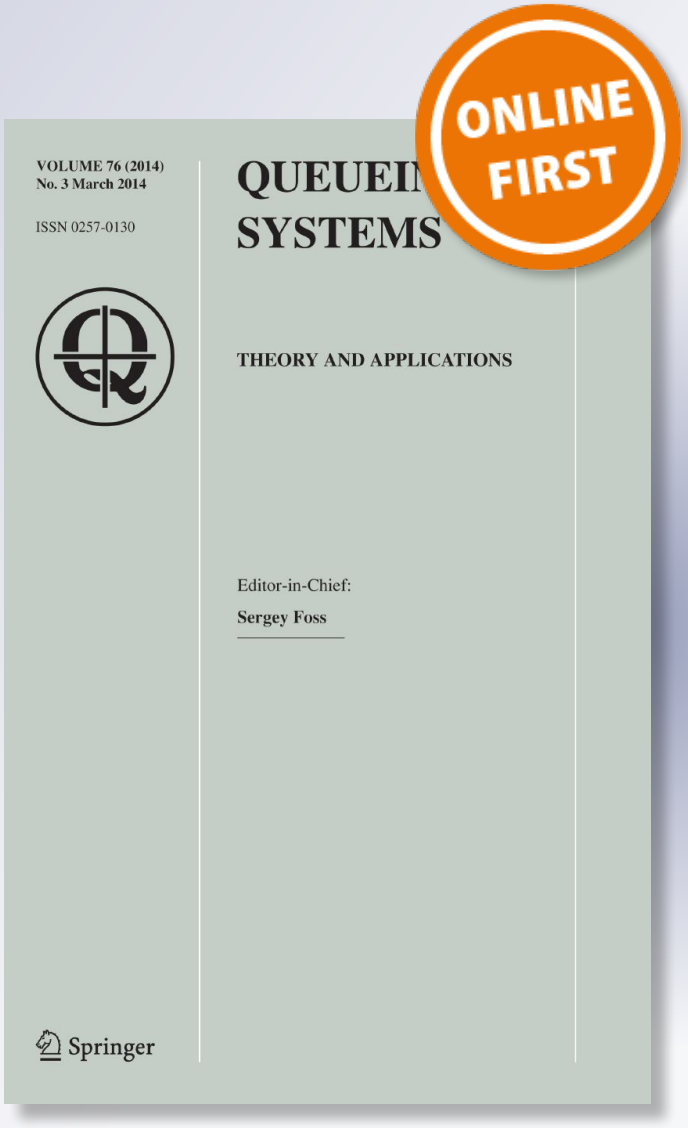
Optimizing buffer size for the retrial queue: two state space collapse results in heavy traffic

Rami Atar & Anat Lev-Ari

Queueing Systems
Theory and Applications

ISSN 0257-0130

Queueing Syst
DOI 10.1007/s11134-018-9585-y



Springer

Queueing Systems
https://doi.org/10.1007/s11134-018-9585-y

Optimizing buffer size for the retrial queue: two state space collapse results in heavy traffic

Rami Atar[1] · **Anat Lev-Ari**[1]

Received: 27 August 2017 / Revised: 24 June 2018

Abstract
We study a single server queueing model with admission control and retrials. In the heavy traffic limit, the main queue and retrial queue lengths jointly converge to a degenerate two-dimensional diffusion process. When this model is considered with holding and rejection costs, formal limits lead to a free boundary curve that determines a threshold on the main queue length as a function of the retrial queue length, above which arrivals must be rejected. However, it is known to be a notoriously difficult problem to characterize this curve. We aim instead at optimizing the threshold on the main queue length *independently* of the retrial queue length. Our main result shows that in the small and large retrial rate limits, this problem is governed by the *Harrison–Taksar free boundary problem*, which is a Bellman equation in which the free boundary consists of a single point. We derive the asymptotically optimal buffer size in these two extreme cases, as the scaling parameter and the retrial rate approach their limits.

Keywords Retrial queue · Diffusion approximation · Heavy traffic · The Harrison–Taksar free boundary problem · State space collapse

Mathematics Subject Classification 60F17 · 60J60 · 60K25 · 93E20

1 Introduction

The consideration of the single server queue with admission control gives rise to a tradeoff between holding costs and rejection costs. The goal of this paper is to study this tradeoff in the presence of retrials, in the heavy traffic limit. As far as the single

✉ Anat Lev-Ari
anatleva@gmail.com

Rami Atar
atar@ee.technion.ac.il

[1] Viterbi Faculty of Electrical Engineering, Technion–Israel Institute of Technology, 32000 Haifa, Israel

Published online: 13 July 2018

server queue with no retrials is concerned, the optimal tradeoff at the heavy traffic limit is characterized by the *Harrison–Taksar free boundary problem* [16]. This is a Hamilton–Jacobi–Bellman (HJB) equation in one variable (thus a nonlinear ODE) with a free boundary point. This point represents a threshold that one puts on the queue length, which governs the asymptotically optimal (AO) behavior. Namely, it is AO to admit arrivals if and only if the queue length at the time of arrival is below this threshold. One may refer to this threshold as the *optimal buffer size*. In a model with retrials, rejected arrivals may choose to retry entry at a later time. When considered in the heavy traffic limit, this model gives rise, as we show, to a two-dimensional diffusion process driven by a one-dimensional Brownian motion (BM). The two variables correspond to the main queue length and retrial queue length. In this setting, the HJB equation that governs the behavior of an AO admission policy is a fully nonlinear, degenerate elliptic partial differential equation in two variables. The best one may then hope for in such a setting is that there exists a free boundary curve in the two-dimensional space, which serves as a threshold that one puts on the main queue length, whose value may depend on the retrial queue length. It is known to be a notoriously difficult problem to characterize this curve. Moreover, to implement this approach one must use information on the retrial length, which in applications is not observable.

The approach that we take here is to optimize over a single parameter, namely a threshold (or buffer size), which is independent of the system's state. Our main result concerns limits of the retrial rate. It shows that when this rate goes to infinity or to zero, the optimization problem, set in a two-dimensional state space, reduces to the one-dimensional Harrison–Taksar free boundary problem alluded to above (with distinct parameters in the two cases). The assertion of the main result involves asymptotic optimality in a double limit, as the heavy traffic parameter and the retrial rate approach their limits.

To describe the model and results in more detail, consider a $G/G/1$ queue with a finite buffer of size b. When the number of customers in the buffer exceeds the value b, arriving customers are rejected, and with a fixed probability p leave the system. Otherwise, they decide to retry at a later, exponentially distributed, time. We refer to the $G/G/1$ queue as *the main queue*, and call the infinite server station that models retrials *the retrial queue*. The retrial queue has an infinite collection of exponential servers with service rate that we denote by μ; this parameter is the retrial rate. A retrying customer is treated at the main queue as a new arrival.

We look at this model at diffusion scale under a heavy traffic condition and a rescaling of the buffer size. First, we establish joint convergence of the pair of queue lengths to an obliquely reflected, degenerate diffusion process. Then we formulate the optimization problem alluded to above, where the buffer size b is selected so as to minimize a linear combination of holding costs at the main station and rejection count. The diffusion model obtained from the aforementioned weak convergence result is considered with a cost that describes the limit of the cost above. As already mentioned, attempting to fully treat the dynamic control problem leads to the difficult problem of identifying a free boundary curve. We observe that in both the limits $\mu \to 0$ and $\mu \to \infty$, this diffusion optimization problem is considerably simpler, and in particular is governed by the Harrison–Taksar free boundary problem [16]. This is an equation in one variable in which the free boundary consists of a single point, for which effective

numerical procedures exist. The term *state space collapse* (SSC) usually refers to a dimension reduction that occurs when the scaling parameter approaches its limit. However, the dimension reduction identified in this paper is different. It is the diffusion model that undergoes a dimension reduction, and this reduction occurs when the retrial rate parameter tends to a limit. The main result of this paper uses this SSC result for the diffusion model to establish AO buffer size for the queuing model in these two regimes. We also provide simulation results aimed at computing the optimal buffer sizes for general values of μ.

Queuing systems with retrials have long been of interest due to their important role in applications, such as computer networks and call centers. They have been extensively studied since they were first analyzed in [10]. We refer the reader to the survey papers [5–7,12,23,24]. As far as heavy traffic limits are concerned, results on retrial queues were obtained for several models. The paper [22] characterizes diffusion limits for various queuing models, including ones with retrials. The scaling regime differs from ours in terms of the retrial rate parameter as well as scaling the number of servers, and as a result, does not lead to a degenerate diffusion limit as in this paper. The same is true for the papers [2,3]. Other diffusion limit results for retrial queues include [13,14], which studied the $M/M/c$ queue with exponential retrial times. The scaling regime is different than ours, and corresponds to light traffic. The models in these two papers differ in the assumption on the waiting space: in [13] the main station is modeled as a server with no waiting space; customers who find all the servers busy join the retrial queue. In [14], customers who find all the servers busy may join a waiting queue or retry (with probabilities that depend on the state of the queue). In both papers, the diffusion limit is given in terms of an Ornstein–Uhlenbeck process.

Several papers have considered a related model under the limiting cases $\mu \to 0$ and $\mu \to \infty$ of the retrial rate (not in the heavy traffic asymptotics). These include [1,10,11]. In [10], the limit $\mu \to 0$ was studied for an $M/M/c$ loss system; that is, customers that find all the servers occupied are rejected, and with fixed probability retry after an exponential time. It was shown that the limit system is the classical Erlang loss system. In the papers [1,11], the $M/M/c$ loss system with exponential retrial rate was considered in the case $\mu \to \infty$, showing that the queue length distributions convergence to those of the associated multiserver queue without retrials (further background on the relationship between systems with and without retrials can be found in [4]). The phenomenon revealed in these three papers, namely that the limiting cases correspond to systems that do not exhibits retrials, is similar to the one obtained in the present paper. However, there are significant differences: The model under consideration in this paper is different; in particular it has waiting space at the main queue. Moreover, the system is considered under a scaling limit (and thus general service time distributions can be addressed), and furthermore, the main goal of our study is the aspect of optimizing the buffer size. Finally, the paper [21] considers a scaling limit result that combines diffusion scale and the regime $\mu \to 0$. The model is the $M/M/1$ queue with no waiting room and exponential retrials, and, under a light traffic assumption, the diffusion limit is again an Ornstein–Uhlenbeck process.

1.1 Organization of the paper

In Sect. 2 we describe the model and the processes under the diffusion scale. In Sect. 3 we introduce the limit model and prove the first main result—the weak convergence of the prelimit diffusion processes. Section 4 is devoted to study the optimization problem and analyze it with respect to different values of μ. The second main result—the AO in the dual limits is presented and proven. Section 5 presents the simulation to study the behavior of the value function and optimal buffer size for general values of μ.

1.2 Notation

For $x, y \in \mathbb{R}^k$ (k a positive integer), $x \cdot y$ and $\|x\|$ denote the usual scalar product and ℓ_2 norm, respectively. With $\mathbb{R}_+ = [0, \infty)$, for $f : \mathbb{R}_+ \to \mathbb{R}^k$, we let $\|f\|_T = \sup_{t\in[0,T]} \|f(t)\|$. For a Polish space $\mathcal{S}$, we let $\mathbb{C}_{\mathcal{S}}[0, \infty)$ and $\mathbb{D}_{\mathcal{S}}[0, \infty)$ denote the set of continuous and, respectively, RCLL functions $[0, \infty) \to \mathcal{S}$. We endow $\mathbb{D}_{\mathcal{S}}[0, \infty)$ with the Skorohod J_1 topology. We say that a sequence of stochastic processes X_n with sample paths in $\mathbb{D}_{\mathcal{S}}[0, \infty)$ is *C-tight* if it is tight and each subsequential limit of it is a process that has sample paths in $\mathbb{C}_{\mathcal{S}}[0, \infty)$, a.s. We write $X_n \Rightarrow X$ for convergence in distribution. We use notation such as $X(t)$ and X_t interchangeably for stochastic processes X and $t \in \mathbb{R}_+$.

2 The model and its diffusion scaling

In this section, we introduce the elements of the model and the stochastic processes involved, and then introduce the diffusion scaling under the heavy traffic condition.

2.1 The model

Our goal is to study a finite buffer $G/G/1$ queue with retrials (Fig. 1). We model it by considering two stations: one, which is referred to as *the main station*, has a single server with general service time distribution and a finite buffer. The other, referred to as *the retrial station*, has an infinite number of exponential servers with identical service rate parameter. Each customer arriving to find the main station's buffer full leaves the system with probability $0 < p < 1$. If it does not leave the system (which happens with probability $q = 1 - p$) it is routed to the retrial station. Every departure from the retrial station is routed back to the main station and is treated as a new incoming customer. This is equivalent to stating that every rejected customer that decides to retry (with probability q) does so according to an exponential clock (with a fixed parameter).

