PROPERTIES OF THE GITTINS INDEX WITH APPLICATION TO OPTIMAL SCHEDULING

Samuli Aalto[1], Urtzi Ayesta[2,3], Rhonda Righter[4]

[1] Aalto University School of Science and Technology, Finland

[2] BCAM – Basque Center for Applied Mathematics, Spain

[3] IKERBASQUE – Basque Foundation for Science, Spain

[4] University of California at Berkeley, USA

November 11, 2010

Contact author Samuli Aalto, E-mail: samuli.aalto@tkk.fi

Short title GITTINS INDEX APPLIED TO OPTIMAL SCHEDULING

Abstract

We consider the optimal scheduling problem for a single-server queue without arrivals. We allow preemptions, and our purpose is to minimize the expected flow time. The optimal non-anticipating discipline is known to be the Gittins index policy, which, however, is defined in an implicit way. Until now, its general behaviour in this specific problem has been characterized only in a few special cases. In this paper, we give as complete a characterization as possible. It turns out that the optimal policy always belongs to the family of Multi-Level-Processor-Sharing (MLPS) disciplines.

1 Introduction

Consider a single-server queue with n jobs at time 0. Jobs are served according to a work conserving and non-anticipating scheduling discipline π that allows preemptions. Let Π denote the family of such disciplines. Let S_j denote the random service time of job j (a.k.a. processing time). In addition, let C_j^π denote the random completion time of job j under the scheduling policy π.

We assume that jobs are statistically identical, i.e., that their unknown service times S_j can be modelled as being independent and identically distributed random variables, with common distribution function $F(x) = P\{S \leq x\}$, $x \geq 0$, which has a finite mean $E[S] < \infty$. The tail function is denoted by $\overline{F}(x) = 1 - F(x)$. We assume that $\overline{F}(x) > 0$ for all $x \geq 0$. In addition we assume that the service time distribution has a continuous density function $f(x)$, $x \geq 0$. The hazard rate $h(x)$, $x \geq 0$, is defined by

$$h(x) = \frac{f(x)}{\overline{F}(x)} = \frac{f(x)}{\int_x^\infty f(y)\,\mathrm{d}y}. \tag{1}$$

The continuity of the hazard rate function $h(x)$ is inherited from the density function $f(x)$. Our third assumption concerning the service time distribution is that the hazard rate is piecewise monotonic. To rule out pathological cases we assume that in any finite interval the direction of $h(x)$ changes only a finite number of times. Otherwise the distribution is arbitrary. The assumption that $h(x)$ is piecewise monotonic is not very restrictive, but there do exist continuous functions that are not monotonic anywhere; an example is the van der Waerden function.

In addition, we define, for all $x \geq 0$,

$$H(x) = \frac{\int_x^\infty f(y)\,\mathrm{d}y}{\int_x^\infty \overline{F}(y)\,\mathrm{d}y} = \frac{\overline{F}(x)}{\int_x^\infty \overline{F}(y)\,\mathrm{d}y}. \tag{2}$$

The function $H(x)$ is related to the mean remaining service time (a.k.a. mean residual lifetime) as follows:

$$E[S - x \mid S > x] = \frac{\int_x^\infty \overline{F}(y)\,\mathrm{d}y}{\overline{F}(x)} = \frac{1}{H(x)}. \tag{3}$$

Throughout the paper we use the terms *increasing* and *decreasing* in their weak form so that the corresponding functions need *not* be strictly increasing or decreasing. Strict monotonicity is expressed explicitly.

The optimal non-anticipating scheduling discipline $\pi^* \in \Pi$ minimizes the expected sum of completion times of all jobs,

$$\sum_{j=1}^{n} E[C_j^{\pi^*}] = \min_{\pi \in \Pi} \sum_{j=1}^{n} E[C_j^{\pi}]. \tag{4}$$

In other words, the objective is to minimize the expected flow time. In scheduling terminology, this is a stochastic version of the scheduling problem $1 \mid prmp \mid \sum C_j$, see, e.g., [7].

For the present problem, the optimal non-anticipating discipline is known to be the Gittins index policy [5, 10]. Unfortunately, the Gittins index policy is defined in a highly implicit way. To find out how the optimal policy operates, one is, in general, urged to completely fix the service time distribution (including the numerical values for all the free parameters describing the distribution).

Until now, the Gittins index policy for this specific problem has been characterized only in few special cases, see [3]. If the service time distribution belongs to the New-Better-than-Used-in-Expectation (NBUE) class, i.e., $H(x) \geq H(0)$ for all $x \geq 0$, the Gittins index policy coincides with any non-preemptive scheduling discipline, e.g., First-Come-First-Served (FCFS) [9]. For the Decreasing-Hazard-Rate (DHR) class, for which $h(x)$ is decreasing for all $x \geq 0$, the Gittins index policy is equal to the Foreground-Background (FB) discipline [11, 8], which in this static setting without arrivals operates just like the Processor-Sharing (PS) discipline. The third example relates to distributions in which jobs initially behave like those in the NBUE class, but after receiving a certain amount of service, they behave like those in the DHR class. In this case, the Gittins index policy is a two-level priority policy called FCFS+FB [1, 3]. We note that the service time distribution class given in the third example is a generalization of the class for which the hazard rate is bitonic, i.e., first increasing and then decreasing [5].

In this paper, we give as complete a characterization as possible for the Gittins index policy in this stochastic single-server scheduling problem without arrivals and under the assumption that all jobs start with the same amount of attained service (a.k.a. age). It turns out that the optimal policy is characterized by a sequence of intervals $0, 1, 2, \ldots$ defined by thresholds $0 = \alpha_0 < \alpha_1 < \alpha_2 < \cdots$. At the start of interval i, all jobs have the same age,

α_i, and then either (i) jobs are taken one at a time and each is served until it completes or its age increases to α_{i+1}, or (ii) jobs are served together (using processor-sharing) until such point as their common age is α_{i+1} (with some jobs dropping out, as they are completed). Intervals of types (i) and (ii) alternate. Thus, the optimal policy always belongs to the family of Multi-Level-Processor-Sharing (MLPS) disciplines, defined in [6]. Recent results concerning the MLPS diciplines in the single-server scheduling problem with arrivals are summarized in [2].

To enable the characterization, we derive several new properties of the Gittins index itself in this job scheduling setting. Notably, we prove that if the hazard rate function of the service time distribution is continuous and piecewise monotonic, the corresponding Gittins index function has the same properties. In addition, the Gittins index is increasing in an interval whenever the hazard rate is increasing in the same interval or the mean residual lifetime function is decreasing in the interval. When the hazard rate is decreasing in an interval, there are three different possible alternatives: (i) the Gittins index is decreasing in the interval, (ii) the Gittins index is first decreasing and then increasing in the interval, and (iii) the Gittins index is increasing in the interval. Any other (non-monotonic or even non-regular) behaviour can be ruled out.

