Abandonment vs. Blocking in Many-Server Queues: Asymptotic Optimality in the QED Regime

Ananda Weerasinghe* and Avishai Mandelbaum†
Iowa State University and Technion-Israel Institute of Technology

May 15, 2013

Abstract

We consider a controlled queueing system of the $G/M/n/B+GI$ type, with many servers and impatient customers. The queue-capacity B is the control process. Customers who arrive to a full queue are blocked and customers who wait too long in the queue abandon. We study the tradeoff between blocking and abandonment, with cost accumulated over a random, finite time-horizon, which yields a queueing control problem (QCP).

In the many-server Quality and Efficiency-Driven (QED) regime, we formulate and solve a diffusion control problem (DCP) that is associated with our QCP. The DCP solution is then used to construct asymptotically optimal controls (of the threshold type) for QCP. A natural motivation for our QCP is telephone call centers, hence the QED regime is natural as well. QCP then captures the tradeoff between busy signals and customer abandonment, and our solution specifies an asymptotically optimal number of trunk-lines.

Keywords: Many-server queues, call centers, Halfin-Whitt (QED) heavy-traffic regime, local-time process, diffusion processes and approximations, asymptotic optimality.
AMS Subject Classifications: 93E20, 60H30.
Abbreviated Title: Abandonment vs. Blocking.

*Research partially supported by Army Research Office Grant W 911NF0710424.

†Research partially supported by BSF Grants 2005175 and 2008480, ISF Grant 1357/08, the Technion funds for promotion of research and sponsored research and the Statistics and Applied Mathematical Sciences Institute (SAMSI) of the NSF.

1 Introduction

We consider a queueing system with a single customer class and a fixed number of independent but statistically identical servers. The customers leave the system after completing their service. They are served according to a "First-Come-First-Serve"(FCFS) discipline. When all the servers are busy catering to customers, incoming new customers form a queue. The *queue capacity* is a system manager's choice, with infinite capacity also allowed. The system manager can choose the queue capacity to be a non-negative integer-valued, time-dependent random variable, which may depend on past history as well as current state of the system. Incoming customers are *blocked* (rejected) if arriving when the queue is full at capacity. Once in queue, they abandon if their patience expires, which happens after random times, iid across customers. (The system manager must remove waiting customers from the queue, at times when queue capacity is reduced below current queue-length. In such a situation, there is an associated cost, which is proportional to the number of removed customers.)

In our Queueing Control Problem (QCP), the only control available to the system manager is the queue-capacity. This capacity can depend on past history subject to some constraints. It is also possible to view our QCP as an admission control problem of a queueing system, where the manager rejects the incoming customers when the current capacity of the system is full. The objective is to minimize a cost functional which is trading off blocking against abandonments: the larger the queue-capacity the less blocked customers which, in turn, leads to longer queues, hence more waiting and thus more abandonment. In fact, we also allow revenues from completed services, which provides an incentive against losing customers (via either blocking or abandonment). Mathematically, however, profit maximization turns out equivalent to cost minimization.

1.1 Motivation: Call Centers in the QED Regime

The motivation for our study is telephone call centers: being blocked amounts to encountering a *busy-signal*, waiting in queue entails seizing a *trunk-line* (typically "listening to music"), and abandoning is *hanging-up* (prior to being served). In practice, due to diminishing technological costs but also the lack of a rationalized tradeoff, call centers opt for the option of infinite-trunking: essentially no busy-signals - namely all customers join the queue. Our goal here is to identify the circumstances under which such practice is optimal - it is not always so - and to quantify its losses when optimality fails.

Operationally, call centers can be viewed as queueing systems [9]. Due to their increasing significance in Western Economies, call centers have given rise to ample research which has demonstrated that asymptotic queueing theory, as the number of servers increases, is well suited for their analysis. Typically, call centers sign a long-term lease of trunk lines with a telecommunications provider. They may not have the freedom to change the capacity of their trunk lines too often. However, a call center can always refuse incoming calls when there are too many customers queueing up for service. In concert with this state of affairs, we analyze the above QCP in a heavy-traffic many-server regime. Formally, we shall consider a sequence of queueing models in the "Quality and Efficiency-Driven" (QED) many-server heavy-traffic regime, which is characterized by condition (2.1) below; here, the number of servers n and the

arrival rate λ_n jointly increase indefinitely so that a "square-root staffing-rule" prevails: $n \approx R_n + \beta\sqrt{R_n}$, where $R_n = \lambda_n/\mu$ is the system's offered-load; for related research, we refer to [13, 5, 10].

The QED regime (2.1) was first formalized by Halfin and Whitt [13], who established both steady-state and process (diffusion) limits for the GI/M/n model. In [10], abandonments were added and [32] addressed the heavy traffic approximation for the G/GI/n model. Massey and Wallace [27] allowed finite queueing-capacity and [19, 18] and [29] analyzed the M/M/n/B+M queue with both blocking and abandonment. Recently [6] established, for the G/G/n+GI queue, an asymptotic relationship between the number of abandoning customers and the integral of queue length in diffusion-scale - we use their results here. For the G/GI/n+GI queue, diffusion approximations for both queue length and virtual waiting times were developed in [26]. Our setup for the abandonment mechanism is adapted from [6] and [26]. Additional related work appears in [1], [2], [5], [31] and [37]. Our mathematical framework follows [1], [2] and [29], the latter being a very useful survey with an extensive reference list.

1.2 Related work and our contribution

In a recent article [20], Kocaga and Ward also consider an admission control problem for the GI/M/n/B+M queue, with the blocking and customer abandonment features. They minimize a *long-run* expected average cost functional, via an admission-control policy that chooses queue capacity that depends on the total number of customers currently in the system. When the arrivals are Poisson and the number of servers is finite, they formulate the problem as a Markov Decision Problem(MDP), for which an algorithm is developed to compute a threshold-type optimal policy. They derive an approximating Diffusion Control Problem (DCP) within the Halfin-Whitt (QED) regime, and obtain a static threshold-type optimal control policy for the DCP. Extensive numerical experiments show that the optimal threshold for the DCP and its value function are excellent approximations for those of the MDP's, even when the number of servers is small, at the level of 10-20. (This accuracy has been repeatedly confirmed, since first identified in [5]; recently, it has found mathematical grounding in [34, 35].) In general, computations related to QCPs are numerically taxing; hence, the application of diffusion approximations via DCPs provides a useful elegant method for understanding optimal solutions of QCPs and their qualitative behavior.

In our model, queue capacity may depend on the current state as well as the past history of the system. Following the formulation of the problem in [20], ours can also be interpreted as an admission control problem, but the long-run average cost minimization problems, such as in [20], address only steady-state performance, while the problem addressed here depends on transient behavior of the system as well. Hence the optimal value in [20] does not depend on the initial state and is also insensitive to costs incurred over a short time horizon. In call center operations, faced with a time-varying arrival rate, it is common practice to divide the day into short time periods and use a model with a constant arrival rate in each of these short time periods. Therefore, it is important to introduce a cost functional on each of these short time periods which is sensitive to transient behavior and dependent on initial data. In fact, steady state costs, such as those in [20] seem less relevant here. Motivated by this application, we model such a time period by a random interval $[0, \tau]$ where τ is exponentially distributed with a parameter γ and is independent of system dynamics. We introduce a cost functional

over this random time interval $[0, \tau]$ which is equivalent to an infinite-time horizon discounted cost functional (2.60) (see Sections 2.5 and 2.6). This cost structure is indeed sensitive to transient behavior, as manifested by our value function being a function of the initial state. Mathematical analysis of the corresponding second-order HJB equation in (3.7), associated with our value function, is more challenging than that of the long-run average cost problems. (The HJB equation for the latter problems can be reduced to a first-order equation.) The asymptotically optimal lower bound for the cost functional of the QCPs is the value function of the DCP, as established in Theorems 4.1 and 4.2. Indeed, this value function depends on the initial data, as seen in (3.26) and (3.27).

Further contributions are as follows: In our queueing model, we incorporate history dependent queue-capacity choices. A QCP is formulated with path-dependent queue-capacity choices to minimize an infinite-horizon discounted cost functional. Next, we derive an approximating DCP, which is equipped with the same cost functional, within the Halfin-Whitt (QED) regime. Admissible processes of this DCP are quite general, and the corresponding stochastic differential equation (SDE) exhibits a non-linear drift coefficient. We also allow path-dependent reflection barriers in this SDE. In Theorem 3.7, we obtain an explicit "threshold-type" Markovian optimal strategy for the DCP. Then, in Theorem 4.1, we show that the value function of the DCP is an asymptotic lower bound for the value functions of the scaled QCP's, with general path-dependent queue-capacity policies. These general policies may include non-threshold type path-dependent queue-capacity policies as well. Finally, using the optimal threshold of the DCP, we construct a sequence of constant (static) queue-capacity policies for the scaled QCP's and then establish their asymptotic optimality in Theorem 4.2. Since our proposed asymptotic optimal policies are with constant queue-capacities, we need only weak convergence results for diffusion scaled queue-lengths with constant queue-capacities. This result was obtained in Theorem 2.2 and, in the case of Markovian abandonment with finite queue-capacity, it is available in [29]. We also strengthen Theorem 2.2 by establishing convergence of the cost functionals.

1.3 Survey of Results

The paper is organized as follows: We introduce our queueing model, related weak convergence results and the cost structure in Section 2. We allow a general arrival process (no Markovian assumptions), subject to the assumptions (2.1), (2.3) and (2.4).

Our model is of the G/M/n/B+GI type. There are n exponential servers; the queue capacity is finite or infinite, which is time-dependent, random and at the discretion of a system controller. (If the controller starts with infinite capacity, then it will be kept infinite throughout.) The abandonment process is general, namely (im)patience is assumed to be random with a distribution function F, and customers abandon only while queueing. Motivated by profit maximization in large call centers, which involve various cost factors in their maintenance, we seek to optimize the cost functional (2.60).

Let the number of servers n tend to infinity, with the arrival-rate and queue-capacity scaled as in (2.1) and (2.42) respectively. Weak convergence of the relevant processes under Markovian abandonment was already established for the systems with a constant queue-capacity (see [29] and references therein). In particular, the total-population process converges to a diffusion process with an upper *reflection barrier* when the normalized queue-capacity converges to this barrier. Moreover, the process of blocked

customers in the n^{th} system converges weakly to the local-time process of the diffusion at its reflection barrier. Here we use the recent work of [6] and generalize the weak convergence results to allow general patience-time distributions. We also provide a proof that the expected cost for the n^{th} system, when properly scaled as in (2.60), converges to that of the limiting diffusion process.

In Section 3, we formulate the Diffusion Control Problem (DCP) arising from our Queueing Control Problem (QCP). This turns out to be a singular control problem. For queueing systems in conventional heavy traffic, where the arrival and service rates both converge to the same finite value, an admission control problem that is similar to ours is analyzed in [36]; stochastic control problems associated with admission control and the control of service rates were addressed recently in [11] and [12]. In contrast, our DCP has a non-linear drift coefficient and state space on $(-\infty, \infty)$ while in those articles, the state space was $[0, \infty)$ since the diffusion-scaled queue-length converges to a non-negative diffusion process under conventional heavy-traffic regime. Non-linearity of the drift here leads to technical difficulties in verifying the optimality of our candidate policy.

We are able to derive a complete solution to the DCP associated with our QCP. We solve it in Theorem 3.7, which is the first of our two main results. Specifically, we show that when the lost profit per rejected customer is greater than or equal to the threshold value in (3.23), the optimal control is the null process, namely there is no blocking of customers; the corresponding state-process for the DCP is then a diffusion on the entire real line. If, on the other hand, the lost profit per rejected customer is less than this threshold, the optimal state-process is a reflected diffusion with an upper reflection barrier, and the optimal control is its local-time process at this barrier. The reflection barrier serves as the optimal scaled queue capacity for DCP. Our results also show that the value function of DCP is a smooth convex solution of the associated Hamilton-Jacobi-Bellman equation.

In Section 4, we obtain convergence of the expected value of the cost functional of the n^{th} system to that of the limiting diffusion. Our methods are influenced by those in [2]. In particular, the moment condition (2.4) guarantees that our results remain valid for general arrival processes. (For diffusion approximations of general arrival processes, we refer to [21] and [37].) Our DCP solution then yields an asymptotically optimal policy for the original queueing system, or QCP. This is proved in Theorems 4.1 and 4.2, which is our second set of main results. More specifically, and in complete analogy to DCP, we show that when the lost profit per rejected customer is greater than or equal to the threshold (3.23), having no blocked customers (no busy signals in a call center) is an asymptotically optimal policy. Conversely, if the lost profit per rejected customer is less than this threshold, the asymptotically optimal queue-capacity for the n^{th} system is a finite constant, as given in (4.5).

Remark: In our model, we have not imposed any costs that arise from *changing* queue-capacity, which is relevant for policies in which the latter is time-dependent. When customers are rejected by such queue-capacity changes, it leads to additional revenue losses (such as giving away coupons or future concessions) since they were kicked out while waiting patiently in the queue. This, in turn, will add additional costs to the cost funtional. But our optimal strategy for DCP, as well as the asymptotically optimal strategies for QCPs, are associated with *constant* queue-capacities and, therefore, such scenarios are avoided. Consequently, our policies still remain (asymptotically) optimal if there were additional costs incurred each time the system controller changes the

queue-capacity.

1.4 Busy Signal or Abandonment?

Our results yield insights on the tradeoff between busy-signal and abandonment in call centers. Indeed, the threshold p_0 in (3.23) can be interpreted as the discounted cost of abandonment: it consists of the cost of waiting until abandonment plus the cost for the abandonment itself. As expected, when busy-signal costs exceed abandonment costs, it is optimal to have ample trunk-lines that render the busy-signal phenomenon negligible. But introducing a busy signal could also be beneficial (blocking costs dominated by abandonment costs). In this case, the queue-capacity is to be dimensioned proportionally to the square-root of the number of servers n. Thus, the number of trunk-lines to be deployed is $n + z_p\sqrt{n}$; here z_p is the optimal reflection boundary obtained for DCP in Section 3, which is characterized by (3.34) (Numerical computations of z_p are given in Section 3.4). With this dimensioning, our asymptotically optimal queueing systems operate in the QED regime, as analyzed in [18, 27, 19, 29] for the case of Poisson arrivals. Hence, the approximations in these references, covering operational performance measures (extensively in [19]), can be readily employed. In particular, the fraction of customers that encounter a busy-signal, and the fraction of those who abandon, are both of order $1/\sqrt{n}$.

To gain insight into the existence of such a threshold p_0, consider an M/M/n+M queue with an infinite buffer. The waiting time for a customer who, upon arrival, faces an "extremely long" queue is the minimum of two independent exponential random variables, with parameters θ (for abandonment) and γ (for observation period) respectively. The cost of waiting per unit time is a and the mean waiting time is $\frac{1}{(\theta+\gamma)}$. Hence $p_0 = \frac{a}{(\theta+\gamma)}$ represents the expected cost for this customer. If the rejection cost $p > p_0$, it is reasonable to let this customer wait in the queue regardless of the queue-length at the customer's arrival. On the other hand, if $p < p_0$, it is reasonable to reject the customer by maintaining a finite queue-capacity.

In summary, our solution yields a simple, asymptotically optimal design rule for the queueing model, which is quite easy to implement in a call center setting. The threshold p_0 (and the corresponding queue-capacity $\sqrt{n}z_p$ when $p < p_0$) can be easily calculated from system parameters.

Notation.
We denote the function space of real-valued right-continuous functions with left limits, defined on $[0,\infty)$, by $\mathbf{D}[0,\infty) \equiv \mathrm{D}([0,\infty),\mathbf{R})$. Similarly, the function space of real-valued right-continuous functions with left limits, defined on $[0,T]$ is denoted by $\mathbf{D}[0,T]$. The function space of R^k valued, right-continuous functions with left limits, defined on $[0,\infty)$ is denoted by $\mathbf{D}^k[0,T]$, where $k \geq 1$. These function spaces are endowed with the standard Skorokhod J_1 topology. The identity function is denoted by e, where $e(t) \equiv t$ for all t. The uniform norm on an interval $[0,T]$ for a function f in $\mathbf{D}[0,T]$ is defined by

$$||f||_T = \sup_{0\leq t\leq T} |f(t)|,$$

for any $T > 0$. Similarly, for a process X with paths in $\mathbf{D}[0,T]$, the norm $||X||_T$ is defined by $||X||_T = \sup_{0\leq t\leq T} |X(t)|$. Throughout, we use $\Rightarrow$ to denote weak convergence

of processes. We also follow the convention that the infimum of an empty set is infinity. For any real number x, x^+ represents $\max\{0, x\}$ and x^- is $\max\{0, -x\}$. For any two real numbers a and b, $a \wedge b$ stands for $\min\{a, b\}$.

2 The Queueing Control Problem (QCP)

2.1 The Stochastic Model

All our stochastic processes are defined on a complete probability space $(\Omega, \mathfrak{F}, P)$. We have a sequence of queueing systems indexed by $n = 1, 2, \cdots$, where the parameter n represents the number of servers in the n^{th} system. All these systems operate under the First Come First Served (FCFS) discipline. Their service times are iid, exponentially distributed with mean $\frac{1}{\mu}$ and independent of the arrival process. Incoming customers are impatient and they leave the system if they wait too long in the queue. This abandonment mechanism works as follows: with each customer, there is an associated clock. This clock rings after a random time, which has the distribution function G. If the clock rings while the customer is waiting in the queue then the customer abandons the system. Otherwise, the clock is ignored and the customer gets served. These clocks are all iid and independent of the arrival and service processes, as well as the history of the system up to that time. Such a description of abandonment is reasonable with invisible queues, for example, those in telephone call centers.

Services and impatience do not vary with the number of servers n, but the arrival process does. To denote the dependence on n, for the n^{th} system, n will appear as a subscript in all associated parameters and processes. Specifically, let (λ_n) be a sequence of positive real numbers which satisfies

$$\lim_{n\to\infty} \frac{\lambda_n - \mu n}{\sqrt{n}} = -\mu\beta, \tag{2.1}$$

where β is a constant. For the n^{th} queueing system, we allow a general arrival process A_n with sample paths in the function space $\mathbf{D}[0, \infty)$, with $A_n(0) = 0$, and which satisfies assumptions (2.3) and (2.4) below. To this end, introduce the normalized arrival process $\hat{A}_n$ by

$$\hat{A}_n(t) = \frac{A_n(t) - \lambda_n t}{\sqrt{n}}, \quad t \geq 0. \tag{2.2}$$

We make two assumptions on this normalized arrival process. First, we assume that

$$\hat{A}_n \Rightarrow \sigma_1 W_1 \ , \tag{2.3}$$

as n tends to infinity, where σ_1 is a positive constant and W_1 is a standard Brownian motion. Second, we also assume the following moment condition:

$$E[\sup_{[0,T]} |\hat{A}_n(t)|^2] \leq K(1 + T^m) \ , \tag{2.4}$$

for two positive constants K and m. Both of these constants are assumed independent of T and n.

It is easy to construct a renewal-type arrival process A_n that satisfies both assumptions (2.3) and (2.4) above. For example, let (v_i) be a sequence of iid positive random

variables with $E[v_i] = 1$ and $\mathrm{Var}(v_i) = c^2 > 0$. Let (λ_n) be a sequence of positive real numbers that satisfies (2.1). Introduce the process A_n by

$$A_n(t) = \sup\{m \geq 0 : \sum_{j=1}^{m} v_j \leq \lambda_n t\} \ ,$$

for $t \geq 0$. Here $\sum_{j=1}^{0} v_j$ is considered zero. Then $\hat{A}_n$, defined according to (2.2), satisfies (2.3) with $\sigma_1^2 = c^2\mu$, and it also satisfies (2.4). For details, we refer to Lemma 2 of [2] and [21].

Next we introduce two processes Φ_n and Ψ_n, capturing the state of the queueing systems. Let $\Phi_n(t)$ be the number of customers waiting in queue at time t, and let $\Psi_n(t)$ be the number of customers being served at time t. Then, both quantities $\Phi_n(t)$ and $\Psi_n(t)$ are non-negative; since n represents the number of servers in the n^{th} system, it follows that

$$0 \leq \Psi_n(t) \leq n \ , \tag{2.5}$$

for all $t \geq 0$. It is also convenient to assume that the initial values $\Phi_n(0)$ and $\Psi_n(0)$ are deterministic. Let $Q_n(t)$ be the total number of customers in the n^{th} system at time t. Then clearly

$$Q_n(t) = \Phi_n(t) + \Psi_n(t) \ , \tag{2.6}$$

for all $t \geq 0$. Here $Q_n(0)$ represents the number of customers in the n^{th} system at time zero; we assume that their patience is infinite and they will not abandon the system. Such an assumption is not restrictive, as explained by Lemma 1, in Section 2 of [26]. We also postulate that, in the presence of queueing customers, all the servers must be busy ("work-conservation"). Hence the identities

$$\Phi_n(t) = (Q_n(t) - n)^+ \quad \text{and} \quad \Psi_n(t) = n - (Q_n(t) - n)^- = Q_n(t) \wedge n$$

must hold, for all $t \geq 0$. Recall that the service times of customers were taken to be exponentially distributed with mean $\frac{1}{\mu}$, where μ is a given constant. To represent the service completions during the interval $[0, t]$ for the n^{th} system, we begin with a unit-intensity Poisson process $\{S_n(t) : t \geq 0\}$. We assume that the process S_n is independent of the arrival process A_n. Then the number of service completions by all the servers during $[0, t]$ is represented by $S_n(\mu \int_0^t \Psi_n(s)ds)$.

As indicated, customers abandon the queue according to the patience-time distribution G. To represent abandoning customers in our model, we let $G_n(t)$ denote the number of customers who have abandoned the system during $[0, t]$. Since the initial customers have infinite patience, they will not make any contribution to the process G_n. For the i^{th} arrival, let t_i^n represent the arrival time, v_i^n represent the service time and d_i^n represent the patience-time of the customer. We assume that the sequences (t_i^n), (v_i^n) and (d_i^n) are independent of each other, (v_i^n) is iid with $exp(\mu)$ and the sequence (d_i^n) also iid with the probability distribution function G. Each abandoning customer reduces profit margins, as explained later in this section. We assume that the distribution function G satisfies $G(0) = 0$, and it is right-differentiable at zero with a positive right derivative

$$\theta = \lim_{x \to 0+} \frac{G(x)}{x} < \infty. \tag{2.7}$$

Similar assumptions were made in the recent works of [6] and [26], which analyze many-server queueing systems with general patience-time distributions.

