Probability in the Engineering and Informational Sciences, 9, 1995, 239–254. Printed in the U.S.A.

A GLOBALLY GATED POLLING SYSTEM WITH A DORMANT SERVER

S. C. BORST
CWI, P.O. Box 94079
1090 GB Amsterdam, The Netherlands

We study a globally gated polling system with a dormant server, which makes a halt at its home base when there are no customers present in the system. We derive an explicit expression for the cycle time distribution as well as for the waiting-time distribution at each of the queues. As a justification of the dormant server policy, we show the waiting time at each of the queues to be smaller (in the increasing-convex-ordering sense) than in the ordinary nondormant server case.

1. INTRODUCTION

A polling system basically consists of several queues attended by a single common server. The service discipline prescribes which customers are to be served during a visit to a queue. In other words, it dictates to the server when to move from one queue to another. The server routing dictates to the server from which queue to which queue to move. Moving from one queue to another typically requires a nonzero switch-over time.

Polling systems arise quite naturally in modeling situations in which several users compete for service from a single common server. Thus, polling systems have found a variety of applications in the areas of computer systems, telecommunication networks, manufacturing, and maintenance (cf. Takagi [19,20], Levy and Sidi [15]). Motivated by the variety of applications, numerous studies have been devoted to the analysis of polling systems (cf. Takagi [19,20]).

 0269-9648/95 $11.00 + .10

In the present paper, we study a globally gated polling system with a *dormant* server, which makes a halt at its home base when there are no customers present in the system. In the polling literature, the server is usually assumed never to idle—in other words, to be switching when not working. In particular, the server is assumed to be switching when there are no customers present in the system. As a rare exception, Eisenberg [9] considered a two-queue model with either alternating priority (the exhaustive service discipline at both queues) or strict priority, in which the server remains idling at a queue when there are no customers present in the system. Eisenberg [10] studied a model with an arbitrary number of queues and the exhaustive service discipline at all queues, in which the server does *not* idle. In a recent study [11], Eisenberg showed, however, how an adapted version of the solution method in his earlier work [10] may be used to analyze a model in which the server makes a halt at some of the queues when the system is empty. The outline of the solution method in Eisenberg [11] may also be used to treat a similar model with the gated service discipline.

Gersht and Marbukh [12] considered a two-queue model with two types of disciplines for switching from one queue to another. For both types of disciplines, they showed that for some region of the system parameters the discipline that minimizes the mean waiting cost inserts forced idle periods. Liu, Nain, and Towsley [17] identified polling policies, allowing idling as a possible action, that stochastically minimize the total amount of work in the system at an arbitrary epoch. They found that optimal policies are exhaustive, greedy, and also, in symmetrical systems, patient; i.e., the server should neither switch nor idle when at a nonempty queue, and in symmetrical systems the server should remain idling at a queue when the entire system is empty.

Gupta and Srinivasan [13] derived explicit expressions for the waiting-time distribution in a model similar to that in Eisenberg [11] by using an approach based on the concept of "descendant sets." They showed that while a patient server policy is generally better in the sense of a reduction of the amount of work in the system, cases do exist where a roving server strategy is better. Blanc and Van der Mei [3] used the power-series algorithm to analyze the performance of a system in which the server may be allowed to make a halt at a queue when the entire system is empty. They found that the performance may improve considerably by allowing the server to make a halt at a queue, especially in light traffic. Borst [5] derived a pseudo-conservation law for a general polling model with a dormant server and used it to compare the dormant and the nondormant server case. Furthermore, a heuristic criterium is proposed in Borst [5] for selecting the queues at which the server should make a halt so as to minimize the mean total amount of work in the system.

One reason why in the polling literature the server is usually assumed never to idle may be that the option of idling in general slightly complicates the operation of the system. If at all technically feasible, some mechanism is needed to control the server and to keep track of the customers present in the system. Consequently, the option of idling in general also slightly complicates the mathemat-

ical analysis of the system. Another reason may be that the option of idling will have the biggest impact in light traffic, when the performance will be satisfactory anyhow.

Quite often, however, there are very sound reasons for letting the server stop switching when no customers are present in the system. In many situations some mechanism to control the server and to keep track of the customers present in the system is needed anyhow. The option of idling then arises quite naturally. In manufacturing and maintenance environments, e.g., one usually requires already some kind of supporting system to schedule the jobs. In such situations it makes sense to let the server make a halt at a queue when the entire system is empty rather than to let the server needlessly circle around. One option is then allowing the server to make a halt at all of the queues, i.e., to stop switching as soon as the entire system is empty. Another option is allowing the server only to make a halt at some of the queues (thus, possibly forcing the server to keep switching for a while), e.g., at a queue that functions as home base or at the queue where a new customer is most likely to arrive. The latter option may be recognized in the dynamic control of traffic lights. When there are no vehicles waiting, typically the main stream is given passage, until a waiting vehicle of a crossing stream is detected.

In many situations there are, moreover, significant cost involved in switching. In manufacturing and maintenance applications, the switch-over usually represents the change-over from one type of jobs to another, which may involve labor cost, material cost, or transportation cost. In such situations a potential saving in switching cost is an additional reason for letting the server stop switching when no customers are present in the system.

The option of idling especially arises quite naturally in case of the globally gated service discipline, recently introduced by Boxma, Levy, and Yechiali [6]. The globally gated service discipline operates as follows. Suppose the server arrives at its home base. Then, all the customers in the system are marked instantaneously and the server immediately starts a tour along the queues. During this tour only the marked customers are served. The service of customers that meanwhile arrive to the system is deferred until the next tour along the queues. Boxma, Weststrate, and Yechiali [7] proposed the globally gated service discipline to be used by a repair crew, in charge of the maintenance activities at several installations. As indicated in Boxma et al. [7], under the globally gated service discipline it does not make sense to start a tour along the queues when there are no customers present in the system. In the present paper, we therefore consider a globally gated polling system with a dormant server, which makes a halt at its home base when there are no customers present in the system. The globally gated service discipline then operates as follows. Suppose again the server arrives at its home base. If there are customers present in the system, they are all marked instantaneously and the server starts a tour along the queues, acting as described earlier. If no customers are present in the system, the server remains idling at its home base, awaiting a new customer to arrive at one of the queues. As soon as a new customer arrives, it is marked instantaneously and

the server starts a tour along the queues. During this tour only the newly arrived customer is served. The service of customers that meanwhile arrive to the system is again deferred until the next tour along the queues.