The paper is organized as follows. Section 2 introduces the Gittins index and the corresponding index policy. Some prior results related to the Gittins index are given in Section 3. Sections 4 and 5 include our main contribution. In Section 4 we derive new properties of the Gittins index needed for the general characterization of the Gittins index policy presented in Section 5, which also includes some illustrative numerical examples. Section 6 concludes the paper.

2 Gittins index policy

The optimal non-anticipating scheduling discipline $\pi^* \in \Pi$ that minimizes the expected flow time is the policy that computes, separately for each job, an index based on the age of the job and then chooses any job with the *highest index*. The index, defined below, is called the Gittins index, and the corresponding policy the *Gittins index policy*. Gittins derived this result as a by-product of his ground-breaking results on the multi-armed bandit problem [5].

A multi-armed bandit is a finite collection of finite-state Markov processes of which exactly one (the chosen one) is evolving at a time while the other bandits are frozen. The Gittins index that determines the optimal policy to choose the bandits is, for each bandit, a function of the state of the bandit. In the job scheduling setting, Markovian bandits are replaced by jobs with stochastic service requirements, and the state of the job is described by a continuous variable indicating its age. Therefore, in this setting, the Gittins index has some special properties, which we will describe in the paper.

Independently of Gittins, Sevcik [10] studied the optimal scheduling problem in a single-server queue without arrivals, and proved the optimality of the Smallest-Rank policy. However, the two policies are the same (as they should be) since the reciprocal of the Sevcik rank is equal to the Gittins index in this setting.

For the definition of the Gittins index, an auxiliary function $J(x,\Delta)$, $x,\Delta \geq 0$, is needed, which is called the *efficiency function* and defined, for $\Delta > 0$, by

$$J(x,\Delta) = \frac{\int_x^{x+\Delta} f(y)\,\mathrm{d}y}{\int_x^{x+\Delta} \overline{F}(y)\,\mathrm{d}y} = \frac{\overline{F}(x) - \overline{F}(x+\Delta)}{\int_x^{x+\Delta} \overline{F}(y)\,\mathrm{d}y}. \tag{5}$$

In addition, let

$$J(x,0) = \frac{f(x)}{\overline{F}(x)} = h(x), \quad J(x,\infty) = \frac{\overline{F}(x)}{\int_x^{\infty} \overline{F}(y)\,\mathrm{d}y} = H(x). \tag{6}$$

Note that $J(x,\Delta)$ is continuous with respect to both of the arguments x and Δ. Note also that, for any $\Delta > 0$,

$$J(x,\Delta) = \frac{P\{S - x \leq \Delta \mid S > x\}}{E[\min\{S-x,\Delta\} \mid S > x]}. \tag{7}$$

Thus, for a job that has attained service x and is assigned Δ units of service, the efficiency function $J(x,\Delta)$ is the ratio between (i) the probability that the job will complete within a service quantum of Δ and (ii) the expected service time during this quantum Δ.

The *Gittins index* $G(x)$, $x \geq 0$, is defined by

$$G(x) = \sup_{\Delta \geq 0} J(x,\Delta). \tag{8}$$

In addition, for any $x \geq 0$, we define the *optimum quantum of service* by

$$\Delta^*(x) = \sup\{\Delta \geq 0 \mid G(x) = J(x,\Delta)\}. \tag{9}$$

Note that $J(x, \Delta^*(x)) = G(x)$ by definition. From this discussion we have the following.

Lemma 1 *$G(x) \geq \max\{h(x), H(x)\}$ for all x, and if $G(x) > h(x)$ then $\Delta^*(x) > 0$.*

3 Prior results

In this section we recall from [3] some prior results for the Gittins index $G(x)$ and the related functions $h(x)$, $H(x)$, $J(x, \Delta)$, and $\Delta^*(x)$.

Lemma 2 *[3, Lemma 1] Function $H(x)$ is strictly increasing [strictly decreasing] at x if and only if $H(x) > h(x)$ [$H(x) < h(x)$] at x. Thus function $H(x)$ has a critical point at x if and only if $H(x) = h(x)$ at x.*

Lemma 3 *[3, Corollary 1] If $\Delta^*(x) > 0$, then $G(y) \geq G(x)$ for all $y \in [x, x + \Delta^*(x))$.*

Lemma 4 *[3, Lemmas 7 and 8] If $G(y) \leq G(x)$ [$G(y) \geq G(x)$] for all $y \in [x, \infty)$, then $G(x) = h(x)$ [$G(x) = H(x)$].*

The following lemma is based on the formula

$$J(x, \Delta) = \frac{\int_x^{x+\Delta} f(y)\,\mathrm{d}y}{\int_x^{x+\Delta} \overline{F}(y)\,\mathrm{d}y} = \frac{\int_x^{x+\Delta} h(y)\overline{F}(y)\,\mathrm{d}y}{\int_x^{x+\Delta} \overline{F}(y)\,\mathrm{d}y}, \tag{10}$$

and the proof is similar to that of [3, Proposition 1].

Lemma 5 *Let $a < b$. If $h(x)$ is strictly decreasing for all $x \in (a, b)$, then $J(x, \Delta)$ is strictly decreasing (with respect to Δ) for all $x \in (a, b)$ and $\Delta \in [0, b - x]$. The lemma remains true if "strictly decreasing" is replaced on both sides by "decreasing", "constant", "increasing", or "strictly increasing".*

The last lemma is easily proved as [3, Proposition 2].

Lemma 6 *Let $a < b$. Now $H(a) < H(b)$ if and only if $J(a, b - a) < J(a, \infty)$. The lemma remains true if "$<$" is replaced on both sides by "$\leq$", "$=$", "$\geq$", or "$>$".*

4 Properties of the Gittins index

In this section we derive several new properties of the Gittins index itself. We start in Section 4.1 with the relationships between the hazard rate $h(x)$ and the Gittins index $G(x)$, while Section 4.2 concerns the relationships between functions $H(x)$ and $G(x)$.

4.1 Relationships between $h(x)$ and $G(x)$

Since the hazard rate curve $h(x)$ is continuous and piecewise monotonic, it consists of a (possibly infinite) number of contiguous intervals where the hazard rate is alternately increasing and strictly decreasing. Next we consider the behaviour of the Gittins index $G(x)$ and the related optimum service quantum $\Delta^*(x)$ within these intervals.

Implications of the monotonicity of the hazard rate

Proposition 1 *Let $a < b \leq \infty$. If $h(x)$ is increasing for all $x \in (a,b)$, then $G(x)$ is increasing and $\Delta^*(x) \geq b - x > 0$ for all $x \in (a,b)$.*

Proof Assume that $h(x)$ is increasing for all $x \in (a,b)$. By Lemma 5, we have, for all $x \in (a,b)$,

$$\Delta^*(x) \geq b - x > 0. \tag{11}$$

Let then $a < x < y < b$. By (11), we have $x < y < b \leq x + \Delta^*(x)$. Thus, by Lemma 3, we conclude that $G(y) \geq G(x)$. □

Proposition 2 *Let $a < b \leq \infty$. If $h(x)$ is strictly decreasing for all $x \in (a,b)$, then there is $c \in [a,b]$ such that*

(i) $G(x)$ is strictly decreasing and $\Delta^*(x) = 0$ for all $x \in (a,c)$, and

(ii) $G(x)$ is increasing and $\Delta^*(x) \geq b - x > 0$ for all $x \in (c,b)$.