The stochastic processes which we now introduce are all defined on a probability space $(\Omega, \mathcal{F}, P)$. We denote by E the expectation w.r.t. P. The parameters and processes are indexed by $n \in \mathbb{N}$. An important aspect of the scaling that we consider is that, while the arrival and service rate at the main station are accelerated in the usual fashion which leads to heavy traffic approximations, we do not accelerate the retrial

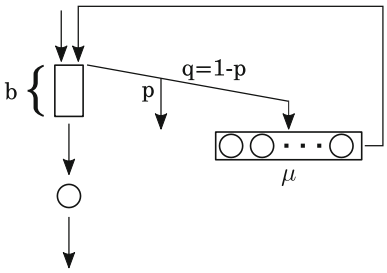


Fig. 1 The finite buffer $G/G/1$ queue with retrials

clock. Thus the infinite collection of servers at the retrial station have a fixed service rate.

We start with the external arrivals, and model them in terms of a renewal processes. Let $\{IA(l), l \in \mathbb{N}\}$ be an independent and identically distributed (IID) sequence of positive random variables (RVs) with $E[IA(1)] = 1$ and $\sigma_{IA}^2 = \text{var}(IA(1)) < \infty$. Let $\lambda^n > 0$ be the reciprocal of the mean inter-arrival time for the n-th system. Then the number of external arrivals up to time t is given by

$$A^n(t) = A(\lambda^n t), \text{ where } A(t) = \sup\left\{l \geq 0 : \sum_{k=1}^{l} IA(k) \leq t\right\}, \quad t \geq 0.$$

The main station also admits internal arrivals, namely the customers that depart from the retrial station. We denote the counting process of departures from the retrial station by $D_2^n(t)$. Hence the counting process for the total number of arrivals to the main station (including external and internal arrivals) is given by

$$A_1^n(t) = A^n(t) + D_2^n(t).$$

For the service process of the main station, consider an IID sequence of positive RVs $\{ST(l), l \in \mathbb{N}\}$, with $E[ST(1)] = 1$ and $\sigma_{ST}^2 = \text{var}(ST(1)) < \infty$. Let $\mu_1^n > 0$ be the reciprocal of the mean service time at the n-th system. Then the number of service completions at the main station by the time the server has devoted t units of time is given by

$$S^n(t) = S(\mu_1^n t), \text{ where } S(t) = \sup\left\{l \geq 0 : \sum_{k=1}^{l} ST(k) \leq t\right\}, \quad t \geq 0.$$

Note that the above is different from the actual number of customers to complete service by time t. To introduce the latter, let $T^n(t)$ denote the cumulative time the server works during $[0, t]$. Then the number of customer completions (or departures) up to time t is given by

$$D_1^n(t) = S^n(T^n(t)). \tag{1}$$

The buffer size depends on n, and we denote it by b^n. Each time the number of customers reaches the size b^n, the system manger rejects arrivals. Denote by $C^n(t)$ the counting process for customer rejections. Consider an IID sequence of $\{0, 1\}$-valued RVs $\{\xi_i\}$, with $P(\xi_1 = 1) = q$. These are used to model retrials: The i-th rejected customer is rerouted to the retrial station if and only if $\xi_i = 1$. Thus, the counting process for arrivals into the retrial station (which is the number of rejected customers that reroute) is given by

$$A_2^n(t) = G(C^n(t)) = (G \circ C^n)(t), \tag{2}$$

where we define

$$G(u) = \sum_{i=1}^{\lfloor u \rfloor} \xi_i, \qquad u \in \mathbb{R}_+. \tag{3}$$

As for the queue length processes, denote by $X^n(t)$ the number of customers in the main station (including the customer in service) and by $R^n(t)$ the number of customers in the retrial station. These processes satisfy the balance equations

$$X^n(t) = X^n(0) + A_1^n(t) - C^n(t) - D_1^n(t), \tag{4}$$

and

$$R^n(t) = R^n(0) + A_2^n(t) - D_2^n(t). \tag{5}$$

Also, T^n and X^n satisfy the relation $T^n(t) = \int_0^t 1_{\{X^n(s)>0\}} \mathrm{d}s$. Moreover, the retrial clocks are assumed to be exponential with mean $1/\mu$, hence D_2^n and R^n satisfy the relation

$$D_2^n(t) = N\left(\int_0^t \mu R^n(s)\mathrm{d}s\right), \tag{6}$$

where N is a standard Poisson process.

The primitive data A, S, G, and N are assumed to be mutually independent. Also, the initial condition $(X^n(0), R^n(0))$ is independent of the primitive data.

2.2 The heavy traffic condition and diffusion scaling

The parameters satisfy the following assumptions: There exist $\lambda, \mu_1 \in (0, \infty)$ and $\hat{\lambda}, \hat{\mu}_1 \in \mathbb{R}$ such that, as $n \to \infty$,

$$\frac{\lambda^n}{n} \to \lambda, \qquad \frac{\mu_1^n}{n} \to \mu_1,$$
$$\hat{\lambda}^n = \frac{\lambda^n - n\lambda}{\sqrt{n}} \to \hat{\lambda}, \qquad \hat{\mu}_1^n = \frac{\mu_1^n - n\mu_1}{\sqrt{n}} \to \hat{\mu}_1.$$

We assume that the system is in heavy traffic, in the sense that $\lambda = \mu_1$. Also, the buffer size is assumed to scale like $\sqrt{n}$. More precisely, $b^n = \lceil b\sqrt{n} \rceil$ for some $b > 0$.

The diffusion-scaled queue length processes are defined as

$$\hat{X}^n(t) = n^{-1/2} X^n(t), \qquad \hat{R}^n(t) = n^{-1/2} R^n(t). \tag{7}$$

The scaled initial condition $(\hat{X}^n(0), \hat{R}^n(0))$ is assumed to converge in distribution to some deterministic $(x, r) \in \mathbb{R}^2_+$.

2.3 Model equations via the Skorohod map on [0, *b*]

It is instrumental to formulate the model equations on the basis of the Skorohod map on an interval. The *Skorokhod problem on* $[0, b]$ is the problem of finding, for any $\psi \in \mathcal{D}[0, \infty)$, a triplet $(\phi, \eta_l, \eta_u) \in (\mathcal{D}[0, \infty))^3$ such that

1. $\phi(t) = \psi(t) + \eta_l(t) - \eta_u(t)$, $\phi(t) \in [0, b]$ for all t.
2. η_l, η_u are nonnegative and non-decreasing and one has

$$\int_{[0,\infty)} \mathbb{I}_{\{\phi(t)>0\}} \mathrm{d}\eta_l(t) = 0, \qquad \int_{[0,\infty)} \mathbb{I}_{\{\phi(t)<b\}} \mathrm{d}\eta_u(t) = 0.$$

It is well-known that the Skorohod problem on $[0, b]$ has a unique solution. The solution map, called the *Skorohod map*, is denoted by $\Gamma_{0,b}$. Thus $(\phi, \eta_l, \eta_u) = \Gamma_{0,b}(\psi)$. Existence, uniqueness and several properties of this map can be found in [19]. A specific important property is the Lipschitz continuity with respect to the sup norm; that is, there exists a constant c_Γ such that, for any $T > 0$ and any $\psi, \tilde{\psi} \in \mathcal{D}[0, \infty)$, writing $(\phi, \eta_l, \eta_u) = \Gamma_{0,b}(\psi)$ and $(\tilde{\phi}, \tilde{\eta}_l, \tilde{\eta}_u) = \Gamma_{0,b}(\tilde{\psi})$,

$$\|\phi - \tilde{\phi}\|_T + \|\eta_l - \tilde{\eta}_l\|_T + \|\eta_u - \tilde{\eta}_u\|_T \le c_\Gamma \|\psi - \tilde{\psi}\|_T. \tag{8}$$

Moreover, it follows from the explicit representation of the map in [19] that there is one Lipschitz constant c_Γ that is valid for all $b \in (0, \infty)$.

To make the connection to our model we introduce some additional diffusion-scaled processes, namely

$$\hat{A}^n(t) = \frac{A^n(t) - \lambda^n t}{\sqrt{n}}, \qquad \hat{S}^n(t) = \frac{S^n(t) - \mu_1^n t}{\sqrt{n}}, \qquad \hat{D}_1^n(t) = \hat{S}^n(T^n(t)), \tag{9}$$

$$\hat{L}^n(t) = \frac{\mu_1^n}{\sqrt{n}} \left(t - T^n(t)\right), \qquad \hat{A}_2^n(t) = \frac{A_2^n(t)}{\sqrt{n}}, \qquad \hat{C}^n(t) = \frac{C^n(t)}{\sqrt{n}}, \tag{10}$$

as well as the process

$$e^n(t) = \frac{N\left(\int_0^t \mu R^n(s)\mathrm{d}s\right) - \int_0^t \mu R^n(s)\mathrm{d}s}{\sqrt{n}} \tag{11}$$

that will serve as an error term, and finally, the constants $\hat{y}^n = \hat{\lambda}^n - \hat{\mu}_1^n$. Using these definitions in (4) and (5), the diffusion-scaled processes are seen to satisfy

$$\begin{cases} \hat{X}^n(t) = \hat{X}^n(0) + \hat{y}^n t + \hat{A}^n(t) - \hat{D}_1^n(t) + e^n(t) + \int_0^t \mu \hat{R}^n(s)\mathrm{d}s - \hat{C}^n(t) + \hat{L}^n(t), \\ \hat{R}^n(t) = \hat{R}^n(0) - e^n(t) + \hat{A}_2^n(t) - \int_0^t \mu \hat{R}^n(s)\mathrm{d}s. \end{cases} \tag{12}$$

The constraint $0 \le X^n(t) \le b^n$ that is satisfied by X^n can be written as $0 \le \hat{X}^n(t) \le \hat{b}^n$, where we define $\hat{b}^n = n^{-1/2} b^n$. Note that $\hat{b}^n = b + e_b^n$, where $e_b^n = (\lceil b\sqrt{n} \rceil - b\sqrt{n})/\sqrt{n} \to 0$ as $n \to \infty$. The processes $\hat{C}^n$, $\hat{L}^n$ have nonnegative, non-decreasing sample paths, and they satisfy

$$\int_{[0,\infty)} \mathbb{I}_{\{\hat{X}_t^n > 0\}} \mathrm{d}\hat{L}_t^n = 0, \qquad \int_{[0,\infty)} \mathbb{I}_{\{\hat{X}_t^n < \hat{b}^n\}} \mathrm{d}\hat{C}_t^n = 0. \tag{13}$$

As a result, we have the relations

$$\begin{cases} (\hat{X}^n, \hat{L}^n, \hat{C}^n) = \Gamma_{0,\hat{b}^n}(\hat{Z}^n), \\ \hat{Z}^n(t) = \hat{X}^n(0) + \hat{y}^n t + \hat{A}^n(t) - \hat{D}_1^n(t) + e^n(t) + \int_0^t \mu \hat{R}^n(s)\mathrm{d}s, \\ \hat{R}^n(t) = \hat{R}^n(0) - e^n(t) + \hat{A}_2^n(t) - \int_0^t \mu \hat{R}^n(s)\mathrm{d}s. \end{cases} \tag{14}$$

3 The diffusion limit

Let $\hat{y} = \lim_{n\to\infty} \hat{y}^n = \hat{\lambda} - \hat{\mu}_1$, $\sigma^2 = \lambda(\sigma_{IA}^2 + \sigma_{ST}^2)$ and let W be a $(\hat{y}, \sigma^2)$-BM. In this section we prove that the diffusion-scaled processes converge to the solution of the following set of equations:

$$\begin{cases} (X, L, C) = \Gamma_{0,b}(Z), \\ Z(t) = x + W(t) + \int_0^t \mu R(s)\mathrm{d}s, \\ R(t) = r + qC(t) - \int_0^t \mu R(s)\mathrm{d}s. \end{cases} \tag{15}$$

In particular, by the definition of $\Gamma_{0,b}$, one has $X = Z + L - C$. This system of equations can be seen as an SDE with reflection in $[0, b] \times \mathbb{R}_+$ (note that $R(t) \ge 0$ since the equation $e^{\mu t} R(t) = e^{\mu t} r + q \int_0^t e^{\mu s} \mathrm{d}C(s)$ holds and C is non-decreasing). To write this equation for the processes $Y_t = (X_t, R_t)^T$, $M_t = (L_t, C_t)^T$, set the directions of reflection to be d_1 on the part $\{0\} \times \mathbb{R}_+$ of the boundary, and d_2 on the part $\{b\} \times \mathbb{R}_+$ of the boundary, where

$$d_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad d_2 = \begin{pmatrix} -1 \\ q \end{pmatrix}$$

(see Fig. 2).

