Queueing Syst (2013) 74:65–104
DOI 10.1007/s11134-012-9326-6

On scheduling a multiclass queue with abandonments under general delay costs

Barış Ata · Mustafa H. Tongarlak

Received: 2 January 2012 / Revised: 3 July 2012 / Published online: 7 September 2012
© Springer Science+Business Media, LLC 2012

Abstract We consider a multiclass queueing system with abandonments and general delay costs. A system manager makes dynamic scheduling decisions to minimize long-run average delay and abandonment costs. We consider the three types of delay cost: (i) linear, (ii) convex, and (iii) convex–concave, where the last one corresponds to settings where customers may have a particular deadline in mind but once that deadline passes there is increasingly little difference in the added delay. The dynamic control problem for the queueing system is not tractable analytically. Therefore, we consider the system in the conventional heavy traffic regime and study the approximating Brownian control problem (BCP). We observe that the approximating BCP does not admit a pathwise solution due to abandonments. In particular, the celebrated $c\mu$ rule and its extension, the generalized $c\mu$ rule, which is asymptotically optimal under convex delay costs with no abandonments, are not optimal in this case. Consequently, we solve the associated Bellman equation, which yields a dynamic index policy (derived from the value function) as the optimal control for the approximating BCP. Interpreting that control in the context of the original queueing system, we propose practical policies for each of the three cases considered and demonstrate their effectiveness through a simulation study.

Keywords Dynamic control of multiclass queueing systems · Abandonments · Heavy traffic analysis · General delay costs

Mathematics Subject Classification 60K25 · 90B36 · 90B22 · 68M20

B. Ata · M.H. Tongarlak (✉)
Kellogg School of Management, Northwestern University, Evanston, IL 60208, USA
e-mail: m-tongarlak@kellogg.northwestern.edu

1 Introduction

This paper studies dynamic scheduling of a multiclass queue with abandonments in the heavy traffic regime. The objective is to minimize the average delay and abandonment costs. We consider three types of delay cost: (i) linear delay costs, (ii) convex delay costs, and (iii) convex–concave delay costs. A linear delay cost models the situation in which customers' marginal delay cost is constant. A convex delay cost is appropriate for situations where customers have a general idea on their desirable "deadline" and longer delays than their deadline are increasingly unattractive. Delays shorter than a customer's implicit deadline are attractive but not unduly so.

Convex–concave delay costs were first introduced by Ata and Olsen [4], and further analyzed by Akan et al. [1] and Ata and Olsen [5]. A convex–concave, or "S-shaped", cost curve models the situation where customers have a particular deadline in mind but once that deadline has passed, the longer delays are increasingly less unattractive, i.e. in that range the delay cost function is concave. Indeed, Leclerc et al. [40] argue through various behavioral experiments that the shape of the delay cost function depends on the context effects. It follows from their experiments that the delay cost function is concave in the range where the delay is relatively long, while it can be convex otherwise.

As shown in [4], for convex–concave delay costs the convex hull of the delay cost function serves as a lower bound to the system costs. The key idea behind designing good policies for the convex–concave delay costs is to asymptotically approach the cost incurred by the convex hull of the delay cost function.

Our work extends the existing work in several significant ways and makes the following contributions. First, in each of the cases considered, by solving a Bellman equation, we provide a novel dynamic index policy, which crucially depends on the derivative of the value function (or the shadow price of workload). Second, we highlight the important role abandonments play in controlling queueing systems. Indeed, under abandonments, the familiar $c\mu$ rule, cf. [19, 32, 35], and a generalization of that for convex delay costs ($Gc\mu$ rule), cf. [57], are not optimal. This is one of the important insights of our paper. To see why these are no longer (asymptotically) optimal, note that the (asymptotic) optimality of $c\mu$ or $Gc\mu$ policies (under no abandonments) hinges on the following two simple observations: (i) the evolution of the workload (i.e. hours of work remaining for the server) is independent of the scheduling discipline used as long as the scheduling policy is work-conserving; (ii) given a particular workload level in the system, the goal is to distribute the workload among various job classes so as to minimize the delay cost rate. The former assertion is no longer true under abandonments, because jobs of various classes may abandon at different rates, and hence, it makes a difference whether the workload is kept in one class versus another. In other words, the inclusion of abandonments may turn a "greedy" scheduling policy into a "nongreedy" one. While in a greedy scheduling policy, the server gives priority to the jobs with the highest immediate cost rate, this is not necessarily the case under a nongreedy policy. Under a nongreedy policy the server may serve jobs that do not have the highest immediate cost rate, but perhaps have a high abandonment rate and thus allows the system to forgo future abandonment costs. Namely, abandonments crucially change the system behavior and calls for

a more sophisticated style of analysis to characterize the optimal scheduling policy, which we undertake in this paper.

Our third contribution is to propose practical policies for the original queueing system. We test their performance against benchmarks (including $c\mu$, $Gc\mu$ rules etc.) and show that our proposed policies can offer significant benefits. Fourth, the convex–concave case presents additional challenges, which we overcome using the convex hull approach developed in [4]. The contribution of our paper (in the case of convex–concave delay costs) over Ata and Olsen [4], Akan et al. [1] and Ata and Olsen [5] is that it extends those to incorporate the abandonment behavior which is a crucial model element in many practical settings. Finally, we also provide a novel method for constructing a solution to the Bellman equation arising in the analysis, which may be of interest in its own right.

The rest of the paper is structured as follows. Section 2 reviews the related literature. Section 3 introduces the scheduling problem. The approximating Brownian control problem is introduced in Sect. 4 and solved in Sect. 5. Section 6 proposes policies for the original queueing system, which are tested in Sect. 7. Section 8 concludes. Appendix A provides a formal derivation of the approximating Brownian control problem, while Appendix B provides various auxiliary results and the technical proofs.

2 Literature review

There is a vast literature that considers the analysis, design and control of queueing networks; see [18] for a classical survey and [55] for a more recent survey. An important stream of research uses heavy traffic approximations to study scheduling problems in a dynamic stochastic environment; see, for example, [3, 10, 11, 16, 25–31, 36–39, 42, 46–52, 56, 57, 59–66]. Also see [33] for an overview of due-date improvement policies, which relates to design and control of queueing systems.

The objective of minimizing delay costs plays a prominent role in the literature on scheduling multiclass queueing systems. In the case of linear delay cost rates, the $c\mu$ rule assigns static priority levels to jobs in increasing order of their index $c_k\mu_k$. This rule minimizes the delay cost in systems with Poisson arrivals and linear delay cost rates, cf. [19, 32, 35]. This result is extended to convex and convex–concave delay costs in an asymptotic sense: In heavy traffic, a dynamic version of the $c\mu$ rule minimizes convex delay costs asymptotically, see [57]; similarly, a dynamic cost-balancing policy based on convex-hull functions minimizes the delay cost rate incurred under convex–concave delay costs; see [1]. Both Van Mieghem [57] and Akan et al. [1] are based on the following simple observations: (i) the evolution of the workload (as measured by hours of work remaining in the system for the server) is independent of scheduling discipline used (as long as the scheduling policy is work-conserving); (ii) given a particular workload level in the system, both policies strive to distribute the workload among various job classes so as to minimize the instantaneous delay cost rate.[1] Combining these two features leads to an asymptotically

[1] Akan et al. [1] also establish the incentive compatibility of their proposed scheduling rule when customers are strategic.

optimal pathwise policy, i.e. independent of the second order problem data; e.g. variance of interarrival and service times.

The recent survey [58] provides a thorough overview of the literature on the analysis of queueing systems with abandonments. In addition to the references covered in [58], recent work by Atar et al. [8, 9] prove optimality of $c\mu/\theta$ index policy for overloaded queues with abandonments. Overloaded queues are of interest due to their applicability to various important problems, including deceased donor organ transplant waiting lists; see for example [7]. Atar et al. [8, 9] focus on fluid scale optimality for overloaded queues, whereas our focus is on diffusion scale optimality for critically loaded systems for which such static priority rules will not be optimal as will be discussed next.

As mentioned earlier the key to the results of [1, 57] was that the evolution of the workload was independent of the scheduling policy. This is no longer true under abandonments. Jobs of various classes may abandon at different rates and they reduce the workload in the system at different rates. Hence, it makes a difference whether the workload is kept in one class versus another. In other words, abandonments crucially change the system behavior and call for a more sophisticated style of analysis to characterize the optimal scheduling policy. More specifically, the inclusion of abandonments may turn a "greedy" scheduling policy into a "nongreedy" one. While in a greedy scheduling policy, the server gives priority to the jobs with the highest immediate cost rate, this is not necessarily the case under a nongreedy policy. Under a nongreedy policy the server may serve jobs that do not have the highest immediate cost rate, but perhaps have a high abandonment rate and thus allow the system to forgo future abandonment costs. This observation underscores the contribution of our work. Indeed, abandonments require comparing the immediate cost of current actions with the future impact of these actions through the dynamic programming approach, which helps assess the "shadow price" of workload displacements. In what follows, we solve the associated Bellman equation and propose a policy based on that solution, which makes the current-future cost trade-off optimally. Interestingly, the proposed policy crucially depends on the value function derived from the Bellman equation and the second order problem data, hence, is not a pathwise solution.

The asymptotic analysis of our problem lends itself to a drift rate control problem on the positive real line. There have been several related papers in recent years which consider controlling the drift rate of a diffusion. Ata et al. [6] study the drift rate control of a reflected diffusion on a bounded interval under general costs of drift control. The authors derive closed-form expressions for the optimal policy and various other quantities of interest. Ata [2] applies a similar framework to an order fulfillment problem in make-to-order manufacturing and characterizes the optimal admission control problem as a nested-threshold policy, where explicit formulas are derived for the thresholds.

Two closely related papers [21, 54] consider scheduling for parallel server systems. Rubino and Ata [54] consider a general parallel server system with abandonments and admission control. There are linear holding costs, abandonment penalties, costs for turning jobs away, and delay constraints, which are replaced with upper bounds on queue lengths in the asymptotic formulation. Rubino and Ata [54] solve the limiting Brownian control problem under the long-run average cost formulation.

The authors then propose a policy for the general parallel server system and show its effectiveness through simulations.

Ghamami and Ward [21] consider a similar setting (though the authors consider a discounted objective). The authors take the analysis to the next level and provide a proof of asymptotic optimality for their proposed policy. It turns out that proving the asymptotic optimality of their proposed policy is challenging because the workload moves between different classes too frequently. Indeed, the usual state-space collapse result between the scaled queue length and the workload process does not seem to hold. Nonetheless, Ghamami and Ward [21] are able to prove the asymptotic optimality of their proposed policy using novel proof ideas.

In other related work, Ormeci-Matoglu and Vande Vate [45] consider the drift rate control problem with changeover costs and establish the optimality of band control policies. Ghosh and Weerasinghe [22] consider the optimal drift rate control and choice of the optimal buffer size. The authors solve a limiting version of this problem explicitly, and propose a policy for the original system based on this policy. Ghosh and Weerasinghe [22] also establish the asymptotic optimality of this policy in the heavy traffic limit; also see [15].

Among the drift rate control problems referenced above Rubino and Ata [54] is closest to our work; both papers consider ergodic control. However, this paper differs from Rubino and Ata [54] in several important ways. Firstly, Rubino and Ata [54] consider only linear delay costs whereas we consider linear, convex, and convex–concave delay costs. This difference leads to substantially different structural insights. Secondly, Rubino and Ata [54] considers the drift rate control problem on a bounded interval which simplifies the solution of the Bellman equation considerably. In contrast, we consider the problem on the entire positive real line, which necessitates a new approach to solve the Bellman equation. Lastly, Rubino and Ata [54] also has an admission control capability whereas we do not allow that to facilitate comparison with the familiar $c\mu$ and the generalized $c\mu$ ($Gc\mu$) scheduling rules.

3 The model

We consider a multiclass queue serving K classes of delay sensitive jobs, who differ in their delay costs. A system manager makes sequencing decisions, choosing the order in which jobs are processed. A class k job has a delay cost of $c_k(\tau)$ associated with an experienced delay of τ. The delay cost function $c_k(\cdot)$ is increasing (for all k), i.e. shorter delays are more desirable. In what follows, we will consider three cases: (i) c_k is linear; (ii) c_k is convex; (iii) c_k is convex–concave.

An important feature of the model is that jobs in the system may abandon, resulting in a penalty of a_k for the system manager per abandoned class k job. We assume that each class abandons after an exponentially distributed amount of time with rate γ_k. (Each abandonment takes place independently of all other abandonments, service completions and arriving jobs.) We assume that the job at the head of each queue does not abandon. Then denoting the number of class k jobs in the system at time t by $Q_k(t)$ for $k = 1, \ldots, K$ and letting $N_1(\cdot), \ldots, N_K(\cdot)$ be K independent, rate-one Poisson processes, the cumulative number of class k jobs abandoning up to time t, denoted by $\Gamma_k(t)$, is given by

$$\Gamma_k(t) = N_k\left(\int_0^t \gamma_k\big[Q_k(s) - 1\big]^+ ds\right), \quad k = 1, \ldots, K.$$

The vector-valued process $\Gamma = (\Gamma_k)$ will be called the abandonment process.

The sequencing decisions take the form of cumulative control process. We restrict attention to head-of-the-line and non-idling (or work-conserving) policies. In particular, let $T_k(t)$ be the cumulative time that the server spends on serving class k jobs up to time t. Then the vector process $T = (T_k)$ denotes the system manager's sequencing policy. Clearly, $T(\cdot)$ is nondecreasing and satisfies

$$0 \le \sum_{k=1}^{K}\big[T_k(t) - T_k(s)\big] \le (t - s), \quad 0 \le s \le t < \infty. \tag{1}$$

For concreteness, we assume that the system is empty initially; and class k jobs arrive at the system at rate λ_k according to a Poisson process $\{A_k(t) : t \ge 0\}$. Similarly, associated with each class k, there is a Poisson process $\{S_k(t) : t \ge 0\}$ with rate μ_k where $S_k(t)$ denotes the number class k jobs served up to time t if the server were continuously serving class k jobs during $[0, t]$. The mean processing time of a class k job is $m_k = 1/\mu_k$.

Given the system manager's sequencing policy T, the queue length process Q_k for class k evolves as follows:

$$Q_k(t) = A_k(t) - S_k\big(T_k(t)\big) - \Gamma_k(t), \quad t \ge 0. \tag{2}$$

The vector-valued process $Q = (Q_k)$ will be called the queue length process. Also let $L(t)$ denote the cumulative amount of time the server k is idle during $[0, t]$. Then

$$L(t) = t - \sum_{k=1}^{K} T_k(t), \quad t \ge 0. \tag{3}$$

To facilitate future analysis, we define the workload $W(t)$ in the system as

$$W(t) = \sum_{k=1}^{K} m_k Q_k(t). \tag{4}$$

A sequencing policy T must satisfy the following:

$$T \quad \text{is non-anticipating with respect to } Q, \tag{5}$$

$$T \quad \text{is non-decreasing and continuous with } T(0) = 0, \tag{6}$$

$$L \quad \text{is non-decreasing and continuous with } L(0) = 0, \tag{7}$$

$$L(t) \quad \text{increases only if } W(t) = 0, \tag{8}$$

$$Q_k(t) \ge 0, \quad t \ge 0,\ k = 1, \ldots, K. \tag{9}$$

Let $\tau_k(t)$ denote the delay experienced by a class k customer arriving at time t. Note that $\tau_k(t)$ is a rather complex function of the policy employed. Although we do not attempt to express it explicitly, in what follows we adopt a simple approximation via the snapshot principle of Reiman [53].

We define $C(t)$ as the cumulative delay cost experienced by the jobs arriving until t plus the abandonment costs during $[0, t]$, i.e.

$$C(t) = \sum_{k=1}^{K} \int_0^t \lambda_k c_k\big(\tau_k(s)\big)\, ds + \sum_{k=1}^{K} a_k \Gamma_k(t), \quad t \geq 0. \tag{10}$$

Then the system manager's problem can be stated as follows: Choose the scheduling policy T so as to

$$\min \varlimsup_{t\to\infty} \frac{1}{t}\mathbb{E}\big[C(t)\big] \quad \text{subject to (1)–(3) and (5)–(10).}$$

In what follows, we consider the following cases. Firstly, we assume that $c_k(\cdot)$ is linear. Secondly, we assume that $c_k(\cdot)$ is strictly convex with $c_k(0) = c_k'(0) = 0$ for all k. Finally, we assume that the delay cost $c_k(\cdot)$ is convex–concave which corresponds to $c_k(\cdot)$ being convex on an interval $[0, d_k]$ and concave on $[d_k, \infty)$ for $k = 1, \ldots, K$. The basic idea is that d_k represents a customer's deadline and he is increasingly more impatient up to this deadline and increasingly more tolerant of additional delays once the deadline has passed. We assume in the convex–concave case that $\lim_{x\to\infty} c_k'(x) = c > 0$, and that $c_k(0) = c_k'(0) = 0$. We also make the following technical assumption in the convex case: For some $M < \infty$,

$$\lim_{x\to\infty} c_k'(x) \leq M. \tag{11}$$

Note that this assumption is made for purely technical reasons and is needed only in Lemma 2. (All other results can be generalized to the case $M = \infty$.) More importantly, (11) does not change the structure of the policies proposed or the insights.