We allow the system manager to choose a *queue-capacity* for the n^{th} system. When the queue is full, incoming customers will be rejected. Each rejected customer incurs a loss in profits. For the n^{th} system, a non-negative integer-valued stochastic process $m_n(\cdot)$ represents the process of controlled queue capacity. One is also allowed to assign the value $m_n(t) = \infty$, for all $t \geq 0$, which corresponds to infinite queue-capacity; in this case, no customers are ever rejected. This queue-capacity process m_n is the only control at the disposal of the manager of the n^{th} system. The choice of a large queue-capacity reduces blocking of customers and increases profit margins. But, at the same time, such a large capacity is likely to give rise to long queues, which will increase the number of abandonments from the system, leading to a loss of income. This tradeoff, between blocking and abandonment, naturally gives rise to a cost minimization, or a profit maximization problem, which is the underlying theme of our paper.

We impose the following constraints on the controlled queue-capacity process m_n. It has piecewise constant paths that are right-continuous with left limits (RCLL). At any time t, the random variable $m_n(t)$ can be infinite or it may take non-negative integer values. Furthermore, the control $m_n(t)$ is allowed to depend on the current state, as well as the whole history of the system up to time t. Therefore, we assume that the process m_n is adapted to the information filtration $(\mathcal{F}_{n,t})_{t\geq 0}$, given below in (2.13). In addition, the process m_n adheres to the following conditions:

(i) For each n, $m_n(0)$ is a non-negative, non-random quantity, which can be finite or $+\infty$. (2.8)

(ii) If $m_n(0) = +\infty$, then $m_n(t) \equiv +\infty$ for all $t \geq 0$. (2.9)

(iii) If $m_n(0) < \infty$, then $m_n(t) < \infty$, for all $t \geq 0$, and it satisfies the following growth condition: there exists $\delta_0 > 0$ such that, for all $0 < \delta < \delta_0$,

$$E[\sup_{|t-s|<\delta} |m_n(t) - m_n(s)|] \leq \sqrt{n}\, p(T)[\rho(\delta) + f(n)] \ , \tag{2.10}$$

where the supremum is taken over all $s, t \in [0, T]$ such that $|t-s| < \delta$. Here $p(\cdot)$ is a positive polynomial function; ρ is a positive, bounded continuous function defined on $[0, \infty)$, which satisfies $\lim_{r\to 0} \rho(r) = 0$; the function f is non-negative and satisfies $\lim_{n\to\infty} f(n) = 0$.

The above conditions will be mainly used in Section 4 to obtain a convergent subsequence of the normalized queue-capacity process $\frac{m_n}{\sqrt{n}}$ in the function space $\mathbf{D}[0, \infty)$. Loosely speaking, the above conditions impose a change of queue-capacity that at any time t will be at most of order $\sqrt{n}$. Consequently, we do not allow infinite changes of queue-capacity. In the case where the controlled queue-capacity m_n is restricted to be a deterministic time-dependent function, all our results prevail with (2.10) replaced by the following weaker assumption:

$$|m_n(t) - m_n(s)| \leq \sqrt{n}\, p(T)[\rho(|t - s|) + f(n)] \ .$$

Note that assumptions on the control allow constant m_n policies. At a jump point of m_n, say $t = t_0$, if $\Phi_n(t_0-) > m_n(t_0)$ then a number of customers that equals

$[\Phi_n(t_0-)-m_n(t_0)]$ are to be removed from the queue; this is in order to reach the queue-length of $\Phi_n(t_0) = m_n(t_0)$. Such a removal incurs a blocking cost that is proportional to $[\Phi_n(t_0-) - m_n(t_0)]^+$. (Our cost structure is formalized later.)

Using the above description, we obtain the following equation for the total customer population Q_n in the system (see Theorem 2.6 of [4], also the discussion in Section 7.3 and equation (114) of [29]):

$$Q_n(t) = Q_n(0) + A_n(t) - S_n(\mu \int_0^t \Psi_n(s)ds) - G_n(t) - U_n(t) \ , \tag{2.11}$$

where

$$U_n(t) = \int_0^t I_{[Q_n(s)=n+m_n(s)]}(s)dA_n(s) + \sum_{0\le s\le t} [Q_n(s-) - (n + m_n(s))]^+. \tag{2.12}$$

$G_n(t)$ is the number of customers that abandon the queue during $[0, t]$ and $U_n(t)$ represents the cumulative number of customers blocked or removed by the system manager during $[0, t]$. When the queue is full, new arrivals are blocked and the total number of such blocked customers during $[0, t]$ is represented by the integral term of (2.12). Furthermore, at a jump point s of the queue-capacity process, the manager changes the queue-capacity from $m_n(s-)$ to $m_n(s)$; if $n + m_n(s) < Q_n(s-) < n + m_n(s-)$ holds, then a number of customers that equals the quantity $[Q_n(s-) - (n + m_n(s))]^+$ will be removed from the queue. The last term of (2.12) represents the number of such removed customers during $[0, t]$. Hence, if the queue-capacity is a constant, then the last term of (2.12) vanishes.

The Skorokhod mapping on the interval $[0, n + m_n(\cdot)]$ (see the next subsection, as well as Theorem 2.6 of [4]), guarantees that U_n is the unique non-decreasing non-negative process that enforces the constraint $Q_n(t) \le n + m_n(t)$, at all $t \ge 0$ (see also [29]).

For each n fixed, let us introduce the filtration $\mathcal{F}_n = \{\mathcal{F}_{n,t} : t \ge 0\}$ by

$$\mathcal{F}_{n,t} = \sigma(A_n(s), S_n(\mu \int_0^s \Psi_n(r)dr), G_n(s), \Phi_n(s), \Psi_n(s) : 0 \le s \le t) \ , \tag{2.13}$$

completed by all the null sets. This $\sigma-$algebra represents all the information available to the system manager at time t. Furthermore, the reflection mapping defined on $[0, n + m_n]$ guarantees that the processes Q_n and U_n are adapted to the filtration $\mathcal{F}_n$; for details, see Theorem 7.4 of [29].

2.2 Basic Estimates.

Consider the fluid-scaled arrival process $\bar{A}_n$ defined by

$$\bar{A}_n(t) = \frac{A_n(t)}{n} \ , \tag{2.14}$$

for all $t \ge 0$. By (2.1), it is evident that $\lim_{n\to\infty} \frac{\lambda_n}{n} = \mu > 0$. Hence, for any $T > 0$, by (2.1) we have

$$\lim_{n\to\infty} ||\bar{A}_n - \mu e||_T = 0 \quad \text{in probability,} \tag{2.15}$$

where $e(t) \equiv t$ is the identity function. Moreover, by (2.4), we can also obtain

$$\lim_{n\to\infty} E[||\bar{A}_n - \mu e||_T^2] = 0. \tag{2.16}$$

Introduce the normalized state process X_n, normalized abandonment process G_n and the associated non-decreasing process V_n by

$$X_n(t) = \frac{Q_n(t) - n}{\sqrt{n}} \ , \tag{2.17}$$

$$\hat{G}_n(t) = \frac{G_n(t)}{\sqrt{n}} \ , \tag{2.18}$$

and

$$V_n(t) = \frac{U_n(t)}{\sqrt{n}} \ , \tag{2.19}$$

for all $t \geq 0$; here (Q_n, U_n) satisfies (2.11) and (2.12), with a queue-capacity process m_n. Then, following the derivation of Theorem 7.6 of [29], one can write the normalized state process X_n in terms of a martingale representation with respect to the filtration $(\mathcal{F}_{n,t})$ in (2.13), as described below. (Throughout the following discussion, we assume that $X_n(0)$ is deterministic, $\lim_{n\to\infty} X_n(0) = x$ exists and x is finite.) From (2.11) and (2.12), we write

$$\begin{aligned} X_n(t) = X_n(0) + \hat{A}_n(t) - M_n(t) - \hat{G}_n(t) + \frac{(\lambda_n - \mu n)t}{\sqrt{n}} \\ + \int_0^t \mu X_n^-(s)ds - V_n(t) \ , \end{aligned} \tag{2.20}$$

for all $t \geq 0$. Here,

$$M_n(t) = \frac{1}{\sqrt{n}}[S_n(\mu \int_0^t \Psi_n(s)ds) - \mu \int_0^t \Psi_n(s)ds] \ , \tag{2.21}$$

for all $t \geq 0$. This representation is similar to the equation (126) of [29]. The scaled process M_n is a right-continuous, square-integrable martingale with respect to the filtration $(\mathcal{F}_{n,t})$, and its predictable quadratic variation is given by

$$\langle M_n \rangle(t) = \frac{\mu}{n} \int_0^t \Psi_n(s)ds = \frac{\mu}{n} \int_0^t (Q_n(s) \wedge n)ds. \tag{2.22}$$

It also holds that

$$E[\sup_{0\leq t\leq T} \langle M_n \rangle(t)] \leq \mu T, \tag{2.23}$$

for all $T > 0$. In particular, $E[\langle M_n \rangle(t)] < \infty$, for all $t \geq 0$.

All the above martingale results are analogous to the proof of Theorem 7.6 of [29]. Moreover, the optional quadratic variation process (square-bracket process) of M_n is given by

$$[M_n, M_n](t) = \frac{1}{n} S_n(\mu \int_0^t \Psi_n(s)ds), \tag{2.24}$$

for all $t \geq 0$. Thus, $[M_n, M_n](T) \leq \frac{1}{n} S_n(\mu n T)$ and hence using Doob's inequality, we also obtain

$$E[\sup_{[0,T]} |M_n(t)|^2] \leq CT \ , \tag{2.25}$$

for $t \leq T$, where $C > 0$ is a generic constant, independent of $T > 0$, and $l \geq m + 1$. Following the proofs of Theorems 7.2 and 7.7 of [29], we also have the following weak convergence result in $\mathbf{D}[0, \infty)$:

$$M_n(\cdot) \Rightarrow \sqrt{\mu}\, W_2 \ , \tag{2.26}$$

where W_2 is a standard Brownian motion, independent of W_1 in (2.3).

We now establish the following lemma, which describes some useful basic estimates.

Lemma 2.1. *Assume (2.1), (2.3) and (2.4), together with the convergence of the initial values* $\lim_{n\to\infty} X_n(0) = x$. *Then the following results hold:*

(i)

$$\sup_{n\geq 1} E[||X_n||_T^2] \leq C_1(1 + T^k), \tag{2.27}$$

(ii)

$$\sup_{n\geq 1} E[\hat{G}_n(T)^2] \leq C_2(1 + T^k), \ \textit{and} \tag{2.28}$$

(iii)

$$\sup_{n\geq 1} E[V_n(T)^2] \leq C_3(1 + T^k), \tag{2.29}$$

where the constants $C_i > 0$, for $i = 1, 2, 3$, $k > 1$ are generic constants, independent of T as well as the sequence of queue-capacity processes (m_n).

Proof. Introduce the process ζ_n by

$$\zeta_n(t) = x_n + \hat{A}_n(t) - M_n(t) + \frac{(\lambda_n - \mu n)t}{\sqrt{n}}, \tag{2.30}$$

for all $t \geq 0$, where $X_n(0) = x_n$. Notice that when $X_n(t)$ takes large positive values, the processes G_n and V_n may increase and it helps X_n to decrease towards ζ_n. Similarly, if $X_n(t)$ takes large negative values, the term $\int_0^t \mu X_n^-(s)ds$ may influence X_n to increase. Therefore, we are able to obtain an upper bound for $||X_n - \zeta_n||_T$ as described below. First, by (2.1), (2.3) and (2.25), it is evident that

$$\sup_{n\geq 1} E[||\zeta_n||_T^2] \leq C(1 + T^k), \tag{2.31}$$

where $C > 0$ and $k \geq 1$ are generic constants independent of T. Hence, $||\zeta_n||_T$ is finite a.s. and we intend to show that $||X_n||_T \leq 3||\zeta_n||_T$, almost surely, using a pathwise argument. Introduce the process Y_n by

$$Y_n(t) = X_n(t) - \zeta_n(t) = \mu \int_0^t X_n^-(s)ds - \hat{G}_n(t) - V_n(t), \tag{2.32}$$

for all $t \geq 0$. Notice that Y_n has sample paths of bounded variation. Fix ω in the probability space Ω so that $M \equiv M(\omega) = ||\zeta_n||_T$ is finite. Suppose that $Y_n(t) > 2M$;

then $X_n(t) > M$ and thus $\Delta Y_n(t) = -(\Delta G_n(t) + \Delta V_n(t)) \leq 0$, for $0 \leq t \leq T$. Similarly, if $Y_n(t) < -2M$, then $X_n(t) < -M$ and hence $\Delta Y_n(t) = \mu X_n^-(t)\Delta t > 0$. Let H be the non-negative, twice continuously-differentiable function given by

$$H(x) = \begin{cases} (|x| - 2M)^4 & \text{if } |x| > 2M, \\ 0 & \text{if } |x| \leq 2M \end{cases}$$

Notice that $H'(x) < 0$ on $(-\infty, -2M)$, $H'(x) = 0$ on $[-2M, 2M]$ and $H'(x) > 0$ on $(2M, +\infty)$. Since Y_n has paths of bounded variation, we obtain $H(Y_n(t)) = \int_0^t H'(Y_n(s))dY_n(s) \leq 0$. Consequently, $|Y_n(t)| \leq 2M$ for all $0 \leq t \leq T$. This in turn yields that $||X_n||_T \leq 3||\zeta_n||_T$ a.s., which together with (2.31), implies (2.27) in part (i).

For part (ii), since G_n and V_n are non-negative, non-decreasing processes, by (2.32), we obtain

$$0 \leq \hat{G}_n(t) + V_n(t) \leq \mu \int_0^t X_n^-(s)ds + |Y_n(t)|,$$

for all $0 \leq t \leq T$. Hence, using the discussion in the proof of part (i), we deduce that $0 \leq \hat{G}_n(T) + V_n(T) \leq (3\mu T + 2)||\zeta_n||_T$. This, combined with (2.31), yields parts (ii) and (iii), which completes the proof. □

2.3 The Skorokhod Map

We now summarize several important properties of the Skorokhod map defined in the function space $\mathbf{D}[0, \infty)$. These properties are essential for later computations.

For a given function κ in $\mathbf{D}[0, \infty)$, the Skorokhod map $\Gamma_\kappa : \mathbf{D}[0, \infty) \to \mathbf{D}[0, \infty)$ with upper reflection boundary κ, is defined by

$$\Gamma_\kappa(f)(t) = f(t) - \sup_{[0,t]}(f(s) - \kappa(s))^+, \tag{2.33}$$

for each f in $\mathbf{D}[0, \infty)$. Let $V_f(t) = \sup_{[0,t]}(f(s) - \kappa(s))^+$. Then the pair $(\Gamma_\kappa(f), V_f)$ is called the *Skorokhod decomposition* of the function f, with respect to the upper reflection boundary κ. Clearly, $\Gamma_\kappa(f)(t) \leq \kappa(t)$, for all f in $\mathbf{D}[0, \infty)$ and for all $t \geq 0$.

We shall refer to the function κ as the *reflection curve*, or *reflection boundary*. First we assume that κ is a non-negative function in $\mathbf{D}[0, \infty)$. Later we discuss the situation that κ is a random process. For properties of the reflection map with a time-dependent reflection boundary, we refer to [4]. In the case κ is identically zero, the upward reflection map is analyzed in [14] and [22].

Using the representation (2.33), a direct computation yields

$$\sup_{[0,T]} |\Gamma_\kappa(f)(t) - \Gamma_\kappa(g)(t)| \leq 2||f - g||_T \ , \tag{2.34}$$

for any f and g in $\mathbf{D}[0, \infty)$, and all $T > 0$. Hence, the Skorokhod map, with respect to any reflection curve κ in $\mathbf{D}[0, \infty)$, is Lipschitz continuous under the sup norm in $\mathbf{D}[0, T]$ for each $T > 0$. In our computations in Section 4, we make use of the fact that the Lipschitz constant 2 is independent of the reflection curve. Let 0 represent the zero function, which is identically zero on $[0, \infty)$. Then, by (2.33), for any non-negative κ in

$\mathbf{D}[0,\infty)$, $\Gamma_\kappa(0)$ is identically zero. Hence, for a given κ in $\mathbf{D}[0,\infty)$, this fact combined with (2.34), yields

$$\sup_{[0,T]} |\Gamma_\kappa(f)(t)| \leq 2||f||_T \ , \tag{2.35}$$

for all f in $\mathbf{D}[0,\infty)$ and $T > 0$.

The following monotonicity property is also useful for our computations, and it readily follows from Propositions 3.4 and 3.5 of [4]: Let κ, f and g be in $\mathbf{D}[0,\infty)$ and $h(t) = f(t) - g(t)$, $t \geq 0$. If h is a non-negative, non-decreasing function in $\mathbf{D}[0,\infty)$, then

$$\Gamma_\kappa(f)(t) \geq \Gamma_\kappa(g)(t), \tag{2.36}$$

for all $t \geq 0$.

Next, let (κ_n) be a convergent sequence of non-negative functions in $\mathrm{D}[0,T]$ such that $\lim_{n\to\infty} \sup_{[0,T]} |\kappa_n(t) - \kappa(t)| = 0$, for some κ in $\mathrm{D}[0,T]$. Then, using (2.33), we obtain

$$\begin{aligned}\sup_{0\leq t\leq T} |\Gamma_{\kappa_n}(f)(t) - \Gamma_\kappa(f)(t)| &\leq \sup_{0\leq t\leq T} |\sup_{0\leq s\leq t} (f(s) - \kappa_n(s))^+ - \sup_{0\leq s\leq t} (f(s) - \kappa(s))^+| \\ &\leq \sup_{0\leq t\leq T} |\kappa_n(s) - \kappa(s)|.\end{aligned}$$

Hence, we conclude that

$$\lim_{n\to\infty} \sup_{0\leq t\leq T} |\Gamma_{\kappa_n}(f)(t) - \Gamma_\kappa(f)(t)| = 0, \tag{2.37}$$

for each $T > 0$, whenever (κ_n) converges uniformly to κ on $[0,T]$.

Let $(W(t))_{t\geq 0}$ be a Brownian motion in our complete probability space $(\Omega, \mathfrak{F}, P)$, and let $(\mathfrak{F}_t^W : t \geq 0)$ be the usual complete filtration generated by it. We pick a Lipschitz continuous drift function $b : \mathbb{R} \to \mathbb{R}$ with a Lipschitz constant $C > 0$. Consider a non-negative valued stochastic process $(\kappa(t))_{t\geq 0}$ in $\mathbf{D}[0,\infty)$, which is adapted to the Brownian filtration $(\mathfrak{F}_t^W : t \geq 0)$. Clearly, the definition (2.33) implies that, for any stochastic process $(Y(t))_{t\geq 0}$ in $\mathbf{D}[0,\infty)$ which is also adapted to $(\mathfrak{F}_t^W : t \geq 0)$, the path-wise reflected process $\Gamma_{\kappa(\cdot,\omega)}(Y)(\cdot,\omega)$ is a well-defined process in $\mathbf{D}[0,\infty)$ and is adapted to the Brownian filtration $(\mathfrak{F}_t^W : t \geq 0)$. Since the Lipschitz constant 2 is independent of the reflecting boundary κ in (2.34), we can obtain strong solutions to stochastic differential equations with drift coefficient b, a constant diffusion coefficient σ and a path-wise reflection boundary process $\kappa(\cdot,\omega)$. Here we outline the proof.

Since the function b is Lipschitz continuous with Lipschitz constant C, from (2.34) it follows that

$$\sup_{[0,T]} |b(\Gamma_\kappa(f)(t)) - b(\Gamma_\kappa(g)(t))| \leq 2\,C||f - g||_T,$$

and the constant C is independent of f, g, the process κ and $T > 0$. We begin with the strong solution Z to the equation

$$Z(t) = x + \sigma W(t) + \int_0^t b(\Gamma_\kappa(Z)(s))ds, \tag{2.38}$$

where x is the initial value and $\sigma > 0$ is a fixed constant. By the existence and uniqueness results for stochastic differential equations in Chapter 5 of [30], (2.38) has a unique strong solution in $\mathbf{D}[0,\infty)$, which is adapted to the Brownian filtration $(\mathfrak{F}_t^W$:

$t \geq 0$). Next, we use the non-negative adapted stochastic process $(\kappa(t))_{t\geq 0}$ in $\mathbf{D}[0,\infty)$ and let $X_x(t) = \Gamma_\kappa(Z)(t)$ and $V(t) = Z(t) - X_x(t)$. Then, the pair (X_x, V) is adapted to the Brownian filtration $(\mathfrak{F}_t^W : t \geq 0)$, and is the unique Skorokhod decomposition of the process Z. The processes X_x and V are in $\mathbf{D}[0,\infty)$ and they satisfy

$$X_x(t) = x + \sigma W(t) + \int_0^t b(X_x(s))ds - V(t) \ , \tag{2.39}$$

and

$$\int_0^t I_{[X_x(s)<\kappa(s)]} dV(s) = 0 \ , \tag{2.40}$$

for all $t \geq 0$. The process V is non-negative, non-decreasing with $V(0) = 0$. If $x > \kappa(0)$, X_x has an initial jump to $\kappa(0)$;thereafter, it satisfies (2.39) starting from $\kappa(0)$. In this case, we label $X_x(0-) = x$, $X_x(0) = \kappa(0)$ and $V(0) - V(0-) = x - \kappa(0) > 0$. If the reflection boundary process κ has continuous sample paths, then using the definition (2.33), it follows that the processes X_x and V also have continuous paths on the time interval $(0,\infty)$.

2.4 Weak Convergence.

In this subsection, we use Section 7 of [29], combined with Theorem 2.1 of [6] and Proposition 4.1 of [33] to derive a weak convergence result for properly normalized processes, when the queue-capacity m_n is a constant function. We thus try to remain consistent with the notation of [29].

