Queueing Syst
DOI 10.1007/s11134-017-9537-y

Personalized queues: the customer view, via a fluid model of serving least-patient first

Avishai Mandelbaum[1] · **Petar Momčilović**[2]

Received: 26 June 2014 / Revised: 17 May 2017

Abstract In personalized queues, information at the level of individuals—customers or servers—affects system dynamics. Such information is becoming increasingly accessible, directly or statistically, as exemplified by personalized/precision medicine (customers) or call center workforce management (servers). In the present work, we take advantage of personalized information about *customers*, specifically knowledge of their actual (im)patience while waiting to be served. This waiting takes place in a many-server queue that alternates between over- and underloaded periods, hence a fluid view provides a natural modeling framework. The parsimonious fluid view enables us to parameterize and analyze *partial* information, and consequently calculate and understand the benefits from personalized customer information. We do this by comparing least-patience first (LPF) routing (personalized) against FCFS (relatively info-ignorant). An example of a resulting insight is that LPF can provide significant advantages over FCFS when the durations of overloaded periods are comparable to (im)patience times.

Keywords Multi-server queue · Time-varying queue · Fluid approximation/model · Earliest deadline first

Mathematics Subject Classification 60K25 · 90B22

✉ Petar Momčilović
petar@ise.ufl.edu

Avishai Mandelbaum
avim@tx.technion.ac.il

[1] Faculty of Industrial Engineering and Management, Technion, 3200 Haifa, Israel

[2] Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA

Published online: 21 June 2017

1 Introduction

A modeling paradigm for personalized queues In a *personalized* queueing system, say $M/M/n+M$ [17,47] for concreteness, interarrival times, service durations and (im)patience are still all exponentially distributed, as usual, but their realizations for *individual* customers and servers are assumed known, or partially known, prior to decision making—for example prior to admitting customers into the system or prior to matching them with servers. Personalized information is lacking from classical protocols, for example FCFS or LCFS or random order, which are oblivious to when exactly the next arrival will happen, or who is the least patient among the customers waiting to be served, or who is the fastest server among those available to serve.

Why "paradigm"? Because essentially every queueing model can be "personalized," by making individual realizations of its primitives available to its decision protocols, yet without altering the sub-models (distributions) of these primitives. While there exists ample queueing research that fits this "personalized" scheme, for example assigning high priority to a shortest processing time or to an earliest deadline, we believe that acknowledging a common timely theme across this dispersed research is of value—and hence worthy of the term "paradigm." Furthermore, in existing schemes full information is available, for example, individual service or patience times are known exactly. Yet of great importance is also the practical case of partial information, where knowledge about individual realizations is noisy (cf. triage process in emergency departments, the goal of which is to reduce such noise). A central challenge in our paradigm is thus the trade-off between information availability and performance.

We expect the paradigm of personalized queues to become increasingly practice-relevant with the proliferation of personalized data. One example is [16], which provides empirical support for a personalized server view—individual service durations. A second example is [18], which in fact motivated the present paper. It develops inference tools that enable the personalization of customer impatience in telephone queues. Making this personalized customer information available to discretionary control should yield a reduction in abandonment. Ultimately, one could combine the server and customer views to form a more general manager view: here one takes into account personalized information about both customers and servers.

On abandonment Customer abandonment is an effect that is prevalent in a variety of service systems, from telephone call centers through internet sites to emergency departments. It is typically desirable to reduce the abandonment rate, which often serves as a proxy for service quality and value: through abandoning, a customer is informing the service provider that the value of its service is unworthy of its wait. The terms "abandonment" and "customer impatience" are context dependent. For example, in call centers [40] or emergency departments [19], customer patience is the amount of time that a customer is *willing* or *able* to wait for service; in terror queues [31], abandonments correspond to terror attacks.

Nowhere is the significance of "abandonment" better encapsulated than in mass casualty events (MCEs). During an MCE, customer "patience" is the longest time period that a patient can survive without receiving medical care—an abandonment is thus death [11]. Furthermore, MCEs are incidents where medical resources

(personnel, equipment) are overwhelmed by the number and severity of casualties [44], for example, it is not uncommon that the arrival rate to a hospital emergency department (ED) triples or quadruples during such events. MCE workloads thus impose an extreme strain on hospital resources (under normal circumstances hospitals already operate close to their capacity—hence very long waiting times are ED routine). Consequently, hospitals often maintain emergency plans that facilitate treatment of a large number of casualties (note that despite such plans, medical personnel can experience ethical dilemmas [42] as treatment must still be rationed due to limited resources). Under such circumstances, a strategy that maximizes the number of saved lives is a natural goal—that is, a strategy that minimizes abandonment.

MCEs typify the realities that our models here capture: impatient customers that seek, at rates that are time-varying over a finite time horizon, service that is to be provided by multi-servers so as to minimize abandonment. In this context, personalized information about customers is naturally their *exact* time to abandon—their patience; and a policy that is a natural candidate for minimizing abandonment (and proved to be such in special cases—see [46,49]) is one that assigns the highest priority (non-preemptively) to a customer with the least patience.

Contributions In this paper, we introduce a many-server fluid model. It corresponds to the many-server $G_t/\mathrm{GI}/n+\mathrm{GI}$ queue, but it can and should be viewed on its own merit, namely a model for time-varying many-server queueing system with impatient customers (we take the view that our fluid model and $G_t/\mathrm{GI}/n+\mathrm{GI}$ are alternatives for capturing a given reality, each with its merits and flaws; and focusing on the former renders somewhat of less significance the fact that it can be proved a limit of the latter—such convergence is thus a fact that we do not establish formally here).

Fluid abandons the queue when its waiting times reaches its "patience." In the model with partial information, we assume that full information is available on individual realizations of estimated (random) individual patience times, rather than the patience times themselves. Patience times and their estimates are dependent and characterized by a joint density function. No information on service times is available to the scheduler. Customers (fluid) with shorter estimated patience times are given priority over customers with longer estimated patience times. That is, the (non-preemptive) least-patient first (LPF) policy based on estimated patience times is implemented. Such a model is very natural in the MCE context.

A benchmark for partial information is the model with full information. Although unrealistic in some applications (for example, MCEs), such a model is important as it provides bounds for the performance of models with partial information. In the full information LPF model, the scheduler has full knowledge of individual realizations of customer patience times (and, hence, residual patience times of customers awaiting service at any moment of time). In both the cases of full and partial information, we propose a numerical algorithm for evaluating relevant performance measures (queue length, abandonment rate, etc.) of the fluid model. (For numerical benchmarking and anecdotal interest only, we also consider an algorithm that corresponds to a fluid model operating under most patient first (MPF) routing.)

To be more specific, we focus on time-varying fluid models that alternate between over- and underloaded periods. As we now explain, these are circumstances when

the advantages (fewer customers abandon) of LPF over FCFS can become significant. In comparison, employing LPF instead of FCFS in a many-server queue, in the quality-and-efficiency driven (QED) regime [20,39,50], decreases the probability of abandonment from order $1/\sqrt{n}$ to $1/n$, with n being the number of servers; thus, for n large enough and practically speaking, QED service levels under FCFS are already too high to warrant a dramatic improvement (though, theoretically, $1/\sqrt{n}$ and $1/n$ do indeed differ significantly). Similarly, implementing LPF in a permanently overloaded queue does not yield significant results since a constant fraction of customers abandon regardless of the policy (unless service-time realizations can be taken into account). On the other hand, when over- and underloaded time intervals are present, a personalized policy can harmlessly shift the load in time (by delaying customers with long patience times), which effectively reduces overloaded periods that cause the abandonment.

One should note that the behavior of LPF differs from that of a multi-class system with *static* priorities. Indeed, the latter cannot mimic LPF, under which the "priority" of a customer awaiting service *continuously* increases as its remaining patience decreases with time.

A comment on terminology Readers would recognize that LPF policy has been traditionally referred to as earliest deadline first (EDF). Such terminology connotes system-imposed deadlines that are common in computing/communication and production/manufacturing systems—these operate mostly in steady state with a few servers [21,43,55]. In contrast, our LPF terminology fits patience which is inherently a personal characteristic of the customer—and this is prevalent in service systems with *time-varying* arrival rates and *many* servers.

Organization Our paper is organized as follows. Next we provide a brief literature review, which is followed by a specification of our fluid model in Sect. 3. The LPF policy under full information is considered in Sect. 4, accompanied by a corresponding numerical algorithm. Our model and algorithm for the case of partial information appear in Sect. 5. Based on numerical examples, we discuss various insights in Sect. 6. The paper concludes in Sect. 7 with some further observations and commentary.

2 Literature review

Support for fluid models of time-varying many-server stochastic queueing systems was provided by [37,38,41,48]. Analyses of many-server fluid models with abandonments have mostly focused on systems operating under the FCFS policy. In particular, a stationary model was studied in [56]. Formal fluid limits for this model were established in [30] by extending results for the model without abandonment from [32]. In [36], a network of fluid models was considered, while a system with time-varying capacity was investigated in [37]. A numerical algorithm for evaluating sample paths of FCFS many-server fluid models with constant capacities was proposed in [27]. A fluid limit of a multi-class many-server queueing network with abandonment and feedback is studied in [28].

Early analyses of EDF can be found in [22–24]. There exist several variants of this policy (preemptive/non-preemptive, etc.), with some shown to satisfy optimality

properties. In particular, the non-preemptive version is optimal for feasibility [13] (that is, if a collection of jobs can be scheduled in a way that ensures all the jobs complete by their deadline, the EDF policy will schedule this collection of jobs so they all complete by their deadline). In [49] it was argued that the EDF policy maximizes the expected number of customers that meet their deadlines, within the class of work-conserving non-preemptive policies, in the $M/G/1+G$ queue. Stability and optimality (under various cost functions) of EDF in single-server systems were examined in [46]. In the case when all customers are served (no abandonment), the EDF policy minimizes the lateness and tardiness of the jobs that are in the system at an arbitrary time [54], as well as any convex function of the average tardiness [45]. EDF scheduling was studied in the context of conventional heavy traffic, both without [14,33] and with abandonments [34]. A fluid limit of a heavily loaded EDF $M/M/1$ queue was considered in [12]. Fluid limits of $G/G/1+G$ queues under EDF were investigated in [6].

Our partial information framework relates to studies of multi-class systems where customer classes can be estimated/predicted [3,4]. Such models are considered under the assumption that one is capable of achieving certain classification rates. For example, this happens with nurses in emergency departments, who can estimate urgencies of patient conditions with reasonable accuracy. Typically, a Bayesian view is adopted, where classes are characterized by probability distributions of service/patience times rather than realizations associated with individual patients; for example, [35]. Such queueing models have been used to capture the triage process in emergency departments [51,52]. This multi-class approach and our framework have a common feature—customer characteristics are estimated/predicted based on data available at the customer's arrival time. In addition, one can also extract some information about individual customers based on their behavior in the system. For example, differentiation among customers present in the waiting room can be obtained by considering their (current) waiting durations (even in the case when all customers belong to the same class). In general, two customers that spent different amounts of time in the waiting room have different probabilities of abandoning the system (consider the conditional distribution of patience). This approach has been exploited in [7], where the authors argue for priority scheduling based on waiting times of customers present in the waiting room. Informally, when the hazard function of patience is increasing (decreasing), priority should be given to customers that spent more (less) time in the waiting room.

Finally, we remark that the trade-off between information availability and queueing performance has been examined in [53], albeit in a different context. The authors consider an overloaded single-server queue with admission control: the service and arrival rates are $1-p$ and $\lambda \in (1-p, 1)$, respectively. Under the constraint that jobs can be rejected up to a rate p, the authors analyzed a policy that minimizes average queue length, as a function of the time-window during which information on future arrivals is available.

3 The fluid model

A flow of fluid (deterministic divisible quantity) arrives at a system that consists of an unlimited waiting space and a service facility with a fixed finite processing

capacity $s > 0$ (throughout the paper we follow the notation and conventions of [56]). Let $Q(t)$ and $B(t)$ be the amount of fluid awaiting service and obtaining service at time $t \geq 0$, respectively. The total fluid inflow over an interval $[0, t]$ is $\Lambda(t)$, where Λ is an absolutely continuous function with $\Lambda(t) = \int_0^t \lambda(x)\,\mathrm{d}x$, $t \geq 0$; $\{\lambda(t),\ t \geq 0\}$ is a time-dependent arrival rate function. At time t, arriving fluid either enters the service facility, if there is space available ($B(t) < s$), or joins the waiting room otherwise ($B(t) = s$). The system satisfies the standard work-conservation condition: the queue is non-empty if and only if there exists no spare capacity.