Proof Assume that $h(x)$ is strictly decreasing for all $x \in (a,b)$. By Lemma 5, we deduce that $J(x,\Delta)$ is strictly decreasing (with respect to Δ) for all $x \in (a,b)$ and $\Delta \in (0, b-x)$. Thus, for all $x \in (a,b)$,

$$\Delta^*(x) = 0 \quad \text{or} \quad \Delta^*(x) \geq b - x. \tag{12}$$

If $\Delta^*(x) = 0$ for all $x \in (a, b)$, then, by definition, $G(x) = h(x)$ and, thus, $G(x)$ is strictly decreasing for all $x \in (a, b)$. In this case the claims are valid with choice $c = b$.

Assume then that there is $x \in (a, b)$ such that $\Delta^*(x) \geq b - x$. By Lemma 3, $G(y) \geq G(x)$ for all $y \in (x, b)$. In addition, since the hazard rate is strictly decreasing, we have, for all $y \in (x, b)$,

$$G(y) \geq G(x) \geq h(x) > h(y), \tag{13}$$

implying that $\Delta^*(y) > 0$ for all $y \in (x, b)$ by Lemma 1. Thus, by (12), $\Delta^*(y) \geq b - x > 0$ for all $y \in (x, b)$, so from Lemma 3 $G(y)$ is increasing for all $y \in (x, b)$ and we have $c = \inf\{x \in [a, b) \mid \Delta^*(x) \geq b - x\}$. □

Note that if $c = b$ in the previous proposition, then $G(x)$ is strictly decreasing in the whole interval, whereas $c = a$ means that $G(x)$ is increasing in the interval.

Continuity of the Gittins index

It follows from the continuity of $h(x)$ that the Gittins index in this job scheduling context is a continuous function. Our later examples will show that it may not be differentiable, even when $h(x)$ is. On the other hand, whenever $h(x)$ is strictly decreasing for all x, the Gittins index is equal to the hazard rate, $G(x) = h(x)$ for all x, by Lemma 5, the proof of which does not utilize the continuity of $h(x)$. We thus have an example where the Gittins index is not continuous if the hazard rate is not required to be continuous.

Theorem 1 *The Gittins index $G(x)$ is continuous and piecewise monotonic for all $x \geq 0$.*

Proof The proof is presented in Appendix. □

Characterization of the monotonicity of the Gittins index

It follows from the previous theorem that the curve $G(x)$ consists of a (possibly infinite) number of contiguous intervals where $G(x)$ is alternately increasing and strictly decreasing. Furthermore, Propositions 1 and 2 imply that $\Delta^*(x) > [=]\, 0$ in an interval whenever $G(x)$ is increasing [strictly decreasing] in the interval. Thus, we get the following two corollaries.

Corollary 1 *Let $a < b \leq \infty$. $G(x)$ is increasing for all $x \in (a,b)$ if and only if $\Delta^*(x) > 0$ for all $x \in (a,b)$.*

Corollary 2 *Let $a < b \leq \infty$. $G(x)$ is strictly decreasing for all $x \in (a,b)$ if and only if $\Delta^*(x) = 0$ for all $x \in (a,b)$. In this case $G(x) = h(x)$ for all $x \in (a,b)$.*

Also the case that $G(x)$ is decreasing can be characterized.

Proposition 3 *Let $a < b \leq \infty$. $G(x)$ is decreasing for all $x \in (a,b)$ if and only if $G(x) = h(x)$ for all $x \in (a,b)$.*

Proof 1° Assume first that $G(x)$ is decreasing for all $x \in (a,b)$. Let $x \in (a,b)$. If $\Delta^*(x) = 0$, then, by definition, $G(x) = J(x,0) = h(x)$.

Now assume that $\Delta^*(x) > 0$. By Lemma 3 and the assumption made above, we have, for all $y \in (x, x + \Delta^*(x))$,

$$G(y) = G(x). \tag{14}$$

Let $0 < \epsilon < \min\{\Delta^*(x), b - x\}$, and define

$$p = \frac{\int_x^{x+\epsilon} \overline{F}(t)\,\mathrm{d}t}{\int_x^{x+\Delta^*(x)} \overline{F}(t)\,\mathrm{d}t}. \tag{15}$$

Note that $p \in (0,1)$. Now

$$\begin{aligned} G(x) &= J(x, \Delta^*(x)) \\ &= pJ(x,\epsilon) + (1-p)J(x+\epsilon, \Delta^*(x) - \epsilon) \\ &\leq pJ(x,\epsilon) + (1-p)G(x+\epsilon). \end{aligned} \tag{16}$$

On the other hand, $J(x,\epsilon) \leq G(x)$ and, by (14), $G(x+\epsilon) = G(x)$. Thus, we conclude that $J(x,\epsilon) = G(x)$. Since this is true for any $0 < \epsilon < \min\{\Delta^*(x), b - x\}$, we have $h(x) = J(x,0) = \lim_{\epsilon \to 0+} J(x,\epsilon) = G(x)$.

2° Assume now that $G(x) = h(x)$ for all $x \in (a,b)$. By definition, we have, for all $a < x < y < b$,

$$J(x, y-x) \leq h(x). \tag{17}$$

Consider then what happens if $h(x) < h(y)$ for some $a < x < y < b$. Since the hazard rate is continuous, there are $x < c < d < y$ such that $h(t) > h(c)$ for all $t \in (c, d)$. Thus,

$$J(c, d-c) = \frac{\int_c^d h(t)\overline{F}(t)\,\mathrm{d}t}{\int_c^d \overline{F}(t)\,\mathrm{d}t} > h(c), \tag{18}$$

which contradicts (17). Thus, we conclude that $h(x)$ is decreasing for all $x \in (a, b)$. □

By combining the results of Corollary 1 and Proposition 3, we get the following corollary.

Corollary 3 *Let $a < b \leq \infty$. $G(x)$ is constant for all $x \in (a, b)$ if and only if $G(x) = h(x)$ and $\Delta^*(x) > 0$ for all $x \in (a, b)$.*

Consider finally the case that the Gittins index is strictly increasing in a finite interval.

