Queueing Syst (2012) 71:247–292
DOI 10.1007/s11134-012-9311-0

Asymptotic independence of queues under randomized load balancing

Maury Bramson · Yi Lu · Balaji Prabhakar

Received: 28 February 2011 / Revised: 1 January 2012 / Published online: 6 June 2012

Abstract Randomized load balancing greatly improves the sharing of resources while being simple to implement. In one such model, jobs arrive according to a rate-αN Poisson process, with $\alpha < 1$, in a system of N rate-1 exponential server queues. In Vvedenskaya et al. (Probl. Inf. Transm. 32:15–29, 1996), it was shown that when each arriving job is assigned to the shortest of D, $D \geq 2$, randomly chosen queues, the equilibrium queue sizes decay doubly exponentially in the limit as $N \to \infty$. This is a substantial improvement over the case $D = 1$, where queue sizes decay exponentially.

The reasoning in Vvedenskaya et al. (Probl. Inf. Transm. 32:15–29, 1996) does not easily generalize to jobs with nonexponential service time distributions. A modularized program for treating randomized load balancing problems with general service time distributions was introduced in Bramson et al. (Proc. ACM SIGMETRICS, pp. 275–286, 2010). The program relies on an ansatz that asserts that, for a randomized load balancing scheme in equilibrium, any fixed number of queues become independent of one another as $N \to \infty$. This allows computation of queue size distributions and other performance measures of interest.

In this article, we demonstrate the ansatz in several settings. We consider the least loaded balancing problem, where an arriving job is assigned to the queue with the smallest workload. We also consider the more difficult problem, where an arriving

M. Bramson (✉)
School of Mathematics, University of Minnesota, Minneapolis, MN 55455, USA
e-mail: bramson@math.umn.edu

Y. Lu
Department of Electrical and Computer Engineering, University of Illinois, Urbana, IL 61801, USA
e-mail: yilu4@illinois.edu

B. Prabhakar
Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA
e-mail: balaji@stanford.edu

job is assigned to the queue with the fewest jobs, and demonstrate the ansatz when the service discipline is FIFO and the service time distribution has a decreasing hazard rate. Last, we show the ansatz always holds for a sufficiently small arrival rate, as long as the service distribution has 2 moments.

Keywords Load balancing · Join the shortest queue · Join the least loaded queue · Asymptotic independence

Mathematics Subject Classification 60K25 · 68M20 · 90B15

1 Introduction

Randomized load balancing is a canonical method for efficiently sharing resources among different jobs that is often simple to implement. For example, it is commonly used in hash tables in data switches for looking up the addresses of incoming packets at high speed; this application was first modeled and analyzed by Azar et al. [1]. In the dynamic version of randomized load balancing, jobs arrive at a bank of N queues, with each arriving job being assigned to a server so as to reduce the long-term backlog in the system. Dynamic randomized load balancing is often referred to as the *supermarket model*.

We are interested here in two load balancing policies for the supermarket model. In each case, jobs arrive at the bank of N servers according to a rate-αN Poisson process, with $\alpha < 1$. The servers each employ the same service discipline (such as FIFO) and the service times are IID with a given arbitrary distribution $F(\cdot)$ having mean 1. As throughout this article, service at each queue is assumed to be non-idling. The *join the shortest queue* policy SQ(D) assigns each arrival to the shortest of D queues chosen independently and uniformly at random, where the shortest queue means the queue with the least number of jobs. When the arrival is instead assigned to the queue with the smallest amount of remaining work, or *workload*, we refer to the policy as *join the least loaded queue* and write LL(D). In both cases, the D queues are chosen without replacement (from among the $\binom{N}{D}$ possible sets). Ties are assumed to be broken randomly, with the arriving job being assigned with equal probability to each of the queues.

When the service times are exponentially distributed, it is not difficult to show that the underlying Markov process is positive recurrent and a unique equilibrium distribution exists. Vvedenskaya et al. [19] analyzed the equilibrium distribution under the SQ(D) policy, with replacement, and found that, for fixed D, $D \geq 2$, as the number of queues N goes to infinity, the limiting probability that the number of jobs in a given queue is at least k is $\alpha^{(D^k-1)/(D-1)}$. This is a substantial improvement over the case $D = 1$, where the corresponding probability is α^k.

The model with exponentially distributed service times was also studied by Mitzenmacher [17]. Its path space evolution was studied by Graham [11] who moreover showed that, starting from independent initial states, as $N \to \infty$, the queues of the limiting process evolve independently. Luczak and McDiarmid [15] showed that the length of the longest queue scales as $(\log\log N)/\log D + O(1)$. Certain generalizations have also been explored. Martin and Suhov [16] studied the supermarket

mall model where each node in a Jackson network is replaced by M parallel servers, and a job joins the shortest of D randomly chosen queues at the node to which it is directed. Luczak and McDiarmid [14] studied the maximum queue length of the original supermarket model when the service speed scales linearly with the number of jobs in the queue.

Little theoretical work has been done on the supermarket model with nonexponentially distributed service times. In this setting, the positive Harris recurrence of the Markov process underlying the supermarket model is no longer obvious. (Since the state space will typically be uncountable, positive Harris recurrence rather than positive recurrence is needed.) In particular, for the SQ(D) policy, jobs might be assigned to short queues where the remaining work is high, which can cause service inactivity after queues with many jobs, but low remaining work, empty. If the system can be "tricked" too often in this manner, it is conceivable that it is unstable although $\alpha < 1$ and the service time has mean 1. Moreover, for general service distributions, the evolution of the supermarket model with the SQ(D) policy will be influenced by the service discipline, which complicates analysis.

Foss and Chernova [10] demonstrated positive Harris recurrence for the supermarket model, for given N, under the FIFO service discipline and general service times. In particular, for given N, a unique equilibrium distribution $\mathcal{E}^{(N)}$ exists. Bramson [4] extended this to general service disciplines and showed uniform bounds, in N, on the tails of $\mathcal{E}^{(N)}$ at each queue. (Both works pertain to a more general setting for arrivals and the rule for selecting the D queues.) Fluid limits are employed as the main tool in [10] and an appropriate Lyapunov function is employed in [4].

For general service times, one wishes to analyze the limiting behavior of the equilibria $\mathcal{E}^{(N)}$, at a given queue, as $N \to \infty$. In Bramson et al. [5], a modularized program is developed for this purpose and relies on an ansatz that asserts that, in equilibrium, any fixed number of queues become independent of one another as $N \to \infty$. This allows computation of queue size distributions and other quantities of interest. Employing the ansatz, it is shown in Bramson et al. [6] that the limiting equilibrium distribution will sometimes have a doubly exponential tail, but that other behavior is also possible, depending on the service discipline and the tail of the service distribution $F(\cdot)$.

In this article, we will demonstrate this independence ansatz under several settings. We first do so for LL(D) policies; this requires no additional assumptions. We next consider SQ(D) policies, which we are only able to analyze when the service discipline is FIFO and the service time distribution has a decreasing hazard rate (DHR). This includes heavy-tailed service distributions and is shown in [6] to lead to interesting phenomena. Last, we show the ansatz holds for a sufficiently small arrival rate, with no assumptions on the policy for selecting a queue, as long as the service distribution has 2 moments. The demonstration of the ansatz in the general setting, without any restrictions, appears to be a difficult problem.

This article is organized as follows. In Sect. 2, we state the ansatz precisely and then state the main results corresponding to the above cases, with independence for the LL(D) policies being demonstrated in Theorem 2.1, independence for the FIFO SQ(D) policy being demonstrated in Theorem 2.2, and independence for small arrival rates being demonstrated in Theorem 2.3. The first two proofs are based on a monotonicity argument that states the process starting from the empty state is dominated

by the process starting from any other state. This is then employed to show uniform convergence as $t \to \infty$, in N, to the corresponding equilibria, when observation of the state is restricted to a fixed number of queues. The third proof employs branching-like reasoning to construct a supermartingale, from which this uniform convergence in N again follows.

In Sect. 3, we provide basic background on the properties of the state space and Markov process that underly the different supermarket models. Section 4 develops the monotonicity argument mentioned above, and Sect. 5 applies it to demonstrate uniform convergence for the LL(D) and SQ(D) models. In Sect. 6, uniform convergence is also demonstrated for general policies and small enough arrival rates. Rather than monotonicity, a martingale argument is applied there. Section 7 shows for all three models that, for large N, near independence persists over small times when the queues are independent in the initial state. In Sect. 8, the main results of Sects. 5–7 are applied to demonstrate Theorems 2.1–2.3.

2 Main results

We state the ansatz and the main results of the article, Theorems 2.1–2.3, and briefly discuss their proofs. For this, we need to introduce some terminology.

Each result is stated in terms of the limit, as $N \to \infty$, of Markov processes $X^{(N)}(t)$, $t \geq 0$, underlying supermarket models with N queues. Such a Markov process is defined on an appropriate state space $S^{(N)}$ that is a product of state spaces $S^{(1)}$ corresponding to each queue. In order to avoid technical details, we postpone until Sect. 3 the construction of $S^{(N)}$ and $X^{(N)}(\cdot)$. At this point, we require only limited specifics, namely that a state $x \in S^{(N)}$ is given by descriptors, including the number of jobs z^n at each queue n, $n = 1, \ldots, N$; the residual service times $v^{n,i}$, $n = 1, \ldots, N$ and $i = 1, \ldots, z^n$, for each of the jobs currently in the system; and the amount of service already received $s^{n,i}$, $n = 1, \ldots, N$ and $i = 1, \ldots, z^n$, by the jobs.

We denote by $\mathcal{E}^{(N,N')}$ the projection of the equilibrium measure $\mathcal{E}^{(N)}$ onto the first N' queues. (Since $X^{(N)}(t)$ is exchangeable when $X^{(N)}(0)$ is, the choice of queues will not matter.) We say that a service discipline for the supermarket model is *local* if the amount of service, at a given queue n, that is assigned to each of the jobs currently there, is a function only of the state of the process at n (e.g., involving terms such as z^n, $v^{n,i}$, $i = 1, \ldots, z^n$, and $s^{n,i}$, $i = 1, \ldots, z^n$). This assumption on $X^{(N)}(\cdot)$ will be needed to ensure the independence of individual queues as $N \to \infty$ in the ansatz.

We need to describe the evolution of individual queues for the limiting process, as $N \to \infty$. For this, we construct a process $X^{\mathcal{H}}(t)$, $t \geq 0$, on $S^{(1)}$, as follows. Let $\mathcal{H}$ denote a probability measure on $S^{(1)}$, which we refer to as the *environment* of the process $X^{\mathcal{H}}(\cdot)$; we refer to $X^{\mathcal{H}}(\cdot)$ as the *cavity process*. We define $X^{\mathcal{H}}(\cdot)$ so that potential arrivals arrive according to a rate-$D\alpha$ Poisson process. When such a *potential arrival* to the queue occurs at time t, $X^{\mathcal{H}}(t-)$ is compared with the states of $D-1$ independent random variables with law $\mathcal{H}$; we refer to these $D-1$ states at a potential arrival as the *comparison states*. Choosing from among these D states, the job is assigned by following the same policy as for the corresponding supermarket model. (For instance, if the SQ(D) policy is employed, then the job is assigned to

the state with the fewest number of jobs.) If the job has chosen the state $X^{\mathcal{H}}(t-)$ at the queue, it then immediately joins the queue; otherwise, the job immediately leaves the system. In either case, the independent $D-1$ states employed for this purpose are immediately discarded. Jobs have the same service distribution and are served according to the same local service discipline as for the corresponding supermarket model. We note that when $X^{\mathcal{H}}(t)$ has measure $\mathcal{H}$ (i.e., the same measure as the comparison states), a potential arrival will choose the queue with probability $1/D$, and so arrivals to the queue occur at rate α. When the environment is a function of t, in which case we write $\mathcal{H}(t)$, we refer to it as the *environment process*; $X^{\mathcal{H}(\cdot)}(\cdot)$ is then defined as above.

When a process $X^{\mathcal{H}}(\cdot)$, with environment $\mathcal{H}$, is stationary with the equilibrium measure $\mathcal{H}$ (i.e., $X^{\mathcal{H}}(t)$ has the distribution $\mathcal{H}$ for all t), we say that $\mathcal{H}$ is an *equilibrium environment*. One can think of an equilibrium environment as being the restriction of an equilibrium measure for the corresponding supermarket model, viewed at a single queue, when "the total number of queues N is infinite". When a process $X^{\mathcal{H}(\cdot)}(\cdot)$, with environment process $\mathcal{H}(\cdot)$, at every time t has distribution $\mathcal{H}(t)$, we say that $\mathcal{H}(\cdot)$ is an *equilibrium environment process*.

We now state the ansatz. Here, $\xrightarrow{v}$ on $S^{(N')}$ denotes convergence in total variation with respect to an appropriate metric $d^{N'}(\cdot,\cdot)$ on $S^{(N')}$. (The metrics will be specified in Sect. 3.)

Ansatz *Consider the supermarket model, with N queues, operating under the SQ(D) or LL(D) policy for fixed D, and possessing a local service discipline that is the same at all queues. Jobs are assumed to have an arbitrary service time distribution $F(\cdot)$, with mean* 1, *and arrivals to the system are Poisson and occur at rate $\alpha < 1$. Then* (a) *for each N',*

$$\mathcal{E}^{(N,N')} \xrightarrow{v} \mathcal{E}^{(\infty,N')} \quad \text{as } N \to \infty, \tag{2.1}$$

where $\mathcal{E}^{(\infty,N')}$ is the N'-fold product of $\mathcal{E}^{(\infty,1)}$. Moreover, (b) *$\mathcal{E}^{(\infty,1)}$ is the unique equilibrium environment for this supermarket model.*

We state the versions of the ansatz that we are able to demonstrate. Theorem 2.1 states that the ansatz always holds for the least loaded policy. Since the choice of service discipline has no effect on which queue an arriving job is directed to, the robustness of this result is not surprising.

Theorem 2.1 *Suppose the assumptions of the ansatz are satisfied for the supermarket model operating under the LL(D) policy. Then the conclusions* (a) *and* (b) *in the ansatz hold.*

The ansatz is considerably more difficult to demonstrate for the supermarket model satisfying the shortest queue policy, and most cases remain open. The next result, Theorem 2.2, demonstrates the ansatz under the FIFO service discipline, for a service distribution $F(\cdot)$ having decreasing hazard rate $h(\cdot)$ (i.e., $h(s) = F'(s)/\bar{F}(s)$ is nonincreasing in s, where $\bar{F}(s) = 1 - F(s)$).

Theorem 2.2 *Suppose the assumptions of the ansatz are satisfied for the supermarket model operating under the SQ(D) policy. Suppose moreover that the service discipline is FIFO and that $F(\cdot)$ has decreasing hazard rate. Then the conclusions* (a) *and* (b) *in the ansatz hold.*

The proofs of Theorems 2.1 and 2.2 employ similar arguments, which we summarize briefly here. Each case utilizes a preordering among the states at a given queue. Under such a preordering, if the states at all of the queues for one initial state dominate those at another initial state, the processes can be coupled so that this condition persists at all times. Since the empty state is dominated by all other states, this implies the distribution of the process starting from the empty state is increasing over time, and therefore converges to an equilibrium distribution. By employing a suitable metric and the uniform bounds from [4] on the equilibrium measures over all N, it will follow that this convergence is uniform in N.

On the other hand, for large enough N, the process started from the empty state will have nearly independent queues over a fixed time interval. By the above uniform convergence of the process, for large enough N and appropriate t, this process will be, at time t, both close to its equilibrium measure and have nearly independent queues. Letting both N and t go to infinity, it will follow that the sequence of equilibrium measures indexed by N converges to a product measure that is the unique equilibrium environment specified in Part (b) of the ansatz.

Theorem 2.3 implies that, for a sufficiently small arrival rate, the conclusions of the ansatz hold irrespective of the service discipline as long as the service distribution has 2 moments. Its proof does not require the SQ(D) or LL(D) policy but only that, after the set of D queues is selected, an arriving job be assigned to one of them according to a fixed rule involving only the states at these D queues, and not depending on N, with the assignment being made in an exchangeable manner (i.e., with the labeling of the queues playing no role). We refer to a model with such a policy as a *generalized supermarket model*.

Theorem 2.3 *Suppose the assumptions of the ansatz are satisfied for the generalized supermarket model and that its service distribution $F(\cdot)$ has 2nd moment $\theta < \infty$. For $\alpha \leq 1/(2\sqrt{D(D \vee \theta)})$, the conclusions* (a) *and* (b) *in the ansatz hold.*

The proof of Theorem 2.3 compares the process corresponding to this model to that for an M/G/1 queue, with Poisson arrival rate $D\alpha$, and with the same service distribution. The latter process is used, together with a martingale argument, to provide a lower bound on the rate at which the original process converges to its equilibrium measure, which does not depend on N. Theorem 2.3 will follow from this uniform convergence and reasoning similar to that employed for Theorems 2.1 and 2.2.

The method of proof that was employed in Theorems 2.1 and 2.2 unfortunately does not apply to many important service disciplines, such as processor sharing and preemptive LIFO. A major part of the difficulty is the absence of a natural preordering between states that is preserved over time, in contrast to the above cases.

3 Markov process background

In this section, we provide a more detailed description of the construction of the Markov processes $X^{(N)}(\cdot)$ that underly the different versions of the supermarket models we consider. Related material for queueing networks is given in Bramson [3] and, for a general family of join the shortest queue networks, in Bramson [4]. Because of the similarity of these settings, we present a summary here and refer the reader to [4] for more detail.

The state space $S^{(N)}$ will be defined somewhat differently for the three models, depending on how much information we wish to record. In the LL(D) and generalized supermarket model settings, we define $S^{(N)}$ to be the set

$$\left(\mathbb{Z}^2 \times \mathbb{R}^3\right)^{\infty} \tag{3.1}$$

subject to the following constraints. Only a finite number of the 5-tuples of coordinates are nonzero, with each 5-tuple corresponding to a particular job in the system. The first coordinate n, $n = 1, \ldots, N$, corresponds to the queue of the job; the next coordinate i, $i = 1, \ldots, z^n$, where z^n is the number of jobs at the queue, gives its rank at the queue based on the time of arrival there, with "older" jobs receiving a lower rank. The third coordinate ℓ, $\ell \geq 0$, is the age of the job (and is used to determine the second coordinate); the fourth coordinate v, $v > 0$, is the residual service time; and the last coordinate r, $r \in [0, 1]$, is the current rate of service for the job. Since the discipline is assumed to be non-idling, the sum of the last coordinates for all jobs at a given nonempty queue must equal 1. The 5-tuples are ordered in increasing order in terms of first the first coordinate, and then the second coordinate (so that distinct points in $S^{(N)}$ correspond to distinct states). The coordinates ℓ, v, and r can be labeled in terms of the first two coordinates (e.g., $\ell^{n,i}$ denotes the age of the (n, i)th job). Depending on the service discipline, it may not be necessary to record as much information regarding the state, in which case various coordinates of S may be omitted; alternatively, coordinates can also be added when relevant.

For the SQ(D) model, less information is required because of the FIFO service discipline. In this setting, we define $S^{(N)}$ to be the set

$$\left(\mathbb{Z} \times \mathbb{R}^2\right)^N. \tag{3.2}$$

Here, the first coordinate z^n, $n = 1, \ldots, N$, corresponds to the number of jobs at the nth queue; the second coordinate s^n, $s^n \geq 0$, is the amount of time the oldest job there has already been served; and the last coordinate v^n, $v^n > 0$, is the residual service time. (When $z^n = 0$, set the other two coordinates equal to 0.) One typically omits the second coordinate; in our setting, it will be used in conjunction with the decreasing hazard rate of the service distribution.

In the proof of Theorem 2.2 for the SQ(D) model, we will employ the spaces $S_r^{(N)}$ obtained by omitting some of the information from $S^{(N)}$. The space given in (3.2) is replaced by $S_r^{(N)} = (\mathbb{Z} \times \mathbb{R})^N$, where the coordinate v corresponding to the residual service time is suppressed. In addition, the coordinate s corresponding to the amount of time the oldest job has been served is truncated at s_∞, with jobs receiving more service being assigned this value, where s_∞ is the first value of s at which $\inf_{s \geq 0} h(s)$

is attained. (Recall that the hazard rate $h(\cdot)$ is decreasing.) Note that, when the service distribution is exponential, $s_\infty = 0$.