Unfortunately, none of the cases is analytically tractable. Moreover, the convex–concave case presents additional challenges, which we overcome using a convex hull approach. To be specific, let h_k denote the convex hull of c_k, i.e. h_k is the maximal convex function $h_k \leq c_k$. Clearly, $h_k = c_k$ in the case of linear or convex delay costs. In the case of convex–concave delay cost, defining

$$b_k = \inf\big\{x \geq 0 : c_k'(x) \geq c\big\},$$

it is easy to see that

$$h_k(x) = \begin{cases} c_k(x) & \text{for } x < b_k, \\ (x - b_k)c + c_k(b_k) & \text{otherwise.} \end{cases}$$

Note that in this case the convex hull h_k is linear beyond b_k and has slope c. Although $c_k(x) > h_k(x)$ for $x > b_k$, we argue below that using a clever sequencing rule, the system manager can achieve this effective slope of c. To be more specific, the system manager can obtain the convex hull function h_k as the realized delay cost, instead of the higher delay cost c_k.

We will replace c_k by h_k in all three cases and consider the following problem: Choose $T(\cdot)$ so that

$$\min \varlimsup_{t\to\infty} \frac{1}{t}\mathbb{E}\big[H(t)\big] \quad \text{subject to (1)–(3) and (5)–(10),} \tag{12}$$

where

$$H(t) = \sum_{k=1}^{K} \int_0^t \lambda_k h_k\big(\tau_k(s)\big)\, ds + \sum_{k=1}^{K} a_k \Gamma_k(t), \quad t \geq 0,$$

and h_k is a convex increasing function and we allow for the possibility that it corresponds to the convex hull of a convex–concave function.

Note that in the cases of linear and convex delay costs (12) is equivalent to (10), whereas in the case of convex–concave delay costs it corresponds to a lower bounding problem, where the delay cost functions are convex (though they may only be weakly convex for large delays). More importantly, in all cases, the solution to (12) facilitates a solution to (10). So, we next consider (12). Unfortunately, the formulation (12) is still not tractable analytically. Therefore, we will replace it with an approximate, yet far more tractable formulation in the large capacity asymptotic regime, which we describe in the next section.

4 Approximating Brownian control problem

In deriving Brownian approximations, one considers a sequence of closely related systems indexed by a parameter, whose formal limit is the approximating Brownian control problem. More specifically, consider a sequence of systems indexed by $n = 1, 2, \ldots$; a superscript n will be attached to the quantities of interest corresponding to the nth system. The asymptotic regime we focus on is the one where the server speeds up proportionally to n in the nth system. More specifically, we assume

$$\mu_k^n = n\mu_k. \tag{13}$$

We also assume that the arrival rates grow with n such that

$$\lambda_k^n = n\lambda_k - \eta_k\sqrt{n}, \tag{14}$$

where $\eta_k > 0$, and that

$$\sum_{k=1}^{K} \lambda_k/\mu_k = 1, \tag{15}$$

which is the usual heavy traffic assumption, cf. [24]. Assumptions (13)–(15) lead to a large, balanced-flow system for n large. For such systems, the workload in the system is expected to be of order $1/\sqrt{n}$. Thus, we scale the delay costs as follows:

$$c_k^n(\cdot) = \frac{c_k^n(\sqrt{n}\cdot)}{\sqrt{n}} \quad \text{for all } k, n.$$

It is easy to check that the convex hull $h_k^n(\cdot)$ of $c_k^n(\cdot)$ is given by the following:

$$h_k^n(\cdot) = \frac{h_k^n(\sqrt{n}\cdot)}{\sqrt{n}} \quad \text{for all } k, n. \tag{16}$$

Note that it is optimal to serve jobs within a class on a FCFS basis, because h_k is convex for all k. Then by the snapshot principle of Reiman [53], we can approximate the actual delay experienced by a class k customer at time t by

$$\tau_k^n(t) \simeq \frac{Q_k^n(t)}{\lambda_k^n} \quad \text{for all } k, t, \tag{17}$$

which becomes accurate in the heavy traffic limit.

We also introduce the centered allocation process Y^n for the nth system as

$$Y_k^n(t) = \frac{\lambda_k^n}{\mu_k^n} t - T_k^n(t), \tag{18}$$

whose (scaled) formal limit will arise as the control process in the approximating Brownian control problem.

The original problem of interest, cf. (12), can be viewed as a specific element of this sequence of problems determined by the particular choice of the parameter n. The underlying assumption of the Brownian approximations is that the system parameter corresponding to the original problem is large enough so that various (scaled) performance-relevant processes of the original system can be approximated by the corresponding processes of the Brownian control problem. One arrives at the approximating Brownian control problem by scaling various processes and taking the formal limit as $n \to \infty$, which is outlined in Appendix A.

Defining B_k as a $(-\eta_k, \sigma_k)$ Brownian motion, where $\sigma_k = \sqrt{2\lambda_k}$, for $k = 1, \ldots, K$, the Brownian control problem can be stated as follows: Choose the control processes $\hat{Y}$, $\hat{L}$ so as to

$$\min \varlimsup_{t\to\infty} \frac{1}{t} \mathbb{E}\left[\int_0^t \sum_{k=1}^K \left(\lambda_k h_k \left(\frac{\hat{Q}_k(s)}{\lambda_k} \right) + a_k \gamma_k \hat{Q}_k(s) \right) ds \right] \tag{19}$$

subject to

$$\hat{Q}_k(t) = B_k(t) - \int_0^t \gamma_k \hat{Q}_k(s)\, ds + \mu_k \hat{Y}_k(t) \geq 0 \quad \text{for all } k,\ t \geq 0, \tag{20}$$

$$\hat{L}(t) = \sum_{k=1}^K \hat{Y}_k(t), \quad t \geq 0, \tag{21}$$

$$\hat{L} \quad \text{is nondecreasing and continuous with } \hat{L}(0) = 0, \tag{22}$$

$$\hat{L} \quad \text{increases only if } \hat{Q}_k = 0 \text{ for all } k, \tag{23}$$

$$\hat{L}, \hat{Y} \quad \text{are nonanticipating with respect to } B_k \ (k = 1, \ldots, K). \tag{24}$$

The process $\hat{Y}$ is the formal limit of the (scaled) sequence of controls Y^n as $n \to \infty$. Similarly, the queue length process $\hat{Q}$ and the cumulative idleness process $\hat{L}$ are formal limits of their (scaled) counterparts in the sequence of systems considered.

Next, we advance a one-dimensional workload formulation, which is equivalent to (19)–(24). To this end, we first formulate a reduced Brownian control problem by replacing (20) with (26)–(27) below, relaxing (20). Then viewing the processes $\hat{Q}$ and $\hat{L}$ as controls, the reduced Brownian control problem, which has the one-dimensional state descriptor $\hat{W} = \{\hat{W}(t) : t \geq 0\}$, can be stated as follows: Choose a policy $(\hat{Q}, \hat{L})$ that is nonanticipating with respect to B so as to

$$\min \varlimsup_{t\to\infty} \frac{1}{t} \mathbb{E}\left[\int_0^t \sum_{k=1}^K \left(\lambda_k h_k \left(\frac{\hat{Q}_k(s)}{\lambda_k} \right) + a_k \gamma_k \hat{Q}_k(s) \right) ds \right] \tag{25}$$

subject to

$$\hat{W}(t) = B(t) - \int_0^t \left(\sum_{k=1}^K \gamma_k m_k \hat{Q}_k(s) \right) ds + \hat{L}(t), \tag{26}$$

$$\hat{W}(t) = \sum_{k=1}^K m_k \hat{Q}_k(t), \tag{27}$$

$$\hat{Q}(t) \geq 0, \tag{28}$$

$$\hat{L} \quad \text{is nondecreasing and continuous with } \hat{L}(0) = 0, \tag{29}$$

$$\hat{L} \quad \text{increases only when } \hat{W} = 0, \tag{30}$$

where $B(t) = \sum_{k=1}^K m_k B_k(t)$ for $t \geq 0$ whose drift rate is $-\eta = -\sum_{k=1}^K m_k \eta_k$ and infinitesimal variance is $\sigma^2 = \sum_{k=1}^K 2\lambda_k m_k^2$. The following proposition is immediate and establishes the equivalence of these formulations from which it is clear that the two formulations have the same objective.

Proposition 1 *The Brownian control problem stated in* (19)–(24) *is equivalent to the reduced Brownian control problem* (25)–(30) *in the following sense. Every feasible policy* $(\hat{Q}, \hat{L})$ *for the reduced Brownian control problem yields a policy* $(\hat{Y}, \hat{L}\,)$ *for the Brownian control problem with the same cost*; *and for every feasible policy* $(\hat{Y}, \hat{L})$ *for the Brownian control problem, there is a policy* $(\hat{Q}, \hat{L})$ *for the reduced Brownian control problem which has lower* (*or the same*) *cost*.

Next, we further simplify the reduced Brownian control problem to arrive at the workload formulation. To that end, define

$$\mathcal{A}(x) = \left\{ q \geq 0 : \sum_{k=1}^K m_k q_k = x \right\},$$

$$\theta(q) = \eta + \sum_{k=1}^K \gamma_k m_k q_k, \tag{31}$$

$$g(q) = \sum_{k=1}^K \left[\lambda_k h_k \left(\frac{q_k}{\lambda_k} \right) + a_k \gamma_k q_k \right]. \tag{32}$$

A policy for the workload formulation consists of a process $\hat{L}(\cdot)$ and a workload configuration function $q : \mathbb{R}_+^2 \to \mathbb{R}_+^K$, which describes how the workload is distributed among various classes[2] at all times. An admissible policy $(\hat{L}, q)$ must jointly satisfy

$$\hat{W}(t) = B(t) - \int_0^t \theta\big(q\big(s, W(s)\big)\big)\, ds + \hat{L}(t) \geq 0, \quad t \geq 0, \tag{33}$$

[2]We do allow randomized workload configuration functions, that is, an admissible workload configuration function q may be sample-path dependent. This dependence, however, is suppressed for notational brevity. Moreover, we construct an optimal workload configuration function in Sect. 5, which is stationary and deterministic.

$$q\big(s, \hat{W}(t)\big) \in \mathcal{A}\big(\hat{W}(t)\big), \quad t \geq 0, \tag{34}$$

$$\hat{L} \quad \text{is nondecreasing and continuous with } \hat{L}(0) = 0, \tag{35}$$

$$\int_0^\infty \hat{W}(s)\, d\hat{L}(s) = 0. \tag{36}$$

The workload problem can then be stated as follows: Choose $\hat{L}$ and $q(\cdot, \cdot)$ so as to

$$\min \varlimsup_{t\to\infty} \frac{1}{t}\mathbb{E}\left[\int_0^t g\big(q\big(s, \hat{W}(s)\big)\big)\, ds\right] \quad \text{subject to} \quad (33)\text{–}(36). \tag{37}$$

In the workload formulation, the system manager tracks the workload process $\hat{W}$ and makes sure it is nonnegative using the control $\hat{L}$. Given the workload $\hat{W}(t)$ at time t, she also decides how to distribute that workload among the various classes $k = 1, \ldots, K$ so that $\sum_{k=1}^K m_k \hat{Q}_k(t) = \hat{W}(t)$ which then results in an effective holding cost rate of $g(\hat{Q}(t)) = \sum_{k=1}^K \lambda_k h_k(\hat{Q}_k(t)/\lambda_k) + \sum_{k=1}^K a_k \gamma_k \hat{Q}_k(t)$, where the first term is the delay cost rate while the second term is the cost of abandonments.

The following proposition shows that the workload formulation is equivalent to the reduced Brownian control problem, and hence it is equivalent to the Brownian control problem too by Proposition 1, for purposes of optimal control.

Proposition 2 *The workload formulation* (37) *is equivalent to the reduced Brownian control problem* (25)–(30) *in the following sense*: *Every admissible policy* $(q, \hat{L})$ *of the workload formulation there corresponds to an admissible policy* $(\hat{Q}, \hat{L})$ *for the reduced Brownian control problem and these two policies have the same cost. Similarly, for any admissible policy* $(\hat{Q}, \hat{L})$ *of the reduced Brownian control problem, there exists an admissible policy* $(q, \hat{L})$ *for the workload formulation, and its cost is less than or equal to that of the policy* $(\hat{Q}, \hat{L}\,)$ *for the reduced Brownian control problem.*

Propositions 1 and 2 make it clear that it suffices to solve the workload formulation, which we undertake in the next section, to solve the Brownian control problem.

5 Solving the workload problem

To characterize the optimal policy for the workload problem, we next consider the associated Bellman equation. The following definitions are needed to introduce the Bellman equation:

$$\tilde{\mathcal{A}} = \left\{ y \in \mathbb{R}_+^K : \sum_{k=1}^K m_k y_k = 1 \right\},$$

$$\psi(x, p) = \min_{y\in\tilde{\mathcal{A}}} \left\{ \sum_{k=1}^K \left(\frac{\lambda_k}{x} h_k\left(\frac{y_k x}{\lambda_k}\right) + a_k \gamma_k y_k \right) - p \sum_{k=1}^K \gamma_k m_k y_k \right\}, \tag{38}$$

$$x > 0,$$

$$\bar{p} = \sup\Big\{ p > 0 : \lim_{x\to\infty} \psi(x, p) > 0 \Big\}.$$

Defining $C^2[0,\infty)$ as the space of functions $f:[0,\infty)\to\mathbb{R}$ that are twice continuously differentiable, the Bellman equation can be stated as follows: Find a convex function $f\in C^2[0,\infty)$ and a constant $\beta>0$ such that the following holds:

$$\beta=\min_{q\in\mathcal{A}(x)}\left\{\frac{1}{2}\sigma^2 f''(x)-\theta(q)f'(x)+g(q)\right\}, \tag{39}$$

subject to the boundary conditions

$$f'(0)=0 \quad \text{and} \quad \lim_{x\to\infty} f'(x)=\bar{p}. \tag{40}$$

Here one interprets β as a guess at the minimum average cost and the unknown function f is often called the relative value function in average cost dynamic programming. The Bellman equation is introduced primarily to motivate our solution approach; the properties of the Bellman equation that we require will be proved from first principles. To facilitate our analysis, define

$$\phi(x,v)=\min_{q\in\mathcal{A}(x)}\left\{g(q)-v\theta(q)\right\} \quad \text{for } x\geq 0 \text{ and } v\in\mathbb{R}.$$

Note as an aside that $\phi(x,v)=x\psi(x,v)-\eta v$. Since the Bellman equation (39)–(40) does not involve the unknown function f itself, it is really a first order equation. Defining $C^1[0,\infty)$ as the space functions that are continuously differentiable, setting $v(x)=f'(x)$ for $x\geq 0$, and using the definition of ϕ one can equivalently state the Bellman equation as follows: Choose a non-decreasing function $v\in C^1[0,\infty)$ and a constant β satisfying

$$\beta=\frac{1}{2}\sigma^2 v'(x)+\phi\big(x,v(x)\big) \quad \text{for } x\geq 0, \tag{41}$$

subject to the boundary conditions

$$v(0)=0 \quad \text{and} \quad \lim_{x\to\infty} v(x)=\bar{p}. \tag{42}$$

Solving the Bellman equation (41)–(42) directly appears challenging, because one needs to pin down the average cost β using the second boundary condition $\lim_{x\to\infty} v(x)=\bar{p}$. To the best of our knowledge, there is no direct method for solving (41)–(42). In what follows, we provide a constructive existence proof, which also facilitates computation of the value function and the corresponding optimal policy; it reveals important structural insights too. To this end, we next consider a family of auxiliary Bellman equations, whose solutions will eventually help construct the solution to the Bellman equation (41)–(42) by taking their limit appropriately.

To be specific, we consider a family of auxiliary Bellman equations parametrized by a penalty parameter p; see (43)–(44). Once we solve the auxiliary Bellman equation for each p, we pass to a limit as $p\nearrow\bar{p}$, and characterize the limits of the auxiliary value functions and other related quantities (in Sect. 5.2). To this end, Proposition 6 verifies the monotonicity of various quantities in p so that the limits are well defined as $p\nearrow\bar{p}$. Proposition 7 verifies useful properties of these limits; and the main result (Proposition 8) shows that the limit of the auxiliary value functions (as $p\nearrow\bar{p}$) solve the Bellman equation (41)–(42); also see Corollary 2. We propose a candidate policy based on this solution and verify its optimality in Theorem 1.