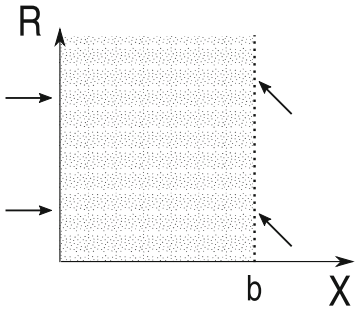


Fig. 2 Reflection directions

Let the drift coefficient, diffusion coefficient and reflection matrix be given by

$$B(y) = \begin{pmatrix} \hat{y} + \mu y_2 \\ -\mu y_2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \sigma \\ 0 \end{pmatrix}, \qquad D = (d_1, d_2) = \begin{pmatrix} 1 & -1 \\ 0 & q \end{pmatrix},$$

for any $y = (y_1, y_2)$, and let $\tilde{W}$ be a 1-dimensional standard BM. Then the SDE is given by

$$\mathrm{d}Y_t = B(Y_t)\mathrm{d}t + \Sigma \mathrm{d}\tilde{W}_t + D\mathrm{d}M_t, \tag{16}$$

with the initial condition $Y_0 = (x, r)^T$. Note that there is no need to specify refection directions on $[0, b] \times \{0\}$ due to the fact that the second component of the drift coefficient $B(y)$, namely $-\mu y_2$, vanishes on this part of the boundary. Thus Y is a two-dimensional diffusion process which is degenerate in the second diffusion coefficient. By [20, Th. 4.3], there exists a unique solution to (16), or equivalently (15) (the result in [20] is stated for a bounded domain, but it is standard, by a localization argument, to argue that existence and uniqueness hold for the unbounded domain considered here).

We now introduce our first main result.

Theorem 3.1 *As* $n \to \infty$,

$$(\hat{X}^n, \hat{R}^n, \hat{L}^n, \hat{C}^n) \Rightarrow (X, R, L, C).$$

Proof We divide the proof into four main steps. Step 1 develops a variation of (14) where the sequence of Skorohod maps $\Gamma_{0,\hat{b}^n}$ is replaced by $\Gamma_{0,b}$. The second and third steps are concerned with tightness of the underlying processes, and the last step combines these results to obtain convergence.

Step 1 $\Gamma_{0,b}$ instead of $\Gamma_{0,\hat{b}_n}$.

We first show that in the set of equations (14) for the prelimit processes, the maps $\Gamma_{0,\hat{b}_n}$ can be changed into the single map $\Gamma_{0,b}$ at the cost of adding some error terms in this set of equations. More precisely, let $X^{*,n}(t) = \hat{X}^n(t) \wedge b$. If we let $e_1^n(t) = \hat{X}^n(t) - X^{*,n}(t)$ then $0 \le e_1^n(t) \le e_b^n$, where we used the fact that $\hat{X}^n(t) \in [0, \hat{b}^n]$ for all t. In particular, the process $e_1^n \Rightarrow 0$ as $n \to \infty$. Next, we can add to the first equation in (12) the term e_1^n to both sides, and obtain an equation for $X^{*,n}$. It would correspond to the second equation in (14), with $\hat{Z}^n$ replaced by $Z^{*,n} = \hat{Z}^n + e_1^n$. We

have achieved the following: The process $X^{*,n}$ satisfies $X^{*,n}(t) \in [0, b]$, for all t, as well as $X^{*,n} = Z^{*,n} - \hat{C}^n - \hat{L}^n$, while, analogously to (13),

$$\int_{[0,\infty)} \mathbb{I}_{\{X_t^{*,n}>0\}} \mathrm{d}\hat{L}_t^n = 0, \qquad \int_{[0,\infty)} \mathbb{I}_{\{X_t^{*,n}<b\}} \mathrm{d}\hat{C}_t^n = 0.$$

While the first identity above is immediate from (13), the second one follows from the fact that $X_t^{*,n} < b$ implies $\hat{X}_t^n < b$, which in turn implies $\hat{X}_t^n < \hat{b}^n$. We can thus write the following set in place of (14):

$$\begin{cases} (X^{*,n}, \hat{L}^n, \hat{C}^n) = \Gamma_{0,b}(Z^{*,n}), \\ Z^{*,n}(t) = \hat{X}^n(0) + \hat{y}^n t + \hat{A}^n(t) - \hat{D}_1^n(t) + e^n(t) + e_1^n(t) + \int_0^t \mu \hat{R}^n(s)\mathrm{d}s, \\ \hat{R}^n(t) = \hat{R}^n(0) - e^n(t) + \hat{A}_2^n(t) - \int_0^t \mu \hat{R}^n(s)\mathrm{d}s. \end{cases} \tag{17}$$

Step 2 The sequence of RVs $\{\hat{R}^n(T)\}_{n\in\mathbb{N}}$ is tight for every T.
Toward proving this statement, fix T. Denote

$$\xi^n(t) = \hat{X}^n(0) + \hat{y}^n t + \hat{A}^n(t) - \hat{D}_1^n(t) + e_1^n(t), \qquad \tilde{Z}^n(t) = \xi^n(t) + \int_0^t \mu \hat{R}^n(s)\mathrm{d}s, \tag{18}$$

and note that $Z^{*,n}(t) = \tilde{Z}^n(t) + e^n(t)$. First, we show that $\|\xi^n\|_T$ is a tight sequence of RVs. By the FCLT for renewal processes, $(\hat{A}^n, \hat{S}^n) \Rightarrow (A, S)$, where A is a $(0, \lambda\sigma_{IA}^2)$-BM, S is a $(0, \mu_1\sigma_{ST}^2)$-BM, and A and S are independent (see §17 of [9]). As a result, $\|\hat{A}^n\|_T$ is a tight sequence of RVs, and so is $\|\hat{S}^n\|_T$. By the definition of $\hat{D}_1^n$ [see (9)] and the fact that $T^n(t) \le t$ for all t, $\|\hat{D}_1^n\|_T \le \|\hat{S}^n\|_T$. Since, by our assumptions, we also have $\hat{X}^n(0) \Rightarrow x$ and $\hat{y}^n \to \hat{y}$, and as we have shown $e_1^n \Rightarrow 0$, it follows that $\|\xi^n\|_T$ forms a tight sequence of RVs.

We use the set of equations (17) to develop a bound on $\hat{R}^n(T)$. It follows from the Lipschitz property (8) of the Skorohod map $\Gamma = \Gamma_{0,b}$ that there exists a constant c_Γ that does not depend on ε, t or n such that

$$\|\hat{C}^n\|_t + \|\hat{L}^n\|_t + \|X^{*,n}\|_t \le c_\Gamma \|Z^{*,n}\|_t \le c_\Gamma \left(\|\tilde{Z}^n\|_t + \|e^n\|_t\right), \qquad t \in [0, T]. \tag{19}$$

In addition, it follows from (2) and (3) that $A_2^n(t) \le C^n(t)$, and from (5) that $R^n(t) \le R^n(0) + A_2^n(t)$. Thus, by the rescaling in (14), we have

$$0 \le \hat{R}^n(t) \le \hat{R}^n(0) + \hat{A}_2^n(t) \le \hat{R}^n(0) + \hat{C}^n(t), \qquad t \in [0, T]. \tag{20}$$

As a result, $\hat{R}^n(t) \le \hat{R}^n(0) + c_\Gamma(\|\tilde{Z}^n\|_t + \|e^n\|_t)$. Using this in the definition of $\tilde{Z}^n$,

$$\begin{aligned} \hat{R}^n(t) &\le \hat{R}^n(0) + c_\Gamma(\|\tilde{Z}^n\|_t + \|e^n\|_t) \\ &\le \hat{R}^n(0) + c_\Gamma \left(\|\xi^n\|_T + \|e^n\|_t + \int_0^t \mu \hat{R}^n(s)\mathrm{d}s\right), \qquad t \in [0, T], \end{aligned} \tag{21}$$

where we have used the fact that $\hat{R}^n \geq 0$. We can now use Gronwall's lemma, by which

$$\hat{R}^n(t) \leq [\hat{R}^n(0) + c_\Gamma(\|\xi^n\|_T + \|e^n\|_t)]e^{c_\Gamma \mu T}, \qquad t \in [0, T]. \tag{22}$$

We have shown that $\|\xi^n\|_T$ are tight RVs. The RVs $\hat{R}^n(0)$ are also tight by the convergence assumption. Let $\varepsilon > 0$ be given. Then there exists d such that $\mathbb{P}(\Omega^{n,1}) > 1 - \varepsilon$, where

$$\Omega^{n,1} = \{\|\xi^n\|_T \vee |\hat{R}^n(0)| < d\}.$$

Fix such d. Let $k = [d + c_\Gamma(d+1)]e^{c_\Gamma \mu T} + 1$ and let

$$\tau^n = \inf\{t : \hat{R}^n(t) \geq k\}$$

be the first time the process $\hat{R}^n$ exceeds k. Also, fix $\varepsilon_0 \in (0, 1)$ and let

$$\Omega^{n,2} = \left\{ \sup_{u \in [0, \mu T(k+1)]} \left\| \frac{N(\sqrt{n}u)}{\sqrt{n}} - u \right\| < \varepsilon_0 \right\}.$$

By the FLLN, $\mathbb{P}(\Omega^{n,2}) > 1 - \varepsilon$ for all sufficiently large n. By the definition of τ^n and the fact that the sample paths of $\hat{R}^n$ are non-decreasing, it follows that $\hat{R}^n(s) \leq k$ for all $s < T \wedge \tau^n$. Since the jumps of the process R^n are of size 1, those of $\hat{R}^n$ are of size $n^{-1/2}$, and so we also have $\hat{R}^n(T \wedge \tau^n) \leq k + n^{-1/2}$. Hence, by the expression (11) for e^n, we see that

$$\text{the bound } \|e^n\|_{T \wedge \tau^n} \leq \varepsilon_0 \text{ holds on the event } \Omega^{n,2}. \tag{23}$$

We now use this in (22). On the event $\Omega^n = \Omega^{n,1} \cap \Omega^{n,2}$, one has

$$\hat{R}^n(T \wedge \tau^n) \leq [d + c_\Gamma(d+1)]e^{c_\Gamma \mu T} = k - 1.$$

Again, by the definition of τ^n, on the event $\{\tau^n \leq T\}$ one has $\hat{R}^n(\tau^n) \geq k$. Combining this with the above display shows that on the event Ω^n one has $\tau^n > T$. As a result, the estimate for $\hat{R}^n(T \wedge \tau^n)$ is valid for $\hat{R}^n(T)$, that is, $\hat{R}^n(T) \leq k - 1$ on Ω^n. Since we have $\liminf_n \mathbb{P}(\Omega^n) \geq 1 - 2\varepsilon$, we have thus shown that given ε there exists k such that $\limsup_n \mathbb{P}(\hat{R}^n(T) > k) \leq 2\varepsilon$. Hence $\hat{R}^n(T)$ is a tight sequence of RVs.

Step 3 The sequence of processes $(\hat{A}^n, \hat{S}^n, \hat{D}_1^n, \hat{A}_2^n, \hat{X}^n, \hat{R}^n, \int_0^{\cdot} \hat{R}^n(s)\mathrm{d}s, \hat{L}^n, \hat{C}^n)$ is C-tight, and $e^n \Rightarrow 0$.

By (23) and the above argument, we can now state that $\|e^n\|_T \leq \varepsilon_0$ holds on the event Ω^n. Since ε and ε_0 are arbitrary, it follows that $\|e^n\|_T \Rightarrow 0$ as $n \to \infty$.

Next, from Step 2, C-tightness of the sequence of processes $\int_0^t \mu \hat{R}^n(s)\mathrm{d}s$ follows immediately.

Also, as already mentioned, $\hat{A}^n$ converges to a BM, thus in particular is C-tight. Similarly, $\hat{S}^n$ is C-tight, hence so is $\hat{D}_1^n$, as follows from its definition as a pathwise time change of $\hat{S}^n$ via the uniformly Lipschitz paths of T^n. This shows that ξ^n, and

in turn $\tilde{Z}^n$, are C-tight. Since we have shown that e^n converge to zero, it follows that $Z^{*,n}$ are also C-tight. In view of the first equation in (17), the Lipschitz property of $\Gamma_{0,b}$ implies the C-tightness of $X^{*,n}$, $\hat{L}^n$ and $\hat{C}^n$. The C-tightness of $\hat{X}^n$ follows from the fact that $e_1^n \Rightarrow 0$.

Regarding the process $\hat{R}^n$, note that, by (2) and (3), we can write $\hat{A}_2^n(t) = \tilde{G}^n(\hat{C}^n(t))$, where $\tilde{G}^n(u) = n^{-1/2}G(n^{1/2}u)$. By the FLLN, for any v, $\sup_{u\in[0,v]} |\tilde{G}^n(u) - qu| \Rightarrow 0$ as $n \to \infty$. Thus, by the C-tightness of $\hat{C}^n$ it follows that $\hat{A}_2^n$ are also C-tight. Using this in the last equation of (17) gives the C-tightness of the processes $\hat{R}^n$.

Step 4 Convergence.

By the definition (10) of $\hat{L}^n$, the tightness of these processes and the fact that $\mu_1^n/\sqrt{n} \to \infty$ imply that $\sup_{t\in[0,T]} |T^n(t) - t| \Rightarrow 0$ as $n \to \infty$. Recalling the definition (9) of $\hat{D}_1^n$ and the convergence of $(\hat{A}^n, \hat{S}^n)$ to (A, S) gives that $(\hat{A}^n, \hat{D}_1^n) \Rightarrow (A, S)$, where A and S are independent. As a result, $\hat{A}^n - \hat{D}_1^n \Rightarrow A - S$. It follows that $\xi^n \Rightarrow x + W$.

We next use the C-tightness obtained in the previous step. We fix a convergent subsequence of $(\xi^n, Z^{*,n}, \hat{A}_2^n, \hat{X}^n, \hat{R}^n, \hat{L}^n, \hat{C}^n)$ and denote its limit by $(x + W, Z, A_2, X, R, L, C)$. In what follows we will argue that this limit satisfies (15). By uniqueness of solutions to this set of equations, the convergence of the whole sequence will follow.