These new spaces are needed, for the SQ(D) model, in order to use the monotonicity relations between pairs of states that were mentioned in the second section. After showing uniform convergence in N to the equilibria of $X^{(N)}(\cdot)$ on $S_r^{(N)}$, it will not be difficult to show the desired uniform convergence for the processes on $S^{(N)}$.

For given $N' \le N$, $S^{(N')}$ is the *projection* of $S^{(N)}$ obtained by restricting nonzero 5-tuples and 3-tuples to the first N' queues. For $x \in S^{(N)}$ as in (3.1), the projection $x' \in S^{(N')}$ of x is the element obtained by omitting 5-tuples with $n > N'$; for $x \in S^{(N)}$ as in (3.2), x' is obtained by omitting the coordinates with $n > N'$. One can also define projections of $S^{(N)}$ onto spaces $S^{(N')}$ corresponding to other subsets of $\{1, \ldots, N\}$ analogously; we will not use these in the article.

We construct metrics $d^{(N)}(\cdot,\cdot)$, with $d^{(N)}(\cdot,\cdot) = (1/N)\sum_{n=1}^N d^{(N),n}(\cdot,\cdot)$, and $d_r^{(N)}(\cdot,\cdot)$, with $d_r^{(N)}(\cdot,\cdot) = (1/N)\sum_{n=1}^N d_r^{(N),n}(\cdot,\cdot)$, for the above spaces. For the metric on $S^{(N)}$ specified by (3.1), and for given $x_1, x_2 \in S^{(N)}$, with the coordinates labeled correspondingly, set

$$d^{(N),n}(x_1,x_2) = \left|z_1^n - z_2^n\right| + \sum_{i=1}^{\infty}\left(\left|\ell_1^{n,i} - \ell_2^{n,i}\right| + \left|v_1^{n,i} - v_2^{n,i}\right| + \left|r_1^{n,i} - r_2^{n,i}\right|\right). \tag{3.3}$$

For the metric on $S^{(N)}$ specified by (3.2), set

$$d^{(N),n}(x_1,x_2) = \left|z_1^n - z_2^n\right| + \left|s_1^n - s_2^n\right| + \left|v_1^n - v_2^n\right|. \tag{3.4}$$

For the metric on $S_r^{(N)}$ obtained from $S^{(N)}$ in (3.2), set

$$d_r^{(N),n}(x_1,x_2) = \left|\left(z_1^n - 1\right)_+ - \left(z_2^n - 1\right)_+\right| + \left|r_1^n - r_2^n\right|. \tag{3.5}$$

Here,

$$r = r(s) = \int_0^\infty \left(\bar F(s+t)/\bar F(s)\right) dt \tag{3.6}$$

is the expected residual service time of a job, given that it has received s units of service, with $r^n = r(s^n)$, $n = 1, \ldots, N$, being the quantities for the corresponding queues, and $(y)_+ = y \vee 0$. (Note that $r(s)$ is increasing in s when $F(\cdot)$ has decreasing hazard rate.)

There is some flexibility in the choice of the metrics here; the above versions will be convenient for our computations. We will employ r rather than s in (3.5) using the monotonicity inherited from the DHR property of the service distribution $F(\cdot)$, which is assumed in Theorem 2.2. One can check that

$$\bar F(t) = e^{-\int_0^t h(s)\,ds} \tag{3.7}$$

and hence

$$r(s) = \int_0^\infty \left(e^{-\int_0^{s'} h_s(s'')\,ds''}\right) ds', \tag{3.8}$$

where $h_s(s'') \stackrel{\text{def}}{=} h(s''+s)$. Setting $r_\infty = \sup_s r(s)$ and employing the monotonicity of $h(\cdot)$, it is therefore not difficult to check that there is a 1 to 1 correspondence between $s \in [0, s_\infty]$ and $r \in [1, r_\infty]$ for $s_\infty < \infty$, and $s \in [0, \infty)$ and $r \in [1, r_\infty)$ for $s_\infty = \infty$.

We also define a pseudometric on $S^{(N)}$, in (3.1), by setting

$$d_r^{(N),n}(x_1, x_2) = \left| w_1^n - w_2^n \right|, \tag{3.9}$$

where w_i^n is the workload at queue n for the state x_i, i.e., w_i^n is the sum of the residual service times at the queue.

One can check that the metrics $d^{(N)}(\cdot,\cdot)$ and $d_r^{(N)}(\cdot,\cdot)$, given in (3.3), (3.4), and (3.5), are separable and locally compact; more detail is given on page 82 of [3]. We equip $S^{(N)}$ and $S_r^{(N)}$ with the standard Borel σ-algebra inherited from $d^{(N)}(\cdot,\cdot)$ and $d_r^{(N)}(\cdot,\cdot)$, which we denote by $\mathscr{S}^{(N)}$ and $\mathscr{S}_r^{(N)}$.

The Markov process $X^{(N)}(t)$, $t \geq 0$, underlying a given model is defined to be the right continuous process with left limits, taking values in $S^{(N)}$ or $S_r^{(N)}$, whose evolution is determined by the model together with the assigned service discipline. We denote the random values of the coordinates $\ell^{n,i}$, $r^{n,i}$, etc., taken by $X^{(N)}(t)$, by $L^{n,i}(t)$, $R^{n,i}(t)$, etc. For the models on $S^{(N)}$ as in (3.1), jobs are allocated service according to rates $R^{n,i}(t)$ that are assumed to be constant in between arrivals and departures of jobs at the queues. Over such an interval, $L^{n,i}(t)$ increases at rate 1, $V^{n,i}(t)$ decreases at rate $R^{n,i}(t)$, and the workload $W^n(t)$ decreases at rate 1. Upon an arrival or departure, rates are reassigned according to the discipline. The standard service disciplines satisfy this property. (The restriction that service rates remain constant between arrivals and departures of jobs is for convenience, and allows one to inductively construct $X^{(N)}(\cdot)$ over increasing times in a simple way.) The evolutions of the Markov processes $X^{(N)}(\cdot)$ corresponding to $S^{(N)}$ as in (3.2) and to $S_r^{(N)}$ are specified similarly. For the model in (3.2), $S^n(t)$ increases at rate 1 and $V^n(t)$ decreases at rate 1 until the departure of the job.

For each of the above processes $X^{(N)}(\cdot)$, when an arrival in the system occurs at time t, the set of D queues that is chosen will be referred to as the *selection set* of the arrival. We will also say that a *potential arrival* occurs then at each of these queues.

Along the lines of page 85 of [3], a filtration $(\mathcal{F}_t^{(N)})$, $t \in [0, \infty]$, can be assigned to $X^{(N)}(\cdot)$ so that $X^{(N)}(\cdot)$ is a piecewise-deterministic Markov process, and hence is Borel right. This implies that $X^{(N)}(\cdot)$ is strong Markov. (We do not otherwise use Borel right.) The reader is referred to Davis [8] for more detail.

We note that, for the SQ(D) model, there is a natural map φ between the sample paths $X^{(N)}(t)$, $t \in [0, \infty)$, for the state spaces $S^{(N)}$ and $S_r^{(N)}$, where $\varphi(x(\cdot))$, for a sample path $x(\cdot)$ on $S^{(N)}$, is defined by omitting the v coordinate for the residual service time. The map is bijective since the residual service time, at a given queue and time, is the remaining time until the next departure there, which is contained in the corresponding sample path $x(\cdot)$ taking values in $S_r^{(N)}$ (although the residual service time depends on the values of the path at later times). The standard coupling on $S_r^{(N)}$, which we construct in the next section, is Markov. Its analog for the SQ(D) model on $S^{(N)}$ is not Markov, however. (This requires some thought.) On the other hand,

for the LL(D) model on $S^{(N)}$, the standard coupling there is Markov; moreover, the corresponding map φ of sample paths from $S^{(N)}$ to $S_r^{(N)}$, obtained by retaining only the workload at each queue, is not bijective. For these reasons, our main computations for the SQ(D) model will be performed on $S_r^{(N)}$, but our main estimates for the LL(D) model will be performed on $S^{(N)}$.

One can show that the SQ(D) supermarket model is Feller, although the LL(D) and generalized supermarket models are not. (We will not need these results.) Convergence in total variation rather than weak convergence is therefore the right medium in which to treat all three models. Convergence in total variation, as in (2.1), means that

$$\lim_{N\to\infty} \sup_{A\in\mathscr{S}^{(N')}} \left|\mathcal{E}^{(N,N')}(A) - \mathcal{E}^{(\infty,N')}(A)\right| = 0.$$

For all of our models, we will in fact employ a somewhat stronger version of convergence in total variation. Consider a sequence of probability measures $\mathcal{M}_k$, $k = 1, 2, \ldots$, defined on the path space $[0, T] \times S^{(N')}$, for given $T, N' > 0$, of right continuous paths with left limits, and corresponding Borel σ-algebra $\mathcal{B}_T^{(N')}$. Then, $\mathcal{M}_k$ converges to $\mathcal{M}_\infty$ in total variation, written $\mathcal{M}_k \xrightarrow{v} \mathcal{M}_\infty$, if

$$\lim_{k\to\infty} \sup_{A\in\mathcal{B}_T^{(N')}} \left|\mathcal{M}_k(A) - \mathcal{M}_\infty(A)\right| = 0.$$

Except when stated otherwise, in the remainder of the article, $X^{(N)}(\cdot)$ will denote the Markov process underlying one of the three supermarket or generalized supermarket models. When two or more processes, e.g., $X_i^{(N)}(\cdot)$, $i = 1, 2$, are employed together, they will correspond to the same model and parameters, differing only in the initial state. When confusion is unlikely, we will sometimes drop the superscript N from quantities such as $X^{(N)}(\cdot)$.

4 Monotonicity for the LL(D) and SQ(D) models

In this section, we introduce the standard coupling for the LL(D) supermarket model and for the SQ(D) supermarket model that is FIFO with DHR. The coupling for the LL(D) model is defined on $S^{(N)}$ and the coupling for the SQ(D) model is defined on $S_r^{(N)}$; in both cases, we will often drop the superscript N for convenience. These couplings induce a monotonicity property for each model that will imply convergence, when starting from the empty state, to a limiting distribution that will also be an equilibrium. Estimates in Sect. 5 show this convergence is uniform in N in an appropriate sense, which will be used in conjunction with Proposition 7.1 to demonstrate Theorems 2.1 and 2.2 in Sect. 8.

The standard coupling The *standard coupling* for the LL(D) supermarket model on S is the pathwise coupling between two copies $X_1(\cdot)$ and $X_2(\cdot)$ of the corresponding Markov process that is defined as follows. For a random permutation $\pi_t = \pi_t^{(N)} = (\pi_t^n)$, $n = 1, \ldots, N$, on $t \geq 0$, each queue n of $X_1(\cdot)$, at time t, is coupled with

queue π_t^n of $X_2(\cdot)$ so that these queues have the same potential arrivals, for each ω, and so that the corresponding arrivals, which are assigned according to the LL(D) policy in each case, have the same service times. Setting π_0 equal to the identity, the permutation π_t is assumed to be constant in between arrivals, where it is updated inductively. An arrival at time t may occur at a given queue n_1 for $X_1(\cdot)$, but at a queue $n_2 \neq \pi_{t-}^{n_1}$ for $X_2(\cdot)$, due to the different workloads in the two systems. (Ties in the workload at queues in the selection set are broken in the same way for each process.) When this occurs, one changes the permutation at time t by setting

$$\pi_t^{n_1} = n_2, \qquad \pi_t^{n_1'} = \pi_{t-}^{n_1}, \qquad \pi_t^n = \pi_{t-}^n \quad \text{for } n \neq n_1, n_1',$$

where n_1' is defined by $\pi_{t-}^{n_1'} = n_2$. That is, the queues in each system where the arrival has just occurred are coupled together, as are the pair of queues previously coupled with them, with all other queues retaining the same coupling. At arrivals where $n_2 = \pi_{t-}^{n_1}$, the permutation remains the same. We denote by $X_{2,\pi}(\cdot)$ the process obtained from $X_2(\cdot)$ by permuting its queues according to $\pi_\cdot$, that is,

$$X_{2,\pi}^n(t) = X_2^{\pi_t^n}(t).$$

The standard coupling, on S_r, for the SQ(D) supermarket model that is FIFO with DHR is defined so that both processes again share the same potential arrival and service time processes. In this setting, $\pi_\cdot$ is always defined to be the identity map, that is, the nth queue of $X_1(\cdot)$ is always coupled with the nth queue of $X_2(\cdot)$, and hence $X_{2,\pi}(\cdot) = X_2(\cdot)$. (This will be needed in the proof of Lemma 4.1 when comparing $S_1^n(t)$ with $S_2^n(t)$.) In addition, the service times of the oldest jobs at a given queue n are coupled so that, when $S_1^n(t) \leq S_2^n(t)$, service for both processes is completed simultaneously at rate $h(S_2^n(t))$ and, independently of this, service for the first process only is also completed at rate $h(S_1^n(t)) - h(S_2^n(t))$. If service for the job in the first process is completed before that in the second process, completion of service continues at rate $h(\cdot)$ for the latter. This coupling relies on the DHR property. Note that, if service commences at a new job for the first process when the corresponding job for the second process is already being served, then completion of service for the new job occurs at a faster rate than for the other job. (This relies again on the DHR property.) Upon a potential arrival, ties in the length of queues in the selection set are broken in the same way for each process, for a given ω.

Extensions of the standard coupling, from 2 to L copies of the processes $X_1(\cdot), \ldots, X_L(\cdot)$, hold for both supermarket models by applying the same reasoning as above. In Sect. 5, the coupling, with $L = 3$, will be employed in one place. The bijection $\varphi(\cdot)$, which was defined at the end of Sect. 3, induces a coupling for the SQ(D) supermarket model on S from the standard coupling on S_r. We will first employ it in Sect. 5, where we will employ the notation $\pi_\cdot$ as well.

For both the LL(D) and SQ(D) models, we define a preorder between pairs of states x_1, x_2 in S or S_r. For the LL(D) model, we say that $x_1^n \leq x_2^n$ if $w_1^n \leq w_2^n$, with $x_1 \leq x_2$ if $w_1^n \leq w_2^n$ for all $n = 1, \ldots, N$. For the SQ(D) model, we require instead that $z_1^n \leq z_2^n$ and $s_1^n \leq s_2^n$; the last condition is equivalent to $r_1^n \leq r_2^n$. (For the LL(D) model, $x_1 \leq x_2$ and $x_2 \leq x_1$ together need not imply $x_1 = x_2$, and so "$\leq$" is not a

partial order whereas, for the SQ(D) model, "$\leq$" is a partial order.) Note that the state $x_1 = 0$, where each queue is empty, satisfies $x_1 \leq x_2$ for any other state x_2. The following lemma states that, if the preorder holds initially, then it persists for all time under the standard coupling.

Lemma 4.1 *For either the LL(D) supermarket model on S, or the SQ(D) model on S_r that is FIFO with DHR, assume that the underlying Markov processes $X_1(\cdot)$ and $X_2(\cdot)$ satisfy $X_1(0) \leq X_2(0)$ and are coupled by the standard coupling π.. Then, for each ω,*

$$X_1(t) \leq X_{2,\pi}(t) \quad \textit{for all } t. \tag{4.1}$$

Proof We apply the standard coupling to each model and argue by contradiction, setting $T = \inf\{t : X_1(t) \not\leq X_{2,\pi}(t)\}$ in each case. We first consider the LL(D) supermarket model.

It is easy to see that $T < \infty$ cannot occur with $X_1(T) \leq X_{2,\pi}(T)$. If it does, then, for small enough $\epsilon > 0$ (depending on ω), there are no arrivals or departures in the system over $(T, T+\epsilon]$. Since $W_1^n(t) \leq W_{2,\pi}^n(t)$ holds at $t = T$, the inequality continues to hold for $t \in (T, T+\epsilon]$, which leads to a contradiction.

Suppose now that $X_1(T) \not\leq X_{2,\pi}(T)$. The inequality $W_1^n(t) \leq W_{2,\pi}^n(t)$ will continue to hold at $t = T$ for all n, except possibly at pairs where there is an arrival for one of $X_1^n(\cdot)$ and $X_{2,\pi}^n(\cdot)$, but not for both. On the other hand, if such an arrival occurs at time T at n_1 for $X_1(\cdot)$ and at $n_2 \neq n_1$ for $X_{2,\pi}(\cdot)$, then

$$W_1^{n_1}(T-) \leq W_{2,\pi}^{n_1}(T-), \qquad W_1^{n_2}(T-) \leq W_{2,\pi}^{n_2}(T-),$$

$$W_1^{n_1}(T-) \leq W_1^{n_2}(T-), \qquad W_{2,\pi}^{n_1}(T-) \geq W_{2,\pi}^{n_2}(T-),$$

where the first line follows from the definition of T and the second line is a consequence of the service discipline. Therefore,

$$W_1^{n_1}(T-) \leq W_{2,\pi}^{n_2}(T-), \qquad W_1^{n_2}(T-) \leq W_{2,\pi}^{n_1}(T-).$$

Denoting by A the service time of the arrival at n_1 and n_2, it follows that

$$W_1^{n_1}(T) = W_1^{n_1}(T-) + A \leq W_{2,\pi}^{n_2}(T-) + A = W_{2,\pi}^{n_1}(T),$$

$$W_1^{n_2}(T) = W_1^{n_2}(T-) \leq W_{2,\pi}^{n_1}(T-) = W_{2,\pi}^{n_2}(T).$$

Consequently, $W_1(T) \leq W_{2,\pi}(T)$, which again contradicts $T < \infty$.

The argument for the SQ(D) supermarket model is the same when $X_1(T) \leq X_2(T)$, but with $Z_i^n(\cdot)$ and $R_i^n(\cdot)$ replacing $W_i^n(\cdot)$. Suppose now that $X_1(T) \not\leq X_2(T)$, and hence $X_1^n(T) \not\leq X_2^n(T)$ for some n. Because of the DHR property and the standard coupling, a departure at n for $X_2(\cdot)$, at time T, can only occur when a departure occurs there for $X_1(\cdot)$, which would contradict the above inequality. On the other hand, if, at time T, an arrival at n occurs for $X_1(\cdot)$, but at some $n' \neq n$ for $X_2(\cdot)$, then

$$\text{either} \quad Z_1^n(T-) < Z_1^{n'}(T-) \quad \text{or} \quad Z_2^n(T-) > Z_2^{n'}(T-),$$

because of the coupling. Since $Z_1^{n'}(T-) \le Z_2^{n'}(T-)$, it follows that $Z_1^n(T-) \le Z_2^n(T-) - 1$, and hence $Z_1^n(T) \le Z_2^n(T)$. Since $S_1^n(t) \le S_2^n(t)$, and hence $R_1^n(t) \le R_2^n(t)$, continues to hold at time T, this implies $X_1^n(T) \le X_2^n(T)$, which again produces a contradiction. Consequently, $T < \infty$ cannot occur for the SQ(D) model as well. □

We will say that two probability measures $\mathcal{E}_1$ and $\mathcal{E}_2$, on S or S_r, satisfy $\mathcal{E}_1 \overset{\mathcal{P}}{\le} \mathcal{E}_2$ if, for some coupling of random variables X_1 and X_2 with these measures, $X_1(\omega) \le X_2(\omega)$ for all ω. Let $\mathcal{E}_i(t)$, $t \ge 0$, $i = 1, 2$, denote two families of measures belonging to processes $X_i(t)$ underlying either the LL(D) or SQ(D) supermarket model. When restated in terms of these measures, Lemma 4.1 implies the following.

Lemma 4.2 *For either the LL(D) supermarket model or the SQ(D) model that is FIFO with DHR, define $\mathcal{E}_i(t)$, $t \ge 0$, $i = 1, 2$, as above. Assume that $\mathcal{E}_i(0)$ are each exchangeable with respect to $n = 1, \ldots, N$. Then $\mathcal{E}_2(t) = \mathcal{E}_{2,\pi}(t)$ for all t. Moreover, if $\mathcal{E}_1(0) \overset{\mathcal{P}}{\le} \mathcal{E}_2(0)$, then*

$$\mathcal{E}_1(t) \overset{\mathcal{P}}{\le} \mathcal{E}_2(t) \quad \textit{for all } t. \tag{4.2}$$

Proof One can choose $X_1(0)$ and $X_2(0)$ with measures $\mathcal{E}_1(0)$ and $\mathcal{E}_2(0)$ so that $X_1(0) \le X_2(0)$. Since $\mathcal{E}_i(0)$ are each exchangeable, one can choose such $X_i(0)$ so that the pair $(X_1(0), X_2(0))$ is also exchangeable. Hence, under the standard coupling, $(X_1(t), X_{2,\pi}(t))$ is exchangeable for each t.