The remainder of this paper is organized as follows. In Section 2 we present a detailed model description. We derive an explicit expression for the Laplace-Stieltjes transform (LST) of the cycle time distribution in Section 3. In Section 4 we obtain the LST of the waiting-time distribution at each of the queues. As a justification of the dormant server policy, we show the waiting time at each of the queues to be smaller (in the increasing-convex-ordering sense) than in the ordinary nondormant server case. In Section 5 we make some concluding remarks.

2. MODEL DESCRIPTION

The model under consideration consists of n queues, $Q_1, \ldots, Q_n$, each of infinite capacity, attended by a single common server S. Customers arrive at the various queues according to independent Poisson processes. Customers arriving at Q_i will also be referred to as type-i customers. Denote by λ_i the arrival rate at Q_i, $i = 1, \ldots, n$. The total arrival rate is $\lambda := \sum_{i=1}^{n} \lambda_i$. Type-$i$ customers require service times $\mathbf{B}_i$, having distribution $B_i(\cdot)$ with LST $\beta_i(\cdot)$, first moment β_i, and second moment $\beta_i^{(2)}$, $i = 1, \ldots, n$. All service times are assumed to be independent. Define the traffic intensity at Q_i as $\rho_i := \lambda_i \beta_i$, $i = 1, \ldots, n$. The total traffic intensity is $\rho := \sum_{i=1}^{n} \rho_i$. Evidently, $\rho < 1$ is a necessary condition for stability. Throughout the paper $\rho < 1$ is assumed to hold.

The server visits the queues in strictly cyclic order, $Q_1, \ldots, Q_n$. Moving from Q_i to Q_{i+1}, where $n + 1$ is to be understood as 1, the server experiences a switch-over time $\mathbf{S}_i$, having distribution $S_i(\cdot)$ with LST $\sigma_i(\cdot)$, first moment s_i, and second moment $s_i^{(2)}$, $i = 1, \ldots, n$. All switch-over times are assumed to be independent. The total switch-over time during a cycle has distribution $S(\cdot)$ with LST $\sigma(\cdot) := \prod_{i=1}^{n} \sigma_i(\cdot)$, first moment $s := \sum_{i=1}^{n} s_i$, and second moment $s^{(2)} := \sum_{i=1}^{n} \sum_{j=1}^{n} s_i s_j + \sum_{i=1}^{n} (s_i^{(2)} - s_i^2)$. The arrival, service, and switch-over processes are assumed to be mutually independent.

The globally gated service discipline operates as follows. Suppose the server is just about to visit Q_1. If there are customers present in the system, they are all marked instantaneously and the server immediately starts visiting $Q_1, \ldots, Q_n$. During the coming cycle only the marked customers are served. At each queue customers are served in order of arrival. The service of customers that meanwhile arrive to the system is deferred until the next cycle. If there are no customers present in the system, the server remains idling at its home base Q_1, awaiting a new customer to arrive at one of the queues. As soon as a new customer arrives, it is marked instantaneously and the server starts visiting $Q_1, \ldots, Q_n$. During the coming cycle, only the newly arrived customer is served. Again, customers that meanwhile arrive to the system are served during the next cycle. During the cycle the server is not allowed to make a halt at a queue when the completion of a service leaves the system empty. In other words, the server is

only allowed to make a halt when the system is empty at the start of a visit to Q_1. In Eisenberg [11] the stopping convention that we adopt here is referred to as the Continue-Cycle-to-Home-Base rule, as opposed to the Jump-Directly-to-Home-Base rule, where the server, on emptying the system, executes a single change-over that takes it directly to its home base. Analogously, the starting convention that we apply here is referred to as the Resume-Cycle rule, as opposed to the Jump-Directly-to-New-Arrival rule, where the server executes a single change-over that takes it directly to the queue receiving the new arrival.

Remark 2.1: For $n = 1$, the model under consideration reduces to a gated vacation model with *single* vacations, whereas the model in the nondormant server case corresponds to a gated vacation model with *multiple* vacations (cf. Takagi [21, pp. 205–213]). The latter model has been analyzed in detail in Takine and Hasegawa [22].

3. THE CYCLE TIME

In this section we relate the cycle time distribution to the joint queue length distribution at the start of the cycle and at the start of the next cycle. Thus, we derive a functional equation for the probability generating function (p.g.f.) of the joint queue length distribution at the start of a cycle and for the LST of the cycle time distribution. By iteratively solving the latter functional equation, we obtain an explicit expression for the LST of the cycle time distribution. The cycle time distribution will play a crucial role in the derivation of the waiting-time distribution at each of the queues in the next section. The approach used here is similar to the approach in Boxma et al. [6] for the ordinary nondormant server case. We first introduce some notation. Denote by $\mathbf{C}^{(m)}$ the length of the mth cycle, i.e., the time from the start of the mth visit to Q_1 to the start of the $(m+1)$th visit to Q_1, $m = 1,2,\ldots$. Denote by $\mathbf{I}^{(m)}$ the length of the mth idling period, i.e., the mth idling time at Q_1 (possibly zero), $m = 1,2,\ldots$. Denote by $\mathbf{B}^{(m)}$ the length of the mth *restricted* cycle, i.e., the mth cycle time minus the mth idling time, $m = 1,2,\ldots$. Denote by $\mathbf{C}$, $\mathbf{I}$, and $\mathbf{B}$ stochastic variables with as distribution the stationary distribution for $m \to \infty$ of $\mathbf{C}^{(m)}$, $\mathbf{I}^{(m)}$, and $\mathbf{B}^{(m)}$, respectively. Let $\alpha(\zeta,\omega) := E(e^{-\zeta\mathbf{I}-\omega\mathbf{B}})$ for $\mathrm{Re}\,\zeta \geq 0$, $\mathrm{Re}\,\omega \geq 0$. Let $\gamma(\omega) := E(e^{-\omega\mathbf{B}})$ for $\mathrm{Re}\,\omega \geq 0$. Denote by $(\mathbf{X}_1,\ldots,\mathbf{X}_n)$ a vector of stochastic variables with as distribution the stationary joint queue length distribution at the start of a cycle. Let $\xi(z) := E(z_1^{\mathbf{X}_1}\ldots z_n^{\mathbf{X}_n})$ for $z = (z_1,\ldots,z_n)$ with $|z_i| \leq 1$, $i = 1,\ldots,n$. Define $\epsilon(z) := \sum_{i=1}^{n} \lambda_i(1 - z_i)$ for $z = (z_1,\ldots,z_n)$ with $|z_i| \leq 1$, $i = 1,\ldots,n$.

