J Stat Phys (2018) 173:1149–1194
https://doi.org/10.1007/s10955-018-2044-7

The Supermarket Model with Bounded Queue Lengths in Equilibrium

Graham Brightwell[1] · **Marianne Fairthorne**[1] · **Malwina J. Luczak**[2]

Received: 26 October 2017 / Accepted: 16 April 2018 / Published online: 28 April 2018

Abstract In the supermarket model, there are n queues, each with a single server. Customers arrive in a Poisson process with arrival rate λn, where $\lambda = \lambda(n) \in (0, 1)$. Upon arrival, a customer selects $d = d(n)$ servers uniformly at random, and joins the queue of a least-loaded server amongst those chosen. Service times are independent exponentially distributed random variables with mean 1. In this paper, we analyse the behaviour of the supermarket model in the regime where $\lambda(n) = 1 - n^{-\alpha}$ and $d(n) = \lfloor n^{\beta} \rfloor$, where α and β are fixed numbers in $(0, 1]$. For suitable pairs (α, β), our results imply that, in equilibrium, with probability tending to 1 as $n \to \infty$, the proportion of queues with length equal to $k = \lceil \alpha/\beta \rceil$ is at least $1 - 2n^{-\alpha+(k-1)\beta}$, and there are no longer queues. We further show that the process is rapidly mixing when started in a good state, and give bounds on the speed of mixing for more general initial conditions.

Keywords Supermarket model · Markov chains · Rapid mixing · Concentration of measure · Load balancing

The research of Malwina Luczak was supported by an EPSRC Leadership Fellowship, Grant references EP/J004022/1 and EP/J004022/2.

✉ Graham Brightwell
g.r.brightwell@lse.ac.uk
http://www.maths.lse.ac.uk/Personal/graham/

Marianne Fairthorne
marianne.fairthorne@googlemail.com

Malwina J. Luczak
mluczak@unimelb.edu.au
https://findanexpert.unimelb.edu.au/display/person450456

[1] Department of Mathematics, London School of Economics, Houghton Street, London WC2A 2AE, UK

[2] School of Mathematics and Statistics, University of Melbourne, Melbourne, Australia

1 Introduction

The supermarket model is a well-studied Markov chain model for a dynamic load-balancing process. There are n servers, and customers arrive according to a Poisson process with rate $\lambda = \lambda(n) < 1$. On arrival, a customer inspects $d = d(n)$ queues, chosen uniformly at random with replacement, and joins a shortest queue among those inspected (in case of a tie, the first shortest queue in the list is joined). Each server serves one customer at a time, and service times are iid random variables, with an exponential distribution of mean 1.

A number of authors [5–9,11–13,17,18,21,23] have studied the supermarket model, as well as various extensions, e.g., to the setting of a Jackson network [15] and to a version with one queue saved in memory [14,20]. There are related ideas in other queueing models, for instance one where one server inspects d queues and serves the longest [1].

Early papers on the supermarket model concentrated on the case where λ and d are held fixed as n tends to infinity. As with other related models (see, e.g. [10,19]), there is a dramatic change when d is increased from 1 to 2: if $d = 1$, the maximum queue length in equilibrium is of order $\log n$, while if d is a constant at least 2, then the maximum queue length in equilibrium is of order $\log \log n / \log d$.

Luczak and McDiarmid [11] prove that, for fixed λ and d, the sequence of Markov chains indexed by n is rapidly mixing: as $n \to \infty$, the time for the system to converge to equilibrium is of order $\log n$, provided the initial state has not too many customers and no very long queue. Also, they show that, for $d \geq 2$, with probability tending to 1 as $n \to \infty$, in the equilibrium distribution the maximum queue length takes one of at most 2 values, and that these values are $\log \log n / \log d + O(1)$.

More recently, there has been interest in regimes where the parameters of the model may vary as n tends to infinity. Fairthorne [6] and Mukherjee et al [21] treat the case where $\lambda < 1$ is fixed and $d = d(n)$ tends to infinity with n. Eschenfeldt and Gamarnik [5] consider the "heavy traffic regime", where $\lambda = \lambda(n)$ tends to 1 from below as $n \to \infty$, and d is held fixed.

In this paper, we study a different regime. We focus on the case where $\lambda = \lambda(n) = 1 - n^{-\alpha}$ and $d = d(n) = \lfloor n^{\beta} \rfloor$, where α and β are fixed constants in (0, 1] with $k - 1 < \alpha/\beta < k$ for some positive integer k. We also require that $2\alpha < 1 + \beta(k-1)$, for reasons that we shall explain after the statement of Theorem 1.1 (see Remark (4)). Our results imply that, in equilibrium, with high probability (i.e., with probability tending to 1 as $n \to \infty$), the proportion of queues of length exactly equal to k is at least $1 - 2n^{-\alpha+(k-1)\beta}$, and there are no longer queues. Our methods actually cover a much broader range of parameter values, but we focus on this case for ease of exposition.

We offer two reasons why such a regime might be of interest: for one, this is a range of parameter values where near-perfect load balancing is achieved, with bounded maximum queue length, even when the system is running at nearly full capacity, and the values of d we obtain thus represent a sufficient amount of resource (in terms of inspection of queue-lengths) required to achieve this load-balancing. From a more theoretical viewpoint, we see our regimes, for the different values of $\lceil \alpha/\beta \rceil$, as possessing a scaling limit as $n \to \infty$, and varying the parameters so that α/β passes through an integer is an example of a phase transition.

To motivate our results, we first give heuristics to indicate what behaviour we might expect. Consider the infinite system of differential equations

$$\frac{dv_j(t)}{dt} = \lambda(v_{j-1}(t)^d - v_j(t)^d) - (v_j(t) - v_{j+1}(t)), \qquad j \geq 1, \tag{1.1}$$

where $v_0(t) = 1$ for all t. For an initial condition $v(0)$ such that $1 \geq v_1(0) \geq v_2(0) \geq \cdots \geq 0$ and $v_j(0) \to 0$ as $j \to \infty$, there is a unique solution $v(t)$ $(t \geq 0)$, with $v(t) = (v_j(t))_{j \geq 1}$, which is such that $1 \geq v_1(t) \geq v_2(t) \geq \cdots \geq 0$ and $v_j(t) \to 0$ as $j \to \infty$, for each $t \geq 0$. It follows from earlier work [7,8,12,13,23] that, with high probability, for each j, the proportion of queues of length at least j at time t stays "close to" $v_j(t)$ over a bounded time interval (or an interval whose length tends to infinity at most polynomially with n), assuming this is the case at time 0.

The system (1.1) has a unique, attractive, fixed point $\pi = (\pi_j)_{j \geq 1}$, such that $\pi_j \to 0$ as $j \to \infty$, given by

$$\pi_j = \lambda^{1+\cdots+d^{j-1}}, \qquad j \geq 1. \tag{1.2}$$

If λ and d are fixed constants, then, in equilibrium, with high probability, the proportion of queues of length at least j is close to π_j for each $j \geq 1$; see [7,8,11,12].

For λ and d functions of n, there is no single limiting differential equation (1.1), but rather a sequence of approximating differential equations, each with their own solutions and fixed points. In this paper, we do not address the question of whether such approximations to the evolution of the process are valid in generality, focussing solely on equilibrium behaviour and the time to reach equilibrium. If $\lambda = 1 - n^{-\alpha}$ and $d = \lfloor n^\beta \rfloor$, and k is an integer with $k - 1 < \alpha/\beta < k$, then

$$\begin{aligned}\pi_k = \lambda^{1+\cdots+d^{k-1}} &\geq (1 - n^{-\alpha})^{(1+o(1))d^{k-1}} \geq 1 - (1 + o(1))n^{-\alpha+(k-1)\beta} \\ &= 1 - o(1), \\ \pi_{k+1} = \lambda^{1+\cdots+d^k} &\leq \exp(-d^k n^{-\alpha}) \leq \exp(-\frac{1}{2} n^{k\beta-\alpha}) = o(1/n).\end{aligned}$$

We will indeed show that, in equilibrium, with high probability, there are no queues of length greater than k, while the proportion of queues with length exactly k tends to 1 as $n \to \infty$. Moreover we show that, for $0 \leq j < k$, the number of queues of length exactly j is very close to $n(\pi_j - \pi_{j+1}) \simeq n^{1-\alpha+j\beta}$.

We also prove results on mixing time to equilibrium. We show that, if we start in a "good" initial state (one without any very long queue, and without too many customers in the system in total), then the mixing time is of order $n^{1+(k-1)\beta} \log n$, which is best possible up to the logarithmic term. We also prove general bounds on the mixing time, in terms of the initial number of customers and the initial maximum queue length, and show that these bounds are also roughly best possible.

We will shortly state our main results precisely, but first we describe the supermarket model more carefully. In fact, we describe a natural discrete-time version of the process, which we shall work with throughout; as is standard, one may convert results about the discrete time version to the continuous model, with the understanding that one unit of time in the continuous model corresponds to about $(1 + \lambda)n$ steps of the discrete model.

A *queue-lengths vector* is an n-tuple $(x(1), \ldots, x(n))$ whose entries are non-negative integers. If $x(j) = i$, we say that queue j has *length* i, or that there are i *customers* in queue j; we think of these customers as in *positions* $1, \ldots, i$ in the queue. We use similar terminology throughout; for instance, to say that a customer *arrives* and *joins queue* j means that $x(j)$ increases by 1, and to say that a customer in queue j *departs* or *is served* means that $x(j)$ decreases by 1. Given a queue-lengths vector x, we write $\|x\|_1 = \sum_{j=1}^n x(j)$ to denote the total number of customers in state x, and $\|x\|_\infty = \max x(j)$ to denote the maximum queue length in state x.

For each $i \geq 0$, and each $x \in \mathbb{Z}_+^n$, we define $u_i(x)$ to be the proportion of queues in x with length at least i. So $u_0(x) = 1$ for all x, and, for each fixed x, the $u_i(x)$ form a non-increasing sequence of multiples of $1/n$, such that $u_i(x) = 0$ eventually. The sequence $(u_i(x))_{i\geq 0}$ captures the "profile" of a queue-lengths vector x, and we shall describe various sets of queue-lengths vectors, and functions of the queue-lengths vector, in terms of the $u_i(x)$.

For positive integers n and d, and $\lambda \in (0, 1)$, we now define the (n, d, λ)-*supermarket process*. This process is a discrete-time Markov chain (X_t), whose state space is the set $\mathbb{Z}_+^n$ of queue-lengths vectors, and where transitions occur at non-negative integer times. Each transition is either a customer *arrival*, with probability $\lambda/(1 + \lambda)$, or a *potential departure*, with probability $1/(1 + \lambda)$. If there is a potential departure, then a queue K is selected uniformly at random from $\{1, \dots, n\}$: if there is a customer in queue K, then they are served and depart the system. If there is an arrival, then d queues are selected uniformly at random, with replacement, from $\{1, \dots, n\}$, and the arriving customer joins a shortest queue among those selected. To be precise, a d-tuple $(K_1, \dots, K_d)$ is selected, and the customer joins queue $k = K_j$, where j is the least index such that $x(K_j)$ is minimal among $\{x(K_1), \dots, x(K_d)\}$.

For $x \in \mathbb{Z}_+^n$, (X_t^x) denotes a copy of the (n, d, λ)-supermarket process (X_t) where $X_0 = x$ a.s. Throughout, we let (Y_t) denote a copy of the process in equilibrium. The processes depend on the parameters (n, d, λ), but we suppress this dependence in the notation. Throughout, we use $(\mathcal{F}_t)$ to denote the natural filtration of the process (X_t). We use the notation $\mathbb{P}(\cdot)$ freely to denote probability in whatever space we work in.

We now state our main results. First, we describe sets of queue-lengths vectors $\mathcal{N}(n, \alpha, \beta)$: our aim is to prove that, for suitable values of α and β, with $d = \lfloor n^\beta \rfloor$, $\lambda = 1 - n^{-\alpha}$ and n sufficiently large, an equilibrium copy of the (n, d, λ)-supermarket process is concentrated in the set $\mathcal{N}(n, \alpha, \beta)$.

For $\alpha, \beta \in (0, 1]$, let $k = \lceil \alpha/\beta \rceil$, and let $\mathcal{N}(n, \alpha, \beta)$ be the set of all queue-lengths vectors x such that: $u_{k+1}(x) = 0$ and, for $1 \leq j \leq k$,

$$\big(1 - \frac{1}{\log n}\big) n^{-\alpha+(j-1)\beta} \leq 1 - u_j(x) \leq \big(1 + \frac{1}{\log n}\big) n^{-\alpha+(j-1)\beta}.$$

So, for $x \in \mathcal{N}(n, \alpha, \beta)$, we have the following.

(a) There are no queues of length $k + 1$ or greater.

(b) For $1 \leq j \leq k$, the number of queues of length less than j is $n(1 - u_j(x))$, which lies between $(1 \pm \frac{1}{\log n}) n^{1-\alpha+(j-1)\beta}$.

(c) In particular, the number of queues of length less than k is at most $(1+\frac{1}{\log n}) n^{1-\alpha+(k-1)\beta} = o(n)$, and so the proportion of queues of length exactly k tends to 1 as $n \to \infty$.

(d) For $1 \leq j \leq k - 1$, the number of queues of length exactly j is $n(u_j(x) - u_{j+1}(x))$, which lies between $(1 \pm \frac{2}{\log n}) n^{1-\alpha+j\beta}$.

Theorem 1.1 *Suppose that $\alpha, \beta \in (0, 1]$ are constants with $k - 1 < \alpha/\beta < k$ for some natural number k, and that $2\alpha < 1 + \beta(k - 1)$. Suppose also that $\lambda = \lambda(n) = 1 - n^{-\alpha}$ and $d = d(n) = \lfloor n^\beta \rfloor$. Then, for n sufficiently large, a copy (Y_t) of the (n, d, λ)-supermarket process in equilibrium satisfies*

$$\mathbb{P}\left(Y_t \notin \mathcal{N}(n, \alpha, \beta)\right) \leq e^{-\frac{1}{4}\log^2 n}.$$

Remarks (1) In fact, our proofs go through essentially unchanged if we demand only that $1 - \lambda(n) = n^{-\alpha+\delta_1(n)}$ and $d(n) = n^{\beta+\delta_2(n)}$, where $\delta_1(n)$ and $\delta_2(n)$ tend to zero as $n \to \infty$, and we replace instances of $n^{-\alpha+(j-1)\beta}$ in the definition of $\mathcal{N}(n, \alpha, \beta)$ by $(1 - \lambda)d^{j-1}$. For ease of exposition, we prefer to stick to definite values of λ and d; however, from now on we allow ourselves to write simply $d = n^\beta$, even though this need not be an integer.

(2) The conclusion of the theorem implies that it is rare for there to be queues of length greater than k in equilibrium, and so in particular it is rare for the last arriving customer to have joined a queue containing k other customers. Theorem 1.1 can thus be used to make statements about the performance of the system in equilibrium in terms of the total waiting time for each customer; we leave the details to the interested reader.

(3) In the case where $\alpha \le \beta$, Theorem 1.1 tells us that, in equilibrium, the maximum queue-length is 1 with high probability, and therefore that it will be extremely rare for an arriving customer to join a non-empty queue. In this case, some of the complexity of our proof can be avoided. This range is also covered by Fairthorne [6], with essentially the same proof and some sharper results, e.g. giving conditions for the maximum queue-length remaining equal to 1 for a time period n^K for fixed K.

(4) We now indicate why the condition $2\alpha < 1 + \beta(k-1)$ in Theorem 1.1 is necessary. For a state in $\mathcal{N}(n, \alpha, \beta)$, the total number of customers in the system is at least $kn - 2n^{1-\alpha+(k-1)\beta}$. If we consider the next $n^{2\alpha}$ steps, the number of arrivals minus the number of potential departures is asymptotically a normal random variable with mean and standard deviation both of order n^{α}. So the probability that the number of arrivals minus the number of departures is at least $3n^{\alpha}$ is bounded away from zero as $n \to \infty$. If $\alpha \ge 1 - \alpha + (k-1)\beta$, then this many excess arrivals would drive the total number of customers in the system over kn, which certainly implies that some queue of length $k+1$ would be created.

(5) If $\alpha \ge 1$ and β is arbitrary, a similar argument shows that, in equilibrium, for each k, the probability that there is a queue of length at least k is bounded away from zero. Indeed, starting from any state, for any $k \in \mathbb{N}$, there is a positive probability that, over the next n^2 transitions, the number of arrivals exceeds the number of departures by at least kn.

(6) For $\lambda < \lambda'$, there is a coupling of the (n, d, λ)- and (n, d, λ')-supermarket processes, so that at each time, each queue in the (n, d, λ)-supermarket process is no longer than in the (n, d, λ')-supermarket process, provided this is true at time 0. So, for instance, if at a given time there are at least m queues with length k in the (n, d, λ)-supermarket process, then there are also at least m queues with length at least k in the (n, d, λ')-supermarket process. If α/β is equal to a positive integer k, and $\alpha < k/(k+1)$ (so that the condition $2\alpha < 1 + (k-1)\beta$ is satisfied), then we can couple with the process for slightly lower, and slightly higher, values of α, to see that the maximum queue length in equilibrium is, with high probability, either k or $k+1$, and that most queues have length either k or $k+1$. Similarly, for $d < d'$, there is a coupling of the (n, d', λ)-supermarket process and the (n, d, λ)-supermarket process such that, for all times $t \ge 0$, and for each j, the number of customers in position at least j in their queue is no higher in the first process than the second (see [7,22]).

Combining these arguments actually gives an essentially complete picture of the maximum queue length in equilibrium for any parameters $\alpha \in (0, 1)$, $\beta > 0$. The regions of the (α, β)-plane not covered by Theorem 1.1 are of the form $E_k = \{(\alpha, \beta) : \alpha < 1, \frac{\alpha}{k} \le \beta \le \frac{2\alpha-1}{k-1}\}$. For a model with parameters in E_k, coupling in d shows that, with high probability, the maximum queue length in equilibrium is at most $k+1$; coupling in λ shows that, with high probability, the maximum queue length in equilibrium is at least k. Moreover, the argument in Remark (4) shows that the value $k+1$ occurs with probability bounded way from zero as $n \to \infty$.

(7) We define the model so that d queues are chosen *with replacement*, so it makes sense to ask what happens if $\beta > 1$. In this case, most arriving customers inspect every queue, and the situation is essentially the same as when $\beta = 1$ (when most arriving customers inspect at least half of the queues), or as when every arriving customer inspects every

queue (the "join the shortest queue" protocol). Our result in this case says that, for $\alpha < 1/2$, the maximum queue length is 1 with high probability in equilibrium. For $\alpha \geq 1/2$, we are in the region E_1 defined in the previous remark: the maximum queue length is either 1 or 2 with high probability in equilibrium, and the value 2 occurs with probability bounded away from 0. For the join the shortest queue protocol and $\lambda = 1 - cn^{-1/2}$, this situation is explored in detail by Eschenfeldt and Gamarnik [4].

(8) The case $\alpha = 1/2$ has been studied in queueing theory under the name of the Halfin-Whitt heavy traffic regime. In this case, Theorem 1.1 applies whenever $\beta < 1/2$ and $1/2\beta$ is not an integer, and the result implies that, in equilibrium, the proportion of queues of length $\lceil 1/2\beta \rceil$ tends to 1 as $n \to \infty$, and with high probability there are no longer queues. For $\beta > 1/2$, the maximum queue length in equilibrium is either 1 or 2 with high probability, and the value 2 occurs with probability bounded away from 0, as in Remark (4).

This is an explicit example of a model where we have a type of scaling limit: as we increase n with $\lambda = 1 - n^{-\alpha}$ and $d = n^{\beta}$, we retain the property that almost all queues have length $k = \lceil \alpha/\beta \rceil$ in equilibrium, with high probability, and the number of shorter queues is of order $n^{1-\alpha+\lfloor \alpha/\beta \rfloor \beta} = o(n)$. As we adjust the parameters so that α/β passes through an integer value, we have a phase transition to a different equilibrium regime.

As mentioned earlier, and explained in more detail in Sect. 2, our results are in line with a more general hypothesis: for a very wide range of parameter values, the maximum queue length of the (n, d, λ)-supermarket model in equilibrium is within 1 of the largest k such that

$$\pi_k = \lambda^{1+d+\cdots+d^{k-1}} > \frac{1}{n}.$$

(Recall that π_k is the "predicted" proportion of queues of length at least k; see (1.2).) This general hypothesis holds when λ and d are constants: see [11]. It is also valid for the range where λ is fixed and $d \to \infty$: see [6], and at least approximately when $\lambda \to 1$ and d is fixed: see [5].

We now state our results concerning "rapid mixing", i.e., rapid convergence to equilibrium. For $x \in \mathbb{Z}_+^n$, let $\mathcal{L}(X_t^x)$ denote the law at time t of the (n, d, λ)-supermarket process (X_t^x) started in state x. Also let Π denote the stationary distribution of the (n, d, λ)-supermarket process.

Theorem 1.2 *Suppose that $\lambda(n) = 1 - n^{-\alpha}$ and $d(n) = n^{\beta}$, where α, β and $k = \lceil \alpha/\beta \rceil$ satisfy the conditions of Theorem 1.1. Let x be a queue-lengths vector in $\mathcal{N}(n, \alpha, \beta)$. Then, for all sufficiently large n and for all $t \geq 0$,*

$$d_{TV}(\mathcal{L}(X_t^x), \Pi) \leq n \left(2e^{-\frac{1}{4}\log^2 n} + 4\exp\left(-\frac{t}{1600kn^{1+(k-1)\beta}} \right) \right).$$

In other words, for a copy of the process started in a state in $\mathcal{N}(n, \alpha, \beta)$, the mixing time is at most of order $n^{1+(k-1)\beta} \log n = o(n^{1+\alpha}) = o(n^2)$. In fact, this upper bound on the mixing time is best possible up to the logarithmic factor: we show that mixing, starting from states in $\mathcal{N}(n, \alpha, \beta)$, requires order at least $n^{1+(k-1)\beta}$ steps.

Theorem 1.3 *Suppose that $\lambda(n) = 1 - n^{-\alpha}$ and $d(n) = n^{\beta}$, where α, β and $k = \lceil \alpha/\beta \rceil$ satisfy the conditions of Theorem 1.1. For all sufficiently large n, there is a state $z \in \mathcal{N}(n, \alpha, \beta)$ such that, for $t \leq \frac{1}{8} n^{1+(k-1)\beta}$,*

$$d_{TV}(\mathcal{L}(X_t^z), \Pi) \geq 1 - 2e^{-\frac{1}{4}\log^2 n}.$$

From states not in $\mathcal{N}(n, \alpha, \beta)$, we cannot expect to have rapid mixing in general. For instance, suppose we start from a state x with number of customers $\|x\|_1 \geq kn$. The expected decrease in the number of customers at each step of the chain is at most $\frac{1-\lambda}{1+\lambda}$, so mixing takes at least of order $(\|x\|_1 - kn)(1-\lambda)^{-1} = (\|x\|_1 - kn)n^\alpha$ steps. Similarly, if we start with one long queue, of length $\|x\|_\infty > k$, then mixing takes at least of order $(\|x\|_\infty - k)n$ steps, to allow time for enough departures from the long queue. This shows that, for instance, if either $\|x\|_1 \geq 2kn$ or $\|x\|_\infty > 2k$, and

$$t \leq \frac{1}{10} \max\big(\|x_1\|n^\alpha, \|x\|_\infty n\big), \tag{1.3}$$

then the total variation distance $d_{TV}(\mathcal{L}(X_t^x, \Pi)$ is near to 1. The next result gives an upper bound on the mixing time for (X_t^x) in terms of $\|x\|_1$ and $\|x\|_\infty$, and shows that (1.3) is best possible up to the constant factor.

Theorem 1.4 *Suppose that α and β satisfy the hypotheses of Theorem 1.1, and let x be any queue-lengths vector with $\|x\|_\infty \leq e^{\frac{1}{4}\log^2 n}$. Then for n sufficiently large and*

$$t \geq 7200\big(kn^{1+\alpha} + \|x\|_1 n^\alpha + \|x\|_\infty n\big),$$

we have $d_{TV}(\mathcal{L}(X_t^x), \Pi) \leq 2e^{-\frac{1}{5}\log^2 n}$.

In the case where the dominant term in the expression above is $kn^{1+\alpha}$, this result is not as sharp as that in Theorem 1.2, since $\alpha > (k-1)\beta$.

The supermarket model is an instance of a model whose behaviour has been comprehensively analysed even though there are an unbounded number of variables that need to be tracked – namely, the proportions $u_i(X_t)$. While what we achieve in this paper is similar to what is achieved by Luczak and McDiarmid in [11] for the case where λ and d are fixed as $n \to \infty$, only some of the techniques of that paper can be used here, as we now explain.

The proofs in [11] rely on a coupling of copies of the supermarket process where the distance between coupled copies does not increase in time. This coupling is, in particular, used to establish concentration of measure, over a long time period, for Lipschitz functions of the queue-lengths vector; this result is valid for any values of (n, d, λ), and in particular in our setting. Fast coalescence of coupled copies, and hence rapid mixing, is shown by comparing the behaviour of the (n, d, λ)-process ($d \geq 2$) with the $(n, 1, \lambda)$-process, which is easy to analyse. This then also implies concentration of measure for Lipschitz functions in equilibrium, and that the profile of the equilibrium process is well concentrated around the fixed point π of the equations (1.1).

