert i and G_T is the gain of the player at the final time. The objective of nature is to choose the set of winning experts to maximize the regret of the player. This problem that has been extensively studied in learning theory [Cover 1967; Cesa-Bianchi and Lugosi 2006; Rakhlin, Shamir, and Sridharan 2012; Gravin, Peres, and Sivan 2016; Haussler, Kivinen, and Warmuth 1995; Littlestone and Warmuth 1994; Vovk 1990], and can also be seen as a discrete time and discrete space robust utility maximization problem similar to [Nutz 2016] for a particular choice of utility function.

For the case of 2 experts, the optimal strategy for the adversary was first described by Cover [Cover 1967] in the 1960s. Recently, for the case with 3 experts, using an ansatz of exponential type for the value function of the game, Gravin et al. showed in [Gravin, Peres, and Sivan 2016] that the so called “comb strategy”, the strategy that consists of choosing the leading and the third leading expert by nature, is optimal. However, the exponential type ansatz for the value of the game does not generalize to larger number of experts.

In this chapter, we follow the setting of [Gravin, Peres, and Sivan 2016], where the maturity of the game is a geometric random variable with parameter $\delta > 0$ and study the game where both the player and nature can use randomized strategies. In this framework, we prove two conjectures stated in [Gravin, Peres, and Sivan 2016] for the game with $N = 4$ experts. We use tools from stochastic analysis and PDE theory to give an explicit expansion of the value function of the game for small $\delta > 0$, which corresponds to long time asymptotics. In Theorem 3.3.1, this expansion allows us to prove that the value of the game, also called best regret, indeed grows as $\frac{\pi}{4\sqrt{2\delta}}$ as conjectured in [Gravin, Peres, and Sivan 2016].

The proof of this result is achieved in two steps. This first step can be found in [Drenska 2017], where, using tools from viscosity theory, the author showed that the rescaled value function (III.4) solves the elliptic PDE (III.5). The second step, which is the main contribution of this chapter, is to explicitly solve this PDE for the case of 4 experts. In order to find this expression, we use the conjectured optimal strategy in [Gravin, Peres, and Sivan 2016], and relate the value function of the control problem (III.6) to the expectation of a functional of an obliquely reflected Brownian motion; see in particular Lemma 3.4.1. This expression is a discounted expected value of the local time that measures the number of times the best two experts’ gains cross each other. Then, using appropriate differentiation of the dynamic programming equation (III.33), we characterize the value of the expectation on two “opposite” faces of the domain of reflection by a system of hyperbolic PDE (III.42) and (III.43). Then, we solve this system of hyperbolic PDE to explicitly compute the value for the conjectured control at the boundary, which then leads to the value in the whole domain. Finally, in Section 3.6, we check that the value given for the conjectured control solve the nonlinear PDE (III.5), which is a simple verification argument proving the optimality of comb strategy, the second conjecture in [Gravin, Peres, and Sivan 2016]. See Theorem 3.3.2. The direct proof of the verification argument is quite tedious. Hence, we came up with a method that relies on Proposition 3.5.4, which is a type of maximum principle for the system of hyperbolic equations (III.53).

From the perspective of control theory, we note that the setting of [Gravin, Peres, and Sivan 2016] is in fact similar to the weak formulation (or feedback/closed loop formulation) of zero-sum games in the sense of [T. Pham and J. Zhang 2014] (see also [Bayraktar, Cosso, and H. Pham 2016]) where the player and nature observe the same source of information, i.e. the path of the gains of the experts and the player. One can also state the game in a Elliott-Kalton sense, see e.g. [Fleming and Souganidis 1989], in which similarly to [Kohn and Serfaty 2010], before taking its decision, nature learns the choice of the player. These two formulations generally lead to different values; see Remark 4.2 in [Bayraktar, Cosso, and H. Pham 2016].

Our expansion is in accordance with well-known results in prediction problems. Indeed, it is known that in the long run, there is an upper bound for the value of regret minimization problems that grows at most as $\sqrt{\frac{T\log(N)}{2}}$, which is achieved by the so-called multiplicative weight algorithms [Cesa-Bianchi and Lugosi 2006]. In this chapter, we compute the exact scaling for the geometric stopping problem which also allows us to directly provide explicit algorithms for nature.

The rest of the chapter is organized as follows. In Section 3.2 we introduce our notation and define the value function of the regret minimization problem. In Section 3.3, we give the main results of the chapter. This result is proven in Section 3.6. The Sections 3.4 and 3.5 are there to provide the methodology used in finding the explicit solution (III.7).

3.2 Statement of the problem

We fix $N \geq 2$ and denote by U the set of probability measures on $\{1, \ldots, N\}$ and by V the set of probability measures on $P(N)$, the power set of $\{1, \ldots N\}$. These sets of probability measures are in fact in bijection with respectively N and 2^N dimensional unit simplexes. We denote by $\{e_i\}_{i=\{1,\ldots,N\}}$ the canonical basis of $\mathbb{R}^N$ and for $J \in P(N)$, e_J stands for $e_J := \sum_{j \in J} e_j$. Similar to [Ichiba, Karatzas, and Shkolnikov 2013], for all $x \in \mathbb{R}^N$, we denote by $\{x^{(i)}\}_{i=1,\ldots,N}$ the ranked coordinates of x with

$$x^{(1)} \leq x^{(2)} \leq \ldots \leq x^{(N)},$$

and define the function

$$\Phi : x \in \mathbb{R}^N \mapsto \max_i x_i = x^{(N)}. \tag{III.1}$$

We assume that a player and nature interact through the evolution of the state of N experts. At time $t \in \mathbb{N}$, the state of the game in hand is described by $\{G^i_s\}_{s=1,\ldots,t-1}$, the history of the gains of each expert $i = 1, \ldots N$ and $\{G_s\}_{s=1,\ldots,t-1}$ the history of the gains of the player. At time step $t \in \mathbb{N}$, observing $\{(G^i_s, G_s) : s = 0, \ldots, t-1\}$, simultaneously, the player chooses $I_t \in \{1, \ldots, N\}$ and nature chooses $J_t \in P(N)$. The gain of each expert chosen by nature increases by 1 i.e.,

$$G^i_t = G^i_{t-1} + 1 \text{ if } i \in J_t$$
$$G^i_t = G^i_{t-1} \text{ if } i \notin J_t.$$

If the player also chooses an expert chosen by nature, then the gain of player also increases i.e.,

$$G_t = G_{t-1} + 1 \text{ if } I_t \in J_t$$
$$G_t = G_{t-1} \text{ if } I_t \notin J_t.$$

The regret of the player at time $t \in \mathbb{N}$ is defined as

$$R_t := \max_{i=1,\ldots,N} G_t^i - G_t.$$

Let T denote the random maturity of the problem. We assume that T is a geometric random variable with parameter $\delta > 0$.

We now convexify the problem by assuming that instead of choosing deterministic J_t and I_t, nature and the player choose randomized strategies. At time t, the player chooses a probability distribution $\alpha_t \in U$ and nature chooses $\beta_t \in V$ that may depend on the observation $\{(G_s^i, G_s) : s = 0, \ldots t-1,\, i = 1, \ldots, N\}$. We denote by $\mathcal{U}$ the set of such sequences $\{\alpha_t\}_{t\in\mathbb{N}}$ and by $\mathcal{V}$ the set of such sequences $\{\beta_t\}$. With some notational abuse, we denote by $I_t \in \{1, \ldots N\}$ the random variable with distribution α_t and $J_t \in P(N)$ the random variable with distribution β_t.

The objective of the player is to minimize his expected regret at time T and the objective of nature is to maximize the regret of the player. Hence we have a zero sum game with the lower and the upper value for the game

$$\sup_{\beta\in\mathcal{V}} \inf_{\alpha\in\mathcal{U}} \mathbb{E}^{\alpha,\beta}\left[R_T\right] \leq \inf_{\alpha\in\mathcal{U}} \sup_{\beta\in\mathcal{V}} \mathbb{E}^{\alpha,\beta}\left[R_T\right]$$

where $\mathbb{E}^{\alpha,\beta}$ is the probability distribution under which we evaluate the regret given the controls

$\alpha = \{\alpha_t\}$ and $\beta = \{\beta_t\}$. We denote by

$$X_t := (X_t^1, \dots, X_t^N) := (G_t^1 - G_t, \dots, G_t^N - G_t) \tag{III.2}$$

the difference between the gain of the player and the experts. The following result, which can be found in [Gravin, Peres, and Sivan 2016; Drenska 2017], establishes the existence of a value for this discrete game.

Proposition 3.2.1. *The game has a value, i.e.,*

$$V^\delta(X_0) := \sup_{\beta \in \mathcal{V}} \inf_{\alpha \in \mathcal{U}} \mathbb{E}^{\alpha,\beta}\left[R_T\right] = \inf_{\alpha \in \mathcal{U}} \sup_{\beta \in \mathcal{V}} \mathbb{E}^{\alpha,\beta}\left[R_T\right]. \tag{III.3}$$

There exists $M > 0$ independent of δ such that for all $\delta > 0$ and $x \in \mathbb{R}^N$ we have that

$$|V^\delta(x) - \Phi(x)| \leq \frac{M}{\sqrt{\delta}}.$$

Additionally, V^δ satisfies the following dynamic programming principle

$$V^\delta(x) = \delta\Phi(x) + (1-\delta) \inf_{\alpha \in U} \sup_{\beta \in V} \sum_J \beta_J \left(V^\delta\left(x + e_J\right) - \alpha(J)\right).$$

Proof. The existence of the value is a direct consequence of the Minimax Theorem and is provided in [Gravin, Peres, and Sivan 2016]. The proof of the rest of the Proposition can be found in [Drenska 2017]. In particular the uniform bound in x is a consequence of [Drenska 2017, Theorem 3]. □

3.2.1 Limiting behavior of V^δ

The main objective of the chapter is to provide an explicit formula for the leading order for the function V^δ for small $\delta > 0$. For this purpose define the rescaled value function:

$$u^\delta : x \in \mathbb{R}^N \mapsto V^\delta \left(\frac{x}{\sqrt{\delta}} \right) \sqrt{\delta}. \tag{III.4}$$

The next result shows that the limiting behavior of the value of the game can be characterized by the value of a stochastic control problem.

Proposition 3.2.2. *As $\delta \downarrow 0$, the function u^δ converges locally uniformly to $u : \mathbb{R}^N \mapsto \mathbb{R}$ which is the unique viscosity solution of the equation*

$$u(x) - \frac{1}{2} \sup_{J \in P(N)} e_J^\top \partial^2 u(x) e_J = \Phi(x) \tag{III.5}$$

in the class of functions with linear growth. Additionally, u admits the stochastic control representation

$$u(x) = \sup_{(\sigma_t)} \mathbb{E} \left[\int_0^\infty e^{-t} \Phi(X_t) dt \right] \tag{III.6}$$

where X is defined by $X_t = x + \int_0^t \sigma_s dW_s$ with W a 1-dimensional Brownian motion and the progressively measurable process (σ_t) satisfies for all t $\sigma_t \in \{e_J : J \in P(N)\}$.

Proof. The fact that u^δ converges to u is a consequence of [Drenska 2017, Theorem 7]. Note also that an analysis of the proof of [Drenska 2017, Theorem 7] and the general methodology of proof in [Barles and Souganidis 1991] allows us to claim that the convergence is in fact locally uniform. The fact that u admits the representation (III.6) is a consequence of uniqueness of viscosity solution of (III.5) with linear growth that is proven in [Crandall, Ishii, and Lions 1992, Theorem 5.1] and the stochastic Perron's method of [Bayraktar and Sîrbu

2013]. □

3.3 Main Results

3.3.1 Explicit solution for 4 experts

The main contribution of the chapter is to provide a method to explicitly solve the PDE (III.5).

Theorem 3.3.1. *With 4 experts, for $x \in \mathbb{R}^4$, the function u is given by the expression*

$$\begin{aligned} u(x) = {} & x^{(4)} - \frac{\sqrt{2}}{4}\sinh(\sqrt{2}(x^{(4)} - x^{(3)})) \\ & + \frac{\sqrt{2}}{2}\arctan\left(e^{\frac{x^{(1)}+x^{(2)}-x^{(3)}-x^{(4)}}{\sqrt{2}}}\right)\cosh\left(\frac{x^{(1)}-x^{(2)}+x^{(3)}-x^{(4)}}{\sqrt{2}}\right) \\ & \cosh\left(\frac{-x^{(1)}+x^{(2)}+x^{(3)}-x^{(4)}}{\sqrt{2}}\right)\cosh\left(\frac{-x^{(1)}-x^{(2)}+x^{(3)}+x^{(4)}}{\sqrt{2}}\right) \\ & + \frac{\sqrt{2}}{2}\operatorname{arctanh}\left(e^{\frac{x^{(1)}+x^{(2)}-x^{(3)}-x^{(4)}}{\sqrt{2}}}\right)\sinh\left(\frac{x^{(1)}-x^{(2)}+x^{(3)}-x^{(4)}}{\sqrt{2}}\right) \\ & \sinh\left(\frac{-x^{(1)}+x^{(2)}+x^{(3)}-x^{(4)}}{\sqrt{2}}\right)\sinh\left(\frac{-x^{(1)}-x^{(2)}+x^{(3)}+x^{(4)}}{\sqrt{2}}\right) \end{aligned} \tag{III.7}$$

Additionally, u is twice continuously differentiable, monotone [1], symmetric in its variables on $\mathbb{R}^4$, satisfy

$$u(x + \lambda(e_1 + e_2 + e_3 + e_4)) = u(x) + \lambda \textit{ for all } x \in \mathbb{R}^4 \textit{ and } \lambda \in \mathbb{R} \tag{III.8}$$

and if J is a maximizer of the Hamiltonian $\sup_{J \in P(N)} e_J^\top \partial^2 u(x) e_J$ then its complement J^c is also a maximizer of the same Hamiltonian.

[1] Monotone here means

$$u(x_1 + y_1, \dots, x_N + y_N) \geq u(x_1, \dots, x_N) \text{ for all } x_i \in \mathbb{R} \text{ and } y_i \geq 0.$$

Moreover,

$$V^{\delta}(0) = \frac{\pi}{4\sqrt{2\delta}} + o\left(\frac{1}{\sqrt{\delta}}\right), \tag{III.9}$$

In fact, u has the following expansion at the origin

$$\begin{aligned} &u(x_1, x_2, x_3, x_4) = \frac{\pi}{4\sqrt{2}} + \frac{1}{4}(x_1 + x_2 + x_3 + x_4) + \\ &\frac{3\pi}{16\sqrt{2}}(x_1^2 + x_2^2 + x_3^2 + x_4^2 - \frac{2}{3}(x_1x_2 + x_1x_3 + x_1x_4 + x_2x_3 + x_2x_4 + x_3x_4)) \\ &+ o(|x|^2). \end{aligned} \tag{III.10}$$

Proof. The proof of this result is provided in Section 3.6 after developing the methodology required to obtain this expression. Note that one can check by hand (or preferably with a computer) that the expression provided at (III.7) solves the equation (III.5) when all x_i are different from each other. Since the set of points $x \in \mathbb{R}^4$ with $x_i = x_j$ for some i, j is of zero Lebesgue measure, this proves that u is an almost everywhere solution of (III.5). However, due to potential discontinuities of the derivatives when two of the $x_i's$ are equal we need to check that the almost everywhere solution of the equation (III.5) defined via this expression is twice continuously differentiable and is therefore a smooth solution. □

Remark 3.3.1. *i)*(III.9) *is the main result for the long time behavior of the regret minimization problem with geometric stopping and is conjectured in [Gravin, Peres, and Sivan 2016]. The optimal regret scales as the square root of the time scale in hand. In this case of geometric stopping $u(0) = \frac{\pi}{4\sqrt{2}}$ gives the term of proportionality between the optimal regret and the stopping time parameter.*

ii) The fact that J^c maximizes $\sup_{J \in P(N)} e_J^{\top} \partial^2 u(x) e_J$ whenever J maximizes this Hamiltonian is a direct consequence of the regularity of u and its translation invariance as in (III.8). *This fact will be useful to us while checking the optimality of comb strategies.*

3.3.2 Asymptotically optimal strategies

Given the value of u, we now describe a family of asymptotically optimal strategies for nature. Inspired by [Gravin, Peres, and Sivan 2016] we give the following definition.

Definition 3.3.1. *(i) We denote*

$$\mathcal{J}^*(x) = \operatorname*{arg\,max}_{J \in P(N)} e_J^\top \partial^2 u(x) e_J, \tag{III.11}$$

the set of maximizers of the Hamiltonian.