As mentioned above, a key step in solving the Bellman equation (41)–(42) is to solve the auxiliary Bellman equation (43)–(44), which is done in Sect. 5.1. To solve the auxiliary Bellman equation (for a fixed penalty parameter p), we consider a family of initial value problems (IVP) parametrized by $\hat{x}$. Proposition 3 shows that the IVP has a unique solution, which strictly increases in the parameter $\hat{x}$, and that for each fixed parameter $\hat{x}$ the solution increases to its maximum over $[0, \hat{x}]$. The last observation is useful in constructing a smooth solution by the smooth-pasting approach. To be more specific, because the derivative at the maximum is zero, pasting the solution to the IVP with a constant function from then on (continuously) yields a smooth solution to the auxiliary Bellman equation. To this end, Proposition 4 and Corollary 1 help pick the "right" parameter $\hat{x} = x(p)$ so that the solution to the IVP for that parameter essentially solves the auxiliary Bellman equation for the penalty parameter p. Indeed, using this solution to the IVP and the parameter $x(p)$, (48) constructs the candidate solution, and Proposition 5 verifies that it indeed solves the auxiliary Bellman equation (for each fixed p).

In our analysis, we also use various lemmas to prove the results. Lemmas 3, 4, and 5 (see Appendix B) establish the elementary properties of ψ and $\bar{p}$ and are used extensively throughout the paper. Lemma 2 (stated before Theorem 1 and proved in Appendix B) is used to show that a term vanishes asymptotically in proving Theorem 1, which in turn verifies the optimality of the proposed policy. Lemma 1 establishes (Lipschitz) continuity of ψ and used in Propositions 4 and 5 to establish the existence of a solution to the IVP and the continuity of that solution in the parameter $\hat{x}$, respectively. It is also used in Proposition 8 to establish a uniform convergence result. Lemmas 6 and 7 establish monotonicity of $\phi(\cdot, p)$ under certain conditions, and are used in Corollary 1 and Proposition 3, respectively.

5.1 A family of auxiliary Bellman equations

For each $p \in (0, \bar{p})$, the auxiliary problem can be stated as follows: Find a nondecreasing $v \in C^1[0, \infty)$, and constants $\beta(p) > 0, x(p) > 0$ such that

$$\beta(p) = \frac{1}{2}\sigma^2 v'(x) + \phi\big(x, v(x)\big), \quad x \in \big[0, x(p)\big], \tag{43}$$

subject to the boundary conditions

$$v(0) = 0 \quad \text{and} \quad v(x) = p \quad \text{for } x \geq x(p). \tag{44}$$

Consider an auxiliary system whereby our original problem is modified so that the system manager can turn away arriving jobs but incurs a rejection penalty of p for doing so per such job. Interestingly, (43)–(44) is the Bellman equation for this problem. Intuitively, as $p \nearrow \bar{p}$, the system manager does not turn away jobs in the limit, which is precisely how we tackle our problem. We solve (43)–(44) for each p, and let $p \to \bar{p}$. Then we prove that the limit yields a solution to the Bellman equation (41)–(42), solving our problem.

The auxiliary problem mentioned in the preceding paragraph is related to the problem studied in [54]. But it has two important differences. First Rubino and Ata [54]

consider only linear delay costs whereas we consider linear, convex, and convex–concave delay costs. This difference leads to substantially different structural insights. Second, Rubino and Ata [54] consider the drift rate control problem on a bounded interval which simplifies the solution to the Bellman equation considerably. In contrast, we consider the problem on the entire positive real line, which necessitates a new approach to solve the Bellman equation.

As a preliminary to solving the auxiliary Bellman equation, define

$$\underline{x}(p) = \inf\left\{x \geq 0 : \psi(x, p) > \frac{\eta p}{x}\right\} \quad \text{for } p \in (0, \bar{p}).$$

Lemma 5 (in Appendix B) characterizes useful properties of $\underline{x}(p)$, ψ, and ϕ.

To construct a solution to the auxiliary Bellman equation (43)–(44) for $p \in (0, \bar{p})$, consider the following initial value problem parametrized by $\hat{x} \geq \underline{x}(p)$, denoted by IVP($\hat{x}$): Find a continuously differentiable function v such that

$$\phi(\hat{x}, p) = \frac{1}{2}\sigma^2 v'(x) + \phi\big(x, v(x)\big) \quad \text{for } x \geq 0 \tag{45}$$

subject to the initial condition

$$v(0) = 0. \tag{46}$$

The following proposition characterizes important properties of the IVP($\hat{x}$).

Proposition 3 *For $p \in (0, \bar{p})$ and $\hat{x} > \underline{x}(p)$, the following hold*:

(i) IVP($\hat{x}$), *stated in* (45)–(46), *has a unique solution, denoted by* $v_{\hat{x}}$.
(ii) $v_{\hat{x}}(x)$ *is strictly increasing in* $\hat{x}$ *for all* $x > 0$.
(iii) $v_{\hat{x}}(\cdot)$ *strictly increases to its maximum on* $[0, \hat{x}]$.

Then replacing the notation $v_{\hat{x}}(\cdot)$ with $v_{\hat{x}}(\cdot; p)$ to emphasize its dependence on p and defining

$$\zeta(\hat{x}; p) = \sup_{0 \leq x \leq \hat{x}} v_{\hat{x}}(x; p), \tag{47}$$

the following proposition characterizes its important properties, which in turn helps us pin down the solution to the auxiliary Bellman equation (43)–(44).

Proposition 4 *For $p \in (0, \bar{p})$, $\zeta(\cdot; p)$ is strictly increasing and continuous on $(\underline{x}(p), \infty)$ with $\zeta(\underline{x}(p); p) = 0$ and $\lim_{\hat{x} \to \infty} \zeta(\hat{x}; p) = \infty$.*

The next corollary follows from Proposition 4.

Corollary 1 *There exists a unique $x(p) > \underline{x}(p)$ such that $\zeta(x(p); p) = p$, and $v_{x(p)}(\cdot)$ is strictly increasing on $(0, x(p))$ with $v'_{x(p)}(x(p); p) = 0$.*

Then letting $\beta(p) = \phi(x(p), p)$ and defining

$$v(x; p) = \begin{cases} v_{x(p)}(x; p), & 0 \leq x \leq x(p), \\ p, & x > x(p), \end{cases} \tag{48}$$

the following proposition solves the auxiliary Bellman equation.

Proposition 5 *The function* $v(\cdot; p)$ *(given in* (48)*) and the constants* $x(p)$ *and* $\beta(p)$ *jointly solve the auxiliary Bellman equation* (43)–(44).

Building on Proposition 5, in the next subsection we vary p and construct a solution to the Bellman equation (41)–(42) as $p \nearrow \bar{p}$.

5.2 Varying p and the solution to the Bellman equation

Firstly, we establish the monotonicity of the solution to the auxiliary Bellman equation.

Proposition 6 *The following hold*:

(i) $x(p)$ *is strictly increasing on* $(0, \bar{p})$.
(ii) $\beta(p)$ *is strictly increasing on* $(0, \bar{p})$.
(iii) $v(x; p_2) > v(x; p_1)$ *for* $0 < p_1 < p_2 < \bar{p}$ *and* $x > 0$.

Then the following limits are well defined (though they may be $+\infty$). Let

$$\beta^* = \lim_{p \to \bar{p}} \beta(p), \qquad \bar{x} = \lim_{p \to \bar{p}} x(p) \quad \text{and} \quad v^*(x) = \lim_{p \to \bar{p}} v(x; p) \quad \text{for } x \geq 0. \tag{49}$$

The next proposition characterizes useful properties of these limits.

Proposition 7 *The following hold*:

(i) $\beta^* < \infty$.
(ii) $\bar{x} = \infty$.
(iii) $v^*(x) < \infty$ *for all* $x \geq 0$, *and* $\lim_{x \to \infty} v^*(x) = \bar{p}$.

Next, we provide a solution to the Bellman equation, for which the following lemma is used crucially.

Lemma 1 $\psi(x, v)$ *is decreasing in* v *and continuous in* (x, v). *Moreover, it is Lipschitz continuous in* v *with Lipschitz constant* $c_L = \max_k \gamma_k$.

The following proposition provides a solution to the Bellman equation.

Proposition 8 *The function* $v^*(\cdot)$ *and the constant* β^* *defined in* (49) *jointly solve the Bellman equation* (41)–(42).

Proof Recall that $\psi(x, v)$ is Lipschitz continuous in v uniformly in x (see Lemma 1) and that $\phi(x, v) = x\psi(x, v) - \eta v$. Therefore, $\phi(x, v)$ is Lipschitz continuous in v (uniformly in x on compact intervals, i.e. when $x \in [0, K]$, $K > 0$). Also note that since $v(\cdot; p)$ is increasing in p by Proposition 6 (iii), that $v^*(x) < \infty$ for all $x \geq 0$, and $v(x, p) \nearrow v^*(x)$ as $p \to \bar{p}$, we conclude by Dini's theorem (see Billingsley [12]) that $v(\cdot; p)$ converges to $v^*(\cdot)$ uniformly over compact intervals.

Next, fix $K > 0$ and let $p_0 < \bar{p}$ be such that $x(p_0) > K$, and note that for $p > p_0$, we have $v(0, p) = 0$ and

$$\beta(p) = \frac{1}{2}\sigma^2 v'(x; p) + \phi\big(x, v(x; p)\big), \quad x \in [0, K],$$

which gives

$$\frac{1}{2}\sigma^2 v(x; p) = \beta(p)x - \int_0^x \phi\big(s, v(s; p)\big)\, ds, \quad x \in [0, K].$$

Then letting $p \to \bar{p}$

$$\frac{1}{2}\sigma^2 v^*(x) = \beta^* x - \lim_{p\to\infty} \int_0^x \phi\big(s, v(s; p)\big)\, ds, \quad x \in [0, K].$$

One can interchange the limit and the integral since $\phi(s, v(s; p))$ converges uniformly in s (on $[0, K]$) as $p \to \infty$, which follows from the uniform convergence of $v(\cdot; p)$ on $[0, K]$ and the Lipschitz continuity of $\phi(s, \cdot)$ uniformly in $s \in [0, K]$. Then we conclude that

$$\frac{1}{2}\sigma^2 v^*(x) = \beta^* x - \int_0^x \phi\big(s, v^*(s)\big)\, ds, \quad x \in [0, K].$$

Since the preceding argument can be repeated for all $K > 0$, we conclude that

$$\frac{1}{2}\sigma^2 v^*(x) = \beta^* x - \int_0^x \phi\big(s, v^*(s)\big)\, ds, \quad x \geq 0,$$

which shows that $v^*(\cdot)$ is continuously differentiable and solves the differential equation below

$$\frac{1}{2}\sigma^2 v'(x) = \beta^* - \phi\big(x, v(x)\big), \quad x \geq 0 \quad \text{subject to} \quad v(0) = 0.$$

Moreover, since $\lim_{x\to\infty} v^*(x) = \bar{p}$ by part (iii) of Proposition 7, $v^*(\cdot)$ and β^* solve the Bellman equation (41)–(42). □

To construct a solution to the Bellman equation (39)–(40), define

$$f(x) = \int_0^x v(u)\, du, \quad x \geq 0.$$

Then the following corollary provides a solution to the Bellman equation formally.

Corollary 2 *The function f and the constant β^* solve the Bellman equation* (39)–(40).

5.3 Proposed solution for the workload formulation and its optimality

Given a solution to the Bellman equation, for every $x \geq 0$, our candidate for the optimal workload configuration is the minimizer of the right-hand side of (39). Let

$$q^*(x) = \operatorname*{argmin}_{q\in\mathcal{A}(x)} \big\{g(q) - \theta(q) f'(x)\big\}, \quad x \geq 0.$$

In Sects. 6.1–6.3, we characterize $q^*(\cdot)$ explicitly for each of the cases considered and propose a scheduling policy based on those characterizations. Note that $q^*(x)$ simply distributes the workload into various classes so as to minimize $g(q)-\theta(q)f'(x)$. The second feature of our candidate policy is that it imposes a reflecting barrier for the workload process at zero by the control L^*. The evolution of the workload process W^* under the candidate policy $(q^*(\cdot), L^*)$ can be described as follows.

$$W^*(t) = B(t) - \int_0^t \theta\big(q^*\big(W^*(s)\big)\big)\,ds + L^*(t), \quad t \geq 0.$$

Moreover, the control L^* and the workload process W^* jointly satisfy

$$\begin{aligned} &W^*(t) \geq 0, \quad t \geq 0, \\ &\int_0^t W^*(s)\,dL^*(s) = 0, \quad t \geq 0, \\ &L^* \quad \text{is continuous and non-decreasing with } L^*(0) = 0. \end{aligned} \tag{50}$$

Note that the candidate policy is stationary. Theorem 1 establishes the optimality of the policy $(q^*(\cdot), L^*)$ using the following technical result proved in the appendix.

Lemma 2 *For any workload process W associated with an admissible policy (q, L), we have*

$$\lim_{t\to\infty} \frac{\mathbb{E}[f(W(t))]}{t} = 0.$$

Theorem 1 *The candidate policy $(q^*(\cdot), L^*)$ associated with the workload configuration function $q^*(\cdot)$ and the reflecting barrier at zero is optimal for the workload problem (12), and it has long-run average cost of β^*.*

Proof The candidate policy $(q^*(\cdot), L^*)$ is admissible because it satisfies (33)–(36). To check its optimality we show that its long-run average cost is β^*, and the long-run average cost associated with any other admissible policy is greater than or equal to β^*.

First, consider the candidate policy $(q^*(\cdot), L^*)$. The cumulative cost incurred up to time $t > 0$ under the candidate policy, denoted by $H^*(t)$, is given by

$$H^*(t) = \int_0^t g\big(q^*\big(W^*(s)\big)\big). \tag{51}$$

A straightforward application of Ito's lemma gives

$$\begin{aligned} \mathbb{E}\big[f\big(W^*(t)\big)\big] = \mathbb{E}\bigg[\int_0^t \Big(-f'\big(W^*(s)\big)\theta\big(q^*\big(W(s)\big)\big) + \frac{1}{2}\sigma^2 f''\big(W^*(s)\big)\Big)\,ds\bigg] \\ + \mathbb{E}\bigg[\int_0^t f'\big(W^*(s)\big)\,dL^*(s)\bigg]. \end{aligned} \tag{52}$$

Combining (51) and (52) gives

$$\begin{aligned}
&\mathbb{E}\big[H^*(t) - \beta^* t + f\big(W^*(t)\big)\big] \\
&\quad = \mathbb{E}\bigg[\int_0^t \Big\{ g\big(q^*\big(W^*(s)\big)\big) - f'\big(W^*(s)\big)\theta\big(q^*\big(W^*(s)\big)\big) \\
&\qquad + \frac{1}{2}\sigma^2 f''\big(W^*(s)\big) - \beta^* \Big\}\, ds\bigg] \\
&\qquad + \mathbb{E}\bigg[\int_0^t f'\big(W^*(s)\big)\, dL^*(s)\bigg].
\end{aligned}$$

Since $W^*(t) \geq 0$ under the candidate policy, the first term on the right-hand side is zero by (39). Also it follows from (50) that L^* increases only when $W^* = 0$, so the second term on the right is zero too since $f'(0) = 0$. Then it also follows from Lemma 2 that

$$\lim_{t\to\infty} \frac{\mathbb{E}[H^*(t)]}{t} = \beta^*.$$

Next, we will show that no admissible policy can achieve a long-run average cost less than β^*. To this end, fix an admissible policy $(q(\cdot), L(\cdot))$. Recall that every admissible policy, and hence L must be continuous (cf. (35)). Then a straightforward application of Ito's lemma, cf. Sect. 4.7 of Harrison [23], gives

$$\begin{aligned}
\mathbb{E}\big[f\big(W(t)\big)\big] = \mathbb{E}\bigg[\int_0^t \Big(-f'\big(W(s)\big)\theta\big(q\big(s, W(s)\big)\big) + \frac{1}{2}\sigma^2 f''\big(W(s)\big)\Big)\, ds\bigg] \\
+ \mathbb{E}\bigg[\int_0^t f'\big(W(s)\big)\, dL(s)\bigg]. \qquad (53)
\end{aligned}$$

The cumulative cost incurred up to time $t > 0$ under the policy (q, L) is given by

$$H(t) = \int_0^t g\big(q\big(s, W(s)\big)\big)\, ds. \qquad (54)$$

Then for $t > 0$, combining (53) and (54) gives the following:

$$\begin{aligned}
&\mathbb{E}\big[H(t) - \beta^* t + f\big(W(t)\big)\big] \\
&\quad = \mathbb{E}\bigg[\int_0^t \Big\{ g\big(q\big(s, W(s)\big)\big) - f'\big(W(s)\big)\theta\big(q\big(s, W(s)\big)\big) \\
&\qquad + \frac{1}{2}\sigma^2 f''\big(W(s)\big) - \beta^* \Big\}\, ds\bigg] \\
&\qquad + \mathbb{E}\bigg[\int_0^t f'\big(W(s)\big)\, dL(s)\bigg]. \qquad (55)
\end{aligned}$$

The right-hand side of (55) is nonnegative. The first term is nonnegative by (39). The second term is zero since $f'(0) = 0$ and Lincreases only when $W = 0$. Thus,

$$\frac{\mathbb{E}[H(t)]}{t} \geq \beta^* - \frac{\mathbb{E}[f(W(t))]}{t}.$$

Then by Lemma 2, we conclude

$$\underline{\lim}_{t\to\infty} \frac{1}{t}\mathbb{E}\big[H(t)\big] \geq \beta^*,$$

proving that no admissible policy (q, L) can achieve a long-run average cost less than β^*. □

6 Proposed policies

Recall that the optimal scheduling policy is determined through $q^*(\cdot)$, which, in turn, is given by

$$q^*(x) = \operatorname*{argmin}_{q \in \mathcal{A}(x)} \big\{ g(q) - \theta(q) v(x) \big\}. \tag{56}$$

Next, in each of the cases considered, we will solve (56) and propose a scheduling policy accordingly.