Toward arguing that the subsequential limit satisfies (15), note first that since the error terms e^n and e_1^n converge to zero, we have the same limit, Z, for $\tilde{Z}^n$, $Z^{*,n}$ and Z^n. Moreover, by (18), the relation $Z = x + W + \int_0^{\cdot} \mu R(s)\mathrm{d}s$ must hold. Taking limits in the first equation of (17) and using the continuity of the map $\Gamma_{0,b}$ gives

$$(X, L, C) = \Gamma_{0,b}(Z).$$

Recall from Step 3 the description of $\hat{A}_2^n(t)$ as $\tilde{G}^n(\hat{C}^n(t))$ and the uniform convergence of $\tilde{G}^n(u)$ to qu. Along with the convergence $\hat{C}^n \Rightarrow C$, this gives $\hat{A}_2^n \Rightarrow qC$. Hence $A_2 = qC$. Using this now in the last equation of (17) gives

$$R(t) = r + qC(t) - \int_0^t \mu R(s)\mathrm{d}s.$$

Hence the subsequential limit processes satisfy (15). As a result, convergence to the unique solution of (15) follows. □

Remark 3.1 Note that the precise size of the term e_b^n, which serves as an upper bound on the error term e_1^n in Step 1 of the proof, does not matter as long as it converges to zero. Consequently, the sequence of rescaled buffer sizes $\hat{b}_n$ could be replaced by any sequence $\tilde{b}_n \to b$.

4 Optimizing the buffer size

In this section, we consider the constrained diffusion model (15) along with a cost of the form

$$E\int_0^\infty e^{-\alpha t}(c_1X(t)+c_2C(t))\mathrm{d}t,$$

which penalizes both queue length at the main station and rejections. One might consider a dynamic control problem where the cost is minimized over control processes C that are adapted to the filtration of the state process $Y=(X,R)$. This is a standard formulation in stochastic control, where the value function can be characterized in terms of a Hamilton–Jacobi–Bellman (HJB) equation [15]. However, this direction has two major drawbacks in the current setting. First, for the queuing model under consideration it is natural to assume that the decision maker cannot observe the number of customers at the retrial queue. Hence a setting where the state process Y is fully observable is not suitable. Second, the HJB equation is in two dimensions, and it is expected (as is almost always the case) that it is not solvable in an explicit form. Our study focuses on a different question, namely the problem of finding the buffer size b that minimizes the above cost. Our results address this problem under the two limits $\mu\to\infty$ and $\mu\to 0$, and in both cases we connect the optimization problem to the Harrison–Taksar free boundary problem [16], which can be seen as a HJB equation in one dimension.

4.1 The optimization problem setting and results

4.1.1 The diffusion optimization problem

Throughout this section, the initial retrial queue is set to zero, that is, $r=0$. Thus the processes X, R, C and L of (15) satisfy

$$\begin{cases}(X,L,C)=\Gamma_{0,b}(Z),\\ Z(t)=x+W(t)+\int_0^t\mu R(s)\mathrm{d}s,\\ R(t)=qC(t)-\int_0^t\mu R(s)\mathrm{d}s,\end{cases}\tag{24}$$

where we recall that W is a $(\hat{y},\sigma^2)$-BM. The dependence of these processes on both the parameters b and μ is important, and so when we wish to emphasize this dependence we denote these processes as $X^{\mu,b}$, etc.

The cost function associated with the problem is

$$J^{\mu,b}_{\mathrm{DOP}}=E\int_0^\infty e^{-\alpha t}\left(c_1X^{\mu,b}(t)+c_2C^{\mu,b}(t)\right)\mathrm{d}t\tag{25}$$

$$=E\int_0^\infty e^{-\alpha t}\left(c_1X^{\mu,b}(t)\mathrm{d}t+\frac{c_2}{\alpha}dC^{\mu,b}(t)\right),\tag{26}$$

where the identity follows by integration by parts (it is not hard to see that $\lim_{t\to\infty} e^{-\alpha t} C_t = 0$, a.s.).

The diffusion optimization problem (DOP) is to minimize the cost:

$$V^{\mu}_{\mathrm{DOP}} = \inf_{b\in(0,\infty)} J^{\mu,b}_{\mathrm{DOP}}. \tag{27}$$

The dependence of these quantities on x, c_1 and c_2 is denoted by writing them as $J^{\mu,b}[x; c_1, c_2]$ and $V^{\mu}_{\mathrm{DOP}}[x; c_1, c_2]$.

4.1.2 The Harrison–Taksar free boundary problem

A related, simpler problem is that of minimizing the cost of (25) when $(X, L, C) = \Gamma_{0,b}(x + W)$. More precisely, let

$$J^{b}_{\mathrm{HT}}[x; c_1, c_2] = E\int_0^{\infty} e^{-\alpha t}[c_1\tilde{X}_t + c_2\tilde{C}_t]\mathrm{d}t, \qquad \text{where } (\tilde{X}, \tilde{L}, \tilde{C}) = \Gamma_{0,b}(x + W), \tag{28}$$

(thus, in particular, $\tilde{X} = x + W + \tilde{L} - \tilde{C}$) and

$$V_{\mathrm{HT}}[x; c_1, c_2] = \inf_{b\in(0,\infty)} J^{b}_{\mathrm{HT}}[x; c_1, c_2]. \tag{29}$$

This is a closely related problem to the one treated by Harrison and Taksar [16] via the Bellman equation:

$$\begin{cases} \left[\frac{1}{2}\sigma^2 f'' + \hat{y} f' - \alpha f + c_1\right] \wedge f' \wedge \left[\frac{c_2}{\alpha} - f'\right] = 0, \quad \text{in } (0, b_0), \\ f'(0) = 0, \qquad f'(b_0) = \frac{c_2}{\alpha}. \end{cases} \tag{30}$$

This is an equation for the unknown pair (f, b_0), where f is a C^2 function defined on $\mathbb{R}_+$, and b_0 is in $(0, \infty)$. It is a free boundary problem in the sense that one of the boundary conditions is given at the point b_0 that is unknown. Following the results of [16], we can state the following: There exists a unique (classical) solution (f, b_0) to Eq. (30). Moreover, f is equal to the function $x \mapsto V_{\mathrm{HT}}[x; c_1, c_2]$ defined by (28), and b_0 is a value of b for which the infimum on the RHS of (29) is attained as a minimum. In particular, Eq. (30) characterizes the optimal buffer size, b_0.

In what follows we denote the value of b_0 that is characterized by the above equation by $b_{\mathrm{HT}}[c_1, c_2]$.

Remark 4.2 To be precise, there are several differences between the above statements and the results obtained in [16]. First, the results of [16] did not include uniqueness of classical solutions to the Bellman equation (30); this point was settled in [8, Proposition 2.2]. Second, the formulation in [16] did not assume that the processes $\tilde{L}$ and $\tilde{C}$ were boundary terms for the Skorohod map the way they are represented in (28), but

considered a broader class of non-decreasing processes. However, the result of [16] states that the optimum is achieved by the tuple $(\tilde{X}, \tilde{L}, \tilde{C}) = \Gamma_{0,b_0}(x + W)$, which is precisely of the form given in (28) (with $b = b_0$). Clearly, this implies that the optimum within this smaller class of processes is also given by $(\tilde{X}, \tilde{L}, \tilde{C}) = \Gamma_{0,b_0}(x + W)$, and so the quantity b_0 characterized by (30) indeed provides the solution to the problem (28). Finally, [16] considers a cost that has an additional term associated with $\tilde{L}$, and consequently obtains an equation that involves two free boundary points rather than one. The fact that in the present setting (with no cost for $\tilde{L}$) the free boundary condition can be replaced by the boundary condition at 0, as we have written it in (30), is proved in [8, Proposition 2.2].

4.1.3 Results

The following result relates the $\mu \to \infty$ asymptotics of V^{μ}_{DOP} to the much simpler object V_{HT}.

Proposition 4.1 *Recall that $p = 1 - q$. One has*

$$\lim_{\mu\to\infty} V^{\mu}_{DOP}[x; c_1, c_2] = V_{HT}\left[x; c_1, \frac{c_2}{p}\right].$$

Based on the above result and the convergence result from Sect. 3, we can show an asymptotic optimality result for the queuing model at the diffusion-scaled limit. To this end consider a cost defined analogously to (25), namely

$$J^{n,\mu,b} = E\int_0^{\infty} e^{-\alpha t}\left(c_1\hat{X}^n(t) + c_2\hat{C}^n(t)\right)\mathrm{d}t, \tag{31}$$

where $(\hat{X}^n, \hat{C}^n)$ is the diffusion-scaled processes defined in Sect. 2. Throughout this section we assume that the retrial queue starts empty, that is, $\hat{R}^n(0) = 0$. In addition, for Theorem 4.2, we also assume that the initial renewal processes $A(t)$ and $S(t)$ have $6 + \varepsilon$ moment, for some small $\varepsilon > 0$. In other words, we assume the RVs $IA(1)$ *and* $ST(1)$ have a finite $6 + \varepsilon$ moment. Let the corresponding value be defined by

$$V^{n,\mu} = \inf_b J^{n,\mu,b}. \tag{32}$$

Theorem 4.2 *Assume the RVs $IA(1)$ and $ST(1)$ have a finite $6 + \varepsilon$ moment, for some $\varepsilon > 0$. One has*

$$\liminf_{\mu\to\infty}\liminf_{n\to\infty} V^{n,\mu} = \limsup_{\mu\to\infty}\limsup_{n\to\infty} V^{n,\mu} = V_{HT}\left[x; c_1, \frac{c_2}{p}\right].$$

Moreover, $b_{HT}[c_1, \frac{c_2}{p}]$ is an AO scaled buffer size, in the sense that, with $b = b_{HT}[c_1, \frac{c_2}{p}]$,

$$\liminf_{\mu\to\infty}\liminf_{n\to\infty} J^{n,\mu,b} = \limsup_{\mu\to\infty}\limsup_{n\to\infty} J^{n,\mu,b} = V_{HT}\left[x; c_1, \frac{c_2}{p}\right].$$

Next we state a result on the DOP at the $\mu \to 0$ limit and its queuing model counterpart.

Proposition 4.2 *One has*

$$\lim_{\mu \to 0} V^{\mu}_{DOP}[x; c_1, c_2] = V_{HT}[x; c_1, c_2].$$

Theorem 4.3 *One has*

$$\liminf_{\mu \to 0} \liminf_{n \to \infty} V^{n,\mu} = \limsup_{\mu \to 0} \limsup_{n \to \infty} V^{n,\mu} = V_{HT}[x; c_1, c_2],$$

and $b_{HT}[c_1, c_2]$ *is an AO scaled buffer size, namely, with* $b = b_{HT}[c_1, c_2]$,

$$\liminf_{\mu \to 0} \liminf_{n \to \infty} J^{n,\mu,b} = \limsup_{\mu \to 0} \limsup_{n \to \infty} J^{n,\mu,b} = V_{HT}[x; c_1, c_2].$$

Proposition 4.1 and Theorem 4.2 are proved in Sect. 4.2, and Proposition 4.2 and Theorem 4.3 are proved in Sect. 4.3.

Remark 4.3 Both Theorems 4.2 and 4.3 relate the model to the Harrison–Taksar problem. An intuitive explanation is as follows: For small μ, rejected customers take a long time to return to the main station. The limiting case corresponds to a model where rejected customers leave and do not come back. Thus Theorem 4.3 merely expresses continuity at $\mu = 0$.

As for the case where μ is large, a rejected customer returns very quickly. If the return time is so short that the system's state does not vary much, it will again be rejected, and thus enter the retrial queue with probability q. Iterating this argument shows that, for large μ, each initial rejection of a customer leads, on average, to $1 + q + q^2 + \cdots = p^{-1}$ rejections. Heuristically, this situation is similar to that of a model without retrials, where the cost associated with rejection is multiplied by p^{-1}. This explains the version with c_2/p in place of c_2 in Theorem 4.2.

4.2 Proofs: the limit as $\mu \to \infty$

First we write some useful relations that follow from the DOP setting. By the first and second equations in (24),

$$X(t) = x + W(t) + \int_0^t \mu R(s)\mathrm{d}s + L(t) - C(t), \tag{33}$$

where W is a $(\hat{y}, \sigma^2)$-BM. A simple manipulation gives $X(t) = x + W(t) - R(t) - pC(t) + L(t)$. Moreover, the solution of the integral equation for R in (24), in terms of C, is given by $R(t) = qC(t) - \mu \int_0^t qC(s)e^{-\mu(t-s)}\mathrm{d}s$. Thus (24) implies the following

relations used in the sequel:

$$\begin{cases} X(t) = x + W(t) - R(t) - pC(t) + L(t), \\ R(t) = qC(t) - \mu \int_0^t qC(s)e^{-\mu(t-s)}\mathrm{d}s. \end{cases} \tag{34}$$

In the sequel, we use the notation $\Delta\xi_{[\sigma,\tau]}$ for $\xi(\tau) - \xi(\sigma)$ to denote the increment of any process ξ over the time interval $[\sigma, \tau]$.