(ii) For all $x \in \mathbb{R}^4$ with $x_{i_1} \leq x_{i_2} \leq x_{i_3} \leq x_{i_4}$, we denote $\mathcal{J}_\mathcal{C}(x) \in P(N)$ the comb strategy which is the control for the problem (III.6) *that consists in choosing the experts i_4 and i_2. We take the convention that if two components x_i and x_j of the points are equal for $i < j$ then the ordering of the point is taken with $x_i \leq x_j$.*

(iii) We denote $\mathcal{J}_\mathcal{C}^b \in \mathcal{V}$ the balanced comb strategy which is the control for nature in game (III.3) *that consists in choosing at $\frac{x}{\sqrt{\delta}} \in \mathbb{R}^4$, $\mathcal{J}_\mathcal{C}(x) \in P(N)$ with probability $\frac{1}{2}$ and $\mathcal{J}_\mathcal{C}^c(x) \in P(N)$ with probability $\frac{1}{2}$.*

Remark 3.3.2. *Note that (ii) defines a control for the control problem* (III.6) *while (iii) defines a control for the game* (III.3)*. Hence the latter depends on δ and x and is scaled to reflect the scaling between the two problems. Additionally, as a consequence of [Gravin, Peres, and Sivan 2016, Claim 1], we have defined $\mathcal{J}_\mathcal{C}^b$ as the unique balanced strategy that can be generated using $\mathcal{J}_\mathcal{C}(x)$.*

One may conjecture that it is asymptotically optimal for nature to choose for all $\frac{x}{\sqrt{\delta}} \in \mathbb{R}^4$ an element in $\mathcal{J}(x)$. However, this conjecture is not true since the strategy is not balanced in the sense of [Gravin, Peres, and Sivan 2016]. Indeed, assume for example that for $x \in \mathbb{R}^4$ $\mathcal{J}^*(x)$ is reduced to a unique subset of cardinality 1, meaning $\mathcal{J}^*(x) = \{J\} = \{\{i\}\}$. In this case, choosing the expert i would be sub-optimal for nature since the player can also guess this control and choose the expert i. It is proven in [Gravin, Peres, and Sivan 2016] that in order to be optimal any strategy of nature has to be balanced. Thanks to the Theorem 3.3.1, the simplest strategy for nature would be to randomize his strategy between the maximizer of the Hamiltonian and its complement.

The main result for asymptotically optimal strategies is the following theorem.

Theorem 3.3.2. *The control $\mathcal{J}_{\mathcal{C}}^{b} \in \mathcal{V}$ is asymptotically optimal for nature, in the sense that*

$$\underline{u}^{\delta}(x) = u(x) + o(1), \tag{III.12}$$

where o is locally uniform in x, and we denote

$$\underline{u}^{\delta}(x) = \sqrt{\delta} \inf_{\alpha \in \mathcal{U}} \mathbb{E}^{\alpha, \mathcal{J}_{\mathcal{C}}^{b}} \left[R_T^{\frac{x}{\sqrt{\delta}}} \right] \tag{III.13}$$

where $R_T^{\frac{x}{\sqrt{\delta}}}$ is the regret of player at time T starting from the state $X_0 = \frac{x}{\sqrt{\delta}}$.

The proof is deferred to Section 3.6.2. We will finish this section with a few remarks.

Remark 3.3.3. *As a sanity check, the expansion of u implies that the Hessian of u is*

$$H = \frac{\pi}{8\sqrt{2}} \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix}$$

and

$$u(0) = \frac{\pi}{4\sqrt{2}} = \frac{1}{2} \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}^{\top} H \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix},$$

where the second equality follows from (III.5) *and the optimality of the comb strategies.*

Remark 3.3.4. *We note that at the leading order it is optimal for nature to choose the controls $\mathcal{J}_C^b$ in the sense that for all family $\alpha^\delta \in \mathcal{U}$ and $\beta^\delta \in \mathcal{V}$ for $\delta > 0$, we have that*

$$\limsup_{\delta \downarrow 0} \sqrt{\delta} \left(\mathbb{E}^{\alpha^\delta, \beta^\delta} \left[R_T^{\frac{x}{\sqrt{\delta}}} \right] - \mathbb{E}^{\alpha^\delta, \mathcal{J}_C^b} \left[R_T^{\frac{x}{\sqrt{\delta}}} \right] \right) \leq 0.$$

This inequality means that up to an error negligible at the leading order, the comb strategy is optimal for nature.

Remark 3.3.5. *i)In the case of 3 experts, [Gravin, Peres, and Sivan 2016] gives the exact value of V^δ based on a "guess and verify approach".The following expression is given for u in [Drenska 2017]*

$$x^{(3)} + \frac{1}{2\sqrt{2}} e^{\sqrt{2}(x^{(2)} - x^{(3)})} + \frac{1}{6\sqrt{2}} e^{\sqrt{2}(2x^{(1)} - x^{(2)} - x^{(3)})},$$

which is obtained by taking a continuum analogue of [Drenska 2017]. Compared to this 3 dimensional counterpart the expression (III.7) *is not a simple sum of exponentials. Instead of guess and verify we needed to directly compute the value of comb strategies.*

ii)It is possible to prove the Theorem 3.3.2 for the particular case of $x = 0$ using a combination of Theorem 3.3.1 and [Gravin, Peres, and Sivan 2016, Theorem 5.1]. However, unlike the proof provided below, such a proof cannot be directly extended to the general case $x \in \mathbb{R}^4$.

Remark 3.3.6. *Note that for all $x \in \mathbb{R}^4$ we have $\partial_i u(x) \in [0,1]$ with $\sum_{i=1}^4 \partial_i u(x) \in [0,1] = 1$. Hence $\{\partial_i u(x)\}_{i=1}^4 \in U$. The claim is direct consequence of* (III.8)*. Thanks to this observation, we can define $\alpha^* \in \mathcal{U}$ via the feedback control : at point $\frac{x}{\sqrt{\delta}} \in \mathbb{R}^4$, the player* *chooses the expert i with probability $\partial_i u(x)$ and define the value*

$$\overline{u}^{\delta}(x) = \sqrt{\delta} \sup_{\beta \in \mathcal{V}} \mathbb{E}^{\alpha^*, \beta} \left[R_T^{\frac{x}{\sqrt{\delta}}} \right]. \tag{III.14}$$

We conjecture that

$$\overline{u}^{\delta}(x) = u(x) + o(1)$$

which would imply that α^* is an asymptotically optimal strategy for the player. The main difficulty one faces to obtain such a result is to obtain locally uniform bounds for $\overline{u}^{\delta}(x)$ when $\delta \downarrow 0$.

3.4 Value for comb strategies

Inspired by the conjecture in [Gravin, Peres, and Sivan 2016], our objective here is to introduce the value of the control problem (III.6) corresponding to comb strategies. Then, in Section 3.5, we develop a methodology to compute this value. Finally, in Section 3.6, we check that the value computed in these sections is a solution to (III.5).

We note that the Sections 3.4 and 3.5 are only included in the chapter to explain how to find the expression (III.7). Indeed, the only rigorous proofs for our results are in Section 3.6. Therefore, in Sections 3.4 and 3.5, we will slightly deviate from mathematical rigor. The purpose of this section is to relate the value given by comb strategies with distributional properties of an obliquely reflected Brownian Motion. Then, we compute and analyze this value in Sections 3.5 and 3.6.

3.4.1 Analysis

The optimal strategy for (III.6) conjectured in [Gravin, Peres, and Sivan 2016] consists in choosing the best and the third best experts. This is a rank based interaction for the evolution of the components of $X^{\mathcal{C}}$, the optimally controlled state. Therefore, for any $x \in \mathbb{R}^4$, it is expected that $X^{i,x,\mathcal{C}}$ solves the following SDE

$$X_t^{i,x,\mathcal{C}} = x^i + \int_0^t \sum_{j=1}^4 \sigma_j^{\mathcal{C}} \mathbf{1}_{\{X_r^{i,x,\mathcal{C}} = X_r^{(j),x,\mathcal{C}}\}} dW_r \text{ for } t \geq 0 \text{ and } i = 1, \dots, 4; \tag{III.15}$$

where $\sigma_4^{\mathcal{C}} = \sigma_2^{\mathcal{C}} = 1$ and $\sigma_3^{\mathcal{C}} = \sigma_1^{\mathcal{C}} = 0$ is the control corresponding to comb strategy.

It is not clear that (III.15) admits a strong solution. In fact, based on [Fernholz et al. 2011, Theorem 4.1], we conjecture that there is no strong solution to (III.15). However, it is expected that the ranked components $X_t^{(i),x,\mathcal{C}}$ are well-defined. Given also the fact that the payoff of the problem is symmetric, we will directly define our value of interest via an obliquely reflected Brownian motion. This procedure also allows a reduction of the dimension of the problem.

We first recall the definition of an obliquely reflected Brownian motion given in [Williams 1995, Definition 2.1].

Definition 3.4.1. *We say that the family of continuous processes $\{\mathcal{Y}_t^y\}_{y \in \mathbb{R}_+^3}$ and probability measures $\{\P^y\}_{y \in \mathbb{R}_+^3}$ is a weak solution to the semimartingale reflected Brownian motion on $\mathbb{R}_+^3$ with covariance matrix Γ and reflection matrix R if*
i) For all $t \geq 0$ and $y \in \mathbb{R}_+^3$

$$\mathcal{Y}_t^y = \mathcal{W}_t^y + R\lambda^y(t)$$

ii) The process $\mathcal{W}_t^y \in \mathbb{R}^3$ is a Brownian motion with covariance matrix Γ under $\P^y$.
iii) λ^y is adapted to the filtration generated by $\mathcal{Y}^y$, $\lambda_0^y = 0$, λ^y is continuous, non decreasing, and

$$\int_0^t \mathbf{1}_{\{\mathcal{Y}_r^{i,y} = 0\}} d\lambda_r^i = \lambda_t^i \mathit{for}\ i = 1, 2, 3.$$

We will denote by $(Y^y)_{y \in \mathbb{R}^3_+} = (Y^{1,y}, Y^{2,y}, Y^{3,y})_{y \in \mathbb{R}^3_+}$ the family with

$$\Gamma := \begin{pmatrix} 1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & 1 \end{pmatrix}, \; R := \begin{pmatrix} 1 & -1/2 & 0 \\ -1/2 & 1 & -1/2 \\ 0 & -1/2 & 1 \end{pmatrix}$$

and $(Y^y_1(0), Y^y_2(0), Y^y_3(0)) = y \in \mathbb{R}^3_+$. These processes have the following semimartingale decomposition for $t \geq 0$,

$$\begin{aligned} dY^{3,y}_t &= dW_t + d\Lambda^3_t - \frac{1}{2} d\Lambda^2_t, \\ dY^{2,y}_t &= -dW_t + d\Lambda^2_t - \frac{1}{2}(d\Lambda^3_t + d\Lambda^1_t), \\ dY^{1,y}_t &= dW_t + d\Lambda^1_t - \frac{1}{2} d\Lambda^2_t, \end{aligned} \tag{III.16}$$

and denote Λ^j_t for $j = 1, 2, 3$ the local time of $Y^y_j \geq 0$ at the origin. Since the matrix $R - I$ is a tridiagonal Toeplitz matrix whose eigenvalues are less than 1 in absolute value, there exists a unique solution to the oblique reflection problem; see [Williams 1995, Theorem 2.1]. However, the existence of solution to (III.15) is not straightforward as discussed above. If a solution to this system existed, then we clearly would have

$$Y^{j-1,y}_t(t) = X^{(j),x,\mathcal{C}}_t - X^{(j-1),x,\mathcal{C}}_t \geq 0 \text{ for } j = 2, 3, 4,$$

with $y = (x^{(2)} - x^{(1)}, x^{(3)} - x^{(2)}, x^{(4)} - x^{(3)}) \in \mathbb{R}^3_+$. Henceforth, we will assume that this is the case. (This is the only non-rigorous part of the derivation. But we should again remark that a rigorous verification of our claims is in Section 3.6 and the arguments here are performed for giving an intuitive construction of the solution.) In the sequel we will denote

$$Y_t^{4,y} = \sum_{j=1}^{4} X_t^{(j),x,\mathcal{C}} = \sum_{j=1}^{4} X_t^{j,x,\mathcal{C}}.$$

3.4.2 Value associated to an obliquely reflected Brownian motion

We now give a lemma that allows us to define our candidate solution to (III.5).

Lemma 3.4.1. *Assume that there exists a weak solution to* (III.15). *Then for all x we have*

$$\mathbb{E}\left[\int_0^\infty e^{-t}\Phi(X_t^{x,\mathcal{C}})dt\right] =\Phi(x) + v(x^{(2)} - x^{(1)}, x^{(3)} - x^{(2)}, x^{(4)} - x^{(3)}) \tag{III.17}$$

where

$$v(y_1, y_2, y_3) := \frac{1}{2}\mathbb{E}\left[\int_0^\infty e^{-t}\Lambda_t^{3,y}dt\right].$$

Proof. We fix $x \in \mathbb{R}^d$ define $y \in \mathbb{R}_+^3$ with $y := (x^{(2)} - x^{(1)}, x^{(3)} - x^{(2)}, x^{(4)} - x^{(3)})$. Thanks to our definitions for all $t \geq 0$,

$$\begin{aligned} Y_t^{1,y} + 2Y_t^{2,y} + 3Y_t^{3,y} &= 3X_t^{(4),x,\mathcal{C}} - X_t^{(3),x,\mathcal{C}} - X_t^{(2),x,\mathcal{C}} - X_t^{(1),x,\mathcal{C}} \\ &= 4X_t^{(4),x,\mathcal{C}} - Y_t^{4,y}. \end{aligned}$$

Thus,

$$\mathbb{E}\left[\int_0^\infty e^{-t}\Phi(X_t^{x,\mathcal{C}})dt\right] = \frac{1}{4}\mathbb{E}\left[\int_0^\infty e^{-t}Y_t^{4,y}dt\right] + \frac{1}{4}\mathbb{E}\left[\int_0^\infty e^{-t}\sum_{k=1}^{3} kY_t^{k,y}dt\right]. \tag{III.18}$$

Note that $Y^{4,y}$ is a martingale, and by differentiation and (III.16)

$$d\left(\sum_{k=1}^{3} kY_t^{k,y}\right) = dW_t + 2d\Lambda_t^{3,y}.$$

Therefore,

$$\mathbb{E}\left[\int_0^\infty e^{-t}\Phi\left(X_t^{x,\mathcal{C}}\right)dt\right] = \sum_{i=1}^{4}\frac{x_i}{4} + \frac{1}{4}\sum_{i=1}^{3} iy_i + \frac{1}{2}\mathbb{E}\left[\int_0^\infty e^{-t}\Lambda_t^{3,y}dt\right]. \tag{III.19}$$

Thus, by the definition of v, and a simple algebraic verification for the first two terms on the right, we have the equality (III.17). □

Remark 3.4.1. *One interpretation of the previous lemma is that the optimal strategy aims to maximize the third component of the local time of a reflected Brownian motion. This is consistent with discrete time problem in the case $N = 2$ or $N = 3$ where the optimal strategies of nature is proven to be maximizer of the number of crossings between the leading and the second leading experts [Cover 1967; Gravin, Peres, and Sivan 2016]. We note that this strategy also maximize the expected value of $\sum_{k=1}^{3} kY_\tau^{k,y}$ where τ is exponentially distributed.*

Proposition 3.4.1. *The function defined by*

$$v : y \in \mathbb{R}_+^3 \mapsto \frac{1}{2}\mathbb{E}\left[\int_0^\infty e^{-t}\Lambda_t^{3,y}dt\right] \tag{III.20}$$

is a viscosity solution of

$$0 = v - \frac{1}{2}\begin{pmatrix}1\\-1\\1\end{pmatrix}^T \partial^2 v \begin{pmatrix}1\\-1\\1\end{pmatrix}, \text{ on } (0,\infty)^3 \tag{III.21}$$

with the reflection conditions

$$\partial_3 v - \frac{1}{2}\partial_2 v = -\frac{1}{2} \textit{ if } y_3 = 0, \textit{ and } (y_1, y_2) \in (0,\infty)^2, \tag{III.22}$$

$$\partial_2 v - \frac{1}{2}(\partial_1 v + \partial_3 v) = 0 \textit{ if } y_2 = 0 \textit{ and } (y_1, y_3) \in (0,\infty)^2, \tag{III.23}$$

$$\partial_1 v - \frac{1}{2}\partial_2 v = 0 \textit{ if } y_1 = 0 \textit{ and } (y_2, y_3) \in (0,\infty)^2. \tag{III.24}$$

Proof. We introduce the auxiliary function

$$\tilde{v}(y) = \frac{1}{4}\sum_{i=1}^{3} i y_i + v(y) \text{ for all } y \in \mathbb{R}_+^3. \tag{III.25}$$

Thanks to (III.18) and (III.19), $\tilde{v}(y) = \frac{1}{4}\mathbb{E}\left[\int_0^\infty e^{-t}\sum_{k=1}^3 kY_t^{k,y}dt\right]$. For all stopping time $\tau \geq 0$, the dynamic programming principle leads to

$$\tilde{v}(y_1, y_2, y_3) = \mathbb{E}\left[\int_0^\tau \frac{e^{-t}}{4}\sum_{k=1}^{3} kY_t^{k,y}dt + e^{-\tau}\tilde{v}(Y_\tau^y)\right].$$

Using the martingality of $\mathcal{Y}^y$ on $(0,\infty)^3$, we obtain that on $(0,\infty)^3$,

$$0 = \tilde{v} - \frac{1}{4}\sum_{k=1}^{3} ky_k - \frac{1}{2}\begin{pmatrix}1\\-1\\1\end{pmatrix}^T \partial^2\tilde{v}\begin{pmatrix}1\\-1\\1\end{pmatrix} = v - \frac{1}{2}\begin{pmatrix}1\\-1\\1\end{pmatrix}^T \partial^2 v\begin{pmatrix}1\\-1\\1\end{pmatrix}.$$

For $\tilde{v}$ the reflection conditions are

$$\begin{aligned}\partial_3\tilde{v} - \frac{1}{2}\partial_2\tilde{v} &= 0, \quad \text{if } y_3 = 0, \text{ and } (y_1, y_2) \in (0,\infty)^2;\\ \partial_2\tilde{v} - \frac{1}{2}(\partial_1\tilde{v} + \partial_3\tilde{v}) &= 0, \quad \text{if } y_2 = 0 \text{ and } (y_1, y_3) \in (0,\infty)^2;\\ \partial_1\tilde{v} - \frac{1}{2}\partial_2\tilde{v} &= 0, \quad \text{if } y_1 = 0 \text{ and } (y_2, y_3) \in (0,\infty)^2.\end{aligned}$$

Thanks to (III.25), this yields (III.21)-(III.24) . □

3.5 Characterization of the value on the reflection boundary

We now characterize the function v via a system of hyperbolic first order PDE.