Moreover, for given t, there is exactly one exchangeable measure on $n = 1, \ldots, N$ for which the distribution on the set of empirical measures obtained from its coordinates is equal to the distribution on the set of empirical measures obtained from $X_{2,\pi}(t)$. Since $X_2(t)$ and $X_{2,\pi}(t)$ are each exchangeable, with the same distribution on the set of their empirical measures, they are themselves equal in distribution, and hence $\mathcal{E}_2(t) = \mathcal{E}_{2,\pi}(t)$.

On the other hand, by Lemma 4.1,

$$X_1(t) \le X_{2,\pi}(t) \quad \text{for all } t.$$

It follows from this and the previous paragraph that $\mathcal{E}_1(t) \overset{\mathcal{P}}{\le} \mathcal{E}_2(t)$ for all t, as desired. □

The empty measure $\mathcal{E}_0 = 0$ and the equilibrium measure $\mathcal{E}_m = \mathcal{E}_m^{(N)}$, for $\alpha < 1$, of an LL(D) or SQ(D) supermarket model, are exchangeable. (To avoid ambiguity here, we employ $\mathcal{E}_m$ rather than the notation $\mathcal{E}$ in the Introduction.) Applying Lemma 4.2 to $\mathcal{E}_1(0) = \mathcal{E}_0$ and $\mathcal{E}_2(0) = \mathcal{E}_1(t_2 - t_1)$ first and then to $\mathcal{E}_1(0) = \mathcal{E}_0$ and $\mathcal{E}_2(0) = \mathcal{E}_m$, we obtain the following results.

Lemma 4.3 *For either the LL(D) supermarket model on S, or the SQ(D) model on S_r that is FIFO with DHR, assume that the underlying Markov process $X(\cdot)$ satisfies*

$X(0) = 0$. *Then, for each* t_1, t_2, *with* $t_1 \le t_2$, *the corresponding measures* $\mathcal{E}(t)$ *satisfy*

$$\mathcal{E}(t_1) \overset{\mathcal{P}}{\le} \mathcal{E}(t_2). \tag{4.3}$$

If $\alpha < 1$ *and* $\mathcal{E}_m$ *is the equilibrium measure, then*

$$\mathcal{E}(t) \overset{\mathcal{P}}{\le} \mathcal{E}_m \quad \textit{for all } t. \tag{4.4}$$

Set $\mathcal{E}(0) = 0$. On account of (4.3) and (4.4) of Lemma 4.3, it will follow that

$$\mathcal{E}(t) \to \mathcal{E}_m^{(N)} \quad \text{as } t \to \infty, \tag{4.5}$$

for "$\to$" defined appropriately. In order to demonstrate Theorem 2.1 and Theorem 2.2, we will in fact need to show that convergence is uniform on N, which will be used to interchange the t and N limits in Sect. 8. For this, we will need the uniform bounds that are given in the following subsection.

Uniform bounds on $\mathcal{E}_m^{(N)}$ For both the LL(D) and SQ(D) supermarket models, we need uniform bounds on the right tails of the corresponding equilibria $\mathcal{E}_m^{(N)}$ that do not depend on N; these bounds rely on results from [4]. For $x \in S^{(N)}$, we set

$$\|x\|^n = w^n, \quad n = 1, \dots, N,$$

for the LL(D) model and, for $x \in S_r^{(N)}$, we set

$$\|x\|^n = \left(z^n - 1\right)_+ + r^n, \quad n = 1, \dots, N,$$

for the SQ(D) model.

Proposition 4.1 *Fix* α, $F(\cdot)$ *and* D, *with* $\alpha < 1$ *and* $F(\cdot)$ *having mean* 1. *For both the LL(D) and SQ(D) supermarket models,*

$$\sup_N \sup_{n \le N} \left\{ \mathcal{E}_m^{(N)} \left(\|X\|^n > M \right) \right\} \to 0 \quad \textit{as } M \to \infty. \tag{4.6}$$

Sketch of proof We note that, since the equilibria $\mathcal{E}_m^{(N)}$ are exchangeable in n for both models, the rate of convergence of the probabilities in (4.6) does not depend on n.

We first consider the proposition for the SQ(D) model. The limit (4.6) will follow from the analogous limits for Z^n and R^n in place of $\|X\|^n$. Since $r = r(s)$, the limit for R^n follows from that for S^n, which is the amount of service already received by the job. This is bounded above by the total service requirement of the job. Therefore, by comparison with the renewal process with distribution $F(\cdot)$, it is not difficult to see that

$$\sup_N \mathcal{E}_r^{(N)} \left(S^n > M \right) \le \int_M^\infty \bar{F}(t)\, dt \le 1. \tag{4.7}$$

(The first inequality is in fact strict since a queue may be empty.) This implies the desired limit for S^n, and hence for R^n.

The limit for Z^n is considerably more difficult, but follows from Corollary 1.2 of Theorem 1.3 in [4], with a little work. The spaces in the corollary contain all the information in $S^{(N)}$, and hence in $S_r^{(N)}$, after appending to the states the amount of time each job has already been served. We refer here to these enriched spaces by $S_e^{(N)}$. As observed below (5.35) in [4], the conclusion (1.24) in the corollary continues to hold on $S_e^{(N)}$ for service disciplines including FIFO. This implies in particular that the equilibria $\mathcal{E}_e^{(N)}$ on $S_e^{(N)}$, for the SQ(D) supermarket model, satisfy

$$\sup_N \mathcal{E}_e^{(N)}\big(Z^n > M\big) \to 0 \quad \text{as } M \to \infty. \tag{4.8}$$

Projecting $S_e^{(N)}$ by removing all information, except for the number of jobs at each queue and the amount of time the oldest job there has already been served, produces $S_r^{(N)}$. Since the evolution of the process depends only on the number of jobs at each queue, the desired bound on Z^n follows immediately from (4.8).

In order to show (4.6) for the LL(D) model, we need to show that, for given n,

$$\sup_N \big\{\mathcal{E}^{(N)}\big(W^n > M\big)\big\} \to 0 \quad \text{as } M \to \infty. \tag{4.9}$$

The spaces in Corollary 1.2 of [4] already contain the information in $S^{(N)}$, and so one does not need to enrich these spaces in the LL(D) setting. There is less work needed here than that for Z^n above, since the uniform stability of the LL(D) model is considerably easier to analyze. By employing the norm in (5.36)–(5.37) of [4], one can show (4.9). □

5 Uniform convergence for the LL(D) and SQ(D) models

In order to demonstrate Theorems 2.1 and 2.2, we will need to demonstrate a variant of (4.5) that is uniform in N. Our first main result for this is Proposition 5.1; the first part of the section is devoted to its proof. We then employ Proposition 5.1 to show a stronger pathwise result, Proposition 5.2, on the original spaces $S^{(N)}$ for both supermarket models; this is done in the second part of the section. Proposition 5.2 will be used in conjunction with Sect. 7 to demonstrate Theorems 2.1 and 2.2 in Sect. 8. In the remainder of the section, we use $S_m^{(N)}$ to denote the spaces $S^{(N)}$ for the LL(D) model and $S_r^{(N)}$ for the SQ(D) model, with $\mathcal{E}_m^{(N)}$ denoting the corresponding equilibria.

Proposition 5.1 states that, for large enough q_0 not depending on N, each system that is started at the empty state will be close to its equilibrium at each time in $t \geq q_0$. For both the LL(D) and SQ(D) models, this is shown with respect to $d_r^{(N)}(\cdot,\cdot)$, which, we recall, is only a pseudometric in the former case.

For a given supermarket model on $S_m^{(N)}$, with $\alpha < 1$, we denote by $X^{(N)}(\cdot)$ the process started from the empty state and by $X_{\mathcal{E}}^{(N)}(\cdot)$ the process started from its equilibrium $\mathcal{E}_m^{(N)}$. We couple these processes by the standard coupling. On account of Lemma 4.1, $X^{(N)}(t) \leq X_{\mathcal{E},\pi}^{(N)}(t)$ for all t and ω. Recall that, in the SQ(D) setting, $\pi.$ is always the identity map.

Proposition 5.1 *Consider, on $S_m^{(N)}$, either the LL(D) supermarket model, or the SQ(D) supermarket model that is FIFO with DHR. Assume the processes $X^{(N)}(\cdot)$ and $X_{\mathcal{E}}^{(N)}(\cdot)$ are defined as above, with $\alpha < 1$ fixed. Then, for each $\gamma > 0$, there exists $q_0 = q_0(\gamma)$ not depending on N such that, for all $t \geq q_0$,*

$$P\left(d_r^{(N),n}\left(X^{(N)}(t), X_{\mathcal{E},\pi}^{(N)}(t)\right) \geq \gamma\right) \leq \gamma \tag{5.1}$$

for all $n = 1, \ldots, N$.

Proposition 5.2 is the analog of Proposition 5.1, but on the original spaces $S^{(N)}$ for both supermarket models. It makes the stronger assertion, for given T and large enough q_1 not depending on N, that, for each $q \geq q_1$ and n, $X^{(N),n}(t) = X_{\mathcal{E},\pi_q}^{(N),n}(t)$ simultaneously for $t \in [q, q+T]$, off of a negligible set of ω; here, $X_{\mathcal{E},\pi_q}^{(N),n}(t) \stackrel{\text{def}}{=} X_{\mathcal{E}}^{(N),\pi_q^n}(t)$, i.e., the permutation of queues for $X_{\mathcal{E}}^{(N)}(\cdot)$ is constant over $[q, q+T]$. (Equation (5.3) also holds with $X_{\mathcal{E},\pi}^{(N),n}(t)$ in place of $X_{\mathcal{E},\pi_q}^{(N),n}(t)$, but we will find the present formulation more convenient.) For the LL(D) supermarket model, we will employ the additional condition that, for given N and appropriate $\epsilon > 0$, the workload W of the equilibrium $\mathcal{E}^{(N)}$ satisfies

$$\mathcal{E}^{(N)}\left(W^{n_i} \in [c_1, c_2], i = 1, 2\right) \leq \left(\mathcal{E}^{(N)}\left(W^{n_1} \in [c_1 - \epsilon, c_2 + \epsilon]\right)\right)^2 + \epsilon \tag{5.2}$$

for each $n_1 \neq n_2$ and $0 \leq c_1 \leq c_2$. We will show in Proposition 7.3 that, for given $\epsilon > 0$, (5.2) is satisfied for large enough N.

Proposition 5.2 *Consider, on $S^{(N)}$, either the LL(D) supermarket model, or the SQ(D) supermarket model that is FIFO with DHR. Assume the processes $X^{(N)}(\cdot)$ and $X_{\mathcal{E}}^{(N)}(\cdot)$ are defined as above.* (a) *Then, for the SQ(D) model, for each $\gamma_1 > 0$ and $T > 0$, there exists $q_1 = q_1(\gamma_1)$ not depending on N such that, for each $q \geq q_1$,*

$$P\left(X^{(N),n}(t) \neq X_{\mathcal{E},\pi_q}^{(N),n}(t) \text{ for some } t \in [q, q+T]\right) \leq \gamma_1, \tag{5.3}$$

for all $n = 1, \ldots, N$. (b) *Assume that, for each N, (5.2) is satisfied for the LL(D) model, with $\epsilon = \epsilon(N) \to 0$ as $N \to \infty$. Then, for each $\gamma_1 > 0$ and $T > 0$, there exists $q_1 = q_1(\gamma_1)$ not depending on N such that (5.3) holds for large enough N, $q \geq q_1$ and all $n = 1, \ldots, N$.*

Demonstration of Proposition 5.1 We first introduce some notation. For two processes $X_1^{(N)}(\cdot)$ and $X_2^{(N)}(\cdot)$ underlying the same supermarket model, we set

$$\psi^{(N),n}(t) = d_r^{(N),n}\left(X_1^{(N),n}(t), X_{2,\pi}^{(N),n}(t)\right) \tag{5.4}$$

and $\psi^{(N)}(t) = (1/N)\sum_{n=1}^N \psi^{(N),n}(t)$.

In the proof of Proposition 5.1, we will employ "truncated" variants $d_{r,L}^{(N)}(\cdot,\cdot)$ of the metric $d_r^{(N)}(\cdot,\cdot)$: for $L \geq 1$, we set $d_{r,L}^{(N)}(x_1, x_2) = \frac{1}{N}\sum_{n=1}^N d_{r,L}^{(N),n}(x_1, x_2)$, with

$$d_{r,L}^{(N),n}(x_1, x_2) = \left|\left(w_1^n \wedge L\right) - \left(w_2^n \wedge L\right)\right|$$

for the LL(D) supermarket model, and

$$d_{r,L}^{(N),n}(x_1,x_2)=\left|\left[\left(\tilde{z}_1^n+r_1^n\right)\wedge L\right]-\left[\left(\tilde{z}_2^n+r_2^n\right)\wedge L\right]\right|,$$

with $\tilde{z}_i^n \stackrel{\text{def}}{=} (z_i^n-1)_+$, for the SQ(D) supermarket model. We will need this truncation because $E[X_{\mathcal{E}}^{(N),n}(0)]$ may be infinite. We set

$$\psi_L^{(N),n}(t)=d_{r,L}^{(N),n}\left(X_1^{(N),n}(t),X_{2,\pi}^{(N),n}(t)\right) \tag{5.5}$$

and $\psi_L^{(N)}(t)=(1/N)\sum_{n=1}^N \psi_L^{(N),n}(t)$. We also set $\|x\|_L^n=\|x\|^n\wedge L$ for both the LL(D) and SQ(D) models. When $x_1^{(N)}\le x_2^{(N)}$, one has

$$d_{r,L}^{(N),n}\left(x_1^{(N)},x_2^{(N)}\right)=\left\|x_2^{(N)}\right\|_L^n-\left\|x_1^{(N)}\right\|_L^n,$$

which is the only setting in which we will employ these truncations.

The following lemma gives lower bounds on the rate of decrease of $E[\psi_L^{(N)}(t)]$ as t increases. In the remainder of the subsection, we will assume that the LL(D) supermarket model is defined on $S^{(N)}$ and the SQ(D) supermarket model is defined on $S_r^{(N)}$.

Lemma 5.1 *Suppose that processes $X_1^{(N)}(\cdot)$ and $X_2^{(N)}(\cdot)$ underlying the LL(D) or SQ(D) supermarket model that is FIFO with DHR satisfy $X_1^{(N)}(0)\le X_2^{(N)}(0)$, and are coupled by the standard coupling. Then, for given $0<a\le b$ and $L>0$,*

$$\begin{aligned}E\left[\psi_L^{(N)}(b)\right]-E\left[\psi_L^{(N)}(a)\right]&\\ \le -\frac{1}{N}\sum_{n=1}^N\int_a^b\Big\{P\left(W_1^{(N),n}(t)=0,W_{2,\pi}^{(N),n}(t)>0\right)&\\ -P\left(\left\|X_2^{(N)}(t)\right\|^n\ge L\right)\Big\}\,dt&\end{aligned} \tag{5.6}$$

for the LL(D) model, and

$$\begin{aligned}E\left[\psi_L^{(N)}(b)\right]-E\left[\psi_L^{(N)}(a)\right]&\\ \le -\frac{1}{N}\sum_{n=1}^N\int_a^b\Big\{P\left(Z_1^{(N),n}(t)=0,Z_2^{(N),n}(t)>0\right)&\\ -2P\left(\left\|X_2^{(N)}(t)\right\|^n\ge L-1\right)\Big\}\,dt&\end{aligned} \tag{5.7}$$

for the SQ(D) model.

Proof In order to obtain (5.6) and (5.7), it suffices to show that, after multiplying by N, the infinitesimal generator of the pair $(X_1^{(N)}(\cdot), X_{2,\pi}^{(N)}(\cdot))$, applied to $\psi_L^{(N)}(\cdot)$, is at most the quantities in the integrands in (5.6) and (5.7) for each coordinate n, and then to apply Dynkin's formula. (See, e.g., Dynkin [9], p. 133. The formula can be obtained here by applying the bounded convergence theorem to $\int_a^b \frac{1}{h}(E[\psi_L^{(N)}(t+h) - \psi_L^{(N)}(t)])\,dt$ as $h \searrow 0$.) When showing (5.6) and (5.7), we avoid the explicit formulas that are needed for a detailed proof.

To see (5.6), note that, since $X_1^{(N)}(\cdot) \le X_{2,\pi}^{(N)}(\cdot)$, $\psi_L^{(N)}(\cdot)$ never increases due to arrivals. Moreover, at each time t, for each n at which $W_1^{(N),n}(t) = 0$ and $W_{2,\pi}^{(N),n}(t) \in (0, L)$, $\psi_L^{(N)}(\cdot)$ decreases at rate $1/N$ due to the service performed at queue n whereas, when $W_{2,\pi}^{(N),n}(t) \ge L$, $\psi_L^{(N)}(\cdot)$ can increase at rate at most $1/N$ due to the decrease of $W_1^{(N),n}(\cdot)$. Applying Dynkin's formula over $[a, b]$, one obtains the bound in (5.6) for $E[\psi_L^{(N)}(b)] - E[\psi_L^{(N)}(a)]$.

To see (5.7), first note that $\psi_L^{(N)}(\cdot)$ does not increase due to a pair of arrivals at the same queue n for the two processes. (Since $r(0) = 1$, $\|X_i^{(N)}\|^n(\cdot)$ increases by 1 upon an arrival, whether or not the queue was empty.) But, if a pair of arrivals occurs at different queues at time t, then

$$\psi_L^{(N)}(t) - \psi_L^{(N)}(t-) \in [0, 1/N]$$

is possible when $\|X_1^{(N)}(t-)\|^n \ge L - 1$, where n is the queue at which the arrival occurs for the first process. On the other hand, for each n at which $Z_1^{(N),n}(t) = 0$ and $\|X_2^{(N)}(t)\|^n \in (0, L)$, $\psi_L^{(N)}(\cdot)$ decreases at rate $1/N$ due to the service performed there whereas, when $\|X_2^{(N)}(t)\|^n \ge L$, $\psi_L^{(N)}(\cdot)$ can increase at rate at most $1/N$ due to the decrease of $\|X_1^{(N)}(\cdot)\|^n$. (Recall again that $r(0) = 1$, and so the start of service of a job upon a departure at the queue does not change $\psi_L^{(N)}(\cdot)$.) Again applying Dynkin's formula and noting that $X_1^{(N)}(\cdot) \le X_2^{(N)}(\cdot)$, one obtains (5.7). □

In Proposition 5.3, we will obtain lower bounds on the integrals of the probabilities involving $W_\cdot^{(N),n}$ and $Z_\cdot^{(N),n}$ on the right side of (5.6) and (5.7) that do not depend on N. For this, we will employ the following lemma, which applies to distribution functions $F(\cdot)$ with decreasing hazard rate. (Recall that $r = r(s)$ is given by (3.6) and there is a 1-1 correspondence between r and s.)

Lemma 5.2 *Suppose that, for given $F(\cdot)$ with decreasing hazard rate, r_1, r_2 satisfy $r_2 - r_1 \ge \delta$, for given $\delta > 0$. Then there exist M_0 and $\epsilon \in (0, 1)$, depending on only $F(\cdot)$ and δ, such that, for some $M \le M_0$,*

$$\frac{\bar{F}(s_2 + M)}{\bar{F}(s_2)} \ge \frac{\bar{F}(s_1 + M)}{\bar{F}(s_1)} + \epsilon. \tag{5.8}$$

Proof Choose s_2' such that $r_2' - r_1 = \delta/2$. Since $r_2 - r_1 \ge \delta$, this implies $r_\infty - r_2' \ge \delta/2$. Then $s_2' \le s_3$, for some s_3 depending only on $F(\cdot)$ and δ. Since $F(\cdot)$ has DHR,

by (3.7), $\bar{F}(s_2'+M)/\bar{F}(s_2')$ is increasing in s_2', and so it suffices to demonstrate (5.8) with s_2' substituted for s_2.