The coupling from [11] also underlies the proofs in the present paper. However, in our regime, comparisons with the $(n, 1, \lambda)$-process are too crude. Thus we cannot show that the coupled copies coalesce quickly enough, until we know something about the profiles of the copies, in particular that their maximum queue lengths are small. Our approach is to investigate the equilibrium distribution first, as well as the time for a copy of the process from a fairly general starting state to reach a "good" set of states in which the equilibrium copy spends most of its time. Having done this, we then prove rapid mixing in a very similar way to the proof in [11].

To show anything about the equilibrium distribution, we would like to examine the trajectory of the vector $u(X_t)$, whose components are the $u_i(X_t)$ for $i \geq 1$. This seems difficult to do directly, but we perform a change of variables and analyse instead a collection of just k functions $Q_1(X_t), \ldots, Q_k(X_t)$. These are linear functions of $u_1(X_t), \ldots, u_k(X_t)$, with the property that the drift of each $Q_j(X_t)$ can be written, approximately, in terms of $Q_j(X_t)$

and $Q_{j+1}(X_t)$ only. Exceptionally, the drift of $Q_k(X_t)$ is written in terms of $Q_k(X_t)$ and $u_{k+1}(X_t)$ (which in fact is usually zero in equilibrium). The particular forms of the Q_j are chosen by considering the Perron–Frobenius eigenvalues of certain matrices M_k derived from the drifts of the $u_j(x)$. Making this change of variables allows us to consider one function $Q_j(X_t)$ at a time, and show that each in turn drifts towards its equilibrium mean (which is derived from the fixed point π of (1.1)), and we are thus able to prove enough about the trajectory of the $Q_j(X_t)$ to show that, starting from any reasonable state, with high probability the chain soon enters a good set of states where, in particular, $u_{k+1}(X_t) = 0$, and so the maximum queue length is at most k. We also show that, with high probability, the chain remains in this good set of states for a long time, which implies that the equilibrium copy spends the vast majority of its time in this set. The argument from [11] about coalescence of coupled copies can be used to show rapid mixing from this good set of states. The drift of the function Q_k to its equilibrium is slower than that of any other Q_j, and its drift rate is approximately $n^{-1-(k-1)\beta}$, which is close to the spectral gap of the Markov chain (X_t), and hence determines the speed of mixing in Theorem 1.2.

The structure of the paper is as follows. In Sect. 2, we expand on the discussion above, and motivate the definitions of the functions $Q_j : \mathbb{Z}_+^n \to \mathbb{R}$, which are fundamental to the proof. In Sect. 3, we give a number of results about the long-term behaviour of random walks with drifts, including several variants on results from [11]. In Sect. 4, we describe the key coupling from [11], and use it to prove some results about the maximum queue length and number of customers. In Sect. 5, we discuss in detail the drifts of the functions Q_j. The proof of Theorem 1.1 starts in Sect. 6, where we show how to derive a slightly stronger result from a sequence of lemmas. These lemmas are proved in Sects. 7–9. We prove our results on mixing times in Sect. 10.

Note this paper is heavily based on a manuscript [3] by the first and third named authors, placed on the arXiv in 2012, but not published in any other outlet. The present paper also incorporates results from the second author's PhD thesis [6]. The results proved in the present paper are in some sense weaker than those in [3] and [6], as, purely for the sake of exposition, we only treat the case where $1 - \lambda(n)$ and $d(n)$ are powers of n, and state our results only in asymptotic form. In a more important sense, our results here are stronger, as they cover essentially best possible ranges of exponents; the key improvement in our methodology compared to [3] is that here we state and use Lemma 3.2 in a form where we get a stronger bound when a function on the state space stays the same with high probability at any step, allowing us to take proper account of the fact that the Q_j for $j < k$ rarely change value. Our intention is to update [3] to incorporate these improvements in our more general setting.

2 Heuristics

In this section, we set out the intuition behind our results and proofs. As before, let (Y_t) be an equilibrium copy of the (n, d, λ)-supermarket process. Guided by the results in [6,11], we start by supposing that, for each $i \geq 1$, $u_i(Y_t)$ is well-concentrated around its expectation u_i, and seeing what that implies about the u_i. For a function F defined on the state space, and a state x, we define the *drift* of F at x to be $\Delta F(x) = \mathbb{E}[F(X_{t+1}) - F(X_t) \mid X_t = x]$, which is independent of t. We have

$$\begin{aligned}\Delta u_i(Y_t) &= \mathbb{E}\left[u_i(Y_{t+1}) - u_i(Y_t) \mid Y_t\right] \\ &= \frac{1}{n(1+\lambda)}\Big[\lambda u_{i-1}(Y_t)^d - \lambda u_i(Y_t)^d - u_i(Y_t) + u_{i+1}(Y_t)\Big]. \qquad (2.1)\end{aligned}$$

To see this, observe that, for $i \geq 1$, conditioned on Y_t, the probability that the event at time $t+1$ is an arrival to a queue of length exactly $i-1$, increasing u_i by $1/n$, is $\frac{\lambda}{1+\lambda}\left(u_{i-1}(Y_t)^d - u_i(Y_t)^d\right)$, while the probability that the event is a departure from a queue of length exactly i, decreasing u_i by $1/n$, is $\frac{1}{1+\lambda}\left(u_i(Y_t) - u_{i+1}(Y_t)\right)$. Note that u_0 is identically equal to 1.

Taking expectations on both sides, and setting them to 0, we see that, since (Y_t) is in equilibrium,

$$0 = \mathbb{E}\left[u_i(Y_{t+1}) - u_i(Y_t)\right] \simeq \frac{1}{n(1+\lambda)}\left[\lambda u_{i-1}^d - \lambda u_i^d - u_i + u_{i+1}\right], \tag{2.2}$$

where the approximations $\mathbb{E}\, u_i(Y_t)^d \simeq u_i^d$ and $\mathbb{E}\, u_{i-1}(Y_t)^d \simeq u_{i-1}^d$ are justified because of our assumption that $u_i(Y_t)$ and $u_{i-1}(Y_t)$ are well-concentrated around their respective means u_i and u_{i-1}.

The system of equations

$$0 = \lambda\pi_{i-1}^d - \lambda\pi_i^d - \pi_i + \pi_{i+1} \quad (i = 1, 2, \dots), \tag{2.3}$$

with $\pi_0 = 1$, has a unique solution with $\pi_i \to 0$ as $i \to \infty$, namely:

$$\pi_i = \lambda^{1+\dots+d^{i-1}} \quad (i = 0, 1, \dots),$$

as in (1.2). See [11] and the references therein for details.

By analogy with [11], and motivated by (2.2), if the $u_i(Y_t)$ are well concentrated, we expect that $u_i \approx \pi_i$, for each i, and moreover that the values of $u_i(Y_t)$ will be close to the corresponding π_i with high probability. In the regime of Theorem 1.1,

$$\log \pi_i = \log(1-(1-\lambda))(1+\dots+d^{i-1}) \simeq -n^{-\alpha+(i-1)\beta},$$

for each $i \geq 1$. As we are assuming that $(k-1)\beta < \alpha < k\beta$, this means that π_i is close to 1 for $i \leq k$, and very close to 0 for $i > k$. In particular, π_{k+1} (which we expect to be the approximate proportion of queues of length greater than k) is much smaller than $1/n$, suggesting that, in equilibrium, the probability that there is a queue of length greater than k is very small.

On the other hand, the fact that π_k is close to 1 suggests that, in equilibrium, most queues have length exactly k. Moreover, $\pi_i^d = 1 - o(1)$ for $i < k$, so that $1 - \pi_i^d \approx d(1-\pi_i)$, whereas $\pi_k^d = o(1)$. We then obtain the following linear approximation to the equations (2.3), written in terms of variables $1-\tilde{u}_1, \dots, 1-\tilde{u}_k$:

$$\begin{aligned}
0 &= \lambda d(1-\tilde{u}_1) + (1-\tilde{u}_1) - (1-\tilde{u}_2),\\
0 &= -\lambda d(1-\tilde{u}_{i-1}) + \lambda d(1-\tilde{u}_i) + (1-\tilde{u}_i) - (1-\tilde{u}_{i+1})\\
&\qquad (2 \leq i \leq k-1),\\
0 &= -\lambda d(1-\tilde{u}_{k-1}) + (1-\tilde{u}_k) - (1-\lambda).
\end{aligned}$$

These linear equations have solution $\tilde{u}$ given by

$$1-\tilde{u}_i = (1-\lambda)(1+(\lambda d)+\dots+(\lambda d)^{i-1}) \quad (i = 1, \dots, k).$$

We then have the further approximation

$$1-\tilde{u}_i \approx (1-\lambda)(\lambda d)^{i-1}, \quad (i = 1, \dots, k),$$

and we aim to show that indeed each $u_i(x)$ is close to the corresponding $\tilde{u}_i$ with high probability in equilibrium.

Ideally, we would seek a single "Lyapunov" function of the $u_i(x)$, which is small when $u_i(x) \approx \tilde{u}_i$ for each i, and larger otherwise, and which has a downward drift outside of a small neighbourhood of $\tilde{u}$: we could then analyse the trajectory of this function to show that $(u_1(x), \ldots, u_k(x))$ stays close to $\tilde{u}$ for a long period. We have been unable to find such a function, and indeed analysing the evolution of the $u_i(X_t)$ directly appears to be challenging. Instead, we work with a sequence of functions $Q_j(x)$, $j = 1, \ldots, k$, each of the form $Q_j(x) = n \sum_{i=1}^{j} \gamma_{j,i}(1 - u_i(x))$, where the $\gamma_{j,i}$ are positive real coefficients. This sequence of functions has the property that the drift of each $Q_j(x)$ can be written (approximately) in terms of $Q_j(x)$ itself and $Q_{j+1}(x)$.

Let us see how these coefficients should be chosen, starting with the special case $j = k$, where we write γ_i for $\gamma_{k,i}$. Consider a function of the form $Q_k(x) = n \sum_{i=1}^{k} \gamma_i(1 - u_i(x))$. As in the argument leading to (2.1), we have that the drift of this function satisfies

$$\begin{aligned}(1+\lambda)\Delta Q_k(x) &= -(1+\lambda)n \sum_{i=1}^{k} \gamma_i \Delta u_i(x) \\ &= -\sum_{i=1}^{k} \gamma_i [\lambda u_{i-1}(x)^d - \lambda u_i(x)^d - u_i(x) + u_{i+1}(x)] \\ &= \sum_{i=1}^{k} \gamma_i [\lambda(1 - u_{i-1}(x)^d) - \lambda(1 - u_i(x)^d) \\ &\quad -(1 - u_i(x)) + (1 - u_{i+1}(x))].\end{aligned}$$

Making the approximations $u_k(x)^d \simeq 0$, and $1 - u_i(x)^d \simeq d(1 - u_i(x))$ for $i = 1, \ldots, k-1$, and rearranging, we arrive at

$$\begin{aligned}(1+\lambda)\Delta Q_k(x) \simeq{}& \gamma_k(1 - \lambda - u_{k+1}(x)) + (\gamma_{k-1} - \gamma_k)(1 - u_k(x)) \\ &+ \sum_{i=1}^{k-1} [\lambda d(\gamma_{i+1} - \gamma_i) - \gamma_i + \gamma_{i-1}](1 - u_i(x)).\end{aligned} \tag{2.4}$$

We set $\gamma_0 = 0$ for convenience of writing the above expression. This calculation is done carefully, with precise inequalities, in Lemma 5.1 below. We would like to choose the γ_i so that the vector

$$\big(\lambda d(\gamma_2 - \gamma_1) - \gamma_1 + \gamma_0, \ldots, \lambda d(\gamma_k - \gamma_{k-1}) - \gamma_{k-1} + \gamma_{k-2}, \gamma_{k-1} - \gamma_k\big) \tag{2.5}$$

of coefficients of the $(1 - u_i)$ in (2.4) is equal to some multiple $-\mu\big(\gamma_1, \ldots, \gamma_{k-1}, \gamma_k\big)$ of the vector with components γ_i, with $\mu > 0$. This would entail

$$(1+\lambda)\Delta Q_k(x) \simeq \gamma_k(1 - \lambda - u_{k+1}(x)) - \frac{\mu Q_k(x)}{n},$$

which in turn would mean that Q_k drifts towards a value of $\gamma_k(1 - \lambda - u_{k+1}(x))n/\mu$. If also $u_{k+1}(x)$ is (nearly) equal to 0, we should obtain that $Q_k(x)$ approaches $\gamma_k(1 - \lambda)n/\mu$ – if Q_k is above this value then it drifts down, whereas if Q_k is below then it drifts up. What we need in order for the vector (2.5) to be a multiple of $\big(\gamma_1, \ldots, \gamma_k\big)$ is for $\big(\gamma_1, \ldots, \gamma_k\big)$ to be a left eigenvector of the $k \times k$ matrix

$$M_k = \begin{pmatrix} -\lambda d - 1 & 1 & 0 & \cdots & 0 & 0 & 0 \\ \lambda d & -\lambda d - 1 & 1 & \cdots & 0 & 0 & 0 \\ 0 & \lambda d & -\lambda d - 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda d - 1 & 1 & 0 \\ 0 & 0 & 0 & \cdots & \lambda d & -\lambda d - 1 & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda d & -1 \end{pmatrix},$$

with eigenvalue $-\mu$, or, equivalently, of the matrix

$$M_k' = M_k + (\lambda d + 1) I_k = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ \lambda d & 0 & 1 & \cdots & 0 & 0 & 0 \\ 0 & \lambda d & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & 0 & \cdots & \lambda d & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda d & \lambda d \end{pmatrix}.$$

The non-negative matrix M_k' has a unique largest "Perron–Frobenius" eigenvalue, with a positive left eigenvector. By inspection, we see that, for $k \geq 2$, this left eigenvector is close to the all-1 vector, with an eigenvalue close to $\lambda d + 1$, so that M_k has largest eigenvalue very close to 0. Recursion shows that a better approximation to the Perron–Frobenius left eigenvector of M_k' is $(\gamma_1, \ldots, \gamma_k)$, where

$$\gamma_i = 1 - \frac{1}{(\lambda d)^i} - \frac{(i-1)}{(\lambda d)^k},$$

for $i = 1, \ldots, k$, and the largest eigenvalue μ of M_k is very close to $-1/(\lambda d)^{k-1}$. We shall see in Lemma 5.1 that this approximation is close enough for our purposes, enabling us to show that, with these choices of the γ_i,

$$(1+\lambda)\Delta Q_k(x) \simeq (1-\lambda) - \frac{Q_k(x)}{n(\lambda d)^{k-1}},$$

and thus $Q_k(x)$ drifts towards a value close to $(1-\lambda)n(\lambda d)^{k-1}$. A further consequence is that, in order for $Q_k(x)$ to move from $(1 \pm 2\varepsilon)(1-\lambda)n(\lambda d)^{k-1}$ to $(1 \pm \varepsilon)(1-\lambda)n(\lambda d)^{k-1}$, it has to travel a distance of $\varepsilon(1-\lambda)n(\lambda d)^{k-1}$ while drifting at rate no greater than $2\varepsilon(1-\lambda)$, and so time of order $n(\lambda d)^{k-1}$ is required. This is then a lower bound on the mixing time from a "good" state to equilibrium, nearly matching that in Theorem 1.2. We make this argument precise at the very end of the paper.

For $1 \leq j < k$, if $Q_j(x) = n \sum_{i=1}^{j} \gamma_{j,i}(1-u_i)$, then a similar analysis reveals that

$$(1+\lambda)\Delta Q_j(x) \simeq \sum_{i=1}^{j} (1-u_i(x)) \left[\gamma_{j,i-1} + \lambda d \gamma_{j,i+1} - (\lambda d + 1)\gamma_{j,i}\right] + (1 - u_{j+1}(x)).$$

(See the proof of Lemma 5.2.) We think of $1 - u_{j+1}(x)$ as an "external" term (which in practice will be very close to $Q_{j+1}(x)/n$), which will determine the value towards which Q_j drifts. We would like the rest of the expression to be a negative multiple of $Q_j(x)$. For this we need $(\gamma_{j,1}, \ldots, \gamma_{j,j})$ to be a left eigenvector of the $j \times j$ matrix

$$M_j = \begin{pmatrix} -\lambda d - 1 & 1 & 0 & \cdots & 0 & 0 & 0 \\ \lambda d & -\lambda d - 1 & 1 & \cdots & 0 & 0 & 0 \\ 0 & \lambda d & -\lambda d - 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda d - 1 & 1 & 0 \\ 0 & 0 & 0 & \cdots & \lambda d & -\lambda d - 1 & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda d & -\lambda d - 1 \end{pmatrix},$$

with eigenvalue $-\mu < 0$ or, equivalently, of the matrix

$$M'_j = M_j + (\lambda d + 1) I_j = \begin{pmatrix} 0 & \lambda d & 0 & \cdots & 0 & 0 & 0 \\ 1 & 0 & \lambda d & \cdots & 0 & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \lambda d & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 & \lambda d \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \end{pmatrix},$$

with eigenvalue $\lambda d + 1 - \mu$. These matrices are tridiagonal Toeplitz matrices, and there is an exact formula for the eigenvalues and eigenvectors. (See, for instance, Example 7.2.5 in [16].) The Perron–Frobenius eigenvalue of M'_j is $2\sqrt{\lambda d}\cos\left(\frac{\pi}{j+1}\right)$, with left eigenvector $\left(\gamma_{j,1}, \ldots, \gamma_{j,j}\right)$ given by

$$\gamma_{j,i} = (\lambda d)^{(j-i)/2} \frac{\sin\left(\frac{i\pi}{j+1}\right)}{\sin\left(\frac{j\pi}{j+1}\right)}.$$

This means that the largest eigenvalue of M_j is $-\lambda d + O(\sqrt{\lambda d})$, so that we obtain

$$(1+\lambda)\Delta Q_j(x) \simeq -\lambda d \frac{Q_j(x)}{n} + \frac{Q_{j+1}(x)}{n} \quad (1 \le j < k),$$

meaning that $Q_j(x)$ will drift to a value close to $Q_{j+1}(x)/\lambda d$. The choices of coefficients ensure that, if the $u_j(x)$ are all near to $\tilde{u}_j \simeq 1 - (1-\lambda)(\lambda d)^{j-1}$, then

$$Q_j(x) = n \sum_{i=1}^{j} \gamma_{j,i}(1 - u_i(x)) \simeq n(1-\lambda) \sum_{i=1}^{j} \frac{\sin\left(\frac{i\pi}{j+1}\right)}{\sin\left(\frac{j\pi}{j+1}\right)} (\lambda d)^{i-1+(j-i)/2},$$

and the top term $i = j$ dominates the rest of the sum, provided λd is large, so $Q_j(x) \simeq (1 - u_j(x))$: this is also true for $j = k$. Thus the relationship $Q_j \simeq Q_{j+1}/\lambda d$ is as we would expect.

This means that, if $Q_{j+1}(X_t)$ remains in an interval around $\tilde{Q}_{j+1} := n(1-\lambda)(\lambda d)^j$ for a long time, then $Q_j(X_t)$ will enter some interval around $\tilde{Q}_j$ within a short time, and stay there for a long time. We can then conduct the analysis for each Q_j in turn, starting with $j = k$, to show that indeed all the $Q_j(X_t)$ quickly become close to $\tilde{Q}_j$, and stay close to $\tilde{Q}_j$ for a long time. This will then imply that the $u_j(X_t)$ all become and remain close to $\tilde{u}_j$.

A subsidiary application of this same technique forms another important step in the proofs (see the proof of Lemma 6.5(1)). If we do not assume that $u_{k+1}(x)$ is zero, but instead build this term into our calculations, we obtain the approximation

$$(1+\lambda)\Delta Q_k(x) \simeq (1-\lambda-u_{k+1}(x)) - \frac{Q_k(x)}{(\lambda d)^{k-1}n}.$$

If $u_{k+1}(X_t)$ remains above $\varepsilon(1-\lambda)$, for some $\varepsilon > 0$, for a long time, this drift equation tells us that Q_k drifts down into an interval whose upper end is below the value $\tilde{Q}_k$, and then each of the Q_j in turn drift down into intervals whose upper ends are below the corresponding $\tilde{Q}_j$, and remain there. For $j = 1$, this means that the number of empty queues is at most $(1-\delta)(1-\lambda)n$, for some positive δ, for a long period of time; this results in a persistent drift down in the total number of customers (since the departure rate is bounded below by $n-(1-\delta)(1-\lambda)n = \lambda n + \delta(1-\lambda)n$ while the arrival rate is λn), and this is not possible.

3 Random Walks with Drifts

In this section, we state some general results about the long-term behaviour of real-valued functions of a Markov chain with bounds on the drift. These are variants of results of Luczak and McDiarmid [11] and Brightwell and Luczak [2], and we do not give the proofs in full detail.

We start with a lemma concerning random walks with a drift, adapted from a result of Luczak and McDiarmid [11]. We have a sequence (R_t) of real-valued random variables; on some "good" event, the jumps $Z_t = R_t - R_{t-1}$ have magnitude at most 1, and expectation at most $-v < 0$. The lemma shows that, on the good event, with high probability, such a random walk, started at some value r_0, hits a lower value r_1 after not too many more than $(r_0 - r_1)/v$ steps.

Lemma 3.1 *Let $\varphi_0 \subseteq \varphi_1 \subseteq \cdots \subseteq \varphi_m$ be a filtration, and let $Z_1, \dots, Z_m$ be random variables taking values in $[-1, 1]$ such that each Z_i is φ_i-measurable. Let $E_0, E_1, \dots, E_{m-1}$ be events where $E_i \in \varphi_i$ for each i, and let $E = \bigcap_{i=0}^{m-1} E_i$. Fix $v \in (0,1)$, and let $r_0, r_1 \in \mathbb{R}$ be such that $r_0 > r_1$ and $vm \geq 2(r_0 - r_1)$. Set $R_0 = r_0$ and, for each integer $t > 0$, let $R_t = R_0 + \sum_{i=1}^{t} Z_i$.*

Suppose that, for each $i = 1, \dots, m$,

$$\mathbb{E}(Z_i \mid \varphi_{i-1}) \leq -v \text{ on } E_{i-1} \cap \{R_{i-1} > r_1\}.$$

Then

$$\mathbb{P}(E \cap \{R_t > r_1 \quad \forall t \in \{1, \dots, m\}\}) \leq \exp\Big(-\frac{v^2 m}{8}\Big).$$

We omit the proof, which is similar to one in [11].

For a discrete-time Markov process (X_t) with state space $\mathcal{X}$, a real-valued function F defined on $\mathcal{X}$, and an element x of $\mathcal{X}$, we define

$$\Delta F(x) := \mathbb{E}[F(X_{t+1}) - F(X_t) \mid X_t = x],$$

and call this the *drift* of F (at x). Similarly, we shall also use the notation $\Delta F(X_t)$ to denote the random variable $\mathbb{E}[F(X_{t+1}) - F(X_t) \mid X_t]$.

The next lemma says that, if the function F has a negative drift of magnitude at least $v > 0$ on a good set $\mathcal{U}$, and makes jumps of size at most 1, then it is unlikely to increase by a large positive value before leaving $\mathcal{U}$.

Lemma 3.2 *Let a, v and p be positive real numbers, with $v \leq p \leq 1$. Let $(X_t)_{t\geq 0}$ be a discrete-time Markov process with state-space $\mathcal{X}$, adapted to the filtration $(\varphi_t)_{t\geq 0}$. Let F be a* *real-valued function on $\mathcal{X}$ such that,* $|F(y)-F(x)| \le 1$ *whenever* $\mathbb{P}(X_{i+1} = y \mid X_i = x) > 0$. *Let $\mathcal{U}$ be a subset of $\mathcal{X}$ such that, for $x \in \mathcal{U}$, $\Delta F(x) \le -v$ and $\mathbb{P}(F(X_{i+1}) = F(X_i) \mid X_i = x) \ge 1-p$. Let $T_{\mathcal{U}} = \inf\{t : X_t \notin \mathcal{U}\}$, and let $T = \inf\{t : F(X_t) - F(X_0) \ge a\}$ Then*

$$\mathbb{P}(T \le T_{\mathcal{U}}) \le \frac{100}{v^2} e^{-va/4p}.$$

Proof (Sketch) We use Theorem 2.5 of [2], applied to the function F. Translated into our setting, that result says that, for all $t \ge 0$, and all $\omega > 0$,

$$\mathbb{P}(\{T_{\mathcal{U}} \ge t\} \cap \{F(X_t) - F(X_0) + vt > \max(\sqrt{\omega p t}, \omega)\} \mid \varphi_0) \le 2e^{-\omega/4}.$$

For each t, we choose $\omega(t) = \frac{v}{2p}(2a+|vt-a|)$. It is easy to verify that $\max(\sqrt{\omega(t)pt}, \omega(t)) < vt + a$ for each t (note that the hypotheses imply that $v \le p$). Therefore

$$\mathbb{P}(T < T_{\mathcal{U}} \mid \varphi_0) \le 2\sum_{t=0}^{\infty} e^{-\omega(t)/4} \le 4e^{-va/4p} \sum_{i=0}^{\infty} e^{-v^2 i/8p} \le \frac{100}{v^2} e^{-va/4p},$$

as desired. □

We now use the two lemmas above to prove a result about real-valued functions of a Markov chain which we shall use repeatedly in our proofs.