Proposition 4 *Let $a < b \leq \infty$. $G(x)$ is strictly increasing for all $x \in (a, b)$ if and only if $G(x) > h(x)$ for all $x \in (a, b)$.*

Proof 1° Assume first that $G(x)$ is strictly increasing for all $x \in (a, b)$. Corollary 1 implies that $\Delta^*(x) > 0$ for all $x \in (a, b)$. In addition, from Corollary 3 we deduce that there is $x \in (a, b)$ such that $G(x) > h(x)$. Define then $c = \inf\{x > a \mid G(x) > h(x)\}$. If $c > a$, then $G(x) = h(x)$ for all $x \in (a, c)$ implying, by Corollary 3 that $G(x)$ is constant for all $x \in (a, c)$, which contradicts our assumption above. Thus, $c = a$ so that $G(x) > h(x)$ for all $x \in (a, b)$.

2° Assume now that $G(x) > h(x)$ for all $x \in (a, b)$. Corollary 1 implies that $G(x)$ is increasing for all $x \in (a, b)$. If $G(x)$ were constant in a subinterval $(c, d) \in (a, b)$, then, by Corollary 3, we would have $G(x) = h(x)$ for all $x \in (c, d)$, which contradicts our assumption above. Thus, we conclude that $G(x)$ is strictly increasing for all $x \in (a, b)$. □

Characterization of the monotonicity of the Gittins index in infinite intervals

Next we show that even stronger results are available when $b = \infty$.

Proposition 5 *Let $a \geq 0$. The following three statements are equivalent:*

(i) *$h(x)$ is strictly decreasing for all $x > a$.*

(ii) *$G(x)$ is strictly decreasing for all $x > a$.*

(iii) $\Delta^*(x) = 0$ *for all* $x > a$.

In this case $G(x) = h(x)$ *for all* $x > a$.

Proof Note first that the last property follows immediately from (iii). In addition, (ii) and (iii) are equivalent by Corollary 2 with $b = \infty$. Thus it remains to prove the equivalence of (i) and (iii).

1° Assume first that $h(x)$ is strictly decreasing for all $x > a$. By Lemma 5, we deduce that $J(x, \Delta)$ is strictly decreasing (with respect to Δ) for all $x > a$ and $\Delta \geq 0$. Thus, $\Delta^*(x) = 0$ for all $x > a$.

2° Assume now that $\Delta^*(x) = 0$ (so that $G(x) = h(x)$) for all $x > a$. It follows from Corollary 2 with $b = \infty$ that $h(x)$ is strictly decreasing for all $x > a$. □

Proposition 6 *Let* $a \geq 0$. *The following three statements are equivalent:*

(i) $h(x)$ *is decreasing for all* $x > a$.

(ii) $G(x)$ *is decreasing for all* $x > a$.

(iii) $G(x) = h(x)$ *for all* $x > a$.

For $a = 0$, *statement (i) says that the service time distribution belongs to DHR.*

Proof Note first that (ii) and (iii) are equivalent by Proposition 3 with $b = \infty$. Thus it remains to prove the equivalence of (i) and (iii).

1° Assume first that $h(x)$ is decreasing for all $x > a$. By Lemma 5, we deduce that $J(x, \Delta)$ is decreasing (with respect to Δ) for all $x > a$ and $\Delta \geq 0$. Thus, $G(x) = h(x)$ for all $x > a$.

2° Assume now that $G(x) = h(x)$ for all $x > a$. It follows from Proposition 3 with $b = \infty$ that $h(x)$ is decreasing for all $x > a$. □

4.2 Relationships between $H(x)$ and $G(x)$

In addition to the hazard rate, the $H(x)$ function (i.e., the inverse of the mean residual lifetime function) is continuous and piecewise monotonic consisting thus of a (possibly infinite) number of contiguous intervals where $H(x)$ is alternately increasing and strictly decreasing. Next we consider the behaviour of the Gittins index in these intervals.

Implications of the monotonicity of function $H(x)$

Proposition 7 *Let $a < b \leq \infty$. If $H(x)$ is increasing for all $x \in (a, b)$, then $G(x)$ is increasing for all $x \in (a, b)$.*

Proof Assume that $H(x)$ is increasing for all $x \in (a, b)$. By Lemma 6, we have, for all $a < x < y < b$,

$$J(x, y - x) \leq J(x, \infty) = H(x), \tag{19}$$

implying that, for all $x \in (a, b)$,

$$\Delta^*(x) \geq b - x. \tag{20}$$

Let then $a < x < y < b$. By (20), we have $x < y < b \leq x + \Delta^*(x)$. Thus, by Lemma 3, we conclude that $G(y) \geq G(x)$. □

Characterization of the monotonicity of the Gittins index in infinite intervals

Proposition 8 *Let $a \geq 0$. The following three statements are equivalent:*

(i) $H(x)$ is increasing for all $x > a$.

(ii) $G(x)$ is increasing for all $x > a$.

(iii) $G(x) = H(x)$ for all $x > a$.

For $a = 0$, statement (i) says that the service time distribution belongs to DMRL.

Proof The equivalence of (i) and (iii) follows directly from Lemma 6. That (ii) implies (iii) follows from Lemma 4, and (i) and (iii) trivially imply (ii). □

Proposition 9 *Let $a \geq 0$. The following three statements are equivalent:*

(i) $H(x) \geq H(a)$ for all $x > a$.

(ii) $G(x) \geq G(a)$ for all $x > a$.

(iii) $G(a) = H(a)$.

For $a = 0$, statement (i) says that the service time distribution belongs to NBUE.

Proof Similar to the proof of Proposition 8. □

Proposition 10 *Let $a \geq 0$. The following four statements are equivalent:*

(i) $h(x)$ *is constant for all* $x > a$.

(ii) $H(x)$ *is constant for all* $x > a$.

(iii) $G(x)$ *is constant for all* $x > a$.

(iv) $G(x) = H(x) = h(x)$ *for all* $x > a$.

For $a = 0$, statement (i) says that the service time distribution is exponential.

Proof The equivalence of (i) and (ii) follows from Lemma 2. The equivalence of (iii) and (iv) follows from Propositions 6 and 8. In addition, (iii) and (iv) together trivially imply (i) and (ii). So it remains to prove that (ii) implies (iii).

If $H(x)$ is constant (and, thus, increasing) for all $x > a$, then $G(x) = H(x)$ (and, thus, constant) for all $x > a$ by Proposition 8. □

5 Characterization of the Gittins index policy

In this section, we fully characterize the Gittins index policy in the static single-server scheduling problem without arrivals, where the policy is known to minimize the expected flow time (4) among the non-anticipating scheduling disciplines. We will show that the optimal Gittins index policy always belongs to the family of Multi-Level-Processor-Sharing (MLPS) disciplines.