Assumption 1 (*Work conservation and finite capacity*) For all $t \geq 0$,

$$(s - B(t))\, Q(t) = 0 \quad \text{and} \quad B(t) \leq s.$$

Let $X(t) = B(t) + Q(t)$ be the total amount of fluid in the system at time t. Then $Q(t) = (X(t) - s)^+$ and $B(t) = s - (s - X(t))^+ = X(t) \wedge s$; here and later, the symbols $\wedge$ and $\vee$ represent the minimum and maximum operators, respectively.

Fluid flows out of the system from either the waiting room—by abandoning, or from the service facility—after being served. Formally, a fraction $F(x)$ of fluid that entered the queue at time t abandons by time $t + x$, provided it has not entered service by then. In addition, a fraction $G(x)$ of any quantity of arriving fluid requires service of at most x time units after entering service. Here the functions F and G are given distribution functions, which are referred to as the abandonment and service distribution, respectively. Denote $\bar{G} := 1 - G$ and $\bar{F} := 1 - F$. A bivariate distribution H will serve as the joint distribution of true and estimated patience times (see Sect. 5).

As in [37], we consider a "smooth" model. Let $\mathbb{C}_p \subseteq \mathbb{D}$ be the set of piecewise-continuous real-valued functions, i.e., functions that have only finitely many discontinuity points in any finite interval, with left and right limits at each discontinuity point (within the interval); here $\mathbb{D}$ is the space of right-continuous functions with left limits. The following assumption implies that the arrival rate λ is bounded over finite intervals [9, p. 122].

Assumption 2 (*Smoothness*) Λ and G are differentiable functions with derivatives λ and g in $\mathbb{C}_p$; the distribution functions F and H have densities f and h.

The generality of the distributions F and G renders $Q(t)$ and $B(t)$ insufficient for capturing the state of the system at time t—a more detailed description is needed, which records the relevant history of fluid in the waiting room and service facility. There are multiple ways to describe the state of fluid awaiting service, which we elaborate on in the next sections. These multiple ways correspond to different models for information and scheduling policies.

As for fluid in service, introduce a two-parameter function B such that $B(t, x)$ is the total quantity of fluid in service at time $t \geq 0$ that has been in service for at most $x \geq 0$ time units; one has $B(t, \infty) = B(t)$. We follow the description of the service facility provided in [37] for a FCFS fluid model. Note that fluid in service obeys the same rules both in the FCFS model and the model we consider. Indeed, the LPF policy (like FCFS) does not use any information about service times. Thus, our focus will be on the description of the waiting room (in Sects. 4, 5). Presently, we provide some basic

description of the service facility for completeness. The remaining three assumptions in this section are from [37]—they ensure that the model of fluid in service is well defined; see [37] for details on how various performance measures can be evaluated. In particular, they imply that $B(t,\cdot)$ admits the representation

$$B(t,x) = \int_0^x b(t,u)\,\mathrm{d}u;$$

here $b(t,x)$ is the density of fluid that spent x time units in service at time t.

Assumption 3 (*Initial fluid in service*) Fluid in service at $t = 0$ satisfies

$$B(0,x) = \int_0^x b(0,u)\,\mathrm{d}u \quad \text{and} \quad B(0) \le s,$$

for some nonnegative integrable $b(0,\cdot) \in \mathbb{C}_p$ such that

$$\sup_{0\le s\le t} \int_0^\infty \frac{b(0,y)\,g(s+y)}{\bar{G}(y)}\,\mathrm{d}y < \infty.$$

Assumption 4 (*Fundamental service evolution equation*) For $t \ge 0$, $x \ge 0$ and $u \ge 0$:

$$b(t+u, x+u) = b(t,x)\frac{\bar{G}(x+u)}{\bar{G}(x)}.$$

Let $A(t) = \int_0^t \alpha(u)\,\mathrm{d}u$ be the total amount of fluid to abandon during the interval $[0,t]$, with $\alpha(t)$ being the abandonment rate at time $t \ge 0$ (it is defined in Sects. 4, 5). Similarly, introduce $E(t)$ to denote the amount of fluid that enters service in $[0,t]$. The total amount of fluid to complete service during the interval $[0,t]$ is denoted $S(t)$. We now deduce the following basic flow conservation equations, which hold for all $t \ge 0$:

$$Q(t) = Q(0) + A(t) - A(t) - E(t) \quad \text{and} \quad B(t) = B(0) + E(t) - S(t). \tag{1}$$

These totals are determined by instantaneous rates [37]:

$$E(t) := \int_0^t \gamma(u)\,\mathrm{d}u, \quad t \ge 0,$$

where $\gamma(t) := b(t,0)$ is the rate at which fluid enters service at time t; and

$$S(t) := \int_0^t \sigma(u)\,\mathrm{d}u, \quad t \ge 0,$$

where $\sigma(t)$ is the service completion rate at time t, and is defined by

$$\sigma(t) := \int_0^\infty b(t,x)\frac{g(x)}{\bar{G}(x)}\,\mathrm{d}x, \quad t \geq 0. \tag{2}$$

Furthermore, in the special case of an initially empty system ($q(0,x) = b(0,x) = 0$, for all $x \geq 0$, or just $X(0) = 0$), the following is known to hold [27]:

$$B(t) = \int_0^t \bar{G}(t-u)\,\mathrm{d}E(u) \tag{3}$$

and

$$E(t) = B(t) + \int_0^t B(t-u)\,\mathrm{d}U(u), \tag{4}$$

where U is the renewal function associated with G, characterized by the renewal equation [5, p. 143]:

$$U(t) = G(t) + \int_0^t U(t-u)\,\mathrm{d}G(u), \tag{5}$$

for $t \geq 0$.

An additional regularity condition will now be imposed to define overloaded and underloaded intervals. An overloaded interval starts at a time t_1 with (i) $Q(t_1) > 0$ or (ii) $Q(t_1) = 0$, $B(t_1) = s$ and $\lambda(t_1) > \sigma(t_1)$, and ends at

$$T_1 := \inf\{u \geq t_1 :\ Q(u) = 0 \quad \text{and} \quad \lambda(u) \leq \sigma(u)\}. \tag{6}$$

An underloaded interval starts at a time t_2 with (i) $B(t_2) < s$ or (ii) $B(t_2) = s$, $Q(t_2) = 0$ and $\lambda(t_2) \leq \sigma(t_2)$, and ends at

$$T_2 := \inf\{u \geq t_2 :\ B(u) = s \quad \text{and} \quad \lambda(u) > \sigma(u)\}.$$

The underloaded interval may contain subintervals that are regarded as critically loaded ($Q(t) = 0$, $B(t) = s$ and $\lambda(t) = \sigma(t)$).

Assumption 5 (*Finitely many switches in finite time*) There are only finitely many switches between overloaded and overloaded intervals in each finite time interval. Each underloaded interval is of positive length.

Remark 1 The last assumption can be eliminated by considering the equivalence of the fluid models in [26,30,37]; see [29] for details. The assumption simplifies the analysis, since one can focus on underloaded and overloaded intervals separately.

To conclude our model specification, it seems worthwhile reviewing its primitives. These are the service and patience time distributions (G and F) and the time-dependent

arrival rate (λ); then the initial states (at time $t = 0$) of the service facility (density $b(0, \cdot)$) and the waiting room (densities $q(0, \cdot)$ and $q(0, \cdot, \cdot)$ for the full and partial information cases—see Sects. 4, 5); note that the partial information framework requires also a joint density of true and estimated patience times (H—see Sect. 5). All other variables/processes are outputs of the model.

4 Full information: a benchmark

In this section, we consider the least-patient first scheduling policy that exploits full information. We define a two-parameter function Q, such that $Q(t, x)$ is the total quantity of fluid in the queue at time $t \geq 0$, with remaining patience at most $x \geq 0$:

$$Q(t, x) = \int_0^x q(t, u)\, \mathrm{d}u \quad \text{and} \quad Q(t, \infty) = Q(t); \tag{7}$$

$q(t, x)$ can be interpreted as the density of fluid awaiting service with the *remaining* patience x at time t (during an overloaded period). The representation (7) is due to Assumptions 6 and 7—see below. Without loss of generality, assume that the overloaded period begins at time 0 and ends at time T that satisfies (6); the value of T need not to be known in advance [37]. The state of the waiting room at time $t = 0$ is defined by fluid density $q(0, \cdot)$:

Assumption 6 (*Initial fluid awaiting service*) In the case of full information, fluid waiting service at $t = 0$ satisfies, for some $q(0, \cdot) \in \mathbb{C}_p$,

$$Q(0, x) = \int_0^x q(0, u)\, \mathrm{d}u \quad \text{and} \quad Q(0) < \infty.$$

4.1 Least-patient first

Under least-patient first scheduling, a quantity of fluid enters service only if no other fluid with lesser remaining patience is present in the waiting room. We define $p_\downarrow(t) \in [0, \infty]$ to be the remaining patience of the least-patient fluid awaiting service:

$$p_\downarrow(t) := \inf\{x \geq 0 : q(t, x) > 0\}; \tag{8}$$

we set $p_\downarrow(t) = \infty$ when $q(t, x) = 0$ for all $x \geq 0$ (the waiting room contains no fluid, $Q(t) = 0$)—hence, $p_\downarrow(T) = \infty$. At time t, the quantity $p_\downarrow(t)$ represents the boundary between remaining patience times of fluid that enters service and fluid that remains in the waiting room. During an overloaded period, the system can be in three states: (i) overloaded with abandonment; (ii) overloaded with no abandonment, and fluid in queue enters service; and (iii) overloaded with no abandonment, and fluid in queue does not enter service. Based on the above definition, we have, for $t \geq 0$,

$$Q(t) = \int_{p_\downarrow(t)}^\infty q(t, x)\, \mathrm{d}x. \tag{9}$$

We remark that the role of $p_\downarrow = \{p_\downarrow(t),\ t \geq 0\}$ in the analysis of the LPF system is similar to that of the boundary waiting time in the FCFS system [27,36]. Informally, $p_\downarrow$ is a key quantity—all relevant functions associated with the model can be derived from it. In general, $p_\downarrow$ need not be a continuous or differentiable function. However, for sufficiently smooth model primitives, $p_\downarrow$ is non-differentiable only at finitely many points on any finite interval (for an illustration, see Example 1 below). Due to the LPF policy, the abandonment rate at time t is defined by

$$\alpha(t) := (q(t,0) - b(t,0))^+; \tag{10}$$

recall that $b(t,0) = \gamma(t)$ is the rate at which fluid enters the service facility. Thus, only fluid with zero remaining patience times that cannot be accommodated in the service facility abandons the system. Note that $q(t,0) = 0$ implies $\alpha(t) = 0$.

A quantity of fluid is present in the waiting room only if its remaining patience (which decreases linearly) does not drop below the boundary value $p_\downarrow$ at any moment from the time of its arrival (not just at the arrival time). In particular, consider a quantity of fluid that arrives at the system with patience x at time u. This fluid is present in the waiting room at time $t \geq u$, if $x - y \geq p_\downarrow(u+y)$, for all $0 \leq y \leq t-u$, i.e., it is not sufficiently impatient on the time interval $[u,t]$; here, $x-y$ is the remaining patience time after y time units spent in the waiting room. Next, let $p_\downarrow(t,u)$ be the initial (at arrival) patience of the least-patient fluid that arrived at time u and is still present in the waiting room at time $t \geq u$ (see Fig. 1). Based on the preceding, the value of $p_\downarrow(t,u)$ is a solution of the following optimization problem: $\min z$, s.t. $z - y \geq p_\downarrow(u+y)$, $\forall y \in [0, t-u]$. Note that the constraint can be rewritten:

$$\begin{aligned} z &\geq \sup_{0 \leq y \leq t-u} \{y + p_\downarrow(u+y)\} \\ &= \sup_{u \leq x \leq t} \{x - u + p_\downarrow(x)\}, \end{aligned}$$

and consequently

$$p_\downarrow(t,u) = \sup_{u \leq x \leq t} \{x - u + p_\downarrow(x)\}; \tag{11}$$

a dual relation holds as well:

$$p_\downarrow(t) = \inf_{0 \leq u \leq t} \{p_\downarrow(t,u) - (t-u)\}. \tag{12}$$

For t such that $\dot{p}_\downarrow(t)$ exists, (11) implies

$$\frac{\partial p_\downarrow(t,u)}{\partial t} = (1 + \dot{p}_\downarrow(t))^+ 1_{\{p_\downarrow(t,u) = t-u+p_\downarrow(t)\}}. \tag{13}$$

(For a function x differentiable at t, we use $\dot{x}(t)$ to denote its derivative at t.)