3.5.1 The value of v for $y_1 = y_3$

We start by characterizing v on the set $y_1 = y_3$.

Proposition 3.5.1. *The value function v satisfies*

$$v(y_1, y_2, y_1) = V(y_1, y_2), \text{ for } y_1 \geq 0,\ y_2 \geq 0,$$

where

$$V(y_1, y_2) := \frac{1}{2}\mathbb{E}\left[\int_0^\infty e^{-t}\Lambda_t^{1,(y_1,y_2)}dt\right], \tag{III.26}$$

and $\Lambda^{1,(y_1,y_2)}$ is the local time at 0 of the first component of the two dimensional obliquely reflected Brownian Motion $(Z_t^{1,(y_1,y_2)}, Z_t^{2,(y_1,y_2)})$ defined by

$$\begin{aligned} dZ_t^{1,(y_1,y_2)} &= dW_t + d\Lambda_t^{1,(y_1,y_2)} - \frac{1}{2}d\Lambda_t^{2,(y_1,y_2)}, \\ dZ_t^{2,(y_1,y_2)} &= -dW_t + d\Lambda_t^{2,(y_1,y_2)} - d\Lambda_t^{1,(y_1,y_2)}. \end{aligned} \tag{III.27}$$

Additionally, for all $y \in \mathbb{R}_+^2$, we have

$$\begin{aligned} V(y_1, y_2) = &\frac{\sqrt{2}}{2}\cosh(\sqrt{2}y_1)\cosh(\sqrt{2}(y_1+y_2))\arctan\left(e^{-\sqrt{2}(y_1+y_2)}\right) \\ &-\frac{\sqrt{2}}{4}\sinh(\sqrt{2}y_1). \end{aligned} \tag{III.28}$$

Proof. If $y_1 = y_3$ it is clear due to the uniqueness of the solution of the oblique reflection problem (III.16) that for all

$$Y_t^{1,y} = Y_t^{3,y} \text{ for all } t \geq 0$$

and the couple $(Y_t^{1,y}, Y_t^{2,y})$ solves the reflection problem (III.27). Thus, (III.26) holds. Additionally, using (III.26) we can derive the following dynamic programming equations for all $(y_1, y_2) \in (0, \infty)^2$,

$$V(y_1, y_2) - \begin{pmatrix} 1 \\ -1 \end{pmatrix}^T \partial^2 V(y_1, y_2) \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 0, \tag{III.29}$$

$$\partial_1 V(0, y_2) - \partial_2 V(0, y_2) = -\frac{1}{2}, \tag{III.30}$$

$$\frac{\partial_1 V(y_1, 0)}{2} - \partial_2 V(y_1, 0) = 0. \tag{III.31}$$

First, we compute the functions

$$V_1(x) := V(x, 0) \text{ and } V_2(x) := V(0, x).$$

Let $y_1 > 0$ and $y_2 > 0$ and define

$$\tau := \tau_1 \wedge \tau_2,$$

where $\tau_1 := \inf\{t \geq 0 : W_t \leq -y_1\}$ and $\tau_2 := \inf\{t \geq 0 : W_t \geq y_2\}$. Then, by the dynamic programming principle

$$\begin{aligned} V(y_1, y_2) &:= \mathbb{E}\left[e^{-\tau} V(\mathcal{Y}_\tau^1, \mathcal{Y}_\tau^2)\right] && \text{(III.32)} \\ &= \mathbb{E}\left[e^{-\tau} \mathbf{1}_{\tau_1 < \tau_2} V_2(y_1 + y_2)\right] + \mathbb{E}\left[e^{-\tau} \mathbf{1}_{\tau_1 > \tau_2} V_1(y_1 + y_2)\right] \\ &= \frac{\sinh(\sqrt{2} y_2)}{\sinh(\sqrt{2}(y_1 + y_2))} V_2(y_1 + y_2) + \frac{\sinh(\sqrt{2} y_1)}{\sinh(\sqrt{2}(y_1 + y_2))} V_1(y_1 + y_2) && \text{(III.33)} \end{aligned}$$

Assuming V is smooth we differentiate this equality in y_1, then in the expression we send

$y_1 \to 0$ for $y_2 > 0$ fixed to obtain

$$\partial_1 V(0, y_2) = \frac{\sqrt{2}}{\sinh(\sqrt{2}y_2)} V_1(y_2) - \frac{\sqrt{2}}{\tanh(\sqrt{2}y_2)} V_2(y_2) + V_2'(y_2).$$

One of the main point of the chapter is the fact that the equality (III.30) allows us to eliminate $\partial_1 V(0, y_2)$ so that we can write a system of differential equations for V_1 and V_2 as follows

$$\partial_2 V_2(0, y_2) - \frac{1}{2} = V_2'(y_2) - \frac{1}{2} = \frac{\sqrt{2}}{\sinh(\sqrt{2}y_2)} V_1(y_2) - \frac{\sqrt{2}}{\tanh(\sqrt{2}y_2)} V_2(y_2) + V_2'(y_2).$$

Similarly, differentiating (III.33) in y_2 and taking the limit as $y_2 \to 0$, we obtain that

$$\partial_2 V(y_1, 0) = -\frac{\sqrt{2}}{\tanh(\sqrt{2}y_1)} V_1(y_1) + \frac{\sqrt{2}}{\sinh(\sqrt{2}y_1)} V_2(y_1) + V_1'(y_1).$$

Additionally, the reflection conditions at (III.31) yield

$$\frac{V_1'(y_1)}{2} = \partial_2 V(y_1, 0) = -\frac{\sqrt{2}}{\tanh(\sqrt{2}y_1)} V_1(y_1) + \frac{\sqrt{2}}{\sinh(\sqrt{2}y_1)} V_2(y_1) + V_1'(y_1).$$

Combining both equalities we find that (V_1, V_2) solves the system

$$-\frac{1}{2} = \frac{\sqrt{2}}{\sinh(\sqrt{2}x)} V_1(x) - \frac{\sqrt{2}}{\tanh(\sqrt{2}x)} V_2(x), \tag{III.34}$$

$$0 = -\frac{\sqrt{2}}{\tanh(\sqrt{2}x)} V_1(x) + \frac{\sqrt{2}}{\sinh(\sqrt{2}x)} V_2(x) + \frac{V_1'(x)}{2}. \tag{III.35}$$

Combining the two equalities we obtain that V_1 is a solution to

$$0 = \frac{1}{\cosh(\sqrt{2}x)} - 2\sqrt{2}\tanh(\sqrt{2}x) V_1(x) + V_1'(x). \tag{III.36}$$

Given the antiderivative of the hyperbolic tangent, the solution to the homogeneous part of

(III.36) is $x \mapsto \cosh^2(\sqrt{2}x)$. Thus, we solve (III.36) under the form

$$V_1(x) = H(x)\cosh^2(\sqrt{2}x),$$

which imposes $H'(x) = \frac{-1}{\cosh^3(\sqrt{2}x)}$. Thus, for some constant C, V_1 is

$$V_1(x) = \left(C - \frac{1}{\sqrt{2}}\arctan\left(\tanh\left(\frac{x}{\sqrt{2}}\right)\right)\right)\cosh^2(\sqrt{2}x) - \frac{\sinh(\sqrt{2}x)}{2\sqrt{2}}.$$

With the choice $C = \frac{\pi}{4\sqrt{2}}$ we obtain that

$$V_1(x) = \frac{1}{\sqrt{2}}\left(\frac{\pi}{4} - \arctan\left(\tanh\left(\frac{x}{\sqrt{2}}\right)\right)\right)\cosh^2(\sqrt{2}x) - \frac{\sinh(\sqrt{2}x)}{2\sqrt{2}} \tag{III.37}$$

is the unique bounded solution to (III.36). Indeed, given the properties of the Gudermannian function, and arctan we have

$$\begin{aligned}\frac{\pi}{4} - \arctan\left(\tanh\left(\frac{x}{\sqrt{2}}\right)\right) &= \frac{\pi}{2} - \arctan\left(e^{\sqrt{2}x}\right)\\ &= \arctan\left(e^{-\sqrt{2}x}\right) = e^{-\sqrt{2}x} + o(e^{-2\sqrt{2}x})\end{aligned}$$

as $x \to \infty$. Thus, as $x \to \infty$,

$$\begin{aligned}&\frac{1}{\sqrt{2}}\left(\frac{\pi}{4} - \arctan\left(\tanh\left(\frac{x}{\sqrt{2}}\right)\right)\right)\cosh^2(\sqrt{2}x) - \frac{\sinh(\sqrt{2}x)}{2\sqrt{2}}\\ &\quad= \frac{1}{\sqrt{2}}\left(e^{-\sqrt{2}x} + o(e^{-2\sqrt{2}x})\right)\left(\frac{1}{4}e^{2\sqrt{2}x} + O(1)\right) - \frac{1}{4\sqrt{2}}\left(e^{\sqrt{2}x} + O(1)\right)\\ &= O(1)\end{aligned}$$

which shows that (III.37) is the unique bounded solution to (III.36). Injecting this into (III.34) and further simplifying we obtain that

$$
\begin{aligned}
V_1(x) &= \frac{1}{\sqrt{2}} \arctan\left(e^{-\sqrt{2}x}\right)\cosh^2(\sqrt{2}x) - \frac{\sinh(\sqrt{2}x)}{2\sqrt{2}},\\
V_2(x) &= \frac{1}{\sqrt{2}} \arctan\left(e^{-\sqrt{2}x}\right)\cosh(\sqrt{2}x).
\end{aligned}
$$

Thanks to (III.33), this finally yields (III.28). □

3.5.2 Deriving a Hyperbolic system to characterize the value on the boundary

We now return to the computation of v defined at (III.20) on $\mathbb{R}^3_+$. In order to compute v on the whole domain we first characterize its value on the boundary of this domain. For this purpose, we define for $x, y \geq 0$,

$$
\begin{aligned}
f(x,y) &= v\left(0, \frac{x}{\sqrt{2}}, \frac{y}{\sqrt{2}}\right), && \text{(III.38)}\\
r_1(x,y) &= v\left(\frac{x}{\sqrt{2}}, 0, \frac{y+x}{\sqrt{2}}\right), && \text{(III.39)}\\
h(x,y) &= v\left(\frac{y}{\sqrt{2}}, \frac{x}{\sqrt{2}}, 0\right) - \frac{1}{2\sqrt{2}}\left(1 + \frac{e^{-2x}}{3}\right), && \text{(III.40)}\\
r_2(x,y) &= v\left(\frac{x+y}{\sqrt{2}}, 0, \frac{x}{\sqrt{2}}\right) - \frac{2}{3\sqrt{2}}e^{-x}. && \text{(III.41)}
\end{aligned}
$$

The next proposition provides a characterization of these functions and allows us to compute the value function everywhere.

Proposition 3.5.2. *The couples* (f, r_1) *and* (h, r_2) *solve the same system of hyperbolic equations on* $(0,\infty)^2$

$$
\begin{aligned}
(\partial_x - 2\partial_y) f(x,y) &= \frac{2}{\tanh x} f(x,y) - \frac{2}{\sinh x} r_1(x,y), && \text{(III.42)}\\
\partial_x r_1(x,y) &= -\frac{2}{\sinh x} f(x,y) + \frac{2}{\tanh x} r_1(x,y), && \text{(III.43)}
\end{aligned}
$$

with the compatibility conditions

$$f(0,y) = r_1(0,y),\; h(0,y) = r_2(0,y)\,\textit{for}\; y > 0,$$

and initial conditions

$$\begin{aligned}
f(x,0) &= \frac{1}{\sqrt{2}}\arctan\left(e^{-x}\right)\cosh(x),\\
r_1(x,0) &= \frac{1}{\sqrt{2}}\arctan\left(e^{-x}\right)\cosh^2(x) - \frac{\sinh(x)}{2\sqrt{2}},\\
h(x,0) &= \frac{1}{\sqrt{2}}\arctan\left(e^{-x}\right)\cosh(x) - \frac{1}{2\sqrt{2}}\left(1 + \frac{e^{-2x}}{3}\right),\\
r_2(x,0) &= \frac{1}{\sqrt{2}}\arctan\left(e^{-x}\right)\cosh^2(x) - \frac{\sinh(x)}{2\sqrt{2}} - \frac{2}{3\sqrt{2}}e^{-x}\,\textit{for}\; x > 0.
\end{aligned}$$

Remark 3.5.1. *In the definition of* h *and* r_2 *the terms* $\frac{1}{2\sqrt{2}}\left(1 + \frac{e^{-2x}}{3}\right)$ *and* $\frac{\sinh(x)}{2\sqrt{2}} - \frac{2}{3\sqrt{2}}e^{-x}$ *are subtracted to eliminate* 1 *in equation* (III.46)*. This allow us to study one system of equation with two different initial condition rather than two systems with the same initial condition.*

Proof. Proceeding similarly as in (III.33), we obtain that for $0 \le y_1 \le y_3$ we have

$$\begin{aligned}
v(y_1,y_2,y_3) = v(0, y_2 + y_1, y_3 - y_1)&\frac{\sinh(\sqrt{2}y_2)}{\sinh(\sqrt{2}(y_1 + y_2))},\\
+ v(y_1 + y_2, 0, y_3 + y_2)&\frac{\sinh(\sqrt{2}y_1)}{\sinh(\sqrt{2}(y_1 + y_2))},
\end{aligned} \tag{III.44}$$

and for $0 \le y_3 \le y_1$,

$$\begin{aligned}
v(y_1,y_2,y_3) = v(y_1 - y_3, y_2 + y_3, 0)&\frac{\sinh(\sqrt{2}y_2)}{\sinh(\sqrt{2}(y_3 + y_2))},\\
+ v(y_1 + y_2, 0, y_3 + y_2)&\frac{\sinh(\sqrt{2}y_3)}{\sinh(\sqrt{2}(y_3 + y_2))}.
\end{aligned} \tag{III.45}$$

Let us first consider the case $0 \le y_1 \le y_3$. Similarly to the proof of (3.5.1), we differentiate

(III.44) in y_1, and send y_1 to 0, and obtain that

$$\partial_1 v(0,y_2,y_3) = \partial_2 v(0,y_2,y_3) - \partial_3 v(0,y_2,y_3) + v(0,y_2,y_3)\frac{-\sqrt{2}}{\tanh(\sqrt{2}y_2)} \\ + v(y_2,0,y_2+y_3)\frac{\sqrt{2}}{\sinh(\sqrt{2}y_2)}$$

Additionally, the reflection conditions (III.24) gives

$$(2\partial_3 - \partial_2)v(0,y_2,y_3) = \frac{2\sqrt{2}}{\sinh(\sqrt{2}y_2)}v(y_2,0,y_3+y_2) - \frac{2\sqrt{2}}{\tanh(\sqrt{2}y_2)}v(0,y_2,y_3)$$

Then we differentiate (III.44) in y_2 and send y_2 to 0 to obtain

$$\partial_2 v(y_1,0,y_3) = \partial_1 v(y_1,0,y_3) + \partial_3 v(y_1,0,y_3) + v(0,y_1,y_3-y_1)\frac{\sqrt{2}}{\sinh(\sqrt{2}y_1)} \\ + v(y_1,0,y_3)\frac{-\sqrt{2}}{\tanh(\sqrt{2}y_1)}.$$