An MLPS discipline π is defined in [6, Sect. 4.7] by a finite set of level thresholds $0 = a_0 < a_1 < \cdots < a_N < a_{N+1} = \infty$ defining $N + 1$ levels, $N \geq 0$. As a slight generalization, we allow an infinite number of levels so that $N \leq \infty$. A job belongs to level n if its age is at least a_{n-1} but less than a_n. Between these levels, a strict priority discipline is applied with the lowest level having the highest priority. Thus, those jobs with age less than a_1 are served first. Within each level n, an internal discipline is applied belonging to $\{\mathrm{FB}, \mathrm{PS}, \mathrm{FCFS}\}$, where, for FCFS, we may have any nonpreemptive discipline. FB refers the Foreground-Background discipline that gives the priority to the youngest job. If there are multiple jobs with the same minimum age, the discipline shares the service capacity evenly among these jobs. We recall from Section 1 that, in our static setting without any arrivals, FB operates, in fact, just like PS.

As already mentioned in the beginning of Section 4.1, the hazard rate curve can be divided into a (possibly infinite) number of contiguous intervals where the hazard rate $h(x)$ is alternately increasing and strictly decreasing. Call them *increasing* and *decreasing intervals* (of $h(x)$), respectively. Let a_n and b_n, respectively, denote the starting point and the ending point of the nth *decreasing interval* of $h(x)$. If $h(x)$ is first increasing, then $b_0 = 0$ and $a_1 > 0$; otherwise $a_1 = 0$. On the other hand, if the number of decreasing intervals is finite (say m) and the last interval is an increasing [decreasing] interval, then $a_n = b_n = \infty$ $[b_{n-1} = a_n = \infty]$ for all $n > m$. For an illustration, see Figure 1.

Proposition 11 *For any n, there is $c_n \in [a_n, b_n]$ such that*

(i) $G(x)$ is strictly decreasing for all $x \in (a_n, c_n)$, and

(ii) $G(x)$ is increasing for all $x \in (c_n, a_{n+1})$.

In addition, $G(x) = h(x)$ for all $x \in (a_n, c_n)$ and any n.

Proof According to Proposition 2, for any decreasing interval $[a_n, b_n)$ of $h(x)$, there is $c_n \in [a_n, b_n]$ such that $G(x)$ is (i) strictly decreasing for all $x \in (a_n, c_n)$, and (ii) increasing for all $x \in (c_n, b_n)$. On the other hand, by Proposition 1, $G(x)$ is increasing for all x belonging to any increasing interval (b_n, a_{n+1}) of $h(x)$. Since $G(x)$ is continuous by Theorem 1, $G(x)$ is increasing for all $x \in (c_n, a_{n+1})$. The last claim that $G(x) = h(x)$ for all $x \in (a_n, c_n)$ follows from Corollary 2. □

Note that if $c_n = a_n$ in the previous proposition, then there is no decreasing part but $G(x)$ is increasing for all $x \in (a_n, a_{n+1})$.

Figure 1, along with Lemma 2, suggests that c_n could be determined from the equality $h(c_n) = H(c_n)$. This is a good approximation and sometimes even exact but, unfortunately, not always, as illustrated by Figure 2, which is zoomed from one of the examples (bottom one) of Figure 1. Figure 2 also demonstrates that even if $h(x)$ is everywhere continuous and differentiable, $G(x)$ may not be everywhere differentiable.

Definition 1 *For any function $g(x)$ defined in $[0, \infty)$, point $c \in (0, \infty)$ is called a* location of a left-strict local minimum *if there is $\epsilon > 0$ such that*

(i) $g(x) > g(c)$ for all $x \in (c - \epsilon, c)$, and

(ii) $g(x) \geq g(c)$ for all $x \in (c, c + \epsilon)$.

Note that c_n defined in Proposition 11, is a location of a left-strict local minimum of $G(x)$ if (and only if) $c_n > a_n$. Similarly a_n is a location of a local maximum of $G(x)$ if (and only if) $c_n > a_n$.

Consider now the characterization of the Gittins index policy in this setting. Let γ_i denote the locations of the *record* local minima of function $G(x)$ defined recursively by letting $\gamma_0 = 0$, and, for $i = 1, 2, \ldots$,

$$\begin{aligned} \gamma_i \quad = \quad & \inf\{c > \gamma_{i-1} \mid G(c) < G(\gamma_{i-1}), \text{ and} \\ & \qquad c \text{ is a location of the left-strict local minima of } G(x)\}. \end{aligned} \tag{21}$$

If the conditions are not satisfied for some γ_i, then we let $\gamma_k = \infty$ for all $k \geq i$. For the characterization, we need still another sequence denoted by α_i, which is defined below separately for the two cases depending on whether $h(x)$ is first increasing or strictly decreasing.

If the hazard rate $h(x)$ is first increasing (so that there is $\delta > 0$ such that $h(x) \geq h(0)$ for all $x \in (0, \delta)$), let $\alpha_1 = \inf\{t > 0 \mid G(t) < G(0)\}$, and, for $k = 1, 2, \ldots$,

$$\begin{aligned} &\alpha_{2k} = \gamma_k, \\ &\alpha_{2k+1} = \inf\{t > \gamma_k \mid G(t) < G(\gamma_k)\}. \end{aligned} \tag{22}$$

Note that for any k there is $j \geq k$ such that $\alpha_{2k} = c_j$, cf. Figures 1 and 3.

But if the hazard rate $h(x)$ is first strictly decreasing (so that there is $\delta > 0$ such that $h(x) < h(0)$ for all $x \in (0, \delta)$), we define, for $k = 1, 2, \ldots$,

$$\begin{aligned} \alpha_{2k-1} &= \gamma_k, \\ \alpha_{2k} &= \inf\{t > \gamma_k \mid G(t) < G(\gamma_k)\}. \end{aligned} \tag{23}$$

In this case, for any k there is $j \geq k$ such that $\alpha_{2k-1} = c_j$.

As before, if the conditions are not satisfied for some α_i, then we let $\alpha_k = \infty$ for all $k \geq i$. In addition, let N denote the highest index i for which α_i is finite,

$$N = \sup\{i = 1, 2, \ldots \mid \alpha_i < \infty\}. \tag{24}$$

For an illustration, see Figure 3, where the three distributions are the same as in Figure 1. These figures illustrate the MLPS nature of the policy, but they also show that even when h, H, and G change direction multiple times, the policy may not change.

Theorem 2

(i) If the hazard rate $h(x)$ is first increasing (so that there is $\delta > 0$ such that $h(x) \geq h(0)$ for all $x \in (0, \delta)$), then the MLPS policy with $N + 1$ levels, thresholds $\alpha_1, \ldots, \alpha_N$ defined in (22), and applying internal disciplines FCFS and FB alternately (starting with FCFS) minimizes the expected flow time (4) among all non-anticipating disciplines Π.

(ii) If the hazard rate $h(x)$ is first strictly decreasing (so that there is $\delta > 0$ such that $h(x) < h(0)$ for all $x \in (0, \delta)$), then the MLPS policy with $N + 1$ levels, thresholds $\alpha_1, \ldots, \alpha_N$ defined in (23), and applying internal disciplines FB and FCFS alternately (starting with FB) minimizes the expected flow time (4) among all non-anticipating disciplines Π.