Our ultimately proposed asymptotically optimal control sequence (m_n) is in fact a sequence of deterministic constant queue-capacities. Therefore, we need the weak convergence result in Theorem 2.2 only for constant queue-capacities (m_n). This will suffice for the asymptotic optimality result in Section 4.

In the proof of the following result, we use Theorem 2.1 of [6], which enables one to approximate the $\hat{G}_n$ process by an integral of the normalized queue length. Then we can follow the proof of Theorem 7.6 of [29], the results of [37], and Proposition 4.1. of [33] to deduce the conclusion.

Theorem 2.2. *Assume (2.1), (2.3) and (2.4), together with the convergence of the initial values*

$$\lim_{n\to\infty} X_n(0) = x, \tag{2.41}$$

where x is a scalar. Further assume that each m_n is a constant, and the following limit exists:

$$\lim_{n\to\infty} \frac{m_n}{\sqrt{n}} = \kappa \ , \tag{2.42}$$

with κ being a constant, satisfying $0 < \kappa \leq \infty$.

Then the sequence of processes $\{X_n : n \geq 0\}$ converges weakly to a diffusion process X_x in the function space $\mathbf{D}[0,\infty)$. This limit process is characterized by the stochastic differential equation

$$X_x(t) = x + \sigma_1 W_1(t) + \sqrt{\mu} W_2(t) - \beta\mu t - \int_0^t (\theta X_x^+(s) - \mu X_x^-(s))ds - U(t), \tag{2.43}$$

over $t \geq 0$. Here W_1 and W_2 are two independent standard one-dimensional Brownian motions; U is the unique non-negative, non-decreasing process in $\mathbf{D}[0,\infty)$, which satisfies $U(0) = 0$ and

$$\int_0^\infty I_{[X(s)<\kappa]}(s)dU(s) = 0. \tag{2.44}$$

When $\kappa = \infty$, the process U vanishes at all times. The constants β and σ_1 are given in (2.1) and (2.3).

Proof. Let $T > 0$ be fixed. We introduce the process ϵ_n by

$$\epsilon_n(t) = \hat{G}_n(t) - \theta \int_0^t X_n^+(s)ds, \tag{2.45}$$

for all $t \geq 0$, where $\theta > 0$ is given in (2.7). We intend to use Theorem 2.1 of [6] to show that $\lim_{n\to\infty} ||\epsilon_n||_T = 0$ in probability. This result was obtained for $G/G/n + GI$ queues in [6] and they allow their arrival process to be time-nonhomogeneous when the queueing model satisfies the assumptions (10), (11), (12) and (13) in [6], as carefully explained there. We make use of their result here.

First, introduce the process E_n of admitted customers to the system by $E_n(t) = A_n(t) - U_n(t)$, $t \geq 0$, where A_n is the arrival process and U_n is the process of rejected customers due to a full buffer. We can consider our system as a $G/G/n + GI$ queue, with E_n as the effective arrival process. Indeed, E_n may not be time-homogeneous, but we can apply results in [6] after verifying their assumptions. We introduce the processes $\bar{E}_n$, $\bar{G}_n$ by $\bar{E}_n(t) = \frac{E_n(t)}{n}$ and $\bar{G}_n(t) = \frac{G_n(t)}{n}$ respectively, for all $t \geq 0$. These processes are non-decreasing with RCLL paths. First, notice that, by (2.28) in Lemma 2.1, $\lim_{n\to\infty} \bar{G}_n(T) = 0$ in probability. Consequently, by (2.15), we have $\lim_{n\to\infty} ||\bar{E}_n - \mu e||_T = 0$, in probability, where e is the identity map. Therefore, it is straightforward to verify the assumption (10) in [6] (one may take $c_T = \frac{\mu}{2}$ there).

Similarly, assumption (11) of [6] is also evident from either (2.15) or (2.16). By part (i) of Lemma 2.1, (2.27) clearly implies that the diffusion scaled queue-length process is stochastically bounded and hence assumption (12) of [6] is verified.

Since we assume that initial customers do not abandon the queue, assumption (13) of [6] trivially holds. Therefore, using Theorem 2.1 of [6], we conclude that $\lim_{n\to\infty} ||\epsilon_n||_T = 0$ in probability.

We can write the state equation for X_n in the form

$$X_n(t) = \zeta_n(t) - \epsilon_n(t) - \int_0^t h(X_n(s))ds - V_n(t), \tag{2.46}$$

for all $t \geq 0$, where $h(x) = \theta x^+ - \mu x^-$ for all x, the process ζ_n as in (2.30) and ϵ_n as above. Moreover,

$$\int_0^\infty I_{[X_n(s)<\kappa_n]}(s)dV_n(s) = 0 \quad \text{a.s.} \tag{2.47}$$

Since $\lim_{n\to\infty} ||\epsilon_n||_T = 0$, in probability, using (2.1), (2.3), and (2.26), we conclude that the process $\zeta_n - \epsilon_n$ converges weakly to $x + \sigma_1 W_1 + \sqrt{\mu} W_2$, in the function space $\mathbf{D}[0,\infty)$, where W_1 and W_2 are two independent standard one-dimensional Brownian motions and the constants β and σ_1 are given in (2.1) and (2.3). The function h is Lipshitz continuous and, therefore, we can use the continuity of the integral representation of

(2.46) in the Skorokhod J_1 topology as in Theorem 7.4 of [29] (see also Section 4 of [33], and Section 4.3 in [24]). Then, one can follow the proofs of Theorems 1.2, 7.6 and 7.7 to obtain the weak convergence of (X_n, V_n) to (X_x, U) in $\mathbf{D}^2[0,\infty)$, where (X_x, U) satisfies (2.43). This completes the proof of the theorem. □

Remark. With the aid of (2.37), the above proof of the theorem can be easily extended to time-dependent reflection barrier processes, but as already noted, the above theorem suffices for our purposes.

2.5 Cost Structure

The cost structure of our model reflects several types of costs, associated with the operation of a telephone call center over a random time period $[0, \tau]$. Here τ is an exponentially distributed random variable with parameter γ, which is independent of system dynamics.

a) Abandonment costs.

With each abandoning customer, the losses for business exceed the loss of the immediate profit per customer, since there is a loss of goodwill and future business. (Under some circumstances, managers may wish to provide future preferential treatment, or discounts or coupons for future deals, in order to regain the goodwill of abandoning customers.)

Recall that $G_n(t)$ is the number of abandoning customers during $[0, t]$. Let $c_{ab} > 0$ be the cost incurred per abandoning customer. Then the expected cost due to abandonments, in the n^{th} system, is given by

$$C_n^A = c_{ab} E[G_n(\tau)]. \tag{2.48}$$

b) Delay costs.

Delay cost arises when the queue length $\Phi_n(t)$ is positive. For the n^{th} system, this cost is given by

$$C_n^D = c_d E[\int_0^\tau \Phi_n(t)dt] \ , \tag{2.49}$$

where c_d is a non-negative constant.

c) Idle server costs.

When there are not enough customers to make full use of the available servers, there could be a cost associated with the idle servers. This can be represented by

$$C_n^I = c_i E[\int_0^\tau (n - \Psi_n(t))dt] \ , \tag{2.50}$$

where c_i is a non-negative constant.

d) Blocking (busy-signal) costs.

Incoming customers are blocked when the total number of customers in the system is $n + m_n$. In addition, if the system controller decides to reduce the queue-capacity to a level below the queue-length at decision time, then some customers must be removed. Such situations lead to lost revenues, and the corresponding cost is represented by

$$C_n^B = c_b E[U_n(\tau)] \ , \tag{2.51}$$

where c_b is a positive constant.

e) 1-800 costs.

Here we introduce a cost that is motivated by 1-800 (toll-free) calls to a call center. To describe the cost incurred by 1-800 calls, let ϵ_0, $0 \le \epsilon_0 \le 1$, be the fraction of such calls received by the system. Then the associated 1-800 cost is given by

$$C_n^{800} = \epsilon_0 \, c_{800} \, E \int_0^\tau [\Phi_n(t) + \Psi_n(t)]dt \ , \tag{2.52}$$

where $c_{800} > 0$ is a constant that represents 1-800 cost per time-unit per occupied telephone-line. In the next section, we shall introduce an income parameter r per call; then, $r\mu = r/(1/\mu)$ represents the average income rate per time-unit of service; on the other hand, $\epsilon_0 c_{800}$ represents the average rate of lost income, due to 1-800 calls. Therefore, throughout we assume that c_{800} is small enough so that it satisfies $r\mu \ge \epsilon_0 c_{800}$.

With the above-described cost elements, the total expected cost for the n^{th} system is given by

$$\begin{aligned} C_n =& \epsilon_0 \, c_{800} \, E \int_0^\tau [\Phi_n(t) + \Psi_n(t)]dt + + c_d E[\int_0^\tau \Phi_n(t)dt] \\ &+ c_i E[\int_0^\tau (n - \Psi_n(t))dt] + c_{ab} E[G_n(\tau)] + c_b E[U_n(\tau)]. \end{aligned} \tag{2.53}$$

2.6 Profit Maximization.

The total expected revenue in the n^{th} system is given by $rE[H_n(\tau)]$, where r is the income per call; for each $t \ge 0$, $H_n(t) = S_n(\mu \int_0^t \Psi_n(s)ds)$, which represents the cumulative number of service completions up to time t. Since τ is independent of system dynamics, using the first martingale in (2.21), we obtain $E[H_n(\tau)] = \mu E \int_0^\tau \Psi_n(s)ds$. Hence, the total expected revenue can be represented by

$$R_n = r\mu E[\int_0^\tau \Psi_n(t)dt] \ . \tag{2.54}$$

It follows from (2.53) and (2.54) that the expected profit in the n^{th} system, $P_n = R_n - C_n$, is given by:

$$\begin{aligned} P_n =& r\mu E[\int_0^\tau \Psi_n(t)dt] - \Big(\epsilon_0 \, c_{800} \, E \int_0^\tau [\Phi_n(t) + \Psi_n(t)]dt \\ &+ c_d E \int_0^\tau \Phi_n(t)dt \\ &+ c_i E[\int_0^\tau (n - \Psi_n(t))dt] + c_{ab} E[G_n(\tau)] + c_b E[U_n(\tau)]\Big). \end{aligned} \tag{2.55}$$

The manager of the n^{th} system seeks to maximize P_n, over all possible choices of admissible queue-capacity processes $m_n(\cdot)$.

Next, we simplify the expression for P_n and discuss its scaled convergence. After straightforward computations, and using $\Psi_n(t) = n - (Q_n(t) - n)^-$, we obtain

$$\frac{P_n}{\sqrt{n}} = \frac{(r\mu - \epsilon_0 c_{800})\sqrt{n}}{\gamma} - E \int_0^\tau [(a_1 X_n^+(t) + b X_n^-(t))dt + c_{ab} \hat{G}_n(\tau) + p \cdot V_n(\tau)] \ , \tag{2.56}$$

where

$$a_1 = c_d + \epsilon_0 c_{800} > 0 \ , \quad b = r\mu + c_i - \epsilon_0 c_{800} \geq 0 \ , \quad p = c_b > 0, \tag{2.57}$$

and $\hat{G}_n$ is given in (2.18). Next, we observe that it is possible to replace the term $c_{ab}E[\hat{G}_n(\tau)]$ by $\theta c_{ab} E\int_0^\tau X_n^+(t)dt$. For this, let $\epsilon_n(t)$ be as in (2.45) in Theorem 2.2, and notice that $\lim_{n\to\infty} ||\epsilon_n||_T = 0$ in probability, as proved therein. Using (2.27) and (2.28) of Lemma 2.1, we have $\sup_{n\geq 1} E[||\epsilon_n||_T^2] \leq C(1+T^k)$, where $C > 0$ and $k > 1$ are generic constants independent of T. Since τ is independent of system parameters, $E||\epsilon_n||_\tau = E\int_0^\infty \gamma e^{-\gamma t}||\epsilon_n||_t dt$. Hence $(||\epsilon_n||_t)$ is uniformly integrable and we conclude that $\lim_{n\to\infty} E[||\epsilon_n||_\tau] = 0$. Consequently, $\lim_{n\to\infty} E[|\hat{G}_n(\tau) - \theta\int_0^\tau X_n^+(t)dt| = 0$. Therefore, we can write

$$\frac{P_n}{\sqrt{n}} = \frac{(r\mu - \epsilon_0 c_{800})\sqrt{n}}{\gamma} - E\int_0^\tau [(aX_n^+(t) + bX_n^-(t))dt + p\cdot V_n(\tau)] + \mathrm{o}(\frac{1}{\sqrt{n}}), \tag{2.58}$$

where

$$a = \theta c_{ab} + c_d + \epsilon_0 c_{800} > 0 \ , \quad b = r\mu + c_i - \epsilon_0 c_{800} \geq 0 \ , \quad p = c_b > 0. \tag{2.59}$$

It follows that the manager's objective, which originally was to maximize profit, is in fact to minimize the cost functional $E\int_0^\tau [(aX_n^+(t)+bX_n^-(t))dt+p\cdot V_n(\tau)]$. Since τ is independent of system dynamics and exponentially distributed with parameter γ, this cost functional can be written as $E\int_0^\infty \gamma e^{-\gamma t}[\int_0^t (aX_n^+(s) + bX_n^-(s))ds + p\cdot V_n(t)]dt$. Using Fubini's theorem, we observe that $E\int_0^\infty \gamma e^{-\gamma t}\int_0^t (aX_n^+(s) + bX_n^-(s))dsdt = E\int_0^\infty e^{-\gamma s}(aX_n^+(s) + bX_n^-(s))ds$ and $E\int_0^\infty \gamma e^{-\gamma t}V_n(t)dt = E\int_0^\infty \gamma e^{-\gamma t}\int_0^t dV_n(s)dt = E\int_0^\infty e^{-\gamma t}dV_n(t)$. Hence, we can write the cost functional in the form

$$J(X_n, V_n, p) = E\int_0^\infty e^{-\gamma t}[(aX_n^+(t) + bX_n^-(t))dt + p\cdot dV_n(t)] \ , \tag{2.60}$$

and the system manager would like to minimize it over all admissible choices of queue-capacity $m_n \geq 0$. In the rest of the article, we use this standard infinite horizon discounted form of the cost functional.

To identify asymptotically optimal policies for this optimization problem, we intend to describe a sequence of processes (X_n, V_n), and associated queue-capacity sequence (m_n), that minimize the limiting cost

$$\liminf_{n\to\infty} J(X_n, V_n, p) \ . \tag{2.61}$$

If a and b are both zero, then it is obviously optimal to choose infinite queue-capacity (i.e. $m_n \equiv \infty$), which makes the above cost functional identically zero. Therefore, we always consider $a > 0$, $b \geq 0$, and this is justified by (2.59) since $b = 0$ is possible, when $r\mu = \epsilon_0 c_{800}$ and $c_i = 0$. In Section 4, we intend to provide an asymptotically optimal sequence of processes (X_n^*, V_n^*), and associated queue-capacities (m_n^*), such that

$$\liminf_{n\to\infty} J(X_n^*, V_n^*, p) \leq \liminf_{n\to\infty} J(X_n, V_n, p) \ , \tag{2.62}$$

when compared against any other feasible sequence of processes (X_n, V_n).

3 The Diffusion Control Problem (DCP)

3.1 Problem Formulation

In this section, we formulate and solve a one-dimensional stochastic control problem for diffusion processes (DCP). This can be considered as the limiting form of the cost minimization problem for the queueing systems (QCP). In Section 4, we shall "translate" the optimal strategy of DCP back into asymptotically optimal strategies for QCP.

Consider a controlled state-process X_x, which is a weak solution to

$$X_x(t) = x + \sigma W(t) - \int_0^t [\beta\mu + h(X_x(s))]ds - U(t) \ , \tag{3.1}$$

where $X_x(0) = x$ is a real number, W is a standard one-dimensional Brownian motion, adapted to a right-continuous filtration $\mathcal{F} = \{\mathcal{F}_t : t \geq 0\}$ on a probability space $(\Omega, \mathfrak{F}, \mathbf{P})$. The σ-algebra $\mathcal{F}_0$ contains all the null sets in $\mathfrak{F}$, and for any $s \geq 0$, $t \geq 0$, the Brownian increment $W(t+s) - W(t)$ is independent of $\mathcal{F}_t$. The parameters $\sigma > 0$, $\mu > 0$ and β are constants; U is a non-negative, non-decreasing right-continuous process which is adapted to the filtration $\mathcal{F}$. In our stochastic control problem, this process U is considered to be the control process. The function h is Lipschitz continuous and is given by

$$h(x) = \begin{cases} \mu x & \text{if } x < 0 \\ \theta x & \text{if } x \geq 0, \end{cases} \tag{3.2}$$

where $\mu > 0$ and $\theta \geq 0$ are constants.

In our analysis of DCP, each vector-valued process (X_x, U, W) satisfying (3.1) and (3.2) is allowed to be defined in its own probability space. In the cases where the process U is identically zero, or U is the local time process of the reflecting diffusion X_x at an upper boundary κ, it is well known (Chapter IV, [16]) that there exists a unique non-explosive strong solution X_x which satisfies (3.1) and (3.2). For the system specified by (3.1), we introduce the cost functional

$$J(X_x, U, p) = E \int_0^\infty e^{-\gamma t}[(aX_x^+(t) + bX_x^-(t))dt + p\, dU(t)] \ , \tag{3.3}$$

where γ, a and p are positive constants, while b is a non-negative constant.

Our diffusion control problem (DCP) is to find an optimal control process U^*, with its corresponding state-process X_x^*, which minimize the above described cost functional. If a and b both vanish, then obviously the zero control process is optimal. Therefore, we assume throughout that $a > 0$ and $b \geq 0$. The results obtained in this section remain valid for any constant $\sigma > 0$ in (3.1). But, to relate (3.1) to the queueing control problem in (2.43), we consider

$$\sigma^2 = \sigma_1^2 + \mu$$

in Section 4. Here σ_1 is given in (2.3). This diffusion control problem belongs to the class of singular optimal control problems, which has been addressed in the literature (see, for example, [8], [38]).

For a given x in $\mathbf{R}$, we call the quintuple $((\Omega, \mathfrak{F}, \mathbf{P}), \mathcal{F}, W, X_x, U)$ *an admissible control system* if

(i) (X_x, U) is a weak solution to (3.1), and

(ii) the cost functional $J(X_x, U, p)$ is finite.

When there is no ambiguity, and with a slight abuse of the notation, we simply use the pair (X_x, U) to represent an admissible control system. Notice that the finiteness of the cost functional implies that the state process is nonexplosive. To define the value function, we introduce the set

$$\mathcal{A}(x) = \{(X_x, U) : (X_x, U) \text{ is admissible}\}\,. \tag{3.4}$$

This set is nonempty for each x in $\mathbf{R}$, since we can obtain a reflecting diffusion process with an upper reflection barrier at a point κ, for which our computations later in this section reveal that the corresponding cost functional is finite. The value function of the diffusion control problem is given by

$$V_p(x) = \inf_{\mathcal{A}(x)} J(X_x, U, p)\ , \tag{3.5}$$

for each x in $\mathbf{R}$.

Next, we describe the formal Hamilton-Jacobi-Bellman (HJB) equation related to our DCP. First, introduce the differential operator $\mathcal{G}$ by

$$\mathcal{G} = \frac{\sigma^2}{2}\frac{d^2}{dx^2} - (\beta\mu + h(x))\frac{d}{dx} - \gamma\ , \tag{3.6}$$

where the constants γ, β, μ and σ, are as described earlier, and the function h appears in (3.2). The formal HJB-equation can now be written as

$$\min\left\{\mathcal{G}F(x) + ax^+ + bx^-\ ,\ p - F'(x)\right\} = 0\ . \tag{3.7}$$

Here, the solution F is a sufficiently smooth function which satisfies the growth condition

$$|F'(x)| \le C_0 \tag{3.8}$$

for all x, where C_0 is a positive constant. Our results show that the value function in (3.5) is a twice continuously differentiable function, which satisfies both (3.7) and (3.8).

3.2 A Verification Lemma

Next, we establish a verification lemma which helps us identify an optimal strategy. It guarantees that any smooth function that satisfies the conditions (3.7) and (3.8) is a lower bound for the value function.

Lemma 3.1. *Let F be a twice continuously differentiable function on $\mathbf{R}$. Assume that F satisfies (3.7), (3.8) and the growth condition $|F(x)| \le C(1 + ax^+ + bx^-)$, for all $x \in \mathbf{R}$, with C being a positive constant. Then $F(x) \le V_p(x)$, for all $x \in \mathbf{R}$, where V_p is the value function defined in (3.5).*

Remark. Assumption (3.8) in fact yields the growth condition assumed in the above lemma, when a and b are both positive. We assume this growth condition to ensure that the conclusion of the lemma remains true when $a > 0$ and $b = 0$.