6.1 The linear delay cost case

Using (31)–(32) and the change of variable $z_k = q_k m_k / x$ for $k = 1, \ldots, K$, (56) leads to

$$z^*(x) = \operatorname{argmin} \left\{ \sum_{k=1}^{K} \left(\frac{c_k + a_k \gamma_k}{m_k} - \gamma_k v(x) \right) z_k : z \geq 0, \sum_{k=1}^{K} z_k = 1 \right\},$$

from which it is clear that the solution is of bang-bang type. Also, defining

$$K^*(v) = \min \left\{ \sum_{k=1}^{K} \left(\frac{c_k + a_k \gamma_k}{m_k} - \gamma_k v \right) z_k : z \geq 0, \sum_{k=1}^{K} z_k = 1 \right\},$$

it is easy to see that $K^*(\cdot)$ is a piecewise linear convex function. Let $0 = s_0 < s_1 < \cdots < s_J < \bar{p}$ (for some $J \leq K$) be $K^*(\cdot)$'s breakpoints and $i_1, \ldots, i_J \in \{1, \ldots, K\}$ be the indices such that

$$K^*(v) = \frac{c_{i_j} + a_{i_j} \gamma_{i_j}}{m_{i_j}} - \gamma_{i_j} v \quad \text{for } v \in (s_{j-1}, s_j].$$

Then define

$$x_j = v^{-1}(s_j), \quad j = 1, \ldots, J,$$

where v^{-1} is the inverse of $v(\cdot)$. Clearly, $0 = x_1 < \cdots < x_J < \infty$, and $z^*_{i_j}(x) = 1$ for $x \in (x_{j-1}, x_j]$.

The scheduling policy we propose strives to keep all workload in class i_j when the workload is between thresholds x_{j-1} and x_j. For concreteness, we propose the following policy: When the workload is in $(x_{j-1}, x_j]$, give the lowest priority to class i_j and prioritize the other classes with respect to their class index (smaller indices have higher priority).

6.2 The convex delay cost case

As in the linear case, the change of variable $z_k = q_k m_k / x$ for $k = 1, \ldots, K$, reduces (56) to finding the maximizer of the following optimization problem:

$$\min \sum_{k=1}^{K} \left\{ \lambda_k h_k \left(\frac{z_k x}{m_k \lambda_k} \right) + \frac{a_k \gamma_k x}{m_k} z_k \right\} - x v(x) \sum_{k=1}^{K} \gamma_k z_k$$

subject to

$$\sum_{k=1}^{K} z_k = 1, \tag{57}$$

$$z \geq 0. \tag{58}$$

By the necessary and sufficient KKT conditions, see [14], z^* is optimal if and only if there exists $\delta_k \geq 0$ for $k = 1, \ldots, K$, and ν such that (57)–(58) hold, and that

$$\delta_k z_k = 0 \quad \text{for } k = 1, \ldots, K,$$

$$\frac{x}{m_k} h_k' \left(\frac{x}{m_k \lambda_k} z_k \right) + x \frac{a_k \gamma_k}{m_k} - x v(x) \gamma_k - \delta_k + \nu = 0.$$

Note that whenever $z_k > 0$, we must have $\delta_k = 0$ so that

$$\frac{1}{m_k} h_k' \left(\frac{x}{m_k \lambda_k} z_k \right) + \frac{a_k \gamma_k}{m_k} - v(x) \gamma_k = -\frac{\nu}{x}.$$

Since the right-hand side is the same for every nonempty class, the left-hand side must be identical for every nonempty class as well. Therefore, we propose the policy that gives priority to the class for which the left-hand side is largest. To be more specific, the server must give priority to the nonempty class for which

$$\frac{1}{m_k} \left(h_k^n \right)' \left(\frac{Q_k^n}{\lambda_k^n} \right) + \frac{a_k \gamma_k}{m_k} - v\left(\hat{W}^n \right) \gamma_k \tag{59}$$

is largest and does not idle unless the system is empty.

Next, we consider two special cases in which (59) reduces to pathwise rules:

Special Case I: Identical abandonment rates, i.e. $\gamma_k = \gamma$ for all k. In this case, the server must give priority to the nonempty class for which

$$\frac{1}{m_k} \left(h_k^n \right)' \left(\frac{Q_k^n}{\lambda_k^n} \right) + \frac{a_k \gamma}{m_k}$$

is largest. Note that the value function no longer plays a role.

Special Case II: Identical abandonment rates and abandonment penalties per hour of processing requirement, i.e. $\gamma_k = \gamma$ and $a_k/m_k = a$ for all k. Then the server must give priority to the nonempty class for which

$$\frac{1}{m_k}\left(h_k^n\right)'\left(\frac{Q_k^n}{\lambda_k^n}\right)$$

is largest, which is the $Gc\mu$ rule.

In the next subsection, we propose a scheduling policy for the convex–concave case, which presents additional challenges. The proposed policy exploits the convex hull approach proposed by Ata and Olsen [4] (also see [5]), and combines the index rule (59) with the policies proposed in [4, 5].

6.3 The convex–concave delay cost case

The idea behind the proposed scheduling policy is to ensure that when the workload is w, a cost rate of

$$\sum_{k=1}^{K}\left\{\lambda_k h_k\left(\frac{q_k^*(w)}{\lambda_k}\right) + a_k\gamma_k q_k^*(w)\right\}$$

is incurred asymptotically. When the workload is small, all classes are kept in the convex cost region and a modified version of the dynamic index rule (introduced for the convex case) is in operation. When the workload is large, however, a different approach is needed. To ensure asymptotic optimality each class is divided into two artificial subclasses so that the cost rate of the convex hull is incurred. In particular, we artificially segment class k into subclasses k_a and k_b, and give "priority" to subclass k_a over subclass k_b while providing a small but positive amount of service capacity to subclass k_b.

It is easiest to describe and implement the proposed policy in a discrete-review framework, see, for example, Ata and Kumar [3] for discrete-review policies for controlling stochastic networks. Here we choose a review-period of length κ^n for each system. Namely, we let

$$\kappa^n = z_1 n^{\alpha_1},$$

where $1/2 < \alpha_1 < 1$ and $z_1 > 0$ is a tuning parameter.

In the nth system, the system manager reviews the system status at times $t_j^n = j\kappa^n$ for $j = 0, 1, 2, \ldots$. At the beginning of each period, she decides how to allocate resources during that period so that she can process jobs present in the system at time t_j^n as prescribed by this plan (and any new arrivals should there be sufficient capacity). To describe the resource allocation decisions in each period, define subperiod lengths

$$\begin{aligned} l_f^n &= z_2 n^{\alpha_2}, \\ l_{k_b}^n &= z_3 n^{\alpha_3}, \quad k = 1, \ldots, K, \\ l_{k_a}^n &= \frac{\lambda_k}{\mu_k}\left(\kappa^n - l_f^n - \sum_{k=1}^{K} l_{k_b}^n\right), \quad k = 1, \ldots, K, \end{aligned}$$

where $1-\alpha_1 > \alpha_3 > \alpha_2$ and $\alpha_2 \in (\frac{1-\alpha_1}{2}, 1-\alpha_1)$ and $z_2, z_3 > 0$ are tuning parameters.

In each period, the server works on classes 1 through K (in that order) sequentially. For each class k, the server spends up to $l^n_{k_a}$ time units working on class k_a first (until she idles because class k_a queue is depleted), then she spends up to $l^n_{k_b}$ time units working on class k_b (until she idles because the queue is depleted). Then the server proceeds to serving class $k+1$ in the same manner. Once the server finishes serving classes $1,\ldots,K$ in this way, the remainder of the period is spent in the "flexible" mode, whereby whenever the server finishes processing a job, she next works on class k for which the following is largest (for the current workload level):

$$\frac{1}{m_k}\left(h^n_k\right)'\left(\frac{Q^n_k(t)}{\lambda^n_k}\right)+\frac{a_k\gamma_k}{m_k}-v\big(\hat{W}^n(t)\big)\gamma_k,$$

where the server works on class k_b if it is nonempty; otherwise she works on class k_a. Clearly, the service policy is nonidling.

The system manager updates her routing decision at the review points. To be more specific, at each review point t^n_j for $j=1,2,\ldots$ the system manager observes the system status. Let $Q_{k_a}(\cdot)$ and $Q_{k_b}(\cdot)$ be the queue-lengths of subclasses k_a and k_b, respectively, so that $Q_k(\cdot)=Q_{k_a}(\cdot)+Q_{k_b}(\cdot)$. She routes the first $\lfloor\mu_k l^n_{k_a}\rfloor$ class k jobs to subclass k_a and routes the next $\lceil\mu_k l^n_{k_b}\rceil$ jobs to subclass k_b. Then the next $[\lfloor\lambda_k b_k\sqrt{n}\rfloor-Q^n_{k_a}(t^n_j)]^+$ jobs are again routed to class k_a. Any further class k arrivals are routed to subclass k_b. Note that this routing policy ensures that $Q^n_{k_a}(t)\le\lambda_k b_k\sqrt{n}+\mu_k\tau^n_{ka}$ at all times. Note that when the backlog is small, each class operates in the convex region of its delay cost curve. Further, most of the workload is kept in subclass a; and subclass b receives a small number of jobs. However, as the backlog in class k increases, it is class k_b that absorbs the added backlog, possibly experiencing quite long delays, while subclass a is kept at a moderate length.

7 A simulation study

In this section, a numerical example with two classes is presented to illustrate the effectiveness of the proposed policies, and the importance of accounting for abandonments in designing scheduling policies. For brevity, we focus attention on the convex delay cost case, and compare our policy (see Sect. 6.2) with the generalized $c\mu$ policy. In the analysis to follow we choose the system parameter $n=100$. In the base case we consider, jobs of each class arrive according to a Poisson process with rate of $\lambda^n_k=95$ per hour, and the processing times are exponentially distributed with a mean of half an hour for both classes so that $\mu^n_k=200$ per hour (for $k=1,2$). Jobs of each class abandon at the rate of $\gamma_k=1$; each abandonment costs $a_k=0.2$. The delay costs are given by

$$c^n_1(x)=x^2\quad\text{and}\quad c^n_2(x)=2x^2.$$

In what follows, we compare the performance of our proposed policy with that of the generalized $c\mu$ rule. Note that the base case we consider corresponds to the special case II of Sect. 6.2 because $\gamma_1=\gamma_2$ and $a_1/m_1=a_2/m_2$. Hence, the two policies are identical. Therefore, in what follows we vary γ_1 and compare the two policies.

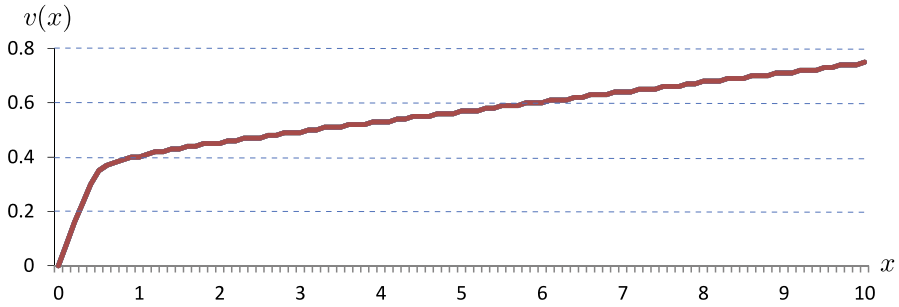


Fig. 1 The value function $v(\cdot)$ for the case of $\gamma_1 = 10$

Table 1 Long-run average costs under generalized $c\mu$ rule and proposed policy as a function of γ_1. The first column corresponds to the base case, in which proposed policy coincides with the generalized $c\mu$ rule. Standard errors of costs are in the order of 0.001, and thus are not shown for expositional clarity

γ_1	1	2	5	10
Generalized $c\mu$	1.934	2.303	3.251	4.383
Proposed policy	1.934	2.212	2.650	3.061

It is straightforward to see that the generalized $c\mu$ rule strives to achieve the following form of state space collapse:

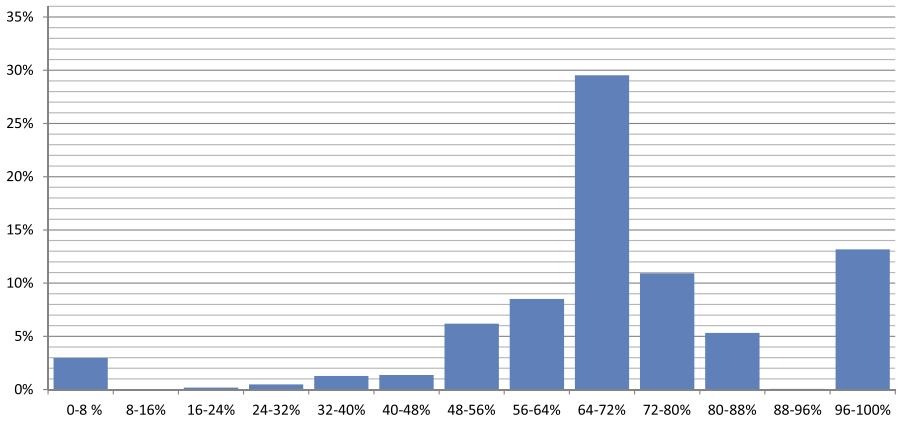
$$Q_1^n \approx 2Q_2^n$$

so that the instantaneous delay cost rate is minimized. Clearly, the generalized $c\mu$ rule prescribes keeping two thirds of the workload in class 1 and one third of it in class 2.