One has

$$\int_0^\infty \left[\frac{\bar{F}(s_2'+t)}{\bar{F}(s_2')} - \frac{\bar{F}(s_1+t)}{\bar{F}(s_1)}\right] dt = r_2' - r_1 = \frac{\delta}{2}. \tag{5.9}$$

Choose M_0 large enough, but depending only on s_3, so that

$$\int_{M_0}^\infty \frac{\bar{F}(s_2'+t)}{\bar{F}(s_2')}\, dt \le \frac{\delta}{4}; \tag{5.10}$$

applying (3.7), one can check that this is possible.

Applying (5.9) and (5.10), one has

$$\int_0^{M_0} \left[\frac{\bar{F}(s_2'+t)}{\bar{F}(s_2')} - \frac{\bar{F}(s_1+t)}{\bar{F}(s_1)}\right] dt \ge \frac{\delta}{4}.$$

Consequently,

$$\frac{\bar{F}(s_2'+M)}{\bar{F}(s_2')} \ge \frac{\bar{F}(s_1+M)}{\bar{F}(s_1)} + \frac{\delta}{4M_0}$$

for some $M \in [0, M_0]$. The lemma follows upon setting $\epsilon = \delta/(4M_0)$. □

Proposition 5.3 obtains bounds that will be applied to the right side of (5.6) and (5.7). The proposition states that for initial states $X_i^{(N)}(0)$, $i = 1, 2$, that are not too close, with the smaller state $X_1^{(N)}(0)$ not being too large, there is a uniform lower bound on the time over which the corresponding process $X_1^{(N)}(\cdot)$ is in the 0 state, but $X_2^{(N)}(\cdot)$ is not.

Proposition 5.3 *Consider either the LL(D) supermarket model, or the SQ(D) supermarket model that is FIFO with DHR, with $\alpha \le 1$. Suppose a pair of underlying processes $X_i^{(N)}(\cdot)$, $i = 1, 2$, are coupled by the standard coupling and satisfy $X_1^{(N),n}(0) \le X_2^{(N),n}(0)$, with $\|X_1^{(N)}(0)\|_r^n \le M_1$ for given n and M_1, and $\psi^{(N),n}(0) \ge \delta$ for given $\delta > 0$. Then, for large enough M_2 and $\epsilon_1 > 0$ depending only on $F(\cdot)$, M_1 and δ,*

$$\int_0^{M_2} P\big(W_1^{(N),n}(t) = 0, W_{2,\pi}^{(N),n}(t) > 0\big)\, dt \ge \epsilon_1 \tag{5.11}$$

for the LL(D) model and, for the SQ(D) model,

$$\int_0^{M_2} P\big(Z_1^{(N),n}(t) = 0, Z_2^{(N),n}(t) > 0\big)\, dt \ge \epsilon_1. \tag{5.12}$$

Proof It is not difficult to show (5.11) by setting $M_2 = M_1 + \delta$, and considering the event on which no potential arrivals occur at n over $[0, M_2]$. On this event,

$W_1^{(N),n}(t) = 0$ but $W_{2,\pi}^{(N),n}(t) = W_2^{(N),n}(t) > 0$ on $[M_1, M_2)$. Since the probability of the event occurring is $\exp\{-\alpha D M_2\} \geq \exp\{-D M_2\}$, this implies the inequality with $\epsilon_1 = \delta \exp\{-D M_2\}$.

In order to show (5.12), we consider the cases where (A) $Z_1^{(N),n}(0) \leq Z_2^{(N),n}(0) - 1$ and where (B) $Z_1^{(N),n}(0) = Z_2^{(N),n}(0)$ and $R_1^{(N),n}(0) \leq R_2^{(N),n}(0) - \delta$ separately.

We consider the case (A) first. Setting $M_2 = 4M_1$, we consider the event A_1 over which (1) no potential arrivals occur at n over $[0, M_2]$, (2) all of the original jobs at n for the first system have departed by time $M_2/2$ and (3) at least one of the original jobs at n for the second system remains there at time M_2. The events in (1) and (2) are independent, with the first event occurring with probability at least $\exp\{-D M_2\}$. The second event occurs with probability at least $1/2$ since $\frac{1}{2}(M_2/2) \geq \|X_1^{(N)}(0)\|^n$, which is the expected service time for the original jobs there. The third event includes the event that the last original job at n for the second system requires service at least M_2, which is independent of the events in (1) and (2) and occurs with probability at least $\exp\{-M_2\}$, since $h(0) = 1$. Consequently, under (A), (5.12) follows with

$$\epsilon_1 = 2M_1 e^{-4DM_1} \cdot \frac{1}{2} \cdot e^{-4M_1} = M_1 e^{-4M_1(D+1)}.$$

We now consider the case (B). We note that $R_i^{(N),n}(0)$, $i = 1, 2$, satisfy the assumptions of Lemma 5.2 with the same δ as in the lemma. Choosing δ, ϵ and M_0 as in the lemma, it follows that, with probability at least ϵ and at some time $M \leq M_0$, the oldest original job of the first system has already been served but the oldest of the second system has not. Also, by time M_0, the probability of there being no arrivals at n in either system is at least $\exp\{-D M_0\}$.

Let T denote the time at which the oldest original job of the first system is served, and let A_2 denote the event where the oldest original job of the second system is not served at T, with $T \leq M_0$, and where, by time M_0, no arrivals at n have occurred. It follows from the previous paragraph that $P(A_2) \geq \epsilon \exp\{-D M_0\}$ and that, on A_2,

$$Z_1^{(N),n}(T) < Z_2^{(N),n}(T).$$

Hence, $X_1^{(N),n}(T) < X_2^{(N),n}(T)$, with

$$\left\|X_1^{(N)}(T)\right\|_r^n \leq \left(\left\|X_1^{(N)}(0)\right\|_r^n \wedge \left\|X_2^{(N)}(T)\right\|_r^n\right) - 1.$$

Consequently, on A_2, the assumptions of the proposition are satisfied at time T, with the data falling under case (A). Application of the bounds obtained in that case then imply that, under (B), (5.12) follows, with $M_2 = M_0 + 4M_1$ and

$$\epsilon_1 = \epsilon e^{-DM_0} M_1 e^{-4M_1(D+1)} \leq \epsilon M_1 e^{-4(M_0 \vee M_1)(2D+1)}.$$

Since (5.12) also holds in case (A) for these new choices of M_2 and ϵ_1, we can employ them there as well. This demonstrates (5.12), and hence the proposition. □

For $M, \epsilon, u > 0$, set

$$L^{X^{(N)}}_{M,\epsilon,u} = \inf\left\{ L : \frac{1}{N}\sum_{n=1}^{N} \int_u^{M+u} P\big(\big\|X^{(N)}(t)\big\|^n \geq L-1\big)\,dt \leq \epsilon \right\},$$

where $X^{(N)}(\cdot)$ is the underlying Markov process of a given supermarket model. The following proposition is a quick consequence of Lemma 5.1 and Proposition 5.3, with $\epsilon_2 = \epsilon_1/2$. It will also be used to demonstrate Proposition 5.2 as well as Proposition 8.5.

Proposition 5.4 *Consider either the LL(D) supermarket model or the SQ(D) supermarket model that is FIFO with DHR, with $\alpha \leq 1$. Suppose a pair of underlying processes $X_i^{(N)}(\cdot)$, $i = 1, 2$, satisfy $X_1^{(N)}(0) \leq X_2^{(N)}(0)$ and are coupled by the standard coupling. Then, for each $\delta > 0$ and M_1, there exists $\epsilon_2 > 0$ such that, for large enough M_2, any $u \geq 0$ and $L \geq \sup_N L^{X_2^{(N)}}_{M_2,\epsilon_2,u}$,*

$$\begin{aligned} &E\big[\psi_L^{(N)}(u+M_2)\big] - E\big[\psi_L^{(N)}(u)\big] \\ &\quad\leq -\frac{\epsilon_2}{N}\sum_{n=1}^{N} P\big(\psi^{(N),n}(u) \geq \delta, \big\|X_1^{(N)}(u)\big\|^n \leq M_1\big). \end{aligned} \tag{5.13}$$

Proposition 5.4 implies that when $X_1^{(N)}(u)$ and $X_2^{(N)}(u)$ are not too close together in the $d_r^{(N)}(\cdot,\cdot)$ metric/pseudometric and $X_1^{(N)}(u)$ is not too large, the distance between the processes in the $d_{r,L}^{(N)}(\cdot,\cdot)$ metric/pseudometric must decrease at least at a specified rate. Proposition 5.1 will follow from this and Proposition 4.1.

Proof of Proposition 5.1 The proofs for the LL(D) and the SQ(D) supermarket models are the same. We first claim that, in place of (5.1), it suffices to show, for given $\gamma > 0$ and $T > 0$, there exists q_0 not depending on N such that, for some $u^{(N)} \leq q_0$,

$$P\big(d_r^{(N),n}\big(X^{(N)}\big(u^{(N)}\big), X^{(N)}_{\mathcal{E},\pi}\big(u^{(N)}\big)\big) \geq \gamma\big) \leq \gamma \tag{5.14}$$

for all $n = 1, \ldots, N$, where $\pi_\cdot = \pi_\cdot^{(N)}$ is the permutation for the standard coupling. This is equivalent to

$$P\big(d_r^{(N),n}\big(X^{(N)}_{\phi}\big(u^{(N)}\big), X^{(N)}_{\mathcal{E}}\big(u^{(N)}\big)\big) \geq \gamma\big) \leq \gamma \tag{5.15}$$

for all $n = 1, \ldots, N$, where $\phi_\cdot = \phi_\cdot^{(N)} = (\pi_\cdot^{(N)})^{-1}$. (It is more convenient to compare $X^{(N),n}_{\phi}(u^{(N)})$ and $X^{(N),n}_{\mathcal{E}}(u^{(N)})$ here, rather than directly comparing $X^{(N),n}(u^{(N)})$ and $X^{(N),n}_{\mathcal{E},\pi}(u^{(N)})$.)

To show (5.15) suffices, let $\tilde{u}^{(N)} = q - u^{(N)}$, for given $q \geq q_0$. It follows from Lemma 4.1 that $X^{(N)}(\tilde{u}^{(N)}) \leq X^{(N)}_{\mathcal{E},\pi}(\tilde{u}^{(N)}))$ or, equivalently,

$$X^{(N)}_{\phi}\big(\tilde{u}^{(N)}\big) \leq X^{(N)}_{\mathcal{E}}\big(\tilde{u}^{(N)}\big). \tag{5.16}$$

Set $X_1^{(N)}(0) = 0$, $X_2^{(N)}(0) = X_\phi^{(N)}(\tilde{u}^{(N)})$ and $X_3^{(N)}(0) = X_{\mathcal{E}}^{(N)}(\tilde{u}^{(N)})$, and denote by $X_1^{(N)}(\cdot)$, $X_2^{(N)}(\cdot)$ and $X_3^{(N)}(\cdot)$ the corresponding processes that evolve according to the same shifted environment starting at time $\tilde{u}^{(N)}$. Since $X_1^{(N)}(0) \le X_2^{(N)}(0) \le X_3^{(N)}(0)$ because of (5.16), it follows from Lemma 4.1 that

$$X_{1,\phi_1}^{(N)}\big(u^{(N)}\big) \le X_{2,\phi_2}^{(N)}\big(u^{(N)}\big) \le X_3^{(N)}\big(u^{(N)}\big), \tag{5.17}$$

where ϕ_1 and ϕ_2 are the inverses of the permutations π_1 and π_2 corresponding to the joint standard couplings of $X_1^{(N)}(\cdot)$, $X_2^{(N)}(\cdot)$ and $X_3^{(N)}(\cdot)$.

Applying (5.17) to (5.15), with $X^{(N)}(\cdot) = X_1^{(N)}(\cdot)$, $X_{\mathcal{E}}^{(N)}(\cdot) = X_3^{(N)}(\cdot)$ and $\phi = \phi_1$, it follows that

$$\begin{aligned} &P\big(d_r^{(N),n}\big(X_{2,\phi_2}^{(N)}\big(u^{(N)}\big), X_3^{(N)}\big(u^{(N)}\big)\big) \ge \gamma\big) \\ &\quad \le P\big(d_r^{(N),n}\big(X_{1,\phi_1}^{(N)}\big(u^{(N)}\big), X_3^{(N)}\big(u^{(N)}\big)\big) \ge \gamma\big) \le \gamma \end{aligned}$$

for all $n = 1, \ldots, N$. Since $X_{2,\phi_2}^{(N)}(u^{(N)}) = X_\phi^{(N)}(q)$ and $X_3^{(N)}(u^{(N)}) = X_{\mathcal{E}}^{(N)}(q)$, where $X^{(N)}(\cdot)$ and $X_{\mathcal{E}}^{(N)}(\cdot)$ are defined on the original environment, this implies (5.1).

In order to show (5.14), we now let $X_1^{(N)}(\cdot)$ denote the process starting from the empty state and $X_2^{(N)}(\cdot)$ the process starting from the equilibrium state $\mathcal{E}_m^{(N)}$. It follows from (4.4) of Lemma 4.3 and Proposition 4.1 that, for given $\gamma > 0$,

$$\frac{1}{N}\sum_{n=1}^{N} P\big(\big\|X_1^{(N)}(u)\big\|^n > M_1\big) \le \frac{\gamma}{2} \tag{5.18}$$

for any u and large enough M_1 not depending on N.

Note that $X_1^{(N)}(\cdot)$ and $X_2^{(N)}(\cdot)$ are exchangeable in $n = 1, \ldots, N$. On account of (5.18), to demonstrate (5.14), and consequently (5.1), it therefore suffices to show that, for given $\gamma > 0$,

$$\frac{1}{N}\sum_{n=1}^{N} P\big(\psi^{(N),n}(u) \ge \gamma, \big\|X_1^{(N)}(u)\big\|^n \le M_1\big) \le \frac{\gamma}{2} \tag{5.19}$$

for some $u \le q$, with q not depending on N.

Assume now that (5.19) fails, for some K, M_2 and N, for each $u = kM_2$, $k = 0, \ldots, K$. We note that, by (4.4) of Lemma 4.3 and Proposition 4.1,

$$L = L_{M,\epsilon} \stackrel{\text{def}}{=} \sup_{N,u} L_{M,\epsilon,u}^{X_2^{(N)}} < \infty \tag{5.20}$$

for each $M, \epsilon > 0$. Choosing M_2 and ϵ_2 as in Proposition 5.4, and setting $M = M_2$, $\epsilon = \epsilon_2$ and $\delta = \gamma$, one has

$$E\big[\psi_L^{(N)}\big((k+1)M_2\big)\big] - E\big[\psi_L^{(N)}(kM_2)\big] \le -\gamma\epsilon_2/2 \tag{5.21}$$

for each $k = 0, \ldots, K$. Summing over k gives

$$E\big[\psi_L^{(N)}\big((K+1)M_2\big)\big] \le E\big[\psi_L^{(N)}(0)\big] - \gamma\epsilon_2(K+1)/2. \tag{5.22}$$

On the other hand, $\psi_L^{(N)}(0) \le L$ by definition. Therefore, the right side of (5.22) is negative for $K \ge 2L/(\gamma\epsilon_2)$, which is not possible. So (5.19) must hold for one of the above choices of u for such K. The proposition therefore follows, with $q_0 = 4LM_2/(\gamma\epsilon_2)$. □

Demonstration of Proposition 5.2 The demonstration of Proposition 5.2 is similar for the LL(D) and SQ(D) supermarket models. In both cases, we will apply Proposition 5.1 to show that (5.3) holds. We will show that, in particular, for given $\gamma_1 > 0$ and $T > 0$,

$$P\big(X^{(N),n}(t) \ne X^{(N),n}_{\mathcal{E},\pi_{q+T}}(t) \text{ for some } t \in [q+T, q+2T]\big) \le \gamma_1 \tag{5.23}$$

for all $q \ge q_0(\gamma)$ and $n = 1, \ldots, N$, with $X^{(N)}(\cdot)$ and $X^{(N)}_{\mathcal{E}}(\cdot)$ both defined on $S^{(N)}$, and where $q_0(\gamma)$ is chosen as in Proposition 5.1 for some $\gamma \ll \gamma_1$. Inequality (5.3) will follow upon setting $q_1(\gamma_1) = q_0(\gamma_0) + T$.

We begin by employing the standard coupling for each model on $S_m^{(N)}$, and denote by $\sigma_q^n = \sigma_q^{(N),n}$ the first time t, $t \ge q$, at which $X^{(N),n}(t) = X^{(N),n}_{\mathcal{E},\pi}(t) = 0$. (Recall that, for the LL(D) model, the coupling for $X^{N)}(\cdot)$ is defined on $S^{(N)}$ whereas, for the SQ(D) model, we employ $S_r^{(N)}$.) Since the queue n is empty at time σ_q^n, it is also empty then for the corresponding state of the SQ(D) model on $S^{(N)}$, given by the bijection φ at the end of Sect. 3. It follows that, for both models defined on $S^{(N)}$, the coupled queues are identical at σ_q^n.

Let $A^{(N),n}_{1,q,T} = \{\sigma_q^n \le q + T\}$. Denote by $A^{(N),n}_{2,q,T}$ the subset of $A^{(N),n}_{1,q,T}$ on which, under the standard coupling, each potential arrival over $[\sigma_q^n, q + 2T]$, at n and its coupled queue, is an arrival at both queues or at neither.

It is not difficult to see for the LL(D) model that, on $A^{(N),n}_{2,q,T}$,

$$X^{(N),n}(t) = X^{(N),n}_{\mathcal{E},\pi}(t) = X^{(N),n}_{\mathcal{E},\pi_{q'}}(t) \quad \text{for } t \in \big[\sigma_q^n, q+2T\big], \tag{5.24}$$

where $q' = \sigma_q^n$. In particular, since the states at n and its coupled queue are equal at time σ_q^n, and the service times of arriving jobs at these queues are identical, their states at future times through $q + 2T$ will be the same; note that on $A^n_{2,q,T}$, for $t \in [\sigma_q^n, q + 2T]$, $\pi^n_{\sigma_q^n} = \pi^n_t$, since the coupling does not change over the interval. For the same reasons, (5.24) also holds for the SQ(D) model on $S_r^{(N)}$. Moreover, employing the bijection $\varphi(\cdot)$, it is not difficult to check that (5.24) also holds for the SQ(D) model on $S^{(N)}$ as well.

In order to demonstrate (5.23), and hence (5.3), it therefore suffices to show

Proposition 5.5 *Assume the processes $X^{(N)}(\cdot)$ and $X^{(N)}_{\mathcal{E}}(\cdot)$ are defined as in Proposition 5.1.* (a) *For the SQ(D) supermarket model, for each $\gamma_1 > 0$ and $T > 0$, there*

exists $q_1 = q_1(\gamma_1)$ not depending on N such that, for each $q \geq q_1$,

$$P\big(\big(A_{2,q,T}^{(N),n}\big)^c\big) \leq \gamma_1 \tag{5.25}$$

for all $n = 1, \ldots, N$. (b) *Assume that, for given N and $\epsilon > 0$, (5.2) is satisfied for the LL(D) supermarket model, with $\epsilon = \epsilon(N) \to 0$ as $N \to \infty$. Then, for each $\gamma_1 > 0$ and $T > 0$, there exists $q_1 = q_1(\gamma_1)$ such that, for large enough N and $q \geq q_1$, (5.25) holds for all $n = 1, \ldots, N$.*

We first obtain upper bounds on $P((A_{1,q,T}^{(N),n})^c)$.

Lemma 5.3 *Assume the processes $X^{(N)}(\cdot)$ and $X_{\mathcal{E}}^{(N)}(\cdot)$ are defined as in Proposition 5.1. For both supermarket models and given $\gamma > 0$,*

$$P\big(\max\big(\|X^{(N)}\|^n(q), \|X_{\mathcal{E},\pi}^{(N)}\|^n(q)\big) > M\big) \leq \gamma \tag{5.26}$$

for all q and large enough M not depending on N and n. Moreover, for $T \geq 2ML$, with $L = 4e^{2DM}\log(1/\gamma)$,

$$P\big(\big(A_{1,q,T}^{(N),n}\big)^c\big) \leq 2\gamma. \tag{5.27}$$

Proof The inequality (5.26) follows from Proposition 4.1, and Lemma 4.1 applied to $X_1^{(N)}(\cdot) = X^{(N)}(\cdot)$ and $X_2^{(N)}(\cdot) = X_{\mathcal{E}}^{(N)}(\cdot)$.