Proof. Let (X_x, U) be an admissible control system. Then the cost functional $J(X_x, U, p)$ is finite and consequently $E\int_0^\infty e^{-\gamma t}dU(t)$ is also finite. Using Fubini's theorem, we observe that

$$\int_0^\infty e^{-\gamma t}dU(t) = \int_{t=0}^\infty \int_{s=t}^\infty \gamma e^{-\gamma s}ds\, dU(t) = \gamma \int_0^\infty e^{-\gamma t}U(t)dt\ .$$

Therefore, one can write

$$J(X_x, U, p) = E\int_0^\infty e^{-\gamma t}[aX_x^+(t) + bX_x^-(t) + \gamma\, p\, U(t)]dt. \tag{3.9}$$

The finiteness of $J(X_x, U, p)$ now implies that $\liminf_{t\to\infty} e^{-\gamma t}E[aX_x^+(t)+bX_x^-(t)+\gamma\, p\, U(t)] = 0$. Therefore, we can find a deterministic sequence (T_n) that is increasing to infinity and

$$\lim_{n\to\infty} e^{-\gamma T_n}E[aX_x^+(T_n) + bX_x^-(T_n) + p\,\gamma U(T_n)] = 0. \tag{3.10}$$

Next we apply the generalized Itô's lemma (see [28], p. 285) to obtain

$$\begin{aligned} F(X_x(T_n))e^{-\gamma T_n} =& F(x) + \sigma\int_0^{T_n} e^{-\gamma s}F'(X_x(s-))dW(s) \\ &+ \int_0^{T_n} e^{-\gamma s}\mathcal{G}F(X_x(s-))ds - \int_0^{T_n} e^{-\gamma s}F'(X_x(s-))dU(s) \\ &+ \sum_{0\le s\le T_n} e^{-\gamma s}[\Delta F(X_x(s)) + F'(X_x(s-))\Delta U(s)]\ , \end{aligned} \tag{3.11}$$

where $\Delta F(X_x(s)) = F(X_x(s)) - F(X_x(s-))$, and $\Delta U(s) = U(s) - U(s-)$. Using assumption (3.8), it is easy to see the quantities

$$E[\sum_{0\le s\le T_n} e^{-\gamma s}[|\Delta F(X_x(s))|]] \quad \text{and} \quad E[\sum_{0\le s\le T_n} e^{-\gamma s}|F'(X_x(s-))|\Delta U(s)]$$

are both bounded by $C_0E[\int_0^\infty e^{-\gamma t}dU(t)]$, where $C_0 > 0$ is a constant. Hence, they are finite. Furthermore, (3.8) also implies that $E\left[\int_0^{T_n} e^{-\gamma s}F'(X_x(s-))dW(s)\right] = 0$. Next, let $\{U^c(t) : t \ge 0\}$ be the continuous part of the process $\{U(t) : t \ge 0\}$. Then, using the above facts, together with (3.11), we obtain

$$\begin{aligned} E[F(X_x(T_n))]e^{-\gamma T_n} =& F(x) + E\int_0^{T_n} e^{-\gamma s}\mathcal{G}F(X_x(s-))ds \\ &- E\int_0^{T_n} e^{-\gamma s}F'(X_x(s-))dU^c(s) \\ &+ E\sum_{0\le s\le T_n} e^{-\gamma s}[\Delta F(X_x(s))]\ . \end{aligned} \tag{3.12}$$

Notice that $\Delta F(X_x(s)) = F'(\zeta(s))(X_x(s) - X_x(s-))$, for some $\zeta(s)$ which lies between $X_x(s)$ and $X_x(s-)$. But $(X_x(s) - X_x(s-)) = -(U(s) - U(s-)) \le 0$. By (3.7), $F'(\zeta(s)) \le p$, and therefore, $\Delta F(X_x(s)) \ge -p\,\Delta U(s)$, for all $s \ge 0$. Consequently, $E\sum_{0\le s\le T_n} e^{-\gamma s}[\Delta F(X_x(s))] \ge -pE\sum_{0\le s\le T_n} e^{-\gamma s}[\Delta U(s)]$. By (3.7), we also obtain that

$-E\int_0^{T_n} e^{-\gamma s}F'(X_x(s-))dU^c(s) \geq -pE\int_0^{T_n} e^{-\gamma s}dU^c(s)$. Using these facts, together with (3.7) and (3.12), we get

$$\begin{aligned} E[F(X_x(T_n))]e^{-\gamma T_n} \geq & F(x) + E\int_0^{T_n} e^{-\gamma s}\mathcal{G}F(X_x(s-))ds - pE\int_0^{T_n} e^{-\gamma s}dU(s) \\ \geq & F(x) - E\int_0^{T_n} e^{-\gamma t}[(aX_x^+(t) + bX_x^-(t))dt + p\,dU(t)] \ . \end{aligned}$$

This implies that

$$E[F(X_x(T_n))]e^{-\gamma T_n} + J(X_x, U, p) \geq F(x) \ . \tag{3.13}$$

Next, we use the assumption $|F(x)| \leq C(1+ax^+ + bx^-)$, for all x, together with (3.10), to conclude that $\lim_{n\to\infty} E[F(X_x(T_n))]e^{-\gamma T_n} = 0$. Hence, (3.13) yields $J(X_x, U, p) \geq F(x)$, for all x, and the conclusion of the lemma follows.

□

3.3 Analysis of the HJB-equation.

In this subsection, we pave the way to construct a solution to the HJB-equation (3.7). This will help obtain an optimal state process, which is a reflected diffusion process on an interval $(-\infty, z_p]$, where z_p is a positive constant, yet to be determined.

We begin with the interval $(-\infty, 0)$, considering the family of solutions to $\mathcal{G}F(x) - bx = 0$ and $F'(x) < p$, for all $x < 0$ and $F(-\infty) = \infty$. By elementary computations, we can write such a solution in the following form:

$$F(x) = kF_\infty(x) + \frac{b}{\gamma+\mu}\left(\frac{\beta\mu}{\gamma} - x\right) \ , \tag{3.14}$$

for all $x \leq 0$, where k is an arbitrary real number, and F_∞ is a bounded solution to the homogeneous equation $\mathcal{G}F(x) = 0$, on the interval $(-\infty, 0)$. The fundamental set of this homogeneous differential equation consists of two linearly independent solutions. We pick the solution F_∞ that satisfies (3.15) and (3.16) below. The other solution grows exponentially fast near negative infinity and therefore, it violates our growth condition (3.8). Hence, it does not make any contribution in our analysis.

First, we construct a solution F_∞ that satisfies $\mathcal{G}F_\infty(x) = 0$, on the interval $(-\infty, 0)$, with the following boundary conditions:

$$\frac{\sigma^2}{2}F_\infty''(x) - (\beta\mu + \mu x)F_\infty'(x) - \gamma F_\infty(x) = 0 \quad \text{for} \quad x < 0, \tag{3.15}$$

$$F_\infty(0) = 1, \ \textit{and} \ \lim_{x\to-\infty} F_\infty(x) = 0. \tag{3.16}$$

In our discussion below, we extend the function F_∞ to $(-\infty, \infty)$, so that it satisfies (3.15) everywhere.

Lemma 3.2. *There is a unique solution F_∞ that satisfies (3.15) and (3.16) above. Furthermore, F_∞ is a strictly positive, strictly convex increasing function on $(-\infty, 0)$, which satisfies $\min\{F_\infty(x), F_\infty'(x), F_\infty''(x)\} > 0$, for all $x \leq 0$, as well as $\lim_{x\to-\infty} F_\infty'(x) = 0$.*

The proof of this lemma is independent of the results in the rest of this section; hence we present it in the appendix.

Remark.

1. The conclusion $F''_\infty(0) > 0$ in the above lemma implies that $\beta\mu F'_\infty(0) + \gamma > 0$, and this will be crucial in our construction of a solution to the HJB-equation.
2. In the case when $\gamma = \mu$, a closed-form solution for F_∞ can be obtained by elementary methods. In fact

$$F_\infty(x) = K\, e^{\frac{\gamma}{\sigma^2}[(x+\beta)^2 - \beta^2]} \int_{-\infty}^{x} e^{-\frac{\gamma}{\sigma^2}(u+\beta)^2} du \ ,$$

where K is a positive constant which is given by $K = (\int_{-\infty}^{0} e^{\frac{-\gamma}{\sigma^2}(u+\beta)^2} du)^{-1}$. In particular, $F'_\infty(0) = \frac{2\gamma\beta}{\sigma^2} + Ke^{-\frac{\gamma}{\sigma^2}\beta^2}$, and this will be useful for numerical computations in Section 3.4.

In terms of the function F_∞ in the above lemma, we now construct a family of functions $\{F_k\}$ on $(-\infty, 0)$, indexed by k running over the non-negative real numbers. Specifically,

$$F_k(x) = kF_\infty(x) + \frac{b}{\gamma+\mu}(\frac{\beta\mu}{\gamma} - x) \ , \quad x < 0, \tag{3.17}$$

where k is an arbitrary non-negative real number. To choose the correct value of k, so that F_k is the desired solution on $(-\infty, 0)$, we must look at the solutions to the HJB-equation (3.7) on the interval $(0, \infty)$.

For each real value k, we begin with the solution H_k to the initial value problem

$$\mathcal{G}H_k(x) + ax = 0, \quad \text{for } x \geq 0 \ , \tag{3.18}$$

with the initial data in agreement with $F_k(0)$ and $F'_k(0)$. Hence,

$$H_k(0) = k + \frac{\beta b\mu}{\gamma(\gamma+\mu)} \qquad \text{and} \qquad H'_k(0) = kF'_\infty(0) - \frac{b}{(\gamma+\mu)} \ . \tag{3.19}$$

From the standard theory of differential equations, there is a unique solution H_k for each k, which also satisfies (3.19) at the origin. Note that the graph of H_k joins smoothly with that of F_k at the origin and, consequently, $H_k(0) = F_k(0)$, $H'_k(0) = F'_k(0)$ and $H''_k(0) = F''_k(0)$. We need to investigate the behavior of H'_k for various values of k. Therefore, we introduce the function

$$G_k(x) = H'_k(x), \qquad \text{for all } x \geq 0. \tag{3.20}$$

By differentiating (3.18), we obtain

$$\frac{\sigma^2}{2}G''_k(x) - (\beta\mu + \theta x)G'_k(x) - (\theta+\gamma)G_k(x) + a = 0, \quad x > 0 \ . \tag{3.21}$$

Using (3.19) and the fact that (3.18) remains valid at the origin, we deduce the initial conditions

$$G_k(0) = kF'_\infty(0) - \frac{b}{(\gamma+\mu)} \quad \text{and} \quad \frac{\sigma^2}{2}G'_k(0) = k(\beta\mu F'_\infty(0) + \gamma) \ . \tag{3.22}$$

Hence, $G_k(0) = F_k'(0)$ and $G_k'(0) = F_k''(0)$. Note that, for a fixed real number k, the above initial value problem uniquely characterizes the solution G_k on the interval $[0, \infty)$. Next, we make two important observations. First, for a given G_k, we can uniquely determine the function H_k by $H_k(x) = H_k(0) + \int_0^x G_k(u)du$, where $H_k(0)$ satisfies

$$\gamma H_k(0) = \frac{\sigma^2}{2} G_k'(0) - \beta\mu G_k(0) = \gamma k + \frac{\beta b \mu}{(\gamma + \mu)} \ .$$

Second, the differential equation (3.21) for G_k has a unique constant solution given by $Y = a/(\theta + \gamma)$, which is independent of k. This quantity plays an important role in determining our optimal strategy. Therefore, we introduce the constant $p_0 > 0$ by

$$p_0 = \frac{a}{\theta + \gamma} \ . \tag{3.23}$$

Since $p_0 > 0$, by (3.22), it is evident that for each $k \geq 0$, the function G_k is not equal to the constant solution p_0. Our next lemma summarizes the properties of the family of functions $\{G_k : k \geq 0\}$.

Lemma 3.3. *Let p_0 be the constant defined in (3.23). For each real number $k \geq 0$, let G_k be the function given in (3.20). Then the following holds:*

(i) $G_k(x)$ is jointly continuous in (k, x), for $k \geq 0$ and $x \geq 0$.

(ii) Let $k > 0$. If $G_k'(\zeta) = 0$ for some $\zeta > 0$, then $x = \zeta$ is a local maximum for G_k and $G_k(\zeta) < p_0$. Furthermore, on the interval $[0, \infty)$, G_k cannot have any local minima and it can have at most one local maximum.

(iii) If $r_1 > r_2 > 0$, then $G_{r_1}(x) > G_{r_2}(x)$ and $G_{r_1}'(x) > G_{r_2}'(x)$, for all $x \geq 0$.

(iv) Let $k > 0$. If $G_k(x_0) \geq p_0$ for some $x_0 \geq 0$, then $G_k'(x) > 0$ for all $x \geq x_0$.

(v) If $k = 0$, then $G_0(0) = \frac{-b}{(\gamma+\mu)} \leq 0$, and G_0 is strictly decreasing on $[0, \infty)$.

Proof. Since the function $G_k(x)$ is the unique solution to the initial value problem described in (3.21) and (3.22), and the initial data in (3.22) is continuous in the variable k, assertion (i) follows from the standard theory of differential equations (see Chapter 5 of [15]).

Next, let $G_k'(\zeta) = 0$, for some $\zeta > 0$. Then, using (3.21), we have

$$\frac{\sigma^2}{2} G_k''(\zeta) = (\theta + \gamma)[G_k(\zeta) - p_0] \ , \tag{3.24}$$

where p_0 is given in (3.23). Let $k > 0$ and suppose that $G_k'(\zeta) = 0$ and $G_k(\zeta) > p_0$, for some $\zeta > 0$. Then, by (3.24), $G_k''(\zeta) > 0$ and $x = \zeta$ is a strict local minimum. But, $G_k'(0) = kF_\infty''(0) > 0$ and, therefore, G_k is strictly increasing in an interval $(0, \delta)$ for some $\delta > 0$. Consequently, G_k has a local maximum at a point x_0 such that $0 < x_0 < \zeta$ and $G_k(x_0) > p_0$. Since $G_k'(x_0) = 0$, then (3.24) implies that $G_k''(x_0) > 0$, and this is a contradiction. Hence $G_k(\zeta) \leq p_0$. If $G_k(\zeta) = p_0$, then we know that $G_k'(\zeta) = 0$ and G_k satisfies (3.21); also recall that $y = p_0$ is a constant solution to (3.21). Hence, by the uniqueness of solutions to (3.21), we conclude that $G_k(x) \equiv p_0$, for all x. This again contradicts the fact that $G_k'(0) > 0$, and hence $G_k(\zeta) < p_0$.

If G_k has a local minimum at some point $y_0 > 0$, then by (3.24), $G_k(y_0) > p_0$. But $G_k'(0) > 0$, and therefore G_k has a local maximum at some point ζ such that

$0 < \zeta < y_0$ and $G_k(\zeta) > p_0$. This contradicts the above proof. Consequently, G_k has no local minima on the interval $[0, \infty)$.

To prove (iii), let $r_1 > r_2 > 0$, and let $P(x) = G_{r_1}(x) - G_{r_2}(x)$, for $x \geq 0$. Recall that $F'_\infty(0) > 0$ and $\frac{\sigma^2}{2} F''_\infty(0) = (\beta\mu F'_\infty(0) + \gamma) > 0$, as in Lemma 3.2. Then, using (3.22), it follows that $P(0) > 0$ and $P'(0) > 0$. Furthermore, P satisfies the homogeneous equation $\frac{\sigma^2}{2} P''(x) - (\beta\mu + \theta x) P'(x) - (\theta + \gamma) P(x) = 0$, for $x > 0$. Hence, P cannot have any positive local maxima. This implies that $P(x) > 0$, for all $x \geq 0$. Therefore, if $P'(\zeta) = 0$, for some $\zeta > 0$, then from the above differential equation, it follows that $P''(\zeta) > 0$ and $x = \zeta$ is necessarily a local minimum. This, together with the fact that $P'(0) > 0$, implies that $P'(x) > 0$, for all $x \geq 0$. Hence, Assertion (iii) follows.

For part (iv), let $G_k(x_0) \geq p_0$, for some $x_0 > 0$. Suppose that $G'_k(x_0) < 0$. Since $G'_k(0) > 0$, it follows that there is a local maximum at a point ζ such that $0 < \zeta < x_0$, and $G_k(\zeta) > p_0$. This contradicts part (ii) above. Therefore, $G'_k(x_0) \geq 0$. Since $G_k(x_0) \geq p_0$, $G'_k(x_0) = 0$ also contradicts part (ii). Hence, $G'_k(x_0) > 0$. Now we can use again part (ii) to conclude that $G'_k(x) > 0$, for all $x \geq x_0$. This yields part (iv).

To prove part (v), notice that $G_0(0) = \frac{-b}{\gamma+\mu}$, $G'_0(0) = 0$ and, by (3.21), $\frac{\sigma^2}{2} G''_0(0) = -[a + \frac{(\gamma+\theta)b}{(\gamma+\mu)}] < 0$. Hence, G_0 is strictly decreasing in an interval $[0, \delta)$, for some $\delta > 0$. Then, using (3.24), similarly to the proof of part (ii), we can show that G_0 is strictly decreasing in $[0, \infty)$. This completes the proof of the lemma.

□

To facilitate the discussion of our next lemma, we make the following observation: Let us choose $k_1 > 0$ such that $k_1 F'_\infty(0) = \frac{a}{\gamma+\theta} + \frac{b}{\gamma+\mu}$. Then, by (3.22), $G_{k_1}(0) = p_0$ and $G'_{k_1}(0) = k_1 F''_\infty(0) > 0$. Now, we can use part (iv) of the above lemma to conclude that $G'_{k_1}(x) > 0$, for all $x > 0$. Using parts (i), (iii) and (iv) of Lemma 3.3, and by a straightforward argument, we conclude that there is an $\epsilon > 0$ and $\delta_\epsilon > 0$ such that $G_k(x) > p_0$, for all $x > \delta_\epsilon$, whenever $k > k_1 - \epsilon$. Then, it also follows that, for each $k > k_1 - \epsilon$, G_k is strictly increasing on the interval $[0, \infty)$. We use these facts in the proof of the next lemma, which is central for constructing our optimal strategy. It shows a dichotomy about the existence of a local maximum for the function G_k.

Lemma 3.4. *Let p_0 be the constant defined in (3.23). Then, there is a finite constant $k_0 > 0$ satisfying the following assertions:*

(i) The function G_{k_0} is strictly increasing on $[0, \infty)$ and $\lim_{x\to\infty} G_{k_0}(x) = p_0$.

(ii) If $k > k_0$, then G_k is strictly increasing on $[0, \infty)$ and $\lim_{x\to\infty} G_k(x) = \infty$.

(iii) If $0 < k < k_0$, then G_k has a unique maximum at a point z_k and $G_k(z_k) < p_0$.

(iv) For each p that satisfies $0 < p < p_0$, there is a constant k_p so that $0 < k_p < k_0$ and $\max_{x\geq 0} G_{k_p}(x) = p$. In particular, this constant k_p is unique.

(v) For each $0 < p < p_0$, let z_{k_p} be the unique maximum point of G_{k_p} which satisfies $G_{k_p}(z_{k_p}) = \max_{x\geq 0} G_{k_p}(x) = p$. Then $0 < p_1 < p_2 < p_0$ implies that $z_{k_{p_1}} < z_{k_{p_2}}$.

Proof. By part (v) of Lemma 3.3, $G_0(0) \leq 0$ and G_0 is strictly decreasing on $[0, \infty)$. If $k > 0$, then $G_k(0) > 0$ and $G'_k(0) > 0$ by (3.22). Using parts (i) and (iii) of Lemma 3.3, we can find $\delta_0 > 0$ such that, for each $0 < k < \delta_0$, the function G_k has a positive strict local maximum. We introduce

$$k_0 = \inf\{k > 0 : G_k \text{ is strictly increasing on } [0.\infty)\}. \tag{3.25}$$

By the discussion after the proof of Lemma 3.3, we know that every $k \geq k_1$ is in the above set and hence it is nonempty. Therefore, k_0 is well defined and $k_0 \leq k_1$. On the other hand, $k_0 \geq \delta_0 > 0$, since each G_k has a local maximum when $0 < k < \delta_0$. Therefore, $0 < \delta_0 \leq k_0 < k_1$. Now, let us consider the function G_{k_0}. Since $G'_{k_0}(0) > 0$, it follows that G_{k_0} is strictly increasing in a neighborhood of the origin. If G_{k_0} has a local maximum at a point $x = \zeta > 0$. Then, by (3.21), $G''_{k_0}(\zeta) < 0$ and it is a strict local maximum. Let us choose $M > 0$ so that $0 < \zeta < M$. Using part (i) of the lemma, we know that G_k approximates G_{k_0} uniformly on the interval $[0, M]$ if k is close to k_0. Therefore, using this, together with part (iii) of Lemma 3.3, we can find a $\delta_1 > 0$ such that, for each k in $[k_0, k_0 + \delta_1)$, G_k has a local maximum. This contradicts the definition of k_0 in (3.25). Hence, G_{k_0} is strictly increasing on $[0, \infty)$.

Now suppose that $G_{k_0}(y) > p_0$, for some y. Then, relying on parts (i) and (iii) of Lemma 3.3, we find a $\delta_2 > 0$ such that $G_k(y) > p_0$, for each $0 < k_0 - \delta_2 < k \leq k_0$. Then, using part (ii) of the same lemma, the function G_k cannot have any local extrema for each such k, and hence G_k is strictly increasing. This, again, contradicts (3.25). Therefore, $G_{k_0}(y) \leq p_0$, for all $y \geq 0$. But, if $G_{k_0}(y_1) = p_0$ for some $y_1 > 0$, then $G_{k_0}(y) > p_0$, for all $y > y_1$ and, then again, we get a contradiction as above. Consequently, the function G_{k_0} is strictly increasing on the interval $[0, \infty)$ and $G_{k_0}(y) < p_0$, for all $y > 0$. Next, we let $l_0 = \lim_{x\to\infty} G_{k_0}(x)$, then $l_0 \leq p_0$. We can write (3.21) in an integral form to obtain

$$\begin{aligned}\frac{\sigma^2}{2}G'_{k_0}(x) + ax =&(\frac{\sigma^2}{2}G'_{k_0}(0) - \beta\mu G_{k_0}(0)) + \beta\mu G_{k_0}(x)\\ &+ \theta x G'_{k_0}(x) + \gamma\int_0^x G_{k_0}(y)dy\ , \quad \text{for } x > 0\ .\end{aligned}$$

Using $G'_{k_0}(x) \geq 0$, and dividing the whole equation by x and letting x tend to infinity, we obtain $a \leq \theta l_0 + \gamma l_0$. Next, by (3.23), we obtain $l_0 \geq p_0$. Consequently, we have $\lim_{x\to\infty} G_{k_0}(x) = p_0$. This completes the proof of part (i).