Observing that $\xi(0,\ldots,0)$ is the probability that no customers are present at the start of a cycle, we find

$$\alpha(\zeta,\omega) = \sigma(\omega)\left[\xi(\beta_1(\omega),\ldots,\beta_n(\omega)) - \left(1 - \frac{\lambda}{\lambda+\zeta}\sum_{i=1}^{n}\frac{\lambda_i}{\lambda}\beta_i(\omega)\right)\xi(0,\ldots,0)\right] \tag{3.1}$$

and

$$\xi(z) = \gamma(\epsilon(z)). \tag{3.2}$$

For a detailed derivation of Eqs. (3.1) and (3.2), we refer to Borst [4]. From Eqs. (3.1) and (3.2),

$$\alpha(\zeta,\omega) = \sigma(\omega)\left[\gamma\left(\sum_{i=1}^{n}\lambda_i(1-\beta_i(\omega))\right) - \gamma(\lambda)\left(1 - \frac{\lambda}{\lambda+\zeta}\sum_{i=1}^{n}\frac{\lambda_i}{\lambda}\beta_i(\omega)\right)\right] \tag{3.3}$$

and

$$\xi(z) = \sigma(\epsilon(z))\left[\xi(\beta_1(\epsilon(z)),\ldots,\beta_n(\epsilon(z))) - \left(1 - \sum_{i=1}^{n}\frac{\lambda_i}{\lambda}\beta_i(\epsilon(z))\right)\xi(0,\ldots,0)\right]. \tag{3.4}$$

Remark 3.1: The evolution of the joint queue length at the start of a cycle, $(\mathbf{X}_1,\ldots,\mathbf{X}_n)$, in fact constitutes a multi-type branching process with state-dependent immigration (cf. Resing [18]). The crucial observation is that the globally gated service discipline satisfies the following property: If there are k_i customers present at Q_i at the start of a cycle, then during the course of the cycle each of these k_i customers will be "replaced" in an independent and identically distributed manner by a random population having p.g.f. $\beta_i(\epsilon(z))$. Adopting the terminology of the theory of multi-type branching processes, the offspring generating functions are given by $f_i(z) = \beta_i(\epsilon(z))$, $i = 1,\ldots,n$, the immigration generating function for the nonzero states is given by $g(z) = \sigma(\epsilon(z))$, and the immigration generating function for the zero state is given by $g(z)h(z)$ with $h(z) = \sum_{i=1}^{n}(\lambda_i/\lambda)f_i(z)$. From the theory of multi-type branching processes, we have

$$\xi(z) = g(z)\,[\xi(f_1(z),\ldots,f_n(z)) - (1-h(z))\xi(0,\ldots,0)], \tag{3.5}$$

which agrees with Eq. (3.4).

Next we solve functional Eq. (3.3). We first derive some preliminary results from Eq. (3.3). Noting that $E(e^{-\zeta\mathbf{I}}) = \alpha(\zeta,0)$ for $\operatorname{Re}\zeta \geq 0$ and $E(e^{-\omega\mathbf{B}}) = \alpha(0,\omega)$ for $\operatorname{Re}\omega \geq 0$,

$$E\mathbf{C} = \frac{s+\gamma(\lambda)/\lambda}{1-\rho}, \qquad E\mathbf{I} = \frac{\gamma(\lambda)}{\lambda}, \tag{3.6}$$

$$E\mathbf{B} = \frac{s+\rho\gamma(\lambda)/\lambda}{1-\rho}, \qquad E\mathbf{B}^2 = \frac{s^{(2)} + \left(2\rho s + \sum_{i=1}^{n}\lambda_i\beta_i^{(2)}\right)E\mathbf{C}}{1-\rho^2}. \tag{3.7}$$

Remark 3.2: We may obtain $E\mathbf{I}$ also directly by observing that

$$E(\mathbf{I}\,|\,\mathbf{I}>0) = \frac{1}{\lambda}, \tag{3.8}$$

while

$$\Pr\{\mathbf{I} > 0\} = \Pr\{(\mathbf{X}_1, \ldots, \mathbf{X}_n) = (0, \ldots, 0)\} = \int_{t=0}^{\infty} e^{-\lambda t} d \Pr\{\mathbf{B} < t\} = \gamma(\lambda). \tag{3.9}$$

Also, we may obtain $E\mathbf{C}$ directly from $E\mathbf{I}$ by observing that $E\mathbf{C} = (s + E\mathbf{I})/(1 - \rho)$.

We now solve functional Eq. (3.3). Obviously, it suffices to find an expression for $\gamma(\omega)$ for $\operatorname{Re}\omega \geq 0$, as substituting such an expression into Eq. (3.3) yields an expression for $\alpha(\zeta, \omega)$. Define $\delta(\omega) := \sum_{i=1}^{n} \lambda_i(1 - \beta_i(\omega))$ for $\operatorname{Re}\omega \geq 0$. Putting $\zeta = 0$ in Eq. (3.3),

$$\gamma(\omega) = \sigma(\omega)\left[\gamma(\delta(\omega)) - \frac{\gamma(\lambda)}{\lambda}\,\delta(\omega)\right], \qquad \operatorname{Re}\omega \geq 0. \tag{3.10}$$

Define recursively

$$\delta^{(0)}(\omega) = \omega, \qquad \operatorname{Re}\omega \geq 0,$$
$$\delta^{(k)}(\omega) = \delta(\delta^{(k-1)}(\omega)), \qquad \operatorname{Re}\omega \geq 0, k = 1,2,\ldots.$$

Iterating Eq. (3.10)

$$\gamma(\omega) = \prod_{k=0}^{M} \sigma(\delta^{(k)}(\omega))\gamma(\delta^{(M+1)}(\omega)) - \frac{\gamma(\lambda)}{\lambda} \sum_{k=0}^{M} \delta^{(k+1)}(\omega) \prod_{l=0}^{k} \sigma(\delta^{(l)}(\omega)), \tag{3.11}$$

for $\operatorname{Re}\omega \geq 0$, $M = 1,2,\ldots$.