The structure of the fluid content awaiting service at time t can be determined from $p_\downarrow$. Consider fluid in the waiting room with remaining patience x at time t. Such

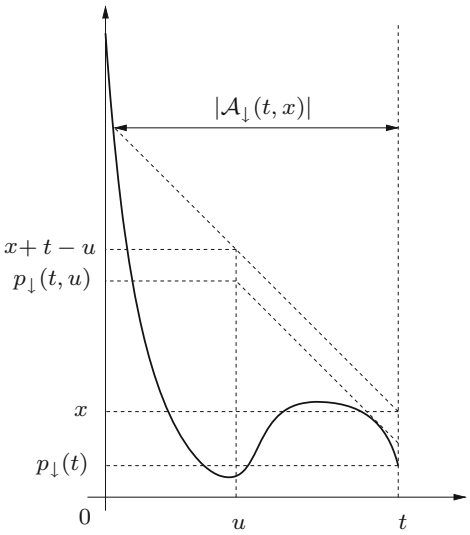

Fig. 1 LPF: an example of $p_\downarrow$. Related quantities $p_\downarrow(t,u)$ and $|\mathcal{A}_\downarrow(t,x)|$ are shown as well

fluid is either present in the system at $t = 0$ or has arrived during the time interval $\mathcal{A}_\downarrow(t,x) = \{u \in [0,t] : \ p_\downarrow(t,u) \le x+t-u\}$. The quantity $|\mathcal{A}_\downarrow(t,x)|$, the Lebesgue measure of $\mathcal{A}_\downarrow(t,x)$, represents the length of a time interval over which fluid with remaining patience x at time t is accumulated in the waiting room (see Fig. 1). The LPF policy and the fact that remaining patience times decrease linearly imply that the density q satisfies the following:

Assumption 7 (*Fundamental LPF evolution equation*) For $t \ge 0$ and $x \ge p_\downarrow(t)$,

$$q(t,x) = q(0,x+t)1_{\{x+t \ge p_\downarrow(t,0)\}} + \int_0^t 1_{\{p_\downarrow(t,u) \le x+t-u\}} \lambda(u)\, f(t-u+x)\, \mathrm{d}u. \tag{14}$$

The assumption is based on the fact that fluid with remaining patience time x at time t must arrive at the system at time $u \in [0,t]$ (or be in the system at time 0) with patience time $x + (t-u)$, and it should not leave the waiting room during the time interval $[u,t]$ (this condition is equivalent to $u \in \mathcal{A}_\downarrow(t,x)$). The first term accounts for fluid in the system initially: fluid with remaining patience $(x+t)$ at time 0 will have remaining patience x at time t, provided it did not leave the waiting room prior to time t (the event $\{x+t \ge p_\downarrow(t,0)\} = \{0 \in \mathcal{A}_\downarrow(t,x)\}$). The integral in (14) accounts for fluid not initially in the system. Assumptions 6 and 7 imply that $q(t,\cdot)$ is right-continuous, for all $t \ge 0$. Given $p_\downarrow$, (11) and (14) can be used to determine $q(t, p_\downarrow(t))$, the density of the "least-patient" fluid awaiting service at time t:

$$\begin{aligned} q(t,p_\downarrow(t)) = {} & q(0,p_\downarrow(t)+t)1_{\{p_\downarrow(t)+t=p_\downarrow(t,0)\}} \\ & + \int_0^t 1_{\{p_\downarrow(t,u)=p_\downarrow(t)+t-u\}} \lambda(u)\, f(t-u+p_\downarrow(t))\, \mathrm{d}u. \end{aligned} \tag{15}$$

Next, we derive an expression for $Q(t)$. To this end, substituting (14) in (9) yields

$$\begin{aligned} Q(t) &= \int_{t+p_\downarrow(t)}^{\infty} q(0,x)1_{\{x\geq p_\downarrow(t,0)\}}\,\mathrm{d}x + \int_{p_\downarrow(t)}^{\infty}\int_{\mathcal{A}_\downarrow(t,x)} \lambda(u)\, f(t+x-u)\,\mathrm{d}u\,\mathrm{d}x \\ &= \int_{p_\downarrow(t,0)}^{\infty} q(0,x)\,\mathrm{d}x + \int_{p_\downarrow(t)}^{\infty}\int_0^t \lambda(u)\, f(t+x-u)\,1_{\{p_\downarrow(t,u)+u-t\leq x\}}\,\mathrm{d}u\,\mathrm{d}x \\ &= \int_{p_\downarrow(t,0)}^{\infty} q(0,x)\,\mathrm{d}x + \int_0^t \int_{p_\downarrow(t,u)+u-t}^{\infty} \lambda(u)\, f(t+x-u)\,\mathrm{d}x\,\mathrm{d}u, \end{aligned} \tag{16}$$

where we used $\{x \geq p_\downarrow(t,0) \geq p_\downarrow(t)+t\}$ [see (11)] and $\{x \geq p_\downarrow(t,u)+u-t \geq p_\downarrow(t)\}$ [see (12)]. Rewriting (16) renders an expression for the total amount of fluid in the waiting room:

$$Q(t) = Q(0) - Q(0, p_\downarrow(t,0)) + \int_0^t \lambda(u)\,\bar{F}(p_\downarrow(t,u))\,\mathrm{d}u, \quad t \geq 0. \tag{17}$$

Note that the fraction $\bar{F}(p_\downarrow(t,u))$ of fluid arriving at the system at time u is present in the waiting room at time $t \geq u$. Thus, the integral in the preceding equality represents the amount of fluid not initially in the system (at time $t=0$) that is in the waiting room at time t.

Combining (1) and (17) yields an equation for $p_\downarrow$. In particular, $p_\downarrow$ is the maximal solution of

$$\int_0^t \lambda(u)\,F(p_\downarrow(t,u))\,\mathrm{d}u + Q(0, p_\downarrow(t,0)) - A(t) = E(t), \tag{18}$$

with $p_\downarrow(t) \geq 0$ and [see (10)]

$$\int_0^t 1_{\{p_\downarrow(u)>0\}}\,\mathrm{d}A(u) = 0. \tag{19}$$

Since, in general, F and $Q(0,\cdot)$ can be constants on certain intervals, there could exist multiple solutions of (18). The function $p_\downarrow$ corresponds to the maximal solution due to (8). The value of $p_\downarrow(0)$ is determined by (8) and $q(0,\cdot)$ (Assumption 6). We note that $p_\downarrow$ appears only on the left-hand side of (18), while the right-hand side is known.

Assumption 8 For sufficiently smooth model primitives, $p_\downarrow(\cdot)$ is non-differentiable only at finitely many points on any finite interval. Moreover, there exists a unique solution $p_\downarrow(\cdot) \geq 0$ of (18) under (19).

The last assumption is motivated by monotonicity (proving existence and uniqueness is beyond the scope of the present paper). In particular, by introducing $\tilde{p}_\downarrow(t) \in \mathbb{R}$ such that $\tilde{p}_\downarrow(t) = p_\downarrow(t) - \alpha(t)$, (18) can be rewritten without constraints [$p_\downarrow(t) \geq 0$ and (19)]:

$$\int_0^t \lambda(u)\,F(\tilde{p}_\downarrow(t,u))\,\mathrm{d}u + Q(0, \tilde{p}_\downarrow(t,0)) - \int_0^t \tilde{p}_\downarrow^-(u)\,\mathrm{d}u = E(t),$$

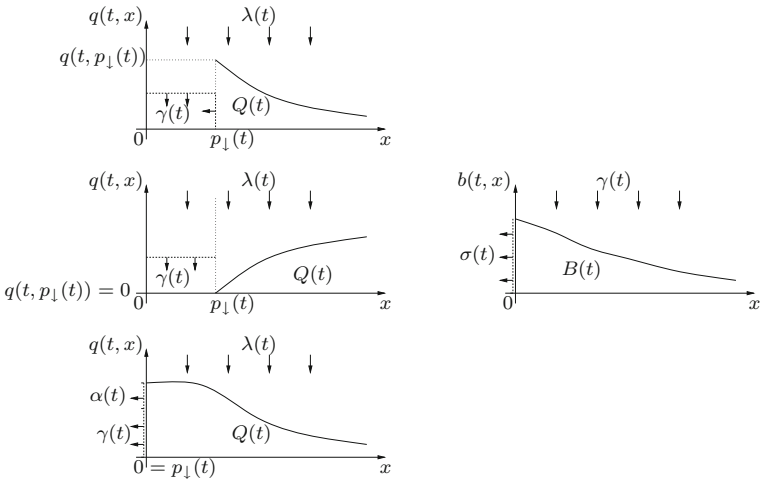

Fig. 2 LPF: examples of densities describing fluid in the waiting room (*left*) and service (*right*) at time t. When $p_\downarrow(t) > 0$ no abandonment occurs. Fluid enters the service facility either directly or via the waiting room

where $\tilde{p}_\downarrow(t,u) := \sup_{u\le x\le t}\{x-u+\tilde{p}^+_\downarrow(x)\}$. Then the left-hand side is monotonic in $\tilde{p}_\downarrow$, which describes both $p_\downarrow$ and α: $p_\downarrow(t) = \tilde{p}^+_\downarrow(t)$ and $\alpha(t) = \tilde{p}^-_\downarrow(t)$; note that $A(t) = \int_0^t \alpha(u)\,\mathrm{d}u = \int_0^t \tilde{p}^-_\downarrow(u)\,\mathrm{d}u$. Our numerical algorithm is based on a discrete-time analogue of (18)—see Sect. 4.4 for details. There, we argue that the above mentioned monotonicity implies existence and uniqueness of a discrete-time version of $p_\downarrow$. In Sect. 4.3, we provide two specific examples that illustrate how (18) characterizes $p_\downarrow$. In the next section, we derive a differential version of (18). It provides some insight into the LPF fluid model.

4.2 Differential version of (18)

A differential version of (18) can be obtained by considering $\gamma(t)$, the rate at which fluid enters the service facility. The case $p_\downarrow(t) = 0$ is straightforward: $\gamma(t) = q(t,0)-\alpha(t)$ [see (10)]. On the other hand, when $p_\downarrow(t) > 0$, fluid entering service does that either directly, or via the waiting room (see Fig. 2). The two cases can be combined into a single equation (for $t \ge 0$ such that $p_\downarrow$ is differentiable at t):

$$\gamma(t) + \alpha(t)1_{\{p_\downarrow(t)=0\}} = \lambda(t)F(p_\downarrow(t)) + (1+\dot{p}_\downarrow(t))^+ q(t, p_\downarrow(t)), \tag{20}$$

where the second term on the right-hand side represents the rate of fluid transfer between the waiting room and the service facility (recall that $q(t,\cdot)$ is right-continuous). Specifically, if $\dot{p}_\downarrow(t) \le -1$, then no fluid in the queue enters service, since $p_\downarrow(t)$ decreases at least as fast as the remaining patience of the least-patient fluid in the waiting room (which decreases at unit rate); the derivative being smaller than -1 is due to external arrivals with patience times smaller than $p_\downarrow(t)$. On the other

hand, when $\dot{p}_\downarrow(t) > -1$, the least-patient fluid in the waiting room (the density of such fluid is $q(t, p_\downarrow(t))$) enters service, since its remaining patience decreases below $p_\downarrow(t)$. Formally, for $t \geq 0$ such that $p_\downarrow(t) > 0$ (no abandonment, $\alpha(t) = 0$) and $\dot{p}_\downarrow(t)$ exists, (18) and (13) result in

$$\begin{aligned}\gamma(t) &= q(0, p_\downarrow(t, 0)) \frac{\partial p_\downarrow(t, 0)}{\partial t} \\ &\quad + \lambda(t) F(p_\downarrow(t)) + \int_0^t \lambda(u) f(p_\downarrow(t, u)) \frac{\partial p_\downarrow(t, u)}{\partial t} \, \mathrm{d}u \\ &= \lambda(t) F(p_\downarrow(t)) + q(0, p_\downarrow(t) + t) (1 + \dot{p}_\downarrow(t))^+ 1_{\{p_\downarrow(t,0) = t + p_\downarrow(t)\}} \\ &\quad + (1 + \dot{p}_\downarrow(t))^+ \int_0^t \lambda(u) f(t - u + p_\downarrow(t)) 1_{\{p_\downarrow(t,u) = t - u + p_\downarrow(t)\}} \, \mathrm{d}u. \end{aligned} \tag{21}$$