Let $k > k_0$. Since $G'_k(0) > 0$, we have $G'_k(x) > 0$, for all $0 < x < \delta$, for some $\delta > 0$. Suppose that $G'_k(\zeta) = 0$, for some $\zeta > \delta$. Then by part (ii) of Lemma 3.3, we have $G_k(\zeta) < p_0$, $G''_k(\zeta) < 0$ and $x = \zeta$ is a strict local maximum. Again, using part (ii) of Lemma 3.3, we conclude that G_k is strictly decreasing in $[\zeta, \infty)$. Since $\lim_{x\to\infty} G_{k_0}(x) = p_0$, and $G_k(0) > G_{k_0}(0)$, we get $G_k(z) = G_{k_0}(z)$, for some $z > 0$. But this contradicts part (i) of Lemma 3.3. Consequently, $G'_k(x) > 0$, for all $x > 0$. Now let $L = \lim_{x\to\infty} G_k(x)$. Suppose that L is finite. Then $\liminf_{x\to\infty} G'_k(x) = 0$. We use the integral form of the differential equation (3.21) to obtain

$$\begin{aligned}\frac{\sigma^2}{2}G'_k(x) + ax =&(\frac{\sigma^2}{2}G'_k(0) - \beta\mu G_k(0)) + \beta\mu G_k(x)\\ &+ \theta x G_k(x) + \gamma\int_0^x G_k(y)dy\ , \quad \text{for } x > 0.\end{aligned}$$

We divide this equation by x and use $\liminf_{x\to\infty} G'_k(x) = 0$ to take the limit along a suitable subsequence to conclude that $L = p_0$. Now consider the function $U(x) = G_k(x) - G_{k_0}(x)$, for $x > 0$. Then $U(x) > 0$, for all $x \geq 0$ and $U'(0) > 0$. Furthermore U satisfies the homogeneous equation

$$\frac{\sigma^2}{2}U''(x) - (\beta\mu + \theta x)U'(x) - (\theta + \gamma)U(x) = 0, \text{ for } x > 0\ .$$

Since $U(0) > 0$, $U'(0) > 0$ and $\lim_{x\to\infty} U(x) = 0$, the function U must have a positive local maximum at a point $\zeta > 0$. Then $U'(\zeta) = 0$, and by the above differential equation, we get $U''(\zeta) > 0$. This is a contradiction, hence L cannot be finite. Consequently, $\lim_{x\to\infty} G_k(x) = \infty$, which completes part (ii) of the lemma.

Next, consider the case $0 < k < k_0$. For each k, we intend to show that the function G_k has a unique maximum on the interval $[0, \infty)$. Suppose that $G_k'(x)$ is nonzero for all $x > 0$. Then $G_k'(x) > 0$, for all $x > 0$, since $G_k'(0) > 0$. Thus G_k is increasing. But $G_k(x) \le G_{k_0}(x) \le p_0$, from which it follows that $\lim_{x\to\infty} G_k(y)$ exists and is finite. Now, following an argument very similar to the proof of part (ii) above, we can obtain a contradiction. Thus, $G_k'(z_k) = 0$, for some $z_k > 0$, and by (3.24), it is a strict local maximum. By part (ii) of Lemma 3.3, there are no local minima, hence $x = z_k$ is the unique positive maximum point. In particular, the inequality $G_k(z_k) < G_{k_0}(z_k) < p_0$ holds. This completes part (iii).

For part (iv), we introduce the function

$$M(k) = \max_{x\ge 0} G_k(x) \ ,$$

for each $k < k_0$. Then $M(k) < p_0$, and using the properties of the functions $\{G_k : 0 \le k < k_0\}$ that have been so far obtained, it follows that the function M is strictly increasing and continuous on the interval $(0, k_0)$. Also, $M(0+) = \frac{-b}{\gamma+\mu}$ and $\lim_{k\to k_0} M(k) = p_0$. Hence, part (iv) follows.

The proof of part (v) is a direct consequence of the above results and part (iii) of Lemma 3.3. This completes our proof of the lemma.

□

By Lemma 3.4, there is a unique constant $k_p > 0$ such that $\max_{x\ge 0} G_{k_p}(x) = p$, for each p. Furthermore, by parts (iv) and (v) of the same lemma, there is a unique point $z_{k_p} > 0$ such that

$$G_{k_p}(z_{k_p}) = p \ .$$

This point $z_{k_p} > 0$ is the threshold used in our optimal strategy, when $0 < p < p_0$. Therefore, we relabel z_{k_p} as z_p, for simplicity. The following proposition describes the behavior of this threshold point z_p, as a function of p.

Proposition 3.5. *Let z_p be the above described threshold point, for $0 < p < p_0$. Consider z_p as a function of p only (while keeping all the other parameters fixed). Then the following prevails:*

(i) When $0 < p < p_0$, z_p is a continuous, strictly increasing function of the variable p.

(ii) $\lim_{p\to p_0^-} z_p = +\infty$.

Proof. By part (v) of Lemma 3.4, it follows that z_p is strictly increasing on the interval $(0, \infty)$. To prove the continuity of the function z_p, let $0 < p < p_0$ and keep p fixed. First we establish the left-continuity at p. We choose a strictly increasing sequence (p_n) of positive numbers such that $\lim_{n\to\infty} p_n = p$. Then, the sequence (z_{p_n}) is also strictly increasing and bounded above by z_p. According to Lemma 3.3 and Lemma 3.4, there is also a corresponding strictly increasing sequence (k_{p_n}) such that $0 < k_{p_n} < k_p < k_0$, for

all n and $G_{k_{p_n}}(z_{p_n}) = p_n$. We let $\Lambda = \lim_{n\to\infty} z_{p_n}$ and $\hat{k} = \lim_{n\to\infty} k_{p_n}$, and then $0 < \Lambda \leq z_p$ and $0 < \hat{k} \leq k_p$. By part (i) of Lemma 3.3, it follows that $G_{\hat{k}}(\Lambda) = p$. Using parts (iii), (iv) and (v) of Lemma 3.4, we have $\hat{k} = k_p$ and $\Lambda = z_p$. This concludes the left-continuity at point p.

To establish the right-continuity at p, let (p_n) be a strictly decreasing sequence such that $p < p_n < p_0$, for all n and $\lim_{n\to\infty} p_n = p$. Similar to the proof of left-continuity, there are two strictly decreasing sequences (k_{p_n}) and (z_{p_n}) such that $0 < k_p < k_{p_n} < k_0$, for all n, (z_{p_n}) is bounded below by z_p, and $G_{k_{p_n}}(z_{p_n}) = p_n$. We let $\hat{k} = \lim_{n\to\infty} k_{p_n}$ and $\Lambda = \lim_{n\to\infty} z_{p_n}$. Using Lemmas 3.3 and 3.4, we have $p_n = G_{k_{p_n}}(z_{p_n}) \geq G_{k_{p_n}}(x) \geq G_{\hat{k}}(x)$ for all x. We let n tend to infinity and obtain $p \geq G_{\hat{k}}(x)$, for all x. This implies that $\hat{k} \leq k_p$. Then a straightforward argument implies that $G_{\hat{k}}(\Lambda) = p$ and, by Lemma 3.4, it follows that $\Lambda = z_p$. This yields the right-continuity of z_p and hence the proof of part (i) is complete.

Since z_p is strictly increasing in p, we let $\Lambda = \lim_{p\to p_0} z_p$. We suppose that Λ is finite. As before, there is a corresponding strictly increasing function k_p such that $0 < k_p < k_0$ and $G_{k_p}(z_p) = p$. We let $\hat{k} = \lim_{p\to p_0} k_p$. By letting p tend to p_0, we obtain $\hat{k} \leq k_0$ and thus $G_{k_0}(\Lambda) \geq G_{\hat{k}}(\Lambda) = p_0$. This contradicts part (i) of Lemma 3.4. Hence we conclude that Λ is infinite, which completes our proof.

□

3.4 Numerical Computation of z_p

For a given value of p, such that $0 < p < p_0$, and given the values of the parameters a, b, γ, β, μ, θ, σ, we would like to numerically compute z_p. A general numerical method is developed in [23] for computing free boundary points, such as z_p, for infinite horizon discounted cost minimization problems, when the drift and diffusion coefficients are linear functions. However, their method is not applicable in our situation since their main Assumptions 2 and 3 are violated here, but we can readily develop a numerical scheme that is based on our results. It is somewhat similar to the procedure developed in [11], Section 5, and therefore, we indicate only the basic steps.

We intend to find the value k_p of the parameter k, and the corresponding function G_{k_p} so that it satisfies part (iv) of Lemma 3.4. Note that, by the same lemma, z_p is the unique point satisfying $G_{k_p}(z_p) = p$ and $G'_{k_p}(z_p) = 0$.

We begin with computing the unique solution F_∞ to the initial value problem (3.15) and (3.16). In particular, we compute $F'_\infty(0)$. When $\mu = \gamma$, a closed-form formula for $F'_\infty(0)$ can be obtained—see Remark 2 after the statement of Lemma 3.2. To initialize the iterative procedure, we choose $\hat{k} > 0$ such that $\hat{k}F'_\infty(0) = p + \frac{b}{\gamma+\mu}$. Therefore, by (3.22), $G_{\hat{k}}(0) = p$ and $G'_{\hat{k}}(0) > 0$. If $G'_{\hat{k}}(x) = 0$ for some x, then by Lemmas 3.3 and 3.4, it is the maximum of $G_{\hat{k}}$ and we have $G_{\hat{k}}(x) > p$. Hence $k_p < \hat{k}$. If $G'_{\hat{k}}(x) > 0$ for all x, then again $k_p < \hat{k}$ holds. Thus, $\hat{k}$ is an upper bound for k_p. By part (v) of Lemma 3.3, it is clear that $k_p > 0$.

In the n^{th} step of the iteration, we let $u(n)$ and $l(n)$ be the upper and lower bounds for k_p. We initialize with $u(0) = \hat{k}$ and $l(0) = 0$. In the $(n+1)^{\text{th}}$ step, we assign $k_{n+1} = \frac{1}{2}(u(n+1)+l(n+1))$ and examine the solution $G_{k_{n+1}}$ to the initial value problem described in (3.21) and (3.22). If $G_{k_{n+1}}(x) > p$ for some x, then we update the bounds

$u(n+1)$ and $l(n+1)$ by $u(n+1) = k_{n+1}$ and $l(n+1) = l(n)$. Otherwise, the above-described lemmas guarantee that $G'_{k_{n+1}}(z_{n+1}) = 0$ for some z_{n+1}, and $G_{k_{n+1}}(z_{n+1})$ is the maximum value of $G_{k_{n+1}}$. We can now compute $G_{k_{n+1}}(z_{n+1})$, which is less than or equal to p. If $G_{k_{n+1}}(z_{n+1}) < p$, then we update $u(n+1) = u(n)$ and $l(n+1) = k_{n+1}$. Since $u(n+1) - l(n+1) = \frac{1}{2}(u(n) - l(n))$ and $0 \le l(n) \le l(n+1) \le k_p \le u(n+1) \le u(n) < \hat{k}$, it follows that the two sequences $(u(n))$ and $(l(n))$ converge to the same limit k_p. Therefore, we can stop the iterative procedure at the first n where $|G_{k_n}(z_n) - p| < \epsilon$. Here $\epsilon > 0$ is an a priori given error bound. Then using part (i) of Lemma 3.3, z_p is well approximated by z_n.

The following numerical computations are obtained by using a Matlab program, based on the above algorithm, with an error bound $\epsilon = 10^{-3}$. Typically, this procedure converges in 16–18 iterations, but may take longer if the value of p is very close to p_0. We chose the value of the parameter β to be between -1 and +1, since this is the range of values that arises in most QED applications.

μ	γ	β	σ	θ	b	a	p_0	p	$\mathbf{z_p}$
1	0.5	-0.2	2	0.8	3	4	3.0769	3	11.2710
1.5	2	0.3	2	1	5	8	2.6667	2	1.8050
2	2	0	3	1.5	4	7	2	1.5	2.2790
3	1	0.8	4	2	1	8	2.6667	2	3.7940
3	2	0.4	3	1	2	12	4	3	2.3960
3	2	0.4	3	1	2	12	4	3.5	3.7790
4	0.2	-0.8	5	3	6	13	4.0625	1	0.9060
5	8	-0.5	3	6	7	20	1.4286	1.4	3.0820

When p gets close to p_0, the value of z_p becomes large, as one expects from the results of Proposition 3.5. Rows 5 and 6 of the above table also verify the monotonicity of z_p in the parameter p.

3.5 A Smooth Solution to the HJB-Equation

With all the necessary technical results in hand, we are now able to exhibit a twice continuously differentiable solution to the HJB-equation (3.7), which satisfies also (3.8) as well as the growth condition $|F(x)| \le C(1 + ax^+ + bx^-)$, for all x in our verification lemma (Lemma 3.1). This solution depends on p and it has different qualitative behavior according to $p < p_0$ vs. $p \ge p_0$, where p_0 is given in (3.23). In both cases, we show that this solution coincides with the value function of the optimal control problem.

First we consider the case $0 < p < p_0$. Let z_p be as in Proposition 3.5, and let k_p be the constant real number that satisfies $0 < k_p < k_0$ and $G_{k_p}(z_p) = p$, as in Lemma 3.4. When $0 < p < p_0$, we can now introduce our candidate solution F_p on $(-\infty, \infty)$ by

$$F_p(x) = \begin{cases} k_p F_\infty(x) + \frac{b}{(\gamma+\mu)}(\frac{\beta\mu}{\gamma} - x) & \text{if } x \le 0, \\ k_p + \frac{b\beta\mu}{\gamma(\gamma+\mu)} + \int_0^x G_{k_p}(u)du & \text{if } 0 \le x \le z_p, \\ F_p(z_p) + p\,(x - z_p) & \text{if } x \ge z_p, \end{cases} \tag{3.26}$$

where the function F_∞ is characterized in Lemma 3.2.

Next we consider the case $p \geq p_0$. Our candidate solution F_{p_0} remains the same for all the values of $p \geq p_0$ and is given by

$$F_{p_0}(x) = \begin{cases} k_0 F_\infty(x) + \frac{b}{\gamma+\mu}(\frac{\beta\mu}{\gamma} - x) & \text{if } x \leq 0, \\ k_0 + \frac{b\beta\mu}{\gamma(\gamma+\mu)} + \int_0^x G_{k_0}(u)du & \text{if } x \geq 0, \end{cases} \tag{3.27}$$

where the constant $k_0 > 0$ and the function G_{k_0} are described in Lemma 3.4. The following theorem establishes that our candidate functions are smooth solutions to the HJB-equation. Each of these functions provides a lower bound for the corresponding value function V_p.

Theorem 3.6. *Let p_0 be defined by (3.23). Then the following results hold:*

(i) For $0 \leq p < p_0$, the function F_p defined in (3.26) is a twice continuously differentiable solution of the HJB-equation (3.7). Furthermore, F_p is a convex function that also satisfies (3.8), and $F_p(x) \leq V_p(x)$, for all x, where V_p is the value function defined in (3.5).

(ii) For $p \geq p_0$, the function F_{p_0} defined by (3.27) is a twice continuously differentiable solution to the HJB-equation (3.7) for each such p. Furthermore, F_{p_0} is a convex function that also satisfies (3.8), and $F_{p_0}(x) \leq V_p(x)$, for all x and every $p \geq p_0$, where V_p is the value function given in (3.5).

Proof. In both parts of the theorem, it is straightforward to check that the function F_p is twice continuously differentiable on $\mathbf{R}$. When $p < p_0$, by Lemma 3.4, $G_{k_p}(z_p) = p$ and $G'_{k_p}(z_p) = 0$. This yields the C^2 property at the point $x = z_p$.

Next, we intend to verify the HJB-equation (3.7) for the case $0 \leq p < p_0$. From our construction of F_p, it follows that $\mathcal{G}F_p(x) + ax^+ + bx^- = 0$, whenever $x \leq z_p$, where the differential operator $\mathcal{G}$ is defined in (3.6). At the point $x = z_p$, this simplifies to $(\beta\mu + \theta z_p)p + \gamma F_p(z_p) = az_p$. When $x > z_p$,

$$\begin{aligned} \mathcal{G}F(x) + ax^+ + bx^- &= -(\beta\mu + \theta x)p - \gamma F_p(x) + ax \\ &= -(\theta + \gamma)(x - z_p)(p - p_0) > 0 \ , \end{aligned}$$

where p_0 is given in (3.23). To obtain the last equality above, we have used $(\beta\mu + \theta z_p)p + \gamma F_p(z_p) = az_p$, and the expression for $F_p(x)$ given in (3.26), for the case $x > z_p$. Consequently, we have

$$\mathcal{G}F(x) + ax^+ + bx^- \begin{cases} = 0 & \text{if } x \leq z_p \ , \\ > 0 & \text{if } x > z_p \ . \end{cases} \tag{3.28}$$

Next, we obtain the bound $F'_p(x) \leq p$, for all x. By Lemmas 3.2, 3.4 and equation (3.26), it follows that

$$F''_p(x) = \begin{cases} k_p F''_\infty(x) & \text{if } x \leq 0, \\ G'_{k_p}(x) & \text{if } 0 < x \leq z_p. \\ 0 & \text{if } x \geq z_p. \end{cases}$$

Therefore, F''_p is non-negative. Hence, F_p is a convex function on $\mathbf{R}$—indeed, it is strictly convex on the interval $(-\infty, z_p)$ and linear on $[z_p, \infty)$. Consequently,

$$F'_p(x) = \begin{cases} < F'_p(p) = p & \text{if } x < z_p \ , \\ > 0 & \text{if } x \geq z_p \ . \end{cases} \tag{3.29}$$

Using (3.28) and (3.30), we conclude that F_p is a C^2-solution to the HJB-equation (3.7), for $0 < p < p_0$. A straightforward argument, using Lemmas 3.2 and 3.4, shows that $|F_p'(x)|$ is bounded and, hence, (3.8) is also satisfied.

Using the fact that the function F_∞ is bounded on $(-\infty, 0]$, as shown in Lemma 3.2 and (3.26), we also obtain the bound $|F_p(x)| \leq C(1 + ax^+ + bx^-)$, where $C > 0$ is a generic constant. Consequently, the function F_p satisfies all the assumptions of Lemma 3.1 and hence, we conclude that $F_p(x) \leq V_p(x)$, for all x. This completes the proof of part (i).

In the case $p \geq p_0$, it is easy to check that F_{p_0} defined in (3.27) is a C^2-solution to the differential equation $\mathcal{G}F_{p_0}(x) + ax^+ + bx^- = 0$, for all x in $\mathbf{R}$. By an argument very similar to that in part (i), we can also establish that $F_{p_0}''(x) > 0$, for all x, and hence it is a convex function. Note that, when $x > 0$, $F_{p_0}'(x) = G_{k_0}(x)$ where G_{k_0} is given in Lemma 3.4. Then, using the convexity of the function F_{p_0} and part (i) of Lemma 3.4, we deduce that $F_{p_0}'(x) < p_0 \leq p$, for all x. Therefore, F_{p_0} is a C^2-solution to the HJB-equation (3.7), for every $p \geq p_0$, and it also satisfies (3.28). The proof of the estimate $|F_{p_0}(x)| \leq C(1 + ax^+ + bx^-)$, for all x, directly follows from (3.27). Consequently, F_{p_0} satisfies all the assumptions of Lemma 3.1. We thus conclude that, for each $p \geq p_0$, the inequality $F_{p_0}(x) \leq V_p(x)$ holds for all x. This completes the proof.

□

3.6 An Optimal Strategy

We are now ready to prove the main theorem of the present section.

Theorem 3.7. *For each $p > 0$, let the cost functional J and the value function V_p be defined by (3.3) and (3.5), respectively.*

(i) Let $0 < p < p_0$, and let point z_p be as in Lemma 3.4. Fix the initial point x; then introduce the reflected diffusion process X_p^* which satisfies the equation

$$X_p^*(t) = x + \sigma W(t) - \int_0^t [\beta\mu + h(X_p^*(s))]ds - U_p^*(t) \ , \qquad \text{for } t > 0, \tag{3.30}$$

where W is a one-dimensional Brownian motion, and U_p^* is the local time process of X_p^* at point z_p. (When $x > z_p$, there is a jump to z_p and, in this case, we take $X_p^*(0-) = x$ and $U_p^*(0-) = (x - z_p)$.) Then (X_p^*, U_p^*) is an optimal strategy and $J(X_p^*, U_p^*, p) = V_p(x)$, for each x. Furthermore, $V_p(x) \equiv F_p(x)$ for all x, where F_p is given by (3.26).

(ii) Let $p \geq p_0$ and fix the initial point x. Consider the process $X_{p_0}^*$ defined by

$$X_{p_0}^*(t) = x + \sigma W(t) - \int_0^t [\beta\mu + h(X_{p_0}^*(s))]ds \ , \tag{3.31}$$

for all $t \geq 0$. The control process $U_{p_0}^*$ is identically zero in this case. Then, for every $p \geq p_0$, $(X_{p_0}^*, U_{p_0}^*)$ is an optimal strategy and $J(X_{p_0}^*, U_{p_0}^*, p) = V_{p_0}(x) = V_p(x)$, for all x. Furthermore, $V_p(x) \equiv F_{p_0}(x)$, for all x, where F_{p_0} is defined in (3.27).

In both cases, the value function V_p satisfies (3.7) and (3.8).

Remark. For the singular optimal control described in part (i), the principle of smooth fit ([25]) holds. Part (ii) of the above theorem shows that the optimal strategy remains the same for all $p \geq p_0$.

Proof. First, we consider the case where the initial point $x \leq z_p$. The existence of a unique solution to (3.30), over all $t > 0$, is well known(Chapter IV, [16]).