Letting $M \to \infty$ in Eq. (3.11), putting $\omega = \lambda$ to obtain $\gamma(\lambda)$, we find

$$\gamma(\omega) = \prod_{k=0}^{\infty} \sigma(\delta^{(k)}(\omega)) - \frac{\prod_{k=0}^{\infty} \sigma(\delta^{(k)}(\lambda))}{\lambda + \sum_{k=0}^{\infty} \delta^{(k+1)}(\lambda) \prod_{l=0}^{k} \sigma(\delta^{(l)}(\lambda))} \times \sum_{k=0}^{\infty} \delta^{(k+1)}(\omega) \prod_{l=0}^{k} \sigma(\delta^{(l)}(\omega)). \tag{3.12}$$

For a detailed proof of the convergence of Eq. (3.12), we refer to Appendix A of Borst [4]. Substituting Eq. (3.12) into Eq. (3.3) yields an expression for $\alpha(\zeta, \omega)$.

4. THE WAITING TIME

In the previous section, we obtained an explicit expression for the LST of the cycle time distribution. In this section we express the waiting-time distribution at each of the queues in terms of the latter LST.

We first introduce some notation. Denote by $\mathbf{W}_i$ the waiting time of an arbitrary type-i customer, $i = 1, \ldots, n$. Let $w_i(\omega) := E(e^{-\omega \mathbf{W}_i})$ for $\operatorname{Re}\omega \geq 0$,

$i = 1, \ldots, n$. For any nonnegative integer-valued stochastic variable $\mathbf{N}$, denote by $\mathbf{V}_i(\mathbf{N})$ the total service time of $\mathbf{N}$ type-i customers, $i = 1, \ldots, n$, so $E(e^{-\omega \mathbf{V}_i(\mathbf{N})}) = E(\beta_i(\omega)^{\mathbf{N}})$, $\mathrm{Re}\,\omega \geq 0$, $i = 1, \ldots, n$. For any nonnegative real-valued stochastic variable $\mathbf{T}$, denote by $\mathbf{A}_i(\mathbf{T})$ the number of type-i customers arriving during a period of length $\mathbf{T}$, $i = 1, \ldots, n$, so $E(y^{\mathbf{A}_i(\mathbf{T})}) = E(e^{-\lambda_i(1-y)\mathbf{T}})$, $|y| \leq 1$, $i = 1, \ldots, n$. Denote by $\mathbf{B}_i$ and $\mathbf{S}_i$ stochastic variables having distribution $B_i(\cdot)$ and $S_i(\cdot)$, respectively.

We now analyze the distribution of the waiting time of an arbitrary type-i customer, by distinguishing whether the customer arrives during a restricted cycle or during an idling period (thus terminating the idling period immediately by initiating a new restricted cycle), in other words, whether the customer sees the server working/switching or idling upon arrival. The waiting time $\mathbf{W}_i^{(\mathrm{B})}$ of an arbitrary type-i customer that arrives during a restricted cycle is composed of the following:

i. the residual lifetime $\mathbf{B}_R$ of the restricted cycle in which it arrives;

ii. the total service time of all customers arriving at $Q_1, \ldots, Q_{i-1}$ during the same restricted cycle;

iii. the total service time of all customers arriving at Q_i during the past lifetime $\mathbf{B}_P$ of the restricted cycle in which it arrives; and

iv. the total switch-over time experienced by the server when moving from Q_1 to Q_i; i.e.,

$$\mathbf{W}_i^{(\mathrm{B})} \overset{d}{=} \mathbf{B}_R + \sum_{j=1}^{i-1} \mathbf{V}_j(\mathbf{A}_j(\mathbf{B}_P + \mathbf{B}_R)) + \mathbf{V}_i(\mathbf{A}_i(\mathbf{B}_P)) + \sum_{j=1}^{i-1} \mathbf{S}_j, \qquad i = 1, \ldots, n. \tag{4.1}$$

From Cohen [8, p. 113],

$$E(e^{-\omega_P \mathbf{B}_P - \omega_R \mathbf{B}_R}) = \frac{1}{E\mathbf{B}} \frac{\gamma(\omega_R) - \gamma(\omega_P)}{\omega_P - \omega_R}, \qquad \mathrm{Re}\,\omega_P \geq 0,\ \mathrm{Re}\,\omega_R \geq 0.$$

So

$$\begin{aligned} E(e^{-\omega \mathbf{W}_i^{(\mathrm{B})}}) &= \prod_{j=1}^{i-1} \sigma_j(\omega) \\ &\quad \times \int_{t_P=0}^{\infty} \int_{t_R=0}^{\infty} e^{-\omega t_R} \prod_{j=1}^{i-1} \{e^{-\lambda_j(1-\beta_j(\omega))(t_P+t_R)}\} e^{-\lambda_i(1-\beta_i(\omega))t_P} \\ &\quad \times d_{t_P, t_R} \Pr\{\mathbf{B}_P < t_P, \mathbf{B}_R < t_R\} \\ &= \prod_{j=1}^{i-1} \sigma_j(\omega) \frac{1}{E\mathbf{B}} \frac{\gamma\left(\sum_{j=1}^{i} \lambda_j(1-\beta_j(\omega))\right) - \gamma\left(\sum_{j=1}^{i-1} \lambda_j(1-\beta_j(\omega)) + \omega\right)}{\omega - \lambda_i(1-\beta_i(\omega))}, \\ &\qquad\qquad i = 1, \ldots, n. \end{aligned} \tag{4.2}$$

The waiting time $\mathbf{W}_i^{(1)}$ of an arbitrary type-i customer that arrives during an idling period is composed solely of the total switch-over time experienced by the server when moving from Q_1 to Q_i; i.e.,

$$\mathbf{W}_i^{(1)} \overset{d}{=} \sum_{j=1}^{i-1} \mathbf{S}_j, \qquad i = 1, \ldots, n. \tag{4.3}$$

So

$$E(e^{-\omega \mathbf{W}_i^{(1)}}) = \prod_{j=1}^{i-1} \sigma_j(\omega), \qquad i = 1, \ldots, n. \tag{4.4}$$

From Eqs. (4.2) and (4.4), observing that an arbitrary customer, irrespective of which type, arrives during a restricted cycle and an idling period with probability $E\mathbf{B}/E\mathbf{C}$ and $E\mathbf{I}/E\mathbf{C}$, respectively,

$$w_i(\omega) = \prod_{j=1}^{i-1} \sigma_j(\omega) \frac{1}{E\mathbf{C}} \times \left[E\mathbf{I} + \frac{\gamma\left(\sum_{j=1}^{i} \lambda_j (1 - \beta_j(\omega)) \right) - \gamma\left(\sum_{j=1}^{i-1} \lambda_j (1 - \beta_j(\omega)) + \omega \right)}{\omega - \lambda_i (1 - \beta_i(\omega))} \right], \tag{4.5}$$

for $\operatorname{Re} \omega \geq 0$, $i = 1, \ldots, n$.