Combining (15) and (21) yields (for t such that $p_\downarrow(t) > 0$ and $\dot{p}_\downarrow(t)$ exists):

$$\gamma(t) = \lambda(t) F(p_\downarrow(t)) + (1 + \dot{p}_\downarrow(t))^+ q(t, p_\downarrow(t)).$$

Furthermore, (20) can be simplified by noting that $\dot{p}_\downarrow(t) < -1$ implies $q(t, p_\downarrow(t)) = 0$. Indeed, recall (15) and assume $1_{\{p_\downarrow(t,u) = p_\downarrow(t) + t - u\}} = 1$ for some $u \in [0, t]$; here u is an integration variable. Under this assumption $\partial p_\downarrow(t, u)/\partial t = \dot{p}_\downarrow(t) + 1 < 0$. However, (13) yields $\partial p_\downarrow(t, u)/\partial t = 0$. Hence, $p_\downarrow(t, u) \neq p_\downarrow(t) + t - u$, and (15) results in $q(t, p_\downarrow(t)) = 0$. Consequently, we have a differential version of (18):

$$\gamma(t) + \alpha(t) 1_{\{p_\downarrow(t) = 0\}} = \lambda(t) F(p_\downarrow(t)) + (1 + \dot{p}_\downarrow(t)) \, q(t, p_\downarrow(t)). \tag{22}$$

4.3 Examples

This subsection contains two examples.

Example 1 ($G_t/M/s + M$ *LPF fluid model*) Consider an initially empty ($b(0, x) = q(0, x) = 0$) fluid model with $\bar{G}(x) = \mathrm{e}^{-\mu x}$, $\bar{F}(x) = \mathrm{e}^{-\theta x}$, and $\lambda(t) = (1+\delta)\mu s$, for some $\delta > 0$. This system evolves through three time periods: (i) during $0 \leq t < t_1$, some spare processing capacity exists (underloaded period); (ii) during $t_1 \leq t < t_2$, there is no spare processing capacity, the queue is nonzero, and no fluid abandonment occurs; and (iii) during $t \geq t_2$, fluid abandonment occurs. Below we derive a detailed description of the evolution. Note that for $0 \leq t \leq t_1$, the departure rate satisfies (see (2) and [37, Proposition 2])

$$\sigma(t) = \int_0^t \lambda(t - u) \, g(u) \, \mathrm{d}u = (1 + \delta)\mu s \left(1 - \mathrm{e}^{-\mu t}\right),$$

and, thus, flow conservation (1) yields ($\Lambda(t_1) = S(t_1) + s$)

$$t_1 = \frac{1}{\mu} \log \frac{1 + \delta}{\delta}. \tag{23}$$

Furthermore, fluid enters service at rate [37]

$$\gamma(t) = \begin{cases} (1+\delta)\mu s, & t < t_1, \\ \mu s, & t \geq t_1, \end{cases} \tag{24}$$

while $\sigma(t) = s$ for $t \geq t_1$. Hence, $B(t) = E(t) - S(t)$ implies

$$B(t) = \begin{cases} (1+\delta)s\left(1 - \mathrm{e}^{-\mu t}\right), & t \leq t_1, \\ s, & t \geq t_1. \end{cases} \tag{25}$$

For $t < t_1$, the system is in an underloaded period, implying that $p_\downarrow(t) = \infty$, for $t \leq t_1$. The function $\{p_\downarrow(t),\ t > t_1\}$ is determined by (18):

$$(1+\delta)\mu s \int_{t_1}^{t} \left(1 - \mathrm{e}^{-\theta p_\downarrow(t,u)}\right) \mathrm{d}u = \mu s(t - t_1) + A(t),$$

where $E(t)$ is defined by (24). The solution of this equation is given by $p_\downarrow(t,u) = t - u + p_\downarrow(t)$, where

$$p_\downarrow(t) = \begin{cases} \infty, & 0 \leq t \leq t_1, \\ \frac{1}{\theta} \log \frac{(1+\delta)\left(1-\mathrm{e}^{-\theta(t-t_1)}\right)}{\delta\theta(t-t_1)}, & t_1 < t \leq t_2, \\ 0, & t \geq t_2, \end{cases} \tag{26}$$

and, consequently, for $t_1 < t < t_2$,

$$\dot{p}_\downarrow(t) = \frac{(1+\theta(t-t_1))\mathrm{e}^{-\theta(t-t_1)} - 1}{\theta(t-t_1)\left(1 - \mathrm{e}^{-\theta(t-t_1)}\right)} > -1, \tag{27}$$

with $\dot{p}_\downarrow(t_1+) = -1/2$; from (26) it follows that

$$p_\downarrow(t_1+) = \lim_{t \downarrow t_1} p_\downarrow(t) = \frac{1}{\theta} \log \frac{1+\delta}{\delta}.$$

Then, t_2 is the root of

$$\frac{1}{\theta} \log \frac{(1+\delta)\left(1 - \mathrm{e}^{-\theta(t-t_1)}\right)}{\delta\theta(t-t_1)} = 0,$$

or

$$\delta\theta(t_2 - t_1) = (1+\delta)\left(1 - \mathrm{e}^{-\theta(t_2-t_1)}\right). \tag{28}$$

Now, (26) and (15) imply

$$q(t, p_{\downarrow}(t)) = \begin{cases} \mu s \theta \delta (t - t_1), & t_1 \le t \le t_2, \\ (1+\delta)\mu s \left(1 - \mathrm{e}^{-\theta(t-t_1)}\right), & t \ge t_2, \end{cases} \tag{29}$$

which in turn yields [see (10)]

$$\alpha(t) = \mu s \left(\delta - (1+\delta)\mathrm{e}^{-\theta(t-t_1)}\right) 1_{\{t \ge t_2\}}; \tag{30}$$

note that $\alpha(t) \to \mu s \delta$, as $t \to \infty$. Based on (26), (27), (29) and (30), it is straightforward to verify that (22) holds. Note that, $p_{\downarrow}(t_1+)$ satisfies $\gamma(t_1+) = \lambda(t_1+)F(p_{\downarrow}(t_1+))$, since $q(t_1+, p_{\downarrow}(t_1+)) = 0$. Finally, observe that $\gamma(t) > \lambda F(p_{\downarrow}(t))$ for $t_1 < t < t_2$ (i.e., fluid enters service both directly and via the waiting room during the time interval (t_1, t_2)). Indeed, for $t_1 < t < t_2$, (24) and (26) imply

$$\begin{aligned} \gamma(t) - \lambda F(p_{\downarrow}(t)) &= \mu s - (1+\delta)\mu s \left[1 - \frac{\delta\theta(t-t_1)}{(1+\delta)\left(1 - \mathrm{e}^{-\theta(t-t_1)}\right)}\right] \\ &= \mu s \delta \left[\frac{\theta(t-t_1)}{1 - \mathrm{e}^{-\theta(t-t_1)}} - 1\right] > 0, \end{aligned}$$

where the inequality is due to $\mathrm{e}^{-x} > 1 - x$, for $x > 0$.

Example 2 ($G_t/D/s + D$ *LPF*: *a non-smooth fluid model*) Although we focus on smooth models, LPF fluid models can be considered under more general conditions. Here the formulation (18) comes to the rescue. For example, suppose $G(x) = 1_{\{x \ge 1/\mu\}}$ and $F(x) = 1_{\{x \ge d\}}$, with initial conditions given by $b(0, x) = \mu s 1_{\{0 \le x < 1/\mu\}}$ and $q(0, x) = 0$. Let the arrival rate satisfy $\lambda(t) = (1+\delta)\mu s$, for some $\delta > 0$. In that case, $b(t, x) = \mu s\, 1_{\{0 \le x < 1/\mu\}}$, and $E(t) = S(t) = \mu s t$, $t \ge 0$. Equation (18) renders an equation for $p_{\downarrow}$:

$$(1+\delta)\mu s \int_0^t 1_{\{p_{\downarrow}(t,u) \ge d\}}\, \mathrm{d}u - A(t) = \mu s t.$$

The solution of the preceding equation is given by

$$p_{\downarrow}(t) = \left(d - \frac{\delta}{1+\delta} t\right)^+, \quad t > 0;$$

thus, no fluid abandonment occurs before time $(1+\delta)d/\delta$. Moreover, $A(t) = \delta\mu s(t - d(1+\delta)/\delta)^+$ and $q(t, x) = (1+\delta)\mu s\, 1_{\{(d - \delta t/(1+\delta))^+ \le x < d\}}$.

4.4 A numerical algorithm

In this subsection, we provide an algorithm (Algorithm 1) for computing relevant functions of the fluid model under LPF. The algorithm is based on an algorithm for

the FCFS system [27]. As in [27], we require that the system is initially ($t = 0$) empty ($E(0) = B(0) = Q(0) = A(0) = 0$, implying $p_{\downarrow}(0) = \infty$); this allows us to utilize (4). The algorithm iteratively computes values of $E(t_i)$, $B(t_i)$, $Q(t_i)$, $A(t_i)$ and $p_{\downarrow}(t_i)$ for $t_i = i\delta$, $i = 1, 2, \ldots, n$, where δ is a time step. The iterative step depends on whether there exists spare capacity in the system.

Our algorithm requires evaluations of several integrals—see (31), (32) and (34). Here, we do not specify a scheme for numerically evaluating those integrals, because multiple methods can be used (based on the partition $0 = t_0 < t_1 < \cdots < t_n$). Note that determining the value of $U(t_i)$ is based on the integral equation (5). Multiple methods for evaluating the integral in (5) can be used as well. For example, using the trapezoidal rule yields

$$U(t_i) \leftarrow \frac{2G(t_i)}{2 - G(t_1)} + \sum_{j=2}^{i} \frac{G(t_j) - G(t_{j-1})}{2 - G(t_1)} \left(U(t_i - t_j) + U(t_i - t_{j-1})\right).$$

The rationale for the algorithm is as follows. Under the first case in the iteration ($B(t_{i-1}) < s$), some capacity is available at $t = t_{i-1}$, and one attempts to evaluate the system state at time $t = t_i$ under the same condition—thus, $E(t_i) \leftarrow \Lambda(t_i) - A(t_{i-1})$ (since there is no abandonment in $[t_{i-1}, t_i]$) and (31), which is based on (3). If it turns out that indeed $B(t_i) < s$, straightforward updates follow. However, if one obtains $B(t_i) = s$, the queue content at time t_i needs to be determined, along with other relevant quantities. To this end, the amount of fluid that entered service by t_i is computed via (32) [see (4)], and the balance equation (1) for the waiting room is utilized—the system of equations (33)–(36) is a discrete-time analogue of (18). Observe that the right-hand side in (33) is known, while the left-hand side depends on $p_{\downarrow}(t_i)$.

The quantity $Q(t_i)$ [evaluated based on (17)] is monotone in $p_{\downarrow}(t_i)$, and $A(t_i) - A(t_{i-1})$ is nonzero only if $p_{\downarrow}(t_i) = 0$. These two facts imply that there exists a maximum solution of (33)–(36). In particular, if $p_{\downarrow}(t_i) = 0$ implies $Q(t_i) + A(t_{i-1}) > \Lambda(t_i) - E(t_i)$, then the solution of (33)–(36) is positive (and unique). Otherwise, the solution is zero and $A(t_i) = \Lambda(t_i) - E(t_i) - Q(t_i)$. Equation (36) is a discrete-time version of (11). The second case in the iteration ($B(t_{i-1}) = s$) follows the same reasoning, except that one first attempts to verify that the system remains overloaded.