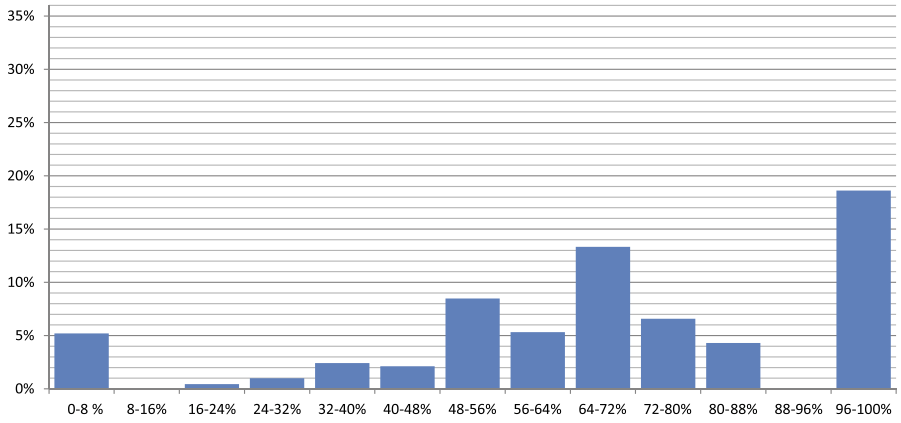
Although our policy coincides with the generalized $c\mu$ policy in the base case, for $\gamma_1 \neq \gamma_2$, it will differ from it as dictated by the index rule given in (59), which requires computing the relative value function $v(\cdot)$. It is straightforward to derive the limiting system parameters (with $n = 100$): $\lambda_1 = \lambda_2 = 1, m_1 = m_2 = 0.5, \sigma^2 = 1, \eta = 0.5, a_1 = a_2 = 0.2, \gamma_2 = 1$; $c_1(x) = 0.1x^2$ and $c_2(x) = 0.2x^2$. We will consider $\gamma_1 \in \{1, 2, 5, 10\}$. Using these parameters, we solve for $v(\cdot)$ numerically; it is shown in Fig. 1 for the case of $\gamma_1 = 10$. Repeating this for different values of γ_1, Table 1 provides the comparison of the long-run average costs under the two policies. Table 2 provides a breakdown of the costs across delay costs versus abandonment costs for each policy. To be specific, Table 1 shows how long-run average costs under generalized $c\mu$ rule and proposed policy change as γ_1 varies. The first column corresponds to the base case, in which $\gamma_1 = \gamma_2 = 1$ and the long-run average costs under the two policies are equal. In the rest of the columns, i.e. for $\gamma_1 \neq \gamma_2$, the proposed policy leads to lower long-run average costs than the generalized $c\mu$ policy. This is because the abandonment cost is much lower when the proposed policy is used while the delay cost is slightly higher; and the difference in abandonment costs is more significant than that of delay costs (see Table 2). Also note that, as the abandonment rate for class 1 increases, the long-run average costs under both policies increase since the

Table 2 Long-run average delay costs and abandonment costs under generalized $c\mu$ rule and proposed policy as a function of γ_1. The first column corresponds to the base case, in which proposed policy coincides with the generalized $c\mu$ rule. Standard errors of costs are in the order of 0.001, and thus are not shown for expositional clarity

γ_1	1	2	5	10
Generalized $c\mu$ delay costs	0.589	0.418	0.231	0.137
Proposed policy delay costs	0.589	0.472	0.411	0.377
Generalized $c\mu$ abandonment costs	1.345	1.885	3.020	4.246
Proposed policy abandonment costs	1.345	1.740	2.239	2.684



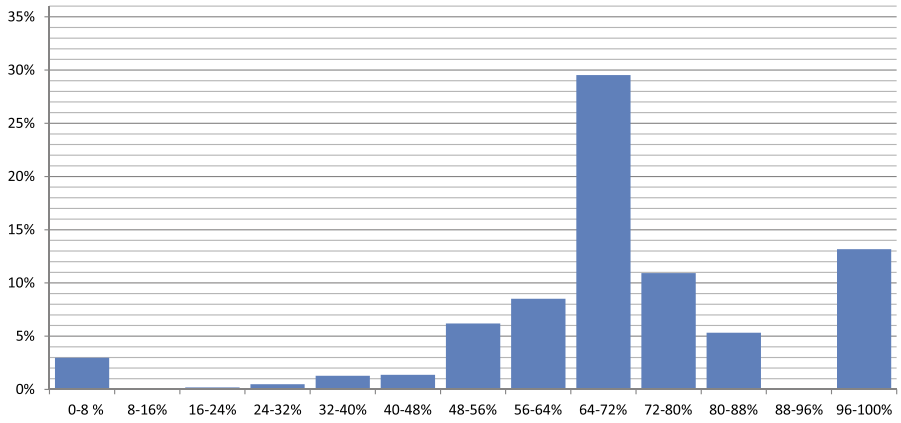
(a) $\gamma_1 = 1$



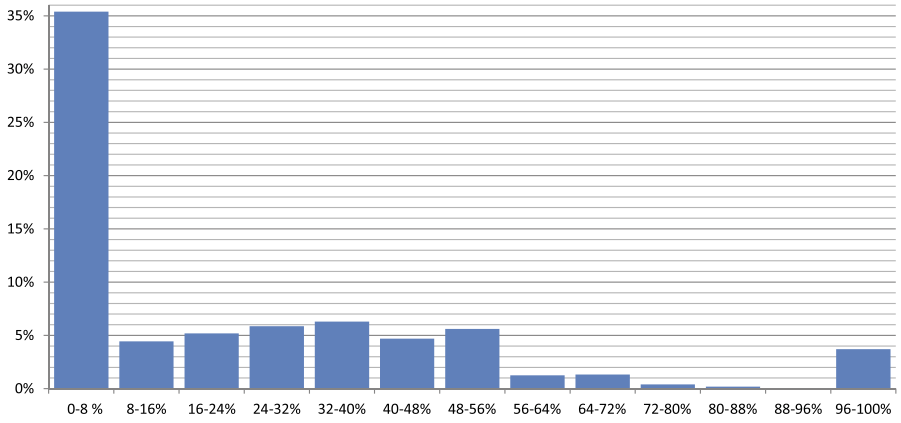
(b) $\gamma_1 = 10$

Fig. 2 The histograms of $Q_1/(Q_1 + Q_2)$ under generalized $c\mu$ policy for $\gamma_1 = 1$ and $\gamma_1 = 10$. The percentage of time there is no workload in the system is not shown in these histograms, and equal to 19.96 % and 32.05 %, respectively, when $\gamma_1 = 1$ and $\gamma_1 = 10$

cost of abandonments outweigh the savings realized in delay costs due to abandonments. It is also useful to look at the histogram of $Q_1/(Q_1 + Q_2)$ under each policy for $\gamma_1 = 1$ and $\gamma_1 = 10$; see Figs. 2 and 3.



(a) $\gamma_1 = 1$



(b) $\gamma_1 = 10$

Fig. 3 The histograms of $Q_1/(Q_1 + Q_2)$ under proposed policy for $\gamma_1 = 1$ and $\gamma_1 = 10$. The percentage of time there is no workload in the system is not shown in these histograms, and equal to 19.96 % and 25.67 %, respectively, when $\gamma_1 = 1$ and $\gamma_1 = 10$

Figure 2 shows how generalized $c\mu$ policy strives to distribute the workload so as to minimize the delay cost rate, i.e. to keep close to two thirds of the workload in class 1 most of the time, regardless of whether the abandonment rates for two classes are equal or not. As the abandonment rate of class 1 increases, more class 1 jobs abandon as they wait. Yet the generalized $c\mu$ policy strives to keep most of the workload in class 1, which leads to more class 1 abandonments, and a shorter class 1 queue than the server wishes (under the generalized $c\mu$ rule). Hence, the server increases the class 1 queue length by reducing its priority, which then leads to even more class 1 abandonments. In summary, ignoring the abandonments leads the generalized $c\mu$ policy to an erroneous cycle of scheduling decisions, causing it to be significantly suboptimal, cf. Table 1.

Our proposed scheduling policy gives priority to the nonempty class for which (59) is largest, thus incorporates abandonment behavior into scheduling considerations. That is, the server may serve jobs that do not have the highest immediate delay costs rate, but perhaps have a high abandonment rate and thus allow the system to forgo future abandonment costs. Therefore, the proposed policy strives to keep close

to two thirds of the workload in class 1, when the abandonment rates for two classes are equal (much like the generalized $c\mu$ policy—Figs. 2a and 3a are identical), and keep a much lower percentage of the workload in class 1, when the abandonment rates for class 1 is higher (see Fig. 3b).

8 Concluding remarks

We study dynamic scheduling of a multiclass queue with abandonments in the heavy traffic regime, where the objective is to minimize long-run average delay and abandonment costs. Three types of delay costs are considered: (i) linear delay costs, (ii) convex delay costs, and (iii) convex–concave delay costs. Since the dynamic control problem for the queueing system is not tractable analytically, we study the system in the conventional heavy traffic regime. Upon observing that the associated approximating Brownian control problem does not admit a pathwise solution due to abandonments, we solve the associated Bellman equation. The solution to the Bellman equation yields a dynamic index policy as the optimal control for the approximating Brownian control problem. We interpret that solution in the context of the original queueing system by proposing practical policies for each of the three cases considered and illustrate their effectiveness through a simulation study.

This paper makes the following contributions to the growing literature on the analysis and control of queueing systems with abandonments: First, in each of the cases considered, it derives a novel dynamic index policy by solving a Bellman equation, which depends on the second order problem data. Second, it highlights the important role abandonments play in controlling the queueing systems by showing that the pathwise policies $c\mu$, the generalized $c\mu$, and the cost balancing policy of Akan et al. [1] are no longer (asymptotically) optimal under abandonments. Third, it proposes simple, hence implementable, policies for the queueing system and illustrates their effectiveness in a simulation study. Fourth, the convex–concave case presents additional challenges which are overcome by the convex hull approach. Finally, it provides a novel method for constructing a solution to the Bellman equation arising in the analysis, which may be of interest in its own right.

Several possible generalizations and related questions are left for future research. First, we expect that the analysis of the convex case can be extended to allow $\lim_{x\to\infty} c_k'(x) = \infty$. It appears that all of our results except for Lemma 2 go through or can easily be extended in that case. However, given its technical nature and that the additional insights are limited, this case is not attempted here. Second, one can extend the analysis of the convex–concave case to allow $\lim_{x\to\infty} c_k'(x) \neq \lim_{x\to\infty} c_j'(x)$ for $k \neq j$. Third, another possibility is to consider general abandonment distributions. We expect that the derivative of the cumulative abandonment time distribution at zero will govern the abandonment behavior along the lines of Dai and He [20] and Mandelbaum and Momcilovic [41]. Fourth, yet another possibility is to consider hazard-rate scaling for the abandonments. We expect the analysis of this case to be challenging as it may lead to non-linear system dynamics.

Indeed, recent work by Kim and Ward [34] has considered this direction for $GI/GI/1+GI$ queue with two customer classes. The authors assume general abandonment distributions; the objective is to minimize long-run average abandonment

costs. Kim and Ward [34] consider this challenging problem in heavy traffic, and derive its approximating Brownian control problem, which has non-linear state dynamics. The authors solve the approximating Brownian control problem by solving the associated Bellman equation. Kim and Ward [34] also propose a policy for the original queueing system based on the solution of the Brownian control problem and illustrate its effectiveness by a simulation study.

Finally, another interesting future research direction is to prove asymptotic optimality of the proposed policies. We expect this to be challenging for especially the linear cost case (and the convex–concave cost case whose convex hull is linear for large delays), because the usual state-space collapse may not hold. The asymptotic optimality proofs rely typically on establishing the weak convergence of the vector-valued (scaled) queue length process, which, in turn, typically involves showing (i) the workload process converges; (ii) a state space collapse result holds whereby the vector-valued queue length process can be computed (or lifted up) from the lower dimensional workload process. Ghamami and Ward [21] observes that under linear delay/holding costs the optimal policy moves the workload between classes too frequently; see Remark 6.5 and the related discussion of Ghamami and Ward [21] for a detailed description of the "chatter" phenomenon. Therefore, for the cases with the linear and convex–concave delay costs, we expect to see a similar challenge. However, one may possibly adopt the approach of Ghamami and Ward [21] to prove the asymptotic optimality of the proposed policies. As a matter of fact, Ghamami and Ward [21] overcome this difficulty by showing that the integrals of the queue length processes converge. We expect that similar ideas can be used here too, which is left for future research. We also expect the asymptotic analysis to be somewhat easier for the case of convex delay cost. Indeed, we expect the usual state-space collapse result hold in that case.

Appendix A: Formal derivation of the approximating Brownian control problem

To facilitate the derivation of the Brownian control problem, note by the functional strong approximations, cf. [17], that

$$A_k^n(t) = \lambda_k^n t + \sqrt{\lambda_k^n}\ \hat{B}_k^n(t) + o\big(\sqrt{n}\big) \quad \text{for all } k, n, \tag{60}$$

$$S_k^n(t) = \mu_k^n t + \sqrt{\mu_k^n}\ \tilde{B}_k^n(t) + o\big(\sqrt{n}\big) \quad \text{for all } k, n, \tag{61}$$

where $\hat{B}_k$, $\tilde{B}_k$ for $(k = 1, \ldots, K)$ are independent standard Brownian motions and $o(\sqrt{n})/\sqrt{n} \to 0$ as $n \to \infty$.

Also, define the following scaled quantities for $n \geq 0$ and $t \geq 0$:

$$\hat{Y}^n(t) = \sqrt{n} Y^n(t), \qquad \hat{L}^n(t) = \sqrt{n} L^n(t),$$

$$\hat{\Gamma}^n(t) = \frac{\Gamma^n(t)}{\sqrt{n}}, \qquad \hat{Q}^n(t) = \frac{Q^n(t)}{\sqrt{n}},$$

$$\hat{W}^n(t) = \frac{W^n(t)}{\sqrt{n}}, \qquad \hat{C}^n(t) = \frac{C^n(t)}{\sqrt{n}}, \quad \text{and} \quad \hat{H}^n(t) = \frac{H^n(t)}{\sqrt{n}}.$$

The following observation facilitates our derivation as well: For $k = 1, \ldots, K$

$$\hat{\Gamma}_k^n(t) = \frac{1}{\sqrt{n}} N_k\left(\sqrt{n}\int_0^t \gamma_k\left[\hat{Q}_k^n(s) - \frac{1}{\sqrt{n}}\right]^+ ds\right), \quad t \geq 0. \tag{62}$$

The strong law of large numbers for Poisson processes implies that $N_k(\sqrt{n}t)/\sqrt{n} \to t$ as $n \to \infty$ almost surely for $t \geq 0$ and $k = 1, \ldots, K$. Therefore, we will replace $\hat{\Gamma}_k^n(t)$ by $\int_0^t \gamma_k \hat{Q}_k^n(s)\, ds$ in deriving the approximating Brownian control problem.

Then substituting (18), (60) and (61) into (2) and replacing $\hat{\Gamma}_k^n(t)$ by $\int_0^t \gamma_k \hat{Q}_k^n(s)\, ds$ in (2), we arrive at the following:

$$\hat{Q}_k^n(t) = -\mu_k t + \sqrt{2\lambda_k + \frac{o(\sqrt{n})}{\sqrt{n}}} B_k(t) - \int_0^t \gamma_k \hat{Q}_k^n(s)\, ds + \mu_k \hat{Y}_k^n(t) + \frac{o(\sqrt{n})}{\sqrt{n}}, \tag{63}$$

where B_k is a standard Brownian motion. Similarly, it follows from (18) that

$$\hat{L}^n(t) = \sum_{k=1}^K \hat{Y}_k^n(t) + \frac{o(\sqrt{n})}{\sqrt{n}}, \tag{64}$$

and (7) and (9) translate into the following under scaling:

$$\hat{L}^n(t) \quad \text{is nondecreasing with } \hat{L}^n(0) = 0, \tag{65}$$

$$\hat{Q}^n(t) \geq 0. \tag{66}$$

Also, using (16), the snapshot principle (17) and substituting $\int_0^t \gamma_k \hat{Q}_k^n(s)\, ds$ for $\hat{\Gamma}_k^n(t)$, the scaled cost function is approximated by

$$\hat{H}^n(t) = \int_0^t \sum_{k=1}^K \left(\lambda_k h_k\left(\frac{\hat{Q}_k^n(s)}{\lambda_k}\right) + a_k \gamma_k \hat{Q}_k^n(s)\right) ds. \tag{67}$$

Moreover, it follows from (4) and (62)–(64) that

$$\hat{W}^n(t) = B(t) - \int_0^t \left(\sum_{k=1}^K \gamma_k m_k \hat{Q}_k^n(s)\right) ds + \hat{L}^n(t) + \frac{o(\sqrt{n})}{\sqrt{n}}, \tag{68}$$

where $B(t) = \sum_{k=1}^K m_k B_k(t)$ for $t \geq 0$.