The reflection conditions (III.23) yields

$$(\partial_1 + \partial_3)v(y_1,0,y_3) = \frac{2\sqrt{2}}{\tanh(\sqrt{2}y_1)}v(y_1,0,y_3) - \frac{2\sqrt{2}}{\sinh(\sqrt{2}y_1)}v(0,y_1,y_3-y_1).$$

Combining both equalities, and write them in $f(x,y)$ and $r_1(x,y)$, we get the desired system:

$$\begin{aligned}(\partial_x - 2\partial_y)f(x,y) &= \frac{2}{\tanh x}f(x,y) - \frac{2}{\sinh x}r_1(x,y),\\ \partial_x r_1(x,y) &= -\frac{2}{\sinh x}f(x,y) + \frac{2}{\tanh x}r_1(x,y).\end{aligned}$$

Let us now consider the case $0 \le y_3 \le y_1$. Following a similar procedure as before, we differentiate (III.45) in y_2, and send y_2 to 0 to obtain

$$\begin{aligned}\partial_2 v(y_1,0,y_3) =&\partial_1 v(y_1,0,y_3)+\partial_3 v(y_1,0,y_3)+v(y_1-y_3,y_3,0)\frac{\sqrt{2}}{\sinh(\sqrt{2}y_3)}\\ &+v(y_1,0,y_3)\frac{-\sqrt{2}}{\tanh(\sqrt{2}y_3)}.\end{aligned}$$

Additionally, the reflection conditions (III.23) gives

$$(\partial_1+\partial_3)v(y_1,0,y_3)=\frac{2\sqrt{2}}{\tanh\left(\sqrt{2}y_3\right)}v(y_1,0,y_3)-\frac{2\sqrt{2}}{\sinh\left(\sqrt{2}y_3\right)}v(y_1-y_3,y_3,0).$$

Then we differentiate (III.45) in y_3 and send y_3 to 0 and obtain

$$\begin{aligned}\partial_3 v(y_1,y_2,0) =&\partial_2 v(y_1,y_2,0)-\partial_1 v(y_1,y_2,0)+v(y_1,y_2,0)\frac{-\sqrt{2}}{\tanh(\sqrt{2}y_2))}\\ &+v(y_1+y_2,0,y_2)\frac{\sqrt{2}}{\sinh(\sqrt{2}y_2)}.\end{aligned}$$

The reflection conditions (III.22) gives

$$(2\partial_1-\partial_2)v(y_1,y_2,0)=1-\frac{2\sqrt{2}v(y_1,y_2,0)}{\tanh(\sqrt{2}y_2)}+\frac{2\sqrt{2}v(y_1+y_2,0,y_2)}{\sinh(\sqrt{2}y_2)}. \tag{III.46}$$

Combining both equalities, and write them in $h(x,y)$ and $r_2(x,y)$, we have the desired system:

$$\begin{aligned}(\partial_x-2\partial_y)h(x,y)&=\frac{2}{\tanh x}h(x,y)-\frac{2}{\sinh x}r_2(x,y),\\ \partial_x r_2(x,y)&=-\frac{2}{\sinh x}h(x,y)+\frac{2}{\tanh x}r_2(x,y).\end{aligned}$$

The compatibility conditions and initial conditions follows form the change of variable described at the beginning of this section and Proposition 3.5.1. □

3.5.3 Solving the Hyperbolic system

Although first order and linear, the system (III.42) can not be directly solved via the method of characteristics since the characteristics for the two equations are not in the same direction. Additionally, we cannot employ methods described in [Tsarev 2007] and [Fusco and Manganaro 1996].

3.5.3.1 Heuristic to find an ansatz of the solution

We first note that if f is given then thanks to (III.43), r solves a linear ODE whose unique solutions that is bounded at infinity is

$$r_1(x,y) = 2\sinh^2(x)\int_x^\infty \frac{f(r,y)}{\sinh^3(r)}dr. \tag{III.47}$$

$\{f(x,0)\}_{x\geq 0}$ being given, we can easily obtain $\{r_1(x,0)\}_{x\geq 0}$ by integration. This allows us to compute $\{\partial_y f(x,0)\}_{x\geq 0}$ by isolating it in (III.42).

Since the system does not depend on y we can differentiate in y. Thus, we can compute $\{\partial_y^2 f(x,0)\}_{x\geq 0}$ with a similar procedure if we start with initial condition $\{\partial_y f(x,0)\}_{x\geq 0}$. Then, we can repeat the procedure to compute several derivatives $\{\partial_y^n f(x,0)\}_{x\geq 0}$.

Additionally thanks to the form of solutions in [Iskenderov and Mamedov n.d.], we expect that the solutions f and r are functions of $x+\frac{y}{2}$ and $\frac{y}{2}$. Combining this with the computation of the derivatives $\{\partial_y^n f(x,0)\}_{x\geq 0}$ we conjecture that

$$\begin{aligned} f(x,y) =& h_1\left(\frac{y}{2}\right)\arctan\left(e^{-x-\frac{y}{2}}\right)\cosh\left(x+\frac{y}{2}\right) \\ &+ h_2\left(\frac{y}{2}\right)\operatorname{arctanh}\left(e^{-x-\frac{y}{2}}\right)\sinh\left(x+\frac{y}{2}\right) + h_3\left(\frac{y}{2}\right) \end{aligned}$$

with the condition

$$h_1(0) = \frac{1}{\sqrt{2}},\ h_2(0) = h_3(0) = 0.$$

3.5.3.2 Solution to the systems

Given the ansatz for f, one can integrate (III.47) to find that r then (III.42) leads to [2]

$$\begin{aligned}
&2\operatorname{arctanh}(e^{-x})\sinh(x)\left(\coth\left(\frac{y}{2}\right)h_2\left(\frac{y}{2}\right)+2h_3\left(\frac{y}{2}\right)+h_1\left(\frac{y}{2}\right)\tanh\left(\frac{y}{2}\right)\right)\\
&+\arctan\left(e^{-x-\frac{y}{2}}\right)\cosh\left(x+\frac{y}{2}\right)\left(2h_1\left(\frac{y}{2}\right)\tanh\left(\frac{y}{2}\right)-h_1'\left(\frac{y}{2}\right)\right)\\
&+\operatorname{arctanh}\left(e^{-x-\frac{y}{2}}\right)\sinh\left(x+\frac{y}{2}\right)\left(2h_2\left(\frac{y}{2}\right)\coth\left(\frac{y}{2}\right)-h_2'\left(\frac{y}{2}\right)\right)\\
&-h_1\left(\frac{y}{2}\right)-h_2\left(\frac{y}{2}\right)-h_3'\left(\frac{y}{2}\right)=0. \qquad \text{(III.48)}
\end{aligned}$$

Setting the second and the third lines to 0, we solve the ODE obtained for h_1 and h_2 with the initial condition to obtain that

$$h_1(y)=\frac{1}{\sqrt{2}}\cosh^2(y) \text{ and } h_2(y)=C\sinh^2(y) \text{ for some constant } C.$$

Injecting this to the first line, the term in parentheses in the first line becomes

$$\left(\frac{C}{2}+\frac{1}{2\sqrt{2}}\right)\sinh\left(y\right)+2h_3\left(\frac{y}{2}\right).$$

This allows us to identify

$$h_3\left(y\right)=-\frac{C\sqrt{2}+1}{4\sqrt{2}}\sinh\left(2y\right).$$

Thus, to satisfy (III.48) we need

$$\frac{1}{\sqrt{2}}\cosh^2(y)+C\sinh^2(y)-\frac{C\sqrt{2}+1}{2\sqrt{2}}\cosh\left(2y\right)=0$$

[2] This computation could be extremely tedious by hand. We have checked the identity with Mathematica V11. The code for this verification and other tedious computations are provided in [Bayraktar, Ekren, and Y. Zhang n.d.].

which is satisfied for $C = \frac{1}{\sqrt{2}}$. Thus, we obtain f as

$$\begin{aligned} f(x,y) = &\frac{1}{\sqrt{2}}\left(\arctan(e^{-x-\frac{y}{2}})\cosh(x+\frac{y}{2})\cosh^2(\frac{y}{2})\right) \\ &+\frac{1}{\sqrt{2}}\left(\operatorname{arctanh}(e^{-x-\frac{y}{2}})\sinh(x+\frac{y}{2})\sinh^2(\frac{y}{2})-\frac{1}{2}\sinh(y)\right). \end{aligned} \tag{III.49}$$

Injecting this expression in (III.47) we obtain

$$\begin{aligned} r_1(x,y) = &\frac{1}{\sqrt{2}}\left(\arctan(e^{-x-\frac{y}{2}})\cosh^2(x+\frac{y}{2})\cosh(\frac{y}{2})\right) \\ &+\frac{1}{\sqrt{2}}\left(\operatorname{arctanh}(e^{-x-\frac{y}{2}})\sinh^2(x+\frac{y}{2})\sinh(\frac{y}{2})-\frac{1}{2}\sinh(x+y)\right). \end{aligned} \tag{III.50}$$

Using the same method we can also solve the system (III.42)-(III.43) with initial condition

$$\left(\frac{1}{2\sqrt{2}}\left(1+\frac{e^{-2x}}{3}\right), \frac{2}{3\sqrt{2}}e^{-x}\right),$$

then using the linearity of the system subtract this from (f, r) to obtain

$$\begin{aligned} h(x,y) = &\frac{1}{\sqrt{2}}\left(\arctan(e^{-x-\frac{y}{2}})\cosh(x+\frac{y}{2})\cosh^2(\frac{y}{2})-\frac{1}{2}-\frac{e^{-2x}}{6}\right) \\ &-\frac{1}{\sqrt{2}}\operatorname{arctanh}(e^{-x-\frac{y}{2}})\sinh(x+\frac{y}{2})\sinh^2(\frac{y}{2}), \end{aligned} \tag{III.51}$$

$$\begin{aligned} r_2(x,y) = &\frac{1}{\sqrt{2}}\left(\arctan(e^{-x-\frac{y}{2}})\cosh^2(x+\frac{y}{2})\cosh(\frac{y}{2})-2\frac{\cosh(x)}{3}\right) \\ &-\frac{1}{\sqrt{2}}\left(\operatorname{arctanh}(e^{-x-\frac{y}{2}})\sinh^2(x+\frac{y}{2})\sinh(\frac{y}{2})-\frac{\sinh(x)}{6}\right). \end{aligned} \tag{III.52}$$

The reader may find in [Bayraktar, Ekren, and Y. Zhang n.d.], the Mathematica code to check that (III.49)-(III.52) provides solutions to the system (III.42) and (III.43). Combining (III.49), (III.50), (III.51) and (III.52), we now give the expression of v.

Proposition 3.5.3. *The function v defined at* (III.20) *is given by*

$$
\begin{aligned}
v(y_1, y_2, y_3) = &- \frac{\sqrt{2}}{4}\sinh(\sqrt{2}y_3) \\
&+ \frac{\sqrt{2}}{2}\arctan\left(e^{-\frac{y_1+2y_2+y_3}{\sqrt{2}}}\right) \\
&\cosh\left(\frac{-y_1+y_3}{\sqrt{2}}\right)\cosh\left(\frac{y_1+2y_2+y_3}{\sqrt{2}}\right)\cosh\left(\frac{y_1+y_3}{\sqrt{2}}\right) \\
&+ \frac{\sqrt{2}}{2}\operatorname{arctanh}\left(e^{-\frac{y_1+2y_2+y_3}{\sqrt{2}}}\right) \\
&\sinh\left(\frac{-y_1+y_3}{\sqrt{2}}\right)\sinh\left(\frac{y_1+2y_2+y_3}{\sqrt{2}}\right)\sinh\left(\frac{y_1+y_3}{\sqrt{2}}\right)
\end{aligned}
$$

Remark 3.5.2. *For reader's convenience we provide in [Bayraktar, Ekren, and Y. Zhang n.d.] the Mathematica code to check that this expression provides a solution to the equations* (III.21) *and* (III.22)-(III.24).

Proof. The proof is a direct consequence of identities (III.44)-(III.45) and (III.49)-(III.52).We inject $f(x, y)$ and $r_1(x, y)$ to obtain v for $0 \leq y_1 \leq y_3$, i.e.

$$v(y_1, y_2, y_3)$$

$$= f(\sqrt{2}(y_1 + y_2)), \sqrt{2}(y_3 - y_1)\frac{\sinh(\sqrt{2}y_2)}{\sinh(\sqrt{2}(y_1 + y_2))}$$

$$+ r_1(\sqrt{2}(y_1 + y_2), \sqrt{2}(y_3 - y_1))\frac{\sinh(\sqrt{2}y_1)}{\sinh(\sqrt{2}(y_1 + y_2))}$$

$$= \frac{\sqrt{2}}{2}\arctan(e^{-\frac{y_1+2y_2+y_3}{\sqrt{2}}})\cosh(\frac{-y_1 + y_3}{\sqrt{2}})\cosh^2(\frac{y_1 + 2y_2 + y_3}{\sqrt{2}})$$

$$\times \operatorname{csch}(\sqrt{2}(y_1 + y_2))\sinh(\sqrt{2}y_1)$$

$$+ \frac{\sqrt{2}}{2}\arctan(e^{-\frac{y_1+2y_2+y_3}{\sqrt{2}}})\cosh^2(\frac{-y_1 + y_3}{\sqrt{2}})\cosh(\frac{y_1 + 2y_2 + y_3}{\sqrt{2}})$$

$$\times \operatorname{csch}(\sqrt{2}(y_1 + y_2))\sinh(\sqrt{2}y_2)$$

$$- \frac{\sqrt{2}}{4}\operatorname{csch}(\sqrt{2}(y_1 + y_2))\sinh(\sqrt{2}y_2)\sinh(\sqrt{2}(-y_1 + y_3))$$

$$- \frac{\sqrt{2}}{4}\operatorname{csch}(\sqrt{2}(y_1 + y_2))\sinh(\sqrt{2}y_1)\sinh(\sqrt{2}(y_2 + y_3))$$

$$+ \frac{\sqrt{2}}{2}\operatorname{arctanh}(e^{-\frac{y_1+2y_2+y_3}{\sqrt{2}}})\operatorname{csch}(\sqrt{2}(y_1 + y_2))\sinh(\sqrt{2}y_2)$$

$$\times \sinh^2(\frac{-y_1 + y_3}{\sqrt{2}})\sinh(\frac{y_1 + 2y_2 + y_3}{\sqrt{2}})$$

$$+ \frac{\sqrt{2}}{2}\operatorname{arctanh}(e^{-\frac{y_1+2y_2+y_3}{\sqrt{2}}})\operatorname{csch}(\sqrt{2}(y_1 + y_2))\sinh(\sqrt{2}y_1)$$

$$\times \sinh(\frac{-y_1 + y_3}{\sqrt{2}})\sinh^2(\frac{y_1 + 2y_2 + y_3}{\sqrt{2}}).$$

And injecting $h(x, y)$ and $r_2(x, y)$ we obtain v for $0 \leq y_3 \leq y_1$

$$
\begin{aligned}
&v(y_1, y_2, y_3) \\
&= \left(h(\sqrt{2}(y_2 + y_3), \sqrt{2}(y_1 - y_3)) + \frac{1}{2\sqrt{2}}(1 + \frac{e^{2\sqrt{2}(y_2+y_3)}}{3}) \right) \frac{\sinh(\sqrt{2}y_2)}{\sinh(\sqrt{2}(y_3 + y_2))} \\
&+ \left(r_2(\sqrt{2}(y_2 + y_3), \sqrt{2}(y_1 - y_3) + \frac{2}{3\sqrt{2}} e^{-\sqrt{2}(y_2+y_3)}) \right) \frac{\sinh(\sqrt{2}y_3)}{\sinh(\sqrt{2}(y_3 + y_2))} \\
&= \frac{\sqrt{2}}{2} \arctan(e^{-\frac{y_1+2y_2+y_3}{\sqrt{2}}}) \cosh^2(\frac{y_1 - y_3}{\sqrt{2}}) \cosh(\frac{y_1 + 2y_2 + y_3}{\sqrt{2}}) \operatorname{csch}(\sqrt{2}(y_3 + y_2)) \sinh(\sqrt{2}y_2) \\
&+ \frac{\sqrt{2}}{2} \arctan(e^{-\frac{y_1+2y_2+y_3}{\sqrt{2}}}) \cosh(\frac{y_1 - y_3}{\sqrt{2}}) \cosh^2(\frac{y_1 + 2y_2 + y_3}{\sqrt{2}}) \operatorname{csch}(\sqrt{2}(y_3 + y_2)) \sinh(\sqrt{2}y_3) \\
&- \frac{\sqrt{2}}{4} \sinh(\sqrt{2}y_3) \\
&- \frac{\sqrt{2}}{2} \operatorname{arctanh}(e^{-\frac{y_1+2y_2+y_3}{\sqrt{2}}}) \operatorname{csch}(\sqrt{2}(y_3 + y_2)) \sinh(\sqrt{2}y_2) \sinh^2(\frac{y_1 - y_3}{\sqrt{2}}) \sinh(\frac{y_1 + 2y_2 + y_3}{\sqrt{2}}) \\
&- \frac{\sqrt{2}}{2} \operatorname{arctanh}(e^{-\frac{y_1+2y_2+y_3}{\sqrt{2}}}) \operatorname{csch}(\sqrt{2}(y_3 + y_2)) \sinh(\sqrt{2}y_3) \sinh(\frac{y_1 - y_3}{\sqrt{2}}) \sinh^2(\frac{y_1 + 2y_2 + y_3}{\sqrt{2}})
\end{aligned}
$$

Note that these expressions can be simplified and combined into one expression on the whole space $0 \leq y_1, y_2, y_3$

$$
\begin{aligned}
&v(y_1, y_2, y_3) \\
&= \frac{\sqrt{2}}{2} \arctan\left(e^{-\frac{y_1+2y_2+y_3}{\sqrt{2}}}\right) \cosh\left(\frac{-y_1 + y_3}{\sqrt{2}}\right) \cosh\left(\frac{y_1 + 2y_2 + y_3}{\sqrt{2}}\right) \cosh\left(\frac{y_1 + y_3}{\sqrt{2}}\right) \\
&- \frac{\sqrt{2}}{4} \sinh(\sqrt{2}y_3) \\
&+ \frac{\sqrt{2}}{2} \operatorname{arctanh}\left(e^{-\frac{y_1+2y_2+y_3}{\sqrt{2}}}\right) \sinh\left(\frac{-y_1 + y_3}{\sqrt{2}}\right) \sinh\left(\frac{y_1 + 2y_2 + y_3}{\sqrt{2}}\right) \sinh\left(\frac{y_1 + y_3}{\sqrt{2}}\right).
\end{aligned}
$$

□

We will close this section by giving a minimum principle for the supersolutions of the system (III.42)-(III.43), which we will need in the next section when proving our main result.

