PERSONALIZED QUEUES: THE CUSTOMER VIEW, VIA LEAST-PATIENT FIRST ROUTING

AVISHAI MANDELBAUM AND PETAR MOMČILOVIĆ

Abstract. In personalized queues, information at the level of individuals – customers or servers – is affecting system dynamics. Such information is becoming increasingly accessible, directly or statistically, as exemplified by personalized 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 under-loaded periods, hence a fluid view provides a natural modeling framework. This parsimonious 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). One resulting insight is that LPF can provide significant advantages over FCFS when the durations of overloaded periods are comparable to (im)patience times.

1. Introduction

A modeling paradigm for personalized queues. In a *personalized* queueing system, say M/M/n+M [40, 12] for concreteness, interarrival times and 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 will the next arrival 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.

And why "Paradigm"? Because essentially every queueing model can be "personalized", by availing individual realizations of its primitives 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 disperse research is of value – hence worthy of the term "Paradigm". Furthermore, in existing schemes full information is available, e.g., individual service or patience times are known exactly. Yet of more interest is the practically relevant case of partial information, where knowledge about individual realizations is noisy (cf. triage process in emergency departments). A central challenge in our paradigm is thus the tradeoff 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 [11], which provides empirical support for a personalized server view – individual service durations. A second example is [13], that in

Date: June 26, 2014.

2000 *Mathematics Subject Classification.* Primary: 60K25; Secondary: 90B22.

Key words and phrases. Multi-server queue, time-varying queue, fluid approximation/model, earliest deadline first.

fact motivated the present paper. It develops inference tools that enable the personalization of customer impatience in telephone queues. Availing this personalized customer information to discretionary control should yield a reduction of 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 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 [34] or emergency departments [14], customer patience is the amount of time that a customer is *willing* or *able* to wait for service; in terror queues [24], 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 [7]. Furthermore, MCEs are incidents where medical resources (personnel, equipment) are overwhelmed by the number and severity of casualties [37]; e.g. 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. (Yet despite such plans, medical personnel can experience ethical dilemmas [35] 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 service at rates that are time-varying over a finite time-horizon – service that is to be catered 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 [42, 39]) 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 $\mathrm{G}_t/\mathrm{GI}/n+\mathrm{GI}$ queue, but it 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 $\mathrm{G}_t/\mathrm{GI}/n+\mathrm{GI}$ 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 – it 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.

In addition to the partial information model, we examine a model with full information. Although in some applications this assumption is not realistic, 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). For comparison sake only, we consider the same model under Most-Patient First (MPF) routing as well. In both 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.

To be more specific, we focus on time-varying fluid models that alternate between over- and under-loaded periods, as these are circumstances when the advantages (less customers abandon) of LPF over FCFS can become significant. In comparison, employing LPF instead of FCFS to a many-server queue, in the Quality-and-Efficiency Driven (QED) regime [15, 43, 33], 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 over-loaded 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 under-loaded time intervals are present, a personalized policy can harmlessly shift the load in time (by delaying customers with long patience times), which effectively reduces over-loaded 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 can not 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 [16, 48, 36]. In contrast, our LPF terminology fits patience that 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 Section 3. LPF policy under full information are considered in Section 4, accompanied with a corresponding numerical algorithm. Our model and algorithm for the case of partial information appear in Section 5. Based on numerical examples, we discuss various insights in Section 6. The paper concludes in Section 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 [32, 41, 30, 31]. Analyses of many-server fluid models with abandonments has mostly focused on systems operating under the FCFS policy. In particular, a stationary model was studied in [49]. Formal fluid limits for this model were established in [23] by extending results for the model without abandonment from [25]. In [29], a network of fluid models was considered, while a system with time-varying capacity was investigated in [30]. A numerical algorithm for evaluating sample paths of FCFS many-server fluid models with constant capacities was proposed in [21]. A fluid limit of a multi-class many-server queueing network with abandonment and feedback is studied in [22].

Early analyses of EDF can be found in [17, 19, 18]. 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 [9] (that is, if a collection of jobs can be scheduled in a way that ensures all the jobs complete by their deadline, the EDF will schedule this collection of jobs so they all complete by their deadline). In [42] 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 [39]. 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 [47], as well as any convex function of the average tardiness [38]. EDF scheduling was studied in the context of conventional heavy traffic, both without [10, 27] and with abandonments [26]. A fluid limit of a heavily loaded EDF M/M/1 queue was considered in [8]. Fluid limits of G/G/1+G queues under EDF were investigated in [5].

Our partial information framework relates to studies of multi-class systems where customer classes can be estimated/predicted [2, 3]. 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 accuracies. Typically, a Bayesian view is adopted where classes are characterized by probability distributions of service/patience times rather than realizations associated with individual patients, e.g. [28]. Such queueing models have been used to capture the triage process in emergency departments [44, 45]. 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 [6], 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 tradeoff between information availability and queueing performance has been examined in [46], 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 to 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 notations and conventions of [49].) Let $Q(t)$ and $B(t)$ be the amount of fluid awaiting service and obtaining service at time t, 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: $(s - B(t))\,Q(t) = 0$, for all $t \geq 0$; in words, the queue is non-empty if and only if there exists no spare capacity. 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 $F(x) = \int_0^x f(u)\,\mathrm{d}u$ 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 $G(x) = \int_0^x g(u)\,\mathrm{d}u$ 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 absolutely-continuous distribution functions, which are referred to as the abandonment and service distribution, respectively; denote $\bar{G} = 1 - G$ and $\bar{F} = 1 - F$.

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 $t \geq 0$. Similarly, let $E(t) = \int_0^t \gamma(u)\,\mathrm{d}u$ be the amount of fluid to enter service in $[0,t]$; here, $\gamma(t)$ is the rate at which fluid enters service at time t. The service completion rate at time t is denoted by $\sigma(t)$; then the total amount of fluid to complete service during the interval $[0,t]$ is $S(t) = \int_0^t \sigma(u)\,\mathrm{d}u$. The existence of rates α, γ and σ is due to the absolute continuity of F, G and Λ. We now deduce the following basic flow-conservation equations, which hold for all $t \geq 0$:

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

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 at 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 section. 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 $t \geq 0$, which will remain so for at most the next $x \geq 0$ time units:

$$B(t,x) = \int_0^x b(t,u)\,\mathrm{d}u \quad \text{and} \quad B(t,\infty) = B(t),$$

for $t \geq 0$, $x \geq 0$; here $b(t,x)$ is the density of fluid in service with the *remaining* service time x at time t – recall that the service distribution is absolutely continuous. (In general, one can consider time-in-service variables instead of remaining service time variables.) We note that $\sigma(t) = b(t,0)$, $t \geq 0$. Initial conditions for the service facility are specified by $b(0,\cdot)$ that satisfies

$$B(0) = \int_0^\infty b(0,u)\,\mathrm{d}u < \infty.$$

Given that the service requirement of fluid entering service is distributed according to G, $\bar{G}(t-u)$ fraction of fluid entering service at time $u \geq 0$ remains in the system by time $t \geq u$; this leads to

$$B(t) = B(0) - B(0,t) + \int_0^t \bar{G}(t-u)\,\mathrm{d}E(u), \tag{2}$$

where the first term accounts for fluid in the system at $t=0$. Note that a quantity of fluid with remaining service time $x+u$ at time $t-u$, has remaining service requirement x at time t for any $u \in [0,t]$. This observation yields, for $x \geq 0$ and $t \geq 0$,

$$b(t,x) = b(0,t+x) + \int_x^{t+x} \gamma(t+x-u)\,\mathrm{d}G(u)$$

and

$$B(t,x) = B(0,t+x) - B(0,t) + \int_0^t \left(\bar{G}(t-u) - \bar{G}(t-u+x)\right)\,\mathrm{d}E(u). \tag{3}$$

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 [21]:

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

where U is the renewal function associated with G, characterized by the renewal equation [4, p. 143]:

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

for $t \geq 0$. Finally, we define $\psi(t,x)$ as the density of fluid with remaining patience $0+$ at time t and initial patience x (either at the arrival time, or at time 0 if fluid is present in the waiting room at that time); here, the $0+$ indicates that fluid is about to leave the waiting room – there is no fluid in the waiting room with remaining patience time 0. Then, $\int_0^t \psi(t,u)\,\mathrm{d}u$ is the amount of fluid about to leave the waiting room at time t.

To conclude our model specification, it seems worthwhile reviewing its primitives. These are the service and patience time distributions (G and F) and the 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 Sections 4 and 5); note that the partial information framework requires also a joint density of true and estimated patience times (H – see Section 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 exploit full information. For comparison sake only, the Most-Patient First policy is analyzed in Appendix A. For these two policies, we define a two-parameter function Q, such that $Q(t,x)$ is the total quantity of fluid in queue at time $t \geq 0$, with patience at most $x \geq 0$:

$$Q(t,x) = \int_0^x q(t,u)\,\mathrm{d}u \quad \text{and} \quad Q(t,\infty) = Q(t), \tag{6}$$

$t \geq 0$, $x \geq 0$; $q(t,x)$ can be interpreted as the density of fluid awaiting service with the *remaining* patience x at time t. Since (6) holds for all $q(t,\cdot)$ that are equal almost everywhere, we let $q(t,\cdot)$ be right-continuous without loss of generality (this property will be used in the rest of the paper).

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:

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

we set $p_\downarrow(t) = \infty$ when $q(t,x) = 0$ for all $x \geq 0$ (the waiting room contains no fluid, $Q(t) = 0$). 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. Based on the above definition, we have

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

The crucial observation is that no abandonment occurs at time t ($\alpha(t) = 0$) if $p_\downarrow(t) > 0$. 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 [29, 21]. 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).

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 to 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 $t \geq u$. This implies that $\bar{F}(p_\downarrow(t,u))$ fraction of fluid arriving to the system at time u is present in the waiting room at time t. 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

$$p_\downarrow(t,u) = \sup_{u \leq x \leq t} \{x - u + p_\downarrow(x)\}. \tag{7}$$

Hence, the total amount of fluid in the waiting room satisfies:

$$Q(t) = Q(0) - Q(0, p_\downarrow(t,0)) + \int_0^t \lambda(u)\, \bar{F}(p_\downarrow(t,u))\, \mathrm{d}u, \quad t \geq 0. \tag{8}$$

A similar argument can be used to determine the structure of the fluid content awaiting service at time t. Consider fluid in the waiting room with remaining patience x at time t. Such fluid is either present in the system at $t=0$ or has arrived during the time period $(t - T_\downarrow(t,x), t]$, where

$$T_\downarrow(t,x) = \inf\{u \in [0,t] : p_\downarrow(t-u) > x+u\}; \tag{9}$$

set $T_\downarrow(t,x) = \infty$ when $p_\downarrow(t-u) \leq x+u$ for all $u \in [0,t]$. The quantity $T_\downarrow(t,x) \wedge t$ represents the length of a time interval over which fluid with remaining patience x at time t is accumulated in the waiting room. Then, the density $q(t,\cdot)$ satisfies, for $x \geq p_\downarrow(t)$,

$$q(t,x) = q(0,x+t) 1_{\{T_\downarrow(t,x) > t\}} + \int_x^{x + t \wedge T_\downarrow(t,x)} \lambda(t+x-u)\, \mathrm{d}F(u). \tag{10}$$

Here, we used the fact that fluid with remaining patience time x at time t must arrive to the system at time $u \in [0,t]$ (or be in the system at time 0 if $u=0$) 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 $T_\downarrow(t,x) > t-u$). The relation (10) can be used to determine $q(t, p_\downarrow(t))$, the density of the "least-patient" fluid awaiting service at time t.

As mentioned earlier, the abandonment rate at time t is positive only when $p_\downarrow(t) = 0$. In that case, one has $\alpha(t) = q(t,0) - \sigma(t)$, namely, fluid with zero remaining patience abandons if it does not enter service. Since $\sigma(t) = b(t,0)$ when the waiting room is not empty (see (1)), it follows that

$$\alpha(t) = (q(t,0) - b(t,0))\, 1_{\{p_\downarrow(t)=0\}}.$$

An additional equation involving $p_\downarrow(t)$ can be obtained by considering $\gamma(t)$, the rate at which fluid enters the service facility. The case $p_\downarrow(t) = 0$ is straightforward: $\gamma(t) = q(t,0) - \alpha(t)$. On the other hand, when $p_\downarrow(t) > 0$, fluid entering service does that either directly, or via the

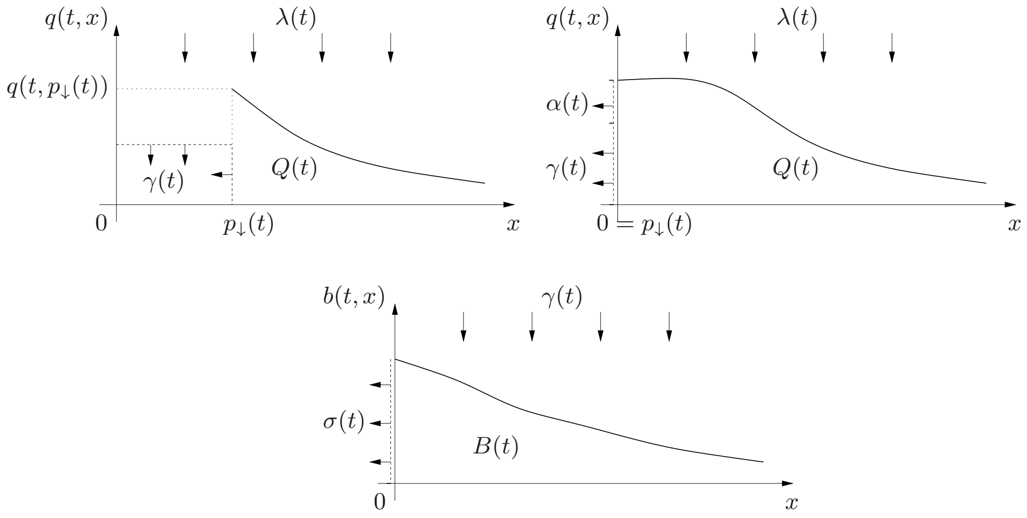

FIGURE 1. LPF: Examples of densities describing fluid in the waiting room (top) and service (bottom) at time t. When $p_\downarrow(t) > 0$ no abandonment occurs. Fluid enters the service facility either directly or via the waiting room.