Lemma 3.3 *Let h, v, c, $\rho \ge 2$, m, $p \le 1$ and s be positive real numbers with $vm \ge 2(c-h)$. Let $(X_t)_{t\ge0}$ be a discrete-time Markov process with state-space $\mathcal{X}$, adapted to the filtration $(\varphi_t)_{t\ge0}$. Let $\mathcal{S}$ be a subset of $\mathcal{X}$, and let F be a real-valued function on $\mathcal{X}$ such that, for all $x \in \mathcal{S}$ with $F(x) \ge h$,*

$$\Delta F(x) \le -v, \text{ and } \mathbb{P}(F(X_{t+1}) \ne F(X_t) \mid X_t = x) \le p,$$

and for all $t \ge 0$, $|F(X_{t+1}) - F(X_t)| \le 1$ a.s. Let T^* be any stopping time, and suppose that $F(X_{T^*}) \le c$ a.s.

Let

$$\begin{aligned} T_0 &= \inf\{t \ge T^* : X_t \notin \mathcal{S}\},\\ T_1 &= \inf\{t \ge T^* : F(X_t) \le h\},\\ T_2 &= \inf\{t > T_1 : F(X_t) \ge h + \rho\}. \end{aligned}$$

Then

(i) $\mathbb{P}(T_1 \wedge T_0 > T^* + m) \le \exp(-v^2 m/8)$;

(ii) $\mathbb{P}(T_2 \le s \wedge T_0) \le \dfrac{100s}{v^2} \exp(-\rho v/8p)$.

When we use the lemma, m will be much smaller than s, with high probability T^* will be much smaller than s, and also $\mathbb{P}(T_0 \le s)$ will be small. In these circumstances, the lemma allows us to conclude that $\mathbb{P}(T_1 > T^* + m)$ and $\mathbb{P}(T_2 \le s)$ are small. This means that, with high probability, $F(X_t)$ decreases from its value at T^* (at most c) to below h in at most a further m steps, and does not increase back above $h + \rho$ before time s. We shall sometimes use the conclusion of (ii) in the weaker form $\mathbb{P}(T_2 \le s < T_0) \le \frac{100s}{v^2} \exp(-\rho v/8p)$. For most uses of part (ii), we shall simply set $p = 1$, but on occasion we need to use the stronger result in cases where the function F rarely changes value.

Proof We start by proving the lemma in the special case where the stopping time T^* is equal to 0.

For (i), we apply Lemma 3.1. The filtration $\varphi_0 \subseteq \varphi_1 \subseteq \cdots \subseteq \varphi_m$ will be the initial segment of the filtration $(\varphi_t)_{t\geq 0}$. For $t \geq 1$, we set $Z_t = F(X_t) - F(X_{t-1})$, so that $R_t := R_0 + \sum_{i=1}^t Z_i = F(X_t)$. For $t \geq 0$, we set E_t to be the event that $T_0 > t$ (i.e., $X_i \in \mathcal{S}$ for all i with $0 \leq i \leq t$), so $E = \bigcap_{i=0}^{m-1} E_i$ is the event that $T_0 \geq m$. We set $r_0 = F(X_0) \leq c$, and $r_1 = h$. We may assume that $r_0 > r_1$; otherwise $T_1 = 0$ and there is nothing to prove.

On the event $E_{i-1} \cap \{R_{i-1} > r_1\}$, we have $X_{i-1} \in \mathcal{S}$ and $F(X_{i-1}) > r_1 = h$, so $\mathbb{E}(Z_i \mid \varphi_{i-1}) = \Delta F(X_{i-1}) \leq -v$. Thus, noting that $vm \geq 2(r_0 - r_1)$ by our assumption on m, we see that the conditions of Lemma 3.1 are satisfied. The event that $R_t > r_1$ for all $t = 1, \ldots, m$ is the event that $T_1 > m$, so

$$\mathbb{P}(T_1 \wedge T_0 > m) \leq \mathbb{P}(\{T_1 > m\} \cap \{T_0 \geq m\}) \leq e^{-v^2 m/8},$$

as required for (i).

We move on to (ii). For each time $r \in \{0, \ldots, s-1\}$, set

$$T(r) = \min\{t \geq 0 : F(X_{r+t}) \notin [h, h+\rho)\}.$$

We say that r is a *departure point* if: $T_1 \leq r$, $F(X_r) \in [h, h+1)$, $F(X_{r+T(r)}) \geq h+\rho$, and $r + T(r) \leq s \wedge T_0$. To say that $T_2 \leq s \wedge T_0$ means that $F(X_t)$ crosses from its value, at most h, at time T_1, up to a value at least $h+\rho$, taking steps of size at most 1, by time $s \wedge T_0$. This is equivalent to saying that there is at least one departure point $r \in [0, s)$. Therefore

$$\begin{aligned}\mathbb{P}(T_2 \leq s \wedge T_0) \leq \sum_{r=0}^{s-1} \mathbb{P}\Big(&\{T_1 \leq r\} \cap \{F(X_r) \in [h, h+1)\} \\ &\cap \{F(X_{r+T(r)}) \geq h+\rho\} \cap \{r + T(r) \leq s \wedge T_0\}\Big) \\ = \sum_{r=0}^{s-1} \mathbb{E}\Big[&\mathbb{1}_{\{T_1 \leq r\}} \mathbb{1}_{\{F(X_r) \in [h,h+1)\}}\, \mathbb{E}\big[\mathbb{1}_{\{F(X_{r+T(r)}) \geq h+\rho\}} \mathbb{1}_{\{r+T(r) \leq s \wedge T_0\}} \mid \varphi_r\big]\Big].\end{aligned}$$

Fix any $r \in [0, s)$. We claim that, for any $h_0 \in [h, h+1)$, on the φ_r-measurable event that $F(X_r) = h_0$, the conditional expectation

$$\mathbb{E}\big[\mathbb{1}_{\{F(X_{r+T(r)}) \geq h+\rho\}} \mathbb{1}_{\{r+T(r) \leq s \wedge T_0\}} \mid \varphi_r\big]$$

is at most $\frac{100}{v^2} e^{-\rho v/8p}$. This will imply that each term of the sum above is at most $\frac{100}{v^2} e^{-\rho v/8p}$, and so that $\mathbb{P}(T_2 \leq s \wedge T_0) \leq \frac{100s}{v^2} \exp(-\rho v/8p)$, as required.

To prove the claim, we use Lemma 3.2. We consider the re-indexed process $(X'_t) = (X_{r+t})$; by the Markov property, this is a Markov chain with the same transition probabilities as (X_t), and initial state $X'_0 = X_r$ with $F(X'_0) = h_0$. We set $\varphi'_i = \varphi_{r+i}$ for each i, so that (X'_i) is adapted to the filtration (φ'_i). We set $a = h + \rho - h_0 \geq \rho - 1 \geq \rho/2$. We also set $\mathcal{U} = \mathcal{S} \cap \{x : F(x) \geq h\}$, $T_{\mathcal{U}} = \inf\{i : X'_i \notin \mathcal{U}\}$, and $T = \inf\{i : F(X_{r+i}) \geq a\}$. Therefore, if $r + T(r) \leq T_0$ and $F(X_{r+T(r)}) \geq h+\rho$, then $T = T(r) \leq T_{\mathcal{U}}$.

For $i \leq T_{\mathcal{U}}$, we have $X'_{i-1} = X_{r+i-1} \in \mathcal{S}$ and $F(X'_{i-1}) \geq h$, and therefore $\Delta F(X'_{i-1}) \leq -v$, and also $\mathbb{P}(F(X'_i) = F(X'_{i-1}) \mid \varphi_{i-1}) \geq 1 - p$. From Lemma 3.2, we now conclude that, on the event $F(X_r) = h_0$,

$$\begin{aligned}&\mathbb{P}\Big(\{F(X_{r+T(r)}) \geq h+\rho\} \cap \{r + T(r) \leq s \leq T_0\} \Big| \varphi_r\Big) \\ &\leq \mathbb{P}\Big(T \leq T_{\mathcal{U}}\Big) \leq \frac{100}{v^2} e^{-va/4p} \leq \frac{100}{v^2} e^{-\rho v/8p},\end{aligned}$$

as required. This completes the proof in the special case where $T^* = 0$.

We now proceed to the general case. Suppose then that the hypotheses of the lemma are satisfied, with stopping time T^*. We apply the result we have just proved to the process $(X'_t) = (X_{T^*+t})$. By the strong Markov property, (X'_t) is also a Markov process, adapted to the filtration $(\varphi'_t)_{t\geq 0} = (\varphi_{T^*+t})_{t\geq 0}$. The condition that $F(X_{T^*}) \leq c$ is equivalent to $F(X'_0) \leq c$. Set:

$$\begin{aligned}
T'_0 &= \inf\{t \geq 0 : X'_t \notin \mathcal{S}\} = \inf\{t \geq 0 : X_{T^*+t} \notin \mathcal{S}\} = T_0 - T^*\\
T'_1 &= \inf\{t \geq 0 : F(X'_t) \leq h\} = \inf\{t \geq 0 : F(X_{T^*+t}) \leq h\} = T_1 - T^*\\
T'_2 &= \inf\{t > T'_1 : F(X'_t) \geq h + \rho\} = \inf\{t > T'_1 : F(X_{T^*+t}) \geq h + \rho\} = T_2 - T^*,
\end{aligned}$$

and note that these are all stopping times with respect to the filtration (φ'_t). The special case of the result (with $T^* = 0$) now tells us that:

$$\begin{aligned}
\text{(i)}\quad \mathbb{P}(T_1 \wedge T_0 > T^* + m) &= \mathbb{P}((T^* + T'_1) \wedge (T^* + T'_0) > T^* + m)\\
&= \mathbb{P}(T'_1 \wedge T'_0 > m)\\
&\leq \exp(-v^2 m/8);\\
\text{(ii)}\quad \mathbb{P}(T_2 \leq s \wedge T_0) &= \mathbb{P}(T^* + T'_2 \leq s \wedge (T^* + T'_0))\\
&\leq \mathbb{P}(T^* + T'_2 \leq (T^* + s) \wedge (T^* + T'_0))\\
&= \mathbb{P}(T'_2 \leq s \wedge T'_0)\\
&\leq \frac{100p}{v^2} \exp(-\rho v/8p).
\end{aligned}$$

In both cases, these are the desired results. □

We also use a "reversed" version of Lemma 3.3 where $\Delta F(x) \geq v$ for all x in some "good" set $\mathcal{S}$ with $F(x) \leq h$. The result and proof are practically identical to Lemma 3.3, changing the directions of inequalities where necessary, and using "reversed" versions of Lemmas 3.1 and 3.2.

The next lemma is a more precise version of Lemma 2.2 in [11]. We omit the proof, which is exactly as in [11], except that we track more carefully the values of the various constants appearing in that proof, and separate out the effects of the two occurrences of δ in that theorem. We will use this result in our proof of Lemma 10.1, showing rapid mixing.

Lemma 3.4 *Let $(\varphi_t)_{t\geq 0}$ be a filtration. Let $Z_1, Z_2, \ldots$ be $\{0, \pm 1\}$-valued random variables, where each Z_i is φ_i-measurable. Let $S_0 \geq 0$ a.s., and for each positive integer j let $S_j = S_0 + \sum_{i=1}^{j} Z_i$. Let $A_0, A_1, \ldots$ be events, where each A_i is φ_i-measurable.*

Suppose that there is a positive integer k_0 and a constant δ with $0 < \delta < 1/2$ such that $\mathbb{P}(Z_i = -1 \mid \varphi_{i-1}) \geq \delta$ on $A_{i-1} \cap \{S_{i-1} \in \{1, \ldots, k_0 - 1\}\}$ and $\mathbb{P}(Z_i = -1 \mid \varphi_{i-1}) \geq 3/4$ on $A_{i-1} \cap \{S_{i-1} \geq k_0\}$. Then, for each positive integer m

$$\mathbb{P}\left(\bigcap_{i=1}^{m}\{S_i \neq 0\} \cap \bigcap_{i=0}^{m-1} A_i\right) \leq \mathbb{P}(S_0 > \lfloor m/16 \rfloor) + 3\exp\left(-\frac{\delta^{k_0-1}}{200k_0} m\right).$$

Several times we shall use the fact that, if Z is a binomial or Poisson random variable with mean μ, then for each $0 \leq \epsilon \leq 1$ we have

$$\mathbb{P}(Z - \mu \leq -\epsilon\mu) \leq e^{-(1/2)\epsilon^2\mu}. \tag{3.1}$$

4 Coupling

We now introduce a natural coupling of copies of the (n, d, λ)-supermarket process (X_t^x) with different initial states x. The coupling is a natural adaptation to discrete time of that in [11]. In this section, we make no assumptions about the values of the parameters n, λ and d.

We describe the coupling in terms of three independent sequences of random variables. There is an iid sequence $\mathbf{V} = (V_1, V_2, \dots)$ of 0–1 random variables where each V_i takes value 1 with probability $\lambda/(1+\lambda)$; $V_i = 1$ if and only if time i is an arrival. Corresponding to every time i there is also an ordered list D_i of d queue indices, each chosen uniformly at random with replacement. Let $\mathbf{D} = (D_1, D_2, \dots)$. Furthermore, corresponding to every time i there is a uniformly chosen queue index $\tilde{D}_i$. Let $\tilde{\mathbf{D}} = (\tilde{D}_1, \tilde{D}_2, \dots)$. At time i, D_i will be used if $Z_i = 1$, and there will be an arrival to the first shortest queue in D_i; otherwise, there will be a departure from the queue with index $\tilde{D}_i$, if that queue is currently non-empty.

Suppose that we are given a realisation $(\mathbf{v}, \mathbf{d}, \tilde{\mathbf{d}})$ of $(\mathbf{V}, \mathbf{D}, \tilde{\mathbf{D}})$. For each possible initial queue-lengths vector $x \in \mathbb{Z}_+^n$, this realisation yields a deterministic process (x_t) with $x_0 = x$: let us write $x_t = s_t(x; \mathbf{v}, \mathbf{d}, \tilde{\mathbf{d}})$. Then, for each $x \in \mathbb{Z}_+^n$, the process $s_t(x; \mathbf{V}, \mathbf{D}, \tilde{\mathbf{D}})$ has the distribution of the (n, d, λ)-supermarket process X_t^x with initial state x. In this way, we construct copies (X_t^x) of the (n, d, λ)-supermarket process for each possible starting state x on a single probability space. When we treat more than one such copy at the same time, we always work in this probability space, and we let $\mathbb{P}(\cdot)$ denote the corresponding coupling measure.

We shall use the following lemma, which is a discrete-time analogue of Lemma 2.3 in [11] and is proved in exactly the same way.

Lemma 4.1 *Fix any triple* $\mathbf{z}, \mathbf{d}, \tilde{\mathbf{d}}$ *as above, and for each queue-lengths vector* x*, write* $s_t(x)$ *for* $s_t(x; \mathbf{z}, \mathbf{d}, \tilde{\mathbf{d}})$*. Then, for each* $x, y \in \mathbb{Z}_+^n$*, both* $\|s_t(x) - s_t(y)\|_1$ *and* $\|s_t(x) - s_t(y)\|_\infty$ *are nonincreasing; and further, if* $0 \le t < t'$ *and* $s_t(x) \le s_t(y)$*, then* $s_{t'}(x) \le s_{t'}(y)$*.*

Given positive real numbers ℓ and b, we set

$$\mathcal{A}_0(\ell, b) = \{x : \|x\|_\infty \le \ell \text{ and } \|x\|_1 \le bn\};$$
$$\mathcal{A}_1(\ell, b) = \{x : \|x\|_\infty \le 3\ell \text{ and } \|x\|_1 \le 3bn\}.$$

We also set

$$\ell^* = (1-\lambda)^{-1} \log^2 n,\ b^* = 2(1-\lambda)^{-1},\ \mathcal{A}_0 = \mathcal{A}_0(\ell^*, b^*),\ \mathcal{A}_1 = \mathcal{A}_1(\ell^*, b^*).$$

Thus a state x is in $\mathcal{A}_0$ if there are at most $2n(1-\lambda)^{-1}$ customers in total, and no more than $(1-\lambda)^{-1} \log^2 n$ in any queue. These requirements are relaxed by a factor of 3 in $\mathcal{A}_1$.

The next result tells us that the (n, d, λ)-supermarket process (Y_t), in equilibrium, is very unlikely to be outside the set $\mathcal{A}_0$, for any d. This is accomplished by proving the result for $d = 1$, when the process is easy to analyse explicitly, and then using coupling in d to deduce the result for all d. Of course, the result is actually extremely weak for all $d > 1$, and later we shall show a much stronger result whenever the various parameters of the model satisfy the conditions of Theorem 1.1; the importance of the lemma below is that it gets us started and enables us to say *something* about where the equilibrium of the process lives.

Lemma 4.2 *Let* (Y_t) *be a copy of the* (n, d, λ)*-supermarket process in equilibrium. Then* $\mathbb{P}(Y_t \notin \mathcal{A}_0) \le 2ne^{-\log^2 n}$*.*

Proof Let $\tilde{Y}$ denote a stationary copy of the $(n, 1, \lambda)$-supermarket process, in which each arriving customer joins a uniform random queue. Then the queue lengths $\tilde{Y}_t(j)$ are independent geometric random variables with mean $\lambda/(1-\lambda)$, where $\mathbb{P}(\tilde{Y}_t(j) = r) = (1-\lambda)\lambda^r$ for $r = 0, 1, 2, \ldots$. Therefore, $\mathbb{P}(\|\tilde{Y}_t\|_\infty \geq r) \leq n\lambda^r$, and also it can easily be checked that $\mathbb{P}\left(\|\tilde{Y}_t\|_1 \geq 2n(1-\lambda)^{-1}\right) \leq e^{-n/4}$.

As mentioned in the remarks after Theorem 1.1, there is a coupling between supermarket processes with different values of d, which can be used to show that the equilibrium copy (Y_t) of the (n, d, λ)-supermarket process, for any d, also satisfies $\mathbb{P}\left(\|Y_t\|_1 \geq 2n(1-\lambda)^{-1}\right) \leq e^{-n/4}$ and $\mathbb{P}(\|Y_t\|_\infty \geq \log^2 n(1-\lambda)^{-1}) \leq n\lambda^{\log^2 n(1-\lambda)^{-1}} \leq ne^{-\log^2 n}$, as required. □

Next we prove a very crude concentration of measure result: if the process (Y_t) in equilibrium is concentrated inside some set $\mathcal{A}_0(\ell, b)$, and we start a copy (X_t^x) of the process at a state $x \in \mathcal{A}_0(\ell, b)$, then the process (X_t^x) is unlikely to leave the larger set $\mathcal{A}_1(\ell, b)$ over a long period of time.

Lemma 4.3 *Let ℓ and b be natural numbers and x a queue-lengths vector in $\mathcal{A}_0(\ell, b)$. Let (Y_t) be a copy of the (n, d, λ)-supermarket process in equilibrium, and let (X_t^x) be a copy started in state x. Then for any natural number s,*

$$\mathbb{P}(\exists t \in [0, s],\ X_t^x \notin \mathcal{A}_1(\ell, b)) \leq \mathbb{P}(\exists t \in [0, s],\ Y_t \notin \mathcal{A}_0(\ell, b)).$$

Proof By Lemma 4.1, we can couple (X_t^x) and (Y_t) in such a way that $\|X_t^x - Y_t\|_1$ and $\|X_t^x - Y_t\|_\infty$ are both non-increasing, and hence that, for each $t \geq 0$,

$$\begin{aligned}\|X_t^x\|_1 &\leq \|X_t^x - Y_t\|_1 + \|Y_t\|_1 \leq \|x - Y_0\|_1 + \|Y_t\|_1\\ &\leq \|x\|_1 + \|Y_0\|_1 + \|Y_t\|_1 \leq bn + \|Y_0\|_1 + \|Y_t\|_1,\end{aligned}$$

and similarly

$$\|X_t^x\|_\infty \leq \ell + \|Y_0\|_\infty + \|Y_t\|_\infty.$$

We deduce that, for each $t \geq 0$,

$$\begin{aligned}\{X_t^x \notin \mathcal{A}_1(\ell, b)\} &= \{\|X_t^x\|_1 > 3bn\} \cup \{\|X_t^x\|_\infty > 3\ell\}\\ &\subseteq \{\|Y_0\|_1 > bn\} \cup \{\|Y_t\|_1 > bn\} \cup \{\|Y_0\|_\infty > \ell\} \cup \{\|Y_t\|_\infty > \ell\}\\ &= \{Y_0 \notin \mathcal{A}_0(\ell, b)\} \cup \{Y_t \notin \mathcal{A}_0(\ell, b)\}.\end{aligned}$$

The result now follows immediately. □

We shall use Lemma 4.3 later for general values of ℓ and b, but for now we note the following immediate consequence of the previous two lemmas. Let $T^\dagger_{\mathcal{A}} = T^\dagger_{\mathcal{A}}(x) = \inf\{t : X_t^x \notin \mathcal{A}_1\}$: this will be an instance of a more general notation we introduce later: when we have a pair of sets $\mathcal{S}_0 \subseteq \mathcal{S}_1$, we will use $T_{\mathcal{S}}$ to denote the first time we enter the inner set, and $T^\dagger_{\mathcal{S}}$ to denote the first time after $T_{\mathcal{S}}$ that we leave the outer one.

Lemma 4.4 *Let x be any queue-lengths vector in $\mathcal{A}_0$. Then, for n sufficiently large,*

$$\mathbb{P}\left(T^\dagger_{\mathcal{A}}(x) \leq e^{\frac{1}{3}\log^2 n}\right) \leq e^{-\frac{1}{2}\log^2 n}.$$

Proof The probability in question is $\mathbb{P}(\exists t \in [0, e^{\frac{1}{3}\log^2 n}],\ X_t^x \notin \mathcal{A}_1)$ which, by Lemma 4.3 and Lemma 4.2, is at most

$$\mathbb{P}(\exists t \in [0, e^{\frac{1}{3}\log^2 n}],\ Y_t \notin \mathcal{A}_0) \leq (e^{\frac{1}{3}\log^2 n} + 1)\,\mathbb{P}(Y_t \notin \mathcal{A}_0^*) \leq 3ne^{-\frac{2}{3}\log^2 n},$$

which, for n sufficiently large, is at most $e^{-\frac{1}{2}\log^2 n}$, as required. □

5 Functions and Drifts

We now start the detailed proofs of our main results.

As explained in Sect. 2, we will consider a sequence of functions Q_k, Q_{k-1}, …, Q_1 defined on the set $\mathbb{Z}_+^n$ of queue-lengths vectors. We now give precise definitions of these functions, along with another function P_{k-1}, and derive some of their properties.

The results in this section will be used in the course of the proof of Theorem 1.1, and we could assume that we are in the regime covered by our theorem; however, for this section all that is necessary is that $\lambda d \geq 16$. In the special case $k = 1$, we need only consider the function $Q_k = Q_1$ and its drift; otherwise we assume that $k \geq 2$.

As in Sect. 2, let Q_k be the function defined on the set $\mathbb{Z}_+^n$ of all queue-lengths vectors by

$$Q_k(x) = n \sum_{i=1}^{k} \gamma_i (1 - u_i(x)),$$

where, for $i = 1, \ldots, k$,

$$\gamma_i = 1 - \frac{1}{(\lambda d)^i} - \frac{i-1}{(\lambda d)^k}.$$

It is also convenient to set $\gamma_0 = 0$. Evidently $\gamma_i < 1$ for each i, an inequality we shall use freely in future. We also note that, provided $\lambda d > 2$,

$$\gamma_{i+1} - \gamma_i = \frac{1}{(\lambda d)^i} - \frac{1}{(\lambda d)^{i+1}} - \frac{1}{(\lambda d)^k}, \tag{5.1}$$

for $i = 0, \ldots, k-1$. Therefore γ_i is increasing in i; also $\gamma_k = 1 - k(\lambda d)^{-k}$.

If $k \geq 2$, we set $P_{k-1}(x) = n \sum_{i=1}^{k-1} (1 - u_i(x))$. Also, for $j = 1, \ldots, k-1$, we let $Q_j(x) = n \sum_{i=1}^{j} \gamma_{j,i} (1 - u_i(x))$, where the coefficients $\gamma_{j,i}$ are given by

$$\gamma_{j,i} = (\lambda d)^{(j-i)/2} \frac{\sin\left(\frac{i\pi}{j+1}\right)}{\sin\left(\frac{j\pi}{j+1}\right)}.$$

Consistent with the expression above, we also define $\gamma_{j,0} = \gamma_{j,j+1} = 0$. It can easily be checked that, for each $i = 1, \ldots, j-1$, and for each $j = 1, \ldots, k-1$,

$$\lambda d \gamma_{j,i+1} + \gamma_{j,i-1} = 2\sqrt{\lambda d} \cos\left(\frac{\pi}{j+1}\right) \gamma_{j,i}.$$

This is equivalent to saying that the $\gamma_{j,i}$ form eigenvectors of the tridiagonal Toeplitz matrices M_j given in Sect. 2, with eigenvalue $-\lambda d - 1 + 2\sqrt{\lambda d} \cos\left(\frac{\pi}{j+1}\right)$.