Proof (i) In the beginning, all jobs by definition have the same age, 0. It follows from the definition (22) that $G(x) \geq G(0)$ for all $x < \alpha_1$. Thus, according to the Gittins index rule, it is optimal to serve jobs one-by-one until they have reached age of α_1. If the service is completed before that, the scheduler starts to serve the next job without any breaks. We see that the proposed MLPS discipline operates like this.

As the result of the first phase, all jobs still in the system have exactly the same age, α_1. Assuming that $\alpha_1 < \infty$, it follows from the definition (22) that $G(x)$ is strictly decreasing for all $\alpha_1 \leq x < \alpha_2$. Thus, according to the Gittins index rule, it is optimal to share the service evenly among all jobs until their common age is α_2. We again see that the proposed MLPS discipline operates like this.

Similar reasoning can be continued since, by the definition (22), $G(x) \geq G(\alpha_{2k})$ for all $\alpha_{2k} \leq x < \alpha_{2k+1}$, and $G(x)$ is strictly decreasing for all $\alpha_{2k+1} \leq x < \alpha_{2k+2}$.

(ii) A similar argument as in (i) can clearly be applied also in this case. □

Note that the internal discipline FCFS may be replaced with any non-preemptive discipline, and the internal discipline FB with PS (in this static setting without arrivals). As special cases, we get the following characterizations.

Corollary 4

(i) If the service time distribution is NBUE, i.e., $H(x) \geq H(0)$ for all $x \geq 0$, then FCFS is optimal.

(ii) If the service time distribution is DHR, i.e., $h(x)$ is decreasing for all $x \geq 0$, then FB is optimal.

(iii) If the service time distribution is NBUE+DHR, i.e., there is $k > 0$ such that $H(x) \geq H(0)$ for all $x \in [0, k)$ and $h(x)$ is decreasing for all $x \in [k, \infty)$, then the two-level MLPS policy FCFS+FB with threshold $\Delta^*(0)$ is optimal.

(iv) If the service time distribution is DHR+IHR, i.e., there is $k > 0$ such that $h(x)$ is decreasing for all $x \in [0, k)$ and increasing for all $x \in [k, \infty)$, then the two-level MLPS policy FB+FCFS with threshold γ_1 is optimal.

6 Conclusions

Independently of each other, Sevcik and Gittins characterized the optimal non-anticipating scheduling policy in a single-server queue with generally distributed service time requirements, which nowadays is known as the Gittins index policy. Remarkably, the Gittins index policy is optimal for both the case without arrivals and the case with Poisson arrivals.

The Gittins index policy assigns an index to each job in the system based on its age and on the service time distribution, and then serves the job with highest index in the system. Under some additional assumptions, the Gittins index policy has a rather simple structure, for instance, if the service time distribution belongs to the NBUE or DHR classes. For the general case, however, it is not straightforward to know a priori how the Gittins index policy operates.

In this paper we provide as complete a characterization as possible for the Gittins index policy for the problem without arrivals. We first show several properties of the Gittins index itself which might be of independent interest. For instance, we show that the Gittins index is a continuous and piecewise monotonic function on the attained service if the hazard rate is such. The study of the index function allows us to derive the main result of the paper, where we show that the Gittins index policy always belongs to the family of Multi-Level-Processor-Sharing (MLPS) disciplines.

In future research, it would be worthwhile to consider the case with arrivals. With arrivals, the optimal policy will be characterized by the same set of thresholds, and within each level the same internal discipline will be used, however, the presence of new arrivals could modify the priorities across levels, that is, it may no longer be the case that the lowest level has the highest priority.

Appendix: Proof of Theorem 1

The claim that $G(x)$ is piecewise monotonic follows immediately from Propositions 1 and 2. Thus, it remains to prove that $G(x)$ is continuous for all x.

1° Consider first an interval (a, b) where the Gittins index $G(x)$ is increasing. Note that Propositions 1 and 2 imply that $\Delta^*(x) \geq b - x > 0$ for all $x \in (a, b)$. In addition, since $G(x)$ is increasing, the limit $G(a^+) = \lim_{x \to a^+} G(x)$ exists. Below we first show that $G(a) \geq G(a^+)$, and then that $G(a) = G(a^+)$, i.e., $G(x)$ is continuous from the right at a.

For any $x \in (a, b)$,

$$G(a) \geq J(a, x + \Delta^*(x) - a) = \frac{\int_a^x f(y)\,\mathrm{d}y + G(x) \int_x^{x+\Delta^*(x)} \overline{F}(y)\,\mathrm{d}y}{\int_a^x \overline{F}(y)\,\mathrm{d}y + \int_x^{x+\Delta^*(x)} \overline{F}(y)\,\mathrm{d}y}. \tag{25}$$

Let $x_n \in (a, b)$ be a decreasing sequence for which $x_n \to a$ as $n \to \infty$. Consequently, $G(x_n) \to G(a^+)$. Since, for any n,

$$0 < \int_{x_1}^{b} \overline{F}(y)\,\mathrm{d}y \leq \int_{x_n}^{x_n+\Delta^*(x_n)} \overline{F}(y)\,\mathrm{d}y \leq \int_0^\infty \overline{F}(y)\,\mathrm{d}y = E[S] < \infty, \tag{26}$$

there is a subsequence $z_k = x_{n_k}$ such that $\int_{z_k}^{z_k+\Delta^*(z_k)} \overline{F}(y)\,\mathrm{d}y$ converges to some finite and positive value c as $k \to \infty$. On the other hand, $\int_a^{z_k} f(y)\,\mathrm{d}y \to 0$ and $\int_a^{z_k} \overline{F}(y)\,\mathrm{d}y \to 0$ as $k \to \infty$. Thus, considering this subsequence z_k, we deduce that

$$G(a) \geq \frac{\int_a^{z_k} f(y)\,\mathrm{d}y + G(z_k) \int_{z_k}^{z_k+\Delta^*(z_k)} \overline{F}(y)\,\mathrm{d}y}{\int_a^{z_k} \overline{F}(y)\,\mathrm{d}y + \int_{z_k}^{z_k+\Delta^*(z_k)} \overline{F}(y)\,\mathrm{d}y} \to G(a^+), \quad \text{as } k \to \infty. \tag{27}$$

Thus, $G(a) \geq G(a^+)$.

If $G(a) > G(a^+)$, then $G(a) > G(a^+) \geq h(a^+) = h(a)$, implying that $\Delta^*(a) > 0$. By Lemma 3, $G(x) \geq G(a)$ for all $x \in (a, a + \Delta^*(a))$ so that $G(a^+) \geq G(a)$, which is a contradiction. Thus, $G(a) = G(a^+)$.

2° Consider still an interval (a, b) where the Gittins index $G(x)$ is increasing. Recall from 1° that $\Delta^*(x) > 0$ for all $x \in (a, b)$. In addition, since $G(x)$ is increasing, the limit $G(b^-) = \lim_{x \to b^-} G(x)$ exists. Next we show that $G(b) = G(b^-)$, i.e., $G(x)$ is continuous from the left at b.