To prove the theorem we use the following lemma.

Lemma 4.1 *For fixed* $0 < b_1 < b_2 < \infty$,

$$\lim_{\mu\to\infty} \sup_{b\in[b_1,b_2]} E \int_0^\infty e^{-\alpha t} \|R^{\mu,b}\|_t \mathrm{d}t = 0. \tag{35}$$

Proof Fix b_1 and b_2 and consider $b \in [b_1, b_2]$. For any $f : \mathbb{R}_+ \to \mathbb{R}$ and $0 < \delta \le t$, write

$$\theta_t(f, \delta) = \inf\{u : |f(s_1) - f(s_2)| \ge \delta,\ 0 \le s_1 \le s_2 \le t,\ s_2 - s_1 = u\}.$$

Step 1 A uniform bound on $C^{\mu,b}(t)$.

We provide a bound on $C^{\mu,b}$ by a certain process that does not depend on μ and b. The dependence of X, C, L on μ and b is suppressed in what follows. By (34) and the positivity of X and R,

$$pC_t \le x + W_t + L_t. \tag{36}$$

Below we provide a bound on L and then use (36) to translate it into a bound on C.

The analysis we provide is pathwise. We fix ω and consider the path $(X, L, C, W) = (X, L, C, W)(\omega)$. An interval $[\sigma, \tau]$ is said to be *admissible* if $0 \le \sigma \le \tau < \infty$, $X(\sigma) = X(\tau) = 0$ and $C(\sigma) = C(\tau)$. An admissible interval is said to be *maximal* if there exists no admissible interval such that $[\sigma, \tau]$ is a proper subset of it. Our argument relies on bounding the number of maximal admissible intervals within $[0, t]$.

Let $[\sigma, \tau]$ be a maximal admissible interval. Using (33) and the monotonicity of the integral term, the increment of L over that interval satisfies

$$\Delta L_{[\sigma,\tau]} \le -\Delta W_{[\sigma,\tau]}. \tag{37}$$

For $t > 0$, denote by N_t the number of maximal admissible intervals $[\sigma, \tau] \subset [0, t]$. On each of them the increment of L is bounded by $2\|W\|_t$. Hence $L_t \le 2N_t\|W\|_t$. Next, recalling that C increases only when $X = b$, it follows that between any two consecutive maximal admissible intervals $[\sigma, \tau]$ and $[\tilde\sigma, \tilde\tau]$, X must reach the value b, for otherwise $[\sigma, \tilde\tau]$ would be an admissible interval, which stands in contradiction to the maximality of $[\sigma, \tau]$ and $[\tilde\sigma, \tilde\tau]$. Thus one has $X(\tau) = 0$, $X(\eta) = b$ for some $\eta \in [\tau, \tilde\sigma]$, and again $X(\tilde\sigma) = 0$. Let $\tilde\eta$ denote the last time when $X = b$ during the interval $[\tau, \tilde\sigma]$. Then on the interval $[\tilde\eta, \tilde\sigma]$, both L and C are flat, and so by (33),

$$-b = \Delta X_{[\tilde\eta,\tilde\sigma]} = \Delta W_{[\tilde\eta,\tilde\sigma]} + \int_{\tilde\eta}^{\tilde\sigma} \mu R(s)\mathrm{d}s \ge \Delta W_{[\tilde\eta,\tilde\sigma]}.$$

Hence $|\Delta W_{[\tilde{\eta},\tilde{\sigma}]}| \geq b \geq b_1$. As a result, the maximal admissible intervals are separated by at least $\theta_t(W, b_1)$. This shows $\theta_t(W, b_1)(N_t - 1) \leq t$. Thus, by (36),

$$\begin{aligned} pC(t) &\leq x + W(t) + L(t) \\ &\leq x + W(t) + 2N_t \|W\|_t \\ &\leq x + W(t) + 2\left(1 + \frac{t}{\theta_t(W, b_1)}\right)\|W\|_t. \end{aligned}$$

Thus

$$C(t) = C^{\mu,b}(t) \leq \zeta(t) := \frac{x}{p} + \frac{3}{p}\left(1 + \frac{t}{\theta_t(W, b_1)}\right)\|W\|_t. \tag{38}$$

This gives a bound on $C(t)$ by the process ζ defined above, which does not depend on μ and $b \in [b_1, \infty)$ (but depends on b_1).

Step 2 Show that there exists a constant $c_0 > 0$ *such that, for all sufficiently small* $\varepsilon_0 > 0$,

$$E\|R\|_t \leq \varepsilon_0 + \int_{\varepsilon_0}^{\infty} P(\tilde{\zeta}(t) \geq \varepsilon\mu)\mathrm{d}\varepsilon, \ \textit{where}\ \tilde{\zeta}(t) = \frac{2\zeta(t)}{\theta_t(W, c_0\varepsilon_0)}. \tag{39}$$

Fix t. Consider the event $\{\|R\|_t \geq \varepsilon\}$. On this event, consider the random times $\tau = \inf\{s : R_s \geq \varepsilon\}$ and $\sigma = \sup\{s < \tau : R_s \leq \varepsilon/2\}$ (recall $R(0) = 0$). It follows from the expression (34) for R that it is impossible for C to be flat in a neighborhood $(\tau - \delta, \tau + \delta)$ of τ, because R is strictly decreasing on any interval on which C is flat and nonzero. A similar statement is valid for σ. Hence these two times are points of increase in C and thus one must have $X_\sigma = X_\tau = b$ on this event. We have $\Delta R_{[\sigma,\tau]} = \varepsilon/2$, $\Delta X_{[\sigma,\tau]} = 0$ and, by (34),

$$0 = \Delta W_{[\sigma,\tau]} - \Delta R_{[\sigma,\tau]} - p\Delta C_{[\sigma,\tau]} + \Delta L_{[\sigma,\tau]}.$$

Also,

$$\Delta R_{[\sigma,\tau]} \leq q\Delta C_{[\sigma,\tau]}.$$

Hence $\Delta W_{[\sigma,\tau]} + \Delta L_{[\sigma,\tau]} \geq 2c_0\Delta R_{[\sigma,\tau]} = c_0\varepsilon$, where $c_0 = (1 + pq^{-1})/2$.

In the case $\Delta L_{[\sigma,\tau]} = 0$, the above argument shows that $\theta_t(W, c_0\varepsilon) \leq \tau - \sigma$. In the case $\Delta L_{[\sigma,\tau]} > 0$, it follows that X hits zero some time in the interval $[\sigma, \tau]$, hence, arguing as in the previous step, W makes a displacement of $b \geq b_1$ in this interval. Hence $\theta_t(W, b_1) \leq \tau - \sigma$. Combining the two cases, we have $\theta_t(W, c_0\varepsilon \wedge b_1) \leq \tau - \sigma$.

Moreover, on the interval $[\sigma, \tau]$, $R \geq \varepsilon/2$. Hence $\int_0^t R_s\mathrm{d}s \geq (\varepsilon/2)\theta_t(W, c_0\varepsilon \wedge b_1)$. Using the last part of (24) and then the first step,

$$\int_0^t R_s\mathrm{d}s \leq \mu^{-1}C_t \leq \mu^{-1}\zeta(t).$$

These two inequalities imply $\zeta(t) \geq \frac{\mu\varepsilon}{2}\theta_t(W, c_0\varepsilon \wedge b_1)$. Hence

$$\begin{aligned} E\|R\|_t &= \int_0^\infty P\left(\|R\|_t \geq \varepsilon\right) \mathrm{d}\varepsilon \\ &\leq \int_0^\infty P\left(\zeta(t) \geq \frac{\mu\varepsilon}{2}\theta_t(W, c_0\varepsilon \wedge b_1)\right) \mathrm{d}\varepsilon \\ &\leq \varepsilon_0 + \int_{\varepsilon_0}^\infty P\left(\zeta(t) \geq \frac{\mu\varepsilon}{2}\theta_t(W, c_0\varepsilon \wedge b_1)\right) \mathrm{d}\varepsilon. \end{aligned}$$

The above is true for any choice of $\varepsilon_0 > 0$. Consider ε_0 for which $c_0\varepsilon_0 < b_1$. Then for any $\varepsilon \in [\varepsilon_0, \infty)$ one has $c_0\varepsilon_0 \wedge b_1 \geq c_0\varepsilon_0$, hence the integral in the above display is bounded from above by the integral $\int_{\varepsilon_0}^\infty P(\zeta(t) \geq \frac{\mu\varepsilon}{2}\theta_t(W, c_0\varepsilon))\mathrm{d}\varepsilon$. This equals $\int_{\varepsilon_0}^\infty P(\tilde{\zeta}(t) \geq \varepsilon\mu)\mathrm{d}\varepsilon$. This shows (39).

Step 3 An estimate of $\tilde{\zeta}$.

We show that, given ε_0, there exists a function $u(t)$, for $t \geq 0$, such that

$$c_u := \int_0^\infty u(t)e^{-\alpha t}\mathrm{d}t < \infty \quad \text{and} \quad E[\tilde{\zeta}(t)] \leq u(t). \tag{40}$$

The function u may depend on ε_0 but not on μ or b. This suffices in order to deduce the result by the following calculation. By (39) we have

$$E\|R\|_t \leq \varepsilon_0 + \frac{E[\tilde{\zeta}(t)]}{\mu} \leq \varepsilon_0 \frac{u(t)}{\mu},$$

hence

$$E\int_0^\infty e^{-\alpha t}\|R\|_t \mathrm{d}t \leq \int_0^\infty e^{-\alpha t}[\varepsilon_0 + \mu^{-1}u(t)]\mathrm{d}t = \alpha^{-1}\varepsilon_0 + \mu^{-1}c_u.$$

Moreover, since u does not depend on μ or b (as long as it lies in $[b_1, \infty)$), we have

$$\sup_{b\in[b_1,\infty)} E\int_0^\infty e^{-\alpha t}\|R^{\mu,b}\|_t \mathrm{d}t \leq \alpha^{-1}\varepsilon_0 + \mu^{-1}c_u.$$

Taking now $\mu \to \infty$ then $\varepsilon_0 \to 0$ proves the result.

To prove (40), we note that

$$E\tilde{\zeta}(t) \leq C\left(E\frac{1}{\theta_t(W, c_0\varepsilon_0)} + E\frac{\|W\|_t}{\theta_t(W, c_0\varepsilon_0)} + E\frac{\|W\|_t}{\theta_t(W, c_0\varepsilon_0)\theta_t(W, b_1)}\right)$$

for some constant C. Using the Cauchy–Schwartz and Hölder inequalities, the second and third terms above are bounded by

$$\{E[\|W\|_t^2]\}^{1/2}\{E[\theta_t(W, c_0\varepsilon_0)^{-2}]\}^{1/2},$$
$$\{E[\|W\|_t^3]\}^{1/3}\{E[\theta_t(W, c_0\varepsilon_0)^{-3}]\}^{1/3}\{E[\theta_t(W, b_1)^{-3}]\}^{1/3},$$

respectively. Now, W is a $(\hat{y}, \sigma^2)$-BM. In the special case where $\hat{y} = 0$, it is a martingale, and thus Burkholder's inequality applies, giving, for $a \geq 2$, $E[\|W\|_t^a] \leq c_a t^{a/2}$. From this it follows for general $\hat{y}$ that $E[\|W\|_t^a] \leq \tilde{c}_a[t^{a/2} + t^a]$. As a result, it suffices to show that, given $\delta > 0$ and $\beta \geq 1$,

$$E[\theta_t(W, \delta)^{-\beta}] \leq C_1 t + C_2, \tag{41}$$

where C_1 and C_2 do not depend on t (but may depend on δ and β).