We arrive at the approximating Brownian control problem by passing to the limit in (63)–(68) formally. Namely, assuming $\hat{Y}^n \to Y$ as $n \to \infty$, we conclude that $\hat{Q}^n \to Q$, $\hat{W}^n \to W$ and $\hat{H}^n \to H$ as $n \to \infty$, where

$$\hat{Q}_k(t) = B_k(t) - \int_0^t \gamma_k \hat{Q}_k(s)\, ds + \mu_k \hat{Y}_k(t), \tag{69}$$

$$\hat{W}(t) = B(t) - \int_0^t \left(\sum_{k=1}^K \gamma_k m_k \hat{Q}_k(s)\right) ds + \hat{L}(t), \tag{70}$$

$$\hat{H}(t) = \int_0^t \left(\sum_{k=1}^{K} \lambda_k h_k \left(\frac{\hat{Q}_k(s)}{\lambda_k} \right) + a_k \gamma_k \hat{Q}_k(s) \right) ds, \tag{71}$$

$$\hat{L}(t) = \sum_{k=1}^{K} \hat{Y}_k(t), \tag{72}$$

$$\hat{Q}(t) \geq 0, \tag{73}$$

$$\hat{L}(t) \quad \text{is nondecreasing with } \hat{L}(0) = 0, \tag{74}$$

and the approximating Brownian control problem can be stated as

$$\min \varlimsup_{t\to\infty} \frac{1}{t} \mathbb{E}\big[\hat{H}(t)\big] \quad \text{subject to} \quad (69)\text{–}(74).$$

Appendix B: Auxiliary results and proofs of technical results

Proof of Lemma 1 It is immediate from (38) that $\psi(x, v)$ is decreasing in v. Also note from (31)–(32) that

$$\psi(x, v) = \min_{y \in \bar{\mathcal{A}}} \left\{ \frac{g(xy) - v\theta(xy)}{x} \right\}, \tag{75}$$

where the minimand is continuous in x, y, and v (for $x > 0$; and defining $\psi(0, v) = 0$ extends continuity everywhere). Consider two pairs (x_1, y_1) and (x_2, y_2), and assume without loss of generality that $\psi(x_1, v_1) \leq \psi(x_2, v_2)$. Clearly, there exist y^1, y^2 such that

$$\psi(x_i, v_i) = \frac{g(x_i y^i) - v_i \theta(x_i y^i)}{x_i}.$$

Then it follows from (75) that

$$\psi(x_1, v_1) \leq \psi(x_2, v_2) \leq \frac{g(x_2 y^1) - v_2 \theta(x_2 y^1)}{x_2}.$$

Thus,

$$\big|\psi(x_1, v_1) - \psi(x_2, v_2)\big| \leq \left| \frac{g(x_2 y^1) - v_2 \theta(x_2 y^1)}{x_2} - \frac{g(x_1 y^1) - v_1 \theta(x_1 y^1)}{x_1} \right|,$$

from which the continuity of ψ follows since g, θ are continuous.

For the Lipschitz continuity of ψ in v, we can repeat the same steps with $x_1 = x_2 = x$, which gives

$$\big|\psi(x_1, v_2) - \psi(x_2, v_1)\big| \leq |v_2 - v_1| \sum_{k=1}^{K} m_k \gamma_k y_k^1 \leq c_L |v_2 - v_1|. \qquad \square$$

Proof of Lemma 2 Let $X(t)$ be the reflected Brownian motion on $[0,\infty)$ with drift rate $-\eta<0$ and infinitesimal variance σ^2. For any admissible policy, we have

$$\mathbb{E}\big[f\big(W(t)\big)\big]\le\mathbb{E}\big[f\big(X(t)\big)\big]$$

because f is monotone and $X(t)$ is stochastically larger than $W^*(t)$, where the latter assertion follows because there are no abandonments involved in the evolution of process $X(\cdot)$. Therefore, it suffices to show that

$$\lim_{t\to\infty}\frac{\mathbb{E}[f(X(t))]}{t}=0.$$

To establish this, recall that $\bar{p}<\infty$. Then since $v^*(x)$ is increasing with $\lim_{x\to\infty}v^*(x)=\bar{p}$, we conclude that $v^*(x)\le\bar{p}$ for all x. Thus, $f(x)\le\bar{p}x$. Then

$$\frac{\mathbb{E}[f(X(t))]}{t}\le\frac{\bar{p}}{t}\mathbb{E}\big[X(t)\big].$$

But we also have (see [23]) that

$$\mathbb{E}\big[X(t)\big]\to\mathbb{E}\big[X(\infty)\big]=\int_0^\infty xe^{-\eta x}\,dx=\frac{1}{\eta}<\infty.$$

Therefore, $\mathbb{E}[f(X(t))]/t\to 0$ as $t\to\infty$. □

The following lemma is immediate from the convexity of h_k (for $k=1,\dots,K$) and establishes that ψ is monotone.

Lemma 3 *$\psi(x,p)$ is strictly increasing in x in the cases of convex or convex–concave delay costs, whereas it is independent of x in the linear delay cost case.*

Lemma 4 $\bar{p}=\min_k\lim_{x\to\infty}(h_k'(x)+a_k\gamma_k)/\gamma_k m_k<\infty$.

Proof of Lemma 4 For notational convenience, let $\alpha=\min_k\lim_{x\to\infty}(h_k'(x)+a_k\gamma_k)/\gamma_k m_k$. For any $p>\alpha$, it is easy to see that $\psi(x,p)<0$. Thus $\bar{p}\le p$ for all $p>\alpha$, which implies $\bar{p}\le\alpha$. Next, we argue that $\bar{p}\ge\alpha-\epsilon$ for $\epsilon>0$. To this end, fix $\epsilon>0$, then there exists x_0 such that

$$\frac{h_k'(x)+a_k\gamma_k}{\gamma_k m_k}>\alpha-\epsilon,\quad x>x_0.$$

Thus $\psi(x,\alpha-\epsilon)>0$ so that $\bar{p}\ge\alpha-\epsilon$ for $\epsilon>0$, from which we conclude that $\bar{p}\ge\alpha$. □

Lemma 5 *The following hold*:

(i) $\underline{x}(p)<\infty$ *for* $p\in(0,\bar{p})$.
(ii) $\psi(\underline{x}(p),p)=\eta p/\underline{x}(p)$.
(iii) $\underline{x}(p)$ *is strictly increasing in* p.
(iv) $\lim_{p\to\bar{p}}\underline{x}(p)=\infty$.
(v) $\psi(x,p)>\eta p/x$ *for* $x>\underline{x}(p)$.
(vi) $\phi(x,p)$ *is strictly increasing in* x *for* $x>\underline{x}(p)$.

Proof of Lemma 5

Part (i) Since $p < \bar{p}$ and ψ is monotone in x, there exist $\epsilon > 0$ sufficiently small and $x_1 < \infty$ sufficiently large such that $\psi(x, p) \geq \epsilon$ for all $x \geq x_1$. Similarly, there exists $x_2 < \infty$ sufficiently large such that $\eta p/x \leq \epsilon/2$ for all $x \geq x_2$. Then letting $x_0 = \max\{x_1, x_2\} < \infty$,

$$\psi(x_0, p) - \frac{\eta p}{x_0} \geq \frac{\epsilon}{2} > 0,$$

and thus, $\underline{x}(p) < x_0 < \infty$ by the monotonicity of $\psi(x, p)$ in x.

Part (ii) Suppose not. Then $\psi(\underline{x}(p), p) > \eta p/\underline{x}(p)$ which contradicts the fact that $\underline{x}(p)$ is the infimum since $\psi(\underline{x}(p) - \epsilon, p) > \eta p/(\underline{x}(p) - \epsilon)$ for $\epsilon > 0$ sufficiently small.

Part (iii) This is clear from the fact that $\psi(x, p) - \eta p/x$ is strictly increasing in x and strictly decreasing in p.

Part (iv) Suppose not. Then there exists $M > 0$ such that $\underline{x}(p) \leq M$ for all $p < \bar{p}$, which implies $\psi(M, \bar{p}) \geq \eta \bar{p}/M$. But then $\psi(2M, \bar{p}) - \eta \bar{p}/(2M) > 0$ because $\psi(x, p) - \eta p/x$ is strictly increasing in x. Thus, we conclude that $\psi(2M, \bar{p} + \epsilon) > \eta(\bar{p} + \epsilon)/2M$ for $\epsilon > 0$ sufficiently small, which contradicts the fact that $\bar{p}$ is the supremum of $p > 0$ such that $\lim_{x\to\infty} \psi(x, p) > 0$. Therefore, $\underline{x}(p) \to \infty$ as $p \to \bar{p}$.

Part (v) This follows from part (ii) and the fact that $\psi(x, p) - \eta p/x$ is strictly increasing in x.

Part (vi) Recall that

$$\phi(x, p) = x\big[\psi(x, p) - \eta p/x\big].$$

Then the result follows from the facts that $\psi(x, p) - \eta p/x \geq 0$ for $x \geq \underline{x}(p)$ and that it is strictly increasing in x.

□

Lemma 6 *Let* $x_2 > x_1 > 0$ *and* $p \in (0, \bar{p})$. *Suppose either* $\phi(x_2, p) > 0$ *or* $\phi(x_1, p) > 0$. *Then* $\phi(x_2, p) > \phi(x_1, p)$.

Proof of Lemma 6 First, assume $\phi(x_2, p) > 0$ and recall that

$$\begin{aligned}\phi(x_2, p) &= \min_{y\in\tilde{\mathcal{A}}} \left\{ \sum_{k=1}^{K} \left[\lambda_k \, h_k\left(\frac{y_k x_2}{\lambda_k}\right) + a_k \gamma_k y_k x_2 \right] - p \sum_{k=1}^{K} \gamma_k m_k y_k x_2 - p\eta \right\} \\ &= x_2 \min_{y\in\tilde{\mathcal{A}}} \left\{ \sum_{k=1}^{K} \left[\frac{\lambda_k}{x_2} \, h_k\left(\frac{y_k x_2}{\lambda_k}\right) + a_k \gamma_k y_k \right] - p \sum_{k=1}^{K} \gamma_k m_k y_k \right\} - p\eta.\end{aligned}$$

let y_k^* be the minimizer of the right-hand side. Then

$$\phi(x_2, p) = x_2 \left\{ \sum_{k=1}^{K} \left[\frac{\lambda_k}{x_2} \, h_k\left(\frac{y_k^* x_2}{\lambda_k}\right) + a_k \gamma_k y_k^* \right] - p \sum_{k=1}^{K} \gamma_k m_k y_k^* \right\} - p\eta,$$

where the first term on the right-hand side is positive because $\phi(x_2, p) > 0$. Thus, we conclude that

$$\begin{aligned}
\phi(x_2, p) &> x_1 \left\{ \sum_{k=1}^{K} \left[\frac{\lambda_k}{x_2} h_k\left(\frac{y_k^* x_2}{\lambda_k} \right) + a_k \gamma_k y_k^* \right] - p \sum_{k=1}^{K} \gamma_k m_k y_k^* \right\} - p\eta \\
&\geq x_1 \left\{ \sum_{k=1}^{K} \left[\frac{\lambda_k}{x_1} h_k\left(\frac{y_k^* x_1}{\lambda_k} \right) + a_k \gamma_k y_k^* \right] - p \sum_{k=1}^{K} \gamma_k m_k y_k^* \right\} - p\eta \\
&\geq x_1 \min_{y \in \tilde{\mathcal{A}}} \left\{ \sum_{k=1}^{K} \left[\frac{\lambda_k}{x_1} h_k\left(\frac{y_k x_1}{\lambda_k} \right) + a_k \gamma_k y_k \right] - p \sum_{k=1}^{K} \gamma_k m_k y_k \right\} - p\eta \\
&= \phi(x_1, p),
\end{aligned}$$

where the first inequality follows since the first term on the right-hand side is positive, the second inequality follows from convexity of $h_k(\cdot)$ and that $h_k(0) = 0$, and the third one follows from the min operation.

Alternatively, assume $\phi(x_1, p) > 0$. Then note that

$$\begin{aligned}
\phi(x_2, p) &= x_2 \min_{y \in \tilde{\mathcal{A}}} \left\{ \sum_{k=1}^{K} \left[\frac{\lambda_k}{x_2} h_k\left(\frac{y_k x_2}{\lambda_k} \right) + a_k \gamma_k y_k \right] - p \sum_{k=1}^{K} \gamma_k m_k y_k \right\} - p\eta \\
&= \frac{x_2}{x_1} x_1 \min_{y \in \tilde{\mathcal{A}}} \left\{ \sum_{k=1}^{K} \left[\frac{\lambda_k}{x_2} h_k\left(\frac{y_k x_2}{\lambda_k} \right) + a_k \gamma_k y_k \right] - p \sum_{k=1}^{K} \gamma_k m_k y_k \right\} - p\eta \\
&\geq \frac{x_2}{x_1} x_1 \min_{y \in \tilde{\mathcal{A}}} \left\{ \sum_{k=1}^{K} \left[\frac{\lambda_k}{x_1} h_k\left(\frac{y_k x_1}{\lambda_k} \right) + a_k \gamma_k y_k \right] - p \sum_{k=1}^{K} \gamma_k m_k y_k \right\} - p\eta \\
&> x_1 \min_{y \in \tilde{\mathcal{A}}} \left\{ \sum_{k=1}^{K} \left[\frac{\lambda_k}{x_1} h_k\left(\frac{y_k x_1}{\lambda_k} \right) + a_k \gamma_k y_k \right] - p \sum_{k=1}^{K} \gamma_k m_k y_k \right\} - p\eta \\
&= \phi(x_1, p),
\end{aligned}$$

where the first inequality follows from convexity of $h_k(\cdot)$ and that $h_k(0) = 0$, whereas the next inequality follows since the first term on the right-hand side is positive and $x_2/x_1 > 1$. □

Lemma 7 *Let v be the unique solution to* IVP($\hat{x}$) *for $\hat{x} > \underline{x}(p)$ and $p \in (0, \bar{p})$. Suppose there exists $x^* \in [0, \hat{x}]$ such that $v'(x^*) = 0$. Then*

$$\phi\big(x, v(x^*)\big) > \phi\big(x^*, v(x^*)\big) \quad \textit{for } x > x^*.$$

Proof of Lemma 7 Recall that $\phi(x, v) = x\psi(x, v) - \eta p$. Since $v'(x^*) = 0$, it follows from (45) that $\phi(x^*, v(x^*)) = \phi(\hat{x}, p) > 0$, which implies $\psi(x^*, v(x^*)) > \eta v(x^*)/x^* > 0$. Then for $x > x^*$, we have $\psi(x, v(x^*)) \geq \psi(x^*, v(x^*)) > 0$ by Lemma 3. Therefore for $x > x^*$,

$$\begin{aligned}
\phi\big(x, v(x^*)\big) &= x\psi\big(x, v(x^*)\big) - \eta v(x^*) \\
&> x^*\psi\big(x^*, v(x^*)\big) - \eta v(x^*) \\
&= \phi\big(x^*, v(x^*)\big).
\end{aligned}$$

□

Proof of Proposition 2 First let $(q, \hat{L})$ be an admissible policy for the workload problem with the associated workload process W, and define $\hat{Q}(t) = q(t, W(t))$. It is straightforward to check that $(\hat{Q}, \hat{L})$ is an admissible policy for the reduced Brownian control problem; and the two policies have the same cost. Next, let $(\hat{Q}, \hat{L})$ be an admissible policy for the reduced BCP. Then choose the workload configuration function q such that $q(t, W(t)) = Q(t)$ for $t \geq 0$. (Recall that we allow the workload configuration function q to depend on the sample path.) Clearly, $(q, \hat{L})$ is an admissible policy for the workload problem, and its cost is less than or equal to that of the policy $(\hat{Q}, \hat{L})$ for the reduced Brownian control problem. □

Proof of Proposition 3 Part (i). It follows from Lemma 1, the fact that $\phi(x, v) = x\psi(x, v) - \eta v$, and Picard's iteration arguments; see pages 89–98 of Boyce and DiPrima [13], that there exists $\delta > 0$ such that we have a unique continuously differentiable solution $v_{\hat{x}}$ on $[0, \delta]$. This result can be extended to the entire interval $[0, K]$ for all $K > 0$ (and hence to $[0, \infty)$) by mimicking the arguments on page 192 of Mandl [43].

Part (ii). Let $\hat{x}_2 > \hat{x}_1 > \underline{x}(p)$. We want to show that $v_{\hat{x}_2}(x) > v_{\hat{x}_1}(x)$ for all $x > 0$, where

$$\frac{1}{2}\sigma^2 v'_{\hat{x}_i}(x) = \phi(\hat{x}_i, p)x - \int_0^x \phi\big(s, v_{\hat{x}_i}(s)\big)\, ds, \quad i = 1, 2. \tag{76}$$

Suppose that $v_{\hat{x}_1}(x) \geq v_{\hat{x}_2}(x)$ for some $x > 0$. Let $x^* = \inf\{x \geq 0 : v_{\hat{x}_1}(x) \geq v_{\hat{x}_2}(x)\}$. If $x^* > 0$, then our hypothesis and the continuity of $v_{\hat{x}_1}$ and $v_{\hat{x}_2}$ guarantee that $v_{\hat{x}_1}(x^*) = v_{\hat{x}_2}(x^*)$, and that $v_{\hat{x}_1}(x) \leq v_{\hat{x}_2}(x)$ on $[0, x^*]$. Then it follows from (76) that

$$\begin{aligned}\frac{1}{2}\sigma^2\big[v_{\hat{x}_2}(x^*) - v_{\hat{x}_1}(x^*)\big] &= \big(\phi(\hat{x}_2, p) - \phi(\hat{x}_1, p)\big)x^* \\ &\quad - \int_0^{x^*} \big[\phi\big(s, v_{\hat{x}_2}(s)\big) - \phi\big(s, v_{\hat{x}_1}(s)\big)\big]\, ds.\end{aligned}$$

Since $\phi(s, \cdot)$ is nonincreasing (by Lemma 1), and $\phi(\hat{x}_2, p) > \phi(\hat{x}_1, p)$ (by part (vi)) of Lemma 5), we have

$$0 = v_{\hat{x}_2}(x^*) - v_{\hat{x}_1}(x^*) \geq \frac{2}{\sigma^2}\big[\phi(\hat{x}_2, p) - \phi(\hat{x}_1, p)\big]x^* > 0,$$

which is a contradiction.