waiting room (see Figure 1). The two cases can be combined into a single equation (for t such that $p_\downarrow(\cdot)$ is differentiable at t):

$$\gamma(t) + \alpha(t)1_{\{p_\downarrow(t)=0\}} = \lambda(t)F(p_\downarrow(t)) + (1+\dot{p}_\downarrow(t))^+ q(t,p_\downarrow(t)), \tag{11}$$

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. (For a function x differentiable at t, we use $\dot{x}(t)$ to denote its derivative at t.) The value of $q(t,p_\downarrow(t))$ can be obtained from (10). The term $(1+\dot{p}_\downarrow(t))^+$ is due to the fact that remaining patience of fluid awaiting service reduces linearly at rate 1; therefore, no fluid leaves the waiting room if $\dot{p}_\downarrow(t) \le -1$.

Finally, observe that $\int_0^t \psi(t,x)\,\mathrm{d}x = \alpha(t) + \gamma(t)$ under LPF (fluid with expiring patience time either enters the service facility or abandons the system), where

$$\psi(t,x) = q(0,x)1_{\{x=t,\, p_\downarrow(t,0)\le x\}} + \lambda(t-x)f(x)1_{\{p_\downarrow(t,t-x)\le x\}},$$

for $0 \le x \le t$; the condition $p_\downarrow(t,t-x) \le x \le t$ ensures that fluid with patience x arriving at time $(t-x)$ does not enter service by time t.

We conclude this subsection with two concrete 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 $\bar{G}(x) = e^{-\mu x}$, $\bar{F}(x) = e^{-\theta x}$, and $\lambda(t) = (1+\delta)\mu s$, for some $\delta > 0$. This system evolves through three time periods: (i) during $0 \le t < t_1$, some spare processing capacity exists; (ii) during $t_1 \le t < t_2$, there is no spare processing capacity, the queue is non zero, and no fluid abandonment occurs; and (iii) during $t > t_2$, fluid abandonment occurs. Below we derive a detailed description of the evolution. Note that for $0 \le t \le t_1$, the departure rate satisfies

$$\sigma(t) = b(t,0) = \int_0^t \lambda(t-u)\,\mathrm{d}G(u) = (1+\delta)\mu s \left(1 - e^{-\mu t}\right),$$

and, thus, flow conservation (1) yields

$$t_1 = \frac{1}{\mu}\log\frac{1+\delta}{\delta} \tag{12}$$

and

$$B(t) = \begin{cases} (1+\delta)s(1-e^{-\mu t}), & t \le t_1, \\ s, & t \ge t_1. \end{cases} \tag{13}$$

Furthermore, fluid enters service at rate

$$\gamma(t) = \begin{cases} (1+\delta)\mu s, & t < t_1, \\ \mu s, & t \ge t_1. \end{cases} \tag{14}$$

Next, (10), (6) and the fact that $p_\downarrow(t)$ is decreasing on $[t_1, t_2]$ can be used to establish, for $t_1 \le t \le t_2$,

$$Q(t) = (1+\delta)\mu s \left(1 - e^{-\theta(t-t_1)}\right) \frac{1}{\theta} e^{-\theta p_\downarrow(t)}.$$

On the other hand, (1), (14) and $\lambda(t) = (1+\delta)\mu s$ result in

$$Q(t) = \mu s \delta (t - t_1),$$

for $t_1 \le t \le t_2$. These two expressions for $Q(t)$ imply

$$p_\downarrow(t) = \begin{cases} \frac{1}{\theta} \log \frac{(1+\delta)(1-e^{-\theta(t-t_1)})}{\delta\theta(t-t_1)}, & t_1 < t \le t_2, \\ 0, & t \ge t_2, \end{cases} \tag{15}$$

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

$$\dot{p}_\downarrow(t) = \frac{(1+\theta(t-t_1))e^{-\theta(t-t_1)} - 1}{\theta(t-t_1)(1-e^{-\theta(t-t_1)})},$$

with $\dot{p}_\downarrow(t_1+) = -1/2$. Note that $p_\downarrow(t_1+)$ satisfies $\gamma(t_1+) = \lambda(t_1+)F(p_\downarrow(t_1+))$ and

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

Then, t_2 is the root of

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

or

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

Observe that $p_\downarrow(\cdot)$ is differentiable for all $t \ge 0$ except t_1 and t_2. Now, (15) and (10) imply

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

which in turn yields

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

note that $\alpha(t) \to \mu s \delta$, as $t \to \infty$. Based on the preceding, it is straightforward to verify that (11) holds. □

Example 2 ($G_t/D/s$+D LPF fluid model). In certain cases, system behavior can be derived from first principles. 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 $\delta > 0$. Straightforward calculations yield $b(t,x) = \mu s \, 1_{\{0 \le x < 1/\mu\}}$ and

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

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

4.2. **A Numerical Algorithm.** In this subsection, we provide an algorithm for computing relevant functions of the fluid model under LPF. The algorithm is based on an algorithm for the FCFS system [21]. As in [21], we require that the system is initially ($t = 0$) empty ($E(0) = B(0) = Q(0) = A(0) = 0$ and $p_{\downarrow}(0) = \infty$); 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 $t_i = i\delta$, $i = 1, 2, \ldots$, where δ is a time step. Values of $Q(i\delta, x)$ and $B(i\delta, x)$, for various x, can be evaluated based on (10) and (3), respectively. The iterative step depends on whether there exists spare capacity in the system:

- If $B(t_{i-1}) < s$, then $E(t_i) := \Lambda(t_i) - A(t_{i-1})$ and
$$B(t_i) := s \wedge \int_0^{t_i} \bar{G}(t_i - u)\, \mathrm{d}E(u). \tag{16}$$
If $B(t_i) < s$, then $p_{\downarrow}(t_i) := \infty$, $Q(t_i) := 0$ and $A(t_i) := A(t_{i-1})$; otherwise,
$$E(t_i) := B(t_i) + \int_0^{t_i} B(t_i - u)\, \mathrm{d}U(u), \tag{17}$$
where U is evaluated based on (5), and $p_{\downarrow}(t_i)$ solves the non-linear equations (18)-(21):
$$Q(t_i) + A(t_i) = \Lambda(t_i) - E(t_i), \tag{18}$$
where
$$Q(t_i) = \int_0^{t_i} \lambda(u)\, \bar{F}(p_{\downarrow}(t_i, u))\, \mathrm{d}u, \tag{19}$$
$$A(t_i) = A(t_{i-1}) + (\Lambda(t_i) - E(t_i) - A(t_{i-1}) - Q(t_i))^+ \, 1_{\{p_{\downarrow}(t_i)=0\}}, \tag{20}$$
and
$$p_{\downarrow}(t_i, u) = \max_{u \le x \le t_{i-1}} \{x - u + p_{\downarrow}(x)\} \vee (t_i - u + p_{\downarrow}(t_i)) \,. \tag{21}$$
- If $B(t_{i-1}) = s$, then $B(t_i) := s$ and $E(t_i)$ is updated according to (17). If $\Lambda(t_i) - \Lambda(t_{i-1}) + Q(t_i - 1) \le E(t_i) - E(t_{i-1})$, then $E(t_i) = \Lambda(t_i) - A(t_{i-1})$, $p_{\downarrow}(t_i) := \infty$, $Q(t_i) := 0$, $A(t_i) := A(t_{i-1})$, and $B(t_i)$ is updated according to (16); otherwise $p_{\downarrow}(t_i)$ solves (18)-(21).

The rationale for the algorithm is as follows. Under the first case in the iteration, 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, $E(t_i) := \Lambda(t_i) - A(t_{i-1})$ (since there is no abandonment in $[t_{i-1}, t_i]$) and (16), which is based on (2). 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 (17) (see (4)), and the balance equation (1) for the waiting room is utilized. Observe that the right-hand side in (18) is known, while the left-hand side depends on $p_{\downarrow}(t_i)$. The quantity $Q(t_i)$ (evaluated based on (8)) is monotone in $p_{\downarrow}(t_i)$, and $A(t_i) - A(t_{i-1})$ is non-zero only if $p_{\downarrow}(t) = 0$. These two facts imply that the values of $Q(t_i)$ and $A(t_i)$ in (18) are unique. The second case in the iteration 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. It is appropriate to think of $h(x, y)$ as the density of arriving fluid with true patience x and estimated (perceived) patience y; then, $H(x, \infty) =$

$F(x)$. The distribution H defines two relevant conditional distributions. For fluid with (true) patience x, the conditional density 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 score of this work, and are left for future research.)

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 (e.g., 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

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

where $(Z,\hat{Z})$ is bivariate normal with $\mathbb{E}Z = \mathbb{E}\hat{Z} = \theta$, $\operatorname{Var}(Z) = \operatorname{Var}(\hat{Z}) = \sigma^2$ and $\operatorname{Cov}(Z,\hat{Z}) = \rho\sigma^2$. 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

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

$x,y \ge 0$. 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 $x \ge 0$, given that the estimated patience is $y \ge 0$, is

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

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 $\sqrt{e^{\sigma^2(1-\rho^2)}-1}$ 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 $t \ge 0$, with patience at most $x \ge 0$ and estimated patience at most y:

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

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

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

Note that, unlike in Section 4, $p_\downarrow(t)$ can take negative values, since it characterizes estimated patience rather than true patience. As remarked earlier, both true and estimated remaining patience decrease linearly at rate one until the corresponding fluid leaves the waiting room. Based on the above definition, it follows that

$$Q(t) = \int_{p_\downarrow(t)}^{\infty} \int_0^{\infty} q(t,x,y)\,\mathrm{d}x\,\mathrm{d}y. \tag{22}$$

However, an additional expression for $Q(t)$ can be derived. The following equation states that fluid that has not entered service or abandoned is present in the waiting room:

$$Q(t) = \int_{p_\downarrow(t,0)}^{\infty} \int_t^{\infty} q(0,x,y)\,\mathrm{d}x\,\mathrm{d}y + \int_0^t \lambda(u)\,\bar{H}(t-u, p_\downarrow(t,u))\,\mathrm{d}u,$$

where $p_\downarrow(t,u)$ is as in (7) and $\bar{H}(x,y) = \int_x^\infty \int_y^\infty h(u,v)\,\mathrm{d}v\,\mathrm{d}u$. The fact that both remaining patience and estimated remaining patience decrease linearly implies the following description of the density of fluid awaiting service:

$$q(t,x,y) = q(0,x+t,y+t)1_{\{T_\downarrow(t,x)>t\}} + \int_0^{t\wedge T_\downarrow(t,y)} \lambda(t-u)\,h(x+u,y+u)\,\mathrm{d}u, \tag{23}$$

for $x \geq 0$ and $y \geq p_\downarrow(t)$, where $T_\downarrow(t,x)$ is as in (9). In order to determine the abandonment rate $\alpha(t)$, one must consider the actual remaining patience rather than the estimated counterpart. In particular, fluid with 0 remaining patience abandons the waiting room:

$$\alpha(t) = \int_{p_\downarrow(t)}^{\infty} q(t,0,y)\,\mathrm{d}y, \tag{24}$$

where (23) defines $q(t,0,y)$. Alternatively, the abandonment rate also satisfies, for $t \geq 0$,

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

The total amount of abandonment by time t, $A(t)$, can be expressed in a couple of ways as well. First, in view of (24), $A(t)$ satisfies

$$A(t) = \int_0^t \int_{p_\downarrow(u)}^{\infty} q(u,0,y)\,\mathrm{d}y\,\mathrm{d}u. \tag{25}$$

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

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

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 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 $q(t,x,p_\downarrow(t))$, leading to

$$\gamma(t)=\lambda(t)\,H(\infty,p_\downarrow(t))+(1+\dot p_\downarrow(t))^+\int_0^\infty q(t,x,p_\downarrow(t))\,\mathrm{d}x,$$

for t such that $\dot p_\downarrow(t)$ is well defined. The above differential equation is an analogue of (11), which holds for the case of full information. Note that $\int_0^t\psi(t,x)\,\mathrm{d}x=\alpha(t)$ under the model described in this section, where

$$\psi(t,x)=1_{\{x=t\}}\int_{p_\downarrow(t,0)}^\infty q(0,x,y)\,\mathrm{d}y+\int_{p_\downarrow(t,t-x)}^\infty\lambda(t-x)\,h(x,y)\,\mathrm{d}y,$$

for $0\le x\le t$.