In what follows we prove (41). First, for any $r_0 > 0$ we have

$$\begin{aligned} E[\theta_t(W, \delta)^{-\beta}] &\leq r_0 + \int_{r_0}^{\infty} P\left(\theta_t(W, \delta)^{-\beta} > r\right) \mathrm{d}r \\ &= r_0 + \int_{r_0}^{\infty} P\left(\theta_t(W, \delta) < r^{-1/\beta}\right) \mathrm{d}r. \end{aligned} \tag{42}$$

Given t and s, we provide an estimate for $P(\theta_t(W, \delta) < s)$. Consider the n_0 subintervals $[t_i, t_{i+1}]$ of $[0, t]$, where $t_i = it/n_0$, $i = 0, 1, \ldots, n_0 - 1$. Denote by Δ_i the oscillation of W within the ith interval, namely $\Delta_i = \sup_{u,v\in[t_i,t_{i+1}]} |W_u - W_v|$. Assume $\frac{t}{n_0} \geq \frac{s}{2}$. Then it follows from the definition of θ that on the event $\{\theta_t(W, \delta) < s\}$ there exists $i \leq n_0 - 1$ such that $\Delta_i \geq \frac{\delta}{3}$. Setting $n_0 = [\frac{2t}{s}]$ and using this argument along with the union bound shows

$$\begin{aligned} &P(\theta_t(W, \delta) < s) \leq n_0 P\left(\Delta_0 \geq \frac{\delta}{3}\right) \leq \\ &n_0 P\left(\max_{v\leq t/n_0} W_v \geq \frac{\delta}{6}\right) + n_0 P\left(\max_{v\leq t/n_0} (-W_v) \geq \frac{\delta}{6}\right). \end{aligned} \tag{43}$$

Denote the centered version of the BM W by $\tilde{W}_t := W_t - \hat{y}t$. It is known [17] that $\Lambda_u := \max_{v\in[0,u]} \tilde{W}_v - \tilde{W}_u$ is equal in distribution to $|\tilde{W}_u|$. Now,

$$\Lambda_u = \max_{v\in[0,u]} (W_v - \hat{y}v) - \tilde{W}_u \geq \max_{v\in[0,u]} W_v - \tilde{W}_u - |\hat{y}|u.$$

Hence

$$P\left(\max_{v\leq t/n_0} W_v \geq \frac{\delta}{6}\right) \leq P\left(\Lambda_{t/n_0} \geq \frac{\delta}{18}\right) + P\left(\tilde{W}_{t/n_0} \geq \frac{\delta}{18}\right) + 1_{\{|\hat{y}|\frac{t}{n_0} \geq \frac{\delta}{18}\}}.$$

Using this in (43) along with the same estimate for $\max_{v\le t/n_0}(-W_v)$, we obtain

$$P(\theta_t(W,\delta) < s) \le 4n_0 P\left(|\tilde{W}_{t/n_0}| \ge \frac{\delta}{18}\right),$$

provided that

$$|\hat{y}|\frac{t}{n_0} < \frac{\delta}{18}. \tag{44}$$

Now, $\tilde{W}_{t/n_0}$ is a normal RV with mean zero and variance $t\sigma^2/n_0$. If N is a $(0,\bar{\sigma}^2)$-normal RV then $P(N > A) \le \frac{\bar{\sigma}}{\sqrt{2\pi}A} e^{-A^2/2\bar{\sigma}^2} \le e^{-A^2/2\bar{\sigma}^2}$, provided $A > \bar{\sigma}$. Hence

$$P\left(|\tilde{W}_{t/n_0}| \ge \frac{\delta}{18}\right) \le 2e^{-\frac{\delta^2 n_0}{648t\sigma^2}}, \tag{45}$$

provided

$$324\sigma\frac{t}{n_0} < \delta^2. \tag{46}$$

Using $\frac{2t}{s} \le n_0 \le \frac{2t}{s} + 1$ we obtain

$$P(\theta_t(W,\delta) < s) \le 8\left(\frac{2t}{s} + 1\right) e^{-\frac{\delta^2}{324\sigma^2 s}}. \tag{47}$$

Moreover, both (44) and (46) hold when $s < s_0$, where s_0 is a constant that depends only on δ and the parameters $\hat{y}$ and σ (not on t). Thus, if we consider the integrand of (42) then we can use the above estimate with $r^{-1/\beta} = s$, and the condition $s < s_0$ is assured to hold if one lets $r_0 = s_0^{-\beta}$. Again, r_0 thus selected depends on δ, $\hat{y}$, σ and β but not on t. The estimate then gives

$$E[\theta_t(W,\delta)^{-\beta}] \le r_0 + \int_{r_0}^{\infty} 8(2tr^{1/\beta} + 1)e^{-\frac{\delta^2}{324\sigma^2}r^{1/\beta}}\,\mathrm{d}r \le C_1 t + C_2.$$

This shows (41). The result follows. □

Lemma 4.2 *One has* $\inf_{\mu\in(0,\infty)} C^{\mu,b}(1) \to \infty$ *a.s., as* $b \to 0$. *Consequently,*

$$\lim_{b\to 0} \inf_{\mu\in(0,\infty)} J^{b,\mu}_{DOP}[x; c_1, c_2] = \infty.$$

Proof Let μ and b be given, and let $\varepsilon > b$. It follows directly from (24) that if, for some $0 \le s < t$, one has $W(t) - W(s) > \varepsilon - b$, then $C^{\mu,b}(t) - C^{\mu,b}(s) > \varepsilon - b$. Let $t_i = \varepsilon^2 i$ for $i = 0, 1, \ldots, n_\varepsilon$, where $n_\varepsilon = [1/\varepsilon^2]$. Let $N(\varepsilon) = \#\{i : \delta_i > \varepsilon, i = 1, \ldots, n_\varepsilon\}$, where $\delta_i = W_{t_i} - W_{t_{i-1}}$. It follows that $C^{\mu,b}(1) \ge (\varepsilon - b)N(\varepsilon)$. Now, $\delta_i \sim \mathcal{N}(\varepsilon^2\hat{y}, \varepsilon^2)$, and so by the LLN and Brownian scaling, $N(\varepsilon)/n_\varepsilon \to p_0$ a.s. as $\varepsilon \to 0$, where

$p_0 = \mathbb{P}(\mathcal{N}(0,1) > 1) > 0$. Selecting $\varepsilon = \varepsilon(b) = b^{1/2}$, we obtain

$$\inf_{\mu\in(0,\infty)} C^{\mu,b}(1) \geq (\varepsilon(b) - b) n_{\varepsilon(b)} \frac{N(\varepsilon(b))}{n_{\varepsilon(b)}} \geq \frac{1}{2} \frac{b^{1/2}}{b} \frac{N(\varepsilon(b))}{n_{\varepsilon(b)}},$$

and the RHS above converges to ∞ a.s. □

Lemma 4.3 *Assume the RVs IA(1) and ST(1) have a finite $6+\varepsilon$ moment, for some small $\varepsilon > 0$. For every $k > 0$ there exists $b_1 > 0$ and n_1 such that, if $b \in (0, b_1)$ and $n > n_1$, then, for all μ,*

$$J^{n,\mu,b} \geq k.$$

Moreover, fix $\mu \in (0,\infty)$. Then given $\varepsilon > 0$ there exist $b_2 \in (0,\infty)$ and n_0 such that, if $b > b_2$ and $n > n_0$, one has

$$J^{n,\mu,b} \geq J^{n,\mu,b_2} - \varepsilon.$$

Proof For the first assertion, recall Eq. (14), by which $(\hat{X}^n, \hat{L}^n, \hat{C}^n) = \Gamma_{0,\hat{b}^n}(\hat{Z}^n)$. Also, recall from (18) that $\tilde{Z}^n(t) = \xi^n(t) + \int_0^t \mu \hat{R}^n(s)\mathrm{d}s$, where $\tilde{Z}^n$ and $\hat{Z}^n$ have the same weak limit. Given $\varepsilon > 0$ and $b < \varepsilon/2$, we have $\hat{b}^n < \varepsilon/2$ for all large n. Write $\mathrm{osc}^+(f;s,t) = \sup_{s\leq v<u\leq t}(f(u) - f(v))$ and $t_i = i\varepsilon^2$ for $i = 0, 1, 2, \ldots, [\varepsilon^{-2}] =: N_\varepsilon$. Recalling that ξ^n converges weakly to a (nondegenerate) BM, and using Brownian scaling, it follows that for all sufficiently large n one has

$$E\sum_{i=1}^{N_\varepsilon} 1_{\{\mathrm{osc}^+(\xi^n;t_{i-1},t_i)>\varepsilon\}} > c\varepsilon^{-2},$$

where $c > 0$ does not depend on ε. Now, it is a property of the Skorohod map that on the event $\mathrm{osc}^+(\xi^n; t_{i-1}, t_i) > \varepsilon$, one has $\hat{C}^n(t_i) - \hat{C}^n(t_{i-1}) > \varepsilon - \hat{b}_n > \varepsilon/2$, regardless of the value of μ. This shows that $E\hat{C}^n(1) > c\varepsilon^{-1}/2$ (for all sufficiently large n, all $b < \varepsilon/2$ and all μ). Since ε is arbitrary, we can achieve $J^{n,\mu,b} \geq k$ for any given k, by selecting sufficiently small ε.

For the second assertion, it suffices to prove that, for fixed μ,

$$\lim_{b_2\to\infty} \limsup_{n\to\infty} \sup_{(b,b')\in(b_2,\infty)^2} |J^{n,\mu,b} - J^{n,\mu,b'}| = 0. \tag{48}$$

It is argued below that, for small enough (fixed) $\varepsilon > 0$, for some $p^n(t)$ that satisfies, for $n > n_0$ and all t, $p^n(t) \leq p(t)$, where p is a polynomial in t, one has

$$\sup_{b>1} E[\hat{C}^{n,\mu,b}(t)^{1+\varepsilon}] \vee E[\hat{X}^{n,\mu,b}(t)^{1+\varepsilon}] \leq p^n(t). \tag{49}$$

Let $\tau^n(b) = \inf\{t : \hat{C}^{n,\mu,b} > 0\}$. Then for every $b \geq b_2$ one has $(\hat{X}^{n,\mu,b}, \hat{C}^{n,\mu,b})(t) = (\hat{X}^{n,\mu,b_2}, \hat{C}^{n,\mu,b_2})(t)$ for all $t \in [0, \tau^n(b_2))$. Hence we can write, for any T and any $(b, b') \in (b_2, \infty)^2$,

$$
\begin{aligned}
&E \int_0^\infty e^{-\alpha t} |\hat{C}^{n,\mu,b} - \hat{C}^{n,\mu,b'}| \mathrm{d}t \\
&\quad \leq E\left[1_{\{\tau^n(b_2) < T\}} \int_0^T e^{-\alpha t} |\hat{C}^{n,\mu,b} - \hat{C}^{n,\mu,b'}| \mathrm{d}t \right] + 2 \int_T^\infty e^{-\alpha t} p^n(t) \mathrm{d}t \\
&\quad \leq P(\tau^n(b_2) < T)^{(1+\varepsilon)/\varepsilon} \sup_{b > b_2} \int_0^T e^{-\alpha t} [E(\hat{C}^{n,\mu,b})^{1+\varepsilon}]^{1/(1+\varepsilon)} \mathrm{d}t + 2 \int_T^\infty e^{-\alpha t} p^n(t) \mathrm{d}t \\
&\quad \leq P(\tau^n(b_2) < T)^{(1+\varepsilon)/\varepsilon} \int_0^T e^{-\alpha t} p^n(t)^{1/(1+\varepsilon)} \mathrm{d}t + 2 \int_T^\infty e^{-\alpha t} p^n(t) \mathrm{d}t \\
&\quad \leq c P(\tau^n(b_2) < T)^{(1+\varepsilon)/\varepsilon} + c \int_T^\infty e^{-\alpha t} p(t) \mathrm{d}t,
\end{aligned}
$$

where c is a finite constant. By similar considerations, the expression on the last line is an upper bound on $E \int_0^\infty e^{-\alpha t} |\hat{X}^{n,\mu,b} - \hat{X}^{n,\mu,b'}| \mathrm{d}t$. Thus

$$
\sup_{(b,b') \in (b_2,\infty)} |J^{n,\mu,b} - J^{n,\mu,b'}| \leq c P(\tau^n(b_2) < T)^{(1+\varepsilon)/\varepsilon} + c \int_T^\infty e^{-\alpha t} p(t) \mathrm{d}t. \tag{50}
$$

Now, for $t < \tau^n(b_2)$ and all $b > b_2$, equations (12) are valid with $\hat{R}^n(t) = 0$. It is routine to obtain from these equations that, given any T, $\lim_{b_2 \to \infty} \limsup_{n \to \infty} P(\tau^n(b_2) < T) = 0$. Thus, if we take $n \to \infty$ in (50) first, then $b_2 \to \infty$ and finally $T \to \infty$, we obtain (48).

It remains to show (49). The argument uses some of the ideas from the proof of Lemma 4.1.