Remark 4.1: Denote by $\mathbf{L}_i$ the queue length at Q_i at an arbitrary epoch, i.e., the number of waiting customers, excluding the customer possibly in service, $i = 1, \ldots, n$. The distribution of $\mathbf{L}_i$ immediately follows from Eq. (4.5) by the distributional form of Little's law: $E(y^{\mathbf{L}_i}) = w_i(\lambda_i(1 - y))$, $|y| \leq 1$, $i = 1, \ldots, n$ (cf. Keilson and Servi [14]).

Remark 4.2: For $n = 1$, using Eqs. (3.6) and (3.10), Eq. (4.5) reduces to

$$w(\omega) = \frac{(1 - \rho)\omega}{\omega - \lambda(1 - \beta(\omega))} \left[\frac{E\mathbf{I}}{s + E\mathbf{I}} + \frac{s}{s + E\mathbf{I}} \frac{1 - \sigma(\omega)}{s\omega} \frac{\gamma(\omega)}{\sigma(\omega)} \right], \tag{4.6}$$

exhibiting the well-known waiting-time decomposition property of $M/G/1$ vacation models.

From Eq. (4.5), using Eqs. (3.6) and (3.7),

$$\begin{aligned} E\mathbf{W}_i &= \left[1 + 2 \sum_{j=1}^{i-1} \rho_j + \rho_i \right] \frac{E\mathbf{B}^2}{2E\mathbf{C}} + \sum_{j=1}^{i-1} s_j \\ &= \left[1 + 2 \sum_{j=1}^{i-1} \rho_j + \rho_i \right] \frac{1}{1 + \rho} \left[\frac{\sum_{j=1}^{n} \lambda_j \beta_j^{(2)}}{2(1 - \rho)} + \frac{\rho s}{1 - \rho} + \frac{s^{(2)}}{2\left(s + \frac{\gamma(\lambda)}{\lambda} \right)} \right] + \sum_{j=1}^{i-1} s_j. \end{aligned} \tag{4.7}$$

Remark 4.3: For $n = 1$, Eq. (4.7) reduces to

$$E\mathbf{W} = [1 + \rho] \frac{E\mathbf{B}^2}{2E\mathbf{C}}$$

$$= \frac{\lambda\beta^{(2)}}{2(1-\rho)} + \frac{\rho s}{1-\rho} + \frac{s^{(2)}}{2\left(s + \frac{\gamma(\lambda)}{\lambda}\right)}, \tag{4.8}$$

which agrees with Takagi [21, p. 213, Eq. (5.40b)].

As a justification of the dormant server policy, we now show the waiting time at each of the queues to be smaller (in the increasing-convex-ordering sense) than in the ordinary nondormant server case. To do so, let us label the variables corresponding to the dormant and the nondormant server case with a circumflex and a tilde, respectively.

From Boxma et al. [6],

$$E\tilde{\mathbf{W}}_i = \left[1 + 2\sum_{j=1}^{i-1} \rho_j + \rho_i\right] \frac{E\tilde{\mathbf{C}}^2}{2E\tilde{\mathbf{C}}} + \sum_{j=1}^{i-1} s_j$$

$$= \left[1 + 2\sum_{j=1}^{i-1} \rho_j + \rho_i\right] \frac{1}{1+\rho} \left[\frac{\sum_{j=1}^{n} \lambda_j \beta_j^{(2)}}{2(1-\rho)} + \frac{\rho s}{1-\rho} + \frac{s^{(2)}}{2s}\right] + \sum_{j=1}^{i-1} s_j. \tag{4.9}$$

Subtracting Eq. (4.9) from Eq. (4.7),

$$E\hat{\mathbf{W}}_i - E\tilde{\mathbf{W}}_i = -\left[1 + 2\sum_{j=1}^{i-1} \rho_j + \rho_i\right] \frac{1}{1+\rho} \frac{\frac{\gamma(\lambda)}{\lambda}}{s + \frac{\gamma(\lambda)}{\lambda}} \frac{s^{(2)}}{2s} \leq 0. \tag{4.10}$$

Proceeding by differentiating the LST of the waiting-time distribution not just once but several times, we may prove that in fact not only the mean waiting times are smaller, but also each of the higher order moments; i.e., $E(\hat{\mathbf{W}}_i^k) \leq E(\tilde{\mathbf{W}}_i^k)$ for any $k \geq 1$. By using coupling techniques, we may however prove that the waiting times are in fact even smaller in the increasing-convex-ordering sense; i.e., $Ef(\hat{\mathbf{W}}_i) \leq Ef(\tilde{\mathbf{W}}_i)$ for any nondecreasing convex function $f(\cdot)$.

LEMMA 4.1:

$$\hat{\mathbf{W}}_i \leq_{\text{icx}} \tilde{\mathbf{W}}_i, \qquad i = 1, \ldots, n;$$

i.e., $Ef(\hat{\mathbf{W}}_i) \leq Ef(\tilde{\mathbf{W}}_i)$, $i = 1, \ldots, n$, *for any nondecreasing convex function* $f(\cdot)$.

PROOF: See the Appendix.

The ordering relation stated in the preceding lemma adds to the modest collection of stochastic ordering results that are known for polling systems so far.

The scarce results that are known exclusively refer to stochastic monotonicity properties of *global* performance measures, like the total amount of work in the system or the cycle time, or refer to monotonicity of quantities like the joint queue length at polling epochs with regard to the parameters of the service discipline or with regard to the underlying stochastic processes. Levy, Sidi, and Boxma [16] showed that the total amount of work in the system is decreasing in the degree of exhaustiveness of the service discipline. Altman, Konstantopoulos, and Liu [2] proved that the cycle time and the joint queue length at polling epochs are stochastically increasing in the arrival rates, service times, and switch-over times. To the best of the author's knowledge, there are, however, no ordering results known at all for the individual waiting times of the nature of the ordering relation stated in the preceding lemma. One might be inclined to conjecture that also the individual waiting times are stochastically decreasing in the degree of exhaustiveness of the service discipline or increasing in the arrival rates, service times, and switch-over times, but such statements have either been disproved by simple counterexamples (cf. Sarkar and Zangwill [1] for example) or have lacked proof so far.