5 Partial information

We now consider the LPF policy under the assumption that only partial information about fluid patience is available. We model partial information by means of a bivariate distribution H, such that $H(x, y) = \int_0^x \int_0^y h(u, v)\, \mathrm{d}v\, \mathrm{d}u$ represents the fraction of arriving fluid with patience at most x and estimated patience at most y; then, $H(x, \infty) = F(x)$.

It is appropriate to think of $h(x, y)$ as the density of arriving fluid with true patience x and estimated (perceived) patience y. The distribution H defines two relevant conditional distributions. For fluid with (true) patience x, the conditional density

Algorithm 1

$B(0) \leftarrow 0, E(0) \leftarrow 0, Q(0) \leftarrow 0, A(0) \leftarrow 0, U(0) \leftarrow 0$ ▷ initialization

$p_{\downarrow}(0) \leftarrow \infty, p_{\downarrow}(0,0) \leftarrow \infty$ ▷ initialization

for $i = 1, \ldots, n$ **do** ▷ iterative step

  if $B(t_{i-1}) < s$ **then**

    $E(t_i) \leftarrow \Lambda(t_i) - A(t_{i-1})$

$$B(t_i) \leftarrow s \wedge \int_0^{t_i} \bar{G}(t_i - u)\, \mathrm{d}E(u) \tag{31}$$

    if $B(t_i) < s$ **then** ▷ empty waiting room

      $p_{\downarrow}(t_i) \leftarrow \infty, Q(t_i) \leftarrow 0, A(t_i) \leftarrow A(t_{i-1})$

    else ▷ $B(t_i) = s$

      evaluate $U(t_i)$ based in (5)

$$E(t_i) \leftarrow B(t_i) + \int_0^{t_i} B(t_i - u)\, \mathrm{d}U(u) \tag{32}$$

      $p_{\downarrow}(t_i) \leftarrow$ maximum solution of nonlinear equations (33)–(36):

$$Q(t_i) + A(t_i) = \Lambda(t_i) - E(t_i) \tag{33}$$

$$Q(t_i) = \int_0^{t_i} \lambda(u)\, \bar{F}(p_{\downarrow}(t_i, u))\, \mathrm{d}u \tag{34}$$

$$A(t_i) = A(t_{i-1}) + \left(\Lambda(t_i) - E(t_i) - A(t_{i-1}) - Q(t_i)\right)^{+} 1_{\{p_{\downarrow}(t_i)=0\}} \tag{35}$$

$$p_{\downarrow}(t_i, t_j) = \max_{j \le k < i} \{t_k - t_j + p_{\downarrow}(t_k)\} \vee \left(t_i - t_j + p_{\downarrow}(t_i)\right), \ j = 0, \ldots, i \tag{36}$$

      evaluate $\{p_{\downarrow}(t_i, t_j)\}_{j=0}^{i}$ based on (36)

      evaluate $Q(t_i)$ based on (34)

      evaluate $A(t_i)$ based on (35)

  else ▷ $B(t_{i-1}) = s$

    $B(t_i) \leftarrow s$

    evaluate $E(t_i)$ based on (32)

    if $\Lambda(t_i) - \Lambda(t_{i-1}) + Q(t_i - 1) \le E(t_i) - E(t_{i-1})$ **then** ▷ empty waiting room

      $E(t_i) \leftarrow \Lambda(t_i) - A(t_{i-1})$

      $p_{\downarrow}(t_i) \leftarrow \infty, Q(t_i) \leftarrow 0, A(t_i) \leftarrow A(t_{i-1})$

      evaluate $B(t_i)$ based on (31)

    else ▷ non-empty waiting room

      $p_{\downarrow}(t_i) \leftarrow$ maximum solution of (33)–(36)

      evaluate $\{p_{\downarrow}(t_i, t_j)\}_{j=0}^{i}$ based on (36)

      evaluate $Q(t_i)$ based on (34)

      evaluate $A(t_i)$ based on (35)

of estimated patience at y is given by $h(x, y)/\int_0^\infty h(x, v)\,\mathrm{d}v$. Similarly, given that estimated patience is equal to y, the conditional density of actual patience at x is given by $h(x, y)/\int_0^\infty h(u, y)\,\mathrm{d}u$. Both conditional distributions can be estimated from (censored) data via statistical analysis (procedures to estimate individual patience times are beyond the scope of this work, and are left for future research. Relevant references include [1,8,10,15,40]).

We focus on a model where patience times are estimated only once—upon arrival. Such a setup does arise in mass casualty events where triage is employed, or in call centers that opt for such protocols. One could also consider models where patience is (re)estimated periodically or continuously. In such models, the scheduling priority (based on re-estimated patience) would change as new information becomes available. The fact that fluid spent a certain amount of time awaiting service provides some information about its patience, since it statistically distinguishes it from fluid that had the same characteristics upon arrival but that has abandoned the system. Additional personalized information could be obtained by proactively acquiring it (for example, obtaining and/or providing information while waiting for a phone service, or via patient reexamination in emergency departments). We also note that our estimates of patience are numbers—a scheme that is appealing since it is straightforward to keep track of such estimates. In a more general setting, probability distributions can be used to describe estimated patience times.

Example 3 (*Partial information*) Let $(\pi, \hat{\pi})$ be a pair of random variables characterizing the true and estimated patience times for an infinitesimally small amount of fluid. Suppose that

$$(\pi, \hat{\pi}) \stackrel{d}{=} \left(\mathrm{e}^{Z}, \mathrm{e}^{\hat{Z}}\right),$$

where $(Z, \hat{Z})$ is bivariate normal with $\mathbb{E}Z = \mathbb{E}\hat{Z} = \theta$, $\mathrm{Var}(Z) = \mathrm{Var}(\hat{Z}) = \sigma^2$ and $\mathrm{Cov}(Z, \hat{Z}) = \rho\sigma^2$. That is, both patience and estimated patience are lognormally distributed with parameters θ and σ ($\pi, \hat{\pi} \sim \ln\mathcal{N}(\theta, \sigma^2)$), and the joint density function is given by

$$h(x, y) = \frac{1}{2\pi\sigma^2 xy\sqrt{1-\rho^2}}\mathrm{e}^{-\frac{(\log x-\theta)^2+(\log y-\theta)^2-2\rho(\log x-\theta)(\log y-\theta)}{2\sigma^2(1-\rho^2)}},$$

$x, y \geq 0$. Under this setup, it is convenient to model dependency between π and $\hat{\pi}$, since it is described by a single parameter (ρ)—the two are independent when $\rho = 0$, and the two are equal when $\rho = 1$. Otherwise, the conditional density of true patience at $x \geq 0$, given that the estimated patience is $y \geq 0$, is

$$h(x \mid y) = \frac{1}{\sqrt{2\pi}\sigma x\sqrt{1-\rho^2}}\mathrm{e}^{-\frac{(\log x-\rho\log y-(1-\rho)\theta)^2}{2\sigma^2(1-\rho^2)}},$$

or equivalently $\pi\,|\{\hat{\pi} = y\} \sim \ln\mathcal{N}(\rho\ln y + (1-\rho)\theta,\ \sigma^2(1-\rho^2))$. The coefficient of variation and mean of this conditional distribution are given by $\sqrt{\mathrm{e}^{\sigma^2(1-\rho^2)}-1}$ and

$y^{\rho} e^{(1-\rho)\theta+\sigma^2(1-\rho^2)/2}$, respectively; that is, the coefficient of variation does not vary with y. For two independent patience times, their order $(<, >)$ is the same as the order of the corresponding estimated patience times with probability

$$\frac{\sqrt{1-\rho^2}}{\pi} \int_0^{\pi/2} \frac{\mathrm{d}x}{1-\rho \sin x};$$

as expected, one obtains $1/2$ and 1, for $\rho = 0$ and $\rho = 1$, respectively.

Introduce a three-parameter function Q such that $Q(t, x, y)$ is the amount of fluid awaiting service at time $t \geq 0$, with patience at most $x \geq 0$ and estimated patience at most $y \in \mathbb{R}$:

$$Q(t, x, y) = \int_{-\infty}^{y} \int_0^x q(t, u, v)\, \mathrm{d}u\, \mathrm{d}v \quad \text{and} \quad Q(t, \infty, \infty) = Q(t).$$

It is appropriate to think of $q(t, x, y)$ as the density of fluid in the waiting room with true remaining patience x and estimated remaining patience y at time t (see Assumptions 9, 10). Note that, in the preceding equation, the integration covers also negative values of estimated remaining patience times. In fact, negative values of such times are feasible, since they decrease linearly over time. This corresponds to situations where a quantity of fluid was supposed to abandon based on an estimate on its arrival, but it remains in the waiting room due to a sufficiently large actual patience time (for example, if actual and estimated remaining patience times are 5 and 1 at $t = 0$, respectively, then those values are 3 and -1 at $t = 2$, respectively). We allow such negative values because they contain relative (ordering) information about estimated patience times of fluid in the waiting room. On the other hand, actual remaining patience times are always nonnegative—fluid abandons from the waiting room as soon as the actual remaining patience time decreases to 0 [see (40)].

Assumption 9 (*Initial fluid awaiting service*) In the case of partial information, the fluid waiting service at $t = 0$ satisfies

$$Q(0, x, y) = \int_{-\infty}^{y} \int_0^x q(0, u, v)\, \mathrm{d}u\, \mathrm{d}v \quad \text{and} \quad Q(0) < \infty,$$

where $q(0, \cdot, \cdot)$ is bounded, with $q(0, x, \cdot) \in \mathbb{C}_p$ for all $x \geq 0$.

In this section, we let $p_{\downarrow}(t)$ be the least *estimated* remaining patience of fluid present in the waiting room:

$$p_{\downarrow}(t) := \inf \left\{ y : \sup_{x \geq 0} q(t, x, y) > 0 \right\};$$

the definition of $p_{\downarrow}(t, u)$ is as in Sect. 4:

$$p_{\downarrow}(t, u) := \sup_{u \leq x \leq t} \{x - u + p_{\downarrow}(x)\}.$$

Note that, unlike in Sect. 4, $p_{\downarrow}(t)$ can take negative values, since it characterizes estimated patience rather than true patience. Based on the above definition, it follows that

$$Q(t) = \int_{p_{\downarrow}(t)}^{\infty} \int_{0}^{\infty} q(t, x, y)\, \mathrm{d}x\, \mathrm{d}y. \tag{37}$$

The fact that both remaining patience and estimated remaining patience decrease linearly at unit rate (until the corresponding fluid leaves the waiting room) motivates the following assumption on the density of fluid awaiting service:

Assumption 10 (*Fundamental LPF evolution equation*) For $t \geq 0$, $x \geq 0$ and $y \geq p_{\downarrow}(t)$:

$$\begin{aligned} q(t, x, y) = {} & q(0, x + t, y + t) 1_{\{y+t \geq p_{\downarrow}(t,0)\}} \\ & + \int_{0}^{t} 1_{\{p_{\downarrow}(t,u) \leq y+t-u\}} \lambda(u)\, h(t - u + x, t - u + y)\, \mathrm{d}u. \end{aligned} \tag{38}$$

An expression for $Q(t)$ can be derived by combining (37) and (38):

$$\begin{aligned} Q(t) = {} & \int_{p_{\downarrow}(t,0)}^{\infty} \int_{t}^{\infty} q(0, x, y)\, \mathrm{d}x\, \mathrm{d}y \\ & + \int_{p_{\downarrow}(t)}^{\infty} \int_{0}^{\infty} \int_{0}^{t} 1_{\{p_{\downarrow}(t,u)+u-t \leq y\}} \lambda(u)\, h(t - u + x, t - u + y)\, \mathrm{d}u\, \mathrm{d}x\, \mathrm{d}y \\ = {} & \int_{p_{\downarrow}(t,0)}^{\infty} \int_{t}^{\infty} q(0, x, y)\, \mathrm{d}x\, \mathrm{d}y \\ & + \int_{0}^{t} \int_{p_{\downarrow}(t)}^{\infty} \int_{t-u}^{\infty} \lambda(u)\, h(x, y + t - u)\, 1_{\{y \geq p_{\downarrow}(t,u)+u-t\}}\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}u. \end{aligned}$$

Thus, one has

$$Q(t) = \int_{p_{\downarrow}(t,0)}^{\infty} \int_{t}^{\infty} q(0, x, y)\, \mathrm{d}x\, \mathrm{d}y + \int_{0}^{t} \lambda(u)\, \bar{H}(t - u, p_{\downarrow}(t, u))\, \mathrm{d}u, \tag{39}$$

where $\bar{H}(x, y) = \int_{x}^{\infty} \int_{y}^{\infty} h(u, v)\, \mathrm{d}v\, \mathrm{d}u$. The first term corresponds to fluid in the waiting room at time $t = 0$, while the second term represents fluid that arrived in $[0, t]$.