If $x^* = 0$, then there exists a sequence $\{x_n\}$ such that $x_n \downarrow 0$ as $n \to \infty$ and $v_{\hat{x}_1}(x_n) \geq v_{\hat{x}_2}(x_n)$. In particular,

$$\frac{v_{\hat{x}_1}(x_n)}{x_n} \geq \frac{v_{\hat{x}_2}(x_n)}{x_n} \quad \text{for } n \geq 1.$$

Because $v_{\hat{x}_1}(0) = v_{\hat{x}_2}(0)$, taking the limit as $n \to \infty$ gives $v'_{\hat{x}_2}(0) \leq v'_{\hat{x}_1}(0)$, which in turn implies $\phi(\hat{x}_2, p) \leq \phi(\hat{x}_1, p)$ by (45), contradicting the fact that $\hat{x}_2 > \hat{x}_1 > \underline{x}(p)$ by part (vi) of Lemma 5.

Therefore, $\hat{x}_2 > \hat{x}_1 > \underline{x}(p)$ implies $v_{\hat{x}_1}(x) < v_{\hat{x}_2}(x)$ for all $x > 0$.

Part (iii) To show that $v_{\hat{x}}$ strictly increases to its maximum on $[0, \hat{x}]$, we proceed in two steps: The first step is to show that $v_{\hat{x}}$ weakly increases to its maximum, that

is, it is not decreasing at any point before it reaches its maximum. Suppose not. Then by continuity of $v_{\hat{x}}$ and its derivative, there exist $x_1, x_2 \in [0, \hat{x}]$ such that

$$x_1 < x_2,$$
$$0 = v'_{\hat{x}}(x_1) < v'_{\hat{x}}(x_2), \tag{77}$$
$$v_{\hat{x}}(x_1) = v_{\hat{x}}(x_2). \tag{78}$$

Then (45) and (77) imply that

$$\phi\big(x_1, v_{\hat{x}}(x_1)\big) = \phi(\hat{x}, p) - \frac{1}{2}\sigma^2 v'_{\hat{x}}(x_1) > \phi(\hat{x}, p) - \frac{1}{2}\sigma^2 v'_{\hat{x}}(x_2) = \phi\big(x_2, v_{\hat{x}}(x_2)\big).$$

Comparing this with (78), we have $\phi(x_1, v_{\hat{x}}(x_1)) > \phi(x_2, v_{\hat{x}}(x_1))$, which contradicts Lemma 7. Therefore, $v_{\hat{x}}$ must increase weakly to its maximum value on $[0, \hat{x}]$.

As the second step, we show that $v_{\hat{x}}$ cannot be constant on any interval. Thus, we conclude that it must strictly increase to its maximum. To see this, suppose that $v_{\hat{x}}$ is constant on some interval $[x_1, x_2]$. Then $v'_{\hat{x}}(x) = 0$ for $x \in [x_1, x_2]$, and therefore, it follows from (45) that $\phi(x, v_{\hat{x}}(x)) = \phi(\hat{x}, p)$ for $x \in [x_1, x_2]$. However, since $\phi(x_1, v_{\hat{x}}(x_1)) = \phi(\hat{x}, p) > 0$ and $v'_{\hat{x}}(x_1) = 0$, one can argue from Lemma 7 that

$$\phi\big(x, v_{\hat{x}}(x_1)\big) > \phi\big(x_1, v_{\hat{x}}(x_1)\big) \quad \text{for } x \in (x_1, x_2]. \tag{79}$$

But then we also have from (45) and $v'_{\hat{x}}(x) = 0$ for $x \in (x_1, x_2]$ that

$$\phi\big(x, v_{\hat{x}}(x_1)\big) = \phi\big(x, v_{\hat{x}}(x)\big) = \phi(\hat{x}, p),$$

which contradicts (79). Thus, $v_{\hat{x}}$ cannot be constant on any interval, and we conclude that it strictly increases to its maximum on $[0, \hat{x}]$. □

Proof of Proposition 4 That $\zeta(\cdot; p)$ is strictly increasing follows from part (ii) of Proposition 3 and (47). Also note from part (ii) of Lemma 5 and the fact that $\phi(x, v) = x\psi(x, v) - \eta v$ that $\phi(\underline{x}(p), p) = 0$. Combining this with the fact that $\phi(0, 0) = 0$ gives $v_{\underline{x}(p)}(\cdot; p) \equiv 0$. Therefore, $\zeta(\underline{x}(p); p) = 0$.

To show that $\lim_{\hat{x}\to\infty} \zeta(\hat{x}; p) = \infty$ for $p \in (0, \bar{p})$, note that

$$\frac{1}{2}\sigma^2 v_{\hat{x}}(x) = \phi(\hat{x}, p)x - \int_0^x \phi\big(s, v_{\hat{x}}(s)\big)\, ds,$$

from which it follows that for $x > 0$ sufficiently small (so that $v_{\hat{x}}(s) \geq 0$ for all $s \in (0, x)$)

$$\zeta(\hat{x}; p) \geq v_{\hat{x}}(x) \geq \frac{\sigma^2}{2}\left[\phi(\hat{x}, p)x - \int_0^x \phi(s, 0)\, ds\right].$$

That is,

$$\zeta(\hat{x}; p) \geq \frac{\sigma^2}{2}\left[\phi(\hat{x}, p)x - \int_0^x \phi(s, 0)\, ds\right]. \tag{80}$$

Moreover, as $\hat{x} \to \infty$, we have $\phi(\hat{x}, p) \to \infty$ because $\phi(\hat{x}, p) = \hat{x}\psi(\hat{x}, p) - \eta p$ and $\lim_{\hat{x}\to\infty} \psi(\hat{x}, p) > 0$ since $p < \bar{p}$. Therefore, the right-hand side of (80) tends to infinity, and hence, $\zeta(\hat{x}, p) \to \infty$.

To prove that ζ is continuous, we first prove that $v_{\hat{x}}(x)$ is continuous in $\hat{x}$, uniformly over compact intervals $[0, K]$, $K > 0$. To this end, let $\hat{x} > \underline{x}(p)$ and $\{\hat{x}_n\}$ be a sequence converging to $\hat{x}$ where $\hat{x}_n \geq \underline{x}(p)$. It suffices to show that $v_{\hat{x}_n}(x) \to v_{\hat{x}}(x)$ as $n \to \infty$ uniformly in x (over compact intervals). Recall that $\psi(x, v)$ is Lipschitz continuous in v uniformly in x (see Lemma 1) and that $\phi(x, p) = x\psi(x, v) - \eta v$. Therefore, $\phi(x, v)$ is Lipschitz continuous in v (uniformly in x when $x \in [0, K]$, i.e. over compact intervals). Pick K sufficiently large so that $\hat{x}_n \leq K < \infty$ for all n. Then we write

$$\begin{aligned}\frac{1}{2}\sigma^2\big|v_{\hat{x}_n}(x) - v_{\hat{x}_m}(x)\big| &\leq \big|\phi(\hat{x}_n, p) - \phi(\hat{x}_m, p)\big|x \\ &\quad + \int_0^x \big|\phi\big(s, v_{\hat{x}_n}(s)\big) - \phi\big(s, v_{\hat{x}_m}(s)\big)\big|\,ds, \\ &\leq \big|\phi(\hat{x}_n, p) - \phi(\hat{x}_m, p)\big|K + c_K \int_0^K \big|v_{\hat{x}_n}(s) - v_{\hat{x}_m}(s)\big|\,ds,\end{aligned}$$

where c_K is the uniform Lipschitz constant of $\phi(x, \cdot)$ for $x \in [0, K]$. Then by Gronwall's inequality, cf. p. 78 of Oksendal [44], it follows that

$$\begin{aligned}\big|v_{\hat{x}_n}(x) - v_{\hat{x}_m}(x)\big| &\leq \frac{2}{\sigma^2}\big|\phi(\hat{x}_n, p) - \phi(\hat{x}_m, p)\big|Kc_K \exp\big\{2c_K x/\sigma^2\big\}, \\ &\leq \frac{2}{\sigma^2}Kc_K \exp\big\{2c_K K/\sigma^2\big\}\big|\phi(\hat{x}_n, p) - \phi(\hat{x}_m, p)\big|.\end{aligned}$$

Therefore the sequence of functions $\{v_{\hat{x}_n}\}$ is a Cauchy sequence (uniformly in $x \in [0, K]$). Then for each $x \in [0, K]$, we have

$$\frac{1}{2}\sigma^2 \lim_{n\to\infty} v_{\hat{x}_n}(x) = \lim_{n\to\infty} \phi(\hat{x}_n, p)x - \lim_{n\to\infty} \int_0^x \phi\big(s, v_{\hat{x}_n}(s)\big)\,ds.$$

One can interchange the limit and the integral since $\phi(s, v_{\hat{x}_n}(s))$ converges uniformly in s as $n \to \infty$, which follows from the uniform convergence of $v_{\hat{x}_n}$ (on $[0, K]$) and the Lipschitz continuity of ϕ uniformly in $s \in [0, K]$. Then since $\phi(\cdot, p)$ is also continuous, the following holds:

$$\frac{1}{2}\sigma^2\tilde{v}(x) = \phi(\hat{x}, p)x - \int_0^x \phi\big(s, \tilde{v}(s)\big)\,ds, \quad x \in [0, K],$$

which shows that $\tilde{v}$ is continuously differentiable and solves the initial value problem IVP($\hat{x}$) on $[0, K]$. By the uniqueness of the solution to the initial value problem IVP($\hat{x}$) it follows that $\tilde{v} = v_{\hat{x}}$. Therefore, $v_{\hat{x}_n} \to v_{\hat{x}}$ as $n \to \infty$ uniformly over compact intervals.

We now combine these results to prove that $\zeta(\cdot; p)$ is continuous. To this end, fix $\hat{x}_1 > \underline{x}(p)$ and let $\epsilon > 0$. Since $v_{\hat{x}}(x)$ is continuous in $\hat{x}$ on compact intervals $[0, K]$ for each $K > 0$, there exists $\delta(K) \in (0, \underline{x}(p))$ such that $|v_{\hat{x}_1}(x) - v_{\hat{x}_2}(x)| < \epsilon/2$ for all $x \in [0, K]$ whenever $|\hat{x}_1 - \hat{x}_2| < \delta(K)$. Also for $K > \underline{x}(p)$, define

$$\underline{\phi}(K, p) = \inf\big\{x \in [0, \hat{x}],\ \hat{x} \in \big[\underline{x}(p), K\big] : \phi\big(x, v_{\hat{x}}(x)\big)\big\},$$

and observe that

$$\underline{\phi}(K, p) < \phi(K, p).$$

Moreover, observe from (45)–(46) that for all $\hat{x} \in (\underline{x}(p), K]$ and $x_1, x_2 \in [0, \hat{x}]$ that

$$v_{\hat{x}}(x_2) - v_{\hat{x}}(x_1) \leq \frac{\sigma^2}{2}\big[\phi(K, p) - \underline{\phi}(K, p)\big](x_2 - x_1). \tag{81}$$

Let K be sufficiently large, i.e. $K \geq 2\hat{x}_1$, and define

$$\hat{\delta}(K) = \min\left\{\delta(K), \frac{\epsilon}{\sigma^2[\phi(K, p), \underline{\phi}(K, p)]}\right\},$$

and consider $\hat{x}_2$ such that $|\hat{x}_1 - \hat{x}_2| < \hat{\delta}(K)$. Consider the following two cases:

Case 1: $\hat{x}_2 < \hat{x}_1$. Then $\zeta(\hat{x}_2) < \zeta(\hat{x}_1)$. Choose x_1^* such that $v_{\hat{x}_1}(x_1^*) = \zeta(\hat{x}_1)$. Then by (81) and definitions of x_1^* and ζ it follows that

$$\zeta(\hat{x}_2) = \max_{0 \leq x \leq \hat{x}_2} v_{\hat{x}_2}(x) \geq v_{\hat{x}_2}\big(x_1^* \wedge \hat{x}_2\big).$$

Then consider the following two subcases:

Case 1a: $x_1^* \leq \hat{x}_2$. Then $\zeta(\hat{x}_2) \geq v_{\hat{x}_2}(x_1^*) \geq v_{\hat{x}_1}(x_1^*) - \epsilon/2 = \zeta(\hat{x}_1) - \epsilon/2$.

Case 1b: $x_1^* > \hat{x}_2$. Then since $x_1^* \in (\hat{x}_2, \hat{x}_1]$, we have by (81) that

$$\zeta(\hat{x}_1) \geq v_{\hat{x}_2}(\hat{x}_2) \geq v_{\hat{x}_1}(\hat{x}_2) - \epsilon/2 \geq v_{\hat{x}_1}\big(x_1^*\big) - \epsilon = \zeta(\hat{x}_1) - \epsilon.$$

Therefore, in either case we have $\zeta(\hat{x}_1) \geq \zeta(\hat{x}_2) - \epsilon$, and combining this with $\zeta(\hat{x}_2) \leq \zeta(\hat{x}_1)$ gives $|\zeta(\hat{x}_2) - \zeta(\hat{x}_1)| < \epsilon$.

Case 2: $\hat{x}_2 > \hat{x}_1$. Then $\zeta(\hat{x}_1) < \zeta(\hat{x}_2)$. Choose x_2^* such that $v_{\hat{x}_2}(x_2^*) = \zeta(\hat{x}_2)$. Then by (81) and definitions of x_1^* and ζ, it follows that

$$\zeta(\hat{x}_1) = \max_{0 \leq x \leq \hat{x}_1} v_{\hat{x}_1}(x) \geq v_{\hat{x}_1}\big(x_2^* \wedge \hat{x}_1\big).$$

Then consider the following two subcases:

Case 2a: $x_2^* \leq \hat{x}_1$. Then $\zeta(\hat{x}_1) \geq v_{\hat{x}_1}(x_2^*) \geq v_{\hat{x}_2}(x_2^*) - \epsilon/2 = \zeta(\hat{x}_2) - \epsilon/2$.

Case 2b: $x_2^* > \hat{x}_1$. Then since $x_2^* \in (\hat{x}_1, \hat{x}_2]$, we have

$$\zeta(\hat{x}_1) \geq v_{\hat{x}_1}(\hat{x}_1) \geq v_{\hat{x}_2}(\hat{x}_1) - \epsilon/2 \geq v_{\hat{x}_2}\big(x_2^*\big) - \epsilon = \zeta(\hat{x}_2) - \epsilon.$$

Therefore, in either case we have $\zeta(\hat{x}_1) \geq \zeta(\hat{x}_2) - \epsilon$, and combining this with $\zeta(\hat{x}_1) \leq \zeta(\hat{x}_2)$ gives $|\zeta(\hat{x}_2) - \zeta(\hat{x}_1)| < \epsilon$.

Combining cases 1 and 2, we conclude that ζ is continuous. □

Proof of Corollary 1 It is clear from Proposition 4 that there exists $x(p)$ such that $\zeta(x(p); p) = p$. Moreover, by Proposition 3, $v_{x(p)}(\cdot)$ increases strictly to its maximum value of p on the interval $[0, x(p)]$. Denote this maximum by x^*. We argue that $x^* = x(p)$. Suppose not, i.e. $x^* < x(p)$. Then

$$\phi\big(x(p), p\big) = \phi\big(x^*, v_{x(p)}\big(x^*\big)\big) = \phi\big(x^*, p\big),$$

which is a contradiction by Lemma 6. Thus $x^* = x(p)$. Moreover, it follows from (44) that $v'(x(p)) = 2/\sigma^2[\phi(x(p), p) - \phi(x(p), v_{x(p)}(x(p))] = 0$. □

Proof of Proposition 5 Note that by construction $v(\cdot; p)$ solve the initial value problem IVP($x(p)$) on $[0, x(p)]$, and by Corollary 1, $v(\cdot; p)$ is continuously differentiable on $[0, \infty)$. Hence the result follows. □

Proof of Proposition 6 Part (i). Suppose not. Then there exist $0 < p_1 < p_2 < \bar{p}$ such that $x(p_2) \le x(p_1)$. Then

$$0 < \phi\big(x(p_2), p_2\big) \le \phi\big(x(p_1), p_2\big) < \phi\big(x(p_1), p_1\big),$$

where the first inequality follows from Lemma 3, part (ii) of Lemma 5 and that $\phi(x, v) = x\psi(x, v) - \eta v$, the second inequality follows from part (vi) of Lemma 5, and the last inequality follows since $\phi(x, \cdot)$ is strictly decreasing. Since $\phi(x(p_2), p_2) < \phi(x(p_1), p_1)$, we can argue as in the proof of part (ii) of Proposition 3 that

$$v_{x(p_1)}(x; p_1) > v_{x(p_2)}(x; p_2) \quad \text{for all } x > 0. \tag{82}$$

Then by definition of $v(\cdot; p)$ it follows that

$$p_1 = v\big(x(p_1), p_1\big) \ge v\big(x(p_2), p_1\big) > v\big(x(p_2), p_2\big) = p_2. \tag{83}$$

To be more specific, the first inequality follows from part (iii) of Proposition 3 and that $x(p_1) \ge x(p_2)$, and the second inequality follows from (82). But (83), i.e. $p_1 > p_2$ is clearly a contradiction. Therefore $x(p)$ is strictly increasing on $(0, \bar{p})$.