We begin with obtaining an upper bound for $E[X_p^*(T)]^2$. Introduce the sequence of stopping times (τ_n), with $n > |x|$, by

$$\tau_n = \inf\{t \geq 0 : |X_p^*(t)| > n\}. \tag{3.32}$$

Using Itô's lemma, through a localization procedure with (τ_n), we obtain

$$E[X_p^*(T \wedge \tau_n)^2] + 2E \int_0^{T \wedge \tau_n} X_p^*(s)(h(X_p^*(s)) + \beta\mu)ds \\ + 2z_p E[U_p^*(T \wedge \tau_n)] = x^2 + \sigma^2 E[T \wedge \tau_n] \ .$$

Since μ and θ are non-negative quantities, by the definition of h in (3.2), it follows that $x(h(x) + \beta\mu) \geq \epsilon x^2 + K$, where ϵ is a non-negative constant and K is a (possibly negative) constant. Then we obtain

$$E[X_p^*(T \wedge \tau_n)^2] + 2\epsilon E \int_0^{T \wedge \tau_n} X_p^*(s)^2 ds \\ + 2z_p E[U_p^*(T \wedge \tau_n)] \leq x^2 + C_1 T,$$

where C_1 is a positive constant. Consequently, we get

$$E[X_p^*(T \wedge \tau_n)^2] \leq x^2 + C_1 T. \tag{3.33}$$

Let F_p be given by (3.26). We apply Itô's lemma to $F_p(X_p^*(T))$. Since $F_p'(z_p) = p$, F_p satisfies (3.8) and (3.28), which yields

$$E[F_p(X_p^*(T))]e^{-\gamma T} = F_p(x) - E \int_0^T e^{-\gamma t}[(aX_p^*(t)^+ + bX_p^*(t)^-)dt + p\, dU_p^*(t)] \ .$$

Using (3.8) again and then (3.33), we have

$$E[|F_p(X_p^*(T))|]e^{-\gamma T} \leq [C_0 + C_2 E|X_p^*(T)|]e^{-\gamma T} \leq [C_0 + C_2\sqrt{x^2 + C_1 T}]e^{-\gamma T},$$

from which we conclude that $\lim_{T\to\infty} E[F_p(X_p^*(T))]e^{-\gamma T} = 0$. Consequently,

$$F_p(x) = E \int_0^T e^{-\gamma t}[(aX_p^*(t)^+ + bX_p^*(t)^-)dt + p\, dU_p^*(t)] \ .$$

Since F_p is finite, (X_p^*, U_p^*) satisfies the requirements for an admissible control policy, with any initial point $x \leq z_p$. This enables us to conclude that $F_p(x) \geq V_p(x)$, for any $x \leq z_p$.

When $x > z_p$, there is an initial jump to z_p, which we represent by taking $X_p^*(0) = x$, $X_p^*(0+) = z_p$, and $U_p^*(0) = (x - z_p)$. Therefore,

$$F_p(z_p) + p(x - z_p) = E \int_0^T e^{-\gamma t}[(aX_p^*(t)^+ + bX_p^*(t)^-)dt + p\, dU_p^*(t)] \ .$$

By (3.26), $F_p(x) = F_p(z_p) + p(x - z_p)$, when $x > z_p$. Hence,

$$F_p(x) = E \int_0^T e^{-\gamma t}[(aX_p^*(t)^+ + bX_p^*(t)^-)dt + dU_p^*(t)]$$

holds when $x > z_p$, and we conclude that $F_p(x) \geq V_p(x)$, for any $x \geq z_p$ as above. Now, we can use part (i) of Theorem 3.6 to get $F_p(x) = V_p(x)$, for all x. Therefore, for any initial point x, the above-described strategy (X_p^*, U_p^*) is an optimal strategy.

The proof of part (ii) is very similar. Let x be any initial point and consider the diffusion process $X_{p_0}^*$ given in (3.31). Let F_{p_0} be the function described in (3.27). Then, as in the proof of Theorem 3.6, this function satisfies $\mathcal{G}F_{p_0}(x) + ax^+ + bx^- = 0$, and $F'_{p_0}(x) < p_0 \leq p$, for all x. We apply Itô's lemma to $F_{p_0}(X_{p_0}^*(t))e^{-\gamma t}$ and use the above results and (3.8) to deduce

$$E[F_{p_0}(X_{p_0}^*(T))]e^{-\gamma T} = F_{p_0}(x) - E \int_0^T e^{-\gamma t}[aX_{p_0}^*(t)^+ + bX_{p_0}^*(t)^-]dt \ .$$

Next (with the control process $U_{p_0}^*$ identically zero), we can derive the estimate (3.33) by a similar argument. Then, since F_{p_0} also satisfies (3.8), we obtain

$$\lim_{T\to\infty} E[F_{p_0}(X_{p_0}^*(T))]e^{-\gamma T} = 0 \ ,$$

similarly to the proof in part (i). Consequently, we have

$$F_{p_0}(x) = E \int_0^T e^{-\gamma t}[aX_{p_0}^*(t)^+ + bX_{p_0}^*(t)^-]dt \ ;$$

$F_{p_0}(x)$ is finite, hence $(X_{p_0}^*, U_{p_0}^*)$ is an admissible strategy, for every $p \geq p_0$, which yields $F_{p_0}(x) \geq V_p(x)$, for all x and all $p \geq p_0$. By part (ii) of Theorem 3.6, we then conclude that, for a given initial point x, $F_{p_0}(x) = V_p(x)$, and $(X_{p_0}^*, U_{p_0}^*)$ is an optimal strategy for all $p \geq p_0$. This completes the proof of the theorem.

□

Remark. The last theorem enables one to uniquely characterize the threshold z_p by

$$z_p = \begin{cases} \inf\{x \in \mathbf{R} : V'_p(x) = p\} & \text{if} \quad 0 < p < p_0, \\ \infty & \text{if} \quad p \geq p_0. \end{cases} \tag{3.34}$$

Here $p_0 = a/(\theta + \gamma)$, as defined in (3.23).

4 Asymptotic Optimality

We now exhibit an asymptotically optimal sequence of processes (X_n, V_n), each characterized via a constant queue-capacity m_n (with $m_n \equiv \infty$ allowed).

4.1 Main Results

Let (X_n, V_n) be the queueing model in (2.20) with constant queue-capacity m_n. Under the assumptions (2.1)–(2.3), (2.41) and (2.42), we can employ Theorem 2.2 to conclude that the sequence of processes (X_n, V_n) converges weakly to a reflected diffusion process

(X_x, U) which is characterized by (2.43). The constant reflection barrier κ is as in (2.42). Note that the Brownian motions W_1 and W_2 in (2.43) are independent and, therefore, one can rewrite (2.43) in the form

$$X_x(t) = x + \sigma W(t) - \beta\mu t - \int_0^t (\theta X_x^+(s) - \mu X_x^-(s))ds - U(t) \ ; \tag{4.1}$$

if $x \leq \kappa$, then $X_x(t) \leq \kappa$, for all $t \geq 0$, where the constant $\kappa > 0$ is given in (2.42). The process U is non-decreasing; it increases only when $X_x(t) = \kappa$ for $t > 0$, $U(0) = 0$, and it satisfies (2.44). (For model completeness, when $x > \kappa$, the state process X_x has an initial jump to the point κ; thereafter, it satisfies (4.1), starting from κ. To represent this, we simply assign $X_x(0-) = x$, $X_x(0) = \kappa$ and $U(0) - U(0-) = x - \kappa > 0$.) Therefore, in general, the following identity holds for all $t \geq 0$:

$$U(t) - U(0) = \int_0^t I_{[X_x(s)=\kappa]}(s)dU(s) \ . \tag{4.2}$$

Next, the diffusion coefficient σ in (4.1) is determined by the coefficients of the independent Brownian motions W_1 and W_2 in (2.43):

$$\sigma^2 = \sigma_1^2 + \mu \ , \tag{4.3}$$

where σ_1 is given in (2.3), and μ is the service rate.

We are now ready to state the two main theorems of this paper. The first theorem shows that the value function of the diffusion control problem given in (3.5), is a lower bound for the sequence of cost functionals $J(X_n, V_n, p)$, when $\lim_{n\to\infty} X_n(0) = x$. The second theorem exhibits an asymptotically optimal sequence of processes (X_n^*, V_n^*), with $\lim_{n\to\infty} X_n^*(0) = x$, for each x.

Theorem 4.1. *[Asymptotic lower bound]*
Assume that the basic assumptions (2.1)–(2.4) and (2.8)–(2.10) hold. For each $p > 0$, let the value function V_p of the diffusion control problem be given by (3.5). For each x in $\mathbf{R}$, let (X_n, V_n) be a sequence of processes that satisfies (2.20), (2.41) where, for each n, the associated cost functional $J(X_n, V_n, p)$ is given in (2.60). Then the following lower bound holds:

$$\liminf_{n\to\infty} J(X_n, V_n, p) \geq V_p(x) \ , \quad \textit{for all } \ x \in \mathbf{R} \ . \tag{4.4}$$

In the next theorem, we use (4.4) jointly with Theorem 2.2, to obtain an asymptotically optimal strategy.

Theorem 4.2. *[Asymptotic optimality]*
Recall p_0, as defined in (3.23), and z_p, the optimal threshold of the diffusion control problem, as given in Theorem 3.7.

(i) Let $0 < p < p_0$. Consider any sequence (X_n^*, V_n^*), with $\lim_{n\to\infty} X_n^*(0) = x$, equipped with a constant queue-capacity sequence (m_n^*) that satisfies

$$\lim_{n\to\infty} \frac{m_n^*}{\sqrt{n}} = z_p \ . \tag{4.5}$$

Then

$$\liminf_{n\to\infty} J(X_n^*, V_n^*, p) = V_p(x) \ , \tag{4.6}$$

for all x, and therefore, the sequence (X_n^*, V_n^*) is asymptotically optimal.

(ii) Let $p \geq p_0$. Consider a sequence (X_n^*, V_n^*) with no blocking ($m_n = \infty$, for all n) and $\lim_{n\to\infty} X_n^*(0) = x$. Then

$$\liminf_{n\to\infty} J(X_n^*, V_n^*, p) = V_p(x) \ , \tag{4.7}$$

for all x; hence the sequence (X_n^*, V_n^*) is asymptotically optimal.

The proofs of the last two theorems will be demonstrated in several steps. In the following discussion, the initial values $X_n(0)$ are deterministic according to our model in Section 2, and they converge to $X_x(0)$. The same conclusions can be made under the assumption $\lim_{n\to\infty} E|X_n(0) - X_x(0)|^2 = 0$, with $X_n(0)$ being random. Throughout the discussion below, the function $\kappa_n(\cdot)$ represents the normalized queue-capacity process $\frac{m_n(\cdot)}{\sqrt{n}}$, for all $n \geq 1$. The process $\kappa_n(\cdot)$ is in $\mathbf{D}[0,\infty)$, representing the reflection barrier of the process X_n.

Lemma 4.3. *Let all the assumptions of Theorem 4.1 hold. Assume also that $\lim_{n\to\infty} \kappa_n(0) = +\infty$, where $\kappa_n(\cdot)$ is the queue-capacity process corresponding to (X_n, V_n). Then*

$$\liminf_{n\to\infty} J(X_n, V_n, p) \geq V_p(x) \ , \quad \textit{for all } \ x \in \mathbf{R} \ .$$

Proof. It suffices to consider the case $\liminf_{n\to\infty} J(X_n, V_n, p) < \infty$. Without loss of generality (by choosing a subsequence, if necessary), we assume that $\lim_{n\to\infty} J(X_n, V_n, p)$ exists. Recall that $\kappa_n(0)$ is non-random. If $\kappa_n(0) = \infty$ for infinitely many n, then by assumption (2.9), it follows that $\kappa_n(t)$ is infinite for all $t \geq 0$ a.s. for each such n, and there is no reflection boundary. In that case, the result follows from the work of [29], and the proof given below can also be greatly simplified. Using (2.20) and (2.45), we write

$$\begin{aligned} X_n(t) = X_n(0) + \hat{A}_n(t) - M_n(t) - \epsilon_n(t) + \frac{(\lambda_n - \mu n)t}{\sqrt{n}} \\ - \int_0^t h(X_n(s))ds - V_n(t) \end{aligned} \ , \tag{4.8}$$

where $h(x) = \theta x^+ - \mu x^-$, for all x, ϵ_n is given in (2.45) and the non-decreasing process V_n satisfies

$$\int_0^t I_{[X_n(s) < \kappa_n(s)]}(s) dV_n(s) = 0 \ , \tag{4.9}$$

for all $t \geq 0$(see section 7.3 of [29]). By (2.3), (2.26), and following the arguments in [29], as well as using $\lim_{n\to\infty} ||\epsilon_n||_T = 0$ in probability (as in the proof of Theorem 2.2), we have

$$\hat{A}_n - M_n - \epsilon_n \Rightarrow \sigma W, \quad \text{in} \quad \mathbf{D}[0,\infty).$$

Here W is a standard Brownian motion with the constant σ given in (4.3). This weak convergence may prevail through a subsequence, but we can relabel it as the original sequence. Using Skorokhod's representation theorem, we can simply assume that $\hat{A}_n - M_n - \epsilon_n$ converges almost surely to σW, in the space $\mathbf{D}[0,\infty)$ equipped with Skorokhod topology. But the limiting process σW has continuous paths and, therefore, this convergence is uniform on finite intervals, almost surely. We let

$$\xi_n(t) = \hat{A}_n(t) - M_n(t) \ , \tag{4.10}$$

for all $t \geq 0$. Then, we have

$$\lim_{n\to\infty} ||\xi_n - \epsilon_n - \sigma W||_T = 0 \ \text{ almost surely,} \tag{4.11}$$

for all $T > 0$. Next, on the same probability space, and using the same Brownian motion W, we consider the strong solution X_x that satisfies

$$X_x(t) = x + \sigma W(t) - \mu\beta t - \int_0^t h(X_x(s))ds \ , \tag{4.12}$$

for each $t \geq 0$. Here, the function h is given by $h(x) = \theta x^+ - \mu x^-$, for all x. Since h is Lipschitz continuous, the above equation has a path-wise unique, non-exploding strong solution X_x, and the process X_x is adapted to the Brownian filtration. In particular, since h is a Lipschitz continuous function, we can have the standard exponential bound

$$E[||X_x||_T^2] \leq Ce^{kT} \ , \tag{4.13}$$

available for each $T > 0$, where C and k are positive constants (see Chapter 5. [17]). From part (ii) of Theorem 3.7, it is evident that $(X_x, 0)$ is an admissible control system for the diffusion control problem in Section 3, where 0 represents the zero process. Consequently, the cost functional $J(X_x, 0, p)$ is finite for each $x \geq 0$.

Let $\epsilon > 0$ be arbitrary. Then we can find $T_0 > 0$ such that

$$E\int_{T_0}^{\infty} e^{-\gamma s} C(X_x(s))ds < \epsilon \ , \tag{4.14}$$

where $C(x) = ax^+ + bx^-$ represents the running cost function. Let Γ_{κ_n} be the Skorokhod map defined in (2.33), corresponding to the upper-reflection boundary process κ_n. Introduce the process Z_n by $Z_n(t) = X_n(t) + V_n(t)$, for all $t \geq 0$, where (X_n, V_n) satisfies (4.8) and (4.9). Then $\Gamma_{\kappa_n}(Z_n)(t) = X_n(t)$, for all $t \geq 0$, and (4.8) can be written as

$$Z_n(t) = x_n + \xi_n(t) - \epsilon_n(t) - \mu\beta_n t - \int_0^t h(\Gamma_{\kappa_n}(Z_n)(s))ds \ , \tag{4.15}$$

for all $t \geq 0$ where $\beta_n = \frac{(\mu n - \lambda_n)}{\sqrt{n}}$ and $X_n(0) = x_n$. Next, we can also find $n_0 > 1$ such that $|x_n - x| + \mu|\beta_n - \beta|T_0 < \epsilon$, for all $n \geq n_0$. Then, by (4.12) and (4.15), we obtain

$$\begin{aligned} |Z_n(t) - X_x(t)| \leq \epsilon + |\xi_n(t) - \epsilon_n(t) - \sigma W(t)| + C\int_0^t |\Gamma_{\kappa_n}(Z_n)(s) - \Gamma_{\kappa_n}(X_x)(s)|ds \\ + C\int_0^t |\Gamma_{\kappa_n}(X_x)(s) - X_x(s)|ds \ , \end{aligned} \tag{4.16}$$

for each $0 \leq t \leq T_0$ and $n \geq n_0$. Next, we let $\hat{\kappa}_n(\omega) = \inf_{[0,T_0]} \kappa_n(s, \omega)$. Using a $\delta > 0$ in (2.10), and a partition $\{0 = t_0 < t_1 < t_2 < \cdots < t_r = T_0\}$, with $|t_{i+1} - t_i| < \delta$, for all i, and r a finite positive integer, we get

$$\sup_{[0,T_0]} |\kappa_n(t) - \kappa_n(0)| \leq r \sup_{|t-s|<\delta} |\kappa_n(t) - \kappa_n(s)| \ ,$$

for any t and s in $[0,T_0]$. Hence, using (2.10), we have $E[\sup_{[0,T_0]}|\kappa_n(t)-\kappa_n(0)|]\leq C_r$, where C_r is a constant independent of n. In particular, $E[|\hat{\kappa}_n-\kappa_n(0)|]\leq C_r$. Consider the random set

$$B_n=\left\{\omega:\ \hat{\kappa}_n(\omega)>\frac{\kappa_n(0)}{4}\right\}\ .$$

Then, using the above estimate, we have

$$P(B_n^c)\leq P\left[|\hat{\kappa}_n(\omega)-\kappa_n(0)|\geq\frac{3}{4}\kappa_n(0)\right]\leq\ \frac{4}{3\kappa_n(0)}E\left[|\hat{\kappa}_n-\kappa_n(0)|\right]\leq\frac{4C_r}{3\kappa_n(0)}\ .$$

Consequently, $\lim_{n\to\infty}P[B_n^c]=0$. We also introduce the set A_n by

$$A_n=\left\{\sup_{[0,T_0]}|X_x(t)|>\hat{\kappa}_n\right\}\ .$$

Then,

$$P[A_n]\leq P\left[\sup_{[0,T_0]}|X_x(t)|>\hat{\kappa}_n,\hat{\kappa}_n>\frac{\kappa_n(0)}{4}\right]+P[B_n^c].$$

Now,

$$P\left[\sup_{[0,T_0]}|X_x(t)|>\hat{\kappa}_n,\hat{\kappa}_n>\frac{\kappa_n(0)}{4}\right]\leq P\left[\sup_{[0,T_0]}|X_x(t)|>\frac{\kappa_n(0)}{4}\right]\ .$$

Using these facts, together with (4.13), we deduce that $\lim_{n\to\infty}P[A_n]=0$.

From (4.16) we obtain

$$\begin{aligned}\|Z_n-X_x\|_t\leq\epsilon+\|\xi_n-\epsilon_n-\sigma W\|_t+&C_1\int_0^t\|Z_n-X_x\|_s ds\\&+C_2\ T_0\ \|X_x\|_{T_0}.I_{A_n}\ ,\end{aligned}\tag{4.17}$$

for all $0\leq t\leq T_0$, where C_1 and C_2 are generic constants. Here we have used (2.35), plus the fact that $\Gamma_{\kappa_n}(X_x)(s)=X_x(s)$ for all $0\leq s\leq T_0$ on the set A_n^c. Using (2.45) and Lemma 2.1, we can easily derive $E[||\epsilon_n||_T^2]\leq C(1+T^k)$, where $C>0$ and $k>1$ are constants independent of n. This, together with the estimates (2.4), (2.25), and the finiteness of $E[\sup_{[0,T_0]}|W(t)|^2]$ implies that $\sup_{n\geq1}E[||\xi_n-\epsilon_n-\sigma W||_{T_0}^2<\infty$. Using this, together with (4.11), we have $\lim_{n\to\infty}E[\|\xi_n-\sigma W\|_{T_0}]=0$. Also, by (4.13) and $\lim_{n\to\infty}P[A_n]=0$, we obtain $\lim_{n\to\infty}E[\|X_x\|_{T_0}.I_{A_n}]=0$. Now, taking the expected value of (4.17), and using these facts with Gronwall's inequality, we conclude that $\lim_{n\to\infty}E[\|Z_n-X_x\|_{T_0}]=0$. Recall that $\Gamma_{\kappa_n}(Z_n)=X_n$, and hence we have

$$\begin{aligned}E\left[\|X_n-X_x\|_{T_0}\right]=&E\left[\|\Gamma_{\kappa_n}(Z_n)-X_x\|_{T_0}\right]\\\leq&E\left[\|\Gamma_{\kappa_n}(Z_n)-\Gamma_{\kappa_n}(X_x)\|_{T_0}\right]+E\left[\|\Gamma_{\kappa_n}(X_x)-X_x\|_{T_0}\right]\\\leq&2\ E\left[\|Z_n-X_x\|_{T_0}\right]+E\left[\|\Gamma_{\kappa_n}(X_x)-X_x\|_{T_0}\right]\ .\end{aligned}$$

In the last line, clearly the first term on the right-hand side converges to zero as n tends to infinity. For the second term, on the set A_n^c notice that $\Gamma_{\kappa_n}(X_x)(t)=X_x(t)$,

for all $0 \le t \le T_0$. Hence $\|\Gamma_{\kappa_n}(X_x) - X_x\|_{T_0} \le 3\,\|X_x\|_{T_0} I_{A_n}$ holds almost surely. Thus, $\lim_{n\to\infty} E[\|\Gamma_{\kappa_n}(X_x) - X_x\|_{T_0}] = 0$ and consequently,

$$\lim_{n\to\infty} E[\|X_n - X_x\|_{T_0}] = 0 \ . \tag{4.18}$$

Let us consider the cost functional $J(X_n, V_n, p)$ in (2.60). We let the running cost $C(x) = ax^+ + bx^-$, for all x, and use (4.18) to obtain

$$\lim_{n\to\infty} E \int_0^{T_0} e^{-\gamma t} |C(X_n(t)) - C(X_x(t))| dt = 0 \ .$$

Consequently,

$$\begin{aligned} \liminf_{n\to\infty} J(X_n, V_n, p) &\ge \liminf_{n\to\infty} E \int_0^{T_0} e^{-\gamma t} C(X_n(t)) dt \\ &= E \int_0^{T_0} e^{-\gamma t} C(X_x(t)) dt \ . \end{aligned} \tag{4.19}$$

By (4.14), $E\int_0^{T_0} e^{-\gamma t} C(X_x(t))dt \ge E\int_0^{\infty} e^{-\gamma t} C(X_x(t))dt - \epsilon$. Hence,

$$E \int_0^{T_0} e^{-\gamma t} C(X_x(t)) dt \ge V_p(x) - \epsilon \ ,$$

where V_p is the value function given in (3.5). Since $\epsilon > 0$ is arbitrary, it follows that $\liminf_{n\to\infty} J(X_n, V_n, p) \ge V_p(x)$. This completes the proof. □

To prove the above lemma in the case $\limsup_{n\to\infty} \kappa_n(0)$ is finite, we first need the following technical result.