Proposition 3.5.4. *Let $F, R : [0, \infty)^2 \mapsto \mathbb{R}$ be functions that are continuous on their domain* *and continuously differentiable in the interior of their domain. Assume that for all* $x, y \geq 0$*,*

$$F(x,0) \geq 0, \; F(0,y) \geq 0, \; \liminf_{r^2+s^2\to\infty} F(r,s) \geq 0, \; \textit{and} \; \lim_{r\to\infty} R(r,y) = 0.$$

Assume also that F, R *are supersolution of* (III.42)-(III.43) *in the sense*

$$(\partial_x - 2\partial_y)F(x,y) \leq \frac{2}{\tanh x} F(x,y) - \frac{2}{\sinh x} R(x,y) \tag{III.53}$$

$$\partial_x R(x,y) \leq -\frac{2}{\sinh x} F(x,y) + \frac{2}{\tanh x} R(x,y) \tag{III.54}$$

Then $F(x,y) \geq 0$ *and* $R(x,y) \geq 0$ *for all* $x, y \geq 0$*.*

Proof. To obtain a contradiction we first assume that F is negative at some point on its domain. Therefore, by the values of this function on the boundary of the domain, its minimum on $[0,\infty)^2$ is achieved and there exists $(x_0, y_0) \in (0,\infty)^2$ and $\delta > 0$ such that

$$\inf_{x,y\in[0,\infty)} F(x,y) = F(x_0,y_0) = -\delta < 0.$$

Thanks to (III.54) we can write

$$\partial_x R(x,y) = -\frac{2}{\sinh x} F(x,y) + \frac{2}{\tanh x} R(x,y) - P(x,y)$$

for some $P \geq 0$ and continuous. We solve this ODE to obtain similarly to (III.47) that

$$\begin{aligned} R(x,y) &= \sinh^2(x) \int_x^\infty \frac{2F(r,y)}{\sinh^3(r)} + \frac{P(r,y)}{\sinh^2(r)} dr \geq 2\sinh^2(x) \int_x^\infty \frac{F(r,y)}{\sinh^3(r)} dr \\ &\geq 2 \inf_{r\in[x,\infty]} F(r,y) \sinh^2(x) \int_x^\infty \frac{1}{\sinh^3(r)} dr. \end{aligned} \tag{III.55}$$

We have the identity

$$2\sinh^2(x)\int_x^\infty \frac{1}{\sinh^3(r)}dr = \cosh(x) - 2\operatorname{arctanh}(e^{-x})\sinh^2(x) \in [0,1] \text{ for all } x > 0.$$

Thus,

$$R(x_0,y_0) \geq 2\inf_{r\in[x_0,\infty]} F(r,y_0)\sinh^2(x_0)\int_{x_0}^\infty \frac{1}{\sinh^3(r)}dr \geq \inf_{r\in[x_0,\infty]} F(r,y_0) = -\delta \quad \text{(III.56)}$$

where the last inequality is due to the fact that

$$\inf_{r\in[x_0,\infty]} F(r,y_0) = -\delta < 0.$$

The minimality of F at $(x_0,y_0) \in (0,\infty)^2$ and the differentiability of F (which implies that $\partial_x F(x_0,y_0) = \partial_y F(x_0,y_0) = 0$) combined with (III.53) allows us to claim that

$$\cosh(x_0)F(x_0,y_0) \geq R(x_0,y_0).$$

Then, the inequality (III.56) yields

$$-\delta\cosh(x_0) = \cosh(x_0)F(x_0,y_0) \geq R(x_0,y_0) \geq -\delta$$

which is in contradiction with $x_0 > 0$. Thus, $F \geq 0$. Combining this inequality with (III.55), we obtain that $R \geq 0$. □

3.6 Regularity of u and proof of the main theorems

In this section we use the expression of v to define the candidate solution to the PDE (III.5). Let $\mathcal{W}_4 := \{x \in \mathbb{R}^d : x_1 < x_2 < x_3 < x_4\}$ and define

$$\begin{aligned} U : x \in \mathcal{W}_4 \mapsto & x_4 + v(x_2 - x_1, x_3 - x_2, x_4 - x_3) = x_4 - \frac{\sqrt{2}}{4}\sinh(\sqrt{2}(x_4 - x_3)) \\ & + \frac{\sqrt{2}}{2}\arctan\left(e^{\frac{x_1+x_2-x_3-x_4}{\sqrt{2}}}\right)\cosh\left(\frac{x_1 - x_2 + x_3 - x_4}{\sqrt{2}}\right) \\ & \cosh\left(\frac{-x_1 + x_2 + x_3 - x_4}{\sqrt{2}}\right)\cosh\left(\frac{-x_1 - x_2 + x_3 + x_4}{\sqrt{2}}\right) \\ & + \frac{\sqrt{2}}{2}\operatorname{arctanh}\left(e^{\frac{x_1+x_2-x_3-x_4}{\sqrt{2}}}\right)\sinh\left(\frac{x_1 - x_2 + x_3 - x_4}{\sqrt{2}}\right) \\ & \sinh\left(\frac{-x_1 + x_2 + x_3 - x_4}{\sqrt{2}}\right)\sinh\left(\frac{-x_1 - x_2 + x_3 + x_4}{\sqrt{2}}\right) \end{aligned} \tag{III.57}$$

so that

$$u(x) = U(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}) \text{ for } x \in \mathbb{R}^4. \tag{III.58}$$

We give the following proposition for the regularity of u and U.

Proposition 3.6.1. *U has a C^2 extension to $\bar{\mathcal{W}}_4$ and the extension satisfies for all $x \in \bar{\mathcal{W}}_4$,*

$$\partial_1 U(x_1, x_1, x_3, x_4) = \partial_2 U(x_1, x_1, x_3, x_4), \tag{III.59}$$

$$\partial_2 U(x_1, x_2, x_2, x_4) = \partial_3 U(x_1, x_2, x_2, x_4), \tag{III.60}$$

$$\partial_3 U(x_1, x_2, x_3, x_3) = \partial_4 U(x_1, x_2, x_3, x_3). \tag{III.61}$$

Additionally, u defined by (III.7) *is C^2 on $\mathbb{R}^4$ and U satisfies*

$$0 = U(x) - \Phi(x) - \frac{1}{2} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}^T \partial^2 U(x) \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} \text{ for all } x \in \mathcal{W}_4. \tag{III.62}$$

Remark 3.6.1. *As needed for the smoothness of u, U is symmetric in its variables.*

Proof. The main problem with the existence of the extension of U is the fact that the function $z \mapsto \operatorname{arctanh}(e^z)$ has a singularity at 0. Thus, the C^2 extension a priori only exists whenever all the components are not equal to each other.

For the points where all the components are equal to each other we use the fact that $\operatorname{arctanh}(e^z)\sinh(z) \to 0$ as $z \downarrow 0$. Thus, the last two lines of (III.57) goes to 0 as x converges to a point whose components are equal. This shows that there is a continuous extension of U to $\bar{\mathcal{W}}_4$.

To show that the extension is C^1 it is now sufficient to show that all partial derivatives admits finite limits as we take the limit to the boundary of $\mathcal{W}_4$, in particular, when $x_1 = x_2 = x_3 = x_4$. First, we observe that

$$\begin{aligned} G(x_1, x_2, x_3, x_4) =& x_4 - \frac{\sqrt{2}}{4} \sinh(\sqrt{2}(x_4 - x_3)) \\ &+ \frac{\sqrt{2}}{2} \arctan\left(e^{\frac{x_1+x_2-x_3-x_4}{\sqrt{2}}}\right) \cosh\left(\frac{x_1 - x_2 + x_3 - x_4}{\sqrt{2}}\right) \\ &\cosh\left(\frac{-x_1 + x_2 + x_3 - x_4}{\sqrt{2}}\right) \cosh\left(\frac{-x_1 - x_2 + x_3 + x_4}{\sqrt{2}}\right) \end{aligned} \tag{III.63}$$

is analytic everywhere so we only need to consider the behavior of

$$\begin{aligned}T(x_1,x_2,x_3,x_4) =&\frac{\sqrt{2}}{2}\operatorname{arctanh}\left(e^{\frac{x_1+x_2-x_3-x_4}{\sqrt{2}}}\right)\sinh\left(\frac{x_1-x_2+x_3-x_4}{\sqrt{2}}\right)\\ &\sinh\left(\frac{-x_1+x_2+x_3-x_4}{\sqrt{2}}\right)\sinh\left(\frac{-x_1-x_2+x_3+x_4}{\sqrt{2}}\right)\end{aligned}$$

at a point satisfying $x_1 = x_2 = x_3 = x_4$. By chain rule, the fist order partial derivatives of U are linear combinations of the following 4 terms:

$$\begin{aligned}t_1(x_1,x_2,x_3,x_4) =&\operatorname{arctanh}\left(e^{\frac{x_1+x_2-x_3-x_4}{\sqrt{2}}}\right)\sinh\left(\frac{x_1-x_2+x_3-x_4}{\sqrt{2}}\right)\\ &\sinh\left(\frac{-x_1+x_2+x_3-x_4}{\sqrt{2}}\right)\cosh\left(\frac{-x_1-x_2+x_3+x_4}{\sqrt{2}}\right)\\ t_2(x_1,x_2,x_3,x_4) =&\operatorname{arctanh}\left(e^{\frac{x_1+x_2-x_3-x_4}{\sqrt{2}}}\right)\sinh\left(\frac{x_1-x_2+x_3-x_4}{\sqrt{2}}\right)\\ &\cosh\left(\frac{-x_1+x_2+x_3-x_4}{\sqrt{2}}\right)\sinh\left(\frac{-x_1-x_2+x_3+x_4}{\sqrt{2}}\right)\\ t_3(x_1,x_2,x_3,x_4) =&\operatorname{arctanh}\left(e^{\frac{x_1+x_2-x_3-x_4}{\sqrt{2}}}\right)\cosh\left(\frac{x_1-x_2+x_3-x_4}{\sqrt{2}}\right)\\ &\sinh\left(\frac{-x_1+x_2+x_3-x_4}{\sqrt{2}}\right)\sinh\left(\frac{-x_1-x_2+x_3+x_4}{\sqrt{2}}\right)\\ t_4(x_1,x_2,x_3,x_4) =&\frac{e^{\frac{x_1+x_2-x_3-x_4}{\sqrt{2}}}}{1-e^{\sqrt{2}(x_1+x_2-x_3-x_4)}}\sinh\left(\frac{x_1-x_2+x_3-x_4}{\sqrt{2}}\right)\\ &\sinh\left(\frac{-x_1+x_2+x_3-x_4}{\sqrt{2}}\right)\sinh\left(\frac{-x_1-x_2+x_3+x_4}{\sqrt{2}}\right).\end{aligned}$$

In $\mathcal{W}$, as $x_1 < x_2 < x_3 < x_4$, we have the inequalities

$$0 \geq x_1 - x_2 + x_3 - x_4 \geq x_1 + x_2 - x_3 - x_4$$

$$-x_1 - x_2 + x_3 + x_4 \geq -x_1 + x_2 + x_3 - x_4 \geq x_1 + x_2 - x_3 - x_4.$$

Combined with the equality $|\sinh(x)| = \sinh(|x|)$, these inequalities yield

$$\left|\sinh\left(\frac{x_1 - x_2 + x_3 - x_4}{\sqrt{2}}\right)\right| \leq \left|\sinh\left(\frac{x_1 + x_2 - x_3 - x_4}{\sqrt{2}}\right)\right|$$
$$\left|\sinh\left(\frac{-x_1 + x_2 + x_3 - x_4}{\sqrt{2}}\right)\right| \leq \left|\sinh\left(\frac{x_1 + x_2 - x_3 - x_4}{\sqrt{2}}\right)\right|.$$

Using the observation that $\operatorname{arctanh}(e^z)\sinh(z) \to 0$ as $z \downarrow 0$ one more time, and the limit $\frac{\sinh(z/\sqrt{2})}{1-e^{\sqrt{2}z}} \to \frac{1}{2}$, as $z \downarrow 0$ we can conclude that each of t_1, t_2, t_3, and $t_4 \to 0$ as x converge to a point where components are equal to each other. Thus, we have showed that T has a C^1 extension to $\bar{\mathcal{W}}_4$ and in fact all its first order partial derivatives are 0 on $x_1 = x_2 = x_3 = x_4$.

Similarly, using these observations, one can also show that all the second order partial derivatives of U have continuous extension on $x_1 = x_2 = x_3 = x_4$ and all second order partial derivatives of T are 0 on $x_1 = x_2 = x_3 = x_4$ as well.

We now use the reflection conditions (III.22), (III.23), and (III.24) to show that on the boundaries $x_1 = x_2$, $x_2 = x_3$, $x_3 = x_4$, the first order partial derivatives of U satisfy (III.59), (III.60), and (III.61).

Since $U(x_1, x_2, x_3, x_4) = x_4 + v(x_2 - x_1, x_3 - x_2, x_4 - x_3)$ using (III.24) we obtain that

$$\begin{aligned}
&\partial_1 U(x_1, x_1, x_3, x_4) - \partial_2 U(x_1, x_1, x_3, x_4) \\
&= -\partial_1 v(0, x_3 - x_1, x_4 - x_3) - (\partial_1 v(0, x_3 - x_1, x_4 - x_3) - \partial_2 v(0, x_3 - x_1, x_4 - x_3)) \\
&= -2\partial_1 v(0, x_3 - x_1, x_4 - x_3) + \partial_2 v(0, x_3 - x_1, x_4 - x_3) = 0.
\end{aligned}$$

Using (III.23) we obtain

$$\begin{aligned}
&\partial_2 U(x_1, x_2, x_2, x_4) - \partial_3 U(x_1, x_2, x_2, x_4)\\
&= \partial_1 v(x_2 - x_1, 0, x_4 - x_2) - \partial_2 v(x_2 - x_1, 0, x_4 - x_2)\\
&\qquad - (\partial_2 v(x_2 - x_1, 0, x_4 - x_2) - \partial_3 v(x_2 - x_1, 0, x_4 - x_2))\\
&= \partial_1 v(x_2 - x_1, 0, x_4 - x_2) - 2\partial_2 v(x_2 - x_1, 0, x_4 - x_2) + \partial_3 v(x_2 - x_1, 0, x_4 - x_2) = 0.
\end{aligned}$$

On the other hand (III.22) gives

$$\begin{aligned}
&\partial_3 U(x_1, x_2, x_3, x_3) - \partial_4 U(x_1, x_2, x_3, x_3)\\
&= \partial_2 v(x_2 - x_1, x_3 - x_2, 0) - \partial_3 v(x_2 - x_1, x_3 - x_2, 0) - (1 + \partial_3 v(x_2 - x_1, x_3 - x_2, 0))\\
&= -1 + \partial_2 v(x_2 - x_1, x_3 - x_2, 0) - 2\partial_3 v(x_2 - x_1, x_3 - x_2, 0) = 0.
\end{aligned}$$

Thus, U has a C^2 extension to $\bar{\mathcal{W}}_4$, its first order partial derivatives satisfy (III.59)-(III.61) and the first two order of partial derivatives of T are 0 on $x_1 = x_2 = x_3 = x_4$.