In order to define the abandonment rate $\alpha(t)$, we consider the actual remaining patience rather than the estimated counterpart. In particular, fluid with zero (true) remaining patience abandons the waiting room:

$$\alpha(t) := \int_{p_{\downarrow}(t)}^{\infty} q(t, 0, y)\, \mathrm{d}y, \quad t \geq 0, \tag{40}$$

where (38) specifies $q(t, 0, y)$ in terms of $p_\downarrow$. By substituting the expression for $q(t, 0, y)$ in (40), one concludes that the abandonment rate obeys, for $t \geq 0$,

$$\alpha(t) = \int_{p_\downarrow(t,0)}^{\infty} q(0, t, y)\, \mathrm{d}y + \int_0^t \lambda(u) \int_{p_\downarrow(t,u)}^{\infty} h(t-u, y)\, \mathrm{d}y\, \mathrm{d}u. \tag{41}$$

Consequently, the total amount of abandonment by time t, $A(t)$, can be expressed in a couple of ways: first, in view of (40), it satisfies

$$A(t) = \int_0^t \int_{p_\downarrow(u)}^{\infty} q(u, 0, y)\, \mathrm{d}y\, \mathrm{d}u; \tag{42}$$

second, in view of (41), $A(t)$ can be written as a sum of two terms that correspond to fluid initially in the system and fluid arriving after time $t = 0$:

$$A(t) = \int_0^t \int_{p_\downarrow(x,0)}^{\infty} q(0, x, y)\, \mathrm{d}y\, \mathrm{d}x + \int_0^t \lambda(u) \int_0^{t-u} \int_{p_\downarrow(u+x,u)}^{\infty} h(x, y)\, \mathrm{d}y\, \mathrm{d}x\, \mathrm{d}u. \tag{43}$$

Combining (1), (39) and the preceding equality yields an equation for $p_\downarrow$. In particular, $p_\downarrow$ is the maximal solution of

$$Q(t) + A(t) = Q(0) + \Lambda(t) - E(t), \tag{44}$$

where $Q(t)$ and $A(t)$ are given by (39) and (43). The left-hand side is nonnegative and monotonic in $p_\downarrow$, which motivates the following assumption:

Assumption 11 There exists a unique solution of (44), where (39) and (43) determine $Q(t)$ and $A(t)$, respectively.

The rate $\gamma(t)$ can be characterized by examining the two ways fluid can enter the service facility. At time t, fluid enters service directly (without entering the waiting room) at rate $\lambda(t)\, H(\infty, p_\downarrow(t))$, since no fluid with estimated patience below $p_\downarrow(t)$ is present in the waiting room. On the other hand, the transfer rate between the waiting room and the service facility is proportional to $q(t, x, p_\downarrow(t))$. Formally, taking derivatives in (1), and utilizing (38), (39) and (41) yields

$$\begin{aligned}
\gamma(t) &= \lambda(t) - \alpha(t) - \dot{Q}(t) \\
&= \lambda(t)\, H(\infty, p_\downarrow(t)) + (1 + \dot{p}_\downarrow(t))^+ \int_0^{\infty} q(t, x, p_\downarrow(t))\, \mathrm{d}x \\
&= \lambda(t)\, H(\infty, p_\downarrow(t)) + (1 + \dot{p}_\downarrow(t)) \int_0^{\infty} q(t, x, p_\downarrow(t))\, \mathrm{d}x,
\end{aligned}$$

for $t \geq 0$ such that $\dot{p}_\downarrow(t)$ is well-defined; the last equality follows from the fact that the last integral is equal to 0 when $\dot{p}_\downarrow(t) < -1$ (due to Assumption 10). The above differential equation is an analogue of (22), which holds for the case of full information.

Finally, we note that the numerical algorithm outlined in Sect. 4.4 is applicable to evaluate fluid models under partial information with one modification: $p_{\downarrow}(t_i)$ solves (33), with

$$Q(t_i) = \int_0^{t_i} \lambda(u)\, \bar{H}(t_i - u,\, p_{\downarrow}(t_i, u))\, \mathrm{d}u,$$
$$A(t_i) = A(t_{i-1}) + \int_{t_{i-1}}^{t_i} \int_0^t \lambda(u) \int_{p_{\downarrow}(t,u)}^{\infty} h(t-u, y)\, \mathrm{d}y\, \mathrm{d}u\, \mathrm{d}t.$$

We conclude this section with an example.

Example 4 ($G_t/M/s + M$ *LPF fluid model with no information*) Consider the setup described in Example 1 with $\bar{H}(x, y) = \mathrm{e}^{-\theta(x+y)}$, $x, y \geq 0$. In that case, even though the distribution of the estimated patience times is the same as the distribution of actual patience times, the two random variables are independent. As in Example 1, the queue is empty during the time interval $[0, t_1]$, where t_1 is given by (23). However, unlike in Example 1, the abandonment rate is positive for $t > t_1$. The form of H, (39) and (43) yield, for $t \geq t_1$,

$$A(t) = \theta \int_{t_1}^{t} Q(u)\, \mathrm{d}u.$$

Combining this equality with (1) and (24) results in

$$Q(t) + \theta \int_{t_1}^{t} Q(u)\, \mathrm{d}u = A(t) - E(t) = \delta\mu s(t - t_1),$$

for $t \geq t_1$. The solution of the preceding integral equation is

$$Q(t) = \frac{\delta\mu s}{\theta} \left(1 - \mathrm{e}^{-\theta(t-t_1)}\right) 1_{\{t \geq t_1\}}, \tag{45}$$

and consequently

$$\alpha(t) = \delta\mu s \left(1 - \mathrm{e}^{-\theta(t-t_1)}\right) 1_{\{t \geq t_1\}}.$$

Equating (39) and (45) yields, for $t > t_1$,

$$p_{\downarrow}(t) = \frac{1}{\theta} \log \frac{(1+\delta)\left(1 - \mathrm{e}^{-2\theta(t-t_1)}\right)}{2\delta\left(1 - \mathrm{e}^{-\theta(t-t_1)}\right)}.$$

Note that $\dot{p}_{\downarrow}(t) > -1$, for $t > t_1$, and $p_{\downarrow}(t, u) = t - u + p_{\downarrow}(t)$.

In steady state (as $t \to \infty$), $\lambda(\infty)(1 - \mathrm{e}^{-\theta p_{\downarrow}(\infty)}) = (1-\delta)\mu s$ is the rate at which fluid enters service directly; on the other hand, $\lambda(\infty)\, \mathrm{e}^{-\theta p_{\downarrow}(\infty)}/2 = \delta\mu s$ is the transfer rate of fluid from the waiting room to service—fluid with true/estimated patience times

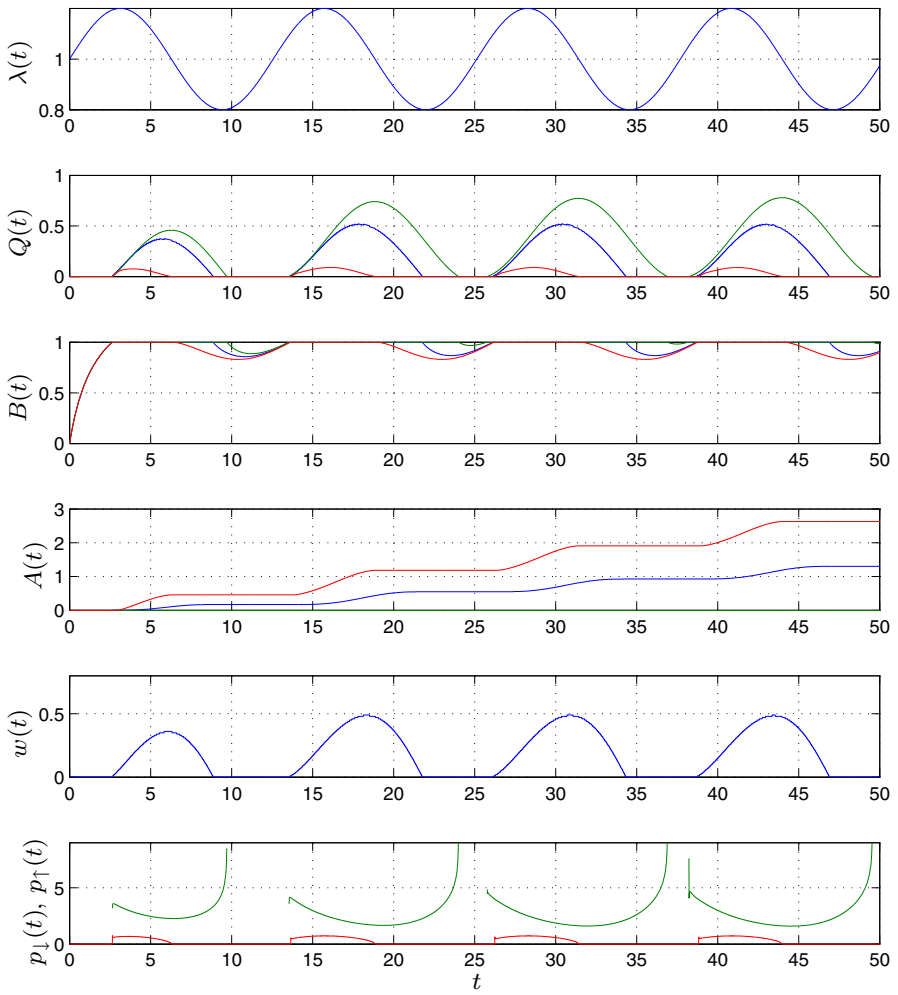

Fig. 3 Performance functions for the system described in Example 5, operating under LPF (*green*), MPF (*red*) and FCFS (*blue*). No abandonment occurs under LPF. The function w represents the waiting time of head-of-line fluid in the waiting room when FCFS is used. The function $p_\uparrow$ represents the remaining patience of the most patient fluid under the MFP policy (Color figure online)

(on arrival) in the set $\{(x, y) : y \geq p_\downarrow(\infty),\ x \geq y - p_\downarrow(\infty)\}$ enters the waiting room and is eventually served.

6 Numerical examples

The proposed numerical algorithm was used for LPF; for FCFS, the algorithm in [27] was used. The time step was set to 0.01, and the trapezoidal rule was used to evaluate integrals. A modification of the proposed algorithms is used to evaluate performance under MPF routing.