We will need some bounds on the sizes of the $Q_j(x)$, for $j < k$. Observe that $\gamma_{j,j} = 1$ for each j, while generally we have

$$1 \leq \frac{\sin(i\pi/(j+1))}{\sin(j\pi/(j+1))} = \frac{\sin(i\pi/(j+1))}{\sin(\pi/(j+1))} \leq i, \tag{5.2}$$

since the sine function is concave on $[0, \pi]$. Thus $(\lambda d)^{(j-i)/2} \leq \gamma_{j,i} \leq i(\lambda d)^{(j-i)/2}$ and therefore

$$Q_j(x) \leq n \sum_{i=1}^{j} i(\lambda d)^{(j-i)/2} \leq \frac{n}{(1 - 1/\sqrt{\lambda d})^2} \leq 2n(\lambda d)^{(j-1)/2}, \tag{5.3}$$

provided $\lambda d \geq 16$. We also note at this point that changing one component $x(\ell)$ of x by ±1 changes $Q_j(x)$ by at most $\gamma_{j,1} = (\lambda d)^{(j-1)/2}$.

It can readily be checked that, for $j \geq 1$, the function

$$f(i) = \sin\left(\frac{i\pi}{j+2}\right) \Big/ \sin\left(\frac{i\pi}{j+1}\right)$$

is increasing over the range $[1, j]$, and so we have, for $1 \leq i \leq j \leq k-2$:

$$\begin{aligned}\frac{\gamma_{j+1,i}}{\gamma_{j,i}} &= \sqrt{\lambda d}\frac{\sin(i\pi/(j+2))\sin(\pi/(j+1))}{\sin(i\pi/(j+1))\sin(\pi/(j+2))}\\ &\leq \sqrt{\lambda d}\frac{\sin(j\pi/(j+2))\sin(\pi/(j+1))}{\sin(j\pi/(j+1))\sin(\pi/(j+2))}\\ &= \sqrt{\lambda d}\frac{\sin(2\pi/(j+2))}{\sin(\pi/(j+2))} \leq 2\sqrt{\lambda d},\end{aligned}$$

using (5.2) for the final inequality. A consequence is that, for $j = 1, \ldots, k-2$, and any $x \in \mathbb{Z}_+^n$,

$$\begin{aligned}\frac{Q_{j+1}(x)}{n} &= (1 - u_{j+1}(x)) + \sum_{i=1}^{j} \gamma_{j+1,i}(1 - u_i(x))\\ &\leq (1 - u_{j+1}(x)) + \sum_{i=1}^{j} 2\sqrt{\lambda d}\gamma_{j,i}(1 - u_i(x))\\ &\leq (1 - u_{j+1}(x)) + 2\sqrt{\lambda d}\frac{Q_j(x)}{n}.\end{aligned} \tag{5.4}$$

For $j = k-1$, we have the stronger inequality that, for any $x \in \mathbb{Z}_+^n$,

$$\frac{Q_k(x)}{n} \leq \sum_{i=1}^{k}(1 - u_i(x)) \leq (1 - u_k(x)) + \frac{Q_{k-1}(x)}{n}. \tag{5.5}$$

We now prove that the drift of the function $Q_k(x)$ is approximately equal to $\frac{1}{1+\lambda}\left(1 - \lambda - u_{k+1}(x) - \frac{1}{(\lambda d)^{k-1}}\frac{Q_k(x)}{n}\right)$.

Lemma 5.1 *If $k \geq 2$, then, for any state $x \in \mathbb{Z}_+^n$,*

$$\begin{aligned}(1+\lambda)\Delta Q_k(x) &\leq \gamma_k\big((1-\lambda) - u_{k+1}(x) + \lambda\exp(-dQ_k(x)/kn)\big)\\ &\quad - \frac{1}{(\lambda d)^{k-1}}\frac{Q_k(x)}{n}\left(1 - \frac{2}{\lambda d}\right),\\ (1+\lambda)\Delta Q_k(x) &\geq \gamma_k\big((1-\lambda) - u_{k+1}(x)\big) - \frac{1}{(\lambda d)^{k-1}}\frac{Q_k(x)}{n}\\ &\quad - \left(\frac{Q_{k-1}(x)}{n}\right)^2\frac{1}{(\lambda d)^{k-3}}.\end{aligned}$$

For $k = 1$, we have

$$\begin{aligned}(1+\lambda)\Delta Q_1(x) &\leq \gamma_1\big((1-\lambda) - u_2(x) + \lambda\exp(-dQ_1(x)/n)\big) - \frac{Q_1(x)}{n},\\ (1+\lambda)\Delta Q_1(x) &\geq \gamma_1\big((1-\lambda) - u_2(x)\big) - \frac{Q_1(x)}{n}.\end{aligned}$$

Proof As in (2.1), we have that, for $i = 1, \dots, k$,

$$\Delta u_i(x) = \frac{1}{n(1+\lambda)} \Big(\lambda u_{i-1}(x)^d - \lambda u_i(x)^d - u_i(x) + u_{i+1}(x) \Big).$$

and that u_0 is identically equal to 1. We deduce that

$$\begin{aligned} \Delta Q_k(x) &= -n \sum_{i=1}^{k} \gamma_i \Delta u_i(x) \\ &= \frac{1}{1+\lambda} \sum_{i=1}^{k} \gamma_i \Big(-\lambda u_{i-1}(x)^d + \lambda u_i(x)^d + u_i(x) - u_{i+1}(x) \Big). \end{aligned}$$

We rearrange the formula above as follows:

$$\begin{aligned} (1+\lambda)\Delta Q_k(x) &= \gamma_k \Big((1-\lambda) + \lambda u_k(x)^d - u_{k+1}(x) + \lambda(1 - u_{k-1}(x)^d) - (1 - u_k(x)) \Big) \\ &\quad + \sum_{i=1}^{k-1} \gamma_i \Big(\lambda(1 - u_{i-1}(x)^d) - \lambda(1 - u_i(x)^d) \\ &\quad - (1 - u_i(x)) + (1 - u_{i+1}(x)) \Big) \\ &= \gamma_k \Big((1-\lambda) + \lambda u_k(x)^d - u_{k+1}(x) \Big) \\ &\quad + \lambda \sum_{i=1}^{k-1} (\gamma_{i+1} - \gamma_i)(1 - u_i(x)^d) - \sum_{i=1}^{k} (\gamma_i - \gamma_{i-1})(1 - u_i(x)). \end{aligned}$$

Here we have used the facts that $\gamma_0 = 0$ and $1 - u_0(x) = 0$.

Now, for $1 \le i \le k$, we have $1 - u_i(x) \le 1 - u_k(x)$ for all x, and $\gamma_i \le 1$. Therefore $Q_k(x) \le nk(1 - u_k(x))$, and hence

$$0 \le u_k(x)^d \le \left(1 - \frac{Q_k(x)}{kn} \right)^d \le \exp(-d Q_k(x)/kn).$$

For $k \ge 2$, in order to estimate the terms constituting the two sums, we note the inequalities $d(1-u) - \binom{d}{2}(1-u)^2 \le 1 - u^d \le d(1-u)$. To obtain our upper bound on $\Delta Q_k(x)$, we apply the inequality $1 - u_i(x)^d \le d(1 - u_i(x))$ for each $i = 1, \dots, k-1$. Using also (5.1), we have

$$\begin{aligned} &\lambda \sum_{i=1}^{k-1} (\gamma_{i+1} - \gamma_i)(1 - u_i(x)^d) - \sum_{i=1}^{k} (\gamma_i - \gamma_{i-1})(1 - u_i(x)) \\ &\le \lambda d \sum_{i=1}^{k-1} (\gamma_{i+1} - \gamma_i)(1 - u_i(x)) - \sum_{i=1}^{k} (\gamma_i - \gamma_{i-1})(1 - u_i(x)) \\ &= - \left[\frac{1}{(\lambda d)^{k-1}} - \frac{2}{(\lambda d)^k} \right] (1 - u_k(x)) \\ &\quad + \sum_{i=1}^{k-1} \left[\frac{\lambda d}{(\lambda d)^i} - \frac{\lambda d}{(\lambda d)^{i+1}} - \frac{\lambda d}{(\lambda d)^k} - \frac{1}{(\lambda d)^{i-1}} + \frac{1}{(\lambda d)^i} + \frac{1}{(\lambda d)^k} \right] (1 - u_i(x)) \end{aligned}$$

$$= -\frac{1}{(\lambda d)^{k-1}} \left[\left(1 - \frac{2}{\lambda d}\right)(1 - u_k(x)) + \sum_{i=1}^{k-1} \left(1 - \frac{1}{\lambda d}\right)(1 - u_i(x)) \right]$$
$$\leq -\frac{1}{(\lambda d)^{k-1}} \frac{Q_k(x)}{n} \left(1 - \frac{2}{\lambda d}\right).$$

This establishes the required upper bound on $(1+\lambda)\Delta Q_k(x)$. The calculation works because the γ_i are the entries of a good approximation to the Perron–Frobenius eigenvector of the matrix M_k defined in Sect. 2.

For the lower bound, the previous calculation, and the bound $1 - u_i(x)^d \geq d(1-u) - \binom{d}{2}(1-u)^2$, lead us to

$$\lambda \sum_{i=1}^{k-1} (\gamma_{i+1} - \gamma_i)(1 - u_i(x)^d) - \sum_{i=1}^{k} (\gamma_i - \gamma_{i-1})(1 - u_i(x))$$
$$\geq -\lambda \binom{d}{2} \sum_{i=1}^{k-1} (\gamma_{i+1} - \gamma_i)(1 - u_i(x))^2$$
$$- \frac{1}{(\lambda d)^{k-1}} \left[\left(1 - \frac{2}{\lambda d}\right)(1 - u_k(x)) + \sum_{i=1}^{k-1} \left(1 - \frac{1}{\lambda d}\right)(1 - u_i(x)) \right]$$
$$\geq -\lambda \binom{d}{2} \sum_{i=1}^{k-1} (\gamma_{i+1} - \gamma_i)(1 - u_i(x))^2 - \frac{1}{(\lambda d)^{k-1}} \frac{Q_k(x)}{n}.$$

Here we used the fact that $1 - 1/(\lambda d) \leq \gamma_i$ for each i.

It remains to show that

$$\lambda \binom{d}{2} \sum_{i=1}^{k-1} (\gamma_{i+1} - \gamma_i)(1 - u_i(x))^2 \leq \left(\frac{Q_{k-1}(x)}{n}\right)^2 \frac{1}{(\lambda d)^{k-3}}.$$

We observe that

$$\left(\frac{Q_{k-1}(x)}{n}\right)^2 = \left(\sum_{i=1}^{k-1} (\lambda d)^{(k-1-i)/2} \frac{\sin\left(\frac{i\pi}{k}\right)}{\sin\left(\frac{(k-1)\pi}{k}\right)} (1 - u_i(x)) \right)^2$$
$$\geq \sum_{i=1}^{k-1} (\lambda d)^{k-1-i} (1 - u_i(x))^2$$
$$\geq (\lambda d)^{k-1} \sum_{i=1}^{k-1} (\gamma_{i+1} - \gamma_i)(1 - u_i(x))^2,$$

which implies the required inequality.

In the special case $k = 1$, the equation for the drift reduces to

$$(1+\lambda)\Delta Q_1(x) = \gamma_1(1 - \lambda - u_2(x)) - \frac{Q_1(x)}{n} + \gamma_1 \lambda u_1(x)^d,$$

and both the required bounds follow immediately. □

We prove a similar result for the functions $Q_j(x)$, $1 \leq j \leq k-1$. Ideally, the drift bounds would be expressed in terms of $Q_j(x)$ itself and $Q_{j+1}(x)$: however, there is a complication.

In the upper bound, there appears a term which can be bounded above by $\lambda\binom{d}{2}\sum_{i=1}^{j}\gamma_{j,i}(1-u_i(x))^2$, and we would like to show that this is small compared with $\lambda d\sum_{i=1}^{j}\gamma_{j,i}(1-u_i(x))$. This is true if $1-u_j(x)\ll 1/d$, but in general we cannot assume this. We bound this term above, very crudely, by

$$\lambda\binom{d}{2}\left(\sum_{i=1}^{k-1}(1-u_i(x))\right)\left(\sum_{i=1}^{j}\gamma_{j,i}(1-u_i(x))\right)=\lambda\binom{d}{2}\frac{P_{k-1}(x)Q_j(x)}{n^2};$$

we use the function P_{k-1} here because its drifts are relatively easy to handle.

Lemma 5.2 *Fix j with $1\le j\le k-1$. For any state $x\in\mathbb{Z}_+^n$, we have*

$$(1+\lambda)\Delta Q_j(x)\le -\lambda d\frac{Q_j(x)}{n}\left(1-\frac{2}{\sqrt{\lambda d}}-\frac{dP_{k-1}(x)}{n}\right)+\frac{Q_{j+1}(x)}{n},$$
$$(1+\lambda)\Delta Q_j(x)\ge -\lambda d\frac{Q_j(x)}{n}\left(1+\frac{2}{\sqrt{\lambda d}}\right)+\frac{Q_{j+1}(x)}{n}.$$

Proof We begin by calculating

$$\begin{aligned}(1+\lambda)\Delta Q_j(x)&=\sum_{i=1}^{j}\gamma_{j,i}\Big(-\lambda u_{i-1}(x)^d+\lambda u_i(x)^d+u_i(x)-u_{i+1}(x)\Big)\\&=\sum_{i=1}^{j}\gamma_{j,i}\Big(\lambda(1-u_{i-1}(x)^d)-\lambda(1-u_i(x)^d)\Big)+\sum_{i=1}^{j}\gamma_{j,i}\Big(-(1-u_i(x))\\&\quad+(1-u_{i+1}(x))\Big).\end{aligned}$$

Rearranging now gives

$$\begin{aligned}(1+\lambda)\Delta Q_j(x)&=\sum_{i=1}^{j}(\gamma_{j,i-1}-\gamma_{j,i})(1-u_i(x))\\&\quad-\lambda\sum_{i=1}^{j}(\gamma_{j,i}-\gamma_{j,i+1})(1-u_i(x)^d)+\gamma_{j,j}(1-u_{j+1}(x)).\end{aligned}$$

Recall that $\gamma_{j,0}=\gamma_{j,j+1}=0$, and note that $\gamma_{j,1}>\gamma_{j,2}>\cdots>\gamma_{j,j}=1$.

As before, we proceed by approximating $1-u_i(x)^d$ by $d(1-u_i(x))$, for $i\le j$. Using first that $1-u_i(x)^d\le d(1-u_i(x))$ for each i, we have

$$\begin{aligned}(1+\lambda)\Delta Q_j(x)&\ge\sum_{i=1}^{j}(\gamma_{j,i-1}-\gamma_{j,i})(1-u_i(x))-\lambda d\sum_{i=1}^{j}(\gamma_{j,i}-\gamma_{j,i+1})(1-u_i(x))\\&\quad+(1-u_{j+1}(x))\\&=\sum_{i=1}^{j}(1-u_i(x))\left[\gamma_{j,i-1}+\lambda d\gamma_{j,i+1}-(\lambda d+1)\gamma_{j,i}\right]+(1-u_{j+1}(x))\\&=-\sum_{i=1}^{j}(1-u_i(x))\gamma_{j,i}\left[\lambda d+1-2\sqrt{\lambda d}\cos\left(\frac{\pi}{j+1}\right)\right]+(1-u_{j+1}(x))\end{aligned}$$

$$
\begin{aligned}
&= -\left[\lambda d + 1 - 2\sqrt{\lambda d}\cos\left(\frac{\pi}{j+1}\right)\right]\frac{Q_j(x)}{n} + (1 - u_{j+1}(x))\\
&\geq -\lambda d\frac{Q_j(x)}{n} + \frac{Q_{j+1}(x)}{n} - 2\sqrt{\lambda d}\frac{Q_j(x)}{n},
\end{aligned}
$$

as claimed. In the last line above, we used (5.4), as well as the inequality $2\sqrt{\lambda d}\cos(\pi/(j+1)) \geq \sqrt{2\lambda d} \geq 1$, valid since $\lambda d \geq 16$.

For the upper bound, we use the facts that $1 - u_{j+1}(x) \leq \frac{Q_{j+1}(x)}{n}$ and $1 - u_i(x)^d \geq d(1 - u_i(x)) - \binom{d}{2}(1 - u_i(x))^2$, to obtain

$$
\begin{aligned}
(1+\lambda)\Delta Q_j(x) &\leq -\left[\lambda d + 1 - 2\sqrt{\lambda d}\cos\left(\frac{\pi}{j+1}\right)\right]\frac{Q_j(x)}{n} + (1 - u_{j+1}(x))\\
&\quad + \lambda\binom{d}{2}\sum_{i=1}^{j}(\gamma_{j,i} - \gamma_{j,i+1})(1 - u_i(x))^2\\
&\leq -\lambda d\frac{Q_j(x)}{n}\left(1 - \frac{2}{\sqrt{\lambda d}}\right) + \frac{Q_{j+1}(x)}{n}\\
&\quad + \frac{P_{k-1}(x)}{n}\lambda\binom{d}{2}\sum_{i=1}^{j}\gamma_{j,i}(1 - u_i(x)).
\end{aligned}
$$

This is the result we require, since $\sum_{i=1}^{j}\gamma_{j,i}(1 - u_i(x)) = Q_j(x)/n$. □

We have a similar result for the function P_{k-1}. For this function, we need only a fairly crude upper bound on the drift, and we omit the simple proof.

Lemma 5.3 *For any state $x \in \mathbb{Z}_+^n$, we have*

$$
(1+\lambda)\Delta P_{k-1}(x) \leq -\frac{\lambda d P_{k-1}(x)}{(k-1)n} + \frac{Q_k(x)}{n}.
$$

6 Hitting Times and Exit Times

At this point, we begin the proof of Theorem 1.1. Accordingly, from now on we fix values of $\alpha, \beta \in (0, 1)$, and a natural number k, satisfying $(k-1)\beta < \alpha < k\beta$ and $2\alpha < 1 + (k-1)\beta$. Throughout the proof, we consider the (n, d, λ)-supermarket model with $\lambda = 1 - n^{-\alpha}$ and $d = n^{\beta}$. (As mentioned in the Introduction, our proofs go through essentially unchanged if we assume only that $1 - \lambda = n^{-\alpha+\delta_1(n)}$ and $d = n^{\beta+\delta_2(n)}$, where $\delta_1(n)$ and $\delta_2(n)$ tend to zero as $n \to \infty$, where we replace the expression $n^{-\alpha+(j-1)\beta}$ in the definition of $\mathcal{N}^{\varepsilon}(n, \alpha, \beta)$ below by $(1 - \lambda)d^{j-1}$.)

We shall actually prove a result stronger than Theorem 1.1, in that we replace the "tolerance" $1/\log n$ in that result by a general function $\varepsilon = \varepsilon(n)$. We assume that $\varepsilon(n) \leq 1/100$, with $1/\varepsilon(n) = o(n^{\delta})$ for every $\delta > 0$, though in fact the proof goes through even if $\varepsilon(n)$ tends to zero as $n^{-\delta}$ for δ sufficiently small (in terms of α and β).

Accordingly, given $\alpha, \beta \in (0, 1)$, and $\varepsilon = \varepsilon(n)$ as above, set $k = \lceil \alpha/\beta \rceil$ as usual, and let $\mathcal{N}^{\varepsilon}(n, \alpha, \beta)$ be the set of queue-lengths vectors x such that $u_{k+1}(x) = 0$ and, for $1 \leq j \leq k$,

$$
(1 - 6\varepsilon)n^{-\alpha+(j-1)\beta} \leq 1 - u_j(x) \leq (1 + 6\varepsilon)n^{-\alpha+(j-1)\beta}.
$$

Theorem 6.1 *Suppose that $\alpha, \beta \in (0, 1]$ are constants with $k-1 < \alpha/\beta < k$ for some natural number k, and that $2\alpha < 1 + \beta(k-1)$. Suppose also that $\lambda = \lambda(n) = 1 - n^{-\alpha}$ and $d = d(n) = n^{\beta}$. Let $\varepsilon = \varepsilon(n) \le 1/100$ be any function such that $\varepsilon(n)^{-1} = o(n^{\delta})$ for every $\delta > 0$. Then, for n sufficiently large, a copy (Y_t) of the (n, d, λ)-supermarket process in equilibrium satisfies*

$$\mathbb{P}\left(Y_t \notin \mathcal{N}^{\varepsilon}(n, \alpha, \beta)\right) \le e^{-\frac{1}{4}\log^2 n}.$$

Moreover, if $X_0 \in \mathcal{N}^{\varepsilon/6}(n, \alpha, \beta)$, then

$$\mathbb{P}\left(X_t \notin \mathcal{N}^{\varepsilon}(n, \alpha, \beta) \text{ for some } t \in [0, e^{\frac{1}{3}\log^2 n}]\right) \le e^{-\frac{1}{4}\log^2 n}.$$

Theorem 1.1 is the case of Theorem 6.1 with $\varepsilon = 1/6 \log n$.

The assumptions of Theorem 6.1 assure us that functions of n such as $\varepsilon^{-1} n^{-\alpha+(k-1)\beta} \log^2 n$ tend to zero, as the dominant term is the strictly negative power of n. We shall use such facts freely throughout the proof, and we shall (sometimes tacitly) assume that n is sufficiently large.

We define a sequence of pairs of subsets of $\mathbb{Z}_+^n$. Each pair consists of a set $\mathcal{S}_0$ in which some inequality holds, and a set $\mathcal{S}_1$ in which a looser version of the inequality holds: we also demand that $\mathcal{S}_0$ and $\mathcal{S}_1$ be subsets of the previous set $\mathcal{R}_1$ in the sequence. Associated with each pair $(\mathcal{S}_0, \mathcal{S}_1)$ in the sequence is a *hitting time*

$$T_{\mathcal{S}} = \inf\{t \ge T_{\mathcal{R}} : X_t \in \mathcal{S}_0\},$$

where $(\mathcal{R}_0, \mathcal{R}_1)$ is the previous pair in the sequence, and an *exit time*

$$T_{\mathcal{S}}^{\dagger} = \inf\{t \ge T_{\mathcal{S}} : X_t \notin \mathcal{S}_1\}.$$

Our aim in each case is to prove that, with high probability, unless the previous exit time $T_{\mathcal{R}}^{\dagger}$ occurs early, $T_{\mathcal{S}}$ is unlikely to be larger than some quantity $m_{\mathcal{S}}$ whose order is polynomial in n. To be precise, if we start in a state in $\mathcal{A}_0(\ell, b)$, then the sum of all the $m_{\mathcal{S}}$ is of order at most the maximum of $bn^{1+\alpha}$ and ℓn, so if ℓ and b are bounded by a polynomial in n, then so are all the $m_{\mathcal{S}}$.

Throughout the proof, we set

$$s_0 = e^{\frac{1}{3}\log^2 n}.$$

We shall also prove that, again with high probability, each exit time $T_{\mathcal{S}}^{\dagger}$ is at least s_0, which is larger than the sum of all the terms $m_{\mathcal{S}}$. For convenience, we shall not be too precise about our error probabilities, and simply declare them all to be at most $1/s_0 = e^{-\frac{1}{3}\log^2 n}$, or some small multiple of $1/s_0$. We will thus prove that, with high probability, we enter each of the sets $\mathcal{S}_0$ in turn, while remaining inside all the earlier sets $\mathcal{S}_1$.

We fix, for the moment, a pair of positive real numbers ℓ and b with $\ell \ge b \ge k$. We set

$$q(\ell, b) = (22k + 72b)\varepsilon^{-1} n^{1+\alpha} + 8\ell n,$$

and we make the (mild) assumption that $\ell \le e^{\frac{1}{4}\log^2 n}$, so that $q(\ell, b) \le s_0/2$.

The first pair of sets in our sequence will be as defined earlier:

$$\mathcal{A}_0 = \mathcal{A}_0(\ell, b) = \{x : \|x\|_\infty \le \ell \text{ and } \|x\|_1 \le bn\},$$
$$\mathcal{A}_1 = \mathcal{A}_1(\ell, b) = \{x : \|x\|_\infty \le 3\ell \text{ and } \|x\|_1 \le 3bn\},$$

and we adopt the hypothesis that $X_0 = x_0$ almost surely, where x_0 is a fixed state in $\mathcal{A}_0 = \mathcal{A}_0(\ell, b)$, so that $T_{\mathcal{A}} := \min\{t \ge 0 : X_t \in \mathcal{A}_0\} = 0$.

For $\ell = \ell^* = n^\alpha \log^2 n$ and $b = b^* = 2n^\alpha$, Lemma 4.4 tells us that indeed the exit time $T^\dagger_{\mathcal{A}} = \inf\{t > 0 : X_t \notin \mathcal{A}_1\}$ is unlikely to be less than s_0. For smaller values of ℓ and b, we do not know this a priori.