We now show that u defined by (III.58) or (III.7) is C^2 on $\mathbb{R}^4$. The smoothness of U and the equalities (III.59)-(III.61) implies that u is C^1. In order to show that u is C^2 we need to show that for any point $x \in \bar{\mathcal{W}}_4$ that has two components x_i, x_j equal, the Hessian of U is symmetric in x_i and x_j. This is implied by the conditions

$$\partial_{1,1}U(x_1,x_1,x_3,x_4) = \partial_{2,2}U(x_1,x_1,x_3,x_4), \tag{III.64}$$

$$\partial_{1,2}U(x_1,x_1,x_3,x_4) = \partial_{2,1}U(x_1,x_1,x_3,x_4), \tag{III.65}$$

$$\partial_{1,3}U(x_1,x_1,x_3,x_4) = \partial_{2,3}U(x_1,x_1,x_3,x_4), \tag{III.66}$$

$$\partial_{1,4}U(x_1,x_1,x_3,x_4) = \partial_{2,4}U(x_1,x_1,x_3,x_4), \tag{III.67}$$

$$\partial_{2,2}U(x_1,x_2,x_2,x_4) = \partial_{3,3}U(x_1,x_2,x_2,x_4), \tag{III.68}$$

$$\partial_{2,3}U(x_1,x_2,x_2,x_4) = \partial_{3,2}U(x_1,x_2,x_2,x_4), \tag{III.69}$$

$$\partial_{2,1}U(x_1,x_2,x_2,x_4) = \partial_{3,1}U(x_1,x_2,x_2,x_4), \tag{III.70}$$

$$\partial_{2,4}U(x_1,x_2,x_2,x_4) = \partial_{3,4}U(x_1,x_2,x_2,x_4), \tag{III.71}$$

$$\partial_{3,3}U(x_1,x_2,x_3,x_3) = \partial_{4,4}U(x_1,x_2,x_3,x_3), \tag{III.72}$$

$$\partial_{3,4}U(x_1,x_2,x_3,x_3) = \partial_{4,3}U(x_1,x_2,x_3,x_3), \tag{III.73}$$

$$\partial_{3,1}U(x_1,x_2,x_3,x_3) = \partial_{4,1}U(x_1,x_2,x_3,x_3), \tag{III.74}$$

$$\partial_{3,2}U(x_1,x_2,x_3,x_3) = \partial_{4,2}U(x_1,x_2,x_3,x_3) \tag{III.75}$$

for $x \in \bar{\mathcal{W}}_4$. Thanks to the smoothness of U on $\bar{\mathcal{W}}_4$, in fact, we only need these equalities for $x \in \mathcal{W}_4$.

Note that for $x \in \mathcal{W}_4$, around each of the points

$$(x_1,x_1,x_3,x_4),\ (x_1,x_2,x_2,x_4),\ \text{and } (x_1,x_2,x_3,x_3)$$

there exists a neighborhood such that the expression defining U is analytical on this neighborhood. Thus, we can apply Schwarz Theorem to obtain (III.65), (III.69) and (III.73). The remaining conditions (III.66), (III.67), (III.70), (III.71), (III.74), and (III.75) on cross derivatives are consequences of differentiation of (III.59)-(III.61). To show (III.64), we differentiate

(III.59) in x_1 then subtract (III.65) to obtain

$$\partial_{1,1}U(x_1,x_1,x_3,x_4) = \partial_{2,2}U(x_1,x_1,x_3,x_4).$$

Repeating the same procedure with (III.60), x_2 and (III.69) then with (III.61), x_3 and (III.73) we obtain (III.68) and (III.72) which concludes the proof. □

3.6.1 Proof of Theorem 3.3.1

The expansion of u, in (III.10), can be found by taking the second order Taylor expansion of G defined in (III.63)[3]. Note that as discussed in the proof of Proposition 3.6.1, the first two derivatives of u and G are equal at 0 and hence the lack of smoothness of the $\operatorname{arctanh}$ does not contribute to the second order derivative at the origin.

We now show that U defined in (III.57) solves (III.5) on $\mathcal{W}_4$ which implies by continuity of the derivatives that u solves the same PDE on $\mathbb{R}^4$. By direct computation[4] we have that for all $x \in \mathcal{W}_4$ we have

$$0 = U(x) - \Phi(x) - \frac{1}{2}\begin{pmatrix}0\\1\\0\\1\end{pmatrix}^T \partial^2 U(x)\begin{pmatrix}0\\1\\0\\1\end{pmatrix},$$

$$0 = U(x) - \Phi(x) - \frac{1}{2}\begin{pmatrix}0\\1\\1\\0\end{pmatrix}^T \partial^2 U(x)\begin{pmatrix}0\\1\\1\\0\end{pmatrix}.$$

[3] The code of the computation is available in [Bayraktar, Ekren, and Y. Zhang n.d.].

[4] The code of the computation is available in [Bayraktar, Ekren, and Y. Zhang n.d.]

The function U also satisfies the equality (III.8). Using its smoothness, we obtain

$$1 = \frac{U(x + \lambda(e_1 + e_2 + e_3 + e_4)) - U(x)}{\lambda} \to \sum_{i=1}^{4} \partial_i U(x) \text{ as } \lambda \to 0. \tag{III.76}$$

Note that $1 = \sum_{i=1}^{4} \partial_i U(x)$ implies

$$\partial^2 U(x)(e_1 + e_2 + e_3 + e_4) = 0 \text{ for all } x \in \mathbb{R}^4.$$

Therefore, for all $J \in P(N)$, we have that

$$e_J^\top \partial^2 U(x) e_J - e_{J^c}^\top \partial^2 U(x) e_{J^c} = (e_J^\top - e_{J^c}^\top) \partial^2 U(x)(e_J + e_{J^c}) = 0.$$

Thus, if J is a maximizer of the Hamiltonian $\sup_{J \in P(N)} e_J^\top \partial^2 u(x) e_J$ then its complement J^c is also a maximizer of the same Hamiltonian. This means that in order to show that the comb strategy (and also the strategy that chooses the second and the third leading expert) is optimal it is sufficient to show that the functions $U_1, ..., U_6$ defined by

$$U_1(x) := U(x) - \Phi(x) - \frac{1}{2} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}^T \partial^2 U(x) \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

$$U_2(x) := U(x) - \Phi(x) - \frac{1}{2} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}^T \partial^2 U(x) \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix},$$

$$U_3(x) := U(x) - \Phi(x) - \frac{1}{2} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}^T \partial^2 U(x) \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix},$$

$$U_4(x) := U(x) - \Phi(x) - \frac{1}{2} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}^T \partial^2 U(x) \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix},$$

$$U_5(x) := U(x) - \Phi(x) - \frac{1}{2} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}^T \partial^2 U(x) \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix},$$

$$U_6(x) := U(x) - \Phi(x) - \frac{1}{2} \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}^T \partial^2 U(x) \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix},$$

are non-negative. We study each term separately. For the first term we have

$$U_1(x) = U(x) - \Phi(x) - \frac{1}{2}\begin{pmatrix}0\\0\\0\\0\end{pmatrix}^T \partial^2 U(x)\begin{pmatrix}0\\0\\0\\0\end{pmatrix} = v(x^{(2)} - x^{(1)}, x^{(3)} - x^{(2)}, x^{(4)} - x^{(3)}) \geq 0$$

due to the definition of v. Additionally we have the following identities for $x \in \mathcal{W}_4$ that can be computed via Mathematica[5].

$$\begin{aligned}
U_4\left(\frac{x}{\sqrt{2}}\right) =& \frac{e^{x_4 - x_2}(e^{2x_1} - e^{2x_3})(e^{2x_3} - e^{2x_2})}{2\sqrt{2}(e^{2(x_1+x_2)} - e^{2(x_3+x_4)})} \geq 0,\\
\frac{\sqrt{2}(U_3(\sqrt{2}x) - U_2(\sqrt{2}x))}{\sinh(2(x_3 - x_4))} =& 1 - \operatorname{arctanh}(e^{x_1+x_2-x_3-x_4})\cosh(x_1 + x_2 - x_3 - x_4)\\
&+ \arctan(e^{x_1+x_2-x_3-x_4})\sinh(x_1 + x_2 - x_3 - x_4) \qquad \text{(III.77)}\\
\frac{\sqrt{2}(U_5(\sqrt{2}x) - U_3(\sqrt{2}x))}{\sinh(2(x_2 - x_3))} =& -\operatorname{arctanh}(e^{x_1+x_2-x_3-x_4})\cosh(x_1 - x_2 - x_3 + x_4)\\
&+ \arctan(e^{x_1+x_2-x_3-x_4})\sinh(x_1 - x_2 - x_3 + x_4) \qquad \text{(III.78)}\\
\frac{\sqrt{2}(U_6(\sqrt{2}x) - U_3(\sqrt{2}x))}{\sinh(2(x_3 - x_1))} =& \operatorname{arctanh}(e^{x_1+x_2-x_3-x_4})\cosh(x_1 - x_2 + x_3 - x_4)\\
&+ \arctan(e^{x_1+x_2-x_3-x_4})\sinh(x_1 - x_2 + x_3 - x_4). \qquad \text{(III.79)}
\end{aligned}$$

Due to $x \in \mathcal{W}_4$, $U_4(x) \geq 0$. Additionally, the function

$$x \geq 0 \mapsto 1 - \operatorname{arctanh}(e^{-x})\cosh(-x) + \arctan(e^{-x})\sinh(-x)$$

is non-positive. Thus

$$U_3 \geq U_2.$$

Finding the sign of the right hand side of (III.78) and (III.79) is equivalent to finding the signs

[5] The code of the computation is available in [Bayraktar, Ekren, and Y. Zhang n.d.]

of

$$-\operatorname{arctanh}(e^{-x})\cosh(-x+y)+\arctan(e^{-x})\sinh(-x+y),\ \text{for } x,y\geq 0$$

and

$$\operatorname{arctanh}(e^{-x})\cosh(-x+y)+\arctan(e^{-x})\sinh(-x+y),\ \text{for } x,y\geq 0.$$

These functions are respectively non-positive and non-negative due to the fact that

$$\operatorname{arctanh}(e^{-x})\geq\arctan(e^{-x})\geq 0 \text{ and} \cosh(x)\geq|\sinh(x)|.$$

Thus

$$U_5\geq U_3 \text{ and } U_6\geq U_3.$$

Finally, to finish the proof of the main theorem, it is sufficient to show that

$$U_2\geq 0. \tag{III.80}$$

To show this inequality, it is more convenient to write U_2 as in terms of v. Thanks to (III.57),

$$U_2(x)=v(x_2-x_1,x_3-x_2,x_4-x_3)-\frac{1}{2}\begin{pmatrix}0\\0\\1\end{pmatrix}^T\partial^2 v(x_2-x_1,x_3-x_2,x_4-x_3)\begin{pmatrix}0\\0\\1\end{pmatrix},$$

and to show (III.80), it is sufficient to show that for all $y_1,y_2,y_3\geq 0$,

$$v_2(y_1,y_2,y_3):=v(y_1,y_2,y_3)-\frac{1}{2}\begin{pmatrix}0\\0\\1\end{pmatrix}^T\partial^2 v(y_1,y_2,y_3)\begin{pmatrix}0\\0\\1\end{pmatrix}\geq 0.$$

Thanks to the smoothness of v on $(0,\infty)^3$ and the fact that the data of (III.21) is constant, we can differentiate (III.21) to obtain that v_2 also solves (III.21). Thanks to the maximum principle for this PDE, in order to show (III.21), it is sufficient to show that $v_2 \geq 0$ for $y_1 = 0$ or $y_2 = 0$ or $y_3 = 0$. Our objective is to use the Proposition 3.5.4. Similarly to (III.38)-(III.41) define

$$\begin{aligned}
\tilde{f}(x,y) &= v_2\left(0, \frac{x}{\sqrt{2}}, \frac{y}{\sqrt{2}}\right), \\
\tilde{r}_1(x,y) &= v_2\left(\frac{x}{\sqrt{2}}, 0, \frac{y+x}{\sqrt{2}}\right), \\
\tilde{h}(x,y) &= v_2\left(\frac{y}{\sqrt{2}}, \frac{x}{\sqrt{2}}, 0\right), \\
\tilde{r}_2(x,y) &= v_2\left(\frac{x+y}{\sqrt{2}}, 0, \frac{x}{\sqrt{2}}\right).
\end{aligned}$$

By direct computation via Mathematica[6], these functions satisfy,

$$\begin{aligned}
(\partial_x - 2\partial_y)\tilde{f}(x,y) &= \frac{2}{\tanh x}\tilde{f}(x,y) - \frac{2}{\sinh x}\tilde{r}_1(x,y), \\
\partial_x \tilde{r}_1(x,y) &= -\frac{2}{\sinh x}\tilde{f}(x,y) + \frac{2}{\tanh x}\tilde{r}_1(x,y), \\
(\partial_x - 2\partial_y)\tilde{h}(x,y) &= \frac{2}{\tanh x}\tilde{h}(x,y) - \frac{2}{\sinh x}\tilde{r}_2(x,y) \\
&\quad + \frac{1}{\sqrt{2}}\left(1 - \operatorname{arctanh}(e^{-x-y/2})\cosh(x+y/2) - \arctan(e^{-x-y/2})\sinh(x+y/2))\right), \\
\partial_x \tilde{r}_2(x,y) &= -\frac{2}{\sinh x}\tilde{h}(x,y) + \frac{2}{\tanh x}\tilde{r}_2(x,y).
\end{aligned}$$

Since the function

$$x \geq 0 \mapsto 1 - \operatorname{arctanh}(e^{-x})\cosh(x) - \arctan(e^{-x})\sinh(x)$$

[6] The code of the computation is available in [Bayraktar, Ekren, and Y. Zhang n.d.]

is non-positive, we have that

$$\begin{aligned}(\partial_x - 2\partial_y)\tilde{f}(x,y) &= \frac{2}{\tanh x}\tilde{f}(x,y) - \frac{2}{\sinh x}\tilde{r}_1(x,y),\\ \partial_x \tilde{r}_1(x,y) &= -\frac{2}{\sinh x}\tilde{f}(x,y) + \frac{2}{\tanh x}\tilde{r}_1(x,y),\\ (\partial_x - 2\partial_y)\tilde{h}(x,y) &\le \frac{2}{\tanh x}\tilde{h}(x,y) - \frac{2}{\sinh x}\tilde{r}_2(x,y),\\ \partial_x \tilde{r}_2(x,y) &= -\frac{2}{\sinh x}\tilde{h}(x,y) + \frac{2}{\tanh x}\tilde{r}_2(x,y).\end{aligned}$$

Thus, to finish the proof of the main result by application of Proposition 3.5.4, we need to control $\tilde{f}$ and $\tilde{h}$ on the boundary of their domain of definition and obtain the limit of $\tilde{r}_1$ and $\tilde{r}_2$ at infinity. Note that $\tilde{r}_1$ and $\tilde{r}_2$ converge to 0 at infinity. By a direct computation[7], we have that

$$\begin{aligned}\tilde{f}(x,y) =& F_f(x,y) + \left(\arctan(e^{-x-\frac{y}{2}}) - e^{-x-\frac{y}{2}} + \frac{e^{-3x-\frac{3y}{2}}}{3}\right) H_f(x,y)\\ &+ \left(\operatorname{arctanh}(e^{-x-\frac{y}{2}}) - e^{-x-\frac{y}{2}} - \frac{e^{-3x-\frac{3y}{2}}}{3}\right) G_f(x,y),\\ \tilde{h}(x,y) =& F_h(x,y) + \left(\arctan(e^{-x-\frac{y}{2}}) - e^{-x-\frac{y}{2}} + \frac{e^{-3x-\frac{3y}{2}}}{3}\right) H_h(x,y)\\ &+ \left(\operatorname{arctanh}(e^{-x-\frac{y}{2}}) - e^{-x-\frac{y}{2}} - \frac{e^{-3x-\frac{3y}{2}}}{3}\right) G_h(x,y),\end{aligned}$$

[7] The code of the computation and the expressions for the functions are available in [Bayraktar, Ekren, and Y. Zhang n.d.].