We use a general upper bound on moments of a centered renewal process from [18, Th. 4]. This result states that if H is a renewal process with inter-arrival distribution that has reciprocal mean λ and possesses a finite m-th moment, for some $m \geq 2$, and one defines the centered diffusion-scaled process $\hat{H}^n(t) = n^{-1/2}(H(nt) - \lambda n t)$, then

$$
E[\|\hat{H}^n\|_t^m] \leq c(1 + t^{m/2}), \tag{51}
$$

where c does not depend on n or t. We can rescale time according to $t = n^a s$, set $N = n^{1+a}$, and then send $a \to \infty$ to obtain from (51) the bound

$$
E[\|\hat{H}^N\|_s^m] \leq c(N^{-m/2} + s^{m/2}), \tag{52}
$$

where, again, c does not depend on N or s. To apply this bound to $\hat{A}^n$ (equivalently, to $\hat{S}^n$), note that $\hat{A}^n(t) = n^{-1/2}(A(\lambda_n t) - \lambda_n t)$ can be written as $c_n \lambda_n^{-1/2} A(\lambda_n t) - \lambda_n t$, where $c_n = (\lambda_n / n)^{1/2}$. Since c_n converges to a positive constant, it follows that the moments of $\hat{A}^n(t)$ admit a similar bound. That is,

$$E[\|\hat{A}^n\|_t^m] \le c(n^{-m/2} + t^{m/2}). \tag{53}$$

First, it follows from the definition of $\hat{A}_2^n$, (2), (3) and (10) that $\hat{C}^n(t) - \hat{A}_2^n(t) = p\hat{C}^n(t) + e_2^n(t)$, where $e_2^n(t) = n^{-1/2}\sum_{i=1}^{C_t}\tilde{\xi}_i$, $\tilde{\xi}_i = -\xi_i + q$. We have $E\tilde{\xi}_i = 0$. The process e_2^n is a martingale and from Burkholder's inequality we have $E[\|e_2^n\|_t^2] \le cE[e_2^n, e_2^n]_t$. The jumps of the process e_2^n are of size $n^{-1/2}$ and their number is bounded by C_t^n, hence the quadratic variation process $[e_2^n, e_2^n]_t$ is bounded by $n^{-1}C(t)$. Thus

$$E[\|e_2^n\|_t^2] \le cn^{-1}EC^n(t) = cn^{-1/2}E\hat{C}^n(t).$$

Let $\tilde{\xi}^n(t) = \hat{X}^n(0) + \hat{y}^n t + \hat{A}^n(t) - \hat{D}_1^n(t)$ and recall the notation ξ^n from (18), by which we have $\xi^n = \tilde{\xi}^n + e_1^n$. By summing both parts of (12), using $\hat{R}^n(0) = 0 \le \hat{R}^n(t)$,

$$\hat{X}^n(t) + p\hat{C}^n(t) \le \tilde{\xi}^n(t) + \hat{L}^n(t) + e_2^n(t). \tag{54}$$

This bound may be viewed as an analogue of (36) in the proof of Lemma 4.1. We provide a bound on $\hat{L}^n$ by the ideas from that proof. The argument that leads to (37) over a maximal admissible interval $[\sigma, \tau]$ (defined similarly) gives $\Delta\hat{L}^n_{[\sigma,\tau]} \le -\Delta\tilde{\xi}^n|_{[\sigma,\tau]} + \Delta e^n|_{[\sigma,\tau]}$. Along the lines leading to the bound (38) we then obtain

$$\hat{X}^n(t) + \hat{C}^n(t) \le c + c\|\tilde{\xi}^n\|_t + c\frac{t\|\xi^n\|_t}{\theta_t(\xi^n, b)}. \tag{55}$$

Now, given $\varepsilon > 0$,

$$E\left[\frac{\|\xi^n\|_t^{1+\varepsilon}}{\theta_t(\xi^n, b)^{1+\varepsilon}}\right] \le E[\|\xi^n\|_t^{2+2\varepsilon}]^{1/2}E[\theta_t(\xi^n, b)^{-2-2\varepsilon}]^{1/2}. \tag{56}$$

To bound the second factor above, use a variation of (42):

$$E[\theta_t(\xi^n, b)^{-\beta}] \le 1 + \int_1^\infty P(\theta_t(\xi^n, b) < r^{-1/\beta})dr. \tag{57}$$

The bound (43) can be adapted. We choose n_0 the same way as in the proof of Lemma 4.1. Clearly (45) is not valid, as the tails do not behave as a Gaussian RV. We replace this estimate by the following bound: In view of (53) and the same bound for the process $\hat{S}^n$, we have

$$E[\|\tilde{\xi}^n(\cdot) - \hat{X}^n(0)\|_t^m] \le c(n^{-m/2} + t^{-m/2}).$$

Hence,

$$P(|\tilde{\xi}^n(t/n_0) - \hat{X}^n(0)| > b) \le P(|\tilde{\xi}^n(t/n_0) - \hat{X}^n(0)| > b) \le b^{-m}(n^{-m/2} + t^{m/2}n_0^{-m/2}).$$

Again, using $\frac{2t}{s} \le n_0 \le \frac{2t}{s} + 1$ gives

$$P(\theta_t(\xi^n, b) < s) \le 8\left(\frac{2t}{s} + 1\right) b^{-m} t^{m/2} n_0^{-m/2} \le cb^{-m} \frac{t}{s} s^{m/2} = cts^{\frac{m}{2}-1}.$$

As before, choosing $s = r^{-1/\beta}$ and $m/2 - 1 = m'$ gives

$$1 + ct \int_1^\infty r^{-m'/\beta} \mathrm{d}r$$

as an upper bound on (57). The exponent $-m'/\beta$ is required to be less than -1 so that $r^{-m'/\beta}$ is integrable, where $\beta = 2 + 2\varepsilon$. This can be achieved by taking $m = 6 + \delta$ for some $\delta > 0$, and the assumption on the finite $6 + \varepsilon$ moments of the renewal processes. Combining this bound with (55) and (56) gives that $E[\hat{X}^n(t) + \hat{C}^n(t)]$ is bounded by a polynomial independent of n. This completes the proof. □

Proof of Proposition 4.1 Clearly, the diffusion model (24) implies that

$$X(t) \in [0, b],\ t \ge 0, \qquad \int_{[0,\infty)} \mathbb{I}_{\{X_t > 0\}} \mathrm{d}L_t = 0, \qquad \int_{[0,\infty)} \mathbb{I}_{\{X_t < b\}} \mathrm{d}C_t = 0.$$

Identities (34) are also valid for the same model. If we consider the first equation in (34) with the above display, we see that the following holds:

$$(X, L, pC) = \Gamma_{0,b}[x + W - R]. \tag{58}$$

Let $b = b_{\mathrm{HT}}[c_1, \frac{c_2}{p}]$. Then $V_{\mathrm{HT}}[x; c_1, \frac{c_2}{p}] = J^b_{\mathrm{HT}}[x; c_1, \frac{c_2}{p}]$. Let (X, L, C, R) denote the processes from the DOP model (24) and $(\tilde{X}, \tilde{L}, \tilde{C})$ those from the HT model (28), where, in both cases, a buffer of size b is used. Then the former tuple satisfies (58), and the latter satisfies

$$(\tilde{X}, \tilde{L}, \tilde{C}) = \Gamma_{0,b}[x + W]. \tag{59}$$

The Lipschitz property (8) implies that for any μ

$$\|X^{\mu,b} - \tilde{X}^b\|_t + \|pC^{\mu,b} - \tilde{C}^b\|_t \le c_\Gamma \|R^{\mu,b}\|_t, \qquad t \ge 0. \tag{60}$$

Hence, by Lemma 4.1,

$$\lim_{\mu\to\infty} E \int_0^\infty e^{-\alpha t} \left(|X_t^{\mu,b} - \tilde{X}_t^b| + |pC_t^{\mu,b} - \tilde{C}_t^b|\right) \mathrm{d}t = 0. \tag{61}$$

By the definition of J_{DOP} and J_{HT} and the above convergence, it follows that

$$\lim_{\mu\to\infty} J^{\mu,b}_{\mathrm{DOP}}[x; c_1, c_2] = J^b_{\mathrm{HT}}\left[x; c_1, \frac{c_2}{p}\right]. \tag{62}$$

Hence

$$\limsup_{\mu\to\infty} V^{\mu}_{\mathrm{DOP}}[x;c_1,c_2] \le J^{b}_{\mathrm{HT}}\left[x;c_1,\frac{c_2}{p}\right] = V_{\mathrm{HT}}\left[x;c_1,\frac{c_2}{p}\right]. \tag{63}$$

For the reverse inequality, note first that as $\mu\to\infty$, the optimal b does not converge to 0. Indeed, by Lemma 4.2 and the bound (63) above, there exists $0 < b_1 < \infty$ such that, for all large μ, $V^{\mu}_{\mathrm{DOP}}[x;c_1,c_2] = \inf_{b\in[b_1,\infty)} J^{b,\mu}_{\mathrm{DOP}}[x;c_1,c_2]$.

Given $\varepsilon > 0$, let $\bar{\mu}$ be so large that (i) $V^{\bar{\mu}}_{\mathrm{DOP}} < \liminf_{\mu\to\infty} V^{\mu}_{\mathrm{DOP}} + \varepsilon$, (ii) $V^{\bar{\mu}}_{\mathrm{DOP}} > J^{\bar{b},\bar{\mu}}_{\mathrm{DOP}} - \varepsilon$ for suitable $\bar{b}\in[b_1,\infty)$, and (iii) $\sup_{b\in[b_1,\infty)} E\int_0^{\infty} e^{-\alpha t}\|R^{b,\bar{\mu}}\|_t \mathrm{d}t < \varepsilon$, where the last item is possible thanks to Lemma 4.1.

As before, let (X, L, C, R) denote the processes from the DOP model and $(\tilde{X}, \tilde{L}, \tilde{C})$ those from the HT model, where b is set to $\bar{b}$ specified above. Then (58) and (59) are valid with $b = \bar{b}$ and hence so is (60). In particular,

$$E\int_0^{\infty} e^{-\alpha t}\left(|X_t^{\bar{\mu},\bar{b}} - \tilde{X}_t^{\bar{b}}| + |pC_t^{\bar{\mu},\bar{b}} - \tilde{C}_t^{\bar{b}}|\right)\mathrm{d}t \le c_\Gamma\varepsilon.$$

As a result,

$$\left|J^{\bar{\mu},\bar{b}}_{\mathrm{DOP}}[x;c_1,c_2] - J^{\bar{b}}_{\mathrm{HT}}\left[x;c_1,\frac{c_2}{p}\right]\right| \le c_3\varepsilon,$$

where c_3 is a constant that depends only on c_1, c_2 and c_Γ. Thus

$$\begin{aligned}\liminf_{\mu\to\infty} V^{\mu}_{\mathrm{DOP}}[x;c_1,c_2] &\ge V^{\bar{\mu}}_{\mathrm{DOP}}[x;c_1,c_2] - \varepsilon \ge J^{\bar{b}}_{\mathrm{HT}}\left[x;c_1,\frac{c_2}{p}\right] - (1+c_3)\varepsilon\\ &\ge V_{\mathrm{HT}}\left[x;c_1,\frac{c_2}{p}\right] - (1+c_3)\varepsilon.\end{aligned}$$

The result follows on taking $\varepsilon\to 0$. □

Proof of Theorem 4.2 Let $\varepsilon > 0$ be given. As in the proof of Proposition 4.1, let $b = b_{\mathrm{HT}}[c_1, \frac{c_2}{p}]$. It follows from (62) that, for all μ_0 sufficiently large,

$$J^{\mu_0,b}_{\mathrm{DOP}}[x;c_1,c_2] \le V_{\mathrm{HT}}\left[x;c_1,\frac{c_2}{p}\right] + \varepsilon.$$

Fix μ_1 so large that the above inequality holds for all $\mu_0 > \mu_1$, and

$$\limsup_{\mu\to\infty}\limsup_{n\to\infty} nV^{n,\mu} < \limsup_{n\to\infty} V^{n\to\infty,\mu_0} + \varepsilon$$

for all $\mu_0 > \mu_1$.

For any n, $J^{n,\mu_0,b}$ is given by (31) as $E\int_0^{\infty} e^{-\alpha t}(c_1\hat{X}^n(t) + c_2\hat{C}^n(t))\mathrm{d}t$ and, by Theorem 3.1, $(\hat{X}^n, \hat{C}^n) \Rightarrow (X, C)$, where the latter processes correspond to the DOP model (24). By an estimate similar to the one from Step 4 in the proof of Theorem

4.1 of [8], the rescaled costs $\int_0^\infty e^{-\alpha t}(c_1\hat{X}^n(t)+c_2\hat{C}^n(t))\mathrm{d}t$ are uniformly integrable. Therefore we may deduce from the weak convergence that $J^{n,\mu_0,b} \to J^{\mu_0,b}_{\mathrm{DOP}}[x;c_1,c_2]$ as $n\to\infty$. Thus

$$\begin{aligned}\limsup_{\mu\to\infty}\limsup_{n\to\infty} V^{n,\mu} &\le \limsup_{n\to\infty} V^{n,\mu_0}+\varepsilon\\ &\le \limsup_{n\to\infty} J^{n,\mu_0,b}+\varepsilon = J^{\mu_0,b}_{\mathrm{DOP}}[x;c_1,c_2]+\varepsilon\\ &\le V_{\mathrm{HT}}\left[x;c_1,\frac{c_2}{p}\right]+2\varepsilon.\end{aligned}$$

Since $\varepsilon>0$ is arbitrary, we obtain

$$\limsup_{\mu\to\infty}\limsup_{n\to\infty} V^{n,\mu} \le \limsup_{\mu\to\infty}\limsup_{n\to\infty} J^{n,\mu,b} \le V_{\mathrm{HT}}\left[x;c_1,\frac{c_2}{p}\right]. \tag{64}$$

Next, given $\varepsilon>0$, let μ_0 be so large that (i) $V^{\mu_0}_{\mathrm{DOP}}[x;c_1,c_2]$ is ε-close to $V_{\mathrm{HT}}[x;c_1,\frac{c_2}{p}]$, and (ii) $\liminf_{\mu\to\infty}\liminf_{n\to\infty} V^{n,\mu}$ is ε-close to $\liminf_{n\to\infty} V^{n,\mu_0}$. Recall the relation (32) and let $\tilde{b}_n$ be a sequence for which

$$\liminf_{n\to\infty} J^{n,\mu_0,\tilde{b}_n} < \liminf_{n\to\infty} V^{n,\mu_0}+\varepsilon.$$

It follows from (64) and Lemma 4.2 that, for some $b_1>0$, one has $\tilde{b}_n\ge b_1>0$ for all n large. Moreover, by Lemma 4.3, we may assume without loss of generality that there exists $b_2\in(b_1,\infty)$ such that $\tilde{b}_n\le b_2$ for all n large. Hence $\tilde{b}_n$ are bounded away from zero and infinity. Consider a subsequence along which $\tilde{b}_n$ converges, and denote its limit by $b\in(0,\infty)$. Then along this subsequence we have the convergence $(\hat{X}^n,\hat{C}^n)\Rightarrow(X,C)$, where the latter is the DOP with buffer size b. In view of Remark 3.1, this is a consequence of Theorem 3.1. Hence, by Fatou's lemma, $\liminf_{n\to\infty} J^{n,\mu_0,\tilde{b}_n}\ge J^{\mu_0,b}_{\mathrm{DOP}}$. Therefore

$$\begin{aligned}\liminf_{\mu\to\infty}\liminf_{n\to\infty} V^{n,\mu} &\ge \liminf_{n\to\infty} V^{n\to\infty,\mu_0}-\varepsilon\\ &= \liminf_{n\to\infty} J^{n,\mu_0,\tilde{b}_n}-\varepsilon\\ &\ge J^{\mu_0,b}_{\mathrm{DOP}}[x;c_1,c_2]-\varepsilon\\ &\ge V^{\mu_0}_{\mathrm{DOP}}[x;c_1,c_2]-\varepsilon\\ &\ge V_{\mathrm{HT}}\left[x;c_1,\frac{c_2}{p}\right]-2\varepsilon.\end{aligned}$$

Sending $\varepsilon\to0$ and combining this bound with (64) gives the result. □

4.3 Proofs: the limit as $\mu \to 0$

Here we prove Proposition 4.2 and Theorem 4.3.