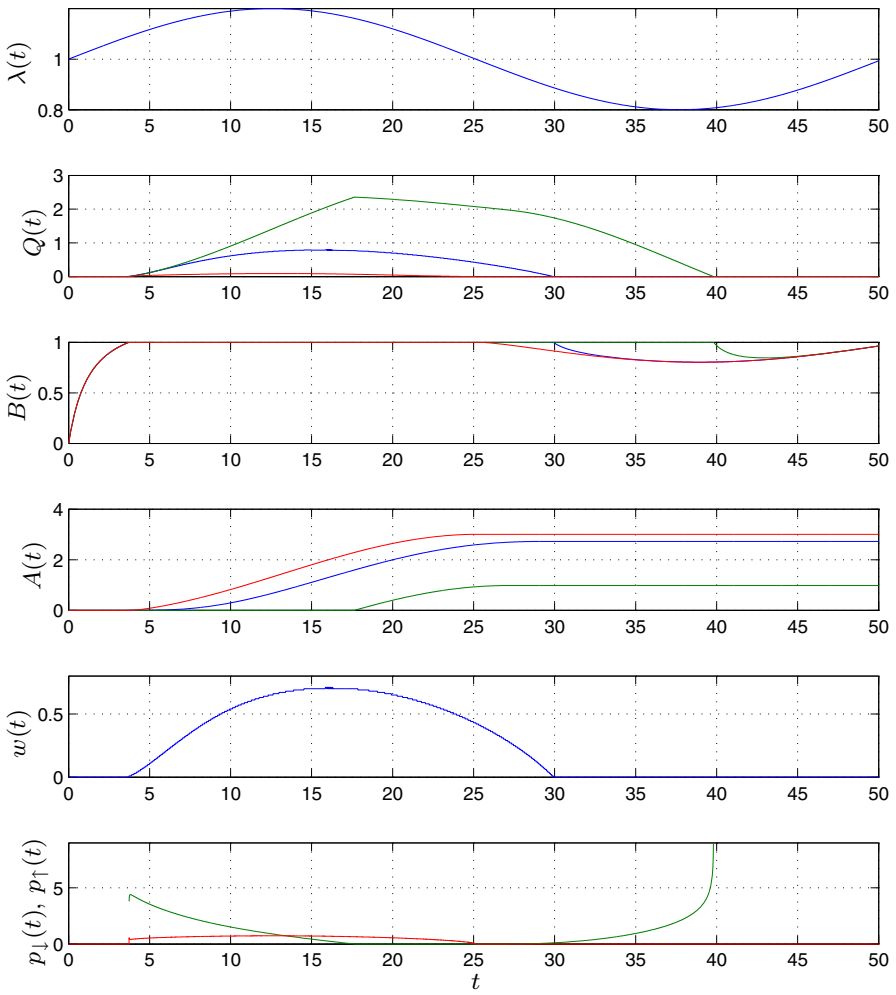

Fig. 4 Performance functions for the system described in Example 5, operating under LPF (*green*), MPF (*red*) and FCFS (*blue*). The function w represents the waiting time of head-of-line fluid in the waiting room when FCFS is used. The function $p_{\uparrow}$ represents the remaining patience of the most patient fluid under the MFP policy (Color figure online)

Example 5 (*LPF and MPF vs. FCFS*) In this example, we compare LPF and MPF to the FCFS policy. Consider an initially empty system with $s = 1$. The service and patience distributions are as follows [27]: $G(t) = (1 - \mathrm{e}^{-2t})/2 + (1 - \mathrm{e}^{-2t/3})/2$ and $F(t) = 1 - \mathrm{e}^{-t} - t\mathrm{e}^{-t}$; the mean service time is 1, while mean patience time is 2. The arrival rate is given by $\lambda(t) = 1 + 0.2\ \sin(t/2)$; hence, the system is critically loaded. In Fig. 3, we show key performance measures of the system.

Observe that no fluid is lost in the system operating under the LPF policy—indeed, the minimum remaining patience of fluid in the waiting room stays strictly positive throughout the considered time interval. Informally, the overloaded period is short

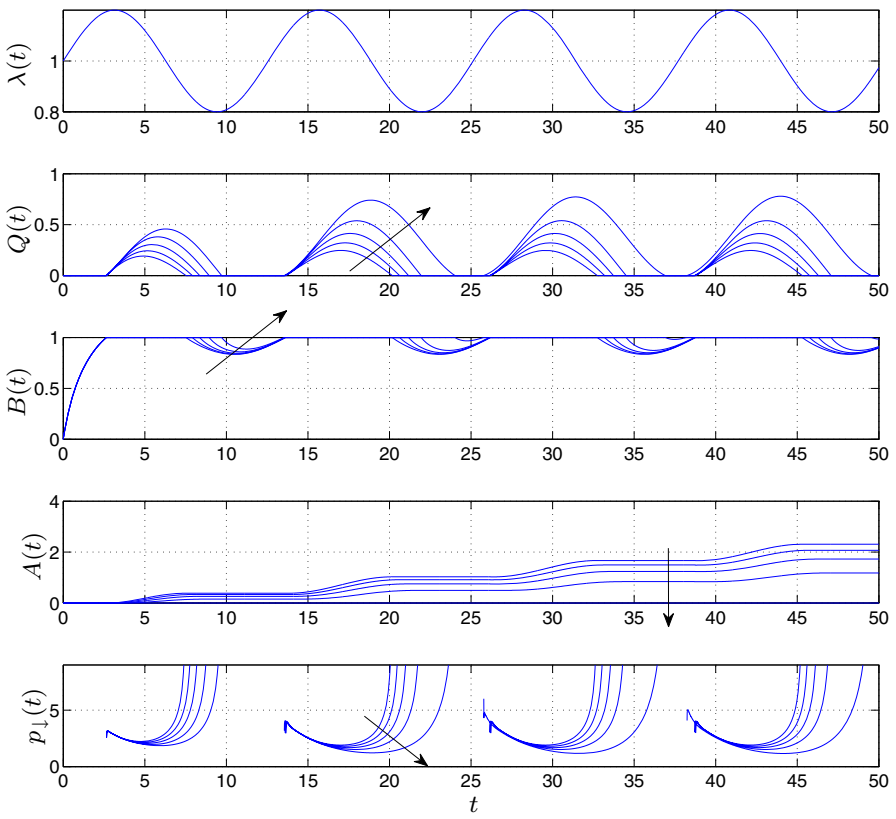

Fig. 5 Performance functions for the system described in Example 6, for $\rho = 0, 0.25, 0.5, 0.75$ and 1. The arrows indicate the direction of increase for ρ

enough and there exists a sufficient amount of fluid with large enough remaining patience times that can be kept in the waiting room in order for fluid with short remaining patience times to be sent to the service facility. As durations of overloaded intervals increase, the amount of fluid with long patience times is not sufficient to avoid abandonments—there does not exist enough fluid in the waiting room that can be delayed without causing abandonments. In Fig. 4, we show the behavior of the system with a modified arrival rate: $\lambda(t) = 1 + 0.2\,\sin(t/8)$, i.e., the frequency of the arrival rate function is decreased by a factor of 4.

Example 6 (*Performance vs. amount of partial information*) As in the previous example, consider an initially empty system with $s = 1$. The service distribution is exponential, $\bar{G}(x) = \mathrm{e}^{-x}$, $x \geq 0$, while the joint distribution of patience and estimated patience is as follows (see Example 3):

$$H(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\log x} \int_{-\infty}^{\log y} \mathrm{e}^{-\frac{u^2+v^2-2\rho uv}{2(1-\rho^2)}}\, \mathrm{d}v\, \mathrm{d}u,$$

$x, y \geq 0$ and $\rho \in (0, 1)$. Then both marginal distributions are lognormal, with parameters (0, 1) (mean $= \sqrt{e} \approx 1.65$ and variance $= (e-1)e \approx 4.47$). Informally, the "level" of information contained in the estimated patience increases with the value of the parameter ρ. In Fig. 5, we plot performance functions for the systems with $\rho = 0, 0.25, 0.5, 0.75$ and 1, where $\rho = 1$ corresponds to the case of full information (for these ρ's, the coefficients of variation of the conditional distributions of patience times are given by ≈ 1.31, ≈ 1.25, ≈ 1.06, ≈ 0.74 and 0, respectively—see Example 3. Given two independent pairs of actual and estimated patience times, the actual patience times are ordered in the same way the corresponding estimated times are ordered with the respective probabilities 0.5, ≈ 0.58, ≈ 0.67, ≈ 0.77 and 1). As evident from the figure, more information leads to less abandonment. Recall that $p_{\downarrow}(t)$ is the least estimated remaining patience of the fluid present in the waiting room at time t; in this example, $p_{\downarrow}(t) > 0$ for all t. Nevertheless, abandonment does occur, since fluid leaves the system based on actual rather than estimated patience times.

7 Concluding remarks

In this final section, we discuss the relationship between our fluid models and a stochastic queueing system with a finite number of servers (G_t/GI/n+GI). As argued in [37] for the FCFS case, deterministic fluid processes can provide approximations for mean values of stochastic queueing processes. In the system with a finite number of servers, arrivals are time-varying (the G_t), service times are i.i.d. (the GI) and customers, equipped with i.i.d. durations of patience, possibly abandon (the +GI); the arrival sequence, service and patience times are mutually independent. The number of servers corresponds to the processing capacity s of the fluid model.

When comparing fluid models with stochastic queueing systems, at least two sources of errors can be identified. First, there exists an error stemming from the deterministic nature of a fluid model, which does not take stochastic fluctuations into account. For example, in the fluid model, it is assumed that customers with patience times below $p_{\downarrow}(t)$ enter the service facility immediately upon arrival. However, such customers enter service only once a customer completes service. It is possible to develop correction terms for fluid functions that take this effect into account. To this end, let $\alpha_n(t)$ be the abandonment rate in an n-server system; for very large values of n, one expects $\alpha_n(t)/n \approx \alpha(t)$, where $\alpha(t)$ is the abandonment rate in the fluid model at time t. An extra term can hence be added to capture some of the present stochastic variability:

$$\frac{\alpha_n(t)}{n} \approx \alpha(t) + \lambda(t)\, F\left(\frac{1}{n\sigma(t)}\right) 1_{\{p_{\downarrow}(t) \in (0,\infty)\}};$$

here, the second term approximates the rate of customers abandoning before a single departure occurs after their arrival. When $p_{\downarrow}(t) \in (0, \infty)$ (non-empty waiting room), the service completion rate is equal to the rate at which customers enter service (due to work conservation). Therefore, $1/(n\sigma(t))$ is an approximation of the mean time between a customer arrival and the next service completion, and $F(1/(n\sigma(t)))$ is an

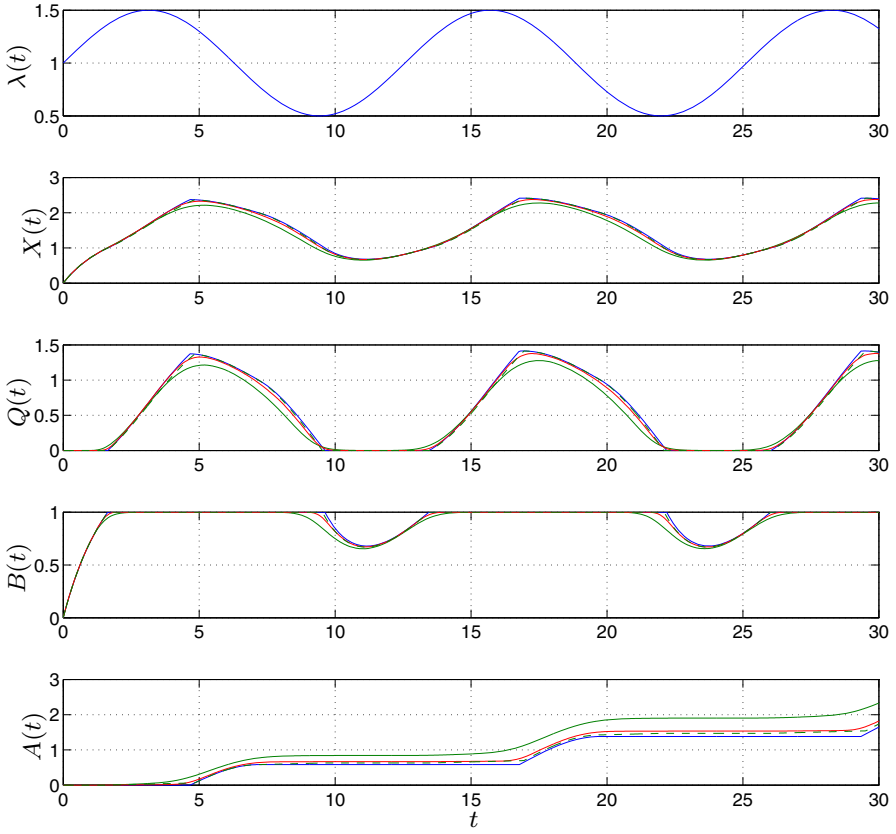

Fig. 6 Performance functions for the systems described in Example 7. For the two finite stochastic systems (50 and 250 servers), empirical averages are plotted (in *green* and *red*, respectively). Performance functions for the fluid model are shown in *blue*. A corrected fluid approximation for the system with 50 servers is shown with *dashed lines* (Color figure online)

approximation of the probability that a customer abandons before a single departure occurs. Such a probability is proportional to $1/n$ and is negligible compared to the leading $\alpha(t)$ term, when n is large. Incorporating this extra term into the fluid model equations leads to a corrected version (that depends on the number of servers) of our numerical algorithm. In particular, we have

$$\tilde{A}(t_i) = \tilde{A}(t_{i-1}) + (\Lambda(t_i) - E(t_i) - \tilde{A}(t_{i-1}) - Q(t_i))^+ 1_{\{p_\downarrow(t_i)=0\}} + \int_{t_{i-1}}^{t_i} \lambda(u)\, F\left(\frac{1}{n\sigma(u)}\right) 1_{\{p_\downarrow(u)\in(0,\infty)\}}\, \mathrm{d}u, \tag{46}$$

where tildes distinguish the present refined abandonment process from its previous version (35). Note that the last term in (46) is proportional to $1/n$ for large n. Hence, as also seen from the example at the end of the section, such a correction term produces only marginal improvements in the quality of the approximation.