If $G(b) < G(b^-)$, then $G(b^-) > G(b) \geq h(b) = h(b^-)$ by the continuity of the hazard rate $h(x)$. By further taking into account that the Gittins index $G(x)$ is increasing, we conclude that there is $c < b$ such that, for all $x \in (c, b)$,

$$J(x, \Delta^*(x)) - J(x, 0) = G(x) - h(x) \geq \frac{1}{2}(G(b^-) - h(b^-)) > 0. \tag{28}$$

But, now it follows from the continuity (in the plane) of the efficiency function $J(x, \Delta)$ that there is $c' < b$ such that $\Delta^*(x) > b - x$ for all $x \in (c', b)$. By Lemma 3, $G(b) \geq G(x)$ for all $x \in (c', b)$, i.e., $G(b) \geq G(b^-)$, which is a contradiction. Thus, $G(b) \geq G(b^-)$.

If $G(b) > G(b^-)$, then $G(b) > G(b^-) \geq h(b^-) = h(b)$, implying that $\Delta^*(b) > 0$. Now

$$G(x) \geq J(x, b + \Delta^*(b) - x) = \frac{\int_x^b f(y)\,\mathrm{d}y + G(b) \int_b^{b+\Delta^*(b)} \overline{F}(y)\,\mathrm{d}y}{\int_x^b \overline{F}(y)\,\mathrm{d}y + \int_b^{b+\Delta^*(b)} \overline{F}(y)\,\mathrm{d}y} \to G(b), \quad \text{as } x \to b^-, \tag{29}$$

which is a contradiction. Thus, $G(b) = G(b^-)$.

Note that 1° and 2° together prove that the Gittins index $G(x)$ is continuous in any interval (a, b) where it is increasing.

3° Consider now an interval (a, b) where the Gittins index $G(x)$ is strictly decreasing. Then, $G(x) = h(x)$ for all $x \in (a, b)$ so that the continuity of $G(x)$ in this interval follows from the assumed continuity of $h(x)$.

4° It remains to consider the points where the "direction" of the Gittins index function $G(x)$ changes from strictly decreasing to increasing (or vice versa).

Consider first a point $c \in (a, b)$ such that $G(x)$ is strictly decreasing for all $x \in (a, c)$ and increasing for all $x \in (c, b)$. Note that $G(x) = h(x)$ for all $x \in (a, c)$ by Proposition 2, implying that $G(c^-) = h(c^-) = h(c)$. The continuity of $G(x)$ from the right follows from 1°. Next we show that $G(c) = G(c^-) = h(c)$, i.e., $G(x)$ is also continuous from the left at c.

Assume that $G(x)$ is not continuous from the left at c. Thus, $G(c) > h(c)$, and, consequently, $\Delta^*(c) > 0$. In addition, we have, for any $x \in (a, c)$,

$$\begin{aligned} h(x) &= G(x) \\ &\geq J(x, c + \Delta^*(c) - x) \\ &= \frac{\int_x^c f(y)\,\mathrm{d}y + \int_c^{c+\Delta^*(c)} f(y)\,\mathrm{d}y}{\int_x^c \overline{F}(y)\,\mathrm{d}y + \int_c^{c+\Delta^*(c)} \overline{F}(y)\,\mathrm{d}y} \\ &= \frac{\int_x^c f(y)\,\mathrm{d}y + G(c)\int_c^{c+\Delta^*(c)} \overline{F}(y)\,\mathrm{d}y}{\int_x^c \overline{F}(y)\,\mathrm{d}y + \int_c^{c+\Delta^*(c)} \overline{F}(y)\,\mathrm{d}y} \to G(c), \quad \text{as } x \to c^-. \end{aligned} \tag{30}$$

Thus, letting $x \to c^-$, we deduce that $h(c) \geq G(c)$, which is a contradiction.

5° Consider finally a point $c \in (a, b)$ such that $G(x)$ increasing for all $x \in (a, c)$ and strictly decreasing for all $x \in (c, b)$. Note that $G(x) = h(x)$ for all $x \in (c, b)$ by Proposition 2, implying that $G(c^+) = h(c^+) = h(c)$. The continuity of $G(x)$ from the left follows from 2°. Next we show that $G(c) = G(c^+) = h(c)$, i.e., $G(x)$ is also continuous from the right at c.

Assume that $G(x)$ is not continuous from the right at c. Thus, $G(c) > h(c)$, and, consequently, $\Delta^*(c) > 0$. In addition, we have, for any $x \in (c, c + \Delta^*(c))$,

$$h(x) = G(x) \geq J(x, c + \Delta^*(c) - x) = \frac{\int_x^{c+\Delta^*(c)} f(y)\,\mathrm{d}y}{\int_x^{c+\Delta^*(c)} \overline{F}(y)\,\mathrm{d}y} \to G(c), \quad \text{as } x \to c^+. \tag{31}$$

Thus, letting $x \to c^+$, we deduce that $h(c) \geq G(c)$, which is a contradiction. □

Acknowledgements

The anonymous referee is acknowledged for the comments that helped in improving the final version of the paper.

References

[1] Aalto, S. & Ayesta, U. (2008). Optimal scheduling of jobs with a DHR tail in the M/G/1 queue. In *Proceedings of ValueTools 2008*, Athens, Greece.

[2] Aalto, S., Ayesta, U., Borst, S., Misra, V., & Núñez-Queija, R. (2007). Beyond Processor Sharing. *ACM Sigmetrics Performance Evaluation Review* 34(4):36-43.

[3] Aalto, S., Ayesta, U., & Righter, R. (2009). On the Gittins index in the M/G/1 queue. *Queueing Systems* 63:437-458.

[4] Gelenbe, E. & Mitrani, I. (1980). *Analysis and Synthesis of Computer Systems.* Academic Press.

[5] Gittins, J.C. (1989). *Multi-armed Bandit Allocation Indices.* Wiley.

[6] Kleinrock, L. (1976). *Queueing Systems, Volume II: Computer Applications.* Wiley.

[7] Pinedo, M.L. (1995). *Scheduling — Theory, Algorithms, and Systems.* Prentice Hall.

[8] Righter, R. & Shanthikumar, J.G. (1989). Scheduling multiclass single server queueing systems to stochastically maximize the number of successful departures. *Probability in the Engineering and Informational Sciences* 3:323-334.

[9] Righter, R., Shanthikumar, J.G., & Yamazaki, G. (1990). On extremal service disciplines in single-stage queueing systems. *Journal of Applied Probability* 27:409-416.

[10] Sevcik, K.C. (1974). Scheduling for minimum total loss using service time distributions. *Journal of the Association for Computing Machinery* 21:66-75.

[11] Yashkov, S.F. (1987). Processor sharing queues: Some progress in analysis. *Queueing Systems* 2:1-17.