The sets we define are dependent on the chosen values of n, α, β and ε, as well as on ℓ and b. For the most part, we drop reference to this dependence from the notation. When we need to vary ε while keeping all other parameters fixed, we shall use the notation (e.g.) $\mathcal{B}^\varepsilon_0$ to emphasise the dependence. We define:

$$\begin{aligned}
\mathcal{B}_0 &= \{x : Q_k(x) \le (1+\varepsilon)n(1-\lambda)(\lambda d)^{k-1}\} \cap \mathcal{A}_1,\\
\mathcal{B}_1 &= \{x : Q_k(x) \le (1+2\varepsilon)n(1-\lambda)(\lambda d)^{k-1}\} \cap \mathcal{A}_1,\\
\mathcal{C}_0 &= \{x : P_{k-1}(x) \le 2kn(1-\lambda)(\lambda d)^{k-2}\} \cap \mathcal{B}_1,\\
\mathcal{C}_1 &= \{x : P_{k-1}(x) \le 3kn(1-\lambda)(\lambda d)^{k-2}\} \cap \mathcal{B}_1,\\
\mathcal{D}_0 &= \{x : Q_{k-1}(x) \le (1+4\varepsilon)n(1-\lambda)(\lambda d)^{k-2}\} \cap \mathcal{C}_1,\\
\mathcal{D}_1 &= \{x : Q_{k-1}(x) \le (1+5\varepsilon)n(1-\lambda)(\lambda d)^{k-2}\} \cap \mathcal{C}_1,\\
\mathcal{E}_0 &= \{x : u_{k+1}(x) \le \varepsilon(1-\lambda) \text{ and } Q_k(x) \ge (1-3\varepsilon)n(1-\lambda)(\lambda d)^{k-1}\} \cap \mathcal{D}_1,\\
\mathcal{E}_1 &= \{x : u_{k+1}(x) \le \varepsilon(1-\lambda) \text{ and } Q_k(x) \ge (1-4\varepsilon)n(1-\lambda)(\lambda d)^{k-1}\} \cap \mathcal{D}_1.
\end{aligned}$$

Next we have a sequence of pairs of sets, indexed by $j = k-1, \dots, 1$:

$$\begin{aligned}
\mathcal{G}^j_0 &= \Big\{x : \Big[1 - \Big(4 + \frac{k-j-1/2}{k}\Big)\varepsilon\Big] n(1-\lambda)(\lambda d)^{j-1} \le Q_j(x)\\
&\le \Big[1 + \Big(4 + \frac{k-j-1/2}{k}\Big)\varepsilon\Big] n(1-\lambda)(\lambda d)^{j-1}\Big\} \cap \mathcal{G}^{j+1}_1,\\
\mathcal{G}^j_1 &= \Big\{x : \Big[1 - \Big(4 + \frac{k-j}{k}\Big)\varepsilon\Big] n(1-\lambda)(\lambda d)^{j-1} \le Q_j(x)\\
&\le \Big[1 + \Big(4 + \frac{k-j}{k}\Big)\varepsilon\Big] n(1-\lambda)(\lambda d)^{j-1}\Big\} \cap \mathcal{G}^{j+1}_1.
\end{aligned}$$

where we declare $\mathcal{G}^k_1$ to be equal to $\mathcal{E}_1$. Finally, departing slightly from our pattern, we define

$$\mathcal{H} = \mathcal{H}_0 = \mathcal{H}_1 = \{x : u_{k+1}(x) = 0\} \cap \mathcal{G}^1_1.$$

In the special case $k = 1$, only the pairs $(\mathcal{B}_0, \mathcal{B}_1)$, $(\mathcal{E}_0, \mathcal{E}_1)$ and $\mathcal{H}$ are defined.

The hitting times and exit times are all defined in accordance with the pattern given. For instance $T_{\mathcal{B}} = \inf\{t : X_t \in \mathcal{B}_0\}$, $T^\dagger_{\mathcal{B}} = \inf\{t > T_{\mathcal{B}} : X_t \notin \mathcal{B}_1\}$, and $T_{\mathcal{C}} = \inf\{t \ge T_{\mathcal{B}} : X_t \in \mathcal{C}_0\}$. We also set $T_{\mathcal{G}^k} = T_{\mathcal{E}}$ and $T^\dagger_{\mathcal{G}^k} = T^\dagger_{\mathcal{E}}$, in accordance with the notion that the set pair $(\mathcal{G}^{k-1}_0, \mathcal{G}^{k-1}_1)$ follows $(\mathcal{E}_0, \mathcal{E}_1)$ in the sequence.

Initially, the sets above all depend on the values of ℓ and b defining the initial pair of sets $(\mathcal{A}_0, \mathcal{A}_1)$, since all the sets are intersected with $\mathcal{A}_1$. However, since states in $\mathcal{H}$ have no queue of length $k+1$ or greater, we have $\mathcal{H} \subseteq \mathcal{A}_0(k,k) \subseteq \mathcal{A}_1(\ell, b)$ for all $\ell, b \ge k$, and so the set $\mathcal{H}$ does not depend on ℓ and b, provided these parameters are each at least k.

We claim that $\mathcal{H}^\varepsilon \subseteq \mathcal{N}^\varepsilon = \mathcal{N}^\varepsilon(n, \alpha, \beta)$. Indeed, if $x \in \mathcal{H}^\varepsilon$, then

$$x \in \mathcal{B}_1 \cap \mathcal{D}_1 \cap \mathcal{E}_1 \cap \mathcal{G}^{k-1}_1 \cap \dots \cap \mathcal{G}^1_1 \cap \{x : u_{k+1}(x) = 0\}.$$

This implies that indeed $u_{k+1}(x) = 0$, and also that all the $Q_j(x)$ are within a factor $1 \pm 5\varepsilon$ of the values $n(1-\lambda)(\lambda d)^{j-1}$. It now follows from (5.4) and (5.5) that, for each $j = 1, \dots, k$,

$$\left|\frac{Q_j(x)}{n} - (1 - u_j(x))\right| \le 2(1-\lambda)(\lambda d)^{\frac{1}{2}+j-2} \le \varepsilon(1-\lambda)(\lambda d)^{j-1},$$

and so $1 - u_j(x)$ is within a factor $1 \pm \frac{11}{2}\varepsilon$ of $n(1-\lambda)(\lambda d)^{j-1}$, so that indeed $x \in \mathcal{N}^\varepsilon$.

We now state a sequence of lemmas. Throughout, we assume that $X_0 = x_0$ a.s., where x_0 is an arbitrary state in $\mathcal{A}_0 = \mathcal{A}_0(\ell, b)$.

Lemma 6.2 *Let* $m_\mathcal{B} = 8k\varepsilon^{-1}n(1-\lambda)^{-1}$.

(1) $\mathbb{P}(T_\mathcal{B} \wedge T^\dagger_\mathcal{A} \ge m_\mathcal{B}) \le 1/s_0$.
(2) $\mathbb{P}(T^\dagger_\mathcal{B} \le s_0 < T^\dagger_\mathcal{A}) \le 1/s_0$.

Lemma 6.3 *For* $k \ge 2$, *let* $m_\mathcal{C} = 8kn(1-\lambda)^{-1}(\lambda d)^{1-k}$.

(1) $\mathbb{P}(T_\mathcal{C} \wedge T^\dagger_\mathcal{B} \ge T_\mathcal{B} + m_\mathcal{C}) \le 1/s_0$.
(2) $\mathbb{P}(T^\dagger_\mathcal{C} \le s_0 < T^\dagger_\mathcal{B}) \le 1/s_0$.

Lemma 6.4 *For* $k \ge 2$, *let* $m_\mathcal{D} = 8\varepsilon^{-1}n(1-\lambda)^{-1}(\lambda d)^{-k/2}$.

(1) $\mathbb{P}(T_\mathcal{D} \wedge T^\dagger_\mathcal{C} \ge T_\mathcal{C} + m_\mathcal{D}) \le 1/s_0$.
(2) $\mathbb{P}(T^\dagger_\mathcal{D} \le s_0 < T^\dagger_\mathcal{C}) \le 1/s_0$.

Lemma 6.5 *Let* $m_\mathcal{E} = m_\mathcal{E}(b) = (13k + 72b)\varepsilon^{-1}n(1-\lambda)^{-1}$.

(1) $\mathbb{P}(T_\mathcal{E} \wedge T^\dagger_\mathcal{D} \ge T_\mathcal{D} + m_\mathcal{E}) \le 1/s_0$.
(2) $\mathbb{P}(T^\dagger_\mathcal{E} \le s_0 < T^\dagger_\mathcal{D}) \le 1/s_0$.

Lemma 6.6 *For* $k \ge 2$, *let* $m_\mathcal{G} = 32k\varepsilon^{-1}n(1-\lambda)^{-1}(\lambda d)^{-1}$. *For* $j = k-1, \dots, 1$, *we have:*

(1) *For* $j = k-1, \dots, 1$, $\mathbb{P}(T_{\mathcal{G}^j} \wedge T^\dagger_{\mathcal{G}^{j+1}} \ge T_{\mathcal{G}^{j+1}} + m_\mathcal{G}) \le 1/s_0$*;*
(2) *For* $j = k-1, \dots, 1$, $\mathbb{P}(T^\dagger_{\mathcal{G}^j} \le s_0 < T^\dagger_{\mathcal{G}^{j+1}}) \le 1/s_0$.

Lemma 6.7 *Let* $m_\mathcal{H} = m_\mathcal{H}(\ell) = n(8\ell + 32\log^2 n)$.

(1) $\mathbb{P}(T_\mathcal{H} \wedge T^\dagger_{\mathcal{G}^1} \ge T_{\mathcal{G}^1} + m_\mathcal{H}) \le 1/s_0$.
(2) $\mathbb{P}(T^\dagger_\mathcal{H} \le s_0 < T^\dagger_{\mathcal{G}^1}) \le 1/s_0$.

We shall postpone the proofs of these lemmas to later sections. For the remainder of this section, we show how the lemmas imply Theorem 6.1. To start with, combining the lemmas gives the following result.

Proposition 6.8 *For any* $x_0 \in \mathcal{A}_0 = \mathcal{A}_0(\ell, b)$, *and a copy* (X_t) *of the process with* $X_0 = x_0$ *a.s., we have*

$$\mathbb{P}(X_t \in \mathcal{H} \text{ for all } t \in [q(\ell, b), s_0]) \ge 1 - \frac{2k+8}{s_0} - \mathbb{P}(T^\dagger_\mathcal{A} \le s_0).$$

Proof The idea is that, with high probability, either the chain (X_t) exits $\mathcal{A}_1(\ell, b)$ before time s_0, or the chain enters each of the sets $\mathcal{B}_0, \dots, \mathcal{H}_0$ in turn, within time $q(\ell, b)$, and does not exit any of the sets $\mathcal{A}_1, \dots, \mathcal{H}_1$ before time s_0, which is what we need.

We assume that $k \ge 2$: if $k = 1$, the proof is very similar and shorter. Consider the following list of events concerning the various stopping times we have defined:

$$E_1 = \{T^\dagger_\mathcal{A} > s_0\}, \quad E_2 = \{T_\mathcal{B} \le m_\mathcal{B}\}, \quad E_3 = \{T^\dagger_\mathcal{B} > s_0\},$$
$$E_4 = \{T_\mathcal{C} \le m_\mathcal{B} + m_\mathcal{C}\}, \quad E_5 = \{T^\dagger_\mathcal{C} > s_0\}, \quad E_6 = \{T_\mathcal{D} \le m_\mathcal{B} + m_\mathcal{C} + m_\mathcal{D}\},$$
$$E_7 = \{T^\dagger_\mathcal{D} > s_0\}, \quad E_8 = \{T_\mathcal{E} \le m_\mathcal{B} + \cdots + m_\mathcal{E}\}, \quad E_9 = \{T^\dagger_\mathcal{E} > s_0\},$$

$$E_{10} = \{T_{\mathcal{G}^{k-1}} \le m_{\mathcal{B}} + \cdots + m_{\mathcal{E}} + m_{\mathcal{G}}\}, \quad E_{11} = \{T^{\dagger}_{\mathcal{G}^{k-1}} > s_0\}, \quad \ldots,$$
$$E_{2k+6} = \{T_{\mathcal{G}^1} \le m_{\mathcal{B}} + \cdots + (k-1)m_{\mathcal{G}}\}, \quad E_{2k+7} = \{T^{\dagger}_{\mathcal{G}^1} > s_0\},$$
$$E_{2k+8} = \{T_{\mathcal{H}} \le m_{\mathcal{B}} + \cdots + (k-1)m_{\mathcal{G}} + m_{\mathcal{H}}\}, \quad E_{2k+9} = \{T^{\dagger}_{\mathcal{H}} > s_0\}.$$

If E_{2k+8} holds, then

$$\begin{aligned} T_{\mathcal{H}} &\le m_{\mathcal{B}} + m_{\mathcal{C}} + m_{\mathcal{D}} + m_{\mathcal{E}} + (k-1)m_{\mathcal{G}} + m_{\mathcal{H}} \\ &= 8k\varepsilon^{-1}n(1-\lambda)^{-1} + 8kn(1-\lambda)^{-1}(\lambda d)^{1-k} \\ &\quad + 8\varepsilon^{-1}n(1-\lambda)^{-1}(\lambda d)^{-k/2} + (13k+72b)\varepsilon^{-1}n(1-\lambda)^{-1} \\ &\quad + 32(k-1)k\varepsilon^{-1}n(1-\lambda)^{-1}(\lambda d)^{-1} + n(8\ell + 32\log^2 n) \\ &\le k\varepsilon^{-1}n(1-\lambda)^{-1}\big(8 + \frac{8\varepsilon}{\lambda d} + \frac{8}{\lambda d} + 13 + \frac{32(k-1)}{\lambda d} \\ &\quad + 32\varepsilon\log^2 n(1-\lambda)\big) + 72b\varepsilon^{-1}n(1-\lambda)^{-1} + 8\ell n \\ &\le \varepsilon^{-1}n(1-\lambda)^{-1}(22k+72b) + 8\ell n \\ &= q(\ell, b), \end{aligned}$$

for sufficiently large n. Therefore, if $E = \bigcap_{j=1}^{2k+9} E_j$ holds, then in particular E_{2k+8} and E_{2k+9} hold, which implies that $X_t \in \mathcal{H}$ for $q(\ell, b) \le t \le s_0$. Thus E is contained in the event $\{X_t \in \mathcal{H}$ for all $t \in [q(\ell, b), s_0]\}$, and it suffices to show that $\mathbb{P}(\overline{E}) \le \frac{2k+8}{s_0} + \mathbb{P}(\overline{E_1})$. We write

$$\mathbb{P}(\overline{E}) = \mathbb{P}(\overline{E_1}) + \sum_{j=2}^{2k+9} \mathbb{P}\left(\overline{E_j} \cap \bigcap_{i=1}^{j-1} E_i\right),$$

and now we see that it suffices to prove that each of the terms $\mathbb{P}\left(\overline{E_j} \cap \bigcap_{i=1}^{j-1} E_i\right)$ is at most $1/s_0$.

We show how to derive the first few of these inequalities from Lemmas 6.2–6.7; first we have

$$\mathbb{P}(\overline{E_2} \cap E_1) = \mathbb{P}(T^{\dagger}_{\mathcal{A}} > s_0,\ T_{\mathcal{B}} > m_{\mathcal{B}}) \le \mathbb{P}(T_{\mathcal{B}} \wedge T^{\dagger}_{\mathcal{A}} \ge m_{\mathcal{B}}) \le 1/s_0$$

by Lemma 6.2(1). Then we have

$$\mathbb{P}(\overline{E_3} \cap E_1 \cap E_2) \le \mathbb{P}(\overline{E_3} \cap E_1) = \mathbb{P}(T^{\dagger}_{\mathcal{B}} \le s_0 < T^{\dagger}_{\mathcal{A}}) \le 1/s_0$$

by Lemma 6.2(2). Next we have, using the fact that $m_{\mathcal{B}} + m_{\mathcal{C}} \le s_0$,

$$\begin{aligned} \mathbb{P}(\overline{E_4} \cap E_1 \cap E_2 \cap E_3) &\le \mathbb{P}(\overline{E_4} \cap E_2 \cap E_3) \\ &= \mathbb{P}(T^{\dagger}_{\mathcal{B}} > s_0,\ T_{\mathcal{B}} \le m_{\mathcal{B}},\ T_{\mathcal{C}} > m_{\mathcal{B}} + m_{\mathcal{C}}) \\ &\le \mathbb{P}(T_{\mathcal{C}} \wedge T^{\dagger}_{\mathcal{B}} > m_{\mathcal{B}} + m_{\mathcal{C}},\ T_{\mathcal{B}} \le m_{\mathcal{B}}) \\ &\le \mathbb{P}(T_{\mathcal{C}} \wedge T^{\dagger}_{\mathcal{B}} > T_{\mathcal{B}} + m_{\mathcal{C}}) \\ &\le 1/s_0, \end{aligned}$$

by Lemma 6.3(1). For $j = 5, \ldots, 2k+9$, the upper bound on $\mathbb{P}\left(\overline{E_j} \cap \bigcap_{i=1}^{j-1} E_i\right)$ follows either as for $j = 3$ or as for $j = 4$: it is important here that $m_{\mathcal{B}} + m_{\mathcal{C}} + m_{\mathcal{D}} + m_{\mathcal{E}} + (k-1)m_{\mathcal{G}} + m_{\mathcal{H}} \le q(\ell, b) \le s_0$. □

We now have the following consequence for an equilibrium copy (Y_t) of the (n, d, λ)-supermarket process.

Corollary 6.9 *$\mathbb{P}(Y_t \in \mathcal{H}$ for all $t \in [0, s_0]) \ge 1 - (4k + 20)/s_0 \ge 1 - e^{-\frac{1}{4}\log^2 n}$, for n sufficiently large.*

Proof Recall the definitions of ℓ^* and b^* in Sect. 4. Set also $q^* = q(\ell^*, b^*)$, and note that $q^* \le s_0/2$, with plenty to spare. From Lemma 4.2, we have that $\mathbb{P}(Y_0 \notin \mathcal{A}_0) \le ne^{-\log^2 n} \le e^{-\frac{1}{3}\log^2 n} = 1/s_0$, since $n \ge 5$. Also, from Lemma 4.4, for a copy (X_t^x) of the process starting in a state $x \in \mathcal{A}_0$, we have that $\mathbb{P}(T_{\mathcal{A}}^\dagger < s_0) \le 1/s_0$. We now have

$$\begin{aligned}
\mathbb{P}(Y_t \notin \mathcal{H} \text{ for some } t \in [0, s_0/2]) &= \mathbb{P}(Y_t \notin \mathcal{H} \text{ for some } t \in [q^*, q^* + s_0/2]) \\
&\le \mathbb{P}(Y_t \notin \mathcal{H} \text{ for some } t \in [q^*, q^* + s_0/2] \mid Y_0 \in \mathcal{A}_0) \\
&\quad + \mathbb{P}(Y_0 \notin \mathcal{A}_0) \\
&\le \mathbb{P}(Y_t \notin \mathcal{H} \text{ for some } t \in [q^*, s_0] \mid Y_0 \in \mathcal{A}_0) \\
&\quad + \mathbb{P}(Y_0 \notin \mathcal{A}_0) \\
&\le \sup_{x \in \mathcal{A}_0^*} \mathbb{P}(X_t^x \notin \mathcal{H} \text{ for some } t \in [q^*, s_0]) + \frac{1}{s_0} \\
&\le \frac{2k+8}{s_0} + \frac{1}{s_0} + \frac{1}{s_0} = \frac{2k+10}{s_0},
\end{aligned}$$

by Proposition 6.8. Hence $\mathbb{P}(Y_t \notin \mathcal{H}$ for some $t \in [0, s_0]) \le (4k + 20)/s_0$. □

The first part of Theorem 6.1 now follows, since we have already noted that $\mathcal{H}^\varepsilon \subseteq \mathcal{N}^\varepsilon$.

We can also use Corollary 6.9 to prove the following more explicit version of Proposition 6.8.

Theorem 6.10 *Suppose that ℓ and b are at least k, and that $q(\ell, b) \le s_0/2$. Let x_0 be any queue-lengths vector in $\mathcal{A}_0(\ell, b)$, and suppose that $X_0 = x_0$ a.s. Then we have, for n sufficiently large,*

$$\mathbb{P}(X_t \in \mathcal{H} \textit{ for all } t \in [q(\ell, b), s_0]) \ge 1 - \frac{6k+28}{s_0} \ge 1 - e^{-\frac{1}{4}\log^2 n}.$$

Proof We apply, successively, Proposition 6.8, Lemma 4.3 and Corollary 6.9 to obtain that

$$\begin{aligned}
\mathbb{P}(X_t \in \mathcal{H} \text{ for all } t \in [q(\ell, b), s_0]) &\ge 1 - \frac{2k+8}{s_0} - \mathbb{P}(T_{\mathcal{A}}^\dagger \le s_0) \\
&= 1 - \frac{2k+8}{s_0} - \mathbb{P}(\exists t \in [0, s_0],\ X_t \notin \mathcal{A}_1(\ell, b)) \\
&\ge 1 - \frac{2k+8}{s_0} - \mathbb{P}(\exists t \in [0, s_0],\ Y_t \notin \mathcal{A}_0(\ell, b)) \\
&\ge 1 - \frac{2k+8}{s_0} - \mathbb{P}(\exists t \in [0, s_0],\ Y_t \notin \mathcal{H}) \\
&\ge 1 - \frac{2k+8}{s_0} - \frac{4k+20}{s_0},
\end{aligned}$$

as required. □

To see the final assertion of Theorem 6.1, suppose that $X_0 = x_0$ a.s., where x_0 is in the set

$$\mathcal{I} = \mathcal{A}_0 \cap \mathcal{B}_0 \cap \mathcal{C}_0 \cap \mathcal{D}_0 \cap \mathcal{E}_0 \cap \bigcap_{j=1}^{k-1} \mathcal{G}_0^j \cap \mathcal{H}_0.$$

Then all the hitting times $T_{\mathcal{B}}$, $T_{\mathcal{C}}$, $T_{\mathcal{D}}$, $T_{\mathcal{E}}$, $T_{\mathcal{G}}^j$ and $T_{\mathcal{H}}$ are equal to 0. In the notation of the proof of Proposition 6.8, this implies that the events E_j for j even occur with probability 1. Also, by Lemma 4.4, $\mathbb{P}(\overline{E_1}) \le 1/s_0$. So following the proof of Proposition 6.8 yields that, for $X_0 = x_0 \in \mathcal{I}$,

$$\mathbb{P}(X_t \in \mathcal{H} \text{ for all } t \in [0, s_0]) \ge 1 - (k+5)/s_0 \ge 1 - e^{-\frac{1}{4}\log^2 n}. \tag{6.1}$$

It can easily be seen that $\mathcal{N}^{\varepsilon/6} \subseteq \mathcal{I}^{\varepsilon}$, and hence this result completes the proof of Theorem 6.1.

7 Proofs of Lemmas 6.2, 6.3 and 6.4

In this section, we prove the first three of the sequence of lemmas stated in the previous section, and also derive tighter inequalities on the drifts of the functions $Q_j(x)$ for $x \in \mathcal{D}_1$. The proofs of the three lemmas are all straightforward applications of Lemma 3.3, and all similar to one another.

Proof of Lemma 6.2

Proof We apply Lemma 3.3. We set $(\varphi_t) = (\mathcal{F}_t)$, the natural filtration of the process, and also: $F = Q_k$, $\mathcal{S} = \mathcal{A}_1$, $p = 1$,

$$h = (1+\varepsilon)(1-\lambda)n(\lambda d)^{k-1}, \quad \rho = \varepsilon(1-\lambda)n(\lambda d)^{k-1},$$

$m = m_{\mathcal{B}} = 8k\varepsilon^{-1}n(1-\lambda)^{-1}$, $s = s_0 = e^{\frac{1}{3}\log^2 n}$ and $T^* = 0$. It is clear that $\rho \ge 2$ and that $Q_k(x) \le c := kn$ for any $x \in \mathbb{Z}_+^n$. We note also that Q_k takes jumps of size at most 1.

Suppose now that $Q_k(x) \ge h$. Then

$$\exp\left(-\frac{dQ_k(X_t)}{kn}\right) \le \exp\left(-\frac{(1-\lambda)(\lambda d)^k}{k}\right) \le \frac{\varepsilon(1-\lambda)}{4}.$$

The final inequality above is true comfortably, as $(1-\lambda)d^k = n^{-\alpha+k\beta} = n^{\delta}$ for some $\delta > 0$.

Hence, by Lemma 5.1, for x with $Q_k(x) \ge h$, we have

$$\begin{aligned}(1+\lambda)\Delta Q_k(x) &\le \beta_k\big((1-\lambda) - u_{k+1}(x) + \lambda\exp(-dQ_k(x)/kn)\big)\\ &\quad - \frac{1}{(\lambda d)^{k-1}}\frac{Q_k(x)}{n}\left(1 - \frac{2}{\lambda d}\right),\\ &\le \beta_k\left((1-\lambda) + \lambda\frac{\varepsilon(1-\lambda)}{4}\right) - (1+\varepsilon)(1-\lambda)(1-\varepsilon/5)\\ &\le (1-\lambda)\left[1 + \frac{\varepsilon}{4} - (1+3\varepsilon/4)\right] = -(1-\lambda)\frac{\varepsilon}{2}.\end{aligned}$$

So $\Delta Q_k(x) \le -(1-\lambda)\varepsilon/4 := -v$. Note that $m_{\mathcal{B}}v = 2c$.

We have now verified that the conditions of Lemma 3.3 are satisfied, for the given values of the parameters. As in the lemma, we have $T_0 = T_{\mathcal{A}}^{\dagger}$, $T_1 = \inf\{t : Q_k(X_t) \le h\}$ and $T_2 = \inf\{t > T_1 : Q_k(X_t) \ge h + \rho\}$.