5. CONCLUDING REMARKS

We studied a globally gated polling system with a dormant server, which makes a halt at its home base when no customers are present in the system. As a justification of the dormant server policy, we showed the waiting time at each of the queues to be smaller (in the increasing-convex-ordering sense) than in the ordinary nondormant server case.

In the present paper, we allowed the server only to make a halt at its home base and only when there are no customers present in the system. In fact, we may also allow the server to make a halt at other queues and in other cases when there are still a few customers present in the system. A first option might be to maintain the service disciplines at the various queues but to decide at the completion of *each visit* whether to switch or to idle, and not only at the completion of a visit that leaves the entire system empty. A second option might be also to drop the service disciplines at the various queues and to decide at the completion of *each service* whether to serve another customer if present, to switch, or to idle, like in Liu et al. [17]. Once having enlarged the freedom of decisions in the operation of the system, it is quite natural to consider the problem of finding a strategy that optimizes the performance of the system. As the enlarged freedom of decisions will considerably complicate the analysis, the chances of exactly solving the problem appear negligible.

Acknowledgments

The author is indebted to O. J. Boxma for several valuable discussions, to M. Eisenberg for giving him in an early stage access to his notes that recently resulted in the work by Eisenberg [11], and to G. M. Koole and J. A. C. Resing for some useful suggestions concerning the proof of Lemma 4.1. Furthermore, the author is grateful to the two anonymous referees for their helpful comments.

References

1. Sarkar, D. & Zangwill, W.I. (1991). Variance effects in cyclic production systems. *Management Science* 37: 444–453.
2. Altman, E., Konstantopoulos, P., & Liu, Z. (1992). Stability, monotonicity and invariant quantities in general polling systems. *Queueing Systems* 11(Special Issue on Polling Models): 35–57.
3. Blanc, J.P.C. & Van der Mei, R.D. (1994). The power-series algorithm applied to polling systems with a dormant server. In J. Labetoulle & J.W. Roberts (eds.), *The fundamental role of teletraffic in the evolution of telecommunications networks, Proceedings of ITC 14*. Amsterdam: North-Holland, pp. 865–874.
4. Borst, S.C. (1993). A polling system with a dormant server. Report BS-R9313. CWI, Amsterdam.
5. Borst, S.C. (1994). A pseudo-conservation law for a polling system with a dormant server. In J. Labetoulle & J.W. Roberts (eds.), *The fundamental role of teletraffic in the evolution of telecommunications networks, Proceedings of ITC 14*. Amsterdam: North-Holland, pp. 729–742.
6. Boxma, O.J., Levy, H., & Yechiali, U. (1992). Cyclic reservation schemes for efficient operation of multiple-queue single-server systems. *Annals of Operations Research* 35: 187–208.
7. Boxma, O.J., Weststrate, J.A., & Yechiali, U. (1993). A globally gated polling system with server interruptions, and applications to the repairman problem. *Probability in the Engineering and Informational Sciences* 7: 187–208.
8. Cohen, J.W. (1982). *The single server queue*, 2nd ed. Amsterdam: North-Holland.
9. Eisenberg, M. (1971). Two queues with changeover times. *Operations Research* 19: 386–401.
10. Eisenberg, M. (1972). Queues with periodic service and changeover times. *Operations Research* 20: 440–451.
11. Eisenberg, M. (1994). The polling system with a stopping server. *Queueing Systems* 18: 387–431.
12. Gersht, A.M. & Marbukh, V.V. (1975). Queueing systems with readjustment. *Engineering Cybernetics* 13: 55–65.
13. Srinivasan, M.M. & Gupta, D. (1993). When should a roving server be patient? Report, University of Tennessee, Knoxville.
14. Keilson, J. & Servi, L.D. (1990). The distributional form of Little's law and the Fuhrmann-Cooper decomposition. *Operations Research Letters* 9: 239–247.
15. Levy, H. & Sidi, M. (1990). Polling systems: Applications, modelling and optimization. *IEEE Transactions on Communications* 38: 1750–1760.
16. Levy, H., Sidi, M., & Boxma, O.J. (1990). Dominance relations in polling systems. *Queueing Systems* 6: 155–171.
17. Liu, Z., Nain, P., & Towsley, D. (1992). On optimal polling policies. *Queueing Systems* 11(Special Issue on Polling Models): 59–83.
18. Resing, J.A.C. (1993). Polling systems and multitype branching processes. *Queueing Systems* 13: 409–426.
19. Takagi, H. (1986). *Analysis of polling systems*. Cambridge, MA: The MIT Press.
20. Takagi, H. (1990). Queueing analysis of polling models: An update. In H. Takagi (ed.), *Stochastic analysis of computer and communication systems*. Amsterdam: North-Holland, pp. 267–318.
21. Takagi, H. (1991). *Queueing analysis*, Vol. 1. Amsterdam: North-Holland.
22. Takine, T. & Hasegawa, T. (1992). On the *M/G/*1 queue with multiple vacations and gated service discipline. *Journal of the Operations Research Society of Japan* 35: 217–235.

APPENDIX: PROOF OF LEMMA 4.1

LEMMA 4.1:

$$\hat{\mathbf{W}}_i \leq_{\text{icx}} \bar{\mathbf{W}}_i, \qquad i = 1, \ldots, n;$$

i.e., $Ef(\hat{\mathbf{W}}_i) \leq Ef(\bar{\mathbf{W}}_i)$, $i = 1, \ldots, n$, *for any nondecreasing convex function* $f(\cdot)$.