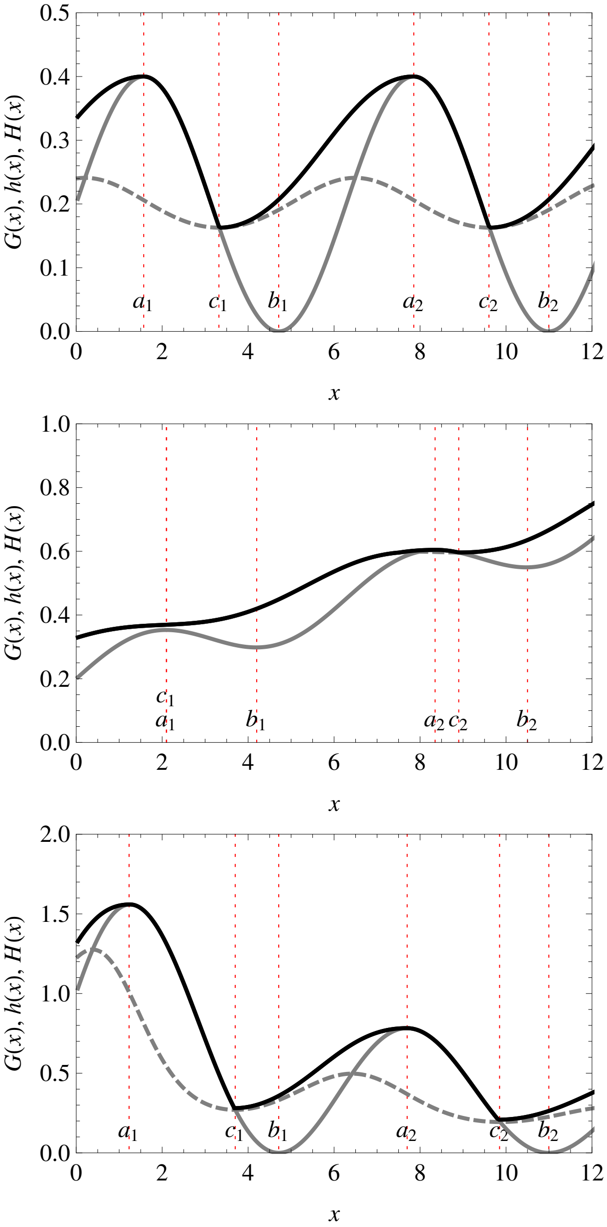

Figure 1: Functions $G(x)$ (solid black), $h(x)$ (solid gray), and $H(x)$ (dashed gray) together with change-over points a_n, c_n and b_n for four service time distributions. *Top*: $h(x) = k(1 + \sin x)$ with $k = 0.2$. *Middle*: $h(x) = k(1 + kx + 2k \sin x)$ with $k = 0.2$. Here $H(x) = G(x)$ except for $x \in (a_2, c_2)$ where they are very close. *Bottom*: $h(x) = (1 + \sin x)/(1 + kx)$ with $k = 0.2$.

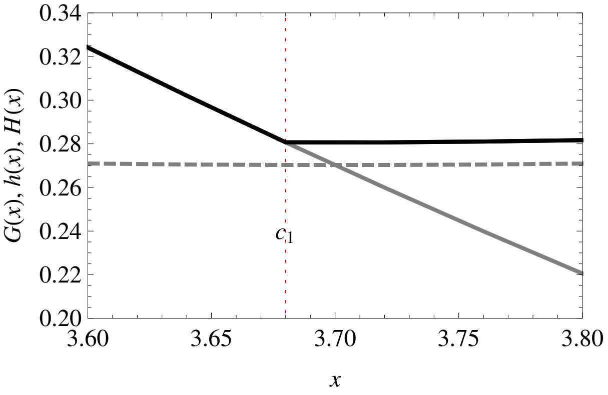

Figure 2: Functions $G(x)$ (solid black), $h(x)$ (solid gray), and $H(x)$ (dashed gray) for the service time distribution $h(x) = (1 + \sin x)/(1 + kx)$ with $k = 0.2$ zoomed from Figure 1 (bottom).

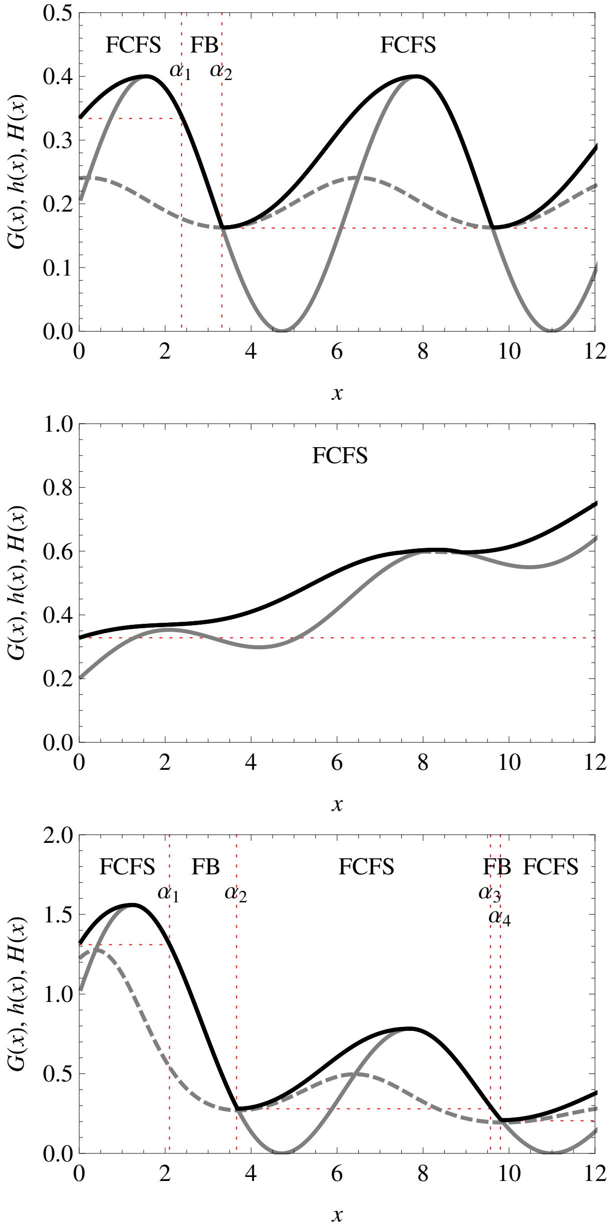

Figure 3: Functions $G(x)$ (solid black), $h(x)$ (solid gray), and $H(x)$ (dashed gray) with thresholds α_n for four service time distributions. *Top*: $h(x) = k(1 + \sin x)$ with $k = 0.2$. *Middle*: $h(x) = k(1 + kx + 2k \sin x)$ with $k = 0.2$. *Bottom*: $h(x) = (1 + \sin x)/(1 + kx)$ with $k = 0.2$.