Part (ii). Let $0 < p_1 < p_2 < \bar{p}$ and consider

$$\phi\big(x(p_1), p_1\big) = \frac{1}{2}\sigma^2 v'_{x(p_1)}(x) + \phi\big(x, v_{x(p_1)}(x)\big) \quad \text{subject to} \quad v_{x(p_1)}(0) = 0, \tag{84}$$

$$\phi\big(x(p_2), p_2\big) = \frac{1}{2}\sigma^2 v'_{x(p_2)}(x) + \phi\big(x, v_{x(p_2)}(x)\big) \quad \text{subject to} \quad v_{x(p_2)}(0) = 0. \tag{85}$$

Suppose $\beta(p_2) = \phi(x_2(p_2), p_2) \le \phi(x(p_1), p_1) = \beta(p_1)$. Then we can argue as in the proof of part (ii) of Proposition 3 that $v_{x(p_2)}(x) \le v_{x(p_1)}(x)$ for all $x > 0$. Then using (84)–(85) and the fact that $\phi(x, v)$ is strictly decreasing in v, we conclude that $v'_{x(p_2)}(x) < v'_{x(p_1)}(x)$ for all $x > 0$. Note, however, that

$$v'_{x(p_1)}\big(x(p_1)\big) = 0 \ge v'_{x(p_2)}\big(x(p_1)\big),$$

which implies $x(p_2) \le x(p_1)$ because $v_{x(p_2)}(\cdot)$ increases strictly to its maximum (at $x(p_2)$) and $v'_{x(p_2)}(x(p_2)) = 0$. But clearly $x(p_2) \le x(p_1)$ contradicts part (i). Thus, $\beta(p_2) > \beta(p_1)$.

Part (iii). Since $\beta(p_2) > \beta(p_1)$ for $0 < p_1 < p_2 < \infty$, this follows along the lines of the proof of part (ii) of Proposition 3. □

Proof of Proposition 7 Part (i). Note that $\beta(p)$ is the long-run average cost in an auxiliary problem where the system manager can turn away arriving jobs, but incurs a rejection penalty of p for doing so per such job. (Given Proposition 5, it is straightforward to verify this along the lines of Theorem 1 of Rubino and Ata [54].) In this auxiliary system, consider the feasible policy which keeps all workload in buffer 1 and never turns away any jobs. Let $\hat{W}$ and $\hat{Q}$ denote the (limiting) workload and queue-length process under this policy ($\hat{Q}_k = 0$ for $k = 2, \ldots, K$). Clearly we have

$$\beta(p) \le \underline{\lim}_{t\to\infty} \frac{1}{t}\mathbb{E}\big[H(t)\big],$$

where

$$H(t) = \int_0^t \left[\lambda_1 h_1 \left(\frac{\hat{W}(s)}{\lambda_1 m_1} \right) + \frac{a_1 \gamma_1}{m_1} \hat{W}(s) \right] ds, \quad t \geq 0.$$

Also consider the reflected Brownian motion $X(t)$ on $[0, \infty)$ with drift rate $\eta < 0$ and infinitesimal variance σ^2. Note that $X(t)$ is stochastically larger than $W(t)$. Thus, by monotonicity of $h_2(\cdot)$

$$\frac{1}{t}\mathbb{E}\big[H(t)\big] \leq \frac{1}{t}\mathbb{E}\left[\int_0^t \left[\lambda_1 h_1 \left(\frac{X(s)}{\lambda_1 m_1} \right) + \frac{a_1 \gamma_1}{m_1} X(s) \right] ds \right],$$

but the right-hand side converges to (see [23]):

$$\mathbb{E}\left[\lambda_1 h_1 \left(\frac{X(\infty)}{\lambda_1 m_1} \right) + \frac{a_1 \gamma_1}{m_1} X(\infty) \right] < \infty,$$

because $X(\infty)$ has an exponential distribution with mean $\sigma^2/2m$. This gives a uniform upper bound on $\beta(p)$. Thus, $\beta^* < \infty$.

Part (ii). Recall that $x(p) > \underline{x}(p)$ by construction, and that $\lim_{p \to \bar{p}} \underline{x}(p) = \infty$ by part (iv) of Lemma 5. Hence the result follows.

Part (iii). Recall that by construction $0 \leq v(x; p) \leq p$ for all $x \geq 0$. Then letting $p \to \bar{p}$ gives

$$0 \leq v^*(x) \leq \bar{p} \quad \text{for all } x \geq 0. \tag{86}$$

Since $\bar{p} < \infty$ by Lemma 4, this proves that $v^*(x) < \infty$ for all $x \geq 0$. Also note that

$$v\big(x(p); p\big) \leq v^*\big(x(p)\big) \leq \bar{p}. \tag{87}$$

Since $v^*(\cdot)$ is nondecreasing, which it inherits from $v(\cdot; p)$, and that $x(p) \nearrow \infty$ as $p \to \bar{p}$, we conclude from (86)–(87) that $\lim_{x \to \infty} v^*(x) = \bar{p}$. □

References

1. Akan, M., Ata, B., Olsen, T.L.: Congestion-based leadtime quotation for heterogeneous customers with convex–concave delay costs: optimality of a cost-balancing policy based on convex hull functions. Oper. Res. (2011). doi:10.1287/opre.1120.1117
2. Ata, B.: Dynamic control of a multiclass queue with thin arrival streams. Oper. Res. **54**(5), 876–892 (2006)
3. Ata, B., Kumar, S.: Heavy traffic analysis of open processing networks with complete resource pooling: asymptotic optimality of discrete review policies. Ann. Appl. Probab. **15**(2), 331–391 (2005)
4. Ata, B., Olsen, T.L.: Near-optimal dynamic leadtime quotation and scheduling under convex–concave customer delay costs. Oper. Res. **57**(3), 753–768 (2009)
5. Ata, B., Olsen, T.L.: Congestion-based leadtime quotation and pricing for revenue maximization with heterogeneous customers. Working Paper, Northwestern University, Evanston, IL (2011)
6. Ata, B., Harrison, J., Shepp, L.: Drift rate control of a Brownian processing system. Ann. Appl. Probab. **15**(2), 1145–1160 (2005)
7. Ata, B., Skaro, A., Tayur, S.: Organjet: overcoming geographical disparities in access to deceased donor kidneys in the United States. Working Paper, Northwestern University, Evanston, IL (2011)
8. Atar, R., Giat, C., Shimkin, N.: The $c\mu/\theta$ rule for many-server queues with abandonment. Oper. Res. **58**(5), 1427–1439 (2010)
9. Atar, R., Giat, C., Shimkin, N.: On the asymptotic optimality of the $c\mu/\theta$ rule under ergodic cost. Queueing Syst. **67**(2), 127–144 (2011)

10. Bell, S., Williams, R.: Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: asymptotic optimality of a threshold policy. Ann. Appl. Probab. **11**, 608–649 (2001)
11. Bell, S., Williams, R.: Dynamic scheduling of a parallel server system in heavy traffic with complete resource pooling: asymptotic optimality of a threshold policy. Electron. J. Probab. **10**, 1044–1115 (2005)
12. Billingsley, P.: Convergence of Probability Measures, 2nd edn. Wiley-Interscience, New York (1999)
13. Boyce, W., DiPrima, R.: Elementary Differential Equations and Boundary Value Problems. Wiley, New York (1992)
14. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
15. Budhiraja, A., Ghosh, A.P., Lee, C.: An ergodic rate control problem for single class queueing networks. SIAM J. Control Optim. **49**, 1570–1606 (2011)
16. Celik, S., Maglaras, C.: Dynamic pricing and lead-time quotation for a multiclass make-to-order queue. Manag. Sci. **54**(6), 1132–1146 (2008)
17. Chen, H., Yao, D.D.: Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization. Springer, New York (2001)
18. Conway, R.W., Maxwell, W., Miller, L.: Theory of Scheduling. Addison-Wesley, Reading (1967)
19. Cox, D., Smith, W.: Queues. Methuen, London (1961)
20. Dai, J., He, S.: Customer abandonment in many-server queues. Math. Oper. Res. **35**(2), 347–362 (2010)
21. Ghamami, S., Ward, A.R.: Dynamic scheduling of an N-System with reneging. Working Paper (2010)
22. Ghosh, A.P., Weerasinghe, A.: Optimal buffer size and dynamic rate control for a queueing network with reneging in heavy traffic. Stoch. Process. Appl. **120**, 2103–2141 (2010)
23. Harrison, J.M.: Brownian Motion and Stochastic Flow Systems. Wiley, New York (1985)
24. Harrison, J.M.: Brownian models of queueing networks with heterogeneous customer populations. In: Fleming, W., Lions, P.L. (eds.) Stochastic Differential Systems, Stochastic Control Theory and Applications. IMA Volumes in Mathematics and its Applications, vol. 10, pp. 147–186. Springer, New York (1988)
25. Harrison, J.M.: Heavy traffic analysis of a system with parallel servers: asymptotic analysis of discrete-review policies. Ann. Appl. Probab. **8**, 822–848 (1998)
26. Harrison, J.M., Wein, L.M.: Scheduling networks of queues: heavy traffic analysis of a simple open network. Queueing Syst. **5**(4), 265–280 (1989)
27. Harrison, J.M., Wein, L.M.: Scheduling networks of queues: heavy traffic analysis of a two-station closed network. Oper. Res. **38**, 1052–1064 (1990)
28. Harrison, J.M., Zeevi, A.: Dynamic scheduling of a multi-class queue in the Halfin–Whitt heavy traffic regime. Oper. Res. **52**, 243–257 (2004)
29. Iglehart, D.L., Whitt, W.: Multiple channel queues in heavy traffic I. Adv. Appl. Probab. **2**(1), 150–177 (1970)
30. Iglehart, D.L., Whitt, W.: Multiple channel queues in heavy traffic II: sequences, networks, and batches. Adv. Appl. Probab. **2**(2), 355–369 (1970)
31. Iglehart, D.L., Whitt, W.: Multiple channel queues in heavy traffic III: random server selection. Adv. Appl. Probab. **2**(2), 370–375 (1970)
32. Kakalik, J.: Optimal dynamic operating policies for a service facility. Technical Report. OR Center, MIT, Cambridge, MA (1969)
33. Keskinocak, P., Tayur, S.: Due date management policies. In: Simchi-Levi, D., Wu, S.D., Shen, Z.M. (eds.) Handbook of Quantitative Supply Chain Analysis: Modeling in the E-Business Era. International Series in Operations Research and Management Science, pp. 485–556. Kluwer Academic, Norwell (2004)
34. Kim, J., Ward, A.R.: Dynamic scheduling of an GI/GI/1 + GI queue with two customer classes. Working Paper, Marshall School of Business, University of Southern California (2011)
35. Klimov, G.P.: Time-sharing service systems I. Theory Probab. Appl. **19**(3), 532–551 (1974)
36. Kocaga, Y.L., Ward, A.R.: Admission control for a multi-server queue with abandonment. Queueing Syst. **6**(3), 275–323 (2010)
37. Kostami, V., Ward, A.R.: Managing service systems with an offline waiting option and customer abandonment. Manuf. Serv. Oper. Manag. **11**(4), 644–656 (2009)
38. Kumar, S.: Two-server closed networks in heavy traffic: diffusion limits and asymptotic optimality. Ann. Appl. Probab. **10**, 930–961 (2000)
39. Laws, C.: Resource pooling in queueing networks with dynamic routing. Adv. Appl. Probab. **24**, 699–726 (1992)

40. Leclerc, F., Schmitt, B.H., Dube, L.: Waiting time and decision making: is time like money? J. Consum. Res. **22**(1), 110–119 (1995)
41. Mandelbaum, A., Momcilovic, P.: Queues with many servers and impatient customers. Math. Oper. Res. (2012). doi:10.1287/moor.1110.0530
42. Mandelbaum, A., Stolyar, A.L.: Scheduling flexible servers with convex delay costs: heavy-traffic optimality of the generalized $c\mu$-rule. Oper. Res. **52**(6), 836–855 (2004)
43. Mandl, P.: Analytic Treatment of One-Dimensional Markov Processes. Springer, Berlin (1968)
44. Oksendal, B.: Stochastic Differential Equations: An Introduction with Applications, 5th edn. Springer, New York (1998)
45. Ormeci-Matoglu, M., Vande Vate, J.: Drift control with changeover costs. Oper. Res. **59**, 427–439 (2011)
46. Plambeck, E., Ward, A.: Optimal control of a high-volume assemble-to-order system. Math. Oper. Res. **31**(3), 453–477 (2006)
47. Plambeck, E., Ward, A.: Optimal control of a high-volume assemble-to-order system with maximum leadtime quotation and expediting. Queueing Syst. **60**(1), 1–69 (2008)
48. Plambeck, E., Kumar, S., Harrison, J.M.: A multiclass queue in heavy traffic with throughput time constraints: asymptotically optimal dynamic controls. Queueing Syst. **39**(1), 23–54 (2001)
49. Randhawa, R.S., Kumar, S.: Usage restriction and subscription services: operational benefits with rational users. Manuf. Serv. Oper. Manag. **10**(3), 429–447 (2008)
50. Randhawa, R.S., Kumar, S.: Multi-server loss systems with subscribers. Math. Oper. Res. **34**(1), 142–179 (2009)
51. Reed, J.E., Tezcan, T.: Hazard rate scaling for the GI/M/N + GI queue. Working paper (2009)
52. Reed, J., 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**(3), 606–644 (2008)
53. Reiman, M.I.: Open queueing networks in heavy traffic. Math. Oper. Res. **9**(3), 441–458 (1984)
54. Rubino, M., Ata, B.: Dynamic control of a make-to-order parallel-server system with cancellations. Oper. Res. **57**(1), 94–108 (2009)
55. Stidham, S.J.: Analysis, design and control of queueing systems. Oper. Res. **50**(1), 197–216 (2002)
56. Stolyar, A.L.: Maxweight scheduling in a generalized switch: state space collapse and workload minimization in heavy traffic. Ann. Appl. Probab. **14**(1), 1–53 (2004)
57. Van Mieghem, J.A.: Dynamic scheduling with convex delay costs: the generalized $c\mu$ rule. Ann. Appl. Probab. **5**(3), 809–833 (1995)
58. Ward, A.: Asymptotic analysis of queueing systems with reneging: a survey of results for fifo, single class models. Surv. Oper. Res. Manag. Sci. **17**(1), 1–14 (2012)
59. Ward, A.R., Glynn, P.W.: A diffusion approximation for a Markovian queue with reneging. Queueing Syst. **43**(1/2), 103–128 (2003)
60. Ward, A.R., Glynn, P.W.: Properties of the reflected Ornstein–Uhlenbeck process. Queueing Syst. **44**(2), 109–123 (2003)
61. Ward, A.R., Glynn, P.W.: A diffusion approximation for a GI/GI/1 queue with balking of reneging. Queueing Syst. **50**(4), 371–400 (2005)
62. Wein, L.M.: Optimal control of a two-station Brownian network. Math. Oper. Res. **15**(2), 215–242 (1990)
63. Wein, L.M.: Scheduling networks of queues: heavy traffic analysis of a two-station network with controllable inputs. Oper. Res. **38**, 1065–1078 (1990)
64. Wein, L.M.: Due-date setting and priority sequencing in a multiclass M/G/1 queue. Manag. Sci. **37**(7), 834–850 (1991)
65. Wein, L.M.: Dynamic scheduling of a multiclass make-to-stock queue. Oper. Res. **40**, 724–735 (1992)
66. Wein, L.M., Veatch, M.: Scheduling a make-to-stock queue: index policies and hedging points. Oper. Res. **44**, 634–647 (1996)