PROOF: We sketch the intuitive idea of the proof. For a detailed technical proof, we refer to Appendix B of Borst [4]. We assume the arrival, service, and switch-over processes in the dormant and nondormant server case to be coupled as follows. In both cases the server experiences exactly the same switch-over times, but—because the dormant and nondormant server cases evolve according to different operational rules—the same switch-over time is not necessarily experienced at the same point in time; i.e., the switch-over times may be shifted in time. Moreover, in the dormant server case, when the server is actually idling, we assume that the server is experiencing a switch-over time, which is, however, immediately interrupted as soon as a new customer arrives, just as if the server would have been idling, awaiting a new customer to arrive. The remainder of the switch-over time is then resumed as soon as the server starts idling again. During one and the same switch-over time, the arrival processes in both cases proceed synchronously; i.e., the same customer arrives at the same relative time (with regard to the switch-over time in question), requiring the same service time. Thus, the server also provides exactly the same service times in both cases, but—because the dormant and nondormant server cases evolve according to different operational rules—the same service time is not necessarily provided at the same point in time. Also, during one and the same service time, the arrival processes in both cases proceed synchronously. So the arrivals in both cases may be shifted in time, however, congruently to the service or switch-over times in which they fall, so that the same customer arrives at the same relative time with regard to the service time or switch-over time in question. By the memoryless property of the Poisson process, the coupling does not affect the stochastic properties of the arrival process. Neither does the coupling affect the stochastic properties of the service and switch-over processes. Thus, we obtain coupled but still marginally unbiased induced stochastic processes (like waiting times and queue lengths) in the dormant and nondormant server cases.

Suppose that at time $t = \mathbf{T}_0$ in both cases the system is empty and the server is at its home base Q_1, just back from switching. The server then starts switching for a time of length $\mathbf{S}_0$. $\mathbf{S}_0, \mathbf{S}_1, \mathbf{S}_2, \ldots$, are independent stochastic variables with common distribution $S(\cdot)$. During the switch-over time $\mathbf{S}_0$, a number of $\mathbf{K}$ customers arrive, let us say $C_1, \ldots, C_{\mathbf{K}}$, at (relative) time $t = \mathbf{A}_1$, $t = \mathbf{A}_1 + \mathbf{A}_2, \ldots, t = \mathbf{A}_1 + \cdots + \mathbf{A}_{\mathbf{K}}$, requiring service times of length $\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_{\mathbf{K}}$. $\mathbf{A}_1, \mathbf{A}_2, \ldots$, are independent exponentially distributed stochastic variables with mean $1/\lambda$. $\mathbf{B}_1, \mathbf{B}_2, \ldots$, are independent stochastic variables with common distribution $\sum_{i=1}^{n} (\lambda_i/\lambda) B_i(\cdot)$. In the dormant server case, at time $t = \mathbf{T}_0 + \mathbf{A}_1$ the server interrupts switching, suspending the remainder $\mathbf{S}_0 - \mathbf{A}_1$ of the switch-over time $\mathbf{S}_0$, and starts a cycle along the queues to serve the newly arrived customer C_1, just as if the server would have remained idling at Q_1 from time $t = \mathbf{T}_0$ on, awaiting the new customer to arrive. At time $t = \mathbf{T}_1$, after $\mathbf{L}_1$ cycles, the system is empty again and the server is back again at its home base Q_1 (these events occurring simulta-

neously for the first time). The server then starts switching, resuming the switch-over time $\mathbf{S}_0$.

At time $t = \mathbf{T}_1 + \mathbf{A}_2$, the server again interrupts switching, suspending the remainder $\mathbf{S}_0 - \mathbf{A}_1 - \mathbf{A}_2$ of the switch-over time $\mathbf{S}_0$, and starts a cycle along the queues to serve the newly arrived customer C_2, again just as if the server would have remained idling at Q_1 from time $t = \mathbf{T}_1$ on, awaiting the new customer to arrive.

In the nondormant server case, at time $t = \mathbf{T}_0 + \mathbf{A}_1$ the server just continues switching, disregarding the newly arrived customer C_1. At time $t = \mathbf{T}_0 + \mathbf{S}_0$, the server finishes switching and starts a cycle along the queues to serve the newly arrived customers $C_1, \ldots, C_{\mathbf{K}}$.

In the dormant server case, at time $t = \mathbf{T}_{\mathbf{K}}$, after $\mathbf{L}_1 + \cdots + \mathbf{L}_{\mathbf{K}}$ cycles, the system is empty and the server is back at its home base Q_1 (these events occurring simultaneously for the **K**th time). The server then starts switching, resuming the switch-over time $\mathbf{S}_0$. At time $t = \mathbf{U}_0$ the server finishes switching, $\mathbf{U}_0 = \mathbf{T}_{\mathbf{K}} + \mathbf{D}$, $\mathbf{D} = \mathbf{S}_0 - \mathbf{A}_1 - \cdots - \mathbf{A}_{\mathbf{K}}$. At time $t = \mathbf{U}_0$ also in the nondormant server case the system is empty and the server just finishes switching. In both cases the server has then experienced exactly the same switch-over times, viz., $\mathbf{S}_0, \mathbf{S}_1, \ldots, \mathbf{S}_{\mathbf{L}_1 + \cdots + \mathbf{L}_{\mathbf{K}}}$, and has provided exactly the same service times, viz., the service times of the customers arriving during $\mathbf{S}_0, \mathbf{S}_1, \ldots, \mathbf{S}_{\mathbf{L}_1 + \cdots + \mathbf{L}_{\mathbf{K}}}$, and of their descendants. Let us say the total number of type-i customers among them is $\mathbf{M}_i$. (Here the descendants of a customer are recursively defined as the customers arriving during its service time or during the service time of one of its descendants.) Concluding, at time $t = \mathbf{U}_0$ in *both* cases the system is empty and the server is at its home base Q_1, just back from switching.

Let $R_i^{(h)}$ be the hth type-i customer served from time $t = \mathbf{T}_0$ on in the dormant server case, $h = 1,2,\ldots$. Denote by $\hat{\mathbf{W}}_i^{(h)}$ and $\bar{\mathbf{W}}_i^{(h)}$ the waiting time of $R_i^{(h)}$ in the dormant and nondormant server case, respectively, $h = 1,2,\ldots$. As the stochastic processes $\{\hat{\mathbf{W}}_i^{(h)}, h = 1,2,\ldots\}$ and $\{\bar{\mathbf{W}}_i^{(h)}, h = 1,2,\ldots\}$ are regenerative with regard to $h = 1$ and $h = \mathbf{M}_i + 1$,