Finally, we note that the numerical algorithm outlined in Section 4.2 is applicable to evaluating fluid models under partial information with one modification: $p_\downarrow(t)$ solves (18), 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 $\bar H(x,y)=e^{-\theta(x+y)}$, $x,y\ge 0$. 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 $[0,t_1]$. Based on the fact that $p_\downarrow(\cdot)$ is not increasing and (22) with (23), one has, for $t\ge t_1$,

$$Q(t)=\frac{(1+\delta)\mu s}{2\theta}e^{-\theta p_\downarrow(t)}\left(1-e^{-2\theta(t-t_1)}\right). \tag{26}$$

Since the actual patience times are exponentially distributed, it follows that $\alpha(t)=\theta Q(t)$ – just as under FCFS (can be verified via (25)). This, together with (1), yields

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

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

$$Q(t)=\frac{\delta\mu s}{\theta}\left(1-e^{-\theta(t-t_1)}\right)1_{\{t\ge t_1\}}, \tag{27}$$

and consequently

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

An expression for $p_\downarrow(t)$, $t\ge t_1$, follows from (26) and (27):

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

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

For LPF and MPF, the proposed numerical algorithm was used; for FCFS, the algorithm in [21] was used. The time step was set to 0.01, and the trapezoidal rule was used to evaluate integrals.

Example 5 (LPF & 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 [21]: $G(t) = (1-e^{-2t})/2+(1-e^{-2t/3})/2$ and $F(t) = 1-e^{-t}-te^{-t}$; 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 Figure 2, 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 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 exists enough fluid in the waiting room that can be delayed without causing abandonments. In Figure 3, 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, $\bar{G}(x) = e^{-x}$, $x \geq 0$, while the joint distribution of patience and estimated patience is as follows (see Example 3):

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

$x, y \geq 0$ 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 parameter ρ. In Figure 4, 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. □

7. Concluding Remarks

In this final section, we discuss the application of fluid models to approximate a stochastic queueing system with a finite number of servers ($\mathrm{G}_t/\mathrm{GI}/n+\mathrm{GI}$). As argued in [30] 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

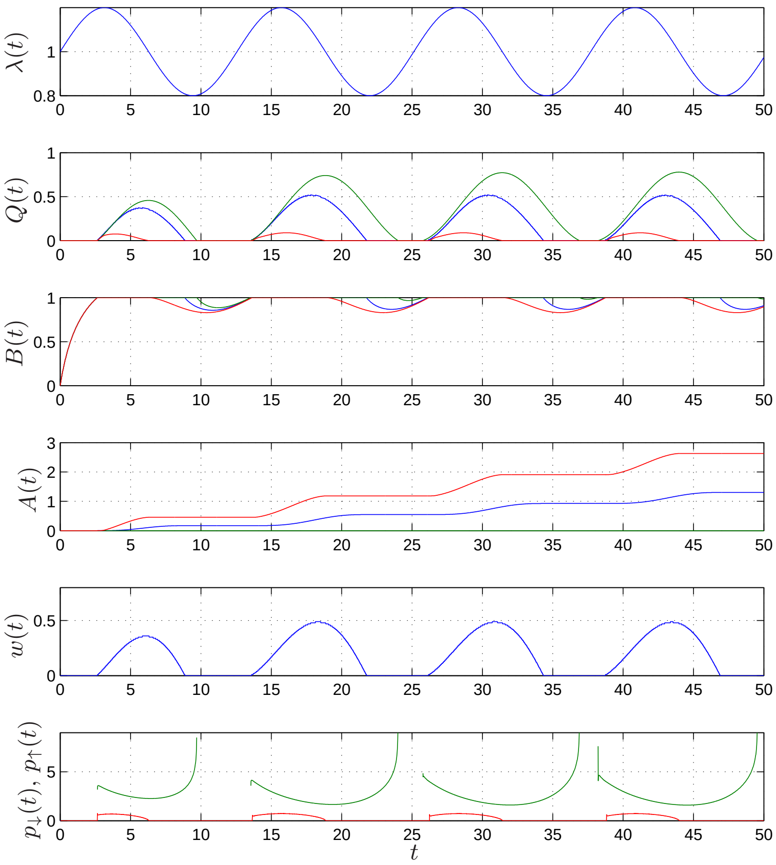

Figure 2. Performance functions for the system described in Example 5, operating under LPF (green), MPF (red) and FCFS (blue). No abandonment occurs under LPF. The function w represents the waiting time of head-of-line fluid in the waiting room when FCFS is used.

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.

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 be added to capture some of the present stochastic variability:

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

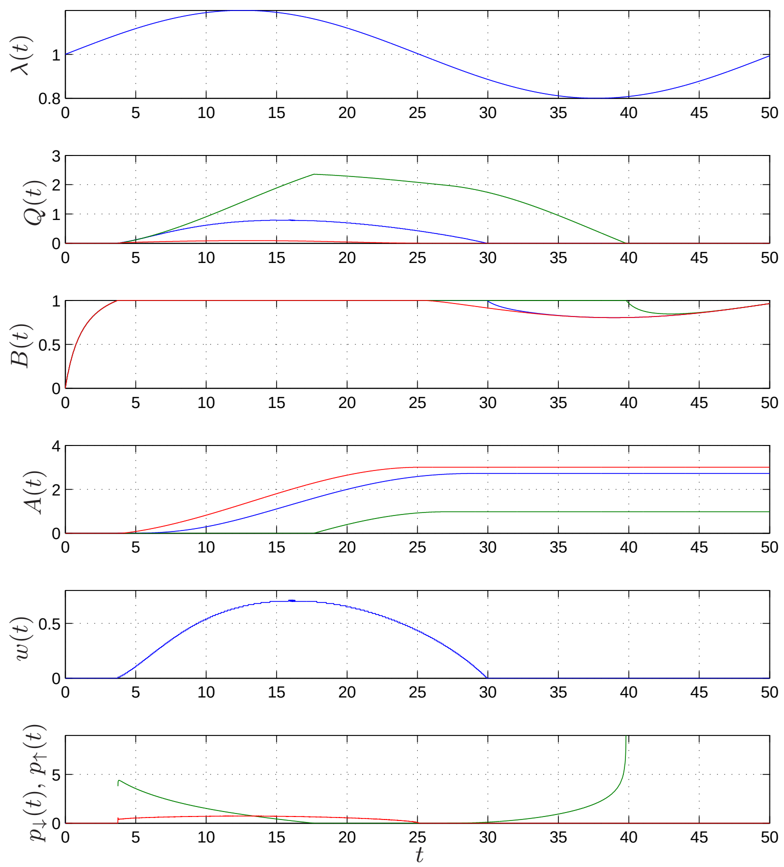

Figure 3. Performance functions for the system described in Example 5, operating under LPF (green), MPF (red) and FCFS (blue). The function w represents the waiting time of head-of-line fluid in the waiting room when FCFS is used.

here, the second term approximates the probability of a customer abandoning 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 fluid-model equations leads to a corrected version (that depends on the number of servers) of our numerical algorithm. In particular, we have

$$\tilde{A}(t_i) = \tilde{A}(t_{i-1}) + (\Lambda(t_i) - E(t_i) - \tilde{A}(t_{i-1}) - Q(t_i))^+ 1_{\{p_\downarrow(t_i)=0\}} + \int_{t_{i-1}}^{t_i} \lambda(u)\, F\left(\frac{1}{n\sigma(u)}\right) 1_{\{p_\downarrow(u)\in(0,\infty)\}}\, \mathrm{d}u, \quad (28)$$

where tildes distinguish the present refined abandonment process from its previous version (20). Note that the last term in (28) is proportional to $1/n$ for large n. Hence, as seen from the example at the end of the section, such a correction term produces only marginal improvements in the quality of the approximation.