Lemma 4.4. *Let κ be a non-negative adapted process, defined over $[0, \infty)$, with continuous sample paths. Let (X_x, V) be a solution to (2.39) and (2.40), with $b(x) = -[\mu\beta + (\theta x^+ - \mu x^-)]$, for all x, and let κ be the upper reflection boundary process. Then, for a given $\epsilon > 0$, there exists $T_0 > 0$ such that*

$$E \int_{T_0}^{\infty} e^{-\gamma t} [C(X_x(t))dt + p \ dV(t)] < \epsilon \ , \tag{4.20}$$

where $C(x) = ax^+ + bx^-$ is the running cost function. Furthermore, T_0 can be chosen so that it does not depend on the reflection boundary process κ.

Proof. Let (X_x, V) be a solution to (2.39) and (2.40), with $b(x) = -[\mu\beta + (\theta x^+ - \mu x^-)]$, for all x, and a continuous reflection boundary process κ. Then by the discussion below (2.40), it follows that the processes X_x and V have continuous sample paths. We apply Itô's lemma via a localization procedure to $X_x(t)^2$, and use the fact that $\int_0^t X_x(s) dV(s) \ge 0$, to obtain

$$E[X_x(t)^2] \le x^2 + \sigma^2 t - 2E \int_0^t X_x(s)(\mu\beta + h(X_x(s))ds \ ,$$

for all $t > 0$, where $h(x) = \theta x^+ - \mu x^-$, for all x. Since μ and β are constants, we can find a large constant $M > 0$ such that $x(\mu\beta + h(x)) > 0$, for all $|x| > M$. The value of M depends only on μ, β and θ. Therefore,

$$\begin{aligned} E[X_x(t)^2] &\leq x^2 + \sigma^2 t - 2E \int_0^t X_x(s)(\mu\beta + h(X_x(s))I_{[-M,M]}(X_x(s))ds \\ &\leq x^2 + C_0 t \ , \end{aligned} \tag{4.21}$$

for all $t > 0$, where C_0 is a positive constant which depends only on the constants μ, θ, β and σ. Note that C_0 is independent of t and the reflection boundary process κ. Next, using (2.39), we can write $V(t) = x + \sigma W(t) - \mu\beta t - \int_0^t h(X_x(s))ds - X_x(t)$, for all $t \geq 0$. Using the inequality $|x| \leq 1 + x^2$, we obtain the estimate

$$E[V(t)] \leq 1 + +C_1\, t + C_2 \int_0^t (1 + x^2 + C_0\, s)ds + (1 + x^2 + C_0\, t) \ .$$

Again, the constants C_0, C_1 and C_2 are independent of the process κ and t. Consequently,

$$E[V(t)] \leq (1 + x^2)(1 + t) + K_1\, t + K_2\, t^2 \ , \tag{4.22}$$

where the positive constants K_1 and K_2 depend only on the constants μ, θ, β and σ, and they do not depend on the reflection boundary process κ and t. Consider the cost functional $J(X_x, V, p)$ associated with (X_x, V), as given in (3.3). Using Fubini's theorem to obtain the representation (3.9) for $J(X_x, V, p)$ and employing the above estimates (4.21) and (4.22), we get

$$J(X_x, U, p) = E \int_0^\infty e^{-\gamma t}[aX_x^+(t) + bX_x^-(t) + \gamma\, p\, V(t)]dt < \infty \ . \tag{4.23}$$

In view of (4.21) and (4.22), we have $E[aX_x^+(t) + bX_x^-(t) + \gamma\, p\, V(t)] \leq K(1 + t^2)$, where $K > 0$ is a generic constant independent of the process κ and t. This constant may depend on the initial value x. Using Fubini's theorem,

$$E \int_T^\infty e^{-\gamma t}[C(X_x(t))dt + p\ \ dV(t)] = \int_T^\infty e^{-\gamma t} E[C(X_x(t))dt + \gamma p\ \ V(t)]dt \ ,$$

for all $T > 0$. Hence, combining the above estimates, we deduce

$$E \int_T^\infty e^{-\gamma t}[C(X_x(t))dt + p\ \ dV(t)] \leq K \int_T^\infty e^{-\gamma t}(1 + t + t^2)dt \ ,$$

for all $T > 0$; here, the generic constant K is independent of the reflection boundary process κ. Now let $\epsilon > 0$ be arbitrary. Then we can find $T_0 > 0$ such that $K \int_{T_0}^\infty e^{-\gamma t}(1 + t + t^2)dt < \epsilon$, and thus $T_0 > 0$ is independent of the reflection boundary process κ. This immediately yields the desired result.

□

The next lemma is needed for completing the proof of Theorem 4.1.

Lemma 4.5. *Let all the assumptions of Theorem 4.1 hold. Also assume that*

$$\limsup_{n\to\infty} \kappa_n(0) < +\infty \ ,$$

where $\kappa_n(\cdot)$ is the queue-capacity process corresponding to (X_n, V_n). Then

$$\liminf_{n\to\infty} J(X_n, V_n, p) \geq V_p(x) \ , \ \ \textit{for all } x \in \mathbf{R} \ .$$

Proof. We first choose a subsequence such that $\lim_{n_k\to\infty} \kappa_{n_k}(0) = \limsup_{n\to\infty} \kappa_n(0) < +\infty$, and for simplicity, identify it with the original sequence (κ_n). It suffices to consider the case $\liminf_{n\to\infty} J(X_n, V_n, p) < \infty$. Without loss of generality (by choosing a subsequence, if necessary), we simply assume that $\lim_{n\to\infty} J(X_n, V_n, p)$ exists and $\lim_{n\to\infty} \kappa_n(0)$ is finite. For a given interval $[0, T]$, using a $\delta > 0$ in (2.10) and a partition $\{0 = t_0 < t_1 < t_2 < \cdots < t_r = T\}$ with $|t_{i+1} - t_i| = \frac{1}{2^m} < \delta$ for all i, and r is a positive integer which is approximately equal to $2^m T$. Then using (2.10), and following a computation similar to the proof of Lemma 4.3, we obtain $\sup_{n\geq 1} E[\sup_{[0,T_0]} |\kappa_n(t) - \kappa_n(0)|] \leq C(1 + T^k)$. Since, $\lim_{n\to\infty} \kappa_n(0)$ is finite, we deduce that

$$\sup_{n\geq 1} E[||\kappa_n||_T] \leq C(1 + T^k) \ , \tag{4.24}$$

for all $T > 0$. Here $C > 0$ and $k \geq 1$ are generic constants, independent of T. By Lemma 2.1, we have the polynomial bounds

$$\sup_{n\geq 1} E[||X_n||_T^2 + ||V_n||_T^2] \ \leq C(1 + T^k), \tag{4.25}$$

for all $T > 0$. Here $C > 0$ and $k \geq 1$ are generic constants independent of T. Let $\epsilon > 0$ be arbitrary. Since the cost function $C(\cdot)$ is of linear growth, the above estimates enable us to find $T_1 > 0$ such that $C\int_{T_1}^{\infty} e^{-\gamma t}(1 + t^k)dt < \epsilon$ and consequently,

$$\sup_{n\geq 1} E\int_{T_1}^{\infty} e^{-\gamma t}[C(X_n(s)) + \gamma p V_n(s)]ds < C\int_{T_1}^{\infty} e^{-\gamma t}(1 + t^k)dt < \epsilon \ , \tag{4.26}$$

where $C(x) = ax^+ + bx^-$, for all x. Hence,

$$\sup_{n\geq 1} |J(X_n, V_n, p) - E\int_0^T e^{-\gamma t}[C(X_n(s)) + \gamma p V_n(s)]ds| < \epsilon, \tag{4.27}$$

for all $T \geq T_1$. Next, we choose $T > \max\{T_0, T_1\}$, with T_0 given in Lemma 4.4, and we restrict our attention to the function space $\mathbf{D}[0, T]$ equipped with Skorokhod topology. Consider the R^2 valued process $Y_n = (\xi_n, \kappa_n)$, for each n. Our aim here is to show that the sequence (Y_n) is relatively compact in $\mathbf{D}[0, T]$, using Corollary 7.4 in Chapter 3 of [7]. Clearly,

$$||Y_n||_T \leq ||\xi_n||_T + ||\kappa_n||_T.$$

We already know that the sequence (ξ_n) converges weakly to σW. These facts, together with (4.24), implies that for a given $\epsilon > 0$, we can find a large $M > 0$ so that

$$\limsup_{n\to\infty} P[\|Y_n\|_T \geq M] < \epsilon \ ,$$

which guarantees that the sequence (Y_n) is stochastically bounded in $\mathbf{D}[0, T]$. Next we introduce the modulus of continuity w'_T in $\mathbf{D}[0, T]$. However, first introduce the function $w(x, A)$ by

$$w(x, A) = \sup_{s,t\in A} |x(s) - x(t)| \ ,$$

for each x in $\mathbf{D}[0, T]$ and $A \subseteq [0, T]$. We now introduce the modulus of continuity by

$$w'_T(x, \delta) = \inf \sup_{1\leq i\leq m} w(x, [t_{i-1}, t_i)) \ ,$$

where the infimum is taken over all partitions $0 \leq t_0 < \cdots < t_m = T$ of $[0,T]$, such that $\min_{0\leq i\leq m}(t_i - t_{i-1}) \geq \delta$. It is straightforward to check that

$$w'_T(x,\delta) = \inf \sup_{1\leq i\leq m} w(x,[t_{i-1},t_i))$$

also holds, where the infimum is taken over all partitions $0 \leq t_0 < \cdots < t_m = T$ of $[0,T]$, such that $\delta \leq t_i - t_{i-1} < 2\delta$, for all $1 \leq i \leq m$.

Now clearly,

$$w'_T(Y_n,\delta) \leq \inf \left(\sup_{1\leq i\leq m} w(\xi_n,[t_{i-1},t_i)) + \sup_{1\leq i\leq m} w(\kappa_n,[t_{i-1},t_i)) \right) ,$$

where the infimum is taken over all partitions $0 \leq t_0 < \cdots < t_m = T$ of $[0,T]$, such that $\delta \leq (t_i - t_{i-1}) < 2\delta$, for all $1 \leq i \leq m$. We can choose a partition $0 \leq s_0 < \cdots < s_m = T$ of $[0,T]$, such that $\delta \leq (s_i - s_{i-1}) < 2\delta$ for all $1 \leq i \leq m$, and $\sup_{1\leq i\leq m} w(\xi_n,[s_{i-1},s_i)) < w'_T(\xi_n,\delta) + \frac{\epsilon}{2}$. In addition,

$$\sup_{1\leq i\leq m} w(\kappa_n,[s_{i-1},s_i)) \leq \sup_{|u-v|<2\delta} |\kappa_n(u) - \kappa_n(v)| ,$$

where u and v are in $[0,T]$. Therefore,

$$w'_T(Y_n,\delta) \leq w'_T(\xi_n,\delta) + \sup_{|u-v|<2\delta} |\kappa_n(u) - \kappa_n(v)| + \frac{\epsilon}{2} .$$

Hence, we have

$$P[w'_T(Y_n,\delta) > \epsilon] \leq P[w'_T(\xi_n,\delta) > \frac{\epsilon}{4}] + \frac{4}{\epsilon} E[\sup_{|u-v|<2\delta} |\kappa_n(u) - \kappa_n(v)|] .$$

Since the sequence (ξ_n) is weakly convergent to σW, there is $\delta > 0$ such that $\limsup_{n\to\infty} P[w'_T(\xi_n,\delta) > \frac{\epsilon}{4}] < \frac{\epsilon}{4}$. We can also obtain the upper bound $\frac{\epsilon}{4}$ for the second term, using (2.10); this implies $\limsup_{n\to\infty} P[w'_T(Y_n,\delta) > \epsilon] < \epsilon$. Thus, the modulus of continuity condition for (Y_n) has been verified. Using Corollary 7.4 in Chapter 3 of [7], we conclude that the sequence (Y_n) is relatively compact in $\mathbf{D}^2[0,T]$. Since, $\limsup_{n\to\infty} ||\epsilon_n||_T = 0$ in probability as in Theorem 2.2, we can conclude that the sequence $(\xi_n - \epsilon_n, \kappa_n)$ is relatively compact in $\mathbf{D}^2[0,T]$. Next, we can choose a subsequence of $(\xi_n - \epsilon_n, \kappa_n)$ which converges weakly to a process (ξ,κ) in $\mathbf{D}^2[0,T]$ and relabel this subsequence by the original sequence. Clearly, the process κ is non-negative. Since we already know that the sequence (ξ_n) is weakly convergent to the process σW in $\mathbf{D}[0,T]$, where σ is a positive constant and W is a standard Brownian motion, we conclude that $\xi = \sigma W$. Let $J(\kappa_n) = \sup_{0\leq t\leq T} |\kappa_n(t) - \kappa_n(t-)|$. Then $J(\kappa_n) \leq \sup_{|t-s|<\delta} |\kappa_n(t) - \kappa_n(s)|$, and as a direct consequence of (2.10), we obtain $\lim_{n\to\infty} E[J(\kappa_n)] = 0$. Hence, using Theorem 10.2 in Chapter 3 of [7], we conclude that the limiting process κ has continuous paths.

Therefore, using the Skorokhod embedding theorem, we can simply assume that the sequence $(\xi_n - \epsilon_n, \kappa_n)$ is almost surely convergent to the process $(\sigma W, \kappa)$, as n

tends to infinity in $\mathbf{D}[0,T]$. Since the limiting process $(\sigma W, \kappa)$ has continuous paths, this convergence is uniform on $[0,T]$ and we have

$$\lim_{n\to\infty} (||\xi_n - \epsilon_n - \sigma W||_T + ||\kappa_n - \kappa||_T) = 0 \ , \tag{4.28}$$

almost surely.

Next, we extend the process κ in (4.28) to $[0,\infty)$ continuously, by simply defining $\kappa(t) = \kappa(T)$ over all $t \geq T$. Then, we can construct the limiting process (X_x, V) on the same probability space so that it is the strong solution to (2.39) and (2.40), where $b(x) = -[\mu\beta + (\theta x^+ - \mu x^-)]$, for all x, and with the continuous reflection boundary process κ that satisfies (4.28). By the discussion below (2.40), it follows that the processes X_x and V have continuous paths and are adapted to the filtration of the Brownian motion. Consequently, (X_x, V) is an admissible process for the diffusion control problem and, by Lemma 4.4, the corresponding cost functional $J(X_x, V, p)$ is finite.

Introduce the process Z by

$$Z(t) = X_x(t) + V(t) \ , \tag{4.29}$$

for all $t \geq 0$. Then $\Gamma_\kappa(Z)(t) = X_x(t)$, for all $t \geq 0$, where Γ_κ is the Skorokhod map given in (2.40). Consider (X_n, V_n) which satisfies (4.25)-(4.27), with the reflection boundary κ_n, $X_n(0) = x_n$ and $\lim_{n\to\infty} x_n = x$. Let $Z_n(t) = X_n(t) + V_n(t)$, for all $t \geq 0$. We use the Lipschitz continuity of $b(x) = -[\mu\beta + (\theta x^+ - \mu x^-)]$, $\lim_{n\to\infty} x_n = x$, $\lim_{n\to\infty} \beta_n = \beta$ and then follow an estimation similar to (4.16) to obtain

$$||Z_n - Z||_t \leq \delta_n + ||\xi_n - \epsilon_n - \sigma W||_t + C \int_0^t |\Gamma_{\kappa_n}(Z_n)(s) - \Gamma_\kappa(Z)(s)| ds \ ,$$

almost surely, where $\lim_{n\to\infty} \delta_n = 0$ and C is the Lipschitz constant. Using (2.34) and (2.37), we obtain the estimate,

$$\begin{aligned}
\int_0^t |\Gamma_{\kappa_n}(Z_n)(s) - \Gamma_\kappa(Z)(s)| ds \leq & \int_0^t |\Gamma_{\kappa_n}(Z_n)(s) - \Gamma_{\kappa_n}(Z)(s)| ds \\
& + \int_0^t |\Gamma_{\kappa_n}(Z)(s) - \Gamma_\kappa(Z)(s)| ds \\
\leq & 2 \int_0^t ||Z_n - Z||_s ds + \int_0^t ||\kappa_n - \kappa||_s ds \ ,
\end{aligned}$$

for each $0 \leq t \leq T$. By combining the above facts, together with $\lim_{n\to\infty} \delta_n = 0$, (4.28) and then using Gronwall's inequality, we obtain

$$\lim_{n\to\infty} ||Z_n - Z||_T = 0 \ \text{ almost surely.} \tag{4.30}$$

On the other hand, using (2.34) and (2.37), we obtain

$$\begin{aligned}
||\Gamma_{\kappa_n}(Z_n) - \Gamma_\kappa(Z)||_T \leq & ||\Gamma_{\kappa_n}(Z_n) - \Gamma_{\kappa_n}(Z)||_T \\
& + ||\Gamma_{\kappa_n}(Z) - \Gamma_\kappa(Z)||_T \\
\leq & 2||Z_n - Z||_T + ||\kappa_n - \kappa||_T \ .
\end{aligned}$$

Hence, using (4.30) and (4.28), we have $\lim_{n\to\infty} ||\Gamma_{\kappa_n}(Z_n) - \Gamma_\kappa(Z)||_T = 0$ a.s. Consequently,

$$\lim_{n\to\infty} ||X_n - X_x||_T = 0, \text{ a.s.} \tag{4.31}$$

This, combined with (4.30), yields

$$\lim_{n\to\infty} ||V_n - V||_T = 0, \text{ a.s.} \tag{4.32}$$

Let $J(X_n, V_n, p)$ be the cost functional in (2.60). Then

$$\begin{aligned}\liminf_{n\to\infty} J(X_n, V_n, p) &\geq \liminf_{n\to\infty} E\int_0^T e^{-\gamma t}[C(X_n(t))dt + \gamma\, p\, V_n(t)]dt \\ &\geq E\int_0^T e^{-\gamma t}[C(X_x(t)) + \gamma\, p\, V(t)]dt\ .\end{aligned}$$

We have used Fatou's lemma to obtain the last inequality. Since (X_x, V) is an admissible control policy for the diffusion control problem in (3.5), with a continuous upper reflection boundary κ, we can use Lemma 4.5 and the fact that $T > T_0$, to conclude that $E\int_0^T e^{-\gamma t}[C(X_x(t)) + \gamma\, p\, V(t)]dt > J(X_x, V, p) - \epsilon \geq V_p(x) - \epsilon$. Hence,

$$\liminf_{n\to\infty} J(X_n, V_n, p) > V_p(x) - \epsilon\ ,$$

where $\epsilon > 0$ is arbitrary. This completes the proof of the lemma.

□

Proof of Theorem 4.1.
The proof clearly follows from Lemmas 4.3 and 4.5.

In the next two results, we prove convergence of the cost functionals, related to constant (finite or infinite) queue-capacities. These two results will then be used to establish Theorem 4.2.

Proposition 4.6. *Let (X_n, V_n) be a state-process equipped with constant queue-capacity m_n, which satisfies the conditions (2.20), (2.41) and (2.42) with a finite constant κ in (2.42). For each integer $n \geq 1$ and $p > 0$, the corresponding cost functional $J(X_n, V_n, p)$ is given by (2.60). Then, $\lim_{n\to\infty} J(X_n, V_n, p) = J(X_x, U, p)$, for each $p > 0$, where $J(X_x, U, p)$ is defined in (3.3).*

Proof. By (2.41) and (2.42), we have $\lim_{n\to\infty} X_n(0) = x$ and $\lim_{n\to\infty} \frac{m_n}{\sqrt{n}} = \kappa > 0$, where κ is a finite constant. Then, by Theorem 2.2, the sequence of processes (X_n, V_n) converges weakly to the process (X_x, U) that satisfies (4.1). To show the convergence of the cost functionals, we use the polynomial growth bounds for $E[||X_n||_T^2]$ and $E[||V_n||_T^2]$ obtained in Lemma 2.1.

Following the application of Fubini's theorem in the derivation of (3.9), we obtain

$$J(X_n, V_n, p) = E\int_0^\infty e^{-\gamma t}[aX_n^+(t) + bX_n^-(t) + p\,\gamma V_n(t)]dt\ . \tag{4.33}$$

Since (X_n, V_n) converges weakly to the process (X_x, U), using Skorokhod's embedding theorem, we may assume that (X_n, V_n) converges almost surely to the process (X_x, U), in some probability space $(\Omega_1, \mathfrak{F}_1, \mathbf{P_1})$. Introduce

$$R_n(t) = aX_n^+(t) + bX_n^-(t) + p\,\gamma V_n(t)\ ,$$

and

$$R_\infty(t) = aX_x^+(t) + bX_x^-(t) + p\,\gamma U(t)\ ,$$

on the space $[0,\infty)\otimes\Omega_1$. Let λ be the finite measure on the collection of Borel subsets B of $[0,\infty)$, defined by $\lambda(B) = \int_B e^{-\gamma t}dt$. Next, consider the product measure $\lambda\otimes P_1$ on the space $[0,\infty)\otimes\Omega_1$, equipped with the product σ-algebra $\mathfrak{B}\otimes\mathfrak{F}_1$. Here $\mathfrak{B}$ is the Borel $\sigma-$algebra on $[0,\infty)$. Then R_n converges to R_∞ almost surely in $\lambda\otimes P_1$, as n tends to infinity. Clearly,

$$0 \le R_n(t) \le C(|X_n(t)| + V_n(t))\ ,$$

where C is a generic constant that does not depend on n and t. Consequently, using Lemma 2.1, we obtain

$$E[||R_n||_T^2] \le C(1+T^k)\ , \tag{4.34}$$

for all $T>0$, where C is a generic constant that does not depend on n and T. By Fatou's lemma, we also have $E[||R_\infty||_T^2] \le C(1+T^k)$. Now, (R_n) converges to R_∞ almost surely with respect to $\lambda\otimes P_1$ and by (4.34), the sequence (R_n) is uniformly integrable on the product space $[0,\infty)\otimes\Omega_1$. Therefore, we conclude that

$$\lim_{n\to\infty} E_{\lambda\otimes P_1}[R_n] = E_{\lambda\otimes P_1}[R_\infty]\ ,$$

with the limit being finite. The proof of the proposition is now complete. □

Next, we treat the case where the queue-capacity $m_n(t)$ is infinite (no blocking) for all $t\ge 0$. In this case, the process V_n in (2.20) is identically zero and, for this reason, we relabel (X_n, V_n) by $(X_n, 0)$. Similarly, for the limiting process, the reflecting barrier κ in (4.1) and (4.2) is infinite. Therefore, the process U in (4.1) is identically zero. Accordingly, we relabel (X_x, U) in (4.1) by $(X_x, 0)$ in the following discussion.