To show (5.27), we first note that the probability of there being no potential arrivals, and hence no arrivals, over $(q, q + 2M]$ is at least e^{-2DM} for both models, where M is chosen as in (5.26). Denote by A the intersection of the events where this holds, but the event in (5.26) does not. Then, on A,

$$W^{(N),n}(q + 2M) = W_{\mathcal{E},\pi}^{(N),n}(q + 2M) = 0$$

for the LL(D) model, whereas for the SQ(D) model,

$$P\big(Z^{(N),n}(q + 2M) = Z_{\mathcal{E},\pi}^{(N),n}(q + 2M) = 0 \,|\, A\big) \geq \frac{1}{2},$$

by applying Markov's inequality to the expected workload at time q. Applying these displays together with (5.26), it follows that, for either model,

$$P\big(X^{(N),n}(q + 2M) = X_{\mathcal{E},\pi}^{(N),n}(q + 2M) = 0\big) \geq \frac{1}{2}(1 - \gamma)e^{-2DM}.$$

We repeat this argument at the times $M\ell$, $\ell = 1, \ldots, L - 1$, noting that $2ML \leq T$. It follows that the probability $X^{(N),n}(q + 2M\ell) = X_{\mathcal{E},\pi}^{(N),n}(q + 2M\ell) = 0$ fails at each of these times, and hence that the event $A_{1,q,T}^{(N),n}$ fails, is at most

$$\gamma + \left(1 - \frac{1}{2}e^{-2DM}\right)^L \leq \gamma + \exp\left\{-\frac{L}{2}e^{-2DM}\right\} \leq 2\gamma,$$

for $L = 4e^{2DM}\log(1/\gamma)$, which implies (5.27). □

Let $\mathcal{C}_{q,T}^{(N),n}$ denote the set of pairs (t, n'), $t \in [q, q+2T]$ and $n' = 1, \ldots, N$, $n' \neq n$, such that a potential arrival occurs at time t with selection set that includes n and n'. It is easy to see that, for any N, n, q, and T,

$$E\big[\big|\mathcal{C}_{q,T}^{(N),n}\big|\big] = 2\alpha(D-1)T. \tag{5.28}$$

This equality will be used to bound the probability of $A_{3,q,T}^{(N),n} = A_{1,q,T}^{(N),n} \cap (A_{2,q,T}^{(N),n})^c$ for both supermarket models. The argument for the SQ(D) model is simpler, so we show it first.

Lemma 5.4 *Assume that $X^{(N)}(\cdot)$, $X_{\mathcal{E}}^{(N)}(\cdot)$ and $q_0(\gamma)$ are defined as in Proposition 5.1. Then, for the SQ(D) model, for each $\gamma > 0$, $T > 0$ and $q \geq q_0(\gamma)$,*

$$P\big(A_{3,q,T}^{(N),n}\big) \leq 2\alpha DT\gamma \tag{5.29}$$

for all N and n.

Proof We first note that, for the SQ(D) model,

$$d_r^{(N),n}\big(X^{(N)}(t-), X_{\mathcal{E}}^{(N)}(t-)\big) < 1,$$

for given n, implies that $Z^{(N),n}(t-) = Z_{\mathcal{E}}^{(N),n}(t-)$, since $X^{(N)}(t-) \leq X_{\mathcal{E}}^{(N)}(t-)$ on account of the standard coupling. (Recall that $\pi_\cdot$ is the identity map here.) But, the standard coupling guarantees that an arrival cannot occur at one of the two queues $n \neq n'$ for $X^{(N)}(\cdot)$ and at the other queue for $X_{\mathcal{E}}^{(N)}(\cdot)$, at a given time t, if the coupled pair of queues, in both cases, have the same number of jobs at $t-$. So, $A_{3,q,T}^{(N),n}$ can only occur if, for some $(t, n') \in \mathcal{C}_{q,T}^{(N),n}$, $d_r^{(N),n}(X^{(N)}(t-), X_{\mathcal{E}}^{(N)}(t-)) \geq 1$. Since (5.1) is satisfied for $t \in [q, q+2T]$, it follows from (5.28) that $P(A_{3,q,T}^{(N),n}) \leq 2\alpha(D-1)T\gamma$, which implies (5.29). □

The upper bound on $P(A_{3,q,T}^{(N),n})$ requires some work. We will need to use the condition (5.2) as well as the following lemma, which gives an upper bound on the density of the equilibrium measures.

Lemma 5.5 *For the LL(D) supermarket model, with $\lambda < 1$ and any N, the equilibrium measure $\mathcal{E}^{(N)}$ satisfies*

$$\mathcal{E}^{(N)}\big(W^{(N),n} \in [c, c+\delta)\big) \leq eD\delta \tag{5.30}$$

for each n, c, and $\delta \in (0, 1/D]$.

Proof Consider the process $X^{(N)}(\cdot)$ with initial distribution $\mathcal{E}^{(N)}$. Let U_d denote the expected number of times over $(0, \delta]$ that the workload $W^{(N),n}(\cdot)$ at queue n has decreased from at least c to strictly less than c (through service) and let U_i denote the expected number of times over $(0, \delta]$ that $W^{(N),n}(\cdot)$ has increased from strictly less

than c to at least c (through the arrival of a job at n). Also, let A denote the event on which there are no potential arrivals over $(0, \delta]$ at n.

Since $X^{(N)}(0)$ and $X^{(N)}(\delta)$ have the same distribution,

$$U_d = U_i.$$

On the other hand,

$$\begin{aligned} U_d &\geq P\big(W^{(N),n}(0) \in [c, c+\delta);\, A\big) = P(A)\mathcal{E}^{(N)}\big(W^{(N),n} \in [c, c+\delta)\big) \\ &\geq e^{-1}\mathcal{E}^{(N)}\big(W^{(N),n} \in [c, c+\delta)\big) \end{aligned}$$

since $\delta \leq 1/D$, whereas

$$U_i \leq E\big[\#\text{ of potential arrivals over } (0, \delta] \text{ at } n\big] \leq D\delta.$$

It follows from the above three equations that

$$\mathcal{E}^{(N)}\big(W^{(N),n} \in [c, c+\delta)\big) \leq eD\delta,$$

as desired. □

We now bound $P(A^{(N),n}_{3,q,T})$ for the LL(D) model.

Lemma 5.6 *Assume that $X^{(N)}(\cdot)$, $X^{(N)}_{\mathcal{E}}(\cdot)$ and $q_0(\gamma)$ are defined as in Proposition 5.1 for the LL(D) supermarket model, and that (5.2) is satisfied for given N and $\epsilon > 0$, with $\epsilon = \epsilon(N) \to 0$ as $N \to \infty$. Then, for appropriate $h_T(\cdot,\cdot)$, and each $\gamma > 0$, $T > 0$ and $q \geq q_0(\gamma)$,*

$$P\big(A^{(N),n}_{3,q,T}\big) \leq h_T(\gamma, N) \tag{5.31}$$

for all $n = 1, \ldots, N$, with

$$\lim_{\gamma \searrow 0} \limsup_{N\to\infty} h_T(\gamma, N) = 0 \quad \textit{for each } T. \tag{5.32}$$

Proof For given $\gamma > 0$, suppose N is large enough so that $\epsilon(N) < \gamma$. It follows from (5.2) and Lemma 5.5 that, for each n and c,

$$\begin{aligned} &\mathcal{E}^{(N)}\big(W^{(N),n}, W^{(N),n'} \in [c-\gamma, c+\gamma]\big) \\ &\quad \leq \big(\mathcal{E}^{(N)}\big(W^{(N),n} \in [c-2\gamma, c+2\gamma)\big)\big)^2 + \gamma \\ &\quad \leq 4eD\gamma\mathcal{E}^{(N)}\big(W^{(N),n} \in [c-2\gamma, c+2\gamma)\big) + \gamma. \end{aligned} \tag{5.33}$$

Choose $K = K(\gamma)$ such that for all N and n, $\mathcal{E}^{(N)}(W^{(N),n} \geq K\gamma) \leq \gamma$. Summing (5.33) over intervals of the form $[k\gamma, (k+4)\gamma)$, it follows that

$$\begin{aligned}
&\mathcal{E}^{(N)}\big(\big|W^{(N),n} - W^{(N),n'}\big| \le \gamma\big) \\
&\quad \le 4eD\gamma \sum_{k=0}^{K-1} \mathcal{E}^{(N)}\big(W^{(N),n} \in \big[k\gamma, (k+4)\gamma\big)\big) + K\gamma + \mathcal{E}^{(N)}\big(W^{(N),n} \ge K\gamma\big) \\
&\quad \le (16eD + K + 1)\gamma. \qquad (5.34)
\end{aligned}$$

Choose q as in Proposition 5.1. Then, by the proposition,

$$P\big(W_{\mathcal{E},\pi}^{(N),n}(t) - W^{(N),n}(t) \ge \gamma \text{ or } W_{\mathcal{E},\pi}^{(N),n'}(t) - W^{(N),n'}(t) \ge \gamma\big) \le 2\gamma$$

for any pair n, n', and $t \in [q, q+2T]$. The same inequality holds with t replaced by $t-$, since the probability of an arrival at time t is 0. Together with (5.34), this implies

$$\begin{aligned}
&P\big(\big[W^{(N),n}(t-), W_{\mathcal{E},\pi}^{(N),n}(t-)\big] \cap \big[W^{(N),n'}(t-), W_{\mathcal{E},\pi}^{(N),n'}(t-)\big] \ne \emptyset\big) \\
&\quad \le (16eD + K + 3)\gamma \qquad (5.35)
\end{aligned}$$

for any pair $n \ne n'$ and $t \in [q, q+2T]$.

The standard coupling guarantees that the event $A_{3,q,T}^{(N),n}$ cannot occur at time t unless the event in (5.35) is violated for some $(t-, n')$, with $(t, n') \in \mathcal{C}_{q,T}^{(N),n}$. It therefore follows, from (5.28) and (5.35) that, for any $\epsilon(N) < \gamma$,

$$P\big(A_{3,q,T}^{(N),n}\big) \le 16\alpha DT(2eD + K + 3)\gamma. \qquad (5.36)$$

Set $h_q(\gamma, N) = 1$ if $\epsilon(N) \ge \gamma$ and set $h_q(\gamma, N)$ equal to the right side of (5.36) if $\epsilon(N) < \gamma$. Since the right side of (5.36) goes to 0 as $\gamma \searrow 0$, both (5.31) and (5.32) follow. □

The proof of Proposition 5.5 follows quickly from Lemmas 5.3, 5.4, and 5.6.

Proof of Proposition 5.5 For given $\gamma > 0$ and $T > 0$,

$$P\big(\big(A_{2,q,T}^{(N),n}\big)^c\big) \le P\big(\big(A_{1,q,T}^{(N),n}\big)^c\big) + P\big(A_{3,q,T}^{(N),n}\big).$$

It follows from Lemmas 5.3, 5.4, and 5.6 that, for the SQ(D) model, this is at most $2(1 + DT)\gamma$ and, for the LL(D) model, at most $2\gamma + h_T(\gamma, N)$, for q chosen as in Proposition 5.1 (and not depending on N or n), where

$$\lim_{\gamma \searrow 0} \limsup_{N \to \infty} h_T(\gamma, N) = 0 \quad \text{for each } T. \qquad (5.37)$$

Setting $\gamma = \gamma_1/(2(1 + DT))$ implies (5.25) for the SQ(D) model. On the other hand, (5.37) implies that for appropriate $\gamma \le \gamma_1/4$ and large enough N, $h_T(\gamma, N) \le \gamma_1/2$, which implies (5.25) for the LL(D) model as well. □

6 Uniform convergence for generalized supermarket models

In this section, we demonstrate Propositions 6.1 and 6.4. Proposition 6.1 is the analog of Proposition 5.2, but for generalized supermarket models rather than for the LL(D) and SQ(D) models; it will be employed in Sect. 8 to demonstrate the first part of Theorem 2.3. Proposition 6.4 is a modification of Proposition 6.1 that will be employed in Sect. 8 to show uniqueness of the equilibrium environment in Theorem 2.3. Recall that, for generalized supermarket models, the only requirement in the selection rule is that, after the D queues in the selection set have been chosen, the arriving job is assigned to one of these queues based only on the states at these D queues and in an exchangeable manner (that does not depend on N). Rather than requiring $\alpha < 1$ as before, we require here the stronger $\alpha \leq 1/(2\sqrt{D(D \vee \theta)})$, where θ is the second moment of the service distribution $F(\cdot)$ (which has mean 1).

We consider processes $X^{(N)}(t)$ and $X_{\mathcal{E}}^{(N)}(t)$, $t \geq 0$, corresponding to a generalized supermarket model whose initial states are the empty state and the equilibrium state $\mathcal{E}^{(N)}$, and that are coupled using a variant of the standard coupling. Here, the nth queue of $X^{(N)}(\cdot)$ will always be coupled with the nth queue of $X_{\mathcal{E}}^{(N)}(\cdot)$. We will require that, for each ω, the processes share the same arrival and selection set processes, and arriving jobs for the two processes share the same service times. When each pair of states at the coupled queues for the selection set of an arrival are the same, arrivals are assigned to queues in the same manner for each process; also, when the states at a given queue are identical for the two processes, jobs are served in the same manner in each case. At queues where the states are not identical, we allow any coupling since the choice does not affect the proof.

We will show the processes in the above coupling become close at large times. For this, we employ the following notation. Consider the queue n at time t, and set $L^{(N),n}(t) = 0$ if the coupled processes are identical at n whereas, if the processes are not identical at n, set $L^{(N),n}(t)$ equal to $\frac{1}{10}$ plus the maximum of the two workloads there. Set $L^{(N)}(t) = \sum_{n=1}^{N} L^{(N),n}(t)$. We denote by $K(t)$ the number of queues n, $n = 1, \ldots, N$, at time t at which the two processes are not identical. We refer to these queues as *discrepancies*; $K^{(N)}(t)$ is then the *number of discrepancies*.

The main result in this section is Proposition 6.1, which is the analog of Proposition 5.2. In Sect. 5, we employed the monotonicity comparisons from Sect. 4 to demonstrate Proposition 5.2. Here, we employ a martingale argument involving $K^{(N)}(\cdot)$ and $L^{(N)}(\cdot)$ to demonstrate Proposition 6.1.

Proposition 6.1 *Consider the coupled generalized supermarket model processes $X^{(N)}(\cdot)$ and $X_{\mathcal{E}}^{(N)}(\cdot)$, as given above. Assume that $\alpha \leq 1/(2\sqrt{D(D \vee \theta)})$. Then, for each $\gamma_1 > 0$ and $T > 0$, there exists $q_1 = q_1(\gamma_1)$ not depending on N such that, for $q \geq q_1$,*

$$P\big(X^{(N),n}(t) \neq X_{\mathcal{E}}^{(N),n}(t) \text{ for some } t \in [q, q+T]\big) \leq \gamma_1, \tag{6.1}$$

for all $n = 1, \ldots, N$.

Demonstration of Proposition 6.1 In order to demonstrate Proposition 6.1, we will compare the generalized supermarket model with the M/G/1 queue having arrival

rate $D\alpha$ and the same service distribution $F(\cdot)$ as the generalized supermarket model, with mean 1 and finite second moment θ. We denote by w_* the expected workload in equilibrium for this queue; $w_* < \infty$ because $\theta < \infty$. Let $W^{(N)}(\cdot)$, $W_{\mathcal{E}}^{(N)}(\cdot)$, $W_*(\cdot)$ and $W_{*,\mathcal{E}}(\cdot)$ denote the workloads corresponding to $X^{(N)}(\cdot)$, $X_{\mathcal{E}}^{(N)}(\cdot)$, the M/G/1 queue $X_*(\cdot)$ with $X_*(0) = 0$, and the M/G/1 queue in equilibrium. Recall that a family I of random variables Y_i, $i \in I$, is uniformly integrable if

$$\lim_{M\to\infty} \sup_{i\in I} E\big[|Y_i|; |Y_i| > M\big] = 0.$$

The following lemma gives upper bounds on $E[W^{(N)}(t)]$ and $E[W_{\mathcal{E}}^{(N)}(t)]$.

Lemma 6.1 *For the above $M/G/1$ queue,*

$$w_* = \alpha\theta/\big(2(1-\alpha)\big). \tag{6.2}$$

Moreover, $W^{(N),n}(t)$ is uniformly integrable over all t, N, and n, and

$$E\big[W^{(N),n}(t)\big] \le w_*, \qquad E\big[W_{\mathcal{E}}^{(N),n}(t)\big] \le w_*. \tag{6.3}$$

Proof The equality (6.2) follows, with a little computation, from the Pollaczek–Khinchin formula and Little's law.

To see (6.3), couple the process $X^{(N)}(\cdot)$, at a given queue n, with $X_*(\cdot)$, where arrivals at the M/G/1 queue are coupled to potential arrivals at queue n for $X^{(N)}(\cdot)$ so that the service times of the jobs are also the same. Then

$$W^{(N),n}(t) \le W_*(t) \quad \text{for all } t.$$

Also, under the obvious coupling,

$$W_*(t) \le W_{*,\mathcal{E}}(t) \quad \text{for all } t.$$

The first part of (6.3) and the uniform integrability of $W^{(N),n}(t)$ follow from these inequalities; the second part of (6.3) follows from these inequalities and a form of the dominated convergence theorem. □

Our main step in the demonstration of Proposition 6.1 is given by the following proposition.

Proposition 6.2 *Consider the coupled generalized supermarket model processes $X^{(N)}(\cdot)$ and $X_{\mathcal{E}}^{(N)}(\cdot)$, and the discrepancy process $K^{(N)}(\cdot)$ as above. Assume that $\alpha \le 1/(2\sqrt{D(D\vee\theta)})$. Then* (a)

$$M^{(N)}(u) \stackrel{\text{def}}{=} L^{(N)}(u) + \frac{1}{2}\alpha^2 \int_0^u K^{(N)}(t)\,dt \tag{6.4}$$

is a supermartingale with respect to the filtration generated by $X^{(N)}(\cdot)$ and $X_{\mathcal{E}}^{(N)}(\cdot)$, and (b) *for each $\gamma > 0$, there exists $u^{(N)} = u^{(N)}(\gamma)$, with $u^{(N)} \le \theta/\alpha^2\gamma^2$, such that*

$$P\big(K^{(N)}\big(u^{(N)}\big)/N \ge \gamma\big) \le \gamma. \tag{6.5}$$

Proof The reasoning for part (b) is quick, if one assumes part (a). Since $X^{(N)}(0) = 0$, it follows from (6.3) that

$$E\big[M^{(N)}(0)\big] \le N\left(w_* + \frac{1}{10}\right). \tag{6.6}$$

By the optional sampling theorem, for each $u > 0$,

$$\frac{1}{2}\alpha^2 E\left[\int_0^u K^{(N)}(t)\,dt\right] \le E\big[M^{(N)}(u)\big] \le E\big[M^{(N)}(0)\big] \le N\left(w_* + \frac{1}{10}\right).$$

It follows that, for $u \ge 2(w_* + \frac{1}{10})/(\alpha^2\gamma^2)$,

$$P\big(K^{(N)}\big(u^{(N)}\big)/N \ge \gamma\big) \le \gamma \quad \text{for some } u^{(N)} \in [0, u].$$

By (6.2), this bound on u is at most $2\theta/(\alpha\gamma^2) + 1/(10\alpha^2\gamma^2) \le \theta/(\alpha^2\gamma^2)$, and so (6.5) follows.

In order to show $M^{(N)}(\cdot)$ is a supermartingale, it suffices to show the infinitesimal generator of the pair $(X^{(N)}(\cdot), X_{\mathcal{E}}^{(N)}(\cdot))$ applied to $L^{(N)}(t)$ is at most $-\frac{1}{2}\alpha^2 K^{(N)}(t)$ at each t and then to apply Dynkin's formula. To obtain the bound, we first claim that $L^{(N)}(t)$ decreases at rate at least

$$K^{(N)}(t) \tag{6.7}$$

due to the performed service, and that it increases at rate at most

$$\alpha D(1.1 + w_*)K^{(N)}(t) \tag{6.8}$$

due to arrivals.