If we set $\mu = 0$ in the DOP then the DOP and the HT problem become equivalent. We show that the $\mu \to 0$ limit is identical to setting $\mu = 0$. Proposition 4.2 is therefore a continuity result for V^{μ}_{DOP}.

Since x, c_1 and c_2 are fixed, we omit these from the notation.

Lemma 4.4

$$\lim_{\mu \to 0} \sup_{b \in (0,\infty)} |J^{\mu,b}_{DOP} - J^{b}_{HT}| = 0.$$

Proof Given b and μ, let (X, L, C, R) and $(\tilde{X}, \tilde{L}, \tilde{C})$ denote the processes from the DOP model (24) and the HT model (28), respectively. It follows from (8) that

$$\|X - \tilde{X}\|_t + \|L - \tilde{L}\|_t + \|C - \tilde{C}\|_t \le c_\Gamma \mu \int_0^t R_s \mathrm{d}s, \qquad t \ge 0, \tag{65}$$

where we recall that c_Γ does not depend on b. To obtain a bound on the RHS of the above display we again use the Lipschitz property of the Skorokhod map and (24), by which

$$C_t \le c_\Gamma \left(x + \|W\|_t + \mu \int_0^t R_s \mathrm{d}s \right).$$

Also, by (24), $R_t \le C_t$ for all $t \ge 0$. Hence $C_t \le c_\Gamma (x + \|W\|_t + \mu \int_0^t C_s \mathrm{d}s)$, and so, by Gronwall's lemma,

$$R_t \le C_t \le c_\Gamma (x + \|W\|_t) e^{c_\Gamma \mu t}.$$

As a result,

$$E \int_0^\infty e^{-\alpha t} (\|X - \tilde{X}\|_t + \|C - \tilde{C}\|_t) \mathrm{d}t \le \gamma(\mu) := E \int_0^\infty c_\Gamma^2 \mu t (x + \|W\|_t) e^{(c_\Gamma \mu - \alpha)t} \mathrm{d}t.$$

Note that $\gamma(\mu)$ does not depend on b. Hence

$$\sup_{b \in (0,\infty)} |J^{\mu,b}_{\mathrm{DOP}} - J^{b}_{\mathrm{HT}}| \le (c_1 + c_2)\gamma(\mu).$$

Since $\gamma(\mu) \to 0$ as $\mu \to 0$, the result follows. □

Proof of Proposition 4.2 Let $b = b_{\mathrm{HT}}[c_1, c_2]$. Then, by Lemma 4.4,

$$\limsup_{\mu \to 0} V^{\mu}_{\mathrm{DOP}} \le \limsup_{\mu \to 0} J^{\mu,b}_{\mathrm{DOP}} = J^{b}_{\mathrm{HT}} = V_{\mathrm{HT}}.$$

Next, for a lower bound on $\liminf_{\mu\to 0} V^{\mu}_{\mathrm{DOP}}$, let $\varepsilon > 0$ be given. Fix $\mu_0 > 0$ be so small that (i) $V^{\mu_0}_{\mathrm{DOP}} < \liminf_{\mu\to 0} V^{\mu}_{\mathrm{DOP}} + \varepsilon$, and (ii) $\sup_{b\in(0,\infty)} |J^{\mu_0,b}_{\mathrm{DOP}} - J^{b}_{\mathrm{HT}}| < \varepsilon$, where in (ii) we used Lemma 4.4. Let b_0 be chosen so that $V^{\mu_0}_{\mathrm{DOP}} > J^{\mu_0,b_0}_{\mathrm{DOP}} - \varepsilon$. Then

$$\liminf_{\mu\to 0} V^{\mu}_{\mathrm{DOP}} \geq V^{\mu_0}_{\mathrm{DOP}} - \varepsilon \geq J^{\mu_0,b_0}_{\mathrm{DOP}} - 2\varepsilon \geq J^{b_0}_{\mathrm{HT}} - 3\varepsilon \geq V_{\mathrm{HT}} - 3\varepsilon.$$

The result follows on taking $\varepsilon \to 0$. □

Proof of Theorem 4.3 Based on Proposition 4.2, the proof of Theorem 4.3 is similar to the proof of Theorem 4.2 based on Proposition 4.1. Hence the details are omitted. □

5 Simulation results

Here we describe the results of simulation runs that show the dependence of the value V and the optimal barrier b on the retrial rate μ.

Figure 3a, b shows the dependence of the buffer size and, respectively, the value on μ as it ranges between 0 and 5000. Figure 3c–f shows a closer look at extreme cases: around zero (specifically, for μ in $[0, 0.01]$) and around 5000 (specifically, in the interval $[4990, 5000]$). Each graph consists of three lines corresponding to three values of p. The values are $p = 0.1, 0.5, 0.9$.

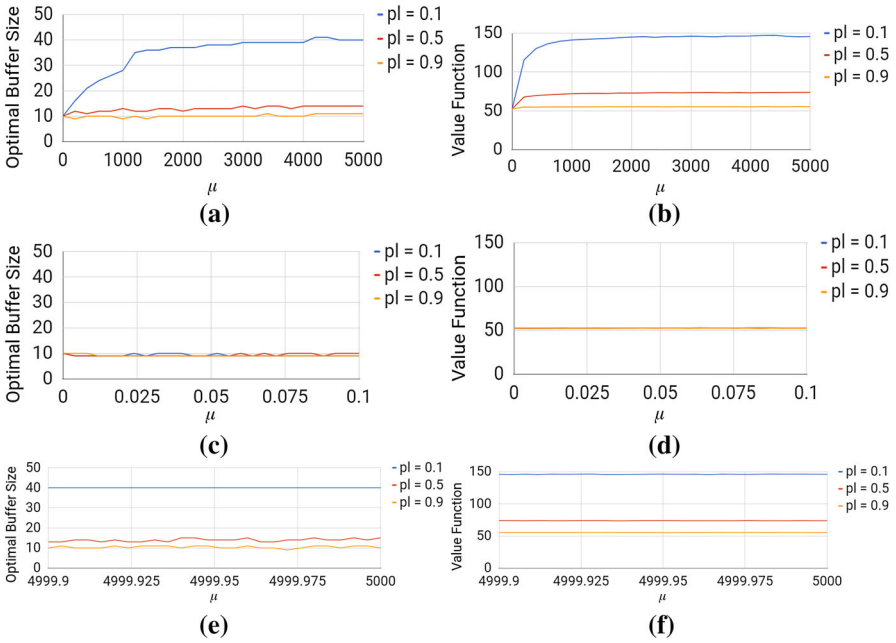


Fig. 3 Optimal buffer size and value

The graphs suggest that b and V increase with μ, whereas they decrease with p. The monotonicity of V can be explained heuristically. As μ grows, retrying customers return faster to the main station. Because the cost is discounted, this contributes to more rejection cost as well as the queue length cost, resulting in a greater overall cost. The monotonicity of b is less obvious. To reduce the rejection costs caused by an increase in retrials, it is clear that it is profitable to increase the buffer size. However, this also contributes to an increased queue length. The monotonicity suggested by the graphs at Fig. 3a indicates that the former factor is more significant than the latter. Similarly, as p decreases, the fraction of retrials increases, and a similar heuristic may be valid.

Another property demonstrated by the graphs is that V and b converge to the same limit for μ small, while their large μ limits depend on p. This agrees with our results from Sect. 4, where we showed that the $\mu \to 0$ limit is characterized by the Harrison–Taksar problem with parameters (c_1, c_2), and the $\mu \to \infty$ limit corresponds to this problem at $(c_1, c_2/p)$. It is thus clear that the former asymptotics should not depend on p, while the latter should.

Acknowledgements This research was supported in part by the ISF (Grant 1315/12).

References

1. Abramov, V.M.: Analysis of multiserver retrial queueing system: a martingale approach and an algorithm of solution. Ann. Oper. Res. **141**(1), 19–50 (2006)
2. Anisimov, V., Atadzhanov, K.L.: Diffusion approximation of systems with repeated calls and an unreliable server. J. Math. Sci. **72**(2), 3032–3034 (1994)
3. Anisimov, V.V.: Switching stochastic models and applications in retrial queues. Top **7**(2), 169 (1999)
4. Artalejo, J., Falin, G.: Standard and retrial queueing systems: a comparative analysis. Revista Matematica Complutense **15**(1), 101–129 (2002)
5. Artalejo, J.R.: Accessible bibliography on retrial queues. Math. Comput. Modell. **30**(3–4), 1–6 (1999)
6. Artalejo, J.R.: A classified bibliography of research on retrial queues: progress in 1990–1999. Top **7**(2), 187–211 (1999)
7. Artalejo, J.R.: Accessible bibliography on retrial queues: progress in 2000–2009. Math. Comput. Modell. **51**(9), 1071–1081 (2010)
8. Atar, R., Shifrin, M.: An asymptotic optimality result for the multiclass queue with finite buffers in heavy traffic. Stoch. Syst. **4**(2), 556–603 (2014)
9. Billingsley, P.: Convergence of Probability Measures. Wiley Series in Probability and Statistics: Probability and Statistics, 2nd edn. Wiley, New York (1999)
10. Cohen, J.: Basic problems of telephone traffic theory and the influence of repeated calls. Philips Telecommun. Rev. **18**(2), 49–100 (1957)
11. Falin, G.: Asymptotic investigation of fully available switching systems with high repetition intensity of blocked calls. Mosc. Univ. Math. Bull. **39**(6), 72–77 (1984)
12. Falin, G.: A survey of retrial queues. Queueing Syst. **7**(2), 127–167 (1990)
13. Falin, G.: A diffusion approximation for retrial queueing systems. Theory Probab. Its Appl. **36**(1), 149–152 (1992)
14. Falin, G., Artalejo, J.: Approximations for multiserver queues with balking/retrial discipline. OR Spectr. **17**(4), 239–244 (1995)
15. Fleming, W.H., Soner, H.M.: Controlled Markov Processes and Viscosity Solutions. Stochastic Modelling and Applied Probability, vol. 15, 2nd edn. Springer, New York (2006)
16. Harrison, J.M., Taksar, M.I.: Instantaneous control of Brownian motion. Math. Oper. Res. **8**(3), 439–453 (1983)

17. Karatzas, I., Shreve, S.E.: Brownian Motion and Stochastic Calculus. Graduate Texts in Mathematics, vol. 113, 2nd edn. Springer, New York (1991)
18. Krichagina, E.V., Taksar, M.I.: Diffusion approximation for $GI/G/1$ controlled queues. Queueing Syst. **12**(3–4), 333–367 (1992)
19. Kruk, L., Lehoczky, J., Ramanan, K., Shreve, S.: An explicit formula for the Skorokhod map on $[0, a]$. Ann. Probab. **35**(5), 1740–1768 (2007)
20. Lions, P.-L., Sznitman, A.-S.: Stochastic differential equations with reflecting boundary conditions. Commun. Pure Appl. Math. **37**(4), 511–537 (1984)
21. Lukashuk, L.: Diffusion approximation and filtering for a queueing system with repeats. Cybern. Syst. Anal. **26**(2), 253–264 (1990)
22. Mandelbaum, A., Massey, W.A., Reiman, M.I.: Strong approximations for Markovian service networks. Queueing Syst. **30**(1), 149–201 (1998)
23. Phung-Duc, T.: Retrial Queueing Models: A Survey on Theory and Applications. Applied Stochastic Models in Business and Industry (**to appear**)
24. Yang, T., Templeton, J.G.: A survey on retrial queues. Queueing Syst. **2**(3), 201–233 (1987)