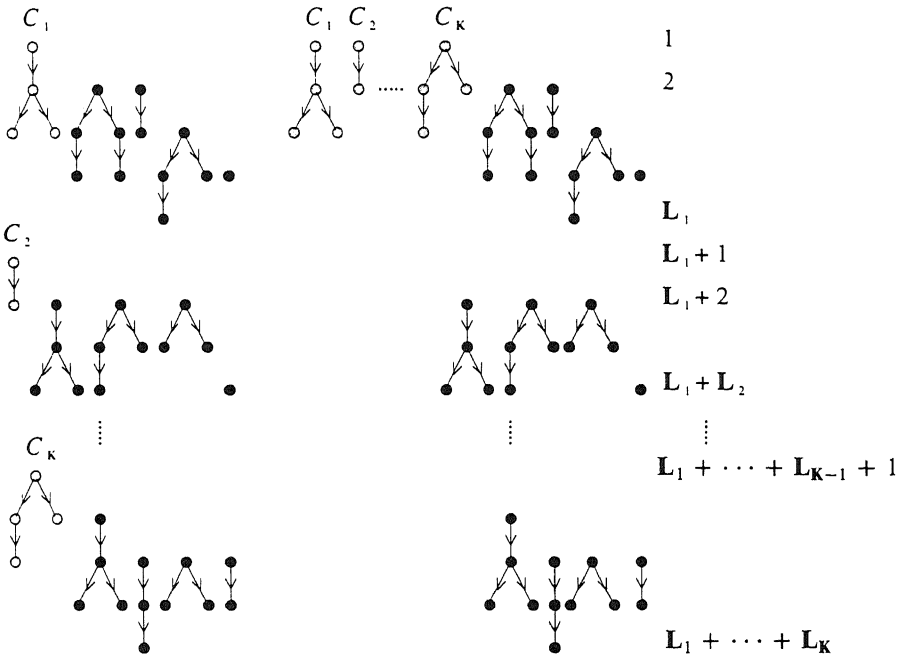
$$Ef(\hat{\mathbf{W}}_i) = \frac{1}{E\mathbf{M}_i} E\left(\sum_{h=1}^{\mathbf{M}_i} f(\hat{\mathbf{W}}_i^{(h)})\right) \tag{A.1}$$

and

$$Ef(\bar{\mathbf{W}}_i) = \frac{1}{E\mathbf{M}_i} E\left(\sum_{h=1}^{\mathbf{M}_i} f(\bar{\mathbf{W}}_i^{(h)})\right). \tag{A.2}$$

Consider now Figure 1, representing the customer offspring process in the dormant and nondormant server cases. In both cases dots at the same (horizontal) level correspond to customers served in the same cycle. An arc indicates that the customer at the head arrives during the service time of the customer at the tail. A dot without any incoming arc represents a customer that arrives during a switch-over time or, in the dormant server case, during an idling period. A dot without any outgoing arc represents a customer requiring a service time during which no single customer arrives.

To prove that Eq. (A.2) majorizes Eq. (A.1), we need to introduce some additional terminology. In the dormant server case, the interval from time $t = \mathbf{T}_{k-1}$ to time $t = \mathbf{T}_k$, comprising the cycles $\mathbf{L}_1 + \cdots + \mathbf{L}_{k-1} + 1$ to $\mathbf{L}_1 + \cdots + \mathbf{L}_k$, is referred to as the kth *busy interval*, $k = 1, \ldots, \mathbf{K}$. Customers arriving during $\mathbf{S}_0$ (thus interrupting $\mathbf{S}_0$ in the dormant server case), together with their descendants, are called *primary* customers (cor-



The dormant server case. The nondormant server case.

FIGURE 1. The customer offspring process.

responding to the grey dots in Figure 1). The remaining customers, i.e., customers arriving during $\mathbf{S}_1,\ldots,\mathbf{S}_{\mathbf{L}_K}$, together with their descendants, are called *secondary* customers (the black dots).

With the busy intervals as background, the dormant and nondormant server cases differ in the service of primary customers, but not in the service of secondary customers. In the dormant server case, the service of primary customers is balanced over the $\mathbf{K}$ busy intervals. In the nondormant server case, the service of primary customers is concentrated in the first busy interval. The service of secondary customers occurs in both cases in corresponding cycles. So in the nondormant server case, the primary customers all bother one another and all bother the same secondary customers. To make the latter intuitive idea precise, observe that the waiting time of every customer is composed of switch-over time, service time of primary customers, and service time of secondary customers. For any waiting time $\mathbf{W}_i^{(h)}$, denote by $\mathbf{V}_{i,k}^{(h)}$ the share constituted by service time of primary customers that are descendants of $C_k, k = 1,\ldots,\mathbf{K}$. Denote by $\mathbf{V}_i^{(h)}$ the remaining share in $\mathbf{W}_i^{(h)}$, i.e., the share constituted by switch-over time and service time of secondary customers.

For a *primary* customer $R_i^{(h)}$, we have

$$\hat{\mathbf{V}}_i^{(h)} + \mathbf{Z}_i^{(h)} \overset{d}{=} \tilde{\mathbf{V}}_i^{(h)}, \qquad \mathbf{Z}_i^{(h)} \geq 0. \tag{A.3}$$

Assuming that $R_i^{(h)}$ is a descendant of C_k,

$$\hat{\mathbf{V}}_{i,k}^{(h)} = \tilde{\mathbf{V}}_{i,k}^{(h)}, \qquad \hat{\mathbf{V}}_{i,m}^{(h)} = 0, \qquad m \neq k. \tag{A.4}$$

For a *secondary* customer $R_i^{(h)}$, we have

$$\hat{\mathbf{V}}_i^{(h)} = \tilde{\mathbf{V}}_i^{(h)}. \tag{A.5}$$

Let $H_{i,l}$ be the index set of the secondary type-i customers served in the lth cycle in the dormant server case, $l = 1, \ldots, \mathbf{L}_1 + \cdots + \mathbf{L}_{\mathbf{K}}$. Let $H_{i,kl}$ be the index set of the secondary type-i customers served in the $\mathbf{L}_1 + \cdots + \mathbf{L}_{k-1} + l$th cycle in the dormant server case if $l \leq \mathbf{L}_k$, $k = 1, \ldots, \mathbf{K}$. If $l > \mathbf{L}_k$, let $H_{i,kl}$ be the empty set. We then have

$$\sum_{h \in H_{i,kl}} \hat{\mathbf{V}}_{i,k}^{(h)} \overset{d}{=} \sum_{h \in H_{i,l}} \tilde{\mathbf{V}}_{i,k}^{(h)}, \qquad \sum_{h \in H_{i,kl}} \hat{\mathbf{V}}_{i,m}^{(h)} = 0, \qquad m \neq k. \tag{A.6}$$

Relations (A.3)-(A.6) constitute the key elements in proving that Eq. (A.2) majorizes Eq. (A.1). For a detailed comparison, we refer to Appendix B of Borst [4]. ■