The bound in (6.7) is clear. For the bound given by (6.8), note that a discrepancy can only be created or increased at a queue when a potential arrival occurs at the queue and there already is a discrepancy at one of the queues in the corresponding selection set. This implies that discrepancies are created or increased in the system at rate at most

$$\alpha D K^{(N)}(t) \tag{6.9}$$

in the system. On the other hand, when this occurs at a queue n, $L^{(N),n}(t)$ increases by at most $Y + W^{(N),n}(t) + \frac{1}{10}$ when the discrepancy is created, and by Y when the discrepancy is increased, where Y is an independent random variable having distribution $F(\cdot)$. (Note that when there is no discrepancy at n at time t, $W^{(N),n}(t) = W_{\mathcal{E}}^{(N),n}(t)$.) Taking expectations and employing (6.3), the expected increase in the workload will be at most $1.1 + w_*$. Multiplication of this by the bound in (6.9) produces the desired bound in (6.8) on the rate of increase due to arrivals. Subtracting the bounds in (6.8) and (6.7) shows that the rate of change of $L^{(N)}(t)$ is at most

$$\big(\alpha D(1.1 + w_*) - 1\big)K^{(N)}(t).$$

The demonstration of part (a) will be complete once we show that

$$\alpha D(1.1 + w_*) - 1 + \frac{1}{2}\alpha^2 \le 0. \tag{6.10}$$

Since $\alpha \le 1/(2\sqrt{D(D\vee\theta)})$ and $D \ge 2$, one automatically has $\alpha \le 1/4$ and $\alpha D \le 1/2$. Because of (6.2) and $\theta \ge 1$, the left side of (6.10) equals

$$\alpha D\left(1.1 + \frac{\alpha\theta}{2(1-\alpha)}\right) - 1 + \frac{1}{2}\alpha^2 \le \alpha D(1.1+\alpha\theta) - 1 \le \alpha D(1+\alpha\theta) - 0.95.$$

It therefore suffices to check that $\alpha D(1+\alpha\theta) - 0.95 \le 0$, which is equivalent to

$$\alpha \le \frac{1}{2\theta}\left[\sqrt{1+4(0.95)\theta/D} - 1\right]. \tag{6.11}$$

One can show (6.11) by considering the cases $4\theta/D \le 2.5$ and $4\theta/D > 2.5$ separately. Setting $f(x) = \sqrt{1+x}$, with $x = 4\theta/D \le 2.5$, one uses $f'(x) \ge 0.265$ for $x \in [0, 2.5]$ whereas, for $x = 4\theta/D > 2.5$, one uses $f(x) - 1 \ge 0.55\sqrt{x}$. Employing $\alpha \le 1/(2\sqrt{D(D\vee\theta)})$, the bound follows in both cases. □

The following somewhat stronger version of Proposition 6.2 follows with a bit of work.

Proposition 6.3 *Under the same assumptions as in Proposition 6.2, for each $\gamma > 0$ and some $q_0 = q_0(\gamma)$ not depending on N,*

$$P\left(K^{(N)}(t)/N > \gamma\right) \le \gamma \tag{6.12}$$

for all $t \ge q_0$.

Proof It follows from Lemma 6.1 that $L^{(N),n}(t)$ is uniformly integrable over all t, N and n. In particular, for each $\epsilon > 0$, there exists a $\delta > 0$ so that, for A with $P(A) < \delta$, $E[L^{(N),n}(t); A] < \epsilon$ (see, e.g., Chung [7], p. 96). On account of the second part of Proposition 6.2, for given $\gamma' > 0$,

$$P\left(L^{(N),n}\left(u^{(N)}(\gamma')\right) \ne 0\right) \le 2\gamma'$$

for each N and n. It follows from this and the uniform integrability of $W^{(N),n}(t)$, $t \in \mathbb{R}_+$, that, for given $\epsilon > 0$ and small enough γ' not depending on N,

$$E\left[L^{(N),n}\left(u^{(N)}(\gamma')\right)\right] \le \epsilon.$$

It therefore follows from the first part of Proposition 6.2 that, for $t \ge u^{(N)}$,

$$E\left[L^{(N),n}(t)\right] \le \epsilon$$

for each N and n. Summing over n, since $L^{(N),n}(t) \ge \frac{1}{10}$ when $L^{(N),n}(t) \ne 0$, one obtains

$$E\left[K^{(N)}(t)/N\right] \le 10\epsilon.$$

Hence,

$$P\left(K^{(N)}(t)/N \ge \gamma\right) \le \gamma$$

for $\gamma = \sqrt{10\epsilon}$, which implies (6.12). □

We now employ Proposition 6.3 to complete the proof of Proposition 6.1. The reasoning is similar, but simpler, than that employed in the proof of Proposition 5.2 for the SQ(D) model.

Proof of Proposition 6.1 For given q, N and n, we set $B_{1,q}^{(N),n} = \{L^{(N),n}(q) = 0\}$. Denote by $B_{2,q,T}^{(N),n}$ the subset of $B_{1,q}^{(N),n}$ on which, under the standard coupling, each potential arrival over $[q, q+T]$, for the coupled queues at n, is an arrival at both queues or at neither. As in (5.24), it is not difficult to see that, on $B_{2,q,T}^{(N),n}$,

$$X^{(N),n}(t) = X_{\mathcal{E}}^{(N),n}(t) \quad \text{for } t \in [q, q+T]. \tag{6.13}$$

In order to demonstrate (6.1), it therefore suffices to show that, for each $\gamma_1 > 0$, there exists $q_1(\gamma_1)$ not depending on N such that, for each $q \geq q_1(\gamma_1)$,

$$P\big(\big(B_{2,q,T}^{(N),n}\big)^c\big) \leq \gamma_1 \tag{6.14}$$

for all n.

We note that, by (6.12) with $t = q$ and $q \geq q_0(\gamma)$,

$$P\big(\big(B_{1,q}^{(N),n}\big)^c\big) \leq 2\gamma. \tag{6.15}$$

On the other hand, denoting by $\mathcal{C}_{q,T}^{(N),n}$ the set of pairs (t, n'), with $t \in [q, q+T]$ and $n' \neq n$, such that a potential arrival occurs at time t with selection set that includes n and n', then, for any N, n, q, and T,

$$E\big[\big|\mathcal{C}_{q,T}^{(N),n}\big|\big] = \alpha(D-1)T, \tag{6.16}$$

which is the analog of (5.28). It thus follows that, as in Lemma 5.4, for $q \geq q_0(\gamma)$ with $q_0(\gamma)$ as in Proposition 6.3,

$$P\big(B_{1,q}^{(N),n} \cap \big(B_{2,q,T}^{(N),n}\big)^c\big) \leq DT\gamma. \tag{6.17}$$

Together with (6.15), (6.17) implies that

$$P\big(\big(B_{2,q,T}^{(N),n}\big)^c\big) \leq (2+DT)\gamma.$$

Setting $\gamma = \gamma_1/(2+DT)$ implies (6.14) with $q_1(\gamma_1) = q_0(\gamma)$. □

Statement and demonstration of Proposition 6.4 In order to show uniqueness of the equilibrium environment in Sect. 8 for generalized supermarket models, we will employ a variant of Proposition 6.1, with $X^{(N)}(\cdot)$ being compared with $X_{\tilde{\mathcal{E}}}^{(N)}(\cdot)$ rather than with $X_{\mathcal{E}}^{(N)}(\cdot)$, where $X_{\tilde{\mathcal{E}}}^{(N)}(\cdot)$ is the process whose initial state has i.i.d. coordinates, with the distribution being given by some equilibrium environment $\tilde{\mathcal{E}}$ for the generalized supermarket model. Note that, unlike $X_{\mathcal{E}}^{(N)}(\cdot)$, the distribution of $X_{\tilde{\mathcal{E}}}^{(N)}(\cdot)$ is not constant over time.

Proposition 6.4 *Consider coupled generalized supermarket model processes* $X^{(N)}(\cdot)$ *and* $X^{(N)}_{\tilde{\mathcal{E}}}(\cdot)$, *as given above. Assume that* $\alpha \le 1/(2\sqrt{D(D \vee \theta)})$. *Then, for each* $\gamma_1 > 0$ *and* $T > 0$, *there exists* $q_1 = q_1(\gamma_1)$, *not depending on* N, *such that, for* $q \ge q_1$,

$$P\big(X^{(N),n}(t) \ne X^{(N),n}_{\tilde{\mathcal{E}}}(t) \text{ for some } t \in [q, q+T]\big) \le \gamma_1, \tag{6.18}$$

for all $n = 1, \ldots, N$.

The demonstration of Proposition 6.4 is very similar to that of Proposition 6.1 on account of the bounds given in the following lemma. Here, $X^{\tilde{\mathcal{E}}}(\cdot)$ denotes the stationary cavity process on S with environment $\tilde{\mathcal{E}}$ and $X_{*,\mathcal{E}}(\cdot)$ denotes the stationary process for the corresponding M/G/1 queue, with $X^{\tilde{\mathcal{E}}}$ and $X_{*,\mathcal{E}}$ being used for random vectors with the corresponding equilibrium measures. Also, $X_{*,\tilde{\mathcal{E}}}(\cdot)$ denotes the process for the M/G/1 queue with initial distribution given by $\tilde{\mathcal{E}}$. We let $W^{\tilde{\mathcal{E}}}(\cdot)$, $W_{*,\mathcal{E}}(\cdot)$, etc., denote the corresponding workloads.

Lemma 6.2 *If* $\alpha < 1$, *then every equilibrium environment* $\tilde{\mathcal{E}}$ *of a generalized supermarket model satisfies*

$$P\big(W^{\tilde{\mathcal{E}}} \ge y\big) \le P(W_{*,\mathcal{E}} \ge y) \quad \text{for all } y. \tag{6.19}$$

Moreover,

$$P\big(W^{(N),n}_{\tilde{\mathcal{E}}}(t) \ge y\big) \le P(W_{*,\mathcal{E}} \ge y) \quad \text{for all } y \tag{6.20}$$

and all t, N *and* $n = 1, \ldots, N$, *and hence*

$$E\big[W^{(N),n}_{\tilde{\mathcal{E}}}(t)\big] \le w_* \quad \text{for all } t. \tag{6.21}$$

Proof For given n, couple $X^{\tilde{\mathcal{E}}}(\cdot)$ and $X_{*,\tilde{\mathcal{E}}}(\cdot)$ so that

$$W^{\tilde{\mathcal{E}}}(t) \le W_{*,\tilde{\mathcal{E}}}(t) \quad \text{for all } t. \tag{6.22}$$

This is possible since $X^{\tilde{\mathcal{E}}}(0)$ and $X_{*,\tilde{\mathcal{E}}}(0)$ have the same distribution, and arrivals for $X_{*,\tilde{\mathcal{E}}}(\cdot)$ can be coupled with potential arrivals for $X^{\tilde{\mathcal{E}}}(\cdot)$. Since $\alpha < 1$, the M/G/1 queue is positive recurrent, and so

$$W_{*,\tilde{\mathcal{E}}}(t) \xrightarrow{\mathcal{D}} W_{*,\mathcal{E}}. \tag{6.23}$$

Therefore, since $\tilde{\mathcal{E}}(t)$ is invariant in t for the cavity process, (6.19) follows from (6.22) and (6.23).

For given N and n, we now couple $X^{(N),n}_{\tilde{\mathcal{E}}}(\cdot)$ with $X_{*,\mathcal{E}}(\cdot)$ so that

$$W^{(N),n}_{\tilde{\mathcal{E}}}(t) \le W_{*,\mathcal{E}}(t) \quad \text{for all } t. \tag{6.24}$$

On account of (6.19), this is possible at $t = 0$ and, by coupling arrivals for $X_{*,\mathcal{E}}(\cdot)$ with potential arrivals at n for $X_{\tilde{\mathcal{E}}}^{(N),n}(\cdot)$, (6.24) holds for all t. This implies (6.20). Inequality (6.21) follows immediately from (6.20). □

Proof of Proposition 6.4 The argument is the same as that for Proposition 6.1, with the only difference being that (6.21) of Lemma 6.2 is used to justify the analogs of (6.6) and (6.3), which are employed in the proofs of part (b) of Proposition 6.1 and in Proposition 6.3, respectively. The rest of the argument for Propositions 6.1 and 6.3, and the proof of Proposition 6.1 itself are not affected, since the transition rules of the processes are the same. □

7 Local independence for small times

In Sects. 5 and 6, we demonstrated the convergence, as $t \to \infty$, of the processes $X^{(N)}(t)$, with $X^{(N)}(0) = 0$, that underly the different supermarket models. In this section, for fixed T and N', we demonstrate the convergence, as $N \to \infty$, of the restriction of $X^{(N)}(t)$, $t \in [0, T]$, to the coordinates $1, \ldots, N'$, for all generalized supermarket models. For this, we employ the convergence $\xrightarrow{v}$ on $[0, T] \times S^{(N')}$ that was introduced at the end of Sect. 3. Rather than assuming $X^{(N)}(0) = 0$, we will assume that $X^{(N),n}(0)$ is i.i.d. over $n = 1, \ldots, N$, and that the corresponding distribution does not depend on N. The results in this section are related to those in Graham [11] and Graham and Méléard [12] on the propagation of chaos.

Most of the section is devoted to demonstrating the following result. At the end of the section, we will justify the inequality (5.2) that was used in Proposition 5.2 for the LL(D) supermarket model.

Proposition 7.1 *Suppose that for the processes $X^{(N)}(\cdot)$, $N \in \mathbb{Z}^+$, underlying a generalized supermarket model, $X^{(N)}(0)$ is i.i.d., and the distribution does not depend on N. For $T > 0$ and $N' \leq N$, let $\mathcal{M}_T^{(N,N')}$ denote the probability measures on $[0, T] \times S^{(N')}$ induced by the first N' coordinates of $X^{(N)}(t)$, with $t \in [0, T]$. Then*

$$\mathcal{M}_T^{(N,N')} \xrightarrow{v} \mathcal{M}_T^{(\infty,N')} \quad \text{as } N \to \infty, \tag{7.1}$$

for some probability measure $\mathcal{M}_T^{(\infty,N')}$, where $\mathcal{M}_T^{(\infty,N')}$ is the N'-fold product of $\mathcal{M}_T^{(\infty,1)}$.

In order to demonstrate (7.1), we reinterpret $X^{(N)}(\cdot)$ in terms of a branching process in reversed time. For this, we employ $\mathcal{I}_T^{(N,N')}(u)$, for $u \in [0, T]$ and $N' \leq N$, which we refer to as the *influence process* of $\{1, \ldots, N'\}$; $\mathcal{I}_T^{(N,N')}(\cdot)$ is the right continuous, piecewise constant process on subsets of $\{1, \ldots, N\}$, with $\mathcal{I}_T^{(N,N')}(0) = \{1, \ldots, N'\}$, that is nondecreasing and can increase at a time u only if there is a potential arrival at time $t = T - u$ and queue n, with $n \in \mathcal{I}_T^{(N,N')}(u-)$, in which case we set

$$\mathcal{I}_T^{(N,N')}(u) = \mathcal{I}_T^{(N,N')}(u-) \cup A, \tag{7.2}$$

where A is the selection set containing n. Also, denote by $t_1 < t_2 < \cdots < t_K$ the (random) arrival times in the system over $(0, T]$, set $u_k = T - t_k$, for $k = 1, \ldots, K$, and denote by B_k the corresponding selection sets. (We will exclude realizations where arrival times are not distinct, which only occur on a set of probability 0.) We refer to the selection sets B_k with $\mathcal{I}_T^{(N,N')}(u_k-) \cap B_k \neq \emptyset$ as *intersecting selection sets* for the triple (T, N, N'); this condition is equivalent to $B_k \subseteq \mathcal{I}_T^{(N,N')}(u_k)$.

For $t \in [0, T]$ and $n' \leq N'$, $X^{(N),n'}(t)$ is determined by the intersecting selection sets for (T, N, N'), the service times of the corresponding arrivals and the initial values $X^{(N),n}(0)$, with $n \in \mathcal{I}_T^{(N,N')}(T)$. One can check this by arguing inductively going backward in time starting at time T, and first noting that, if B_{k_1} is the last selection set containing a given $n' \leq N'$ before time t, then $X^{(N),n'}(t)$ is determined by $X^{(N),n_1}(t_{k_1})$, for $n_1 \in B_{k_1}$, and the service time of the arrival there. We refer to the intersecting selection sets together with the service times of the corresponding arrivals and the above values of $X^{(N),n}(0)$ as the *underlying data* for (T, N, N').

Set $I_T^{(N,N')}(u) = |\mathcal{I}_T^{(N,N')}(u)|$. One can couple $I_T^{(N,N')}(\cdot)$ with a continuous time D-ary branching process $I_{T,\infty}^{(N,N')}(\cdot)$ that branches at rate αD for each parent, with $I_{T,\infty}^{(N,N')}(0) = N'$, so that

$$I_T^{(N,N')}(u) \leq I_{T,\infty}^{(N,N')}(u) \quad \text{for all } u \in [0, T]. \tag{7.3}$$

In this coupling, when B_k is an intersecting selection set,

$$I_{T,\infty}^{(N,N')}(u_k) = I_{T,\infty}^{(N,N')}(u_k-) + D - 1;$$

$I_{T,\infty}^{(N,N')}(\cdot)$ also increases at times corresponding to births for the additional parents not included in $\mathcal{I}_T^{(N,N')}(\cdot)$. Note that $I_T^{(N,N')}(T) \neq I_{T,\infty}^{(N,N')}(T)$ only when, for some k,

$$q(u_k) \stackrel{\text{def}}{=} \left|B_k \cap \mathcal{I}_T^{(N,N')}(u_k-)\right| \geq 2. \tag{7.4}$$

We then say the selection set B_k is *deficient* with *deficiency* ℓ if $q(u_k) = \ell + 1$.

One has the following bounds on the first and second moments of $I_{T,\infty}^{(N,N')}(T)$.

Lemma 7.1 *For the above process* $I_{T,\infty}^{(N,N')}(\cdot)$,

$$E\left[I_{T,\infty}^{(N,N')}(T)\right] = N' \exp\left\{\alpha(D-1)DT\right\} \tag{7.5}$$

and

$$E\left[\left(I_{T,\infty}^{(N,N')}(T)\right)^2\right] \leq 2D\left(N'\right)^2 \exp\left\{2(D-1)DT\right\}. \tag{7.6}$$

Proof Since $I_{T,\infty}^{(N,N')}(\cdot)$ is a continuous time D-ary branching process that branches at rate αD per parent, with $I_{T,\infty}^{(N,N')}(0) = N'$, (7.5) is a standard result from branching process theory (see, e.g., Athreya and Ney [2] or Harris [13]). Equation (7.6) follows

after some computation using generating functions, as in Theorem 6.1 on page 103 of [13]. □

Set $G_T^{(N,N')} = \{\omega : I_T^{(N,N')}(T) \neq I_{T,\infty}^{(N,N')}(T)\}$. The following proposition shows that, for given T, N', and large N, the event $G_T^{(N,N')}$ is unlikely.

Proposition 7.2 *For $G_T^{(N,N')}$ as defined above,*

$$P\big(G_T^{(N,N')}\big) \leq \frac{2D(N')^2}{N} \exp\big\{2(D-1)DT\big\}. \tag{7.7}$$

Proof We will show that

$$P\big(G_T^{(N,N')} \,\big|\, \sigma\big(I_{T,\infty}^{(N,N')}(T)\big)\big) \leq \big(I_{T,\infty}^{(N,N')}(T)\big)^2/N. \tag{7.8}$$

Employing (7.8), together with (7.6), implies that

$$P\big(G_T^{(N,N')}\big) \leq \frac{1}{N} E\big[\big(I_{T,\infty}^{(N,N')}(T)\big)^2\big] \leq \frac{2D(N')^2}{N} \exp\big\{2(D-1)DT\big\},$$

as claimed.