Second, discrepancy between the fluid functions and sample means of stochastic processes also arises because of the nonlinear nature of the system. Indeed, averaging sample paths of stochastic processes could produce a bias relative to their fluid functions due to Jensen's inequality. The reader is referred to [2,25] for examples of studies of such biases in queueing contexts. Somewhat different results are obtained depending on whether one considers the total number in the system or, separately, the number awaiting service and being served.

Our numerical experiments indicate that errors due to Jensen's inequality play a more prominent role than the errors discussed earlier. In fact, one expects (based on the CLT) that errors due to nonlinearities are of the order $1/\sqrt{n}$. In general, better results are obtained when a system does not operate in the critical regime (QED), during which the queue size is close to 0 (Quality) and the service facility is close to full (Efficiency).

Example 7 (*Fluid approximation*) Consider initially empty LPF systems with $\lambda(t) = 1 + 0.5\sin(t/2)$, and unit-rate exponential service and patience times. In addition to the fluid model, we consider two corresponding queues with 50 and 250 servers (arrival processes are Poisson with rates $50\lambda(t)$ and $250\lambda(t)$, respectively). In Fig. 6, we plot relevant functions for the three systems. In particular, functions (X, Q, B, A) corresponding to the fluid model are shown in blue. For the two finite systems with $n = 50$ (green) and $n = 250$ (red), we plot empirical averages (scaled by the number of servers) of queueing stochastic processes, based on 10,000 and 1000 replications, respectively. For the system with 50 servers, we also plot the corrected fluid approximation [based on (46)] as well. As seen in the figure, this approximation provides only a minor improvement, which is consistent with our discussion that led to the example.

Acknowledgements The work of A. M. has been partially supported by BSF Grants 2008480 and 2014180, ISF Grants 1357/08 and 1955/15 and by the Technion funds for promotion of research and sponsored research. Some of the research was funded by and carried out while A. M. was visiting the Statistics and Applied Mathematical Sciences Institute (SAMSI) of the NSF; the Department of Statistics and Operations Research (STOR), the University of North Carolina at Chapel Hill; the Department of Information, Operations and Management Sciences (IOMS), Leonard N. Stern School of Business, New York University; and the Department of Statistics, The Wharton School, University of Pennsylvania—the wonderful hospitality of all four institutions is gratefully acknowledged and truly appreciated. The work of P. M. has been partially supported by the NSF Grant CMMI-1362630 and the BSF Grant 2014180. Finally, the authors thank the 2012 SAMSI Working Group on Data-Based Patient Flow in Hospitals, which provided an encouraging forum for our research as it evolved. In particular, Jamol Pender suggested, during a SAMSI meeting, the MPF policy as a benchmark.

References

1. Aksin, Z., Ata, B., Emadi, S., Su, C.-L.: Structural estimation of callers' delay sensitivity in call centers. Manag. Sci. **59**(12), 2727–2746 (2013)
2. Altman, E., Jiménez, T., Koole, G.: On the comparison of queueing systems with their fluid limits. Probab. Eng. Inf. Sci. **15**(2), 165–178 (2001)
3. Argon, N., Ziya, S.: Priority assignment under imperfect information on customer type identities. Manuf. Serv. Oper. Manag. **11**(4), 674–693 (2009)
4. Argon, N., Ziya, S., Righter, R.: Scheduling impatient jobs in a clearing system with insights on patient triage in mass casualty incidents. Probab. Eng. Inf. Sci. **22**(3), 301–332 (2008)

5. Asmussen, S.: Applied Probability and Queues, 2nd edn. Springer, New York (2003)
6. Atar, R., Biswas, A., Kaspi, H.: Fluid limits of $G/G/1+G$ queues under the non-preemptive earliest-deadline-first discipline. Math. Oper. Res. **40**(3), 683–702 (2015)
7. Bassamboo, A., Randhawa, R.: Using estimated patience levels to optimally schedule customers. Preprint (2013)
8. Batt, R., Terwiesch, C.: Waiting patiently: an empirical study of queue abandonment in an emergency department. Manag. Sci. **61**(1), 39–59 (2015)
9. Billingsley, P.: Convergence of Probability Measures, 2nd edn. Wiley, New York (1999)
10. Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., Zhao, L.: Statistical analysis of a telephone call center: a queueing-science perspective. J. Am. Stat. Assoc. **100**, 36–50 (2005)
11. Cohen, I., Mandelbaum, A., Zychlinski, N.: Minimizing mortality in a mass casualty event: fluid networks in support of modeling and management. IIE Trans. **46**(7), 728–741 (2014)
12. Decreusefond, L., Moyal, P.: Fluid limit of a heavily loaded EDF queue with impatient customers. Markov Process. Relat. Fields **14**(1), 131–158 (2008)
13. Dertouzos, M.: Control robotics: the procedural control physical processes. In: Proc. IFIP Congress, Stockholm (1974)
14. Doytchinov, B., Lehoczky, J., Shreve, S.: Real-time queues in heavy traffic with earliest-deadline-first queue discipline. Ann. Appl. Probab. **11**(2), 332–378 (2001)
15. Feigin, P.: Analysis of customer patience in a bank call center. Working Paper (2006)
16. Gans, N., Liu, N., Mandelbaum, A., Shen, H., Ye, H.: Service times in call centers: agent heterogeneity and learning with some operational consequences. In: Berger, J., Cai, T., Johnstone, I. (eds.) Borrowing Strength: Theory Powering Applications—A Festschrift for Lawrence D. Brown, Collections, vol. 6, pp. 99–123. Institute of Mathematical Statistics, Beachwood (2010)
17. Garnett, O., Mandelbaum, A., Reiman, M.: Designing a call center with impatient customers. Manuf. Serv. Oper. Manag. **4**(3), 208–227 (2002)
18. Ghebali, R.: Real-time prediction of the probability of abandonment in call centers. Master's Thesis, Technion—Israel Institute of Technology, Haifa (2012)
19. Green, L., Soares, J., Giglio, J., Green, R.: Using queueing theory to increase the effectiveness of emergency department provider staffing. Acad. Emerg. Med. **13**(1), 61–68 (2006)
20. Halfin, S., Whitt, W.: Heavy-traffic limits for queues with many exponential servers. Oper. Res. **29**(3), 567–588 (1981)
21. Hong, J., Tan, X., Towsley, D.: A performance analysis of minimum laxity and earliest deadline scheduling in a real-time system. IEEE Trans. Comput. **38**(12), 1736–1744 (1989)
22. Jackson, J.: Some problems in queueing with dynamic priorities. Nav. Res. Log. Q. **7**(3), 235–249 (1960)
23. Jackson, J.: Waiting time distribution for queues with dynamic priorities. Nav. Res. Log. Q. **9**(1), 31–36 (1960)
24. Jackson, J.: Queues with dynamic priority discipline. Manag. Sci. **8**(1), 18–34 (1961)
25. Jiménez, T., Koole, G.: Scaling and comparison of fluid limits of queues applied to call centers with time-varying parameters. OR Spect. **26**(3), 413–422 (2004)
26. Kang, W.: Existence and uniqueness of a fluid model for many-server queues with abandonment. Oper. Res. Lett. **42**(6–7), 478–483 (2014)
27. Kang, W., Pang, G.: Computation and properties of fluid models of time-varying many-server queues with abandonment. Preprint (2013)
28. Kang, W., Pang, G.: Fluid limit of a multiclass many-server queueing network with abandonment and feedback. Preprint (2013)
29. Kang, W., Pang, G.: Equivalence of fluid models for $G_t/GI/N+GI$ queues. Preprint (2014)
30. Kang, W., Ramanan, K.: Fluid limits of many-server queues with reneging. Ann. Appl. Probab. **20**(6), 2204–2260 (2010)
31. Kaplan, E.: Terror queues. Oper. Res. **58**(4), 773–784 (2010)
32. Kaspi, H., Ramanan, K.: Law of large numbers limits for many-server queues. Ann. Appl. Probab. **21**(1), 33–114 (2011)
33. Kruk, L., Lehoczky, J., Shreve, S.: Accuracy of state space collapse for earliest-deadline-first queues. Ann. Appl. Probab. **16**(2), 516–561 (2006)
34. Kruk, L., Lehoczky, J., Ramanan, K., Shreve, S.: Heavy traffic analysis for EDF queues with reneging. Ann. Appl. Probab. **21**(2), 484–545 (2011)

35. Li, D., Glazebrook, K.: A Bayesian approach to the triage problem with imperfect information. Eur. J. Oper. Res. **215**(1), 169–180 (2011)
36. Liu, Y., Whitt, W.: A network of time-varying many-server fluid queues with customer abandonment. Oper. Res. **59**(4), 835–846 (2011)
37. Liu, Y., Whitt, W.: The $G_t/GI/s_t + GI$ many-server fluid queue. Queueing Syst. Theory Appl. **71**(4), 405–444 (2012)
38. Liu, Y., Whitt, W.: A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading. Oper. Res. Lett. **40**(5), 307–312 (2012)
39. Mandelbaum, A., Momčilović, P.: Queues with many servers and impatient customers. Math. Oper. Res. **37**(1), 41–64 (2012)
40. Mandelbaum, A., Zeltyn, S.: Data-stories about (im)patient customers in tele-queues. Queueing Syst. Theory Appl. **75**(2–4), 115–146 (2013)
41. Mandelbaum, A., Massey, W., Reiman, M.: Strong approximations for Markovian service networks. Queueing Syst. Theory Appl. **30**(1–2), 149–201 (1998)
42. Merin, O., Ash, N., Levy, G., Schwaber, M., Kreiss, Y.: The Israeli field hospital in Haiti—ethical dilemmas in early disaster response. N. Engl. J. Med. **362**(11), e38 (2010)
43. Mieghem, J.V.: Due date scheduling: asymptotic optimality of generalized longest queue and generalized largest delay rules. Oper. Res. **51**(1), 113–122 (2003)
44. Mistovich, J., Hafen, B., Karren, K.: Prehospital Emergency Care. Prentice Hall Health, Englewood Cliffs (2000)
45. Moyal, P.: Convex comparison of service disciplines in real-time queues. Oper. Res. Lett. **36**(4), 496–499 (2008)
46. Moyal, P.: On queues with impatience: stability, and the optimality of earliest deadline first. Queueing Syst. Theory Appl. **75**(2–4), 211–242 (2013)
47. Palm, C.: Methods of judging the annoyance caused by congestion. Tele **2**, 1–20 (1953)
48. Pang, G., Whitt, W.: Two-parameter heavy-traffic limits for infinite-server queues. Queueing Syst. Theory Appl. **65**(4), 325–364 (2010)
49. Panwar, S., Towsley, D., Wolf, J.: Optimal scheduling policies for a class of queues with customer deadlines to the beginning of service. J. ACM **35**(4), 832–844 (1988)
50. Reed, J.: The $G/GI/N$ queue in the Halfin–Whitt regime. Ann. Appl. Probab. **19**(6), 2211–2269 (2009)
51. Saghafian, S., Hopp, W., Oyen, M.V., Desmond, J., Kronick, S.: Patient streaming as a mechanism for improving responsiveness in emergency departments. Oper. Res. **60**(5), 1080–1097 (2012)
52. Saghafian, S., Hopp, W., Oyen, M.V., Desmond, J., Kronick, S.: Complexity-augmented triage: a tool for improving patient safety and operational efficiency. Manuf. Serv. Oper. Manag. **16**(3), 329–345 (2014)
53. Spencer, J., Sudan, M., Xu, K.: Queueing with future information. Ann. Appl. Probab. **24**(5), 2091–2142 (2014)
54. Stoyenko, A., Georgiadis, L.: On optimal lateness and tardiness scheduling in real-time systems. Computing **47**(3–4), 215–234 (1992)
55. Wein, L.: Due-date setting and priority sequencing in a multiclass $M/G/1$ queue. Manag. Sci. **37**(7), 834–850 (1991)
56. Whitt, W.: Fluid models for multiserver queues with abandonments. Oper. Res. **54**(1), 37–54 (2006)