where as $x^2 + y^2 \to \infty$,

$$\begin{aligned}
F_f(x,y) &:= \frac{e^{-2(2x+y)}\big(4\cosh(y) + 6\operatorname{csch}(2x+y)\sinh^2(x) - \sinh(y)\big)}{24\sqrt{2}} = o(1),\\
H_f(x,y) &:= \frac{3\cosh(x-\frac{y}{2}) + 6\cosh(x+\frac{y}{2}) - 5\cosh(x+\frac{3y}{2})}{16\sqrt{2}} = o(e^{5x+\frac{5y}{2}}),\\
G_f(x,y) &:= \frac{3\sinh(x-\frac{y}{2}) - 6\sinh(x+\frac{y}{2}) - 5\sinh(x+\frac{3y}{2})}{16\sqrt{2}} = o(e^{5x+\frac{5y}{2}}),\\
F_h(x,y) &:= \frac{e^{-2(2x+y)}\left(4 + 3\coth(2x+y) - 3\cosh(y)\operatorname{csch}(2x+y)\right)}{24\sqrt{2}} = o(1),\\
H_h(x,y) &:= \frac{(-1+3\cosh(y))\cosh(x+\frac{y}{2})}{8\sqrt{2}} = o(e^{5x+\frac{5y}{2}}),\\
G_h(x,y) &:= \frac{(1+3\cosh(y))\sinh(x+\frac{y}{2})}{8\sqrt{2}} = o(e^{5x+\frac{5y}{2}}).
\end{aligned}$$

Given also the expansions at 0

$$\arctan(x) = x - \frac{x^3}{3} + O(x^5), \quad \operatorname{arctanh}(x) = x + \frac{x^3}{3} + O(x^5),$$

we have that

$$\lim_{x^2+y^2\to\infty} \tilde{f}(x,y) = \lim_{x^2+y^2\to\infty} \tilde{h}(x,y) = 0.$$

Additionally,

$$\begin{aligned}
&\tilde{f}(x,0) = \tilde{h}(x,0) = v_2(0, x/\sqrt{2}, 0)\\
&= \frac{1}{8\sqrt{2}}\big(2\arctan(e^{-x})\cosh(x) - 4\operatorname{arctanh}(e^{-x})\sinh(x) + \tanh(x)\big),\\
&\tilde{f}(0,y) = \frac{5}{8\sqrt{2}}e^{y}(-1+\coth(y))\sinh^2(y) + \frac{1}{16\sqrt{2}}e^{y}(-1+\coth(y))\sinh(y)\\
&\Big(\arctan(e^{-y/2})(9\cosh(y/2) - 5\cosh(3y/2) - \operatorname{arctanh}(e^{-y/2})(9\sinh(y/2) + 5\sinh(3y/2))\Big),\\
&\tilde{h}(0,y) = \frac{1}{16\sqrt{2}}e^{y}(-1+\coth(y))\sinh(y)\\
&\Big(\arctan(e^{-y/2})(\cosh(y/2) + 3\cosh(3y/2)) + \operatorname{arctanh}(e^{-y/2})(\sinh(y/2) - 3\sinh(3y/2))\Big).
\end{aligned}$$

These functions are all non-negative. Direct application of Proposition 3.5.4 then yields

$$\tilde{f}, \tilde{h}, \tilde{r}_1, \tilde{r}_2 \geq 0 \text{ on } [0, \infty)^2.$$

Thus, for all $x \in \mathcal{W}_4$ we have

$$U(x) - \frac{1}{2} \sup_{J \in P(N)} e_J^\top \partial^2 U(x) e_J = \Phi(x).$$

Thanks to the smoothness and symmetry of u, we obtain (III.5).

3.6.2 Proof of Theorem 3.3.2

We first prove the asymptotics for $\underline{u}^\delta$. This function satisfies the dynamic programming principle

$$\begin{aligned}\underline{u}^\delta(x) &= \delta\Phi(x) \\ &\quad + \frac{1-\delta}{2} \inf_{\alpha \in \mathcal{U}} \Big(\underline{u}^\delta(x + \sqrt{\delta} e_{\mathcal{J}_\mathcal{C}(x)}) - \alpha(\mathcal{J}_\mathcal{C}(x)) + \underline{u}^\delta(x + \sqrt{\delta} e_{\mathcal{J}_\mathcal{C}^c(x)}) - \alpha(\mathcal{J}_\mathcal{C}^c(x)) \Big) \\ &= \delta\Phi(x) + \frac{1-\delta}{2} \Big(\underline{u}^\delta(x + \sqrt{\delta} e_{\mathcal{J}_\mathcal{C}(x)}) + \underline{u}^\delta(x + \sqrt{\delta} e_{\mathcal{J}_\mathcal{C}^c(x)}) - 1 \Big)\end{aligned}$$

This is equivalent to

$$\underline{u}^\delta(x) = \Phi(x) + \frac{1-\delta}{2\delta} \Big(\underline{u}^\delta(x + \sqrt{\delta} e_{\mathcal{J}_\mathcal{C}(x)}) + \underline{u}^\delta(x + \sqrt{\delta} e_{\mathcal{J}_\mathcal{C}^c(x)}) - 1 - 2\underline{u}^\delta(x) \Big).$$

Similarly to u and V^δ,

$$\underline{u}^\delta(x + \sqrt{\delta}\lambda \sum_{i=1}^{4} e_i) = \underline{u}^\delta(x) + \lambda \text{ for all } \lambda \in \mathbb{R}.$$

Thus,

$$\underline{u}^{\delta}(x+\sqrt{\delta}e_{\mathcal{J}_{\mathcal{C}}^{c}(x)})-1=\underline{u}^{\delta}(x+\sqrt{\delta}e_{\mathcal{J}_{\mathcal{C}}^{c}(x)}-\sqrt{\delta}\sum_{i=1}^{4}e_i)=\underline{u}^{\delta}(x-\sqrt{\delta}e_{\mathcal{J}_{\mathcal{C}}(x)})$$

and the DPP becomes

$$\underline{u}^{\delta}(x)=\Phi(x)+\frac{1-\delta}{2\delta}\left(\underline{u}^{\delta}(x+\sqrt{\delta}e_{\mathcal{J}_{\mathcal{C}}(x)})+\underline{u}^{\delta}(x-\sqrt{\delta}e_{\mathcal{J}_{\mathcal{C}}(x)})-2\underline{u}^{\delta}(x)\right).$$

Due the fact that $\mathcal{J}_{\mathcal{C}}^{b}$ is balanced, $\underline{u}^{\delta}$ in fact does not depend on $\alpha \in \mathcal{U}$. Thus, by choosing a particular control we can prove similarly to the proof of [Drenska 2017, Theorem 7] that $\underline{u}^{\delta}$ converges to the unique viscosity solution of the equation

$$f(x)-\frac{1}{2}e_{\mathcal{J}_{\mathcal{C}}(x)}^{\top}\partial^{2}f(x)e_{\mathcal{J}_{\mathcal{C}}(x)}=\Phi(x)$$

with linear growth. Note that thanks to (III.62), u also solves this PDE and has linear growth. Thus, comb strategies are asymptotically optimal and $\underline{u}^{\delta}(x)\to u(x)$ as $\delta \downarrow 0$.

3.7 Concluding Remarks

Using a system of first order hyperbolic PDE, (III.42)-(III.43), we characterize and compute the expectation (III.20) of the third component of the local time of an obliquely reflected Brownian motion in the first octant. Then, using a maximum principle in Proposition 3.5.4, we show that this value provides a solution to the Hamilton-Jacobi-Bellman equation (III.5) that characterizes the long time behavior of a regret minimization problem with 4 experts. Finally, we prove that, as conjectured in [Gravin, Peres, and Sivan 2016], comb strategies are asymptotically optimal for nature.

We conjectured that this methodology can be performed for $N \geq 5$ experts, and the comb strategy would be optimal. However, numerical experiments in [Chase 2019] suggests that when $N = 5$, the comb strategy is not asymptotically for nature. Instead, the strategy of picking the first and third experts together with probability 1/2 and picking the second, fourth, and fifth experts together with probability 1/2 performs strictly better in the numerical experiments.

We also mention that one can use our methodology to study the parabolic version of (III.5) which corresponds to the long-time behavior of the game with deterministic stopping. In this case, we expect the system (III.42)-(III.43) to have an additional time dependence. The optimality of the comb strategy is proved for the parabolic version in [Bayraktar, Ekren, and X. Zhang 2020].

CHAPTER IV

Conclusion

Online learning and sequential decision-making under uncertainty apply to a wide variety of problems. In this thesis, we investigated two such problems with different models, and used different analysis techniques for each. In the first problem, we considered stochastic uncertainties and used many stochastic analysis techniques to bear, and in the second we studied adversarial uncertainties and used PDE methods.

In Chapter II, we considered the problem of learning an optimal dispatch policy for an $M/M/1$ queue with unknown arrival and service rates. We focused on designing an online learning algorithm that operates in batches where each batch has an explicit exploration phase that can be skipped using an adaptive criterion, and an exploitation phase with extending duration. We analyzed and compared the asymptotic performance of our algorithm with the optimal algorithm that has knowledge of the model statistics by coupling the queue-length process under the proposed algorithm with the queue-length process of a system using the optimal dispatch policy. Our algorithm achieves $O(1)$ regret when optimal algorithms use non-zero admission thresholds and $O(\ln^{1+\epsilon}(N))$ for $\epsilon > 0$ regret when 0 is one of the optimal admission thresholds where N is the number of arrival customers. We also discussed the finite time-horizon performance of our algorithm through numerical experiments by choosing different values for the hyper-parameters of the algorithm.

In Chapter III, we studied the problem of finding the optimal policy for nature to pick experts in an adversarial manner with 4 experts and at all times before a geometrically distributed stopping time. We took a game theoretical point of view, adopted PDE results from [Drenska 2017], and proved the optimality of the "comb strategy" in which nature randomizes uniformly between two options: picking the best and third-place experts together , and picking the second and fourth place experts together, where the experts are ranked based on their accumulated reward. We also showed that the regret grows as $O(\sqrt{1/\delta})$ where δ is the parameter of the geometric stopping time when nature and the player both perform optimally.

There are a few future directions to pursue based on the problems we studied in this thesis. Some of them are briefly discussed below.

4.1 Non-parametric learning in queuing systems

In Chapter II, we assumed Poisson arrivals and exponentially distributed service requirements for the customers. The memoryless property of the exponential distribution and many properties of the Poisson process are crucial to our analysis. However, as arrival processes and service requirements are not this specific in the real-world applications of our model, generalizations are merited.One possible future direction is to adapt our learning algorithm to general arrival processes and service-time distributions or to design a non-parametric learning algorithm that does not make assumptions about the arrival processes or the service-time distributions. The analysis of the performance of any learning algorithm would be more involved in this case since the analysis in [Naor 1969] no longer holds. This problem has received attention—see [Oz 2022]—under a different information structure where only the queue-length is observed by arrivals. Under this setting, the analytical optimal strategy for this problem is still unknown and may be time-varying; see [Oz 2022] for details.

4.2 Learning-based allocation in multiple server queuing systems

Our analysis focused on a single-queue single-server setting with statistically homogeneous customer service requirements. A generalization would be studying either the dispatch or scheduling problem for a system with multiple servers and multiple queues, and where the customers are heterogeneous with statistically different service requirements.

4.3 Asymptotic optimal strategy of the player in the 4 expert setting

In Chapter III, we focused on studying the asymptotically optimal strategy of the adversarial nature. In Remark $3.3.6$, we conjectured that α^* is an asymptotically optimal strategy for the player in the 4 experts setting. However, proving this conjecture is still an open problem.

4.4 Asymptotic optimal strategy of nature in the 5 or more expert settings

Numerical experiments in [Chase 2019] have shown evidence that the strategy of choosing the best and third place experts together with probability 1/2 and choosing the second, fourth, and fifth place experts together with probability 1/2 where the experts are ranked based on their accumulated rewards outperforms the “comb strategy” for the adversarial nature in the 5 expert setting. However, finding the asymptotic optimal strategy of the adversary nature in cases with 5 or more experts is still an open problem.

Bibliography

Adler, Saghar, Mehrdad Moharrami, and Vijay Subramanian (2022). “Learning a Discrete Set of Optimal Allocation Rules in Queueing Systems with Unknown Service Rates.” In: DOI: 10.48550/ARXIV.2202.02419. URL: https://arxiv.org/abs/2202.02419.

Agrawal, Shipra and Randy Jia (2022). “Learning in structured MDPs with convex cost functions: improved regret bounds for inventory management.” In: *Oper. Res.* 70.3, pp. 1646–1664. ISSN: 0030-364X. DOI: 10.1287/opre.2022.2263. URL: https://doi.org/10.1287/opre.2022.2263.

Atar, Rami, Eyal Castiel, and Yonatan Shadmi (2022). “Scheduling in the high uncertainty heavy traffic regime.” In: DOI: 10.48550/ARXIV.2204.05733. URL: https://arxiv.org/abs/2204.05733.

Balsubramani, Akshay (2015). *Sharp Finite-Time Iterated-Logarithm Martingale Concentration*. arXiv: 1405.2639 [math.PR].

Barles, G. and P. E. Souganidis (1991). “Convergence of approximation schemes for fully nonlinear second order equations.” In: *Asymptotic Anal.* 4.3, pp. 271–283. ISSN: 0921-7134.

Bayraktar, Erhan, Andrea Cosso, and Huyên Pham (2016). “Robust feedback switching control: Dynamic programming and viscosity solutions.” In: *SIAM J. Control Optim.*

54.5, pp. 2594–2628. ISSN: 0363-0129. DOI: 10.1137/15M1046903. URL: https://doi.org/10.1137/15M1046903.

Bayraktar, Erhan, Ibrahim Ekren, and Xin Zhang (2020). “Finite-time 4-expert prediction problem.” In: *Comm. Partial Differential Equations* 45.7, pp. 714–757. ISSN: 0360-5302. DOI: 10.1080/03605302.2020.1712418. URL: https://doi.org/10.1080/03605302.2020.1712418.

— (2021). “Prediction against a limited adversary.” In: *J. Mach. Learn. Res.* 22, Paper No. 72, 33. ISSN: 1532-4435.

— (2022). *A PDE approach for regret bounds under partial monitoring*. arXiv: 2209.01256 [math.PR].

Bayraktar, Erhan, Ibrahim Ekren, and Yili Zhang (2020). “On the asymptotic optimality of the comb strategy for prediction with expert advice.” In: *Ann. Appl. Probab.* 30.6, pp. 2517–2546. ISSN: 1050-5164. DOI: 10.1214/20-AAP1565. URL: https://doi.org/10.1214/20-AAP1565.

— (n.d.). “Mathematica Appendix.” In: (). https://sites.google.com/site/ibrahimekren/.

Bayraktar, Erhan, H. Vincent Poor, and Xin Zhang (2021). “Malicious Experts Versus the Multiplicative Weights Algorithm in Online Prediction.” In: *IEEE Transactions on Information Theory* 67.1, pp. 559–565. DOI: 10.1109/TIT.2020.3025866.

Bayraktar, Erhan and Mihai Sîrbu (2013). “Stochastic Perron’s method for Hamilton-Jacobi-Bellman equations.” In: *SIAM J. Control Optim.* 51.6, pp. 4274–4294. ISSN: 0363-0129. DOI: 10.1137/12090352X. URL: https://doi.org/10.1137/12090352X.

Benosman, M. (2016). *Learning-Based Adaptive Control: An Extremum Seeking Approach – Theory and Applications*. Elsevier Science. ISBN: 9780128031513. URL: https://books.google.com/books?id=SXcZBgAAQBAJ.

Bertsekas, Dimitri (2019). *Reinforcement learning and optimal control*. Athena Scientific.

Buyukkoc, C., P. Varaiya, and J. Walrand (1985). “The $c\mu$ rule revisited.” In: *Adv. in Appl. Probab.* 17.1, pp. 237–238. ISSN: 0001-8678. DOI: 10.2307/1427064. URL: https://doi.org/10.2307/1427064.

Cesa-Bianchi, Nicolò and Gábor Lugosi (2006). *Prediction, learning, and games*. Cambridge University Press, Cambridge, pp. xii+394. ISBN: 9780521841085;0521841089. DOI: 10.1017/CBO9780511546921. URL: https://doi.org/10.1017/CBO9780511546921.

Chase, Zachary (2019). *Experimental Evidence for Asymptotic Non-Optimality of Comb Adversary Strategy*. arXiv: 1912.01548 [cs.GT].

Chen, Xinyun, Yunan Liu, and Guiyu Hong (2022). *An online learning approach to dynamic pricing and capacity sizing in service systems*. arXiv: 2009.02911 [math.PR].

Chen, Ying and John J. Hasenbein (Oct. 2020). “Knowledge, congestion, and economics: Parameter uncertainty in Naor’s model.” English. In: *Queueing Systems* 96.1-2. Copyright - © Springer Science+Business Media, LLC, part of Springer Nature 2020; Last updated - 2020-12-22, pp. 83–99. URL: https://proxy.lib.umich.edu/login?url=https://www.proquest.com/scholarly-journals/knowledge-congestion-economics-parameter/docview/2471741635/se-2.

Choudhury, Tuhinangshu et al. (2021). “Job Dispatching Policies for Queueing Systems with Unknown Service Rates.” In: *Proceedings of the Twenty-Second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*. MobiHoc ’21. Shanghai, China: Association for Computing Machinery, pp. 181–190. ISBN: 9781450385589. DOI: 10.1145/3466772.3467047. URL: https://doi.org/10.1145/3466772.3467047.