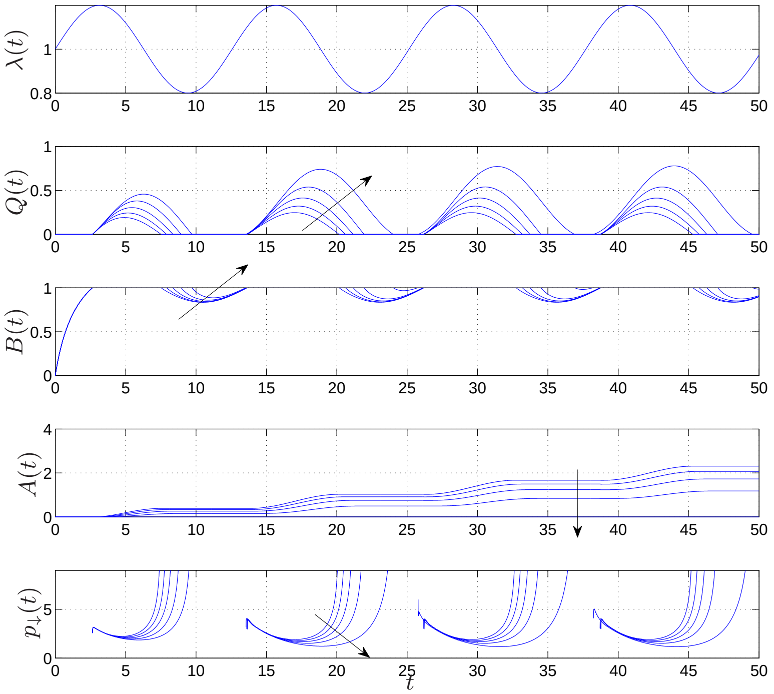

FIGURE 4. Performance functions for the system described in Example 6, for $\rho = 0, 0.25, 0.5, 0.75$ and 1. The arrows indicate the direction of increase for ρ.

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 produces a bias relative to their fluid functions due to Jensen's inequality. The reader is referred to [1, 20] 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 $1/\sqrt{n}$-order. 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 $100\lambda(t)$, respectively). In Figure 5, 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 10000 and 1000 replications, respectively. For the system with 50 servers, we also plot the corrected fluid approximation (based on (28)) as well. However, as it can be seen in the figure, this approximation provides only a minor improvement; this is consistent with our discussion in this section. □

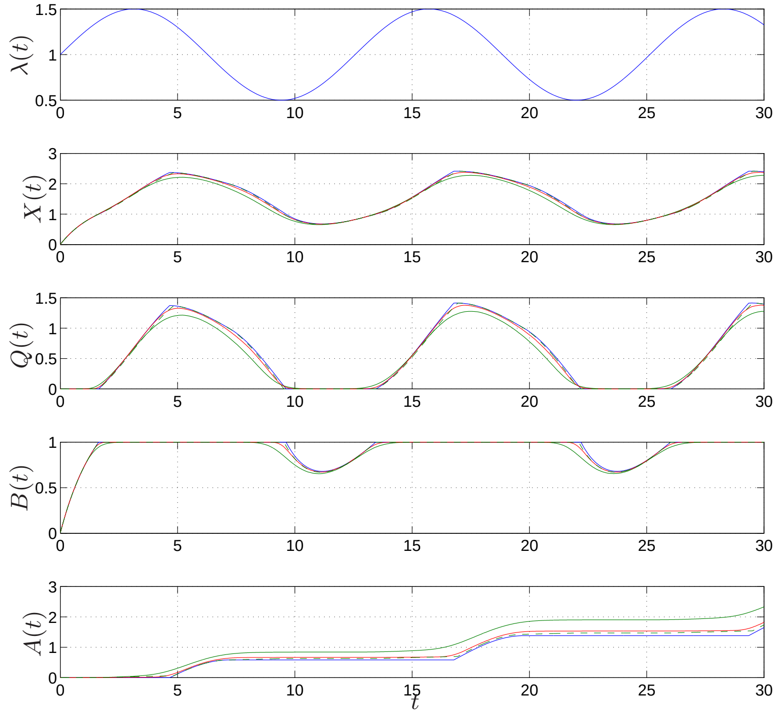

Figure 5. Performance functions for the three systems described in Example 7. For the two finite stochastic systems (50 and 250 servers), empirical averages are plotted (in green and red, respectively). A corrected fluid approximation for the system with 50 servers is shown with dashed lines.

Acknowledgments

The work of A.M. has been partially supported by BSF Grants 2005175 and 2008480, ISF Grant 1357/08 and by the Technion funds for promotion of research and sponsored research. Some of the research was 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 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] E. Altman, T. Jiménez, and G. Koole. On the comparison of queueing systems with their fluid limits. *Probab. Engin. Inform. Sc.*, 15(2):165–178, 2001. 7

[2] N. Argon and S. Ziya. Priority assignment under imperfect information on customer type identities. *Manufacturing Service Oper. Management*, 11(4):674–693, 2009. 2

[3] N. Argon, S. Ziya, and R. Righter. Scheduling impatient jobs in a clearing system with insights on patient triage in mass casualty incidents. *Probab. Engin. Inform. Sci.*, 22(3):301–332, 2008. 2

[4] S. Asmussen. *Applied Probability and Queues.* Springer-Verlag, New York, NY, 2nd edition, 2003. 3

[5] R. Atar, A. Biswas, and H. Kaspi. Fluid limits of G/G/1+G queues under the non-preemptive earliest-deadline-first discipline. Preprint, 2013. 2

[6] A. Bassamboo and R. Randhawa. Using estimated patience levels to optimaly schedule customers. Preprint, 2013. 2

[7] I. Cohen, A. Mandelbaum, and N. Zychlinski. Minimizing mortality in a mass casualty event: Fluid networks in support of modeling and management. Preprint, 2013. 1

[8] L. Decreusefond and P. Moyal. Fluid limit of a heavily loaded EDF queue with impatient customers. *Markov Process. Related Fields*, 14(1):131–158, 2008. 2

[9] M. Dertouzos. Control robotics: The procedural control physical processes. In *Proc. IFIP Congress*, Stockholm, Sweden, August 1974. 2

[10] B. Doytchinov, J. Lehoczky, and S. Shreve. Real-time queues in heavy traffic with earliest-deadline-first queue discipline. *Ann. Appl. Probab.*, 11(2):332–378, 2001. 2