In order to show (7.8), we first note that $G_T^{(N,N')}$ can only occur if the event in (7.4) occurs for some k. Moreover, conditioning on $n \subseteq B_k \cap \mathcal{I}_T^{(N,N')}(u_k-)$, for given n, the other $D-1$ queues in B_k occur with equal probability. Abbreviating $I = I_{T,\infty}^{(N,N')}(T)$, some thought therefore shows that the left side of (7.8) is dominated by the probability that, starting from N' distinct queues and sequentially choosing each of the remaining $I - N'$ queues randomly from among all N queues, at least two of the I queues are the same. This latter probability can be written as

$$1 - \prod_{j=0}^{I-N'-1} \big(1 - \big(N' + j\big)/N\big) \leq 1 - (1 - I/N)^I \leq I^2/N, \tag{7.9}$$

which implies (7.8). □

One can extend the process $\mathcal{I}_T^{(N,N')}(\cdot)$ to a process $\mathcal{I}_{T,\infty}^{(N,N')}(\cdot)$ defined on subsets of $\mathbb{Z}^+$ (rather than $\{1,\ldots,N\}$), with

$$\mathcal{I}_T^{(N,N')}(u) \subseteq \mathcal{I}_{T,\infty}^{(N,N')}(u), \qquad \mathcal{I}_T^{(N,N')}(u) = \mathcal{I}_{T,\infty}^{(N,N')}(u) \cap \{1,\ldots,N\}$$

and $|\mathcal{I}_{T,\infty}^{(N,N')}(u)| = I_{T,\infty}^{(N,N')}(u)$, for $u \in [0,T]$, where $\mathcal{I}_{T,\infty}^{(N,N')}(u)$ is nondecreasing in u, and $I_{T,\infty}^{(N,N')}(\cdot)$ is the branching process defined earlier. We assume that the restriction of $\mathcal{I}_{T,\infty}^{(N,N')}(\cdot)$ to $\{1,\ldots,N\}$ satisfies the analog of (7.2) for the intersecting selection sets. When the intersecting selection set at $u = u_k$ has deficiency ℓ and

$$\mathcal{I}_{T,\infty}^{(N,N')}(u-) - \{1,\ldots,N\} = \{N+1,\ldots,N+L\}, \tag{7.10}$$

for some L, we set

$$\mathcal{I}_{T,\infty}^{(N,N')}(u) = \mathcal{I}_{T}^{(N,N')}(u) \cup \{N+1,\ldots,N+L+\ell\}. \tag{7.11}$$

We furthermore endow each $n \in \mathbb{Z}^+ - \{1,\ldots,N\}$ with a rate-αD Poisson point process; when an event occurs at time u and site $n \in \mathcal{I}_{T,\infty}^{(N,N')}(u-) - \{1,\ldots,N\}$, with (7.10) holding for some L, we define $\mathcal{I}_{T,\infty}^{(N,N')}(u)$ as in (7.11), with $\ell = D-1$. One can check that this construction for $\mathcal{I}_{T,\infty}^{(N,N')}(\cdot)$ satisfies the properties given at the beginning of the paragraph. In essence, $\mathcal{I}_{T,\infty}^{(N,N')}(\cdot)$ extends $\mathcal{I}_{T}^{(N,N')}(\cdot)$ to $\mathbb{Z}^+$ from $\{1,\ldots,N\}$ so that the additional sites that are added in $\mathbb{Z}^+ - \{1,\ldots,N\}$ ensure that $|\mathcal{I}_{T,\infty}^{(N,N')}(\cdot)|$ is a branching process.

We also extend $X^{(N),n}(0)$ from $n = 1,\ldots,N$ to $n \in \mathbb{Z}^+$ so that $X^{(N),n}(0)$ are i.i.d.

By employing the same construction as outlined in the paragraph beginning below (7.2), the process $\mathcal{I}_{T,\infty}^{(N,N')}(\cdot)$, the service times at intersecting selection sets (including the sets intersecting $\mathbb{Z}^+ - \{1,\ldots,N\}$, which we also refer to as selection sets), and $X^{(N),n}(0)$, with $n \in \mathcal{I}_{T,\infty}^{(N,N')}(T)$, together define a right continuous process $X_{T,\infty}^{(N,N')}(t)$ with left limits and taking values in $S^{(N')}$, for $t \in [0,T]$. Note that, on $(G_T^{(N,N')})^c$,

$$X_{T,\infty}^{(N,N'),n}(t) = X_T^{(N),n}(t) \quad \text{for } t \in [0,T],\ n \le N'. \tag{7.12}$$

We denote by $\mathcal{M}_{T,\infty}^{(N,N')}$ the probability measure on $[0,T] \times S^{(N')}$ induced by $X_{T,\infty}^{(N,N')}(t)$, for $t \in [0,T]$. The following lemma shows $\mathcal{M}_{T,\infty}^{(N,N')}$ does not depend on N.

Lemma 7.2 *For the measures $\mathcal{M}_{T,\infty}^{(N,N')}$ defined above and $N_1, N_2 \ge N'$,*

$$\mathcal{M}_{T,\infty}^{(N_1,N')} = \mathcal{M}_{T,\infty}^{(N_2,N')}. \tag{7.13}$$

Proof The influence processes $\mathcal{I}_{T,\infty}^{(N_i,N')}(\cdot)$, $i = 1,2$, each induce the same continuous time D-ary branching process that was introduced before (7.3), with both branching processes starting with N' ancestors. The processes $\mathcal{I}_{T,\infty}^{(N_i,N')}(\cdot)$ include corresponding ancestral trees, where the line of descent of each individual is explicitly given; the labelling of the coordinates for the respective trees differs for $N_1 \ne N_2$.

One can couple the influence processes so that, for all $u \in [0,T]$ and $n \in \mathbb{Z}^+$,

$$\mathcal{I}_{T,\infty}^{(N_1,N')}(u) \cap \{\pi_1^n\} = \mathcal{I}_{T,\infty}^{(N_2,N')}(u) \cap \{\pi_2^n\}, \tag{7.14}$$

where the random permutations $\pi_i : \mathbb{Z}^+ \to \mathbb{Z}^+$, $i = 1,2$, are defined so that $\pi_i^1, \pi_i^2, \ldots$ orders the queues according to their times of inclusion in $\mathcal{I}_{T,\infty}^{(N_i,N')}(\cdot)$, with lower indexed queues being ordered first in case of ties; note that $\pi_i^n = n$ for $n \le N'$. Service times of arrivals and $X^{(N_i),n}(0)$ can also be coupled since they are each are i.i.d., and are independent of $\mathcal{I}_{T,\infty}^{(N_i,N')}(\cdot)$ and each other.

Denote by $\tilde{\mathcal{I}}_{T,\infty,\pi}^{(N_i,N')}(\cdot)$ and $\tilde{X}_{T,\infty,\pi}^{(N_i),n}(0)$ the ancestral trees and initial data obtained by permuting $\mathcal{I}_{T,\infty}^{(N_i,N')}(\cdot)$ and $X^{(N_i),n}(0)$ according to π_i, and denote by $\tilde{X}_{T,\infty,\pi}^{(N_i,N')}(\cdot)$ the processes induced by them and the permuted service times. By the above coupling, $\tilde{X}_{T,\infty,\pi}^{(N_1,N')}(\cdot)$ and $\tilde{X}_{T,\infty,\pi}^{(N_2,N')}(\cdot)$ have the same law. Moreover, since the policy dictating which queue an arriving job selects does not depend on the labelling of the queues and since $\pi_i^n = n$ for $n \le N'$, $X_{T,\infty}^{(N_i,N')}(\cdot)$ and $\tilde{X}_{T,\infty,\pi}^{(N_i,N')}(\cdot)$, $i = 1, 2$, have the same law. Consequently, $X_{T,\infty}^{(N_1,N')}(\cdot)$ and $X_{T,\infty}^{(N_2,N')}(\cdot)$ have the same law, which implies (7.13). □

Since $\mathcal{M}_{T,\infty}^{(N,N')}$ does not depend on N, we drop the index and set $\mathcal{M}_T^{(\infty,N')} = \mathcal{M}_{T,\infty}^{(N,N')}$.

Proposition 7.1 follows quickly from Proposition 7.2 and Lemma 7.2.

Proof of Proposition 7.1 On $(G_T^{(N,N')})^c$, for given $T > 0$ and $N' \le N$,

$$\mathcal{I}_T^{(N,N')}(u) = \mathcal{I}_{T,\infty}^{(N,N')}(u) \quad \text{for } u \in [0,T],$$

and hence

$$\bigl\{X^{(N),n}(0), n \in \mathcal{I}_T^{(N,N')}(T)\bigr\} = \bigl\{X^{(N),n}(0), n \in \mathcal{I}_{T,\infty}^{(N,N')}(T)\bigr\}.$$

So, by Proposition 7.2 and Lemma 7.2, for any set $A \in B_T^{(N')}$,

$$\begin{aligned}\bigl|\mathcal{M}_T^{(N,N')}(A) - \mathcal{M}_T^{(\infty,N')}(A)\bigr| &= \bigl|\mathcal{M}_T^{(N,N')}(A) - \mathcal{M}_{T,\infty}^{(N,N')}(A)\bigr| \\ &\le \frac{2D(N')^2}{N}\exp\bigl\{2(D-1)DT\bigr\}.\end{aligned}$$

Since the last quantity converges to 0 as $N \to \infty$, this implies (7.1).

We still need to show that $\mathcal{M}_T^{(\infty,N')}$ is the N'-fold product of $\mathcal{M}_T^{(\infty,1)}$, for which we employ the influence process $\mathcal{I}_{T,\infty}^{(N,N')}(\cdot)$ from the proof of Lemma 7.2, with $N = N'$. The evolutions of the different subtrees emanating from each of the N' original ancestors are independent (although the labelling is not). Since the service times of arrivals for the different subtrees are also independent, as are the values $X^{(N'),n}(0)$ taken at the (nonoverlapping) sites $n \in \mathbb{Z}^+$ corresponding to each subtree, it follows that the processes $X_{T,\infty}^{(N',N'),n'}(t)$, $t \in [0,T]$, are independent over $n' = 1, \ldots, N'$. Since the subtrees corresponding to different $n' \in N'$ have the same law, as do the corresponding service times and values $X^{(N'),n}(0)$, $X_{T,\infty}^{(N',N'),n'}(\cdot)$ has the same law as $X_{T,\infty}^{(N',N'),1}(\cdot)$, for each $n' \le N'$. It follows that $\mathcal{M}_{T,\infty}^{(N',N')}$, and hence $\mathcal{M}_T^{(\infty,N')}$, has the desired properties. □

In Proposition 5.2, we required the condition (5.2) for the LL(D) supermarket model. By employing Proposition 7.1 together with Proposition 5.1, (5.2) follows without difficulty.

Proposition 7.3 *The inequality* (5.2) *is satisfied for the LL(D) supermarket model, with* $\epsilon = \epsilon(N) \to 0$ *as* $N \to \infty$.

Proof It suffices to show (5.2) for $n_1 = 1$ and $n_2 = 2$. Let $\mathcal{M}^{(N,N')}(t)$ denote the probability measure on $(S^{(N')}, \mathscr{S}^{(N')})$ at time t induced by the first N' coordinates of $X^{(N)}(t)$ and, for $B \in \mathscr{S}$, denote by B^δ the set of points in S within distance $\delta > 0$ of B, according to the $d_r^{(1),1}(\cdot,\cdot)$ pseudometric given by (3.9). It follows from Proposition 5.1 that, for given $\epsilon > 0$ and large enough t not depending on N,

$$\begin{aligned} &\mathcal{E}^{(N,2)}(B_1 \times B_2) \le \mathcal{M}^{(N,2)}\big(t; B_1^{\epsilon/2} \times B_2^{\epsilon/2}\big) + \epsilon/4, \\ &\mathcal{M}^{(N,1)}\big(t; B_i^{\epsilon/2}\big) \le \mathcal{E}^{(N,1)}\big(B_i^{\epsilon}\big) + \epsilon/4, \end{aligned} \tag{7.15}$$

for all $B_i \in \mathscr{S}$, $i = 1, 2$, where $\mathcal{M}^{(N,N')}(t; B)$ denotes the measure of $B \in \mathscr{S}^{(N')}$ with respect to $\mathcal{M}^{(N,N')}(t)$. (By Lemma 4.2, $\mathcal{E}_\pi(t) = \mathcal{E}(t)$ for each t.)

On the other hand, by Proposition 7.1, for given $\epsilon > 0$ and t, and large enough N,

$$\big|\mathcal{M}^{(N,2)}\big(t; B_1^{\epsilon/2} \times B_2^{\epsilon/2}\big) - \mathcal{M}^{(N,1)}\big(t; B_1^{\epsilon/2}\big) \cdot \mathcal{M}^{(N,1)}\big(t; B_2^{\epsilon/2}\big)\big| \le \epsilon/4. \tag{7.16}$$

Together with (7.15), this implies that, for given $\epsilon > 0$ and large enough t,

$$\mathcal{E}^{(N,2)}(B_1 \times B_2) \le \mathcal{E}^{(N,1)}\big(B_1^{\epsilon}\big) \cdot \mathcal{E}^{(N,1)}\big(B_2^{\epsilon}\big) + \epsilon. \tag{7.17}$$

Setting $B_i = [c_1, c_2]$, $i = 1, 2$, one obtains (5.2) from (7.17). □

8 Demonstration of Theorems 2.1, 2.2 and 2.3

In this section, we combine the main results of Sects. 5–7 to demonstrate Theorems 2.1–2.3. Since the proofs of the different theorems are similar, we combine their arguments while demonstrating each part of the conclusions (a) and (b) in the ansatz. In particular, we break the proof into Propositions 8.1–8.4, where we show convergence to a limit, as in (2.1), in Proposition 8.1; identify the limit as the N'-fold product of some measure $\mathcal{E}^{(\infty,1)}$, in Proposition 8.2; show this measure is an equilibrium environment in Proposition 8.3; and show this measure is the unique equilibrium environment in Proposition 8.4.

In these propositions, we employ the following notation. The measures $\mathcal{E}^{(N,N')}$, $N' \le N$, denote the measures on $S^{(N')}$ of the restriction to the first N' coordinates of the equilibrium measure of the corresponding generalized supermarket model with N queues; $\mathcal{E}_T^{(N,N')}$, $N' \le N$, denotes the corresponding measure, on $[0, T] \times S^{(N')}$, for the stationary process $X_{\mathcal{E}}^{(N)}(\cdot)$, with initial measure $\mathcal{E}^{(N,N')}$, that is restricted to the first N' coordinates. As in Proposition 7.1, $\mathcal{M}_T^{(N,N')}$, $N' \le N$, denote the measures on $[0, T] \times S^{(N')}$ corresponding to the process started from the empty state, and $\mathcal{M}_T^{(\infty,N')}$ denotes their limit as $N \to \infty$; $\mathcal{M}_{q,T}^{(N,N')}$ and $\mathcal{M}_{q,T}^{(\infty,N')}$ will denote the measures on the same space corresponding to the process over $[q, q + T]$, with $q \ge 0$.

Proposition 8.1 states that, as $N \to \infty$, respectively $q \to \infty$, the measures $\mathcal{E}_T^{(N,N')}$, respectively $\mathcal{M}_{q,T}^{(N,N')}$, converge to a limit that is the same in both cases. Consequently, the limiting behavior of $\mathcal{M}_{q,T}^{(N,N')}$, as $N \to \infty$ and $q \to \infty$, does not depend on the order in which the limits are taken. By restricting the limit $\mathcal{E}_T^{(\infty,N')}$ to its marginal $\mathcal{E}^{(\infty,N')}$ at any $t \in [0, T]$, one obtains the limit in (2.1). The proof employs Propositions 5.2, 6.1, and 7.1.

Proposition 8.1 *For given T and N', let $\mathcal{E}_T^{(N,N')}$ and $\mathcal{M}_{q,T}^{(\infty,N')}$ denote the above measures corresponding to a LL(D) supermarket model, a SQ(D) supermarket model that is FIFO with DHR, or a generalized supermarket model, where $\alpha < 1$ is assumed in the first two cases and $\alpha \le 1/(2\sqrt{D(D \vee \theta)})$ is assumed in the third case. Then*

$$\mathcal{E}_T^{(N,N')} \xrightarrow{v} \mathcal{E}_T^{(\infty,N')} \quad \text{as } N \to \infty \tag{8.1}$$

and

$$\mathcal{M}_{q,T}^{(\infty,N')} \xrightarrow{v} \mathcal{E}_T^{(\infty,N')} \quad \text{as } q \to \infty, \tag{8.2}$$

for some probability measure $\mathcal{E}_T^{(\infty,N')}$ on $[0, T] \times S^{(N')}$, whose marginals are invariant in $t \in [0, T]$.

Proof The reasoning for (8.1) and (8.2) is similar. For (8.1), it follows from Propositions 5.2, 6.1, and 7.3 that, for given T, N' and $\epsilon > 0$, and large enough q and N_i, $i = 1, 2$,

$$\left|\mathcal{M}_{q,T}^{(N_i,N')}(A) - \mathcal{E}_T^{(N_i,N')}(A)\right| \le \epsilon \tag{8.3}$$

for any $A \in \mathcal{B}_T^{(N')}$, where $\mathcal{B}_T^{(N)}$ is defined at the end of Sect. 3. (It follows from Lemma 4.2 that $X_{\mathcal{E}}^{(N)}(\cdot)$ and $X_{\mathcal{E},\pi_q}^{(N)}(\cdot)$ have the same law over $[q, q + T]$.) On the other hand, by Proposition 7.1, for given q, T, N' and $\epsilon > 0$, and large enough N_i, $i = 1, 2$,

$$\left|\mathcal{M}_{q,T}^{(N_1,N')}(A) - \mathcal{M}_{q,T}^{(N_2,N')}(A)\right| \le \epsilon. \tag{8.4}$$

Combining (8.3) and (8.4), one obtains

$$\left|\mathcal{E}_T^{(N_1,N')}(A) - \mathcal{E}_T^{(N_2,N')}(A)\right| \le 3\epsilon, \tag{8.5}$$

from which it follows that

$$\mathcal{E}_T^{(\infty,N')}(A) = \lim_{N\to\infty} \mathcal{E}_T^{(N,N')}(A) \tag{8.6}$$

exists for all $A \in \mathcal{B}_T^{(N')}$. One can check that $\mathcal{E}_T^{(\infty,N')}$ defines a probability measure on $([0, T] \times S^{(N')}, \mathcal{B}_T^{(N')})$. This shows (8.1).

To show (8.2), we note that, for given T, N' and $\epsilon > 0$, and large enough q_i, $i = 1, 2$, and N, it follows from Propositions 5.2 and 6.1 that

$$\left|\mathcal{M}_{q_1,T}^{(N,N')}(A) - \mathcal{M}_{q_2,T}^{(N,N')}(A)\right| \le \epsilon \tag{8.7}$$

for any $A \in \mathcal{B}_T^{(N')}$. On the other hand, by Proposition 7.1, for given q_i, $i = 1, 2$, T, N' and $\epsilon > 0$, and large enough N,

$$\left|\mathcal{M}_{q_i,T}^{(N,N')}(A) - \mathcal{M}_{q_i,T}^{(\infty,N')}(A)\right| \le \epsilon. \tag{8.8}$$

Combining (8.7) and (8.8), one obtains

$$\left|\mathcal{M}_{q_1,T}^{(\infty,N')}(A) - \mathcal{M}_{q_2,T}^{(\infty,N')}(A)\right| \le 3\epsilon, \tag{8.9}$$

from which it follows that

$$\tilde{\mathcal{E}}_T^{(\infty,N')}(A) = \lim_{q\to\infty} \mathcal{M}_{q,T}^{(\infty,N')}(A) \tag{8.10}$$

exists for all $A \in \mathcal{B}_T^{(N')}$ and defines a probability measure. This shows (8.2).

We still need to show that

$$\tilde{\mathcal{E}}_T^{(\infty,N')} = \mathcal{E}_T^{(\infty,N')}. \tag{8.11}$$

Note that, on account of (8.3) and (8.8), for given $\epsilon > 0$, large q, and sufficiently larger N,

$$\left|\mathcal{M}_{q,T}^{(\infty,N')}(A) - \mathcal{E}_T^{(N,N')}(A)\right| \le 2\epsilon \tag{8.12}$$

for all $A \in \mathcal{B}_T^{(N')}$. Together with (8.6) and (8.10), (8.12) implies (8.11). Since the marginals of $\mathcal{E}_T^{(N,N')}$ do not depend on $t \in [0, T]$, neither do the marginals of $\mathcal{E}_T^{(\infty,N')}$. □

The proof that $\mathcal{E}_T^{(N,N')}$, and hence $\mathcal{E}^{(N,N')}$, is a product measure follows quickly from the previous proposition and Proposition 7.1.