Cohen, Asaf (2019a). “Asymptotic analysis of a multiclass queueing control problem under heavy traffic with model uncertainty.” In: *Stoch. Syst.* 9.4, pp. 359–391. DOI: 10.1287/stsy.2019.0034. URL: https://doi.org/10.1287/stsy.2019.0034.

Cohen, Asaf (2019b). “Brownian control problems for a multiclass M/M/1 queueing problem with model uncertainty.” In: *Math. Oper. Res.* 44.2, pp. 739–766. ISSN: 0364-765X. DOI: 10.1287/moor.2018.0944. URL: https://doi.org/10.1287/moor.2018.0944.

Cohen, Asaf and Subhamay Saha (2021). “Asymptotic optimality of the generalized $c\mu$ rule under model uncertainty.” In: *Stochastic Process. Appl.* 136, pp. 206–236. ISSN: 0304-4149. DOI: 10.1016/j.spa.2021.03.004. URL: https://doi.org/10.1016/j.spa.2021.03.004.

Cohen, Asaf, Vijay G. Subramanian, and Yili Zhang (2022). “Learning-based Optimal Admission Control in a Single Server Queuing System.” In: arXiv: 2212.11316 [math.OC].

Cover, Thomas M. (1967). “Behavior of sequential predictors of binary sequences.” In: *Trans. Fourth Prague Conf. on Information Theory, Statistical Decision Functions, Random Processes (Prague, 1965)*. Academia, Prague, pp. 263–272.

Cox, David R. and Walter L. Smith (1961). *Queues*. Methuen’s Monographs on Statistical Subjects. Methuen & Co., Ltd., London; John Wiley & Sons, Inc., New York, pp. xii+180.

Crandall, Michael G., Hitoshi Ishii, and Pierre-Louis Lions (1992). “User’s guide to viscosity solutions of second order partial differential equations.” In: *Bull. Amer. Math. Soc. (N.S.)* 27.1, pp. 1–67. ISSN: 0273-0979. DOI: 10.1090/S0273-0979-1992-00266-5. URL: https://doi.org/10.1090/S0273-0979-1992-00266-5.

Csörgő, Miklós (1968). “On the strong law of large numbers and the central limit theorem for martingales.” In: *Trans. Amer. Math. Soc.* 131, pp. 259–275. ISSN: 0002-9947. DOI: 10.2307/1994694. URL: https://doi.org/10.2307/1994694.

Drenska, Nadejda (2017). *A PDE Approach to a Prediction Problem Involving Randomized Strategies*. Thesis (Ph.D.)–New York University. ProQuest LLC, Ann Arbor, MI, p. 77. ISBN: 978-0355-40708-2. URL: http://gateway.proquest.com/openurl?

url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqm&rft_dat=xri:pqdiss:10600500.

Durrett, Richard (2016). *Essentials of stochastic processes*. Springer Texts in Statistics. Third edition [of MR2933766]. Springer, Cham, pp. ix+275. ISBN: 978-3-319-45613-3; 978-3-319-45614-0. DOI: 10.1007/978-3-319-45614-0. URL: https://doi.org/10.1007/978-3-319-45614-0.

Fernholz, E Robert et al. (2011). "Planar Diffusions with Rank-Based Characteristics: Transition Probabilities, Time Reversal, Maximality and Perturbed Tanaka equations." In: *arXiv preprint arXiv:1108.3992*.

Fleming, W. H. and P. E. Souganidis (1989). "On the existence of value functions of two-player, zero-sum stochastic differential games." In: *Indiana Univ. Math. J.* 38.2, pp. 293–314. ISSN: 0022-2518. DOI: 10.1512/iumj.1989.38.38015. URL: https://doi.org/10.1512/iumj.1989.38.38015.

Fusco, D and N Manganaro (1996). "A method for finding exact solutions to hyperbolic systems of first-order PDEs." In: *IMA journal of applied mathematics* 57.3, pp. 223–242.

Gravin, Nick, Yuval Peres, and Balasubramanian Sivan (2016). "Towards optimal algorithms for prediction with expert advice." In: *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*. ACM, New York, pp. 528–547. DOI: 10.1137/1.9781611974331.ch39. URL: https://doi.org/10.1137/1.9781611974331.ch39.

Haussler, David, Jyrki Kivinen, and Manfred K. Warmuth (1995). "Tight worst-case loss bounds for predicting with expert advice." In: *Computational learning theory (Barcelona, 1995)*. Vol. 904. Lecture Notes in Comput. Sci. Springer, Berlin, pp. 69–83. DOI: 10.1007/3-540-59119-2_169. URL: https://doi.org/10.1007/3-540-59119-2_169.

Hazan, Elad (2021). “Introduction to Online Convex Optimization.” In: arXiv: 1909.05207 [cs.LG].

Hoeven, Dirk van der, Nikita Zhivotovskiy, and Nicolò Cesa-Bianchi (2022). “A Regret-Variance Trade-Off in Online Learning.” In: arXiv: 2206.02656 [cs.LG].

Hoi, Steven C.H. et al. (2021). “Online learning: A comprehensive survey.” In: *Neurocomputing* 459, pp. 249–289. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2021.04.112. URL: https://www.sciencedirect.com/science/article/pii/S0925231221006706.

Ichiba, Tomoyuki, Ioannis Karatzas, and Mykhaylo Shkolnikov (2013). “Strong solutions of stochastic equations with rank-based coefficients.” In: *Probab. Theory Related Fields* 156.1-2, pp. 229–248. ISSN: 0178-8051. DOI: 10.1007/s00440-012-0426-3. URL: https://doi.org/10.1007/s00440-012-0426-3.

Ioannou, P.A. and J. Sun (2012). *Robust Adaptive Control*. Dover Books on Electrical Engineering Series. Dover Publications, Incorporated. ISBN: 9780486498171. URL: https://books.google.com/books?id=pXWFY%5C_vbg1MC.

Iskenderov, Nizameddin Sh and Anar A Mamedov (n.d.). “Inverse scattering problem for a hyperbolic system of first order equations on a semi-axis on a first approximation.” In: ().

Jia, Huiwen, Cong Shi, and Siqian Shen (2022). *Online Learning and Pricing for Service Systems with Reusable Resources*. DOI: 10.1287/opre.2022.2381.

Johansen, Søren Glud and Shaler Stidham Jr. (1980). “Control of arrivals to a stochastic input-output system.” In: *Adv. in Appl. Probab.* 12.4, pp. 972–999. ISSN: 0001-8678. DOI: 10.2307/1426752. URL: https://doi.org/10.2307/1426752.

Kahe, Ghasem and Amir Hossein Jahangir (June 2019). “A Self-Tuning Controller for Queuing Delay Regulation in TCP/AQM Networks.” In: *Telecommun. Syst.* 71.2, pp. 215–229. ISSN: 1018-4864. DOI: 10.1007/s11235-018-0526-1. URL: https://doi.org/10.1007/s11235-018-0526-1.

Knudsen, Niels Chr. (1972). "Individual and social optimization in a multiserver queue with a general cost-benefit structure." In: *Econometrica* 40, pp. 515–528. ISSN: 0012-9682. DOI: 10.2307/1913182. URL: https://doi.org/10.2307/1913182.

Kohn, Robert V. and Sylvia Serfaty (2010). "A deterministic-control-based approach to fully nonlinear parabolic and elliptic equations." In: *Comm. Pure Appl. Math.* 63.10, pp. 1298–1350. ISSN: 0010-3640. DOI: 10.1002/cpa.20336. URL: https://doi.org/10.1002/cpa.20336.

Koolen, Wouter M. and Tim van Erven (2015). "Second-order Quantile Methods for Experts and Combinatorial Games." In: arXiv: 1502.08009 [cs.LG].

Krishnasamy, Subhashini, P. T. Akhil, et al. (Dec. 2018). "Augmenting Max-Weight With Explicit Learning for Wireless Scheduling With Switching Costs." In: *IEEE/ACM Transactions on Networking* 26.6, pp. 2501–2514. DOI: 10.1109/tnet.2018.2869874. URL: https://doi.org/10.1109%5C%2Ftnet.2018.286987.

Krishnasamy, Subhashini, Ari Arapostathis, et al. (2018). "On Learning the cμ Rule in Single and Parallel Server Networks." In: DOI: 10.48550/ARXIV.1802.06723. URL: https://arxiv.org/abs/1802.06723.

Krishnasamy, Subhashini, Rajat Sen, et al. (2021). "Learning unknown service rates in queues: A multiarmed bandit approach." In: *Oper. Res.* 69.1, pp. 315–330. ISSN: 0030-364X. DOI: 10.1287/opre.2020.1995. URL: https://doi.org/10.1287/opre.2020.1995.

Landau, I.D. et al. (2011). *Adaptive Control: Algorithms, Analysis and Applications*. Communications and Control Engineering. Springer London. ISBN: 9780857296641. URL: https://books.google.com/books?id=fb1GVyJHeBgC.

Lattimore, Tor and Csaba Szepesvári (2020a). *Bandit Algorithms*. Cambridge University Press. DOI: 10.1017/9781108571401.

— (2020b). *Bandit algorithms*. Cambridge University Press.

Li, Quan-Lin et al. (2019). *An Overview for Markov Decision Processes in Queues and Networks*. arXiv: 1907.10243 [math.OC].

Lippman, Steven A. and Shaler Stidham Jr. (1977). “Individual versus social optimization in exponential congestion systems.” In: *Operations Res.* 25.2, pp. 233–247. ISSN: 0030-364X. DOI: 10.1287/opre.25.2.233. URL: https://doi.org/10.1287/opre.25.2.233.

Littlestone, Nick and Manfred K. Warmuth (1994). “The weighted majority algorithm.” In: *Inform. and Comput.* 108.2, pp. 212–261. ISSN: 0890-5401. DOI: 10.1006/inco.1994.1009. URL: https://doi.org/10.1006/inco.1994.1009.

Mhammedi, Zakaria, Wouter M. Koolen, and Tim van Erven (2019). “Lipschitz Adaptivity with Multiple Learning Rates in Online Learning.” In: arXiv: 1902.10797 [cs.LG].

Naor, Pinhas (1969). “The Regulation of Queue Size by Levying Tolls.” In: *Econometrica* 37.1, pp. 15–24. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1909200 (visited on 11/20/2022).

Neely, Michael J. (2010). “Introduction.” In: *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Cham: Springer International Publishing, pp. 1–14. ISBN: 978-3-031-79995-2. DOI: 10.1007/978-3-031-79995-2_1. URL: https://doi.org/10.1007/978-3-031-79995-2_1.

Neely, Michael J., Scott T. Rager, and Thomas F. La Porta (2012). “Max-Weight Learning Algorithms for Scheduling in Unknown Environments.” In: *IEEE Transactions on Automatic Control* 57.5, pp. 1179–1191. DOI: 10.1109/TAC.2012.2191874.

Nutz, Marcel (2016). “Utility maximization under model uncertainty in discrete time.” In: *Math. Finance* 26.2, pp. 252–268. ISSN: 0960-1627. DOI: 10.1111/mafi.12068. URL: https://doi.org/10.1111/mafi.12068.

Oz, Binyamin (Apr. 2022). “Optimal admission policy to an observable M/G/1 queue.” English. In: *Queueing Systems* 100.3-4. Copyright - © The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022; Last updated - 2022-11-30, pp. 477–479. URL: https://www.proquest.com/scholarly-journals/optimal-admission-policy-observable-m-g-1-queue/docview/2672489021/se-2.

Pham, Triet and Jianfeng Zhang (2014). "Two person zero-sum game in weak formulation and path dependent Bellman-Isaacs equation." In: *SIAM J. Control Optim.* 52.4, pp. 2090–2121. ISSN: 0363-0129. DOI: 10.1137/120894907. URL: https://doi.org/10.1137/120894907.

Rakhlin, Alexander, Ohad Shamir, and Karthik Sridharan (2012). "Relax and Randomize: From Value to Algorithms." In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira et al. Curran Associates, Inc., pp. 2141–2149. URL: http://papers.nips.cc/paper/4638-relax-and-randomize-from-value-to-algorithms.pdf.

Salodkar, Nitin et al. (2008). "An on-line learning algorithm for energy efficient delay constrained scheduling over a fading channel." In: *IEEE Journal on Selected Areas in Communications* 26.4, pp. 732–742. DOI: 10.1109/JSAC.2008.080514.

Shwartz, Adam and Armand M. Makowski (1986). "An optimal adaptive scheme for two competing queues with constraints." In: *Analysis and optimization of systems (Antibes, 1986)*. Vol. 83. Lect. Notes Control Inf. Sci. Springer, Berlin, pp. 515–532. DOI: 10.1007/BFb0007586. URL: https://doi.org/10.1007/BFb0007586.

Smith, Wayne E. (1956). "Various optimizers for single-stage production." In: *Naval Res. Logist. Quart.* 3, pp. 59–66. ISSN: 0028-1441. DOI: 10.1002/nav.3800030106. URL: https://doi.org/10.1002/nav.3800030106.

Stahlbuhk, Thomas, Brooke Shrader, and Eytan Modiano (2021). "Learning algorithms for minimizing queue length regret." In: *IEEE Trans. Inform. Theory* 67.3, pp. 1759–1781.

ISSN: 0018-9448. DOI: 10.1109/TIT.2021.3054854. URL: https://doi.org/10.1109/TIT.2021.3054854.

Sutton, Richard S and Andrew G Barto (2018). *Reinforcement learning: An introduction*. MIT press.

Takagi, Hideaki and Ahmed M. K. Tarabia (2009). "Explicit probability density function for the length of a busy period in an $M/M/1/K$ queue." In: *Advances in queueing theory and network applications*. Springer, New York, pp. 213–226. DOI: 10.1007/978-0-387-09703-9_12. URL: https://doi.org/10.1007/978-0-387-09703-9_12.

Tsarev, Sergey P (2007). "On factorization and solution of multidimensional linear partial differential equations." In: *Computer Algebra 2006: Latest Advances in Symbolic Algorithms*. World Scientific, pp. 181–192.

Vershynin, Roman (2018). *High-dimensional probability*. Vol. 47. Cambridge Series in Statistical and Probabilistic Mathematics. An introduction with applications in data science, With a foreword by Sara van de Geer. Cambridge University Press, Cambridge, pp. xiv+284. ISBN: 978-1-108-41519-4. DOI: 10.1017/9781108231596. URL: https://doi.org/10.1017/9781108231596.

Vovk, Volodimir G. (1990). "Aggregating Strategies." In: *Proceedings of the Third Annual Workshop on Computational Learning Theory*. COLT '90. Rochester, New York, USA: Morgan Kaufmann Publishers Inc., pp. 371–386. ISBN: 1-55860-146-5. URL: http://dl.acm.org.proxy.lib.umich.edu/citation.cfm?id=92571.92672.

Wainwright, Martin J. (2019). *High-dimensional statistics*. Vol. 48. Cambridge Series in Statistical and Probabilistic Mathematics. A non-asymptotic viewpoint. Cambridge University Press, Cambridge, pp. xvii+552. ISBN: 978-1-108-49802-9. DOI: 10.1017/9781108627771. URL: https://doi.org/10.1017/9781108627771.

Walton, Neil and Kuang Xu (2021). “Learning and Information in Stochastic Networks and Queues.” In: DOI: 10.48550/ARXIV.2105.08769. URL: https://arxiv.org/abs/2105.08769.

Williams, Ruth J (1995). “Semimartingale reflecting Brownian motions in the orthant.” In: *IMA Volumes in Mathematics and its Applications* 71, pp. 125–125.

Xiong, Naixue et al. (2010). “A novel self-tuning feedback controller for active queue management supporting TCP flows.” In: *Information Sciences* 180.11, pp. 2249–2263. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2009.12.001. URL: https://www.sciencedirect.com/science/article/pii/S0020025509005258.

Yang, Zixian, R. Srikant, and Lei Ying (Apr. 2023). “Learning While Scheduling in Multi-Server Systems With Unknown Statistics: MaxWeight with Discounted UCB.” In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. Ed. by Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent. Vol. 206. Proceedings of Machine Learning Research. PMLR, pp. 4275–4312. URL: https://proceedings.mlr.press/v206/yang23d.html.

Zhang, Honggang et al. (June 2003). “A Self-Tuning Structure for Adaptation in TCP/AQM Networks.” In: *SIGMETRICS Perform. Eval. Rev.* 31.1, pp. 302–303. ISSN: 0163-5999. DOI: 10.1145/885651.781068. URL: https://doi.org/10.1145/885651.781068.

Zhong, Yueyang, John R. Birge, and Amy Ward (2022). “Learning the Scheduling Policy in Time-Varying Multiclass Many Server Queues with Abandonment.” In: DOI: 10.2139/ssrn.4090021. URL: http://dx.doi.org/10.2139/ssrn.4090021.