It need not be the case that $T_1 = T_{\mathcal{B}}$, since X_{T_1} need not be in $\mathcal{A}_1$. However, we do have $T_1 \wedge T^\dagger_{\mathcal{A}} = T_{\mathcal{B}} \wedge T^\dagger_{\mathcal{A}}$ and thus

$$\begin{aligned}\mathbb{P}(T_{\mathcal{B}} \wedge T^\dagger_{\mathcal{A}} > m_{\mathcal{B}}) &= \mathbb{P}(T_1 \wedge T^\dagger_{\mathcal{A}} > m_{\mathcal{B}})\\ &\le \exp(-v^2 m_{\mathcal{B}}/8)\\ &= \exp(-\varepsilon kn(1-\lambda)/16) \le 1/s_0.\end{aligned}$$

Also the events $T_2 \le s_0 < T^\dagger_{\mathcal{A}}$ and $T^\dagger_{\mathcal{B}} \le s_0 < T^\dagger_{\mathcal{A}}$ coincide, so we have

$$\begin{aligned}\mathbb{P}(T^\dagger_{\mathcal{B}} \le s_0 < T^\dagger_{\mathcal{A}}) &\le \mathbb{P}(T_2 \le s_0 < T^\dagger_{\mathcal{A}})\\ &\le \frac{100s}{v^2}\exp(-\rho v/8)\\ &= \frac{100s_0}{v^2}\exp(-\varepsilon^2(1-\lambda)^2 n(\lambda d)^{k-1}/32)\\ &= \frac{100s_0}{v^2}\exp(-\varepsilon^2\lambda^{k-1}n^{1-2\alpha+(k-1)\beta}/32)\\ &\le 1/s_0,\end{aligned}$$

as required. Here we used that $1-2\alpha+(k-1)\beta > 0$. □

Proof of Lemma 6.3

Proof Again we apply Lemma 3.3 to the Markov process (X_t) with its natural filtration. Set $F = P_{k-1}$, $\mathcal{S} = \mathcal{B}_1$, $p = 1$,

$$h = 2kn(1-\lambda)(\lambda d)^{k-2},\ \rho = kn(1-\lambda)(\lambda d)^{k-2},$$

$m = m_{\mathcal{C}} = 8kn(1-\lambda)^{-1}(\lambda d)^{1-k}$, and $s = s_0$. Set $T^* = T_{\mathcal{B}}$. It is again clear that $\rho \ge 2$, that P_{k-1} takes jumps of size at most 1, and that $P_{k-1}(x) \le c := kn$ for all $x \in \mathbb{Z}^n_+$. Here $T_0 = T^\dagger_{\mathcal{B}}$, $T_1 = \inf\{t \ge T_{\mathcal{B}} : P_{k-1}(X_t) \le h\}$, and $T_2 = \inf\{t > T_1 : P_{k-1}(X_t) \ge h + \rho\}$.

For $x \in \mathcal{B}_1$ with $P_{k-1}(x) \ge h$, we have $Q_k(x) \le (1+2\varepsilon)n(1-\lambda)(\lambda d)^{k-1}$ and so, by Lemma 5.3,

$$\begin{aligned}(1+\lambda)\Delta P_{k-1}(x) &\le -\frac{\lambda d P_{k-1}(x)}{(k-1)n} + \frac{Q_k(x)}{n}\\ &\le -2\lambda d(1-\lambda)(\lambda d)^{k-2} + (1+2\varepsilon)(1-\lambda)(\lambda d)^{k-1}\\ &\le -\frac{1}{2}(1-\lambda)(\lambda d)^{k-1}.\end{aligned}$$

We conclude that, for such x, $\Delta P_{k-1}(x) \le -\frac{1}{4}(1-\lambda)(\lambda d)^{k-1} := -v$. Note that $m_{\mathcal{C}} v = 2c$.

As in the previous lemma, it need not be the case that $T_1 = T_{\mathcal{C}}$, since X_{T_1} need not be in $\mathcal{B}_1$, so we may have $T_{\mathcal{C}} > T_1$. However, we do have $T_1 \wedge T^\dagger_{\mathcal{B}} = T_{\mathcal{C}} \wedge T^\dagger_{\mathcal{B}}$. From Lemma 3.3, we obtain that

$$\begin{aligned}\mathbb{P}(T_{\mathcal{C}} \wedge T^\dagger_{\mathcal{B}} > T_{\mathcal{B}} + m_{\mathcal{C}}) &= \mathbb{P}(T_1 \wedge T_0 > T_{\mathcal{B}} + m_{\mathcal{C}})\\ &\le \exp(-v^2 m_{\mathcal{C}}/8)\\ &= \exp(-kn(1-\lambda)(\lambda d)^{k-1}/16) \le 1/s_0.\end{aligned}$$

Similarly, the events $T_2 \le s_0 < T^\dagger_{\mathcal{B}}$ and $T^\dagger_{\mathcal{C}} \le s_0 < T^\dagger_{\mathcal{B}}$ coincide, and so, for $k \ge 2$,

$$\begin{aligned}\mathbb{P}(T_{\mathcal{C}}^{\dagger} \le s_0 < T_{\mathcal{B}}^{\dagger}) &= \mathbb{P}(T_2 \le s_0 < T_0)\\ &\le \frac{100 s_0}{v^2} \exp(-\rho v/8)\\ &= \frac{100 s_0}{v^2} \exp(-kn(1-\lambda)^2(\lambda d)^{2k-3}/32)\\ &\le \frac{100 s_0}{v^2} \exp\left(-k\lambda^{2k-3} n^{1-2\alpha+(k-1)\beta+(k-2)\beta}/32\right)\\ &\le 1/s_0,\end{aligned}$$

as required. □

Sketch of Proof of Lemma 6.4

Proof The basic plan for this proof is the same as for the previous two lemmas, but here we have to take account of the fact that Q_{k-1} can take jumps of size up to $(\lambda d)^{(k-2)/2}$, and accordingly we apply Lemma 3.3 to the "scaled" function $F(x) = Q'_{k-1}(x) = Q_{k-1}(x)/(\lambda d)^{(k-2)/2}$.

Apart from this, the proof is identical in structure to that of Lemma 6.3, and we give only the key calculation. For $x \in \mathcal{C}_1$ with $Q'_{k-1}(x) \ge h = (1+4\varepsilon)n(1-\lambda)(\lambda d)^{(k-2)/2}$, we have $Q_k(x) \le (1+2\varepsilon)n(1-\lambda)(\lambda d)^{k-1}$, $P_{k-1}(x) \le 3kn(1-\lambda)(\lambda d)^{k-2}$ and $Q_{k-1}(x) \ge (1+4\varepsilon)n(1-\lambda)(\lambda d)^{k-2}$. Thus, by Lemma 5.2 with $j = k-1$, we have

$$\begin{aligned}(1+\lambda)\Delta Q_{k-1}(x) &\le -\lambda d \frac{Q_{k-1}(x)}{n}\left(1 - \frac{2}{\sqrt{\lambda d}} - \frac{dP_{k-1}(x)}{n}\right) + \frac{Q_k(x)}{n},\\ &\le -\lambda d(1+4\varepsilon)(1-\lambda)(\lambda d)^{k-2}\left(1 - \frac{2}{\sqrt{\lambda d}} - 3kd(1-\lambda)(\lambda d)^{k-2}\right)\\ &\quad +(1+2\varepsilon)(1-\lambda)(\lambda d)^{k-1}\\ &\le -\varepsilon(1-\lambda)(\lambda d)^{k-1}.\end{aligned}$$

Thus, for such x, the drift in the scaled chain satisfies $\Delta Q'_{k-1}(x) \le -\frac{1}{2}\varepsilon(1-\lambda)(\lambda d)^{k/2} := -v$. Now $Q'_{k-1}(x) \le c := 2n$ for all x by (5.3), and $m_{\mathcal{D}} v = 2c$.

It is now straightforward to derive the result. □

A queue-lengths vector $x \in \mathcal{D}_1$ satisfies the three inequalities:

$$Q_k(x) \le (1+2\varepsilon)n(1-\lambda)(\lambda d)^{k-1}, \tag{7.1}$$

$$P_{k-1}(x) \le 3kn(1-\lambda)(\lambda d)^{k-2},$$

$$Q_{k-1}(x) \le (1+5\varepsilon)n(1-\lambda)(\lambda d)^{k-2}; \tag{7.2}$$

in fact the second of these is redundant, as $P_{k-1}(x) \le Q_{k-1}(x) \le 2n(1-\lambda)(\lambda d)^{k-2}$ for all $x \in \mathbb{Z}_+^n$. Substituting these bounds into the bounds of Lemmas 5.1 and 5.2, we obtain the following.

Lemma 7.1 *For $x \in \mathcal{D}_1$, we have*

$$\begin{aligned}(1+\lambda)\Delta Q_k(x) &\le \beta_k(1-\lambda-u_{k+1}(x)) - \frac{Q_k(x)}{n(\lambda d)^{k-1}}\\ &\quad + \exp(-dQ_k(x)/kn) + \frac{\varepsilon}{6}(1-\lambda),\\ (1+\lambda)\Delta Q_k(x) &\ge \beta_k(1-\lambda-u_{k+1}(x)) - \frac{Q_k(x)}{n(\lambda d)^{k-1}} - \frac{\varepsilon}{6}(1-\lambda),\end{aligned}$$

and, for $1 \le j \le k-1$,

$$(1+\lambda)\Delta Q_j(x) \le -\lambda d \frac{Q_j(x)}{n}\Big(1 - \frac{\varepsilon}{25k}\Big) + \frac{Q_{j+1}(x)}{n},$$
$$(1+\lambda)\Delta Q_j(x) \ge -\lambda d \frac{Q_j(x)}{n}\Big(1 + \frac{\varepsilon}{50k}\Big) + \frac{Q_{j+1}(x)}{n}.$$

8 Proof of Lemma 6.5

This section is devoted to the rather more complex proof of Lemma 6.5. First, we prove a statement stronger than part (1) of the lemma. We set

$$\begin{aligned}\mathcal{K} &= \Big\{x : u_{k+1}(x) \le \varepsilon(1-\lambda) \text{ and } Q_k(x) \ge n\big(1 - \frac{\varepsilon}{3}\big)(1-\lambda)(\lambda d)^{k-1}\Big\} \cap \mathcal{D}_1;\\ W_{\mathcal{K}} &= \inf\{t \ge T_{\mathcal{D}} : X_t \in \mathcal{K}\}.\end{aligned}$$

Note that $\mathcal{K} \subseteq \mathcal{E}_0$, so to prove Lemma 6.5(1) it suffices to prove that

$$\mathbb{P}(W_{\mathcal{K}} \wedge T^{\dagger}_{\mathcal{D}} \ge T_{\mathcal{D}} + m_{\mathcal{E}}) \le 1/s_0.$$

We prove this result on the assumption that $T_{\mathcal{D}} = 0$ (i.e., that $x_0 \in \mathcal{A}_0 \cap \mathcal{B}_0 \cap \mathcal{C}_0 \cap \mathcal{D}_0$). The general case follows immediately by applying the result for $T_{\mathcal{D}} = 0$ to the shifted process $(X'_t) = (X_{T_{\mathcal{D}}+t})$, using the strong Markov property. So our task is to show that $\mathbb{P}(W_{\mathcal{K}} \wedge T^{\dagger}_{\mathcal{D}} \ge m_{\mathcal{E}}) \le 1/s_0$, where $W_{\mathcal{K}} = \inf\{t \ge 0 : X_t \in \mathcal{K}\}$, whenever $X_0 = x_0$ a.s., for any $x_0 \in \mathcal{A}_0 \cap \mathcal{B}_0 \cap \mathcal{C}_0 \cap \mathcal{D}_0$.

We define the following further sets, hitting times and exit times. We set

$$\begin{aligned}\mathcal{L}_1^{k+1} &= \mathcal{D}_1 \setminus \mathcal{K}\\ &= \Big\{x : u_{k+1}(x) > \varepsilon(1-\lambda) \text{ or } Q_k(x) < n\big(1 - \frac{\varepsilon}{3}\big)(1-\lambda)(\lambda d)^{k-1}\Big\} \cap \mathcal{D}_1,\end{aligned}$$

$W_{\mathcal{L}^{k+1}} = 0$ and $W^{\dagger}_{\mathcal{L}^{k+1}} = \inf\{t \ge 0 : X_t \notin \mathcal{L}_1^{k+1}\} = W_{\mathcal{K}} \wedge T^{\dagger}_{\mathcal{D}}$. Also, for $j = k, \ldots, 1$, let

$$\begin{aligned}\mathcal{L}_0^j &= \Big\{x : Q_j(x) \le n(1-\lambda)(\lambda d)^{j-1}(1 - \frac{\varepsilon}{6} - \frac{j\varepsilon}{6k})\Big\} \cap \mathcal{L}_1^{j+1};\\ \mathcal{L}_1^j &= \Big\{x : Q_j(x) \le n(1-\lambda)(\lambda d)^{j-1}\big(1 - \frac{\varepsilon}{6} - \frac{j\varepsilon}{6k} + \frac{\varepsilon}{24k}\big)\Big\} \cap \mathcal{L}_1^{j+1};\\ W_{\mathcal{L}^j} &= \inf\{t \ge W_{\mathcal{L}^{j+1}} : X_t \in \mathcal{L}_0^j\};\\ W^{\dagger}_{\mathcal{L}^j} &= \inf\{t \ge W_{\mathcal{L}^j} : X_t \notin \mathcal{L}_1^j)\}.\end{aligned}$$

Our goal is to show that $\mathbb{P}(W^{\dagger}_{\mathcal{L}^{k+1}} < m_{\mathcal{E}}) \ge 1 - 1/s_0$. If $x_0 \in \mathcal{K}$, then $W^{\dagger}_{\mathcal{L}^{k+1}} = 0$ and we are done, so we may assume that $x_0 \notin \mathcal{K}$, and hence that $x_0 \in \mathcal{L}_1^{k+1}$. Thus Lemma 6.5(1) follows from the proposition below.

Proposition 8.1 *Let* x_0 *be any queue-lengths vector in* $\mathcal{L}_1^{k+1}$. *For a copy* (X_t) *of the* (n, d, λ)-*supermarket process with* $X_0 = x_0$ *a.s., we have*

$$\mathbb{P}(W^{\dagger}_{\mathcal{L}^{k+1}} \ge m_{\mathcal{E}}) \le 1/s_0.$$

For the proof of Proposition 8.1, we fix a state $x_0 \in \mathcal{L}_1^{k+1}$, and work with a copy (X_t) of the (n, d, λ)-supermarket process where $X_0 = x_0$ a.s.

Our general plan for proving Proposition 8.1 is as follows. We suppose that the process (X_t) stays inside $\mathcal{L}_1^{k+1} = \mathcal{D}_1 \setminus \mathcal{K}$ over the interval $[0, m_{\mathcal{E}})$, with the aim of showing that this event has low probability. Observe that, if $x \in \mathcal{L}_1^{k+1} \setminus \mathcal{L}_0^k$, then $u_{k+1}(x) > \varepsilon(1-\lambda)$ and $Q_k(x) > n(1-\frac{\varepsilon}{3})(1-\lambda)(\lambda d)^{k-1}$. This "excess" in u_{k+1} would result in a downward drift in $Q_k(X_t)$, so if the process does not exit $\mathcal{L}_1^{k+1}$ quickly, then it enters $\mathcal{L}_0^k$ quickly, and stays in $\mathcal{L}_1^k$ throughout the interval $[0, m_{\mathcal{E}})$: i.e., $W_{\mathcal{L}^k}$ is small and $W^\dagger_{\mathcal{L}^k}$ is large, with high probability. This means that $Q_k(X_t)$ maintains a "deficit" compared to $\tilde{Q}_k := n(1-\lambda)(\lambda d)^{k-1}$ until time $m_{\mathcal{E}}$. A deficit in $Q_k(X_t)$ would lead to a deficit in each $Q_j(X_t)$ in turn, compared to $\tilde{Q}_j := n(1-\lambda)(\lambda d)^{j-1}$, for $j = k-1, k-2, \dots, 1$: each $W_{\mathcal{L}^j}$ is small, and $W^\dagger_{\mathcal{L}^j}$ is large, with high probability. Finally, a deficit in $Q_1(X_t)$ compared to $\tilde{Q}_1 = n(1-\lambda)$ is unsustainable, as this would lead to a drift down in the total number of customers over a long enough time interval to empty the entire system of customers. This would entail exiting the set $\mathcal{B}_1 \supseteq \mathcal{L}_1^{k+1}$, a contradiction.

Lemma 8.2 (1) $\mathbb{P}(W_{\mathcal{L}^k} \wedge W^\dagger_{\mathcal{L}^{k+1}} \geq 12k\varepsilon^{-1}n(1-\lambda)^{-1}) \leq 1/6s_0$.
(2) $\mathbb{P}(W^\dagger_{\mathcal{L}^k} < m_{\mathcal{E}} \leq W^\dagger_{\mathcal{L}^{k+1}}) \leq 1/12s_0$.

Proof We apply Lemma 3.3 to the process (X_t), with its natural filtration, and the function $F = Q_k$. We set $h = (1-\frac{\varepsilon}{3})n(1-\lambda)(\lambda d)^{k-1}$ and $\rho = \frac{\varepsilon}{24k}n(1-\lambda)(\lambda d)^{k-1} \geq 2$. We also set $\mathcal{S} = \mathcal{L}_1^{k+1}$ and $T^* = 0$. We note that $Q_k(x) \leq c := kn$ for every x, and we take $m = 12k\varepsilon^{-1}n(1-\lambda)^{-1}$, and $s = m_{\mathcal{E}} - 1$. Then $T_0 = W^\dagger_{\mathcal{L}^{k+1}}$, $T_1 = \inf\{t : Q_k(X_t) \leq h\}$ and $T_2 = \inf\{t > T_1 : Q_k(X_t) \geq h + \rho\}$, as in the lemma.

For $x \in \mathcal{L}_1^{k+1}$ with $Q_k(x) > h$, we have $u_{k+1}(x) > \varepsilon(1-\lambda)$ and $x \in \mathcal{D}_1$. So Lemma 7.1 applies, and we have

$$\begin{aligned}(1+\lambda)\Delta Q_k(x) &\leq \beta_k(1-\lambda-u_{k+1}(X_t)) - \frac{Q_k(x)}{n(\lambda d)^{k-1}} + \exp(-dQ_k(x)/kn) + \frac{\varepsilon}{6}(1-\lambda)\\ &\leq (1-\lambda)(1-\varepsilon) - (1-\lambda)\big(1-\frac{\varepsilon}{3}\big) + \frac{\varepsilon}{6}(1-\lambda) + \frac{\varepsilon}{6}(1-\lambda)\\ &= -\frac{1}{3}\varepsilon(1-\lambda).\end{aligned}$$

So $\Delta Q_k(x) \leq -\frac{1}{6}\varepsilon(1-\lambda) := -v$ for such x. Note that $mv = 2c$. Hence we may apply Lemma 3.3.

As in earlier lemmas, we have $T_1 \wedge W^\dagger_{\mathcal{L}^{k+1}} = W_{\mathcal{L}^k} \wedge W^\dagger_{\mathcal{L}^{k+1}}$, so we obtain

$$\begin{aligned}\mathbb{P}(W_{\mathcal{L}^k} \wedge W^\dagger_{\mathcal{L}^{k+1}} > m) &= \mathbb{P}(T_1 \wedge T_0 > m)\\ &\leq \exp(-v^2 m/8)\\ &= \exp(-\varepsilon kn(1-\lambda)/24) < 1/6s_0.\end{aligned}$$

Also the events $W^\dagger_{\mathcal{L}^k} < m_{\mathcal{E}} \leq W^\dagger_{\mathcal{L}^{k+1}}$ and $T_2 < m_{\mathcal{E}} \leq W^\dagger_{\mathcal{L}^{k+1}}$ coincide, and the second is equivalent to $T_2 \leq s < W^\dagger_{\mathcal{L}^{k+1}}$ (since $s = m_{\mathcal{E}} - 1$). So

$$\begin{aligned}\mathbb{P}(W^\dagger_{\mathcal{L}^k} < m_{\mathcal{E}} \leq W^\dagger_{\mathcal{L}^{k+1}}) &= \mathbb{P}(T_2 \leq s < T_0)\\ &\leq \frac{100s}{v^2}\exp(-\rho v/8)\end{aligned}$$

$$\begin{aligned} &= \frac{100s}{v^2}\exp(-\varepsilon^2 n(1-\lambda)^2(\lambda d)^{k-1}/1152k) \\ &< 1/12s_0, \end{aligned}$$

as required. □

The next lemma states that, if the process stays in some set $\mathcal{L}_1^{j+1}$ for a long time, then it quickly enters the "next" set $\mathcal{L}_0^j$, and stays in $\mathcal{L}_1^j$ for a long time.

Lemma 8.3 *For each* $j = k-1, \ldots, 1$,

(1) $\mathbb{P}(W^\dagger_{\mathcal{L}^{j+1}} \wedge W_{\mathcal{L}^j} > W_{\mathcal{L}^{j+1}} + \varepsilon^{-1}n(1-\lambda)^{-1}) \le 1/3ks_0$.
(2) $\mathbb{P}(W^\dagger_{\mathcal{L}^j} < m_{\mathcal{E}} \le W^\dagger_{\mathcal{L}^{j+1}}) \le 1/3ks_0$.

Proof (Sketch) This proof is very similar to that of earlier lemmas, and we mention only a few points. As in Lemma 6.4, we apply Lemma 3.3 to the scaled process $Q'_j(x) = Q_j(x)/(\lambda d)^{(j-1)/2}$. The key step is to show that, for $x \in \mathcal{L}_1^{j+1}$ with $Q'_j \ge h = n(1-\lambda)(\lambda d)^{(j-1)/2}(1 - \frac{\varepsilon}{6} - \frac{j\varepsilon}{6k})$, we have $\Delta Q'_j(x) \le -\frac{\varepsilon}{24k}(1-\lambda)(\lambda d)^{(j+1)/2} := -v$. The proof now proceeds as earlier ones.

For part (2) of the lemma, we set $\rho = \frac{\varepsilon}{24k}n(1-\lambda)(\lambda d)^{(j-1)/2}$. We make use of the fact that the value of Q'_j only changes if either (i) the event is an arrival, and some queue of length at most $j-1$ is inspected, or (ii) the event is a departure from some queue of length at most j. From any state $x \in \mathcal{L}_1^j$, the probability of (i) is at most $d(1-u_j(x)) \le dQ_j(x)/n \le (1-\lambda)d^j$, and the probability of (ii) is at most $(1-u_{j+1}(x)) \le Q_{j+1}(x)/n \le (1-\lambda)d^j$. Hence we may apply Lemma 3.3(ii), with $p = 2(1-\lambda)d^j$.

$$\begin{aligned} \mathbb{P}(W^\dagger_{\mathcal{L}^j} < m_{\mathcal{E}} \le W^\dagger_{\mathcal{L}^{j+1}}) &\le \frac{100m_{\mathcal{E}}}{v^2}\exp(-\rho v/8p) \\ &= \frac{100m_{\mathcal{E}}}{v^2}\exp\left(-\frac{\varepsilon^2\lambda^j}{4608k^2}n(1-\lambda)\right) \\ &\le 1/3ks_0. \end{aligned}$$

We now prove a hitting time lemma for $\|X_t\|_1$, the total number of customers in the system at time t. Let $W_{\mathcal{M}} = \min\{t \ge W_{\mathcal{L}^1} : \|X_t\|_1 = 0\}$.

Lemma 8.4

$$\mathbb{P}(W^\dagger_{\mathcal{L}^1} \wedge W_{\mathcal{M}} > W_{\mathcal{L}^1} + 72b\varepsilon^{-1}n(1-\lambda)^{-1}) \le 1/12s_0.$$

Proof We apply Lemma 3.3(i) to the chain (X_t), with the filtration $(\mathcal{F}_t)$, and the function $F(x) = \|x\|_1$, which takes jumps of size at most 1. Since $\mathcal{A}_1(\ell, b) \supseteq \mathcal{L}_1^1$, we have $\|X_0\|_1 \le c := 3bn$. We also set $\mathcal{S} = \mathcal{L}_1^1$, $T^* = W_{\mathcal{L}^1}$, $h = 0$ and $m = 72b\varepsilon^{-1}n(1-\lambda)^{-1}$.

Note that $\|X_{t+1}\|_1 - \|X_t\|_1$ is equal to $+1$ if the event at time t is an arrival, with probability $\lambda/(1+\lambda)$, and equal to -1 if the event is a potential departure from a non-empty queue, with probability $u_1(X_t)/(1+\lambda)$, so the drift $\Delta\|x\|_1$ is equal to $\frac{1}{1+\lambda}(\lambda - u_1(x))$. For $x \in \mathcal{L}_1^1$, we have

$$1 - u_1(x) = \frac{Q_1(x)}{n} \le (1-\lambda)\left(1 - \frac{\varepsilon}{6} - \frac{\varepsilon}{6k} + \frac{\varepsilon}{24k}\right) \le (1-\lambda)\left(1 - \frac{\varepsilon}{6}\right).$$

Hence, for $x \in \mathcal{L}_1^1$, $(1+\lambda)\Delta\|x\|_1 = (1-u_1(x)) - (1-\lambda) \le -\frac{\varepsilon}{6}(1-\lambda)$, and so $\Delta\|x\|_1 \le -\frac{\varepsilon}{12}(1-\lambda) := -v$. Note that $vm = 2c$.