[11] N. Gans, N. Liu, A. Mandelbaum, H. Shen, and H. Ye. Service times in call centers: Agent heterogeneity and learning with some operational consequences. In J. Berger, T.T. Cai, and I. Johnstone, editors, *Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown*, volume 6 of *Collections*, pages 99–123. Institute of Mathematical Statistics, Beachwood, OH, 2010. 1

[12] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing Service Oper. Management*, 4(3):208–227, 2002. 1

[13] R. Ghebali. Real-time prediction of the probability of abandonment in call centers. Master's thesis, Technion – Israel Institute of Technology, Haifa, Israel, 2012. 1

[14] L.V. Green, J. Soares, J.F. Giglio, and R.A. Green. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Acad. Emerg. Med.*, 13(1):61–68, 2006. 1

[15] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.*, 29(3):567–588, 1981. 1

[16] J. Hong, X. Tan, and D. Towsley. A performance analysis of minimum laxity and earliest deadline scheduling in a real-time system. *IEEE Trans. Comput.*, 38(12):1736–1744, 1989. 1

[17] J. Jackson. Some problems in queueing with dynamic priorities. *Naval Res. Log. Quart.*, 7(3):235–249, 1960. 2

[18] J. Jackson. Waiting time distribution for queues with dynamic priorities. *Naval Res. Log. Quart.*, 9(1):31–36, 1960. 2

[19] J. Jackson. Queues with dynamic priority discipline. *Management Sci.*, 8(1):18–34, 1961. 2

[20] T. Jiménez and G. Koole. Scaling and comparison of fluid limits of queues applied to call centers with time-varying parameters. *OR Spectrum*, 26(3):413–422, 2004. 7

[21] W. Kang and G. Pang. Computation and properties of fluid models of time-varying many-server queues with abandonment. Preprint, 2013. 2, 3, 4.1, 4.2, 6, 5

[22] W. Kang and G. Pang. Fluid limit of a multiclass many-server queueing network with abandonment and feedback. Preprint, 2013. 2

[23] W. Kang and K. Ramanan. Fluid limits of many-server queues with reneging. *Ann. Appl. Probab.*, 20(6):2204–2260, 2010. 2

[24] E. Kaplan. Terror queues. *Oper. Res.*, 58(4):773–784, 2010. 1

[25] H. Kaspi and K. Ramanan. Law of large numbers limits for many-server queues. *Ann. Appl. Probab.*, 21(1):33–114, 2011. 2

[26] Ł. Kruk, J. Lehoczky, K. Ramanan, and S. Shreve. Heavy traffic analysis for EDF queues with reneging. *Ann. Appl. Probab.*, 21(2):484–545, 2011. 2

[27] Ł. Kruk, J. Lehoczky, and S. Shreve. Accuracy of state space collapse for earliest-deadline-first queues. *Ann. Appl. Probab.*, 16(2):516–561, 2006. 2

[28] D. Li and K. Glazebrook. A Bayesian approach to the triage problem with imperfect information. *Eur. J. Oper. Res.*, 215(1):169–180, 2011. 2

[29] Y. Liu and W. Whitt. A network of time-varying many-server fluid queues with customer abandonment. *Oper. Res.*, 59(4):835–846, 2011. 2, 4.1

[30] Y. Liu and W. Whitt. The G_t/GI/s_t+GI many-server fluid queue. *Queueing Syst. Theory Appl.*, 71(4):405–444, 2012. 2, 7

[31] Y. Liu and W. Whitt. 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. 2

[32] A. Mandelbaum, W. Massey, and M. Reiman. Strong approximations for Markovian service networks. *Queueing Syst. Theory Appl.*, 30(1-2):149–201, 1998. 2

[33] A. Mandelbaum and P. Momčilović. Queues with many servers and impatient customers. *Math. Oper. Res.*, 37(1):41–64, 2012. 1

[34] A. Mandelbaum and S. Zeltyn. Data-stories about (im)patient customers in tele-queues. *Queueing Syst. Theory Appl.*, 75(2-4):115–146, 2013. 1

[35] O. Merin, N. Ash, G. Levy, M.J. Schwaber, and Y. Kreiss. The Israeli field hospital in Haiti – Ethical dilemmas in early disaster response. *N. Engl. J. Med.*, 362(11):e38, 2010. 1

[36] J. Van Mieghem. Due date scheduling: Asymptotic optimality of generalized longest queue and generalized largest delay rules. *Oper. Res.*, 51(1):113–122, 2003. 1

[37] J. Mistovich, B. Hafen, and K. Karren. *Prehospital Emergency Care*. Prentice Hall Health, 2000. 1

[38] P. Moyal. Convex comparison of service disciplines in real-time queues. *Oper. Res. Lett.*, 36(4):496–499, 2008. 2

[39] P. Moyal. On queues with impatience: Stability, and the optimality of Earliest Deadline First. *Queueing Syst. Theory Appl.*, 75(2-4):211–242, 2013. 1, 2

[40] C. Palm. Methods of judging the annoyance caused by congestion. *Tele*, 2:1–20, 1953. 1

[41] G. Pang and W. Whitt. Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Syst. Theory Appl.*, 65(4):325–364, 2010. 2

[42] S. Panwar, D. Towsley, and J. Wolf. Optimal scheduling policies for a class of queues with customer deadlines to the beginning of service. *J. ACM*, 35(4):832–844, 1988. 1, 2

[43] J. Reed. The G/GI/N queue in the Halfin-Whitt regime. *Ann. Appl. Probab.*, 19(6):2211–2269, 2009. 1

[44] S. Saghafian, W.J. Hopp, M.P. Van Oyen, J.S. Desmond, and S.L. Kronick. Patient streaming as a mechanism for improving responsiveness in emergency departments. *Oper. Res.*, 60(5):1080–1097, 2012. 2

[45] S. Saghafian, W.J. Hopp, M.P. Van Oyen, J.S. Desmond, and S.L. Kronick. Complexity-based triage: A tool for improving patient safety and operational efficiency. Preprint, 2013. 2

[46] J. Spencer, M. Sudan, and K. Xu. Queueing with future information. Preprint, 2013. 2

[47] A. Stoyenko and L. Georgiadis. On optimal lateness and tardiness scheduling in real-time systems. *Computing*, 47(3-4):215–234, 1992. 2

[48] L. Wein. Due-date setting and priority sequencing in a multiclass M/G/1 queue. *Management Sci.*, 37(7):834–850, 1991. 1

[49] W. Whitt. Fluid models for multiserver queues with abandonments. *Oper. Res.*, 54(1):37–54, 2006. 2, 3