Proposition 4.7. *For each integer $n\ge 1$, consider the process $(X_n,0)$ which satisfies (2.20) and (2.41) with $m_n(t)=\infty$ for all $t\ge 0$. The associated cost functional $J(X_n,0,p)$ is given by (2.60) for each $p>0$. Consider $(X_x,0)$ which satisfies (4.1) with $\kappa=\infty$ and the process U identically zero. Then $\lim_{n\to\infty} J(X_n,0,p) = J(X_x,0,p)$ where $J(X_x,0,p)$ is defined in (3.3).*

Proof. By Theorem 2.2, we know that under the above assumptions, the process X_n converges weakly to the process X_x in the function space $\mathbf{D}[0,\infty)$, as n tends to infinity. By Lemma 2.1, we have the polynomial growth bound

$$\sup_{n\ge 1} E[||X_n||_T^2] \le C(1+T^k)\ , \tag{4.35}$$

where $C>0$ and $k\ge 1$ are generic constants that do not depend on T. The rest of the proof is very similar to that of Proposition 4.6 and therefore, it is omitted. □

Using the last two propositions, we are now able to complete the proof of Theorem 4.2.

Proof of Theorem 4.2.

Proof. To prove part (i), first let $0 < p < p_0$, where p_0 is given in (3.23). We consider a sequence of processes (X_n^*, V_n^*) that satisfies (2.20) and (2.41), with the associated constant queue capacity sequence (m_n^*). We assume that the sequence (m_n) satisfies $\lim_{n\to\infty} \frac{m_n}{\sqrt{n}} = z_p$, where z_p is the optimal threshold point given in Theorem 3.7. Then, by Theorem 2.2, (X_n^*, V_n^*) converges weakly to the reflected diffusion process (X_x^*, U_p^*), which has the upper reflection barrier at point z_p. From Theorem 3.7, (X_x^*, U_p^*) is an optimal strategy for the diffusion control problem in (3.5). Therefore, $J(X_x^*, U_p^*, p) = V_p(x)$, for all x, where V_p is the value function given in (3.5). On the other hand, by Proposition 4.6, we have $\lim_{n\to\infty} J(X_n^*, V_n^*, p) = J(X_x^*, U_p^*, p)$. Consequently, $\lim_{n\to\infty} J(X_n^*, V_n^*, p) = V_p(x)$, and using Theorem 4.1, we conclude that the sequence (X_n^*, V_n^*) is asymptotically optimal. This completes the proof of part (i).

In part (ii), we consider the case $p \geq p_0$. Let (X_n^*, V_n^*) be a sequence that satisfies (2.20) and (2.41) with the associated queue-capacity $m_n(t) \equiv \infty$, for all $t \geq 0$ and all n. Then, V_n^* is identically zero and (X_n^*, V_n^*) converges weakly to $(X_x^*, 0)$, where X_x^* satisfies (4.1), with the process U identically zero. In Theorem 3.7, we have proved that X_x^* is the optimal process for the diffusion control problem, for every $p \geq p_0$. Therefore, $J(X_x^*, 0, p) = V_p(x)$, for all x and all $p \geq p_0$. Combining this with Proposition 4.7, we obtain $\lim_{n\to\infty} J(X_n^*, V_n^*, p) = J(X_x^*, 0, p) = V_p(x)$, for all x and all $p \geq p_0$. Using Theorem 4.1, we conclude that the sequence (X_n^*, V_n^*) is asymptotically optimal, which completes the proof. □

5 Appendix

Proof of Lemma 3.2

Let $x \leq 0$ and consider the diffusion process Y characterized by

$$Y(t) = x + \sigma W(t) - \mu \int_0^t (\beta + Y(s))ds \ , \quad t \geq 0 \ , \tag{5.1}$$

where W is a standard Brownian motion. Next, we introduce the stopping time τ_0 by

$$\tau_0 = \inf\{t \geq 0 : Y(t) = 0\}. \tag{5.2}$$

We intend to show that the function F_∞ of (3.15) and (3.16) has the stochastic representation $F_\infty(x) = E[e^{-\gamma\tau_0}|Y(0) = x]$, for every $x \leq 0$. To this end, introduce

$$\tilde{F}_\infty(x) = E[e^{-\gamma\tau_0}|Y(0) = x] \ , \tag{5.3}$$

for all $x \leq 0$. For each $N \geq |x|$, we also introduce the stopping time

$$\tau_N = \inf\{t \geq 0 : Y(t) = -N\}. \tag{5.4}$$

To construct the function F_∞, we begin with a sequence of functions (F_n). Let F_n be the unique solution to the boundary value problem

$$\frac{\sigma^2}{2} F_n''(x) - (\beta\mu + \mu x)F_n'(x) - \gamma F_n(x) = 0 \ , \quad \text{for } \ x < 0 \ ,$$

$$F_n(0) = 1 \ , \quad \text{and} \ \ F_n(-n) = 0 \ .$$

Then, with the use of Itô's lemma, one can easily verify that

$$F_n(x) = E_x[e^{-\gamma\tau_0} I_{[\tau_0<\tau_n]}] \ , \quad \text{for} \ \ -n \le x \le 0 \ .$$

Using the scale function associated with the diffusion Y, it follows that (see [17], Section 5 of Chapter 5)

$$P[\tau_n < \tau_0 | Y(0) = x] = \frac{\int_x^0 e^{\frac{\mu}{\sigma^2}(y^2+2\beta y)} dy}{\int_{-n}^0 e^{\frac{\mu}{\sigma^2}(y^2+2\beta y)} dy} \ .$$

Therefore, $P[\tau_n < \tau_0 | Y(0) = x]$ is decreasing to zero as n tends to infinity. Consequently, the sequence $(F_n(x))$ is increasing to the function $\tilde{F}_\infty$. We fix the interval $[x, 0]$ and integrate the differential equation for F_n twice on this interval to obtain an integral equation for F_n. Then we use $0 \le F_n(y) \le \tilde{F}_\infty(y) \le 1$, on $[-n, 0]$, and $\lim_{n\to\infty} F_n(y) = \tilde{F}_\infty(y)$, together with the bounded convergence theorem, to conclude that the function $\tilde{F}_\infty$ also satisfies the same integral equation. By differentiating it twice, we observe that $\tilde{F}_\infty$ also satisfies (3.15), together with the boundary condition $\tilde{F}_\infty(0) = 1$.

The stochastic representation (5.3) also implies that $\tilde{F}_\infty$ is increasing and $\tilde{F}'_\infty(x) \ge 0$ on the interval $(-\infty, 0]$. Consequently, $\lim_{x\to-\infty} \tilde{F}_\infty(x) = L_0$ exists, with $0 \le L_0 < 1$. Furthermore, if $\tilde{F}'_\infty(\xi) = 0$ for some $\xi < 0$, then by (3.15), $\tilde{F}''_\infty(\xi) > 0$. Hence, $x = \xi$ is a strict local minimum. This is a contradiction since $\tilde{F}_\infty$ is increasing. Hence $\tilde{F}'_\infty(x) > 0$, for all $x < 0$ and as a consequence, $\tilde{F}_\infty$ is strictly increasing on $(-\infty, 0]$.

Our next step is to prove that $\lim_{x\to-\infty} \tilde{F}_\infty(x) = 0$. We consider the process Z defined by $Z(t) = Y(t) + \beta$, for all $t \ge 0$, where Y is given in (3.17). Then Z is an Ornstein-Uhlenbeck process that satisfies

$$Z(t) = (x + \beta) + \sigma W(t) - \mu \int_0^t Z(s) ds \ , \quad t \ge 0 \ . \tag{5.5}$$

For each $y > x + \beta$, we introduce the stopping time $\tilde{\tau}_y$ by

$$\tilde{\tau}_y = \inf\{t \ge 0 : Z(t) \ge y\} \ . \tag{5.6}$$

Then τ_0, defined in (5.2), is identical to $\tilde{\tau}_\beta$, and $\tilde{F}_\infty(x) = E[e^{-\gamma\tilde{\tau}_\beta} | Z(0) = x + \beta]$, for all $x < \min\{-\beta, 0\}$. Using the strong Markov property, we obtain

$$E[e^{-\gamma\tilde{\tau}_0} | Z(0) = x + \beta] = E[e^{-\gamma\tilde{\tau}_\beta} | Z(0) = x + \beta] E[e^{-\gamma\tilde{\tau}_0} | Z(0) = \beta] \ , \ \text{ if } \ \beta < 0 \ ,$$

and

$$E[e^{-\gamma\tilde{\tau}_\beta} | Z(0) = x + \beta] = E[e^{-\gamma\tilde{\tau}_0} | Z(0) = x + \beta] E[e^{-\gamma\tilde{\tau}_\beta} | Z(0) = 0] \ , \ \text{ if } \ \beta > 0 \ .$$

Therefore, to reach the desired conclusion, it suffices to demonstrate that

$$\lim_{y\to-\infty} E[e^{-\gamma\tilde{\tau}_0} | Z(0) = y] = 0 \ .$$

For this, consider the process Z that is characterized by $Z(t) = y + \sigma W(t) - \mu \int_0^t Z(s) ds$. Then it is well known that, via a random time change, one can write

$$Z(t) = e^{-\mu t}[y + B(\frac{\sigma^2}{2\mu}(e^{2\mu t} - 1))] \ , \quad t \ge 0 \ ,$$

where B is another Brownian motion. Next, we introduce a collection of stopping times (T_y) with respect to this Brownian motion. Specifically, for each $y < 0$, let

$$T_y = \inf\{t \geq 0 : B(t) = -y\}.$$

This enables one to derive a relationship between $\tilde{\tau}_0$ and T_y, namely

$$\tilde{\tau}_0 = \frac{1}{2\mu} \log[(\frac{2\mu}{\sigma^2})T_y + 1] \ .$$

The distribution of the Brownian stopping time T_y is well known (see [3]), and it will help us compute the limit $\lim_{y\to-\infty} E[e^{-\gamma\tilde{\tau}_0}|Z(0) = y]$.

Observe that $e^{-\gamma\tilde{\tau}_0} = [(\frac{2\mu}{\sigma^2})T_y + 1]^{-\frac{\gamma}{2\mu}}$. It follows that

$$\lim_{y\to-\infty} E[e^{-\gamma\tilde{\tau}_0}|Z(0) = y] = \lim_{y\to-\infty} E[\ \frac{1}{[(\frac{2\mu}{\sigma^2})T_y + 1]^{\frac{\gamma}{2\mu}}} \mid B(0) = 0\ ] \ .$$

Using the bounded convergence theorem, we notice that the limit on the right-hand side vanishes since $\lim_{y\to-\infty} T_y = \infty$ almost surely.

Consequently, $\lim_{x\to-\infty} \tilde{F}_\infty(x) = 0$. Now it is clear that $\tilde{F}_\infty$ satisfies (3.15) and (3.16). The uniqueness of solutions to (3.15) and (3.16) can be established by the fact that the difference of two solutions to (3.15) cannot have any positive local maxima. Therefore, $\tilde{F}_\infty$ is identical to F_∞, as characterized via the initial value problem (3.15) and (3.16). It also has the stochastic representation (5.3).

In our next step, we show that $\lim_{x\to-\infty} F'_\infty(x) = 0$ and the function F_∞ is strictly convex. First we extend the function F_∞ to $(-\infty, \infty)$, so that it satisfies the differential equation (3.15) everywhere on $(-\infty, \infty)$. Since F_∞ is strictly increasing on $(-\infty, 0]$, by (3.16) it is clear that $\liminf_{x\to-\infty} F'_\infty(x) = 0$. Thus, we can choose a sequence (y_n) strictly decreasing to $-\infty$, such that $y_{n+1} < y_n < 0$ and $0 < F'_\infty(y_{n+1}) < F'_\infty(y_n)$, for all n. Consequently, there is a point ξ_n such that $y_{n+1} < \xi_n < y_n$ and $F''_\infty(\xi_n) > 0$. Note that the sequence (ξ_n) is also strictly decreasing to $-\infty$.

Let $z = \inf\{x \geq 0 : \ F'_\infty(x) \leq 0\ \}$. If z is finite, then $F'_\infty(z) = 0$ and $F'_\infty(x) > 0$, for all $x < z$, and consequently, $F_\infty(z) \geq F_\infty(0) = 1$. By (3.15), Y_n has paths of bounded variation $F''_\infty(z) > 0$, and hence F_∞ has a strict local minimum at the point $x = z$, which is a contradiction. Therefore, z cannot be finite and $F'_\infty(x) > 0$, for all x in $(-\infty, \infty)$.

Next we consider any point $x_1 > 0$, such that $x_1+\beta > 0$. Then $F_\infty(x_1) > F_\infty(0) = 1$ and $F'_\infty(x_1) > 0$. Using (3.15), we also obtain $F''_\infty(x_1) > 0$. Now introduce the function $H(x) = F''_\infty(x)$, on the interval $[\xi_n, x_1]$. Then, by differentiating (3.15), we see that

$$\frac{\sigma^2}{2}H''(x) - (\beta\mu + \mu x)H'(x) - (\gamma + 2\mu)H(x) = 0 \ , \quad \text{on } \ [\xi_n, x_1] \ ,$$

$H(\xi_n) > 0$ and $H(x_1) > 0$. Let $\xi_n \leq c \leq x_1$ so that $H(c) = \min_{[\xi_n, x_1]} H(x)$. Suppose that $H(c) \leq 0$, then $\xi_n < c < x_1$ and $H'(c) = 0$. If $H(c) = 0$, then by the uniqueness of the solution to the above differential equation, it follows that H is identically zero. If $H(c) < 0$, again by the above differential equation, we have $H''(c) < 0$ and $x = c$ is a strict local maximum; this is a contradiction. Hence $H(c) > 0$ and, consequently,

$F''_\infty(x) > 0$ on the interval $[\xi_n, x_1]$. But one can choose x_1 arbitrarily large and the sequence (ξ_n) is decreasing to $-\infty$. We thus conclude that $F''_\infty(x) > 0$, on the interval $(-\infty, \infty)$. This, together with the fact that $\liminf_{x\to-\infty} F'_\infty(x) = 0$, implies $\lim_{x\to-\infty} F'_\infty(x) = 0$. Hence, F_∞ is a strictly convex function that satisfies all the conclusions of Lemma 3.2. This completes the proof.

Acknowledgements. The research of A.W. has been supported in part by the Army Research Office under grant no. W 911NF0710424. The work of A.M. has been partially supported by BSF Grants 2005175 and 2008480, ISF Grant 1357/08 and by the Technion funds for promotion of research and sponsored research. Some of the research was carried out while A.M. was visiting the Statistics and Applied Mathematical Sciences Institute (SAMSI) of the NSF, and the Department of Statistics and Operations Research (STOR), the University of North Carolina at Chapel Hill - the hospitality of both institutions are gratefully acknowledged. We would also like to thank former graduate student Ju Ming, at the Mathematics Department of Iowa State University, for his help in preparing a Matlab program for the numerical computations in Section 3.4. We are grateful to Lillian Bluestein, of the Faculty of Industrial Engineering and Management, Technion, for her technical assistance in preparation of this manuscript. Finally, we thank the referees for valuable suggestions that has led to a significantly improved manuscript.

References

[1] Atar, R., *Scheduling control for queueing systems with many servers:Asymptotic optimality in heavy traffic* , Ann. Appl. Probab., 15 (2005), pp. 2606-2650.

[2] Atar, R., Mandelbaum, A. and Reiman, M. I., *Scheduling a multi class queue with many exponential servers:Asymptotic optimality in heavy traffic* , Ann. Appl. Probab., 14 (2004), pp. 1084-1134.

[3] Borodin, A. N. and Salminen, P., *Handbook of Brownian Motion-Facts and Formulae*, Second edition, Birkhäuser-Verlag (2002).

[4] Burdzy, K., Kang, W. and Ramanan, K., *The Skorokhod problem in a time-dependent interval*, Stoch. Proc. and their Appl., 119 (2009), pp. 428-452.

[5] Borst, S., Mandelbaum, A. and Reiman, M. I., *Dimensioning large call centers*, Oper. Res., 52 (2004), pp. 17-34.

[6] Dai, J. G. and He, S., *Customer abandonment in many-server queues,* Math. Oper. Res., 35 (2010), pp. 347-362.

[7] Ethier, S. N. and Kurtz, T. G., *Markov Processes: Characterization and Convergence*, Wiley, New York (1986).

[8] Fleming, W. H. and Soner, H. M., *Controlled Markov Processes and Viscosity Solutions,* Second edition, Springer-Verlag, New York (2006).

[9] Gans, N., Koole, G., and Mandelbaum, A., *Telephone call centers: tutorial, review and research prospects*, Manuf. Service Oper. and Mgmt., 5 (2003), pp. 79-141.

[10] Garnett, O., Mandelbaum, A. and Reiman, M. I., *Designing a telephone call center with impatient customers,* Manuf. Service Oper. and Mgmt., 4 (2002), pp. 208-227.

[11] Ghosh, A. P. and Weerasinghe, A. P., *Optimal buffer size for a stochastic processing network in heavy traffic,* Queueing Systems, 55 (2007), pp. 147-159.

[12] Ghosh, A. P. and Weerasinghe, A. P., *Optimal buffer size and dynamic rate control for a queueing network in heavy traffic,* Stoch. Proc. and their Appl. 120 (2010), pp. 2103-2141.

[13] Halfin, S. and Whitt, W., *Heavy traffic limits for queues with many exponential servers*, Oper. Res. 29 (1981), pp. 567-588.

[14] Harrison, J. M., *Brownian Motion and Stochastic Flow Systems*, Wiley Publications, New York (1985).

[15] Hartman, P., *Ordinary Differential Equations*, Wiley Publications, New York (1964).

[16] Ikeda, N. and Watanabe, S., *Stochastic Differential Equations and Diffusion Processes*, North-Holland Publishers, Amsterdam (1981).

[17] Karatzas, I. and Shreve, S.E., *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York (1988).

[18] Khudyakov, P., *Designing a Call Center with an IVR (Interactive Voice Response)*, Technion M.Sc. Thesis (2006). (Downloadable at http://iew3.technion.ac.il/serveng/References/thesis_polyna.pdf)

[19] Khudyakov, P., Feigin, P., and Mandelbaum, A., *Designing a Call Center with an IVR (Interactive Voice Response)*, Queueing Systems, 66 (2010), pp. 215-237.

[20] Kocaga, Y. L. and Ward, A., *Admission control for a multi-server queue with abandonment*, Queueing Systems, 65 (2010), pp. 275-323.

[21] Krichagina, E. V. and Taksar, M., *Diffusion approximation for GI/G/1 controlled queues*, Queueing Systems Theory Appl., 12 (1992), pp. 333-367.

[22] Kruk, L., Lehoczky, J., Ramanan,K. and Shreve, S., *An explicit formula for the Skorokhod map on [0,a]*, Ann. Appl. Probab., (2007), pp. 669-682.

[23] Kumar, S. and Muthuraman, M., *A numerical method for solving singular stochastic control problems*, Oper. Res., 52 (2004), pp. 563-582.

[24] Lee, C. and Weerasinghe, A., *Convergence of a queueing system in heavy traffic with general patience-time distributions,* Stoch. Proc. and their Appl., 121, (2011), pp. 2507-2552.

[25] Ma, J., *On the principle of smooth-fit for a class of singular stochastic control problems for diffusions,* SIAM J. Control and Opt., 30 (1992), pp. 975-999.

[26] Mandelbaum, A. and Momcilovic, P., *Queues with many servers and impatient customers,* Math. Oper. Res., 37 (2012), pp. 41-64.

[27] Massey A. W. and Wallace B. R., *An optimal design of the M/M/C/K queue for call centers*, To appear in QUESTA.

[28] Meyer, P. A., *Un cours sur les integrales stochastiques, Seminaire de Probabilities X, Lecture Notes in Math. 511*, Springer, New York (1974).

[29] Pang, G., Talreja, R. and Whitt, W., *Martingale proofs of many-server heavy-traffic limits for Markovian queues,* Probability Surveys, 4 (2007), pp. 193-267.

[30] Protter, P., *Stochastic Differential Equations,* Second edition, Springer-Verlag (2004).

[31] Puhalskii, A. A. and Reiman, M. I., *The multi-class $GI/PH/N$ queue in the Halfin-Whitt regime,* Adv. Appl. Probab., 32 (2000), pp. 564-595.

[32] Reed, J. E., *The $G/GI/N$ queue in the Halfin-Whitt regime,* Ann. Appl. Probab. 19 (2009), pp. 2211–2269.

[33] Reed, J. E. and Ward, A. R., *Approximating the $GI/GI/1+GI$ queue with a nonlinear drift diffusion: hazard rate scaling in heavy traffic,* Math. Oper. Res., 33 (2008), pp. 606–644.

[34] van Leeuwaarden, J., Janssen, A. J. E. M. and Zwart, B., *Gaussian expansions and bounds for the Poisson distribution with application to the Erlang B formula*, Advances in Applied Probability 40, (2008), pp. 122-143 .

[35] van Leeuwaarden, J., Janssen A. J. E. M. and Zwart, B., *Refining square root safety staffing by expanding Erlang C*, Oper. Res., 59 (2011), pp. 1512–1522.

[36] Ward, A. and Kumar, S., *Asymptotically optimal admission control of a queue with impatient customers,* Math. Oper. Res., 33 (2008), pp. 167-202.

[37] Whitt, W., *Heavy traffic limits for the $G/H_2^*/n/m$ queue,* Math. Oper. Res., 30 (2005), pp. 1-27.

[38] Weerasinghe, A., *A bounded variation control problem for diffusion processes*, SIAM J. Control and Opt., 44 (2005), pp. 389-417.

Ananda P. Weerasinghe
396 Carver Hall
Department of Mathematics
Iowa State University
Ames, IA 50011, USA.
ananda@iastate.edu

Avishai Mandelbaum
Faculty of Industrial Engineering
and Management
Technion
Haifa 32000, ISRAEL.
avim@tx.technion.ac.il