Hence we may apply Lemma 3.3(i). With T_0 and T_1 as in that lemma, we have $T_0 = W^\dagger_{\mathcal{L}^1}$ and $T_1 = W_{\mathcal{M}}$, so we conclude that

$$\begin{aligned}\mathbb{P}(W^\dagger_{\mathcal{L}^1} \wedge W_{\mathcal{M}} \geq W_{\mathcal{L}^1} + m) &\leq \exp(-v^2 m/8)\\ &= \exp(-\varepsilon b n(1-\lambda)/16) \leq 1/12 s_0,\end{aligned}$$

as required. □

We now combine Lemmas 8.2, 8.3 and 8.4 to prove Proposition 8.1.

Observe that, for a copy (X_t) of the (n, d, λ)-supermarket process starting in a state $x_0 \in \mathcal{L}^{k+1}_1$, exactly one of the following occurs:

(a) $W^\dagger_{\mathcal{L}^{k+1}} < m_{\mathcal{E}}$,
(b) not (a), and one of $W^\dagger_{\mathcal{L}^k}, W^\dagger_{\mathcal{L}^{k-1}}, \dots, W^\dagger_{\mathcal{L}^1}$ is less than $m_{\mathcal{E}}$,
(c) neither of the above, and $W_{\mathcal{L}^k} > 12k\varepsilon^{-1}n(1-\lambda)^{-1}$,
(d) none of the above, and $W_{\mathcal{L}^j} > W_{\mathcal{L}^{j+1}} + \varepsilon^{-1}n(1-\lambda)^{-1}$ for some $j = k-1, \dots, 1$,
(e) none of the above, and $W_{\mathcal{M}} > W_{\mathcal{L}^1} + 72b\varepsilon^{-1}n(1-\lambda)^{-1}$,
(f) none of the above, and $W_{\mathcal{M}} < m_{\mathcal{E}} \leq W^\dagger_{\mathcal{L}^{k+1}}$.

Indeed, if none of (a)–(e) occurs, then $W^\dagger_{\mathcal{L}^{k+1}} \geq m_{\mathcal{E}}$ since (a) fails, and also

$$\begin{aligned}W_{\mathcal{M}} &= W_{\mathcal{L}^k} + \sum_{j=1}^{k-1}(W_{\mathcal{L}^j} - W_{\mathcal{L}^{j+1}}) + (W_{\mathcal{M}} - W_{\mathcal{L}^1})\\ &\leq 12k\varepsilon^{-1}n(1-\lambda)^{-1} + (k-1)\varepsilon^{-1}n(1-\lambda)^{-1} + 72b\varepsilon^{-1}n(1-\lambda)^{-1}\\ &< (13k + 72b)\varepsilon^{-1}n(1-\lambda)^{-1} = m_{\mathcal{E}}.\end{aligned}$$

We now show that the probability of each of (b)–(f) is small. For (b), Lemmas 8.2(2) and 8.3(2) give that

$$\begin{aligned}&\mathbb{P}(W^\dagger_{\mathcal{L}^k} \wedge W^\dagger_{\mathcal{L}^{k-1}} \wedge \cdots \wedge W^\dagger_{\mathcal{L}^1} < m_{\mathcal{E}} \leq W^\dagger_{\mathcal{L}^{k+1}})\\ &\quad \leq \mathbb{P}(W^\dagger_{\mathcal{L}^k} < m_{\mathcal{E}} \leq W^\dagger_{\mathcal{L}^{k+1}}) + \sum_{j=1}^{k-1}\mathbb{P}(W^\dagger_{\mathcal{L}^j} < m_{\mathcal{E}} \leq W^\dagger_{\mathcal{L}^{j+1}})\\ &\quad \leq \frac{1}{6s_0} + (k-1)\frac{1}{3ks_0} \leq \frac{1}{2s_0},\end{aligned}$$

i.e., the probability of (b) is at most $1/2s_0$. The probability of (c) is at most $1/12s_0$ by Lemma 8.2(1). The probability of (d) is at most $(k-1)\frac{1}{3ks} \leq 1/3s_0$ by Lemma 8.3(1). The probability of (e) is at most $1/12s_0$ by Lemma 8.4. Finally, (f) is not possible, since at time $W_{\mathcal{M}}$ there are no customers in the system, so $Q_k(X_{W_{\mathcal{M}}}) > n$, and thus $W_{\mathcal{M}} \geq T^\dagger_{\mathcal{B}}$, but also $T^\dagger_{\mathcal{B}} \geq W^\dagger_{\mathcal{L}^{k+1}}$ since $\mathcal{L}^{k+1}_1 \subseteq \mathcal{D}_1 \subseteq \mathcal{B}_1$ by definition.

Thus the probability of (a), for a copy of the process starting in a state in $\mathcal{L}^{k+1}_1$, is at least $1 - \frac{1}{2s_0} - \frac{1}{12s_0} - \frac{1}{3s_0} - \frac{1}{12s_0} = 1 - \frac{1}{s_0}$, which is what we need to prove Proposition 8.1, and thus also Lemma 6.5(1).

Now we move to the proof of Lemma 6.5(2), stating that the exit time $T^\dagger_{\mathcal{E}}$ is large with high probability. There are two things to prove here. The first is that, if $X_t \in \mathcal{E}_1$, then it is very unlikely that, at time $t+1$, a customer arrives and creates a queue of length $k+1$. The second is that, once $Q_k(X_t)$ has reached $(1-3\varepsilon)n(1-\lambda)(\lambda d)^{k-1}$, while $u_{k+1}(X_t)$ is at most $\varepsilon(1-\lambda)$, Q_k is unlikely to "cross down against the drift" to $(1-4\varepsilon)n(1-\lambda)(\lambda d)^{k-1}$.

For $t \geq 0$, let L_t denote the event that, at time t, a customer arrives and joins a queue of length at least k (equivalently, the probability that the event is an arrival and that all the selected queues have length at least k). So L_t is the event that $u_j(X_t) > u_j(X_{t-1})$ for some $j \geq k+1$.

Lemma 8.5 *On the event that $X_t \in \mathcal{E}_1$, we have $\mathbb{P}(L_{t+1} \mid \mathcal{F}_t) < e^{-\log^2 n}$.*

Proof From the definition of L_t, we have $\mathbb{P}(L_{t+1} \mid \mathcal{F}_t) = \frac{\lambda}{1+\lambda} u_k(X_t)^d \leq u_k(X_t)^d$. For $x \in \mathcal{E}_1$, we have $Q_k(x) \geq (1-4\varepsilon)n(1-\lambda)(\lambda d)^{k-1}$ and $Q_{k-1}(x) \leq (1+5\varepsilon)n(1-\lambda)(\lambda d)^{k-2} \leq \frac{1}{3}\varepsilon n(1-\lambda)(\lambda d)^{k-1}$. Therefore, by (5.5), we have

$$1 - u_k(x) \geq \frac{Q_k(x)}{n} - \frac{Q_{k-1}(x)}{n} \geq \left(1 - \frac{13}{3}\varepsilon\right)(1-\lambda)(\lambda d)^{k-1} \geq \frac{1}{2}(1-\lambda)d^{k-1}.$$

Hence, on the event that $X_t \in \mathcal{E}_1$,

$$u_k(X_t)^d \leq \left(1 - \frac{1}{2}(1-\lambda)d^{k-1}\right)^d \leq \exp\left(-\frac{1}{2}(1-\lambda)d^k\right) \leq \exp(-\log^2 n),$$

as required. □

Let $U^\dagger = \inf\{t > T_{\mathcal{E}} : u_{k+1}(X_t) > \varepsilon(1-\lambda)\}$ and $V^\dagger = \inf\{t > T_{\mathcal{E}} : Q_k(X_t) < (1-4\varepsilon)n(1-\lambda)(\lambda d)^{k-1}\}$, and note that $T_{\mathcal{E}}^\dagger = T_{\mathcal{D}}^\dagger \wedge U^\dagger \wedge V^\dagger$. We thus have

$$\mathbb{P}(T_{\mathcal{E}}^\dagger \leq s_0 < T_{\mathcal{D}}^\dagger) \leq \mathbb{P}(U^\dagger \leq s_0 \wedge T_{\mathcal{D}}^\dagger \wedge V^\dagger) + \mathbb{P}(V^\dagger \leq s_0 \wedge T_{\mathcal{D}}^\dagger \wedge U^\dagger).$$

We claim that each of these last two probabilities is at most $1/2s_0$. For the first, we may apply Lemma 8.5. Observe that, if $U^\dagger = t+1$, then the event L_{t+1} occurs. We now have:

$$\begin{aligned}\mathbb{P}(U^\dagger \leq s_0 \wedge T_{\mathcal{D}}^\dagger \wedge V^\dagger) &= \sum_{t=0}^{s_0-1} \mathbb{P}(U^\dagger = t+1 \leq T_{\mathcal{D}}^\dagger \wedge V^\dagger) \\ = \sum_{t=0}^{s_0-1} \mathbb{P}(U^\dagger = t+1 \text{ and } X_t \in \mathcal{E}_1) &= \sum_{t=0}^{s_0-1} \mathbb{E}[\mathbb{1}_{\{X_t \in \mathcal{E}_1\}} \mathbb{E}(\mathbb{1}_{\{U^\dagger = t+1\}} \mid \mathcal{F}_t)] \\ &\leq \sum_{t=0}^{s_0-1} \mathbb{E}[\mathbb{1}_{\{X_t \in \mathcal{E}_1\}} \mathbb{E}(\mathbb{1}_{L_{t+1}} \mid \mathcal{F}_t)].\end{aligned}$$

By Lemma 8.5, each term is at most $e^{-\log^2 n}$, and so we have

$$\mathbb{P}(U^\dagger \leq s_0 \wedge T_{\mathcal{D}}^\dagger \wedge V^\dagger) \leq s_0 e^{-\log^2 n} < 1/2s_0,$$

as claimed.

To obtain the other required inequality, we apply the reversed version of Lemma 3.3(ii). We consider the process (X_t), with its natural filtration, the function $F = Q_k$, and the set $\mathcal{S} = \{x : u_{k+1}(x) \leq \varepsilon(1-\lambda)\} \cap \mathcal{D}_1$. We set $h = (1-3\varepsilon)n(1-\lambda)(\lambda d)^{k-1}$ and $\rho = \varepsilon n(1-\lambda)(\lambda d)^{k-1} \geq 2$. We also set $s = s_0$ and $T^* = T_{\mathcal{E}}$. We have $T_0 = \inf\{t \geq T_{\mathcal{E}} : X_t \notin \mathcal{D}_1 \text{ or } u_{k+1}(X_t) > \varepsilon(1-\lambda)\}$, so that $T_0 \geq T_{\mathcal{D}}^\dagger \wedge U^\dagger$ (strict inequality occurs if $T_{\mathcal{D}}^\dagger < T_{\mathcal{E}}$). Also $T_1 = \inf\{t \geq T_{\mathcal{E}} : Q_k(X_t) \geq h\} = T_{\mathcal{E}}$, and $T_2 = \inf\{t > T_{\mathcal{E}} : Q_k(X_t) \leq h - \rho\} = V^\dagger$.

Take $x \in \mathcal{S}$ with $Q_k(x) \leq h$. As $x \in \mathcal{D}_1$, we apply Lemma 7.1 to obtain

$$\begin{aligned}(1+\lambda)\Delta Q_k(x) &\geq \beta_k(1-\lambda-u_{k+1}(x)) - \frac{Q_k(x)}{n(\lambda d)^{k-1}} - \frac{\varepsilon}{6}(1-\lambda)\\ &\geq \beta_k(1-\lambda)(1-\varepsilon) - (1-\lambda)(1-3\varepsilon) - \frac{\varepsilon}{6}(1-\lambda)\\ &\geq (1-\lambda)\left[\left(1-\frac{\varepsilon}{2}\right)(1-\varepsilon) - 1 + 3\varepsilon - \frac{\varepsilon}{6}\right]\\ &\geq \varepsilon(1-\lambda).\end{aligned}$$

This yields $\Delta Q_k(x) \geq \frac{1}{2}\varepsilon(1-\lambda) := v$, for such x.

The reversed version of Lemma 3.3(ii) gives that

$$\begin{aligned}\mathbb{P}(V^\dagger \leq s_0 \wedge T^\dagger_{\mathcal{D}} \wedge U^\dagger) &\leq \mathbb{P}(T_2 \leq s_0 \wedge T_0)\\ &\leq \frac{100 s_0}{v^2}\exp(-\rho v/8)\\ &= \frac{100 s_0}{v^2}\exp(-\varepsilon^2 n(1-\lambda)^2(\lambda d)^{k-1}/16)\\ &\leq 1/2s_0,\end{aligned}$$

as required. This completes the proof of Lemma 6.5.

9 Proofs of Lemmas 6.6 and 6.7

In this section, we prove the final two of our sequence of lemmas.

Proof of Lemma 6.6

Proof Fix j with $1 \leq j \leq k-1$, and consider the state of the process at the hitting time $T_{\mathcal{G}^{j+1}}$. The hitting time $T_{\mathcal{G}^j}$ is the first time $t \geq T_{\mathcal{G}^{j+1}}$ that $Q_j(X_t)$ lies in the interval between $\left[1-(4+\frac{k-j-1/2}{k})\varepsilon\right]n(1-\lambda)(\lambda d)^{j-1}$ and $\left[1+(4+\frac{k-j-1/2}{k})\varepsilon\right]n(1-\lambda)(\lambda d)^{j-1}$. Let B_h be the event that $Q_j(X_{T_{\mathcal{G}^{j+1}}}) > \left[1+(4+\frac{k-j-1/2}{k})\varepsilon\right]n(1-\lambda)(\lambda d)^{j-1}$, and B_ℓ be the event that $Q_j(X_{T_{\mathcal{G}^{j+1}}}) < \left[1-(4+\frac{k-j-1/2}{k})\varepsilon\right]n(1-\lambda)(\lambda d)^{j-1}$.

For part (1) of the lemma, we have to show that, on the event B_h, with high probability $Q_j(X_t)$ enters the interval from above within time $m_{\mathcal{G}}$, and also that, on the event B_ℓ, with high probability $Q_j(X_t)$ enters the interval from below within time $m_{\mathcal{G}}$. These two results are essentially the same, and we give details only for the first. Of course, we have nothing to prove on the event that $Q_j(X_{T_{\mathcal{G}^{j+1}}})$ is already in the interval.

We apply Lemma 3.3(i) to (X_t), with its natural filtration, and the scaled function $F(x) = Q'_j(x) = Q_j(x)/(\lambda d)^{(j-1)/2}$. We take $\mathcal{S} = \mathcal{G}_1^{j+1}$ and $T^* = T_{\mathcal{G}^{j+1}}$. We set

$$h = \left[1+(4+\frac{k-j-1/2}{k})\varepsilon\right]n(1-\lambda)(\lambda d)^{(j-1)/2},$$

and $m = m_{\mathcal{G}} = 32k\varepsilon^{-1}n(1-\lambda)^{-1}(\lambda d)^{-1}$. From (5.3), we have that $Q'_j(x) \leq c := 2n$ for all x. Also $T_0 = T^\dagger_{\mathcal{G}^{j+1}}$ and $T_1 = \inf\{t \geq T_{\mathcal{G}^{j+1}} : Q'_j(X_t) \leq h\}$.

For $x \in \mathcal{G}_1^{j+1}$, we have

$$Q_{j+1}(x) \leq \left[1+(4+\frac{k-j-1}{k})\varepsilon\right]n(1-\lambda)(\lambda d)^j.$$

(This follows from the specification of $\mathcal{G}_1^{j+1}$ for $j < k-1$, and since $\mathcal{G}_1^k = \mathcal{E}_1 \subseteq \mathcal{B}_1$ for $j = k-1$.) If also $Q_j'(x) \ge h$, we have

$$Q_j(x) \ge \left[1 + \left(4 + \frac{k-j-1/2}{k}\right)\varepsilon\right] n(1-\lambda)(\lambda d)^{j-1}.$$

Lemma 7.1 applies since $x \in \mathcal{D}_1$, so

$$\begin{aligned}(1+\lambda)\Delta Q_j(x) &\le -\lambda d \frac{Q_j(x)}{n}\left(1 - \frac{\varepsilon}{25k}\right) + \frac{Q_{j+1}(x)}{n} \\ &\le -\left[1 + \left(4 + \frac{k-j-1/2}{k}\right)\varepsilon\right](1-\lambda)(\lambda d)^j \left(1 - \frac{\varepsilon}{25k}\right) \\ &\quad + \left[1 + \left(4 + \frac{k-j-1}{k}\right)\varepsilon\right](1-\lambda)(\lambda d)^j \\ &\le -\frac{1}{4k}\varepsilon(1-\lambda)(\lambda d)^j,\end{aligned}$$

and so $\Delta Q_j'(x) \le -\frac{1}{8k}\varepsilon(1-\lambda)(\lambda d)^{(j+1)/2} := -v$. Note that $v m_{\mathcal{G}} \ge 2c$.

Lemma 3.3(i) now gives

$$\begin{aligned}\mathbb{P}(T_1 \wedge T_0 > T_{\mathcal{G}^{j+1}} + m_{\mathcal{G}}) &\le \exp(-v^2 m_{\mathcal{G}}/8) \\ &= \exp(-\varepsilon n(1-\lambda)(\lambda d)^{(j+1)/2}/16k) \\ &\le 1/2s_0.\end{aligned}$$

On the $T_{\mathcal{G}^j}$-measurable event B_h, the stopping times $T_1 \wedge T_0$ and $T_{\mathcal{G}^j} \wedge T_0$ coincide, so we have

$$\mathbb{P}(B_h \cap \{T_{\mathcal{G}^j} \wedge T^\dagger_{\mathcal{G}^{j+1}} > T_{\mathcal{G}^{j+1}} + m_{\mathcal{G}}\}) \le 1/2s_0.$$

Essentially exactly the same calculation gives

$$\mathbb{P}(B_\ell \cap \{T_{\mathcal{G}^j} \wedge T^\dagger_{\mathcal{G}^{j+1}} > T_{\mathcal{G}^{j+1}} + m_{\mathcal{G}}\}) \le 1/2s_0,$$

and part (1) of the lemma now follows, for this value of j.

To prove part (2) of the lemma, we need to show that, once X_t has reached $\mathcal{G}_0^j$, and while it remains in $\mathcal{G}_1^{j+1}$, the process is unlikely to leave the set $\mathcal{G}_1^j$ quickly. There are two separate things to prove: that $Q_j(X_t)$ is unlikely to cross against the drift from $\left[1 + (4 + \frac{k-j-1/2}{k})\varepsilon\right] n(1-\lambda)(\lambda d)^{j-1}$ to $\left[1 + (4 + \frac{k-j}{k})\varepsilon\right] n(1-\lambda)(\lambda d)^{j-1}$ before time s_0, and also that $Q_j(X_t)$ is unlikely to cross against the drift from $\left[1 - (4 + \frac{k-j-1/2}{k})\varepsilon\right] n(1-\lambda)(\lambda d)^{j-1}$ to $\left[1 - (4 + \frac{k-j}{k})\varepsilon\right] n(1-\lambda)(\lambda d)^{j-1}$ before time s_0. Again, the two calculations required here are essentially identical, and we shall concentrate on the first.

We apply Lemma 3.3(ii), again for the process (X_t) with its natural filtration, and the scaled function $F(x) = Q_j(x)/(\lambda d)^{(j-1)/2}$. We take the same values of parameters as above, and additionally set $\rho = \frac{\varepsilon}{2k} n(1-\lambda)(\lambda d)^{(j-1)/2}$ and $s = s_0$. As before, we may take $p = 2(1-\lambda)d^j$. Here $T_2 = \inf\{t > T_1 : Q_j'(X_t) \ge h + \rho\}$.

$$\begin{aligned}\mathbb{P}(T_2 \le s_0 < T^{\dagger}_{\mathcal{G}^{j+1}}) &\le \frac{100 s_0}{v^2} \exp(-\rho v/8p)\\ &= \frac{100 s_0}{v^2} \exp(-\varepsilon^2 n(1-\lambda)\lambda^j/256k)\\ &\le 1/2s_0.\end{aligned}$$

Setting $U_2 = \inf\{t > T_1 : Q_j(X_t) \le \Big[1-(4+\frac{k-j}{k})\varepsilon\Big] n(1-\lambda)(\lambda d)^{j-1}\}$, we have, similarly, $\mathbb{P}(U_2 \le s_0 < T^{\dagger}_{\mathcal{G}^{j+1}}) \le 1/2s_0$.

The events $T_2 \wedge U_2 \le s_0 < T^{\dagger}_{\mathcal{G}^{j+1}}$ and $T^{\dagger}_{\mathcal{G}^j} \le s_0 < T^{\dagger}_{\mathcal{G}^{j+1}}$ coincide, so

$$\begin{aligned}\mathbb{P}(T^{\dagger}_{\mathcal{G}^j} \le s_0 < T^{\dagger}_{\mathcal{G}^{j+1}}) &\le \mathbb{P}(T_2 \le s < T_0) + \mathbb{P}(U_2 \le s < T_0)\\ &\le \frac{1}{2s_0} + \frac{1}{2s_0} = \frac{1}{s_0},\end{aligned}$$

as required for part (2) for this value of j. □

Proof of Lemma 6.7

Proof We first prove part (1). For $i = 1, \dots, n$, let N_i be the number of potential departures from queue i over the time period between $T_{\mathcal{G}^1}$ and $T_{\mathcal{G}^1} + m_{\mathcal{H}}$, so N_i is a binomial random variable with parameters $(m_{\mathcal{H}}, 1/n(1+\lambda))$. Recall that L_t is the event that, at time t, a customer arrives and joins a queue of length k or longer, and observe that

$$\mathbb{P}(T_{\mathcal{H}} \wedge T^{\dagger}_{\mathcal{G}^1} \ge T_{\mathcal{G}^1} + m_{\mathcal{H}}) \le \mathbb{P}\left(\bigcup_{t=T_{\mathcal{G}^1}+1}^{T_{\mathcal{G}^1}+m_{\mathcal{H}}} (L_t \cap \{X_{t-1} \in \mathcal{G}^1_1\})\right) + \mathbb{P}\left(\exists i, N_i < 3\ell\right).$$

Indeed, at time $T_{\mathcal{G}^1}$, the process is in $\mathcal{A}_1(\ell, g)$, and so there is no queue with more than 3ℓ customers in it at that time. If there are at least 3ℓ potential departures from each queue over the time interval, and $\bigcup_{t=T_{\mathcal{G}^1}+1}^{T_{\mathcal{G}^1}+m_{\mathcal{H}}} L_t$ does not occur, then by time $T_{\mathcal{G}^1} + m_{\mathcal{H}}$, every queue is reduced to length at most k, and no new queue of length $k+1$ is created before $T_{\mathcal{G}^1} + m_{\mathcal{H}}$.

Now let $(X'_t) = (X_{T_{\mathcal{G}^1}+t})$, $(\mathcal{F}'_t) = (\mathcal{F}_{T_{\mathcal{G}^1}+t})$ and $L'_t = L_{T_{\mathcal{G}^1}+t}$. We have:

$$\begin{aligned}\mathbb{P}\left(\bigcup_{t=T_{\mathcal{G}^1}+1}^{T_{\mathcal{G}^1}+m_{\mathcal{H}}} (L_t \cap \{X_{t-1} \in \mathcal{G}^1_1\})\right) &= \mathbb{P}\left(\bigcup_{t=1}^{m_{\mathcal{H}}} (L'_t \cap \{X'_{t-1} \in \mathcal{G}^1_1\})\right)\\ &\le \sum_{t=1}^{m_{\mathcal{H}}} \mathbb{P}(L'_t \cap \{X'_{t-1} \in \mathcal{G}^1_1\})\\ &= \sum_{t=1}^{m_{\mathcal{H}}} \mathbb{E}\left[\mathbb{1}_{\{X'_{t-1} \in \mathcal{G}^1_1\}} \mathbb{E}[\mathbb{1}_{L'_t} \mid \mathcal{F}'_{t-1}]\right]\\ &\le m_{\mathcal{H}} e^{-\log^2 n} \le 1/2s_0,\end{aligned}$$

where we used the strong Markov property, and Lemma 8.5.

Recall that $m_{\mathcal{H}} = n(8\ell + 32\log^2 n)$, so that the mean μ of each N_i is $m_{\mathcal{H}}/n(1+\lambda) \ge 4\ell + 16\log^2 n$. By (3.1), with $\varepsilon = 1/4$, we have

$$\mathbb{P}(N_i \le 3\ell) \le \mathbb{P}(N_i \le \frac{3}{4}\mu) \le e^{-\mu/32} \le e^{-\frac{1}{2}\log^2 n}$$

for each i. Thus the probability that there are fewer than 3ℓ departures from any queue over the interval from $T_{\mathcal{G}^1}$ to $T_{\mathcal{G}^1} + m_{\mathcal{H}}$ is at most $ne^{-\frac{1}{2}\log^2 n} < 1/2s_0$, and part (1) follows.

For part (2), as above we have

$$\mathbb{P}\left(\bigcup_{t=T_{\mathcal{G}^1}+1}^{T_{\mathcal{G}^1}+s_0} (L_t \cap \{X_{t-1} \in \mathcal{G}_1^1\}) \right) \le s_0 e^{-\log^2 n} \le 1/s_0.$$

Thus $\mathbb{P}(T^\dagger_{\mathcal{H}} \le s_0 < T^\dagger_{\mathcal{G}_1})$ is at most the probability that X_t exits the set $\mathcal{H}_1$ before time $T^\dagger_{\mathcal{G}_1} \wedge s_0$, necessarily by the creation of a new queue of length $k+1$, is at most $1/s_0$, as required. □

10 Rapid Mixing

In this section, we prove our results about rapid mixing of the (n, d, λ)-supermarket process. We continue to assume that the functions $\lambda = \lambda(n) = 1 - n^{-\alpha}$, $d = d(n) = n^\beta$ and $\varepsilon = \varepsilon(n)$ of the model satisfy the conditions of Theorem 6.1. We also assume throughout this section that $b \le \ell \le e^{\frac{1}{4}\log^2 n}$, so that $q(\ell, b) = (22k + 72b)\varepsilon^{-1} n^{1+\alpha} + 8\ell n \le s_0/2$.