Proposition 8.2 *Assume that the LL(D), SQ(D) and generalized supermarket models are as specified in Proposition 8.1, and let $\mathcal{E}_T^{(\infty,N')}$ be the limit given in (8.1). Then $\mathcal{E}_T^{(\infty,N')}$ is the N'-fold product of the measures $\mathcal{E}_T^{(\infty,1)}$.*

Proof By Proposition 7.1, $\mathcal{M}_{q,T}^{(\infty,N')}$ is the N'-fold product of $\mathcal{M}_{q,T}^{(\infty,1)}$ for all q, T and N'. So, the claim follows from (8.2). □

The marginal of $\mathcal{E}_T^{(\infty,1)}$, at a given time t, does not depend on t. We next show that this measure $\mathcal{E}^{(\infty,1)}$ is an equilibrium environment. To show this, we first recall the process $X_{T',\infty}^{(N,N')}(\cdot)$ on $[0, T'] \times S^{(N')}$, from Sect. 7 that was constructed using

the influence process $\mathcal{I}^{(N,N')}_{T',\infty}(\cdot)$ and the corresponding intersecting selection sets, the service times of the corresponding arrivals, and $X^{(N),n}(0)$, $n \in \mathbb{Z}^+$. Here, we set $N = N' = 1$, restrict $X^{(1,1)}_{T',\infty}(\cdot)$ to $t \in [q, q+T]$, with $T = T' - q$, and we write $X^1_{q,T}(\cdot)$ for the corresponding process on $[0, T] \times S$ that is obtained by translating this restriction by q in time. Also, recall that, by Lemma 7.2, the probability measure $\mathcal{M}^{(\infty,1)}_T \stackrel{\text{def}}{=} \mathcal{M}^{(N,1)}_{T,\infty}$, and hence $\mathcal{M}^{(\infty,1)}_{q,T}$, does not depend on N. The evolution over different branches of $\mathcal{I}^{(1,1)}_{T',\infty}(\cdot)$ is independent with the same law, as are the corresponding intersecting selection sets and $X^{(N),n}(0)$. One can therefore check that the process $X^1_{q,T}(\cdot)$ is a cavity process with the equilibrium environment process $\mathcal{M}^1_{q,T}(t)$, $t \in [0, T]$, where $\mathcal{M}^1_{q,T}(t)$ is the marginal at time t of $\mathcal{M}^{(\infty,1)}_{q,T}$, and whose policy is that of the original generalized supermarket model.

Proposition 8.3 *Assume that the LL(D), SQ(D), and generalized supermarket models are as specified in Proposition 8.1, let $\mathcal{E}^{(\infty,1)}_T$ be as in (8.1), and let $\mathcal{E}^{(\infty,1)}$ be its marginal. Then $\mathcal{E}^{(\infty,1)}$ is an equilibrium environment for the corresponding model.*

Proof Fix $T > 0$. In addition to $X^1_{q,T}(\cdot)$, we consider the cavity processes $X^2_{q,T}(\cdot)$ and $X^3_T(\cdot)$ on $[0, T] \times S$, where $X^2_{q,T}(\cdot)$ and $X^3_T(\cdot)$ have the same policy as $X^1_{q,T}(\cdot)$, and $X^2_{q,T}(0) = X^1_{q,T}(0)$, but where $X^2_{q,T}(\cdot)$ and $X^3_T(\cdot)$ have environment process $\mathcal{E}^{(\infty,1)}_T(\cdot)$, with $\mathcal{E}^{(\infty,1)}_T(t) \stackrel{\text{def}}{=} \mathcal{E}^{(\infty,1)}$ for all t, instead of $\mathcal{M}^1_{q,T}(\cdot)$, and $X^3_T(0)$ is distributed accord to $\mathcal{E}^{(\infty,1)}$; note that the law of $X^3_T(\cdot)$ does not depend on q. We will write $\mathcal{M}^2_{q,T}$ and $\mathcal{M}^3_T$ for the measures on $[0, T] \times S$ corresponding to these cavity processes.

We first compare $\mathcal{M}^1_{q,T}$ with $\mathcal{M}^2_{q,T}$. By (8.2), for given T and $\epsilon > 0$, and large enough q,

$$\left|\mathcal{E}^{(\infty,1)}_T(A) - \mathcal{M}^1_{q,T}(A)\right| \le \epsilon$$

for any $A \in \mathcal{B}^{(1)}_T$, and hence

$$\left|\mathcal{E}^{(\infty,1)}(B) - \mathcal{M}^1_{q,T}(t; B)\right| \le \epsilon \tag{8.13}$$

for any $B \in \mathscr{S}$, where $\mathcal{M}^1_{q,T}(t; B)$ is the measure of B with respect to $\mathcal{M}^1_{q,T}(t)$. Recall that $\mathscr{S}$ is the Borel σ-algebra on S.

By employing the maximal coupling (see, e.g., Thorisson [18], p. 107) and (8.13), one can couple $X^1_{q,T}(\cdot)$ and $X^2_{q,T}(\cdot)$ so that potential arrivals occur together and, when a potential arrival occurs, the $D-1$ comparison states are the same for both processes except on a set of probability $\epsilon(D-1)$. On the other hand, the expected number of potential arrivals over $[0, T]$ is DT. Letting $F_{q,T}$ denote the set on $[0, T] \times S$ where, for some potential arrival, the comparison sets for $X^1_{q,T}(\cdot)$ and $X^2_{q,T}(\cdot)$ are not identical, it follows that, under this coupling,

$$P(F_{q,T}) \le \epsilon(D-1)DT.$$

Since $X^2_{q,T}(0) = X^1_{q,T}(0)$, it follows that

$$P\big(X^2_{q,T}(t) \neq X^1_{q,T}(t) \text{ for some } t \in [0,T]\big) \leq \epsilon(D-1)DT$$

and hence, for any $A \in \mathcal{B}^{(1)}_T$,

$$\big|\mathcal{M}^2_{q,T}(A) - \mathcal{M}^1_{q,T}(A)\big| \leq \epsilon(D-1)DT. \tag{8.14}$$

The comparison of $\mathcal{M}^2_{q,T}$ with $\mathcal{M}^3_T$ is simpler. The environment processes for $X^2_{q,T}(\cdot)$ and $X^3_T(\cdot)$ have the same law, and so can be coupled so that the potential arrivals and comparison states are identical. On account of (8.2), their initial states can be coupled so that, for given T and $\epsilon > 0$, and large enough q,

$$P\big(X^3_T(0) \neq X^2_{q,T}(0)\big) \leq \epsilon.$$

It therefore follows that

$$P\big(X^3_T(t) \neq X^2_{q,T}(t) \text{ for some } t \in [0,T]\big) \leq \epsilon$$

and hence, for any $A \in \mathcal{B}^{(1)}_T$,

$$\big|\mathcal{M}^3_T(A) - \mathcal{M}^2_{q,T}(A)\big| \leq \epsilon. \tag{8.15}$$

Together, (8.14) and (8.15) imply that, for given T and $\epsilon > 0$, and large enough q,

$$\big|\mathcal{M}^3_T(A) - \mathcal{M}^1_{q,T}(A)\big| \leq \epsilon\big[(D-1)DT + 1\big] \tag{8.16}$$

for any $A \in \mathcal{B}^{(1)}_T$. On the other hand, for given T and $\epsilon > 0$, and large enough q, (8.2) implies that

$$\big|\mathcal{M}^1_{q,T}(A) - \mathcal{E}^{(\infty,1)}_T(A)\big| \leq \epsilon \tag{8.17}$$

for any $A \in \mathcal{B}^{(1)}_T$. It follows from (8.16) and (8.17) that, for given T and $\epsilon > 0$, and large enough q,

$$\big|\mathcal{M}^3_T(A) - \mathcal{E}^{(\infty,1)}_T(A)\big| \leq \epsilon\big[(D-1)DT + 2\big]$$

for any $A \in \mathcal{B}^{(1)}_T$; letting $\epsilon \searrow 0$, it follows that $\mathcal{M}^3_T = \mathcal{E}^{(\infty,1)}_T$. Since $X^3_T(\cdot)$ has environment process $\mathcal{E}^{(\infty,1)}_T(\cdot)$, it follows that $\mathcal{E}^{(\infty,1)}_T(\cdot)$ is an equilibrium environment process; consequently, $\mathcal{E}^{(\infty,1)}$ is an equilibrium environment, as desired. □

Proposition 8.4 *Assume that the LL(D), SQ(D) and generalized supermarket models are as specified in Proposition 8.1, and let $\mathcal{E}^{(\infty,1)}_T$ be as in (8.1). Then its marginal $\mathcal{E}^{(\infty,1)}$ is the unique equilibrium environment for the corresponding model.*

Demonstration of Proposition 8.4 The remainder of the section is devoted to demonstrating Proposition 8.4, for which we require two preliminary results. The following lemma states that an equilibrium environment process is determined by its initial value.

Lemma 8.1 *Suppose $\mathcal{H}_i(\cdot)$, $i = 1, 2$, are equilibrium environment processes for a generalized supermarket model, with $\mathcal{H}_1(0) = \mathcal{H}_2(0)$. Then $\mathcal{H}_1(t) = \mathcal{H}_2(t)$ for all t.*

Proof Let $X_i(\cdot)$, $i = 1, 2$, denote cavity processes corresponding to $\mathcal{H}_i(\cdot)$, which we couple in the natural fashion so that $X_1(0) = X_2(0)$, potential arrivals are coupled and, at each potential arrival, the maximal coupling is applied to the comparison states. Denote by $p(t)$ the probability that $X_1(t) \neq X_2(t)$, which is at least the distance between $\mathcal{H}_1(t)$ and $\mathcal{H}_2(t)$ in the total variation norm. The rate at which potential arrivals occur is D, and so one can check that

$$p(t) \leq \int_0^t (D-1)Dp(u)\,du.$$

Since $p(0) = 0$, it follows from Gronwall's inequality that $p(t) = 0$ for all t, which implies the claim. □

The following corollary follows from Proposition 7.1 and Lemma 8.1.

Corollary 8.1 *Suppose the processes $X^{(N)}(\cdot)$, $N \in \mathbb{Z}^+$, are as in Proposition 7.1, with the coordinates $X^{(N),n}(0)$ of $X^{(N)}(0)$ having distribution given by an equilibrium environment $\tilde{\mathcal{E}}$ for the corresponding supermarket policy, and let $\mathcal{M}_T^{(\infty,1)}$ be the limit in (7.1). Then $\mathcal{M}_T^{(\infty,1)}(t) = \tilde{\mathcal{E}}$ for all $t \in [0, T]$.*

Proof The same reasoning as above Proposition 8.3 implies that $\mathcal{M}_T^{(\infty,1)}(\cdot)$ is an equilibrium environment process. (The only change is that here, $X^{(N)}(0)$ has distribution $\tilde{\mathcal{E}}$, rather than $X^{(N)}(0) = 0$ as before.) On the other hand, $\tilde{\mathcal{E}}_T(\cdot)$ is also an equilibrium environment process, with $\tilde{\mathcal{E}}_T(t) \stackrel{\text{def}}{=} \tilde{\mathcal{E}}$ for $t \in [0, T]$. Since the marginal at time 0 of $\mathcal{M}_T^{(\infty,1)}$ equals $\tilde{\mathcal{E}}$, it follows from Lemma 8.1 that $\mathcal{M}_T^{(\infty,1)}(t) = \tilde{\mathcal{E}}_T(t)$ for $t \in [0, T]$, which implies the corollary. □

The following proposition is a slight variant of Proposition 5.2. Here, $X_1^{(N)}(\cdot)$ denotes the process with $X_1^{(N)}(0) = 0$ and $X_2^{(N)}(\cdot)$ denotes the process where $X_2^{(N),n}(0)$, $n = 1, \ldots, N$, are i.i.d. with distribution given by an equilibrium environment $\tilde{\mathcal{E}}$. (Note that $X^{(N)}_{\mathcal{E},\pi_q}(q) = X^{(N)}_{\mathcal{E},\pi}(q)$, for π_q as in Proposition 5.2.)

Proposition 8.5 *Consider, on $S^{(N)}$, either the LL(D) supermarket model, or the SQ(D) supermarket model that is FIFO with DHR. Assume the processes $X_1^{(N)}(\cdot)$ and $X_2^{(N)}(\cdot)$ are defined as above. Then, for each $\gamma_1 > 0$, there exists $q = q(\gamma_1)$, with $q \to \infty$ as $\gamma_1 \searrow 0$, such that, for large enough N depending on q,*

$$P\big(X_1^{(N),n}(q) \neq X_{2,\pi}^{(N),n}(q)\big) \leq \gamma_1 \tag{8.18}$$

for all $n = 1, \ldots, N$.

Summary of proof The argument is essentially the same as those employed for Propositions 5.1 and 5.2, with changes in two places.

Since the assumptions in Proposition 5.4 continue to hold in the present setting, the analog of the inequality in (5.13) holds. By employing (5.13), one can show the analog of (5.1) in Proposition 5.1 holds for large enough N depending on q where, as in Proposition 8.5, q is chosen so that $q \to \infty$ as $\gamma_1 \searrow 0$. The reasoning is the same as in the proof of Proposition 5.1, except that, since the supremum in (5.20) might not be finite, (5.20) is replaced by

$$L = L_{M,\epsilon,q} \stackrel{\text{def}}{=} \sup_{N \geq N_q, u \leq q} L_{M,\epsilon,u}^{X_2^{(N)}} < \infty; \tag{8.19}$$

on account of Corollary 8.1, N_q may be chosen to grow sufficiently quickly when q increases so that this supremum is finite.

The same reasoning may be applied as in the proof of Proposition 5.2, starting with (5.23), except that (5.26) needs to be modified by restricting N as above so that $N \geq N_q$, and N_q grows with q. The proofs of Lemmas 5.3–5.6 and Proposition 5.5 then proceed as before. □

The proof of Proposition 8.4 follows quickly from (8.2), Corollary 8.1, and Proposition 8.5.

Proof of Proposition 8.4 For a given equilibrium environment $\tilde{\mathcal{E}}^{(\infty,1)}$, we wish to show $\tilde{\mathcal{E}}^{(\infty,1)} = \mathcal{E}^{(\infty,1)}$, where $\mathcal{E}^{(\infty,1)}$ is the equilibrium environment in Proposition 8.3. To show this, we employ the processes $X_i^{(N)}(\cdot)$, $i = 1, 2$, with the corresponding supermarket policy, where $X_1^{(N)}(0) = 0$ and $X_2^{(N)}(0)$ has i.i.d. coordinates with distribution given by $\tilde{\mathcal{E}}^{(\infty,1)}$. Since $X_1^{(N)}(0)$ and $X_2^{(N)}(0)$ are exchangeable, it follows from Lemma 4.2 that $X_2^{(N)}(t)$ and $X_{2,\pi}^{(N)}(t)$ have the same distribution for each t. Denoting the space-time measures of $X_i^{(N),1}(t)$, $i = 1, 2$, by $\mathcal{M}_i^{(N,1)}$, it therefore follows from Proposition 8.5 that, for the LL(D) and SQ(D) policies and each $\epsilon > 0$, there exists q, with $q \to \infty$ as $\epsilon \to \infty$, such that, for large enough N depending on q,

$$\big|\mathcal{M}_2^{(N,1)}(q; B) - \mathcal{M}_1^{(N,1)}(q; B)\big| \leq \epsilon \tag{8.20}$$

for any $B \in \mathscr{S}$. Moreover, on account of Proposition 6.4, (8.20) also holds for generalized supermarket models, when $\alpha \leq 1/(2\sqrt{D(D \vee \theta)})$.

In all three cases, it follows from Proposition 7.1 that, for given q and large enough N,

$$\big|\mathcal{M}_i^{(N,1)}(q; B) - \mathcal{M}_i^{(\infty,1)}(q; B)\big| \leq \epsilon \tag{8.21}$$

for any $B \in \mathscr{S}$ and $i = 1, 2$. Also, because of (8.2), for large enough q,

$$\left|\mathcal{M}_1^{(\infty,1)}(q; B) - \mathcal{E}^{(\infty,1)}(B)\right| \leq \epsilon. \tag{8.22}$$

Moreover, it follows from Corollary 8.1 that, for each q

$$\mathcal{M}_2^{(\infty,1)}(q) = \tilde{\mathcal{E}}^{(\infty,1)}. \tag{8.23}$$

Combining (8.20), (8.21), (8.22), and (8.23), it follows that, for each $\epsilon > 0$,

$$\left|\tilde{\mathcal{E}}^{(\infty,1)}(B) - \mathcal{E}^{(\infty,1)}(B)\right| \leq 4\epsilon \tag{8.24}$$

for any $B \in \mathscr{S}$. Letting $\epsilon \searrow 0$ implies $\tilde{\mathcal{E}}^{(\infty,1)} = \mathcal{E}^{(\infty,1)}$, as desired. □

Acknowledgements Maury Bramson is supported in part by NSF Grants CCF-0729537 and DMS-1105668. Balaji Prabhakar is supported in part by NSF Grant CCF-0729537 and by a grant from the Clean Slate Program at Stanford University.

References

1. Azar, Y., Broder, A., Karlin, A., Upfal, E.: Balanced allocations. In: Proc. 26th ACM Symp. Theory Comp., pp. 593–602 (1994)
2. Athreya, K.B., Ney, P.E.: Branching Processes. Springer, Berlin (1972)
3. Bramson, M.: Stability of queueing networks. In: École d'Été de Probabilités de Saint-Flour XXXVI—2006. Lecture Notes in Mathematics, vol. 1950. Springer, Berlin (2008)
4. Bramson, M.: Stability of join the shortest queue networks. Ann. Appl. Probab. **21**, 1568–1625 (2011)
5. Bramson, M., Lu, Y., Prabhakar, B.: Randomized load balancing with general service time distributions. In: Proc. ACM SIGMETRICS, pp. 275–286 (2010)
6. Bramson, M., Lu, Y., Prabhakar, B.: Decay of tails at equilibrium for FIFO joint the shortest queue networks. To appear in Ann. Appl. Probab. (2012)
7. Chung, K.L.: A Course in Probability Theory, 2nd edn. Academic Press, New York (1985)
8. Davis, M.H.A.: Markov Models and Optimization. Chapman & Hall, London (1993)
9. Dynkin, E.B.: Markov Processes, vol. 1. Springer, Berlin (1965)
10. Foss, S., Chernova, N.: On the stability of a partially accessible multi-station queue with state-dependent routing. Queueing Syst. **29**, 55–73 (1998)
11. Graham, C.: Chaoticity on path space for a queueing network with selection of the shortest queue among several. J. Appl. Probab. **37**, 198–211 (2000)
12. Graham, C., Méléard, S.: Chaos hypothesis for a system acting through shared resources. Probab. Theory Relat. Fields **100**, 157–173 (1994)
13. Harris, T.E.: The Theory of Branching Processes. Springer, Berlin (1963)
14. Luczak, M., McDiarmid, C.: On the power of two choices: Balls and bins in continuous time. Ann. Appl. Probab. **15**, 1733–1764 (2005)
15. Luczak, M., McDiarmid, C.: On the maximum queue length in the supermarket model. Ann. Probab. **34**, 493–527 (2006)
16. Martin, J.B., Suhov, Y.M.: Fast Jackson networks. Ann. Appl. Probab. **9**, 840–854 (1999)
17. Mitzenmacher, M.: The power of two choices in randomized load balancing. IEEE Trans. Parallel Distrib. Syst. **12**(10), 1094–1104 (2001)
18. Thorisson, H.: Coupling, Stationarity, and Regeneration. Springer, New York (2000)
19. Vvedenskaya, N.D., Dobrushin, R.L., Karpelevich: Queueing system with selection of the shortest of two queues: an asymptotic approach. Probl. Inf. Transm. **32**(1), 15–29 (1996)