Appendix A. Most-Patient First

Our analysis of MPF scheduling is similar to the analysis in the previous subsection. In this case, a quantity of fluid enters service only if no other fluid with greater remaining patience is present in the waiting room. We define $p_{\uparrow}(t) \in [0, \infty]$ to be the remaining patience of the most patient fluid awaiting service:

$$p_{\uparrow}(t) = \sup\{x \geq 0 : \ q(t,x) > 0\};$$

we set $p_{\uparrow}(t) = 0$ when the queue is empty; hence,

$$Q(t) = \int_0^{p_{\uparrow}(t)} q(t,x)\,\mathrm{d}x.$$

Now, consider a quantity of fluid that arrives to the system with patience x at time u. This fluid is present in the waiting room at time $t \geq u$, if $x > t - u$ and $p_{\uparrow}(u + y) \geq x - y$ for all $0 \leq y \leq t - u$. Thus, $F(p_{\uparrow}(t,u))$ fraction of fluid arriving to the system at time u is present in the waiting room at time t, where

$$p_{\uparrow}(t,u) = \inf_{u \leq x \leq t} \{x - u + p_{\uparrow}(x)\}.$$

Furthermore, the total amount of fluid in the waiting room satisfies:

$$Q(t) = Q(0, p_{\uparrow}(t,0) \vee t) - Q(0,t) + \int_0^t \lambda(u)\ \left(F(p_{\uparrow}(t,u)) - F(t-u)\right)^+ \mathrm{d}u; \tag{29}$$

a similar reasoning leads to

$$A(t) = Q(0, p_{\uparrow}(t,0) \wedge t) + \int_0^t \lambda(u)\, F(p_{\uparrow}(t,u) \wedge (t-u))\,\mathrm{d}u. \tag{30}$$

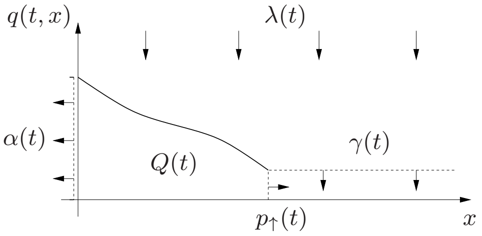

FIGURE 6. MPT: An example of a density describing fluid awaiting service at time t.

As in the LPF case, one can determine the structure of the fluid content awaiting service at time t. Consider fluid in the waiting room with remaining patience x at time t. Such fluid is either present in the system at $t = 0$ or has arrived during the time period $(t - T_\uparrow(t,x), t]$, where

$$T_\uparrow(t,x) = \inf\{u \in [0,t] : p_\uparrow(t-u) < x+u\};$$

set $T_\uparrow(t,x) = \infty$ when $p_\uparrow(t-u) < x+u$ for all $u \in [0,t]$. The quantity $T_\uparrow(t,x) \wedge t$ represents the length of a time interval over which fluid with remaining patience x at time t is accumulated in the waiting room. Then, the density $q(t,\cdot)$ satisfies, for $x \le p_\uparrow(t)$,

$$q(t,x) = q(0,x+t)1_{\{T_\uparrow(t,x)>t\}} + \int_x^{x+t\wedge T_\uparrow(t,x)} \lambda(t+x-u)\,\mathrm{d}F(u). \tag{31}$$

The abandonment rate at time t is $\alpha(t) = q(t,0)$, while for the rate $\gamma(t)$ one has (see Figure 6):

$$\gamma(t) = \lambda(t)\bar{F}(p_\uparrow(t)) + (\dot{p}_\uparrow(t) - 1)^+ q(t, p_\uparrow(t)); \tag{32}$$

here, similarly to the LPF case (see (11)), this equation holds for t that is a differentiability point of $p_\uparrow(\cdot)$. The value of $q(t,p_\uparrow(t))$ can be obtained from (31). No fluid is transferred from the waiting room to the service facility at time t if $\dot{p}_\uparrow(t) < 1$.

Finally, note that $\int_0^t \psi(t,x)\,\mathrm{d}x = \alpha(t)$ under MPF, where

$$\psi(t,x) = q(0,x)1_{\{x=t,\, p_\uparrow(t,0)\ge x\}} + \lambda(t-x)f(x)1_{\{p_\uparrow(t,t-x)\ge x\}},$$

for $0 \le x \le t$; the condition $p_\uparrow(t,t-x) \ge x$ ensures that fluid with patience x arriving at time $(t-x)$ does not enter service by time t.

A numerical algorithm for computing relevant functions of the fluid model under MPF can be obtained by modifying the algorithm in Section 4.2: (i) $p_\downarrow$ is replaced with $p_\uparrow$, (ii) $p_\downarrow(t) := \infty$ is replaced with $p_\uparrow(t) := 0$, and (iii) $p_\uparrow(t)$ solves (18), with

$$Q(t_i) = \int_0^{t_i} \lambda(u)\ (F(p_\uparrow(t_i,u)) - F(t_i-u))^+\ \mathrm{d}u,$$
$$A(t_i) = \int_0^{t_i} \lambda(u)\,F(p_\uparrow(t_i,u) \wedge (t_i-u))\,\mathrm{d}u,$$

and

$$p_\uparrow(t_i,u) = \min_{u \le x \le t_{i-1}} \{x - u + p_\uparrow(x)\} \wedge (t_i - u + p_\uparrow(t_i)).$$

We conclude this appendix with a concrete example.

Example 8 (G$_t$/M/s+M MPF fluid model). Consider the setup described in Example 1. The systems operating under LPF and MPF are equivalent on the time interval $[0,t_1)$, where t_1 is defined by (12) (since no waiting occurs then). Moreover, (13) and (14) hold for the MPF systems as well. The behavior of $p_\uparrow$ is simple in this case: $p_\uparrow(t)$ is a constant for $t \geq t_1$ and it satisfies $\lambda(t)\,\bar{F}(p_\uparrow(t)) = \gamma(t)$ (see (32)). Hence, one has

$$p_\uparrow(t) = \frac{1}{\theta}\log(1+\delta)\,1_{\{t \geq t_1\}}$$

and as a consequence, for $u \leq t$,

$$p_\uparrow(t,u) \wedge (t-u) = \begin{cases} 0, & u < t_1, \\ p_\uparrow(t_1), & t_1 \leq u < t - p_\uparrow(t_1), \\ t-u, & t - p_\uparrow(t_1) \leq u \leq t. \end{cases}$$

Then, from the preceding equality, (29) and (30) it follows that

$$Q(t) = \begin{cases} (1+\delta)\mu s \frac{1}{\theta}\left(1 - e^{-\theta(t-t_1)}\right) - \mu s(t-t_1), & t_1 \leq t \leq t_1 + p_\uparrow(t_1), \\ \mu s \frac{1}{\theta}\left(\delta - \log(1+\delta)\right), & t \geq t_1 + p_\uparrow(t_1), \end{cases}$$

and

$$\alpha(t) = \begin{cases} (1+\delta)\mu s(1 - e^{-\theta(t-t_1)}), & t_1 \leq t \leq t_1 + p_\uparrow(t_1), \\ \mu s\delta, & t \geq t_1 + p_\uparrow(t_1). \end{cases}$$

□

Faculty of Industrial Engineering and Management, Technion, Haifa 3200, Israel
E-mail address: avim@tx.technion.ac.il

Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611
E-mail address: petar@ise.ufl.edu