We say that two queue-lengths vectors are *adjacent* if they differ by one customer in one queue, and we first consider two copies of the process starting in adjacent states in $\mathcal{A}_0(\ell, b)$, coupled according to the coupling referred to in Lemma 4.1. The proof partly follows along the lines of the proof of Lemma 2.6 in [11].

Lemma 10.1 *Let x, y be a pair of adjacent states in $\mathcal{A}_0(\ell, b)$, with $x(j_0) = y(j_0) - 1$ for some queue j_0, and $x(j) = y(j)$ for $j \ne j_0$. Consider coupled copies (X_t^x) and (X_t^y) of the (n, d, λ)-supermarket process, where $X_0^x = x$ and $X_0^y = y$. For n sufficiently large, and all times $t \ge 2q(\ell, b)$, we have*

$$\mathbb{E}\, \|X_t^x - X_t^y\|_1 = \mathbb{P}(X_t^x \ne X_t^y) \le 2e^{-\frac{1}{4}\log^2 n}.$$

Proof By Lemma 4.1, X_t^x and X_t^y are always neighbours or equal, always $X_t^x \le X_t^y$, and if for some time s we have $X_s^x = X_s^y$, then $X_t^x = X_t^y$ for all $t \ge s$. Thus in particular $\mathbb{E}\, \|X_t^x - X_t^y\|_1 = \mathbb{P}(X_t^x \ne X_t^y)$. The probability of coalescence is increasing with t, so we may assume that $t = 2q(\ell, b)$.

Initially, the queue j_0 is unbalanced, i.e., $X_0^x(j_0) \ne X_0^y(j_0)$, and all other queues are balanced. Observe that the index of the unbalanced queue in the coupled pair of processes may change over time. Let W_t denote the longer of the unbalanced queue lengths at time t, if there is such a queue, and let $W_t = 0$ otherwise. The time for the two coupled processes to coalesce is the time T until W_t hits 0.

We first run (X_t^x) and (X_t^y) together using the coupling. Let $T_{\mathcal{H}}^x$ and $T_{\mathcal{H}}^y$ denote the times $T_{\mathcal{H}}$, as defined in Sect. 6, for the two copies of the process, and set $T_{\mathcal{H}}^* = T_{\mathcal{H}}^x \vee T_{\mathcal{H}}^y$. By Theorem 6.10, $T_{\mathcal{H}}^* \le q(\ell, b)$ with probability at least

$$1 - \frac{2(6k + 28)}{s_0} \ge 1 - \frac{1}{3}e^{-\frac{1}{4}\log^2 n}.$$

We now track the performance of the coupling after time $T_{\mathcal{H}}^*$. If the processes have coalesced by time $T_{\mathcal{H}}^*$ (i.e., if $T \le T_{\mathcal{H}}^*$), then we are done. Otherwise, $X_{T_{\mathcal{H}}^*}^x$ and $X_{T_{\mathcal{H}}^*}^y$ are still adjacent, and there is some random index J_0 such that the queue J_0 is unbalanced, i.e.,

$X^x_{T^*_{\mathcal{H}}}(J_0) \neq X^y_{T^*_{\mathcal{H}}}(J_0)$, and all other queues are balanced. Moreover, since $u_{k+1}(x) = 0$ for all $x \in \mathcal{H}$, we have $W_{T^*_{\mathcal{H}}} \leq k$.

We shall use Lemma 3.4 to give a suitable upper bound on $\mathbb{P}(W_t > 0)$. The idea is that, since, with high probability, both copies of the process remain in $\mathcal{H}$ for a long time, the unbalanced queue length W_t will often drop below k, and then there is a chance of going all the way down to 0 before returning to k.

For each $t \geq 0$, let B_t be the event that $X^y_s, X^x_s \in \mathcal{H}$ for all s with $T^*_{\mathcal{H}} \leq s \leq t - 1$. It follows from Theorem 6.10 that $\mathbb{P}(\overline{B}_t) \leq (12k + 56)/s_0 \leq \frac{1}{3}e^{-\frac{1}{4}\log^2 n}$, provided $t \leq s_0$.

Let N_r be the number of jumps of the longer unbalanced queue length in the first r steps after $T^*_{\mathcal{H}}$. Also set $N = N_T$, the total number of these jumps, with $N_T = 0$ if $T \leq T^*_{\mathcal{H}}$. For $j = 1, 2, \ldots$, let T_j be the time of the jth jump after $T^*_{\mathcal{H}}$ if $N \geq j$, and otherwise set $T_j = T^*_{\mathcal{H}} \vee T$. Thus, if $T^*_{\mathcal{H}} < T$, we have $T^*_{\mathcal{H}} < T_1 < \cdots < T = T_N = T_{N+1} = \cdots$. If $T^*_{\mathcal{H}} \geq T$, then all of the T_j are equal to $T^*_{\mathcal{H}}$.

Let $S_0 = y(J_0)\mathbb{1}_{\{T^*_{\mathcal{H}} < T\}} = W_{T^*_{\mathcal{H}}}\mathbb{1}_{\{T^*_{\mathcal{H}} < T\}}$, the longer unbalanced queue length at time $t = T^*_{\mathcal{H}}$ if coalescence has not occurred. For each positive integer j, if $N \geq j$, let $S_j = W_{T_j}$, which is either 0 or the longer of the unbalanced queue lengths at time T_j, immediately after the jth arrival or departure at the unbalanced queue. Also, if $N \geq j$, let Z_j be the ± 1-valued random variable $S_j - S_{j-1}$. For each non-negative integer j, let φ_j be the σ-field $\mathcal{F}_{T_{j+1}-1}$ of all events before time T_{j+1}. Let also A_j be the φ_j-measurable event $B_{T_{j+1}}$, that is the event that $X^y_s, X^x_s \in \mathcal{H}$ for each s with $T^*_{\mathcal{H}} \leq s \leq T_{j+1} - 1$.

We shall use Lemma 3.4. We take the sequences $(\varphi_j)_{j\geq 0}$, $(Z_j)_{j\geq 0}$, $(S_j)_{j\geq 0}$ and $(A_j)_{j\geq 0}$ as defined above, and we set $k_0 = k$ and $\delta = 1/(\lambda d + 1)$. Note first that, at any time $t < T$, the probability, conditioned on $\mathcal{F}_t$, of an arrival to the longer of the unbalanced queues is at most $d\lambda/n(1 + \lambda)$, while the conditional probability of a departure from that queue is $1/n(1 + \lambda)$. Therefore, on the event that $N \geq j$, the probability, conditioned on φ_{j-1}, that the event at time T_j is a departure from the longer unbalanced queue is at least

$$\frac{1/n(1+\lambda)}{1/n(1+\lambda) + d\lambda/n(1+\lambda)} = \frac{1}{1+d\lambda} = \delta.$$

In other words, on the event $N \geq j$ we have $\mathbb{P}(Z_j = -1 \mid \varphi_{j-1}) \geq \delta$.

We now show that, on the event $\{N \geq j\} \cap A_{j-1} \cap \{S_{j-1} \geq k\}$, we have

$$\mathbb{P}(Z_j = -1 \mid \varphi_{j-1}) \geq \frac{3}{4}.$$

To see this, consider a time $t \geq T^*_{\mathcal{H}}$. On the event B_t, we have $X_t \in \mathcal{H} \subseteq \mathcal{E}_1$, and so, by Lemma 8.5, the conditional probability $\mathbb{P}(L_{t+1} \mid \mathcal{F}_t)$ that the event at time $t + 1$ is an arrival to a queue of length k or greater is at most $e^{-\log^2 n}$. In particular, on the event $B_t \cap \{W_{t-1} \geq k\}$, the conditional probability that the event at time $t + 1$ is an arrival joining the longer unbalanced queue is at most $e^{-\log^2 n}$, while the conditional probability that the event at time $t+1$ is a departure from the longer unbalanced queue is $1/n(1+\lambda)$. Therefore, on the event $\{N \geq j\} \cap A_{j-1} \cap \{S_{j-1} \geq k\}$, we have

$$\mathbb{P}(Z_j = -1 \mid \varphi_{j-1}) \geq \frac{1/n(\lambda+1)}{1/n(\lambda+1) + e^{-\log^2 n}} \geq \frac{3}{4}.$$

We have now shown that $S_m - S_0$ can be written as a sum $\sum_{i=1}^m Z_i$ for $\{0, \pm 1\}$-valued random variables Z_i that satisfy the conditions of Lemma 3.4, with $k_0 = k$ and $\delta = 1/(\lambda d + 1)$. (The argument above establishes this for $m \leq N$: for $m > N$, we have set $Z_m = S_m = 0$, which also meets the requirements of the lemma.) Note that $\delta^{-(k-1)} = (\lambda d+1)^{k-1} \le 2d^{k-1}$. Hence, for $m \ge 16k$,

$$\mathbb{P}\left(\bigcap_{i=1}^{m}\{S_i \neq 0\} \cap \bigcap_{i=0}^{m-1} A_i\right) \le \mathbb{P}(S_0 > \lfloor m/16 \rfloor) + 3\exp\left(-\frac{\delta^{k-1}}{200k}m\right)$$
$$\le 0 + 3\exp\left(-\frac{m}{400kd^{k-1}}\right).$$

Here $\mathbb{P}(\cdot)$ refers to the coupling measure in the probability space of Sect. 4, with coupled copies of the process for each possible starting state.

Let $q = q(\ell, b)$ and $m = \lfloor q/4n \rfloor \ge n^{\alpha}$. Since, at each time after $T^*_{\mathcal{H}}$ and before T, a jump in the longer unbalanced queue occurs with probability at least $1/2n$ while the queue is nonempty, we have, by inequality (3.1), $\mathbb{P}(\{T > T^*_{\mathcal{H}} + q\} \cap \{N_q < m\}) \le e^{-q/16n}$. Also,

$$\mathbb{P}\left(\{N_r \ge m\} \cap \bigcup_{i=0}^{m-1} \overline{A_i} \cap \{T^*_{\mathcal{H}} \le q\}\right) \le \mathbb{P}(\overline{B_{2q}}) \le \mathbb{P}(\overline{B_{s_0}}) \le \frac{1}{3}e^{-\frac{1}{4}\log^2 n}.$$

Now we have that

$$\mathbb{P}(T > 2q) \le \mathbb{P}(T^*_{\mathcal{H}} > q) + \mathbb{P}(\{T > T^*_{\mathcal{H}} + q\} \cap \{T^*_{\mathcal{H}} \le q\})$$
$$\le \mathbb{P}(T^*_{\mathcal{H}} > q) + \mathbb{P}(\{T > T^*_{\mathcal{H}} + q\} \cap \{N_q < m\})$$
$$+ \mathbb{P}\left(\{N_q \ge m\} \cap \bigcup_{i=0}^{m-1} \overline{A_i} \cap \{T^*_{\mathcal{H}} \le q\}\right)$$
$$+ \mathbb{P}\left(\{N_q \ge m\} \cap \bigcap_{i=0}^{m-1} A_i \cap \bigcap_{i=1}^{m}\{S_i \neq 0\}\right).$$

To see this, note that $\{N_q \ge m\} \cap \bigcup_{i=1}^{m}\{S_i = 0\} \subseteq \{T \le T^*_{\mathcal{H}} + q\}$. Now we have

$$\mathbb{P}(T > 2q) \le \frac{1}{3}e^{-\frac{1}{4}\log^2 n} + e^{-q/16n} + \frac{1}{3}e^{-\frac{1}{4}\log^2 n}$$
$$+3\exp\left(-\frac{q}{1600kd^{k-1}n}\right)$$
$$\le \frac{2}{3}e^{-\frac{1}{4}\log^2 n} + 4\exp\left(-\frac{n^{\alpha-(k-1)\beta}}{1600k}\right)$$
$$\le e^{-\frac{1}{4}\log^2 n},$$

as required. □

Theorem 10.2 *Let (X_t^x) and (X_t^y) be two copies of the (n, d, λ)-supermarket process, starting in states x and y in $\mathcal{A}_0(\ell, b)$. Then, for n sufficiently large and $t \ge 2q(\ell, b)$, we have*

$$\mathbb{E}\,\|X_t^x - X_t^y\|_1 \le 2bne^{-\frac{1}{4}\log^2 n} \le e^{-\frac{1}{5}\log^2 n}.$$

Proof Given two distinct states x and y in $\mathcal{A}_0(\ell, b)$, we can choose a path $x = z_0, z_1, \dots, z_m = y$ of adjacent states in $\mathcal{A}_0(\ell, b)$ from x down to the empty queue-lengths vector and back up to y, where $m = \|x\|_1 + \|y\|_1 \le 2bn$. By Lemma 10.1, for $t \ge 2q(\ell, b)$,

$$\mathbb{E}\,\|X_t^x - X_t^y\|_1 \le \sum_{i=0}^{m-1} \mathbb{E}\,\|X_t^{z_i} - X_t^{z_{i+1}}\|_1 \le 2bne^{-\frac{1}{4}\log^2 n},$$

as required. □

We saw in Corollary 6.9 that $Y_t \in \mathcal{A}_0(\ell, b)$ with probability at least $1-e^{-\frac{1}{4}\log^2 n}$, whenever $\ell, b \ge k$, where (Y_t) is a copy of the (n, d, λ)-supermarket process in equilibrium. Thus we have the following corollary.

Corollary 10.3 *Take any $\ell, b \ge k$, and let (X_t^x) be a copy of the (n, d, λ)-supermarket process starting in a state $x \in \mathcal{A}_0(\ell, b)$. Also let (Y_t) be a copy in equilibrium. Then, for n sufficiently large and $t \ge 2q(\ell, b)$, we have*

$$d_{TV}(\mathcal{L}(X_t^x), \mathcal{L}(Y_t)) \le 2e^{-\frac{1}{5}\log^2 n}.$$

This now implies Theorem 1.4. We choose $\varepsilon(n) = 1/100$. The hypothesis that $\ell = \|x\|_\infty \le e^{\frac{1}{4}\log^2 n}$, together with $b = \|x\|_1/n \le \ell$, ensures that $q(\ell, b) \le \frac{1}{2}s_0$, and the setting of ε ensures that $q(\ell, b) \le 7200\big(kn^{1+\alpha} + bn^{1+\alpha} + \ell n\big)$.

We now show that mixing actually takes place faster if we start from a "good" state, i.e., a state in $\mathcal{N} = \mathcal{N}^\varepsilon$.

Lemma 10.4 *Let x, y be a pair of adjacent states in $\mathcal{N}^\varepsilon$, with $x(j_0) = y(j_0) - 1$ for some queue j_0, and $x(j) = y(j)$ for $j \ne j_0$. Consider coupled copies (X_t^x) and (X_t^y) of the (n, d, λ)-supermarket process. For n sufficiently large and all times $t \ge 0$, we have*

$$\mathbb{E}\,\|X_t^x - X_t^y\|_1 \le e^{-\frac{1}{4}\log^2 n} + 4\exp\left(-\frac{t}{1600kd^{k-1}n}\right).$$

Proof The proof is nearly identical to that of Lemma 10.1. Instead of starting by running the two copies of the process together until some time T^*, we use the final part of Theorem 6.1, which tells us that, with probability at most $1 - e^{-\frac{1}{4}\log^2 n}$, both X_t^x and X_t^y remain within $\mathcal{N}^{6\varepsilon}$ throughout the interval $0 \le t \le s_0$. We may thus repeat the proof of Lemma 10.1 with T^* and $q = q(\ell, b)$ replaced by 0, and running the second phase for any number t of steps instead of q, and we obtain the result. □

As before, we can use this result to deduce an upper bound on the mixing time, starting from a good state.

Theorem 10.5 *Let (X_t^x) and (X_t^y) be two copies of the (n, d, λ)-supermarket process with starting states x and y in $\mathcal{N}^\varepsilon$. Then, for n sufficiently large and $t \ge 0$, we have*

$$\mathbb{E}\,\|X_t^x - X_t^y\|_1 \le n\left(e^{-\frac{1}{4}\log^2 n} + 4\exp\left(-\frac{t}{1600kd^{k-1}n}\right)\right).$$

Proof (Sketch) Take any two queue-lengths vectors x and y in $\mathcal{N}^\varepsilon$. It is straightforward to show that there is a path between $x = z_0 z_1 \cdots z_m = y$ in $\mathcal{N}^\varepsilon$ of length $m \le 4n(1-\lambda)(\lambda d)^{k-1} \le n$ between x and y. The result now follows as in the proof of Theorem 10.2. □

As before, since Y_0 lies in $\mathcal{H}^\varepsilon \subseteq \mathcal{N}^\varepsilon$ with probability at least $1-e^{-\frac{1}{4}\log^2 n}$, by Corollary 6.9, we may now deduce that the total variation distance $d_{TV}(\mathcal{L}(X_t^x), \mathcal{L}(Y_t))$ is at most

$$
\begin{aligned}
& e^{-\frac{1}{4}\log^2 n} + n\left(e^{-\frac{1}{4}\log^2 n} + 4\exp\left(-\frac{t}{1600kd^{k-1}n}\right)\right) \\
& \le n\left(2e^{-\frac{1}{4}\log^2 n} + 4\exp\left(-\frac{t}{1600kn^{1+(k-1)\beta}}\right)\right)
\end{aligned}
$$

whenever $x \in \mathcal{N}^\varepsilon$. This result is exactly the statement of Theorem 1.2 (where we take $\varepsilon = 1/\log n$: the result would hold if our initial state were in $\mathcal{N}^\varepsilon$ for ε a suitably small constant).

Theorem 1.2 shows that, from states $x \in \mathcal{N}$, we have mixing to equilibrium in time of order $n^{1+(k-1)\beta}\log n$. We finish by proving Theorem 1.3, showing that this bound is approximately best possible.

Note that there is a state z in $\mathcal{I}^\varepsilon \subseteq \mathcal{H}^\varepsilon \subseteq \mathcal{N}^\varepsilon$ with $Q_k(z) \le (1-3\varepsilon)n(1-\lambda)(\lambda d)^{k-1}$. However, we know from Corollary 6.9 that $\mathbb{P}(Y_t \in \mathcal{H}^{\varepsilon/5}) \ge 1 - e^{-\frac{1}{4}\log^2 n}$, so in order for $d_{TV}(\mathcal{L}(X_t^z), \Pi)$ to be small, we need that $Q_k(X_t^z) \ge (1-\varepsilon)n(1-\lambda)(\lambda d)^{k-1}$ with high probability. Set $t = \frac{1}{6}n(\lambda d)^{k-1}$.

For $x \in \mathcal{H}^\varepsilon$, we obtain from Lemma 7.1, with a calculation almost exactly as in Lemma 8.2, that

$$
\begin{aligned}
(1+\lambda)\Delta Q_k(x) &\le (1-\lambda)(1+\varepsilon/6) - \frac{Q_k(x)}{n(\lambda d)^{k-1}} + \exp\left(-dQ_k(x)/kn\right) \\
&\le (1-\lambda)(1+\varepsilon/3-(1-4\varepsilon)) \le 5\varepsilon(1-\lambda),
\end{aligned}
$$

so $\Delta Q_k(x) \le 5\varepsilon(1-\lambda)$ also. We know from (6.1) that, with probability at least $1-e^{-\frac{1}{4}\log^2 n}$, $X_s^z \in \mathcal{H}^\varepsilon$ for all $s = 0, \dots, t-1$, and we also have that $Q_k(x) \le kn$ for every state x. It follows that

$$
\begin{aligned}
\mathbb{E}\, Q_k(X_t^z) &= Q_k(z) + \sum_{s=0}^{t-1} \mathbb{E}\left(\mathbb{E}(\Delta Q_k(X_s^z) \mid \mathcal{F}_s)\right) \\
&\le (1-3\varepsilon)n(1-\lambda)(\lambda d)^{k-1} + 5\varepsilon t(1-\lambda) + kne^{-\frac{1}{4}\log^2 n} \\
&\le (1-2\varepsilon)n(1-\lambda)(\lambda d)^{k-1}.
\end{aligned}
$$

A result from [11] (adapted for discrete time) states that, for some absolute constant c, for any 1-Lipschitz function f, any starting state z, any $t > 0$ and any $u \ge 0$,

$$
\mathbb{P}(|f(X_t^z) - \mathbb{E}\, f(X_t^z)| \ge u) \le ne^{-cu^2/(t+u)}.
$$

Applying this with $f = Q_k$, $t = \frac{1}{6}n(\lambda d)^{k-1}$ and $u = \varepsilon t(1-\lambda)$, we find that

$$
\begin{aligned}
\mathbb{P}(Q_k(X_t^z) > (1-\varepsilon)n(1-\lambda)(\lambda d)^{k-1}) &\le \mathbb{P}(Q_k(X_t^z) - \mathbb{E}\, Q_k(X_t^z) > \varepsilon n(1-\lambda)(\lambda d)^{k-1}) \\
&\le ne^{-c\varepsilon^2 n(1-\lambda)^2(\lambda d)^{k-1}/2} \le 1/s_0.
\end{aligned}
$$

This completes the proof of Theorem 1.3.

References

1. Alanyali, M., Dashouk, M.: Occupancy distributions of homogeneous queueing systems under opportunistic scheduling. IEEE Trans. Inf. Theor. **57**, 256–266 (2011)
2. Brightwell, G., Luczak, M.: Vertices of high degree in the preferential attachment tree. Electron. J. Probab. **17**(14), 1–43 (2012)
3. Brightwell, G., Luczak, M.: The supermarket model with arrival rate tending to one (2012). arXiv:1201.5523
4. Eschenfeldt, P., Gamarnik, D.: Join the shortest queue with many servers. The heavy traffic asymptotics (2015). arXiv:1502.00999
5. Eschenfeldt, P., Gamarnik, D.: Supermarket queueing system in the heavy traffic regime. Short queue dynamics (2016). arXiv:1610.03522
6. Fairthorne, M.: PhD Thesis, London School of Economics (2011)
7. Graham, C.: Chaoticity on path space for a queuing network with selection of the shortest queue among several. J. Appl. Probab. **37**, 198–201 (2000)
8. Graham, C.: Functional central limit theorems for a large network in which customers join the shortest of several queues. Prob. Theor. Relat. Fields **131**, 97–120 (2004)
9. Luczak, M.J.: Concentration of measure and mixing of Markov chains. Discret. Math. Theor. Comp. Sci. (Proc. 5th Colloq. Mathem. Comp. Sci.) 95–120 (2008)
10. Luczak, M.J., McDiarmid, C.: On the power of two choices: balls and bins in continuous time. Ann. Appl. Probab. **15**, 1733–1764 (2006)
11. Luczak, M.J., McDiarmid, C.: On the maximum queue length in the supermarket model. Ann. Probab. **34**, 493–527 (2006)
12. Luczak, M.J., McDiarmid, C.: Asymptotic distributions and chaos for the supermarket model. Electron. J. Probab. **12**, 75–99 (2007)
13. Luczak, M.J., Norris, J.R.: Strong approximation for the supermarket model. Ann. Appl. Probab. **15**, 2038–2061 (2005)
14. Luczak, M.J., Norris, J.R.: Averaging over fast variables in the fluid limit for Markov chains: application to the supermarket model with memory. Ann. Appl. Probab. **23**, 957–986 (2013)
15. Martin, J.B., Suhov, Y.M.: Fast Jackson networks. Ann. Appl. Probab. **9**, 854–870 (1999)
16. Meyer, C.D.: Matrix Analysis and Applied Linear Algebra. SIAM, Philadelphia (2000)
17. Mitzenmacher, M.: Load balancing and density dependent jump Markov processes. In: Proceedings of 37th IEEE Annual Symposium on Foundations of Computer Science (FOCS), pp. 213–222 (1996)
18. Mitzenmacher, M.: The power of two choices in randomized load-balancing. PhD thesis, Berkeley. http://www.eec.harvard.edu michaelm/ (1996)
19. Mitzenmacher, M., Richa, A., Sitaraman, R.: The power of two random choices: a survey of techniques and results. In: Pardalos, P., Rajasekaran, S., Rolim, J. (eds.) Handbook of Randomized Computing: Volume 1, pp. 255–312 (2001)
20. Mitzenmacher, M., Prabhakar, B., Shah, D.: Load-balancing with memory. In: Proceedings of 43rd IEEE Annual Symposium on Foundations of Computer Science (FOCS), pp. 799–808 (2002)
21. Mukherjee, D., Borst, S.C., van Leeuwaarden, J.S.H., Whiting, P.A.: Universality of power-of-d load balancing in many server systems (2016). arXiv:1612.00723v1
22. Turner, S.R.E.: The effect of increasing routing choice on resource pooling. Probab. Eng. Inf. Sci. **12**, 109–124 (1998)
23. Vvedenskaya, N.D., Dobrushin, R.L., Karpelevich, F.I.: Queueing system with selection of the shortest of two queues: an asymptotic approach. Prob. Inf. Transm. **32**, 15–27 (1996)